Feature coding for image classification based on saliency detection and fuzzy reasoning and its application in elevator videos

Xiao Lv* , Dingdong Zou, Lei Zhang and Shangyuan Jia
Chongqing special equipment inspection and research institute
No.5 Furongyuan Road Northern New Chongqing
PEOPLE'S REPUBLIC OF CHINA
lvxiao87@126.com, cq_zdd@126.com, zl_816@163.com, 1458018035@qq.com

Abstract: - Feature coding is an fundamental step in bag-of-words based model for image classification and have drawn increasing attention in recent works. However, there still exits ambiguity problem, and it is also sensitiveness to unusual features. To improve the stability and robustness, we introduce saliency detection and fuzzy reasoning rules to propose an novel coding scheme. In detail, saliency maps generated by saliency detection are first used to divide each image into salient and non-salient region, then a structured dictionary is obtained by combing two separated codebooks in them. Secondly, fuzzy reasoning rules are introduced to choose the most salient and stable codewords to encode. Finally, saliency maps are incorporated into pooling operation named saliency based spatial pooling to introduce spatial information. Experiments on several datasets demonstrate our approach outperforms all other coding methods in image classification. Furthermore, we also apply it into elevator video event classification, which shows the potential application in intelligent elevator video surveillance, such as overload detection, violence detection, video summarization.

Key-Words: - Image classification, feature coding, saliency detection, fuzzy reasoning, elevator video event

1 Introduction

Automatic image classification is one of the most fundamental problems in computer vision and pattern recognition, whose aim is to assign one or more category labels to an image. It has drawn increasing attention from the researchers around the world due to its widespread prospects in a wide range of applications, e.g., image retrieval [1, 36], video retrieval [2], video surveillance [3], human-computer interaction [4], web content analysis [5], and biomedical [6, 37]. There are many approaches proposed for image classification in the literatures. Among them, the bag-of-words (BOW) model [7] and its extensions [8] achieve the state-of-the-art performance in several famous databases, such as Caltech 101 [9], Scenes 15 [10], Caltech 256 [11], and PASCAL VOC [12].

The BOW quantizes local descriptors into discrete visual codewords and counts their occurrence frequencies in the entire image. Then the resulting histogram is used as the image representation. Fig. 1 shows the general framework of the BOW model. It usually comprises of the following common steps: (1) Feature extraction. It extracts images' local features by detectors or dense sampling and then calculates their descriptors, such as Harris detector [13], affine invariant salient region detector [14], SIFT (Scale-Invariant Feature Transform) [15] descriptor, HOG (Histogram of Oriented Gradient) [16] descriptor. (2) Codebook generation. After obtained local descriptors, a codebook is usually needed to represent them. It is typically generated by clustering (e.g., K-means [17]) over a subset of descriptors, which is randomly sampled from all descriptors in database in real application for computational efficiency. (3) Feature coding and pooling. In this step, each local descriptor first activates a number of codewords, and generate a coding vector. Then, all responses on each codeword are integrated into one value by feature pooling. Various coding and pooling strategies will be described in detail in Section 2. The output of this step is a vector whose length is equal to the size of the codebook, namely the final image representation. (4) Classification. Finally, the image representation vectors are sent to a classifier, such as SVM (Support Vector Machine) [18-19] for classification.

Among these steps, feature coding and pooling is the fundamental component, which will greatly influences image classification in terms of both

* Corresponding author.

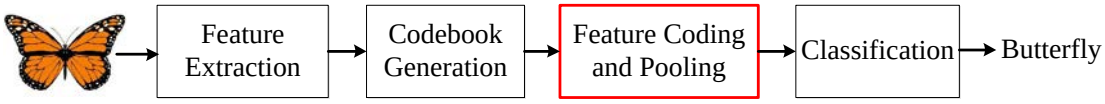


Fig. 1 The framework of the Bag-of-Words based model.

