Unmanned Systems, Vol. 10, No. 4 (2022) 1–16

DOI: 10.1142/S2301385023500164

UNMANNED SYSTEMS
https://www.worldscientific.com/worldscinet/us

Multi-Model Fusion of Encoding Methods-Based Visual Words for Target Recognition in Infrared Images

Billel Nebili*, Atmane Khellal*,‡, Abdelkrim Nemra*, Laurent Mascarilla†

*Ecole Militaire Polytechnique,
UER SAI, Algiers 16111, Algeria

†Laboratoire MIA, Univ. La Rochelle,
Avenue Michel Crépeau, F-17042 La Rochelle Cedex, France

The limited information contained in infrared images present a serious problem, therefore it is necessary to form a powerful feature descriptor that allows extracting the maximum information and describing the image efficiently. To address this challenge, we propose a novel approach named multi-model fusion of encoding methods (MMFEM). First, several encoding methods for Bag Of Visual Words (BOVW) model were evaluated. Then, we fuse the best encoding methods obtained using three levels of fusion: feature-level fusion, decision-level fusion and hybrid-level fusion. Finally, the outputs of the fusion process were used to form a final decision for target recognition in infrared images. Two infrared datasets were employed to evaluate the performance of the proposed approach. The first one is Visible and Infrared Spectrum (VAIS) dataset comprising six categories of ships and the second dataset is a subset of Forward-Looking InfraRed (FLIR) thermal dataset comprising two object categories, vehicles and pedestrians. The proposed approach has exceed the state of the art for both datasets and we have reached 96.96% for FLIR and 71.26% for VAIS in overall classification accuracy.

Keywords: Multi-model fusion; SIFT encoding; bag of visual words; infrared images; VAIS; FLIR.

US

1. Introduction

Recently, the number of applications using target recognition such as human identification [1], maritime navigation [2], agriculture control [3] and driving assistance [4] increased significantly. The development of the artificial intelligence in computer vision field and the performance improvement of the algorithms proposed by the scientific community led to more and more efficient target recognition solutions. The technological developments related to imagery in various spectral bands have also greatly contributed to the development of target recognition, in particular, thermal technology. Most existing target recognition systems generally use the visible electromagnetic spectrum. These images may be sufficient for some applications in controlled environments, but the effectiveness of these systems is limited in uncontrolled ones [5]. The use of thermal sensors can reduce the loss of performance in difficult situations (darkness, low light, etc.) because it works day and night [6]. However, infrared images have specific characteristics that differ from visible images: low signal-to-noise ratio, low resolution, low contrast ratio between targets and background, competitive background clutter, lack of sharp edges and target boundary information are not usually very visible [7]. Therefore, target recognition in infrared imagery is a challenging problem.

Typically, target recognition has two stage process: features extraction and classification. For the first stage, there are two categories of feature extractors: (1) traditional hand-crafted feature extractors such as Scale Invariant Feature Transform (SIFT) [8] and Speeded Up Robust Features (SURF) [9], (2) features learning generally based on Convolutional Neural Networks (CNNs) [10–12]. The former has provided good results for target recognition

Received 14 December 2021; Accepted 14 August 2022; Published xx xx xx.
This paper was recommended for publication in its revised form by editorial board member, Zhi Gao.
Email Addresses: ‡*khe.atmane@gmail.com, atmane.khellal@emp.mdn.dz*

problems in visible images, which has motivated their use for thermal images. For example, Jungling and Arens [13] studied a local feature-based human detector on a thermal dataset where they have used SURF for features extraction.

According to the application field, two types of features can be distinguished: global and local. (1) Global features are based on properties such as color, texture or shape over the entire image. This information can be extracted from the whole image using approaches like: color histogram [14] for color, Gabor filter [15] for texture, etc. (2) Local features are calculated around points or regions of interest. Different types of description algorithms can be used to extract and describe these points/regions: SIFT, SURF, etc. One method of exploiting these local descriptors is the Bag Of Visual Words (BOVW) model also called visual vocabulary [16]. The number of interest points is different for each image, although some classification methods require vectors of fixed size as input. To obtain a vector of fixed dimension, an encoding step allows creating it. This vector is based on the quantification of local descriptors in visual vocabulary.

To improve the encoding step, several suggestions have been proposed. (1) Instead of quantifying local descriptors in a single cluster (hard assignment) [16], the local descriptor can be assigned to several clusters "soft assignment" (e.g. kernel codebook [17], locality-constrained linear coding [18]). (2) By quantifying the difference between the features and the visual words (e.g. fisher kernel [19], vector of locally aggregated descriptors [20]). (3) The integration of local descriptors positions in images into the BOVW model using a spatial pyramidal matching (SPM) technique [21].

However, these suggestions are still insufficient to reach the desired performance, because a single descriptor may not be comprehensive enough to represent an image, so some researchers have further explored descriptors fusion.

In this context, this work aims to propose a solution that increases the reliability and improves the accuracy of target recognition systems in infrared images. Our solution proposes a combination between several global descriptors using the following fusion strategies: feature-level fusion, decision-level fusion and hybrid-level fusion. Descriptors are generated by the different encoding methods. We have used five encoding methods: Histogram Encoding (HE), Kernel CodeBook (KCB) encoding, Locality-constrained Linear Coding (LLC), Fisher Kernel (FK) encoding and Vector of Locally Aggregated Descriptors (VLAD) encoding. This work will be divided in two parts: (1) Evaluation of each encoding method separately to derive the best ones. (2) Fusion the encoding models to boost the classification performances for infrared images. The experiments are carried out on two public datasets: Visible and Infrared Spectrum (VAIS) dataset [22] and FLIR thermal starter dataset [23].

The contributions offered by this paper can be summarized in the following points:

- We proposed a novel multi-level-based approach for automatic target recognition; which integrates both the feature-level and decision-level fusions. Each component of the model is well designed and selected based on the best performance on cross validation set. In the test phase, the measure of generalization performance shows the superiority of our approach compared with the state of the art, according to the best of our knowledge, in both FLIR and VAIS datasets.
- We have proposed a multi-model approach based on multiple feature encoding, where each feature vector is derived from a different feature encoding method. Each method encodes specific properties in the image. That, allows us to take full advantage of the complementary nature between different encoding features and makes our approach more discriminative.
- As supported by the experimental results, the proposed approach is well suited for infrared datasets, which are characterized by small training sets and low spatial resolution. In such situations, even the state-of-the-art CNNs fail to generalize well due to the problem of overfitting.
- According to the best of our knowledge, this paper presents the first comprehensive analysis and exhaustive comparisons of five different encoding methods (HE, KCB, LLC, VLAD, FK) for target recognition in infrared images.

This paper is arranged as follows. The related work is described in Sec. 2. The encoding methods are explained in Sec. 3. In Sec. 4, we will explain the proposed approach (multi-model fusion of encoding methods (MMFEM)). Then, Sec. 5 describes the experiments and results. Section 6 concludes.

2. Related Work

As mentioned above, infrared images contain limited information and they are characterized by small training sets and low spatial resolution, so it is necessary to design a powerful feature representation that allows describing images efficiently. Our work is motivated by many works that are realized to improve the quality of descriptors and consequently improve the performance of target recognition systems in infrared images. Several proposed solutions adopt the encoding of features, data fusion strategies or both. In this section, we will present the main efforts that have been investigated to resolve this problem.

The BOVW model is the core of feature encoding methods, it has been widely applied in infrared images classification. Malpani *et al.* [24] and Akula *et al.* [25] evaluated a

BOVW framework based on SURF descriptor for human and vehicle classification in infrared images, respectively. In [26], an optimized BOVW framework has been designed for object recognition in IR images and compared with the best feature extractors in the BOVW model. Akula *et al.* [7] considered a framework-based BOVW and WingerMSER detector to resolve the problems caused by the specific characteristics of IR images. All these approaches have used hard assignment to encode local features. However, in order to maintain more information about local features many approaches based on soft assignment are proposed. In [27], Zhao *et al.* integrated VLAD encoding into the feature extraction architecture of insulators detection in infrared images. Vadivelu *et al.* [28] developed an approach for human action recognition in thermal images, based on the FK encoding.

The major limitation of these approaches is the use of single feature vector, which makes them less discriminative. Furthermore, there is no doubt that the transition from a uni-descriptor system to multi-descriptor system increases its reliability and robustness. Therefore, several studies have proposed multi-model frameworks based on levels fusion strategies to improve the accuracy of target recognition systems in infrared images.