accuracy and computation cost [20]. Thus, it has drawn increasing attention in recent works, and various good strategies have been proposed in the literatures. Hard assignment (HA) [7-8] is the original coding method in BOW, which assigns descriptors to just one codeword nearest to it. Although is simple, it yields high quantization error. Then, soft assignment (SA) [21] is developed, wherein each descriptor is represented by all the codewords according to their Euclidean distances in Gaussian function. To further improve it, localized soft assignment coding (LSAC) [22] was proposed by introducing locality constraint. Sparse coding (SC) [23] is another novel method to reduce quantization error, which is realized by reconstructing descriptors plus a sparse constraint to the codes. However, it is non-consistent and time consuming. Then, locality-constrained linear coding (LLC) [24] was proposed by incorporating locality constraint into the objective function to encourage similar descriptors have similar codes. While Gao et al. [25] proposed Laplcian sparse coding (LSC) to preserve the consistence of coding. However, it is computationally infeasible. In order to meet the applications of large scale image classification, high dimensional schemes have been proposed, such as Fisher kernel coding (FKC) [26] and super vector coding (SVC) [27]. It usually needs a large quantity of memory. Huang et al. [28] found that saliency is one of the fundamental characteristics of feature coding when combining with Max-pooling and then proposed saliency-based coding (SaC), which performs much better than classic assignment schemes and more efficient than reconstruction based schemes. To improve it, Wu et al. [29] further proposed group salient coding (GSC), wherein the latent structure information of a codebook is explored by grouping neighboring codewords into a group-code. Recently, a novel approach called local similarity global coding (LSGC) [30] was proposed, which uses the local similarities between bases to obtain a nonlinear global similarity measure between local descriptor and bases.

From the above arguments, we can see that there are still some limitations haven't been well solved in previous works. We summarize it in Table 1. In this paper, we propose a coding and pooling scheme with low quantization, non-consistency, computational cost, ambiguity, though introducing saliency detection and fuzzy reasoning rules, which called fuzzy reasoning based salient coding (FRSC). In detail, saliency detection are used to generate saliency maps which are used to divide image into salient and non-salient region, and then combine two separated codebook clustered in them to produce a structured dictionary. Then, fuzzy reasoning rules are introduced to select the most salient and stable codewords to encode. By using it, the underlying manifold structure of descriptors can be well captured. Finally, saliency maps are used again to locate the interest object, which can be used to spatial pooling to incorporate spatial information. Thus, our new improved BOW model can be obtained by combing the above together.

The remainder of this paper is organized as follows. In Section 2, we briefly analyze various coding and pooling schemes. Section 3 presents our coding and pooling approach. Then experimental results on the Caltech 101, Scenes 15, and UIUC Sport databases are provided in Section 4. Finally in Section 5, conclusions are drawn, some future work and applications are discussed.

2 Related Work

In this section, we briefly review commonly used coding and pooling schemes. Let x_i ($x_i \in R^d$) be a d dimensional descriptor, $B_{d\times M} = (b_1, b_2, \ldots, b_M)$ be a codebook with M cluster centers, and u_i ($u_i \in R^d$) be the coding coefficient vector of x_i, e.g., u_{ij} be the response of xi on codeword b_j.

Hard assignment coding: For a local descriptor x_i, only the closest codeword is used for coding, in which Euclidean distance is used.

$$u_{i,j} = \begin{cases} 1, & \text{If } j = \underset{j=1,\cdots,n}{\operatorname{argmin}} \left\| x_i - b_j \right\|_2^2 \\ 0, & \text{others} \end{cases} \quad (1)$$

Soft assignment coding: Each local descriptor is encoded by multiple codewords using the kernel function of distance between descriptors and codewords, such as Gaussian function.

$$u_{i,j} = \frac{\exp\left(-\beta \left\| x_i - b_j \right\|_2^2\right)}{\sum_{k=1}^{m} \exp\left(-\beta \left\| x_i - b_k \right\|_2^2\right)} \quad (2)$$

Table 1 Comparison of previous coding schemes. H: high, M: middle, L: low.

	Quantization error	Non-consistency	Computational cost	Ambiguity
HA[8]	H	L	L	H
SA[21]	L	L	M	H
LSAC[22]	L	L	L	H
SC[23]	L	H	H	H
LLC[24]	L	L	L	H
LSC[25]	L	L	H	H
FKC[26]	L	L	H	H
SVC[27]	L	L	H	H
SaC[28]	H	L	L	H
GSC[29]	L	L	L	H
LSGC[30]	L	L	H	H

where β is the smoothing factor controlling the softness of the assignment, and m$\in$[1, n].

Localized soft assignment coding: It is an improved version of SA. Their difference is that SA encodes each descriptor across all the codewords while LSAC confines it to a local neighborhood around the coded descriptor.

$$u_{i,j}=\frac{\exp\left(-\beta\hat{d}\left(x_i,b_j\right)\right)}{\sum_{l=1}^{n}\exp\left(-\beta\hat{d}\left(x_i,b_l\right)\right)}$$
$$\hat{d}\left(x_i,b_l\right)=\begin{cases}d\left(x_i,b_l\right), \text{If } b_l\in \mathrm{N}_K\left(x_i\right)\\ \propto, \qquad \text{others}\end{cases} \tag{3}$$

Sparse coding: It is a reconstruction based coding, which use sparse constraint to alleviate the quantization error. In detail, it represents a local descriptor by a linear combination of a sparse set of basis vectors by solving an l_1-norm regularized approximation problem, which can be solved by FS (Feature-sign search algorithm) [31] and LD (Lagrange dual) [31].

$$u_i=\arg\min\left\|x_i-Bu_i\right\|_2^2+\lambda\left\|u_i\right\|_1$$
$$s.t.\ \left\|u_i\right\|_1=1 \tag{4}$$

where $\|\bullet\|_1$ denotes the l_1-norm.

Locality-constrained linear coding: Further study [32] found that the locality constraint is more important than the sparse constraint. Thus, LLC was proposed by introducing the locality constraint, which is obtained by minimizing the Euclidean distance between each descriptor and codewords.

$$u_i=\arg\min\left\|x_i-Bu_i\right\|_2^2+\lambda\left\|d_i\odot u_i\right\|_2^2$$
$$s.t.\ \left\|u_i\right\|_1=1 \tag{5}$$
$$d_i=\exp\left(\frac{dist\left(x_i,B\right)}{\sigma}\right)$$

where $dist(x_i, B)$ denotes the Euclidean distance between x_i and b_j, σ is a parameter controlling the weighting vector d_i.

Furthermore, a simplified and fast implementation was proposed to reduce the computation cost.

$$u_i=\arg\min\left\|x_i-\tilde{B}u_i\right\|_2^2$$
$$s.t.\ \left\|u_i\right\|_1=1 \tag{6}$$

where $\tilde{B}$ is K closest codewords to x_i.

Salient coding: Its main idea is employing the difference between the closest codeword and the other K-1 closest codewords to reflect saliency. Thus, a local descriptor can be represented as:

$$\Psi\left(x_i,\tilde{b}_j\right)=1-\frac{\left\|x_i-\tilde{b}_j\right\|_2}{\frac{1}{K-1}\sum_{k\neq j}^{K}\left\|x_i-\tilde{b}_k\right\|_2}$$
$$u_{i,j}=\begin{cases}\Psi\left(x,b_j\right), \text{If } j=\arg\min\limits_{l,l\in n}\left(\left\|x-b_l\right\|^2\right)\\ 0, \qquad \text{others}\end{cases} \tag{7}$$

where $[\tilde{b}_1,\tilde{b}_2,\cdots,\tilde{b}_k]$ is the K closest codewords to x_i.

Group saliency coding: Hard assignment used in SaC is coarse for feature coding. Thus, group coding was introduced in GSC, whose main idea is calculating the saliency response of a group of codewords, and the response is fed back to all the codewords in the group, finally, the maximum of all responses are calculated according to different group sizes.

$$u_{i,j}^{k}=\begin{cases}\Phi^k\left(x_i\right), \text{If } b_j\in g\left(x_i,K\right)\\ 0, \qquad \text{others}\end{cases}$$
$$\Phi^k\left(x_i\right)=\sum_{t=1}^{G+1-k}\left(\left\|x_i-\tilde{b}_{k+t}\right\|_2^2-\left\|x_i-\tilde{b}_k\right\|_2^2\right) \tag{8}$$

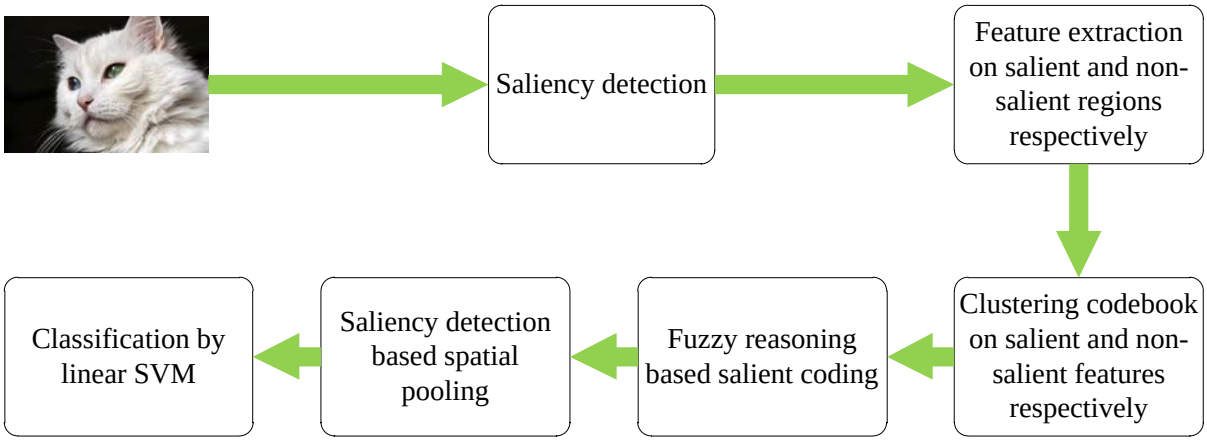


Fig.2 The flowchart of the proposed method.