A novel multi-feature structure fusion using the spectral regression discriminant analysis (SF-SRDA) has been proposed in [29]. To get high performance, Zhang *et al.* [29] used a high dimension feature vector, which implied higher consumption of memory resources and longer time for training and testing. Zhang *et al.* [22] combined the probabilistic outputs of Gnostic Field and VGGNet to improve recognition scores for target classification, however poor performances have been obtained in VAIS dataset. Moreover, Santos and Bhanu [30] proposed architectures that consist of a decision-level fusion of convolutional networks using a probabilistic model for the final classification part. In [31], features fusion architecture called hybrid fusion has been evaluated. In this case, features were extracted from IR and visible images using two CNN branches to produce a multi-spectral feature vector, which is finally fed into a classifier to achieve ship recognition task.

Shi *et al.* [32] combined low-level features obtained by Gabor filters and MS-CLBP with high-level features obtained by deep CNN. Huang *et al.* [33] presented a classification framework containing two fusion strategies: feature-level and decision-level. The two approaches proposed in [32, 33] have included global features in their fusion architectures. However, global features have a few limitations face illumination variation, scaling, sensitivity to noise and failure to identify the important features of the image. So, they are not suitable for some applications as infrared image classification.

As can be seen from the related work, several proposed approaches are based on a single feature vector and single fusion level which makes the classification performance very limited especially with infrared images. Furthermore, few works exploit the hybrid fusion for target recognition in infrared images. Most of them use features fusion and decisions fusion.

Inspired by these observations, a multi-model fusion framework for target recognition in infrared images is proposed in this work, where the three fusion strategies are employed. First, we have used SIFT to extract local features, which have achieved satisfying performances in image classification. Then, various encoding methods have been used to obtain the global feature representation of images. Each method encodes specific properties in the image. KCB and LLC encodings represent features as combination of visual words. FK and VLAD encodings represent differences between features and visual words. This allows us to take full advantage of the complementary nature between different encoding features and makes our approach more discriminative. We further combine encoding methods and the SPM technique to enhance the spatial order structure of local features. After that, three fusion strategies, feature-level fusion, decision-level fusion and hybrid-level fusion are investigated for final classification. We have compared and evaluated the three fusion strategies by combining them with the best performing encoding methods. According to the best of our knowledge, this paper presents the first comprehensive analysis and exhaustive comparisons of five different encoding methods (HE, KCB, LLC, VLAD, FK) for target recognition in infrared images.

Figure 1 shows an overview of the MMFEM approach. It consists on three components: encoding, fusion and decision. Each input image is encoded L times by one of the encoding methods. Thereby, the encoding step will produce L vectors for each input. Next, a fusion strategy is applied to these vectors, where the output of this process is used in the decision component to predict the class of the input image.

3. Encoding Component

The first component of the proposed approach is encoding. The diagram in Fig. 2 shows an overview of the encoding process. It consists of features extraction, quantization,

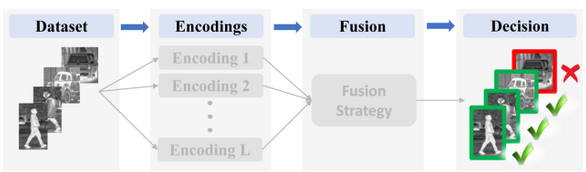

Fig. 1. Overview of the MMFEM approach.

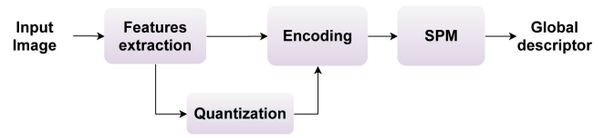

Fig. 2. Overview of the encoding component.

encoding and SPM. First, a dense SIFT is applied to extract features from the input image. Second, a visual vocabulary is build by a clustering algorithm. Then, the extracted local features will be encoded based on the visual vocabulary and with respect to SPM, to obtain at the end a global descriptor of the input image.

3.1. *Features extraction*

To build a global descriptor, a large set of local descriptors are counted and encoded by one of the encoding methods [34]. A simple dense version of SIFT (Pyramid Histogram Of visual Words (PHOW)) is used to extract features of key points to form these local descriptors. PHOW is applied with spatial steps at several resolutions, which are controlled by the width (in pixels) of the spatial bins. In our work, the step is two and the width of spatial bins is defined as 4, 6, 8 and 10 pixels.

3.2. *Quantization*

After the features extraction step, each image is abstracted by several local descriptors. To finally have a global descriptor for each image from these local descriptors and to construct this global descriptor, we need to proceed to the quantization step.

The quantization is based essentially on the categorization of local descriptors into sets called "bag of visuals words". In this section, the "bag of visual words" will be explained and also the methods that allow building it will be described.

3.2.1. *Bag of visual words*

BOVW model consists on describing an image as a set of important key points, obtained by one of clustering algorithms like K-means, Gaussian Mixture Models (GMM), etc. These points represent the centers of the clusters and are called visual words. With a collection of visuals words, we build a visual vocabulary or a codebook [16, 35].

3.2.2. ***K**-means*

One of the most used methods to build the visual codebook is the K-means clustering method [36] because it allows setting the desired codebook size.

3.2.3. *Gaussian mixture models*

The second proposed method of clustering is GMM [37], which has some advantages over K-means. Unlike K-means, GMM takes into account variance, which allows handling very oblong clusters. Moreover, GMM performs soft assignment, while K-means effects hard assignment.

GMM is a weighted sum of K components Gaussian densities as given by

$$p(x \mid \theta) = \sum_{k=1}^{K} p(x \mid \mu_k, \Sigma_k)\pi_k, \quad x \in R^D, \tag{1}$$

where

$$p(x \mid \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^D \det \Sigma_k}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

and $\theta = \{\mu_k, \pi_k, \Sigma_k, k = 1, \ldots, K\}$ is the vector of parameters of the model which are learned by the expectation maximization (EM) algorithm [38] from a training set of descriptors $x_1, x_2, \ldots, x_N$; the means $\mu_k \in R^D$, the mixture weight $\pi_k \in R_+$ and Σ_k the covariance matrix of Gaussian density $p(x \mid \mu_k, \Sigma_k)$. The soft assignments of x_i to K clusters are defined by

$$q_{ki} = \frac{p(x_i|\mu_k, \Sigma_k)\pi_k}{\sum_{j=1}^{K} p(x_i|\mu_j, \Sigma_j)\pi_j}, \quad k = 1, \ldots, K. \tag{2}$$

3.3. *Encoding*

In this section, we briefly survey some popular methods that produce a global representation of an image based on a set of local descriptors and a visual vocabulary. From now, let $X = [x_1, x_2, \ldots, x_N] \in R^{D \times N}$ a set of D-dimensional local descriptors extracted from an image and $U = [\mu_1, \mu_2, \ldots, \mu_K] \in R^{D \times K}$ a codebook with K visual words obtained by GMM for FK encoding and by K-means for other encodings.

3.3.1. *Histogram encoding*

Histogram encoding [16] is the basic encoding approach for an image. Its main idea is to quantify the frequency of local descriptors in a histogram by assigning each local descriptor to the closest visual word. q_i will be the assignment of x_i computed by K-means. The HE will be given by the vector $f_{\text{hist}} \in R^K$ so that

$$[f_{\text{hist}}]_k = |\{i : q_i = k\}|, \quad k = 1, \ldots, K. \tag{3}$$

3.3.2. *Kernel CodeBook encoding*

HE has two main problems: uncertainty and plausibility of codebook. Van Gemert *et al.* [17] proposed two model solutions by using the kernel density estimation [39]. In this work, the model "codeword uncertainty" is used, where the quantization is performed in a soft manner. More specifically, each descriptor x_i is assigned to several visual words with a probability according to their proximity distance. This probability denoted $\bar{a}_k(x_i)$ is the soft assignment of x_i to the cluster μ_k, and it is given by

$$\bar{a}_k(x_i) = \frac{\exp(-\frac{\|x_i-\mu_k\|}{2\sigma^2})}{\sum_{j=1}^{K}\exp(-\frac{\|x_i-\mu_j\|}{2\sigma^2})}, \tag{4}$$

where σ is a smoothing parameter. Therefore, the KCB vector encoding $f_{\mathrm{KCB}} \in R^K$ is computed by

$$[f_{\mathrm{KCB}}]_k = \frac{1}{N}\sum_{i=1}^{N}\bar{a}_k(x_i), \quad k = 1,\ldots,K. \tag{5}$$

3.3.3. *Locality-constrained linear encoding*

For LLC, each descriptor x_i is projected into its local-coordinate system using the locality constraint. The local system is defined by M visual words μ_k closer to x_i, let $b_1,\ldots,b_M$ their indices and denote them as $B = [\mu_{b_1}, \mu_{b_2},\ldots, \mu_{b_M}] \in R^{D\times M}$. LLC encoding uses the following criteria:

$$\min_{C}\sum_{i=1}^{N}\|x_i - Bc_i\|^2 + \beta\|c_i\|^2 \tag{6}$$

s.t. $1^T c_i = 1, \forall\, i$,
where β is a small regularization constant and $C = [c_1, c_2, \ldots, c_N] \in R^{M\times N}$ is the set of codes for X. The LLC vector encoding $f_{\mathrm{LLC}} \in R^K$ will be given by

$$[f_{\mathrm{LLC}}]_k = \begin{cases} \max\limits_{i=1,\ldots,N} c_i(k) & k \le M, \\ 0 & \text{else}, \end{cases} \tag{7}$$

where $k = 1,\ldots,K$.