where $g(x_i, K)$ denotes the K closest codewords of x_i, G is the maximum group size.

A pooling operation is often needed to obtain an image-level representation when the coding responses of all local descriptors are calculated.

Sum pooling:

$$p_j = \sum_{i=1}^{q} u_{i,j} \tag{9}$$

Average pooling:

$$p_j = \frac{1}{q}\sum_{i=1}^{q} u_{i,j} \tag{10}$$

where q is the total number of local descriptors in an image.

Max pooling:

$$p_j = \max_i u_{i,j} \tag{11}$$

The max pooling often gets better performance than sum and average pooling, such as in SC, LLC, SaC, GSC,LSAC. However, its mechanism has not been fully studied in the literature.

3 The Proposed Method

The main components of the proposed approach is composed of three parts: saliency detection based structured codebook generation, fuzzy reasoning based salient coding, and saliency detection based spatial pooling. The overview of the proposed method is shown in Fig.2. The details of these three aspects are presented as follows.

3.1 Saliency detection based codebook generation

As we know, images are usually corrupted by noise and there are often more than one object in an image with different shapes and occlusions, even in the same class. Thus, researchers divide an image into two items which called the correlated (or common) part and the specific (or noisy) part respectively. And the both parts are more robust and discriminative for image classification because it captures complementary attributes in an image. Inspired by these observations, some low-rank based methods [32-33] are proposed. They use low-rank and spares techniques to decompose local features of an image or images within each class into a low-rank part and a sparse part, which represent homogeneousness and diversity respectively. However, they are time consuming. In this paper, we use saliency map generated by saliency detection to decompose images into salient parts and non-salient parts. For computational efficiency, we use efficient saliency detection method in [34]. Then, we extract SIFT descriptors in both parts and cluster by K-means to get two codebooks. Finally, we combine them as a structured codebook to encode original descriptors. Experimental results show that the structured dictionary has comparable representation capability with low-rank based methods, which are presented in Section 4.

3.2 Fuzzy reasoning based salient coding

Recently, saliency based coding methods get satisfactory results due to its efficiency and stable representation. However, they will lose their superiority in performance when codebook size is relatively large. Furthermore, they are also sensitiveness to unusual features, e.g., noisy features. Thus, we present a fuzzy reasoning based coding scheme to solve these problems in this paper.

In saliency based coding, the response of a local descriptor is reflected by saliency degree using K closest codewords selected from the codebook. Then, only the maximum response is preserved while the low responses are suppressed in the later

maximum pooling operation. Therefore, we think that only those largest responses are the meaningful responses. Saliency is used to measure this character in the original saliency based coding. However, if all the *K* closest codewords are near to the local descriptor, they have similar saliency value, it cannot reflect the saliency in this case, because all these *K* closest codewords are needed to represent the local descriptor. Thus, we introduce fuzzy reasoning rules to reflect it.

First, we use d_i to denote the Euclidean distance between the local descriptor and the *i*th closest codeword, and s_i to denote the saliency value of each local descriptor on the *K* closest codewords.

$$s_i = 1 - \frac{d_i}{\frac{1}{K-1}\sum_{k \neq i}^{K} d_k} \tag{12}$$

Take *K*=5 for example, we can define six fuzzy rules which are described as follows:

Rule-1: If s_1 is low, s_2 is low, s_3 is low, s_4 is low, and s_5 is low, then none of the *K* closest codewords can represent the local descriptor independently.

Rule-2: If s_1 is high, s_2 is low, s_3 is low, s_4 is low, and s_5 is low, then only the closest codeword can represent the local descriptor independently.

Rule-3: If s_1 is high, s_2 is high, s_3 is low, s_4 is low, and s_5 is low, then the two closest codewords can represent the local descriptor stably.

Rule-4: If s_1 is high, s_2 is high, s_3 is high, s_4 is low, and s_5 is low, then the three closest codewords can represent the local descriptor stably.