3.3.4. *Fisher Kernel encoding*

The FK encodes additional distribution information about the descriptors. In the HE, the computation of frequency for each visual word encodes the 0-order statistics of the descriptors distribution. However, the FK allows having high-order encoding statistics (first order and second order). The FK is based on the GMM model parameters [40]. Given $\theta = \{\mu_k, \pi_k, \Sigma_k, k = 1,\ldots,K\}$ the parameters learned from the training of GMM in X descriptor. The covariance matrices Σ_k is assumed to be diagonal and is denoted by σ_k^2. Let q_{ki} (Eq. (2)) be the soft assignment of x_i to the Gaussian parameters π_k, μ_k, Σ_k.

The FK vector encoding $f_{\mathrm{FK}} \in R^{2DK}$ will be given by

$$f_{\mathrm{FK}} = [u_1^T, v_1^T, \ldots, u_k^T, v_k^T, \ldots, u_K^T, v_K^T]^T, \tag{8}$$

where

$$u_k = \frac{1}{N\sqrt{\pi_k}}\sum_{i=1}^{N} q_{ki}\frac{(x_i-\mu_k)}{\sigma_k},$$

$$v_k = \frac{1}{N\sqrt{2\pi_k}}\sum_{i=1}^{N} q_{ki}\left[\frac{(x_i-\mu_k)(x_i-\mu_k)}{\sigma_k^2} - I_D\right],$$

where the multiplication between vectors is as a term-by-term operation and $I_D = [1,1,\ldots,1]^T \in R^D$.

The FK output vector is normalized in two steps. First, each component is independently power normalized using the following formula:

$$f(z) = \mathrm{sign}(z)|z|^{\alpha} \tag{9}$$

with $0 \le \alpha \le 1$. Then, the power normalized FK vector is $L2$-normalized.

3.3.5. *Vector of locally aggregated descriptors*

The VLAD encoding can be viewed as a simplification of the FK encoding. As the previous encodings, a visual codebook U is learned by K-means. After that, each descriptor x_i is assigned to μ_k in a hard manner (HE) or a soft manner (KCB). For VLAD encoding, the idea is to accumulate the difference $x_j - \mu_k$ of the vectors x_j assigned to μ_k. As the FK, the power- and $L2$-normalization can be applied. VLAD gives a KD size encoding. Table 1 shows the VLAD algorithm.

3.4. *Spatial pyramid matching*

In image representation, the spatial relationships between patches are very important. These relationships are overlooked by the BOVW model, hence it will be unable to capture shapes or locate objects. SPM allows integrating this lost information during codebook building. The idea of SPM is to partition an image into increasingly fine spatial sub-regions and computing the local features encoding for each sub-region. In our work, the images are divided into 1×1, 2×2 and 3×1 grids (like [34]), so 8 regions in total (Fig. 3). The features of each region are normalized individually using the $l1$ or $l2$ norm ($l1$ for HE and KCB, $l2$ for LLC, VLAD and FK, as suggested in the original publications).

To improve learning performance, an efficient features mapping is computed, using "additives homogeneous kernels" [42]. In this work, we consider Hellinger's kernel

which is defined by: $[\Psi(f)]_i = \sqrt{f_i}$. After kernel mapping, the final entire vector encoding is $l2$-normalized.

The details of the encoding component are illustrated in Fig. 3. A densely sampled SIFT (PHOW) is applied to the whole image. Then, a visual vocabulary is constructed with the extracted features, where we use K-means or GMM. The input image is partitioned into disjoint regions (1×1, 2×2 and 3×1, 8 regions in total). In each region, the extracted features are encoded by one of the encoding methods listed above, so we obtain 8 encoding vectors. The entire vector encoding is obtained by concatenating the 8 vectors to form a long feature vector. Finally, the obtained feature vector is mapped and $L2$-normalized.

Table 1. Computation of the VLAD descriptor [17, 41].

Algorithm VLAD encoding	
For $k = 1, \ldots, K$	
$v_k := 0_D$	
% Hard assignment	% Soft assignment
For $j = 1, \ldots, N$	For $k = 1, \ldots, K$
$k = \arg\min_t \|x_j - \mu_t\|$	$v_k = \sum_{i=1}^{N} \bar{a}_k(x_i)(x_i - \mu_k)$
$v_k := v_k + x_j - \mu_k$	where $\bar{a}_k(x_i)$ is the soft assignment of x_i to μ_k (Eq. (4))
$V = [v_1^T, \ldots, v_K^T]$	
% apply power normalization	
For $t = 1, \ldots, KD$	
$V_t := \text{sign}(V_t)\|V_t\|^{\alpha}$	
% apply $L2$-normalization	
$V := \frac{V}{\|V\|_2}$	

4. Multi-Model Fusion and Decision

4.1. *Multi-model fusion component*

The second component of our proposed approach is fusion. The target recognition task can be improved by fusing information extracted from different models. For this reason, there are three common fusion strategies in the literature, namely, feature-level fusion, decision-level fusion and model-level or hybrid-level fusion [43].

4.1.1. *Feature-level fusion*

Considering that a single feature vector may not be sufficiently complete to represent an image, we have thought of fusing two feature vectors. Each vector is derived from a different encoding with different properties, resulting in a final vector that represents better the information contained in the image.

Figure 4 illustrates the features fusion process of two different encoding methods. Two feature vectors X_1 and X_2 are concatenated to get a single feature vector X. This resulting vector is fed to an SVM classifier.

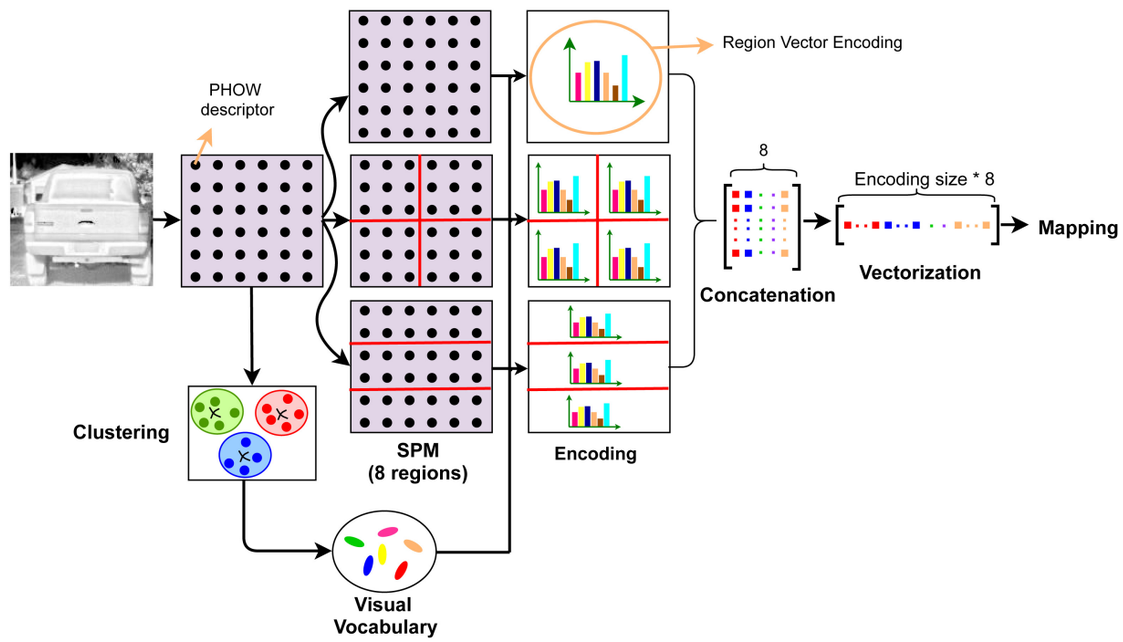

Fig. 3. The architecture of the encoding component based on SPM technique and BOVW model.

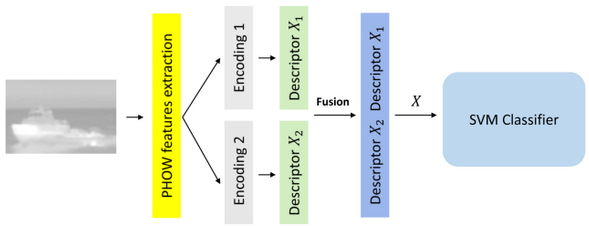

Fig. 4. Feature-level fusion process.

4.1.2. *Decision-level fusion*

Decision-level fusion is a potential approach where the outputs of multiple classifiers are combined independently to generate more reliable decision. These outputs can be label maps or class membership probability maps. In our experiment, we will work with a particular case of decision-level fusion which is the scores fusion.

Figure 5 shows the followed decisions fusion process. Let $S_1 = [S_1^{(1)} \quad \dots \quad S_1^{(k)} \quad \dots S_1^{(m)}]$ and $S_2 = [S_2^{(1)} \dots S_2^{(k)} \dots S_2^{(m)}]$ be the probability vectors of SVM classifier 1 and 2, respectively (m is the number of classes). To compute the final score vector $S_f = [S_f^{(1)} \quad \dots \quad S_f^{(k)} \quad \dots \quad S_f^{(m)}]$, several formulas can be used: MAX, OR, weighted sum, etc. In our work, we have used the weighted sum. This latter is defined by the following formula:

$$S_f = \alpha S_1 + (1-\alpha)S_2 \quad \text{s.t. } \alpha \in [0,1], \tag{10}$$

where α is chosen adaptively based on cross validation. The class with the highest probability in S_f is selected.

4.1.3. *Model-level or hybrid-level fusion*

Hybrid-level fusion combines the advantages of feature-level and decision-level fusion strategies. Various architectures for hybrid fusion have been proposed. Our adopted architecture consists on performing a features fusion of two models and combining its output with the scores of another model that has been processed independently.

Figure 6 shows the adopted hybrid fusion process. We perform a features fusion of two models to get S_1. S_2 score is

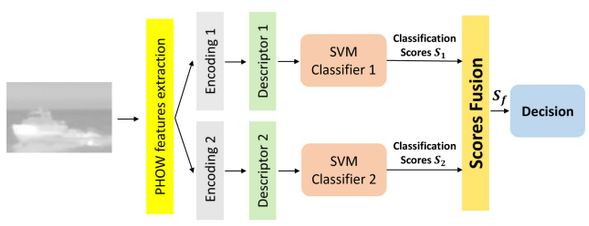

Fig. 5. Decision-level fusion process.

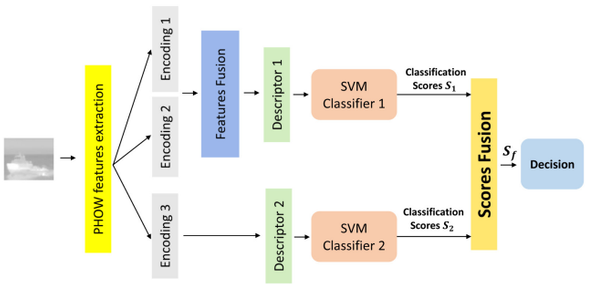

Fig. 6. Hybrid-level fusion process.

obtained by processing a third model independently. S_f is now obtained by applying the scores fusion explained above.

4.2. *Decision component*

The third component in our approach is decision. As shown in Fig. 4, a vector resulting from the concatenation of several encoding vectors will serve as an input for the SVM classifier, so this latter represents the decision part in the features fusion process.

For the two other fusion strategies (decision and hybrid), we have S_f as an input for the decision stage (Figs. 5 and 6). The class with the highest probability in S_f is selected.

5. Experiments and Results

In this section, we will describe the experimental setup and present the results obtained. First, we will briefly provide an overview of the thermal datasets used, VAIS and FLIR. Next, many experiments will be carried out to evaluate each encoding method separately. The encoding methods that show the best performances will be selected for the fusion process. Next, the experiences and results of the MMFEM are reported and discussed. Finally, a comparison of our results is provided with the state of the art in target recognition in both datasets.

5.1. *Datasets*

5.1.1. *FLIR dataset*

The FLIR thermal starter dataset is a multi-spectrum dataset for training and validation of object detection algorithms. It contains 10,228 frames (annotated thermal images and nonannotated RGB images), consisting of four classes: people, bicycles, cars and dogs. This dataset provides the scientific community with a great opportunity to develop safer and more efficient advanced driver assistant

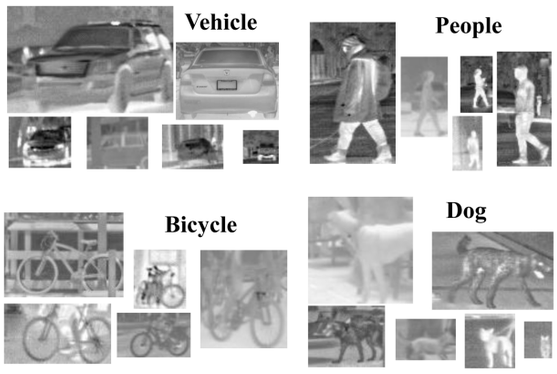

Fig. 7. Examples of FLIR dataset images.

Table 2. Number of training and test samples using the "FLIR dataset".

FLIR		
Class	Training	Test
Pedestrian	7815	5779
Vehicle	12,265	5432

systems (ADAS). Some example images from the FLIR starter thermal dataset are given in Fig. 7.

In our experiments, we have worked only with two classes. By respecting the official split of FLIR dataset, we have created a subdataset like [7] from the FLIR thermal starter dataset, which consists of two classes: vehicles and people categorized as a pedestrian. This dataset (hereby referred "FLIR dataset") comprises more than 30,000 images (Table 2) of different sizes, low contrast and objects are sometimes partially visible, making their detection difficult.

5.1.2. *VAIS dataset*

VAIS is a public dataset for object classification algorithms, consists of visible and infrared images of ships. It contains 2865 images, (1623 visible and 1242 IR), divided into six categories (Fig. 8). In our research, we have used both thermal and visible images for training, and only thermal images for testing (Table 3).

5.1.3. *Data pre-processing*

All FLIR images have been pre-processed to increase contrast. The processing operation consists of mapping the intensity values of an image to new values so that it saturates the bottom 1% and the top 1% of all pixel values.

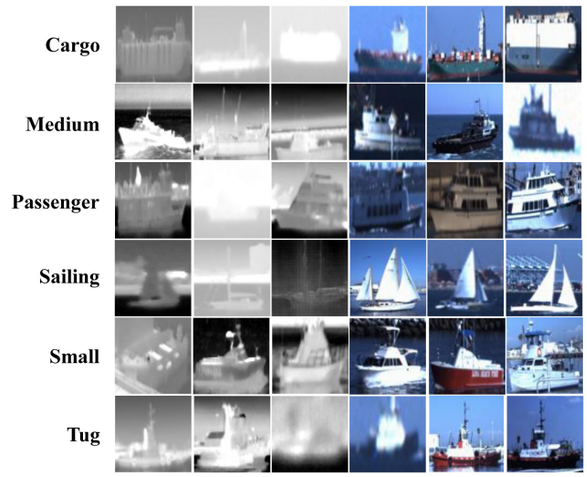

Fig. 8. Examples of VAIS dataset images.

Table 3. Number of training and test samples using the "VAIS dataset".

VAIS			
	Training		Test
Class	RGB	IR	IR
Cargo	103	83	97
Medium	99	62	90
Passenger	78	58	71
Sailing	214	148	151
Small	342	158	225
Tug	37	30	96

For VAIS dataset, all the images have been resized to a dimension of 79×79.

5.2. *Evaluation of the encoding component*

To achieve good performances, we proceed to analyze our system component by component. In this section, the encoding component is evaluated by looking for the best hyper-parameters of each encoding method. Except for FK encoding, which uses GMM to build BOVW, the other methods use K-means. The evaluation process is illustrated in Fig. 9.