Rule-5: If s_1 is high, s_2 is high, s_3 is high, s_4 is high, and s_5 is low, then the four closest codewords can represent the local descriptor stably.

Rule-6: If s_1 is high, s_2 is high, s_3 is high, s_4 is high, and s_5 is high, then all the five closest codewords can represent the local descriptor stably.

Low and high are fuzzy membership functions shown in Eq. 13 and Eq. 14. Both of them are trapezoid shapes and illustrated in Fig. 3.

$$\mathrm{Low}(s) = \begin{cases} 1, & s < a \\ \dfrac{s-b}{a-b}, & a \le s < b \\ 0, & s \ge b \end{cases} \tag{13}$$

$$\mathrm{High}(s) = \begin{cases} 0, & s < a \\ \dfrac{s-a}{b-a}, & a \le s < b \\ 1, & s \ge b \end{cases} \tag{14}$$

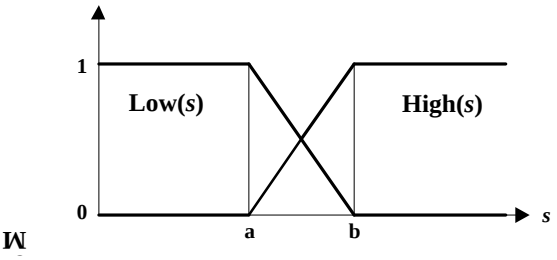


Fig.3. The fuzzy membership functions Low(*s*) and High(*s*).

Note that all the saliency values are normalized to 1 in this paper. And the two parameters *a* and *b* are usually set to 0.2 and 0.8 respectively. Then, let the fuzzy truth value F be defined below:

$$F_1 = \mathrm{Low}(s_1) \cdot \mathrm{Low}(s_2) \cdot \mathrm{Low}(s_3) \cdot \mathrm{Low}(s_4) \cdot \mathrm{Low}(s_5)$$
$$F_2 = \mathrm{High}(s_1) \cdot \mathrm{Low}(s_2) \cdot \mathrm{Low}(s_3) \cdot \mathrm{Low}(s_4) \cdot \mathrm{Low}(s_5)$$
$$F_3 = \mathrm{High}(s_1) \cdot \mathrm{High}(s_2) \cdot \mathrm{Low}(s_3) \cdot \mathrm{Low}(s_4) \cdot \mathrm{Low}(s_5)$$
$$F_4 = \mathrm{High}(s_1) \cdot \mathrm{High}(s_2) \cdot \mathrm{High}(s_3) \cdot \mathrm{Low}(s_4) \cdot \mathrm{Low}(s_5)$$
$$F_5 = \mathrm{High}(s_1) \cdot \mathrm{High}(s_2) \cdot \mathrm{High}(s_3) \cdot \mathrm{High}(s_4) \cdot \mathrm{Low}(s_5)$$
$$F_6 = \mathrm{High}(s_1) \cdot \mathrm{High}(s_2) \cdot \mathrm{High}(s_3) \cdot \mathrm{High}(s_4) \cdot \mathrm{High}(s_5)$$

where product inference engine [35] is used to realize the fuzzy reasoning. After all the fuzzy truth values obtained, we can determine which codewords can be used to encode by the largest fuzzy truth value. Then, the previous coding schemes can be used to encode, here, the SaC is adopted for efficiency. Thus, our FRSC can be defined by:

Case-1:

If $\max\{F_1, F_2, F_3, F_4, F_5, F_6\} = F_2$

$\mathrm{FRSC} = \{c_1, 0, 0, 0, 0\}$;

Case-2:

If $\max\{F_1, F_2, F_3, F_4, F_5, F_6\} = F_3$

$\mathrm{FRSC} = \{c_1, c_2, 0, 0, 0\}$;

Case-3:

If $\max\{F_1, F_2, F_3, F_4, F_5, F_6\} = F_4$

$\mathrm{FRSC} = \{c_1, c_2, c_3, 0, 0\}$;

Case-4:

If $\max\{F_1, F_2, F_3, F_4, F_5, F_6\} = F_5$

$\mathrm{FRSC} = \{c_1, c_2, c_3, c_4, 0\}$;

Case-5:

If $\max\{F_1, F_2, F_3, F_4, F_5, F_6\} = F_1$

or $\max\{F_1, F_2, F_3, F_4, F_5, F_6\} = F_6$

$\mathrm{FRSC} = \{c_1, c_2, c_3, c_4, c_5\}$;