For classification, the well-known algorithm "Support Vector Machine" is used.

5.2.1. *Experiments*

Histogram Encoding. For HE, the most significant parameter is the codebook size. We have varied this

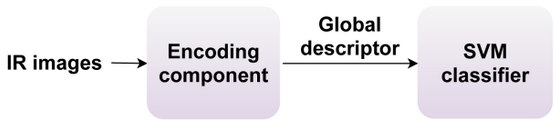

Fig. 9. Evaluation procedure of the encoding component.

Table 4. Classification accuracy of HE using VAIS dataset.

VAIS		
Experiment	Codebook size	Accuracy
HEv50	50	58.89
HEv100	100	60.74
HEv150	150	**61.88**
HEv200	200	**61.88**
HEv400	400	61.59
HEv800	800	59.32

Table 5. Classification accuracy of HE using FLIR dataset.

FLIR		
Experiment	Codebook size	Accuracy
HEf50	50	94.11
HEf100	100	94.25
HEf150	150	94.23
HEf200	200	94.43
HEf400	400	**94.57**
HEf800	800	94.42

parameter between 50 and 800 for the two datasets. The results are presented in Tables 4 and 5.

Kernel CodeBook encoding. Compared to the basic encoding, this encoding performs the quantization in a soft way and also requires defining the smoothing parameter σ, that it was searched in $[50, 100, 150, 200]$. For more speed and efficiency, we did not consider all visual words, but we tried to find the adequate number of the top nearest visual words. It was searched in $[5, 10, 20, 50]$. The details of the experiments and the best results obtained are mentioned in Tables 6 and 7.

Locality-constrained Linear encoding. As KCB encoding, LLC requires defining two hyper-parameters: the adequate number of the top nearest visual words and the constant of regularization β, that have been searched in $[5, 10, 20, 50]$ and $[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$, respectively.

Table 6. Classification accuracy of KCB encoding using VAIS dataset.

VAIS				
Experiment	Codebook size	σ	Top nearest	Accuracy
KCBv50	50	100	20	64.01
KCBv100	100	150	20	65.29
KCBv150	150	150	20	65.58
KCBv200	200	100	50	**66.57**
KCBv400	400	100	50	**66.57**
KCBv800	800	100	50	66.29

Table 7. Classification accuracy of KCB encoding using FLIR dataset.

FLIR				
Experiment	Codebook size	σ	Top nearest	Accuracy
KCBf50	50	100	20	94.90
KCBf100	100	100	20	95.24
KCBf150	150	100	50	95.29
KCBf200	200	100	20	95.43
KCBf400	400	100	20	95.59
KCBf800	800	100	50	**95.75**

Tables 8 and 9 show the best results obtained for each codebook size.

Fisher Kernel. Contrary to other methods, FK is based on GMM algorithm, we will vary the number of visual words in $[32, 48, 64, 96]$ for VAIS and $[32, 48, 64]$ for FLIR. The choice of codebook size for FK is restricted by memory capacity. The final vector resulting from the fisher encoding has a large dimension so it will be reduced by the PCA algorithm. The parameter α of power normalization is set to 0.5. The results are shown in Table 10.

Vector of Locally Aggregated Descriptors. The assignment of visual words is done in two manners: hard

Table 8. Classification accuracy of locality-constrained linear encoding using VAIS dataset.

VAIS				
Experience	Codebook size	β	Top nearest	Accuracy
LLCv50	50	10^{-2}	50	66.15
LLCv100	100	10^{-2}	50	65.86
LLCv150	150	10^{-2}	50	65.86
LLCv200	200	10^{-2}	50	**66.43**
LLCv400	400	10^{-2}	50	**66.43**
LLCv800	800	10^{-3}	50	65.86

Table 9. Classification accuracy of LLC.

FLIR				
Experiment	Codebook size	β	Top nearest	Accuracy
LLCf50	50	10^{-3}	50	94.55
LLCf100	100	10^{-3}	50	95.22
LLCf150	150	10^{-3}	20	95.32
LLCf200	200	10^{-4}	50	95.59
LLCf400	400	10^{-3}	50	95.83
LLCf800	800	10^{-3}	50	**95.94**

and soft. In the soft way, the KCB assignment is done with $\sigma = 100$ and the top 20 nearest visual words. As the FK, PCA is applied to the final vector with the same parameters and the choice of codebook size is also restricted by memory capacity. The power normalization parameter α is also set to 0.5. The results are shown in Tables 9 and 10.

Finally, the best results obtained for the encoding component have been summarized in Table 11.

5.2.2. *Discussion*

From the results presented below, VLAD has shown good performance for both datasets. As expected, HE has delivered low classification accuracy compared to the other methods except for FK in VAIS experiences which suffered from the overfitting problem. Soft quantization methods LLC and KCB have shown almost similar performances with small difference (e.g. KCBv200 versus LLCv200, KCBf800 versus LLCf800). FK and VLAD are the best performing in the FLIR and VAIS experiments, respectively (FKf64, VLADvs96).

By analyzing the results of VLAD and FK in FLIR experiments (after PCA application), we can see that both encodings with a small value of K and a small vector size are significantly better than the other methods with large values of K (e.g. FKf64 versus LLCf800, VLADsf32 versus KCBf400). For VAIS experiments, we can observe the same thing for VLAD (VLADsv96), as opposed to FK, which presented poor classification accuracy caused by the overfitting problem (e.g. FKv32, FKv96). Therefore, on average, VLAD has shown superiority over other encoding methods.

The soft quantization has largely outperformed the hard one. First, KCB and LLC have shown higher accuracy than HE (e.g. HEv200 versus KCBv200, HEf800 versus LLCf800). In addition, soft assignment on VLAD encoding performed better than hard assignment (e.g. VLADsf96 versus VLADhf96, VLADhv64 versus VLADsv64).

The experiments were carried out using a range of codebook sizes. Generally, increasing codebook size leads to high classification accuracy (LLCf800, KCBf800, VLADsv96, VLADsf96, FKf64). However, experiments have shown that a large codebook size does not always give the highest performance (e.g. LLCv200 versus LLCv800, HEf400 versus HEf800).

5.3. *Evaluation of MMFEM*

Now, and after the evaluation of the encoding component, we have methods that are the best performing in both datasets. So, the best models obtained are selected to use in the fusion process. In addition, we select the best models for each encoding to ensure the fusion between all encoding methods. Table 14 shows the selected models for each encoding method.

Features fusion. For the features fusion, we have chosen to combine two models $(L = 2)$. The two vectors each issued from a different encoding component will be concatenated into a single vector (Fig. 4). This latter is fed to a linear SVM classifier.

Table 10. Classification accuracy of FK encoding using FLIR dataset and VAIS dataset.

FLIR									
		Vector size after PCA							
Experiment	Codebook size	32	48	64	96	128	256	512	1024
FKf32	32	95.25	95.65	95.84	95.99	96.26	96.21	96.01	94.98
FKf48	48	95.58	95.96	96.15	96.16	96.19	96.36	96.18	95.65
FKf64	64	95.59	96.03	96.04	96.19	96.14	96.33	**96.42**	95.98
VAIS									
FKv32	32	57.75	**60.17**	60.88	60.03	58.61	59.60		
FKv48	48	57.04	58.75	59.75	59.75	59.46	59.89		
FKv64	64	57.89	58.75	58.75	60.03	59.46	58.46		
FKv96	96	57.75	58.04	58.04	60.1	59.01	59.46		

Table 11. Classification accuracy of VLAD encoding using VAIS dataset.

VAIS								
			Vector size after PCA					
Assignment	Experiment	Codebook size	32	48	64	96	128	256
Hard	VLADhv32	32	59.03	60.03	60.31	60.03	60.17	58.89
	VLADhv48	48	60.31	58.04	57.47	56.19	56.47	57.33
	VLADhv64	64	60.74	58.04	58.75	59.18	59.89	58.32
	VLADhv96	96	60.31	58.75	60.03	59.33	58.46	57.89
Soft	VLADsv32	32	64.01	65.72	65.43	63.44	61.59	61.31
	VLADsv48	48	63.87	65.15	66.15	63.30	62.3	62.3
	VLADsv64	64	64.58	66.71	66.57	63.44	63.58	61.45
	VLADsv96	96	63.3	**68.56**	66.43	61.88	61.88	62.02

Table 12. Classification accuracy of VLAD encoding using FLIR dataset.