$c_i = \mathrm{High}(s_i)$

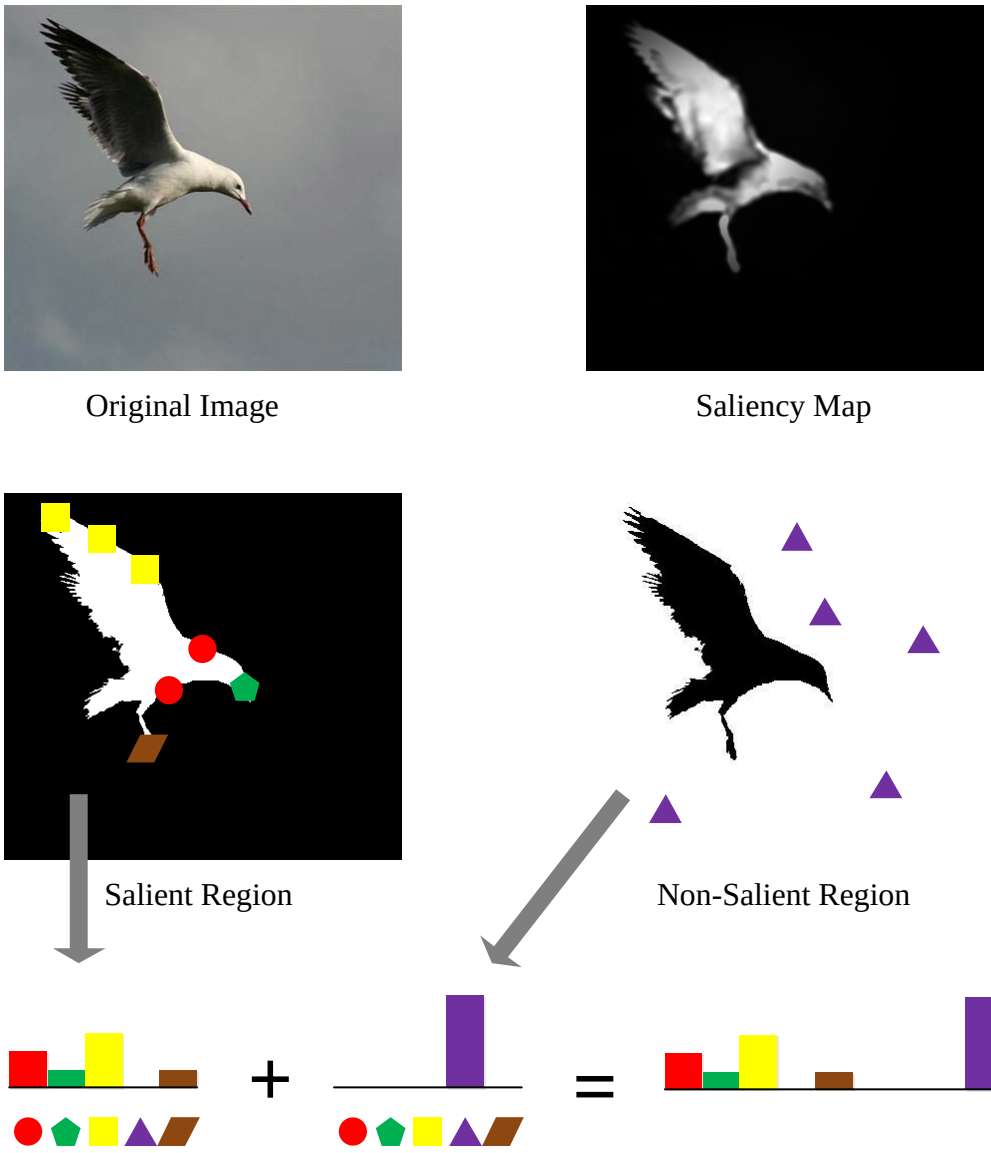


Fig.4. The diagram of our SSP. The colored shapes denote codewords.

Finally, max pooling is used to obtain the final coding responses for each local descriptor.

3.3 Saliency detection based spatial pooling

Current state-of-the-art image classification systems are usually using spatial pyramid matching (SPM) to incorporate the spatial information, in which pools low-level image features over pre-defined coarse spatial bins, such as three levels of 1×2, 2×2, and 4×4. In this paper, we propose a saliency detection based spatial pooling (SSP) approach for image classification. In contrast to SPM pooling, our SSP first extracts the interest object in an image by saliency detection in [34], then pools the coding responses obtained in the previous subsection separately in the salient region (interest object) and the non-salient region (background) to form the image-level representation with spatial information, which is shown in Fig.4. Obviously, our SSP tends to produce more consistent image representation than SPM pooling.