FLIR										
			Vector size after PCA							
Assignment	Experiment	Codebook size	32	48	64	96	128	256	512	1024
Hard	VLADhf32	32	93.87	94.61	94.69	94.88	94.92	94.98	94.66	94.25
	VLADhf48	48	94.22	94.73	94.84	95.04	95.02	94.94	94.84	94.27
	VLADhf64	64	93.81	94.31	94.22	94.76	94.99	94.84	94.68	93.83
	VLADhf96	96	93.88	94.53	94.60	94.86	94.92	94.84	95.12	94.63
Soft	VLADsf32	32	95.72	95.93	95.85	95.98	96.10	95.78	95.50	95.36
	VLADsf48	48	95.75	95.77	95.66	96.05	96.11	96.13	95.74	95.68
	VLADsf64	64	95.86	95.86	96.08	96.34	96.35	96.11	95.86	95.74
	VLADsf96	96	95.98	95.93	96.05	96.36	96.35	**96.39**	96.16	95.61

Decisions fusion. As shown in Fig. 5, the scores of two models S_1 and S_2 are combined. We have tested three formulas to compute the final score vector S_f: $\max(S_1, S_2)$, OR (S_1, S_2) and weighted sum. The weighted sum (Eq. (10)) performed better than others.

As mentioned above, α is fixed by cross validation during training. The training set is randomly partitioned into two subsets: 80% training subset and 20% validation subset. For each α, we run the trained model in the validation subset. α that yields the best accuracy will be selected.

Table 13. Comparison between encoding methods.

	Accuracy	
Method	FLIR	VAIS
HE	94.57	61.88
KCB	95.75	66.57
LLC	95.94	66.43
FK	**96.42**	60.17
VLAD	96.39	**68.56**

The results of features fusion and decisions fusion are mentioned in Tables 15 and 16.

Discussion. The evaluation of HE has shown these limitations even with the fusion of its features with other encoding methods features. We have obtained low classification accuracy, and this is expected, because HE is based on hard assignment.

According to the experiments described above, features fusion in general does not enhance classification accuracy. However, we can see that features fusion (LLC+KCB) for VAIS and (VLAD + FK) for FLIR performs well and improves the classification results.

Based on decisions fusion results, we notice that decisions fusion is quite good, it improves the classification accuracy for nearly all the experiments performed, we have achieved an accuracy of 70.98% for VAIS with

Table 14. The selected models used in the fusion process.

	Selected models	
Methods	VAIS	FLIR
HE	HEv200	HEf400
KCB	KCBv200	KCBf400
	KCBv400	KCBf800
LLC	LLCv200	LLCf400
	LLCv400	LLCf800
VLAD	VLADsv64 (48)	VLADsf96(96)
	VLADsv96 (48)	VLADsf96(256)
FK	FKv32 (48)	FKf48(256)
		FKf64(512)

KCBv200 + VLADsv96(48) and 96.83% for FLIR with FKf48(256) + VLADsf96(96).

So features fusions (KCB + LLC) for VAIS and (FK + VLAD) for FLIR have shown an enhancement. The decisions fusion has proved to be effective especially for the models that have already performed well without fusion (VLAD for VAIS, FK and VLAD for FLIR).

To benefit from the advantages of both fusion strategies, we have proposed to work with the hybrid fusion (Fig. 6). Tables 17 and 18 show the carried out experiments and the obtained results.

All models are enhanced by hybrid fusion. The best accuracy achieved for VAIS is 71.26% and 96.96% for FLIR.

In general, the fusion method is quite effective to improve the performances. The hybrid fusion mostly

Table 15. Features fusion and decisions fusion accuracy using FLIR dataset.

		FLIR			
Method 1	Method 2	Method 1 accuracy	Method 2 accuracy	Features fusion accuracy	Decisions fusion accuracy
HEf400	KCBf400	94.57	95.59	95.29	95.67
	KCBf800	94.57	95.75	95.51	95.75
	LLCf400	94.57	95.83	95.46	95.91
	LLCf800	94.57	95.94	95.75	95.94
	VLADsf96(96)	94.57	96.36	94.88	96.36
	VLADsf96(256)	94.57	96.39	95.22	96.39
	FKf48(256)	94.57	96.36	95.09	96.36
	FKf64(512)	94.57	96.42	95.37	96.42
KCBf400	LLCf400	95.59	95.83	95.92	95.99
	LLCf800	95.59	95.94	96.02	96.15
	VLADsf96(96)	95.59	96.36	95.87	96.48
	VLADsf96(256)	95.59	96.39	96	96.42
	FKf48(256)	95.59	96.36	95.77	96.52
	FKf64(512)	95.59	96.42	95.93	96.55
KCBf800	LLCf400	95.75	95.83	95.91	96.14
	LLCf800	95.75	95.94	95.84	96.09
	VLADsf96(96)	95.75	96.36	95.86	96.56
	VLADsf96(256)	95.75	96.39	95.86	96.49
	FKf48(256)	95.75	96.36	95.89	96.55
	FKf64(512)	95.75	96.42	95.88	96.58
LLCf400	VLADsf96(96)	95.83	96.36	95.91	96.61
	VLADsf96(256)	95.83	96.39	95.98	96.56
	FKf48(256)	95.83	96.36	96.11	96.57
	FKf64(512)	95.83	96.42	96.16	96.57
LLCf800	VLADsf96(96)	95.94	96.36	96.03	96.68
	VLADsf96(256)	95.94	96.39	96.05	96.56
	FKf48(256)	95.94	96.36	96.13	96.66
	FKf64(512)	95.94	96.42	96.16	96.68
FKf48(256)	VLADsf96(96)	96.36	96.36	96.69	**96.83**
	VLADsf96(256)	96.36	96.39	**96.74**	**96.83**
FKf64(512)	VLADsf96(96)	96.42	96.36	96.65	96.80
	VLADsf96(256)	96.42	96.39	96.56	96.75

outperforms the simple concatenation (features fusion) and scores fusion (decisions fusion). Based on the above observations and analysis, we can conclude that the hybrid fusion method can boost classification performance for target recognition in infrared images.

5.4. *Performance comparison*

After evaluating the whole proposed approach, our results were compared with the state-of-the-art methods on VAIS and FLIR datasets, in terms of classification performance. The comparison results for VAIS and FLIR datasets are shown, respectively, in Tables 19 and 20. These results are already published in the original papers of each approach. To make a fair comparison, we kept the same training and test sets during our experiments.

VAIS: We have compared the proposed approach with the reference method [20], where the authors have used the Gnostic Field algorithm, CNN and the combination of both. Moreover, our approach has been compared with the traditional methods (HOG + SVM, LBP + SVM) [29], SFLPP [29, 44] and SRDA [29, 44]. Next, our approach was evaluated against deep CNN-based methods (ELM-CNN [46], hybrid fusion [31] and SF-SRDA [29]). Finally, we fine-tuned some recent pre-trained CNN on VAIS dataset.

Khellal *et al.* [46] showed the limitations of the approaches proposed in the reference method [22]. Khellal *et al.* [46] required an extreme learning machine method to learn CNN features and additional integrated extreme learning machine for classification, which doubles complexity.

To analyze the SF-SRDA [29] method, let us notice that there are two problems with this approach: (1) The feature vectors used to get 70.98% classification accuracy are very large size (VGG-19(relu5-4): 100,352, ResNet-152(pool5): 2048), so they worked with vectors of size 102,352, which implied higher consumption of memory resources and longer time for training and testing. However, our model was able to reach an accuracy of 71.26% with vector sizes (KCBf200: 1600; LLCf200: 1600; VLADsv96(48): 12,288 before PCA, 48 after PCA), so the largest vector size used is 12,288 (which is 1/8 of 100,352). (2) The final vector size that feeds the SVM classifier is "the number of classes minus

Table 16. Features fusion and decisions fusion accuracy using VAIS dataset.

VAIS					
Method 1	Method 2	Method 1 accuracy	Method 2 accuracy	Features fusion accuracy	Decisions fusion accuracy
HEv200	KCBv200	61.88	66.57	65.58	66.71
	KCBv400	61.88	66.57	65.15	67
	LLCv200	61.88	66.43	65.43	67.14
	LLCv400	61.88	66.43	65.86	66.57
	VLADsv64(48)	61.88	66.71	62.16	67.71
	VLADsv96(48)	61.88	68.56	64.85	68.71
	FKv32(48)	61.88	60.17	61.74	63.58
KCBv200	LLCv200	66.57	66.43	**67.99**	67
	LLCv400	66.57	66.43	67	67.71
	VLADsv64(48)	66.57	66.71	67	67.99
	VLADsv96(48)	66.57	68.56	67.56	**70.98**
	FKv32(48)	66.57	60.17	66	66.57
KCBv400	LLCv200	66.57	66.43	67.14	67.71
	LLCv400	66.57	66.43	67.56	67.85
	VLADsv64(48)	66.57	66.71	66.71	67.99
	VLADsv96(48)	66.57	68.56	67.56	69.41
	FKv32(48)	66.57	60.17	66.71	66.57
LLCv200	VLADsv64(48)	66.43	66.71	67	69.13
	VLADsv96(48)	66.43	68.56	67	70.70
	FKv32(48)	66.43	60.17	66.57	67
LLCv400	VLADsv64(48)	66.43	66.71	66.57	68.71
	VLADsv96(48)	66.43	68.56	67.56	69.13
	FKv32(48)	66.43	60.17	66.15	66.43
FKv32(48)	VLADsv64(48)	60.17	66.71	65.29	67
	VLADsv96(48)	60.17	68.56	66	67.85

Table 17. Hybrid fusion accuracy using VAIS dataset.

VAIS				
Features fusion selected methods	Accuracy	Decisions fusion selected method	Accuracy	Hybrid fusion accuracy
KCBv200 + LLCv200	67.99	VLADsv64(48)	66.71	68.28
	67.99	VLADsv96(48)	68.56	**71.26**
KCBv200 + LLCv400	67	VLADsv64(48)	66.71	68.28
	67	VLADsv96(48)	68.56	70.70
KCBv400 + LLCv200	67.14	VLADsv64(48)	66.71	68.71
	67.14	VLADsv96(48)	68.56	70.98
KCBv400 + LLCv400	67.56	VLADsv64(48)	66.71	68.14
	67.56	VLADsv96(48)	68.56	69.13
KCBv200 + VLADsv64(48)	67	VLADsv64(48)	66.71	68.14
	67	VLADsv96(48)	68.56	68.71
LLCv200 + VLADsv64(48)	67	VLADsv64(48)	66.71	69
	67	VLADsv96(48)	68.56	69.13

1" (size 5 for VAIS). For example, if we try to apply it on binary classification (like FLIR 2 classes), the length of the final vector will be 1, which is very small as an input vector for SVM, that it makes the proposed approach specific to multi-class classification.

The pre-trained CNN presented low-performance comparing to our approach, which is highly probable due to the fact that VAIS is a very small-size dataset (1 k images).

FLIR: The comparison was made using two traditional methods (MSER [7, 25] and WingerMSER [7]) and an

Table 18. Hybrid fusion accuracy for FLIR dataset.

FLIR				
Features fusion selected methods	Accuracy	Decisions fusion selected method	Accuracy	Hybrid fusion accuracy
FKf48(256) + VLADsf96(96)	96.69	KCBf800	95.75	96.69
	96.69	LLCf800	95.94	96.75
	96.69	VLADsf96(96)	96.36	96.93
	96.69	VLADsf96(256)	96.39	96.91
	96.69	FKf48(256)	96.36	96.69
	96.69	FKf64(512)	96.42	96.80
FKf48(256) + VLADsf96(256)	96.74	KCBf800	95.75	96.74
	96.74	LLCf800	95.94	96.77
	96.74	VLADsf96(96)	96.36	**96.96**
	96.74	VLADsf96(256)	96.39	96.79
	96.74	FKf48(256)	96.36	96.82
	96.74	FKf64(512)	96.42	96.81
FKf64(512) + VLADsf96(96)	96.65	KCBf800	95.75	96.67
	96.65	LLCf800	95.94	96.76
	96.65	VLADsf96(96)	96.36	96.79
	96.65	VLADsf96(256)	96.39	96.83
	96.65	FKf48(256)	96.36	96.74
	96.65	FKf64(512)	96.42	96.65
FKf64(512) + VLADsf96(256)	96.56	KCBf800	95.75	96.62
	96.56	LLCf800	95.94	96.70
	96.56	VLADsf96(96)	96.36	96.78
	96.56	VLADsf96(256)	96.39	96.68
	96.56	FKf48(256)	96.36	96.74
	96.56	FKf64(512)	96.42	96.62

Table 19. The performance comparison of our approach and the state of the art on FLIR dataset.

FLIR	
Approaches	Test accuracy (%)
MSER [7, 25]	80.0%
WingerMSER [7]	84.6%
AlexNet_conv4 [47]	88.8%
VGG19_conv5_3 [47]	94.1%
HE_SIFT [48]	94.8%
The proposed approach	**96.9%**

Table 20. The performance comparison of our approach and the state of the art on VAIS dataset.

VAIS	
Approaches	Test accuracy (%)
VGG 16	61.62%
VGG 19	59.57%
GoogleNet	54.66%
ResNet 50	62.34%
ResNet 101	63.72%
ResNet 152	62.97%
Gnostic Field + CNN [22]	57.72%
ELM-CNN [46]	61.17%
Hybrid fusion [31]	68.60%
LBP + SVM [29]	56.67%
HOG + SVM [29]	57.18%
SFLPP [29, 44]	65.43%
SRDA [29, 44]	70.56%
SF-SRDA [29]	70.98%
The proposed approach	**71.26%**

approach based on transfer learning [47] (VGGNet and ALexNet). Akula *et al.* [7] used the BOVW model with the basic encoding (HE), which justifies the low classification accuracy obtained. Akula *et al.* [47] fixed the whole part of features extraction, which is pre-trained on visible images. These latter are different from thermal images, which are sensitive to thermal emissions and not to light. Moreover, to achieve an accuracy of 94%, the SVM classifier was fed with a vector size of 100,352, which means a large memory consumption and a large computation time. The best result of our approach was obtained with the size vectors (FKf48 (256): 12,288 before PCA, 256 after PCA; VLADsf96(256): 12,288 before PCA, 256 after PCA; VLADsf96(96): 12,288 before PCA, 96 after PCA), so the largest vector size used is 24,576 (which is 1/4 of 100,352) and the largest vector that was passed to an SVM classifier has a length of 512.

Therefore, the principal conclusion of this experiment is that, according to the best of our knowledge, the proposed approach outperforms the state-of-the-art models in both datasets: VAIS and FLIR.

6. Conclusion

In this paper, we have proposed a classification method for infrared images. The proposed pipeline of our approach is composed of three parts: SIFT features encoding through several methods, multi-model fusion and a decision part. Specifically, feature vectors are assigned to visual words of the codebook through five popular feature encoding methods (HE, KCB, LLC, FK, VLAD). These latter are evaluated separately to select the best-performing ones. The selected models were fused by three strategies: feature-level fusion, decision-level fusion and hybrid-level fusion. The experiments have shown that: (1) VLAD outperforms the other encoding methods for the infrared images classification on both datasets (VAIS and FLIR). (2) Fusion especially decisions fusion and hybrid fusion, improve the classification performances in infrared images.

Finally, a comparison was made with the state of the art in target recognition on FLIR and VAIS. Our approach has achieved 71.26% for VAIS and 96.96% for FLIR, which has exceeded the state-of-the-art.

References

[1] A. Balamurugan and B. Suganya, An efficient real time face expression identification system using SVM, *J. Phys. Conf. Ser.* **1916** (2021) 012229.

[2] M. Teutsch and W. Krüger, Classification of small boats in infrared images for maritime surveillance, in *Int. WaterSide Security Conf.*, Carrara, Italy (IEEE, 2010), pp. 1–7.

[3] J. Su, M. Coombes, C. Liu, Y. Zhu, X. Song, S. Fang, L. Guo and W.-H. Chen, Machine learning-based crop drought mapping system by UAV remote sensing RGB imagery, *Unmanned Syst.* **8**(1) (2020) 71–83.

[4] C. Minwalla, D. Tulpan, N. Belacel, F. Famili and K. Ellis, Detection of airborne collision-course targets for sense and avoid on unmanned aircraft systems using machine vision techniques, *Unmanned Syst.* **4**(4) (2016) 255–272.

[5] N. Sun, F. Jiang, H. Yan, J. Liu and G. Han, Proposal generation method for object detection in infrared image, *Infrared Phys. Technol.* **81** (2017) 117–127.

[6] B. Fahima and N. Abdelkrim, Multispectral visual odometry using SVSF for mobile robot localization, *Unmanned Syst.* **10**(3) (2021) 1–16.