4 Experimental Result

This experiment aims to verify that i) the structured dictionary can improve the classification performance; ii) the proposed fuzzy reasoning based salient coding can produce comparable or even better performance than LLC, GSC, LSC, which are wildly used or the state-of-the-art; iii) the proposed SSP can perform better than SPM. We choose LLC, SaC, GSC, and LSAC for comparison. Note that all of them are efficient coding schemes. The following three datasets are used for test: Caltech 101, Scenes 15, and UIUC Sport. Some example images of these three datasets are shown in Fig.5. We first study the

Caltech 101: cup

Caltech 101: wild cat

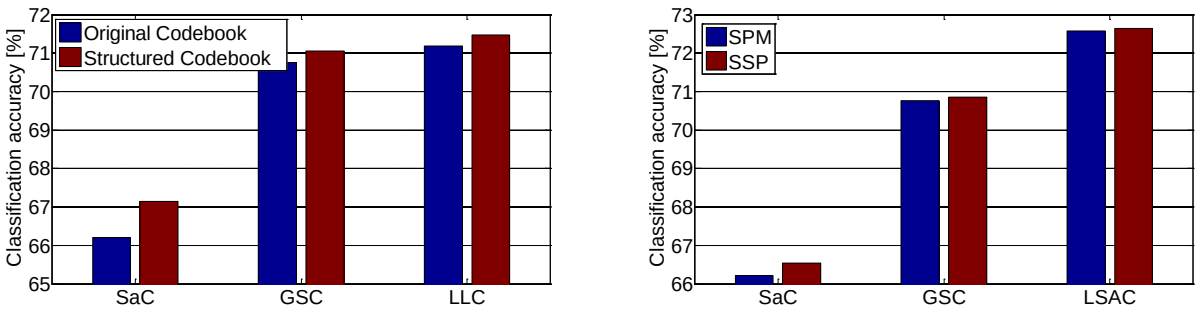
Scenes 15: suburb

Scenes 15: office

UIUC Sport: croquet

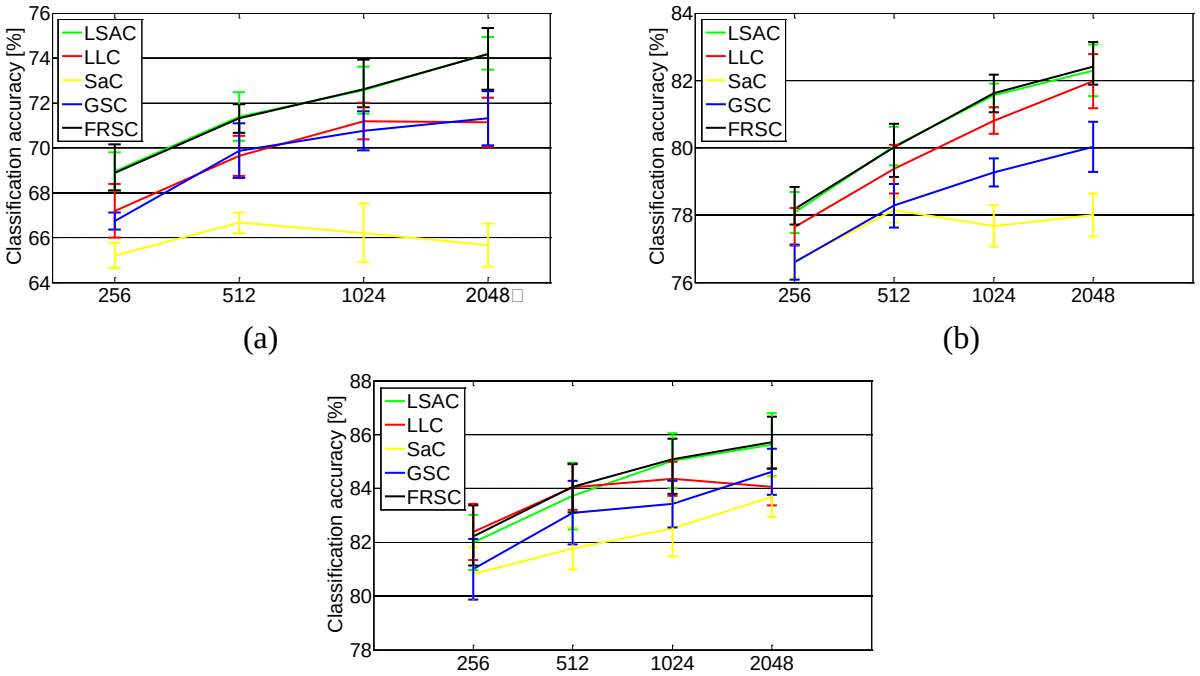
UIUC Sport: sailing

Fig.5 Some example images of the test datasets.