[7] A. Akula, R. Ghosh, S. Kumar and H. K. Sardana, WignerMSER: Pseudo-Wigner distribution enriched MSER feature detector for object recognition in thermal infrared images, *IEEE Sens. J.* **19**(11) (2019) 4221–4228.

[8] D. G. Lowe, Object recognition from local scale-invariant features, in *Proc. 7th IEEE Int. Conf. Computer Vision*, Kerkyra, Greece (IEEE, 1999), pp. 1150–1157.

[9] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, Speeded-up robust features (SURF), *Comput. Vis. Image Underst.* **110**(3) (2008) 346–359.
[10] K. Feng, W. Li, J. Han and F. Pan, Low-latency aerial images object detection for UAV, *Unmanned Syst.* **10**(1) (2022) 57–67.
[11] S. Khan, M. Tufail, M. T. Khan, Z. A. Khan, J. Iqbal and A. Wasim, A novel framework for multiple ground target detection, recognition and inspection in precision agriculture applications using a UAV, *Unmanned Syst.* **10**(1) (2022) 45–56.
[12] L. Dai, Y. Zhu, G. Luo, C. He and H. Lin, A real-time visual tracking approach using sparse autoencoder and extreme learning machine, *Unmanned Syst.* **3**(4) (2015) 267–275.
[13] K. Jungling and M. Arens, Feature based person detection beyond the visible spectrum, in *IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops*, Miami, FL, USA (IEEE, 2009), pp. 30–37.
[14] K. Van De Sande, T. Gevers and C. Snoek, Evaluating color descriptors for object and scene recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9) (2009) 1582–1596.
[15] T. P. Weldon, W. E. Higgins and D. F. Dunn, Efficient Gabor filter design for texture segmentation, *Pattern Recognit.* **29**(12) (1996) 2005–2015.
[16] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray, Visual categorization with bags of keypoints, in *Workshop on Statistical Learning in Computer Vision, ECCV*, Prague (Springer, 2004), pp. 1–2.
[17] J. C. Van Gemert, J.-M. Geusebroek, C. J. Veenman and A. W. Smeulders, Kernel codebooks for scene categorization, in *10th European Conf. Computer Vision*, Marseille, France (Springer, 2008), pp. 696–709.
[18] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang and Y. Gong, Locality-constrained Linear Coding for image classification, in *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, San Francisco, CA, USA (IEEE, 2010), pp. 3360–3367.
[19] F. Perronnin, J. Sánchez and T. Mensink, Improving the fisher kernel for large-scale image classification, in *11th European Conf. Computer Vision*, Crete, Greece (Springer, 2010), pp. 143–156.
[20] H. Jégou, M. Douze, C. Schmid and P. Pérez, Aggregating local descriptors into a compact image representation, in *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, San Francisco, CA, USA (IEEE, 2010), pp. 3304–3311.
[21] S. Lazebnik, C. Schmid and J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in *IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'06)* (IEEE, New York, NY, USA, 2006), pp. 2169–2178.
[22] M. M. Zhang, J. Choi, K. Daniilidis, M. T. Wolf and C. Kanan, VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums, in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, Boston, MA, USA (IEEE, 2015), pp. 10–16.
[23] FLIR thermal starter dataset (2018), available online: https://www.flir.com/oem/adas/adas-dataset-form/.
[24] S. Malpani, C. Asha and A. Narasimhadhan, Thermal vision human classification and localization using bag of visual word, in *IEEE Region 10 Conf. (TENCON)*, Singapore (IEEE, 2016), pp. 3135–3139.
[25] N. V. A. Akula, R. Ghosh, N. Guleria, S. Kumar and H. K. Sardana, Towards an optimal bag-of-features representation for vehicle type classification in thermal infrared imagery, in *Optics and Photonics for Information Processing XII* (SPIE, 2018), pp. 191–207.
[26] N. V. A. Akula and H. K. Sardana, Optimized bag of features framework for object recognition in thermal infrared images, *J. Electron. Imaging* **27**(6) (2018) 1–15.
[27] Z. Zhao, X. Fan, G. Xu, L. Zhang, Y. Qi and K. Zhang, Aggregating deep convolutional feature maps for insulator detection in infrared images, *IEEE Access* **5** (2017) 21831–21839.
[28] S. Vadivelu, S. Ganesan, O. V. R. Murthy and A. Dhall, Thermal imaging based elderly fall detection, in *Computer Vision — ACCV 2016 Workshops*, Taipei, Taiwan (Springer, 2017), pp. 541–553.
[29] E. Zhang, K. Wang and G. Lin, Classification of marine vessels with multi-feature structure fusion, *Appl. Sci.* **9**(10) (2019) 2153.
[30] C. E. Santos and B. Bhanu, Dyfusion: Dynamic IR/RGB fusion for maritime vessel recognition, in *25th IEEE Int. Conf. Image Processing*, Athens, Greece (IEEE, 2018), pp. 1328–1332.
[31] X. Qiu *et al.*, Deep convolutional feature fusion model for multi-spectral maritime imagery ship recognition, *J. Comput. Commun.* **8**(11) (2020) 23.
[32] Q. Shi, W. Li, F. Zhang, W. Hu, X. Sun and L. Gao, Deep CNN with multi-scale rotation invariance features for ship classification, *IEEE Access* **6** (2018) 38656–38668.
[33] L. Huang, W. Li, C. Chen, F. Zhang and H. Lang, Multiple features learning for ship classification in optical imagery, *Multimed. Tools Appl.* **77**(11) (2018) 13363–13389.
[34] K. Chatfield, V. S. Lempitsky, A. Vedaldi and A. Zisserman, The devil is in the details: An evaluation of recent feature encoding methods, in *Proc. British Machine Vision Conf.*, Dundee, Scotland (BMVA, 2011), pp. 76.1–76.12.
[35] J. Yang, Y.-G. Jiang, A. G. Hauptmann and C.-W. Ngo, Evaluating bag-of-visual-words representations in scene classification, in *Proc. Int. Workshop on Workshop on Multimedia Information Retrieval*, Augsburg, Bavaria, Germany (ACM, 2007), pp. 197–206.
[36] S. Lloyd, Least squares quantization in PCM, *IEEE Trans. Inf. Theory* **28**(2) (1982) 129–137.
[37] D. A. Reynolds, Gaussian mixture models, in *Encyclopedia of Biometrics* (Springer, US, 2009), pp. 659–663.
[38] G. J. McLachlan, S. X. Lee and S. I. Rathnayake, Finite mixture models, *Ann. Rev. Stat. Appl.* **6**(1) (2019) 355–378.
[39] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (CRC Press, 1986).
[40] F. Perronnin and C. Dance, Fisher kernels on visual vocabularies for image categorization, in *IEEE Conf. Computer Vision and Pattern Recognition*, Minneapolis, MN, USA (IEEE, 2007), pp. 1–8.
[41] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez and C. Schmid, Aggregating local image descriptors into compact codes, *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(9) (2012) 1704–1716.
[42] A. Vedaldi and A. Zisserman, Efficient additive kernels via explicit feature maps, *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(3) (2012) 480–492.
[43] P. Tzirakis, S. Zafeiriou and B. Schuller, Real-world automatic continuous affect recognition from audiovisual signals, in *Multimodal Behavior Analysis in the Wild*, eds. X. Alameda-Pineda, E. Ricci and N. Sebe (Academic Press, USA, 2019), pp. 387–406.
[44] G. Lin, H. Zhu, X. Kang, C. Fan and E. Zhang, Multi-feature structure fusion of contours for unsupervised shape classification, *Pattern Recognit. Lett.* **34**(11) (2013) 1286–1290.
[45] D. Cai, X. He and J. Han, SRDA: An efficient algorithm for large-scale discriminant analysis, *IEEE Trans. Knowl. Data Eng.* **20**(1) (2008) 1–12.
[46] A. Khellal, H. Ma and Q. Fei, Convolutional neural network based on extreme learning machine for maritime ships recognition in infrared images, *Sensors* **18**(5) (2018) 1490.
[47] A. Akula and H. K. Sardana, Deep CNN-based feature extractor for target recognition in thermal images, in *IEEE Region 10 Conf. (TENCON)*, Kochi, India (IEEE, 2019), pp. 2370–2375.
[48] B. Nebili, A. Khellal and A. Nemra, Histogram encoding of SIFT based visual words for target recognition in infrared images, in *Int. Conf. Recent Advances in Mathematics and Informatics*, Tebessa, Algeria (IEEE, 2021), pp. 1–6.