(a) (b)

Fig.6 Comparison between (a) original codebook and structured codebook in different coding schemes; (b) SPM and SSP under different coding methods with the original codebook.



(a) (b)

(c)

Fig.7 Performance comparison of various coding strategies (a) under different codebook size with original codebook and SPM on the Caltech 101 dataset; (b) under different codebook size with structured dictionary and SSP on the Scenes 15 dataset; (c) under different codebook size with structured dictionary and SSP on the UIUC Sport dataset.

proposed method in the Caltech 101 dataset with an in-depth analysis, including different codebook, coding and spatial pooling, and then combine them together in the other datasets. For fair comparison, all the tested coding methods are implemented in a unified framework. Thus, the consistency of all the configurations other than the coding part can be guaranteed. In our framework, SIFT descriptor is extracted from images on a grid with step size of 6 pixels under 16×16 scale. Codebook is generated by standard K-means clustering, wherein the subset of descriptors used for clustering is randomly sampled from all descriptors in the database. Lib-linear SVM is adopted for efficient classification, wherein the penalty coefficient is set to 1 as most methods did. As the other methods did, we repeat the experiment 10 times with different training and testing sets, then report the average accuracy and the standard deviation as the results. All the tests are conducted in a computer with an Intel Core 2 Duo 1.83 GHz CPU and 2GB RAM.

4.1 Results on the three datasets

Caltech 101: This is wildly used dataset with 9,144 images in 102 classes including a background class, which contains animals, vehicles, flowers, etc., and

with high intra-class appearance shape variability. Each category contains images from 31 to 800. The average image resolution is 300×300 pixels. In this dataset, codebook size is set as 1024 when comparing different part. Fig.6(a) shows the performance of different coding schemes with original codebook and structured codebook respectively. As shown, the structured codebook is slightly better than the original codebook. The classification performance between SSP and SPM under different coding methods with original codebook is shown in Fig.6(b), in which we also find that our SSP is slightly better than SPM. Then we show the results of various coding strategies under different codebook size with original codebook and SPM in Fig.7(a). The proposed FRSC almost performs the same with LSAC, but outperforms the other three coding schemes. We further make a comparison on the computation cost, which is shown in Table 2. We can see that the proposed FRSC is slightly faster than LSAC.

Table 2 Computation cost on Caltech 101 per image.

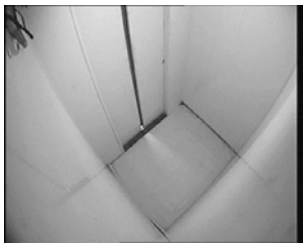
Method	LLC	SaC	GSC	LSAC	FRSC
ms	103	5	32.6	14.2	11.9

Scenes 15: It is a natural scene dataset including 4,485 images fallen into 15 scene categories, the number of image per class ranges from 200 to 400. It contains bedroom, suburb, industrial, kitchen, living room, coast, forest, highway, inside city, mountain, open country, street, tall building, office, and store. Following the standard setting, 100 images per class are used for training and the rest for testing. Combing our structured codebook, fuzzy reasoning coding, and SSP together, classification accuracy under different codebook size is compared in Fig.7(b). As seen, our FRSC is slightly better than LSAC when with large (2048) codebook size, and outperforms the others.

UIUC Sport: It is a sport event dataset, which contains 8 categories including badminton, bocce, croquet, polo, rock climbing, rowing, sailing, and snowboarding. It contains 1792 images and the size of each class varies from 137 to 250. The image resolution is higher than the above two datasets. We randomly choose 70 images for training and the remainder for testing. Comparison results are shown in Fig.7(c). Again, the proposed FRSC gets best performance, although it still obtains similar result with LSAC.

4.2 Application in elevator videos

Video event classification is also an important computer vision problem. In this paper, we further extend our FRSC into event classification in elevator video. We first select some videos from elevator, including empty, full loading, violence. Some example video frames are shown in Fig.8. Each video event contains 120 frames. In our experiment, half of the frames in each video are used for training and the other half for testing. SIFT descriptors are extracted for each frame on a dense grid, every 4 pixels and for 5 scale levels. To incorporate spatial information, the linear version of SPM kernel with three levels of 1×1, 2×2, and 4×4 is adopted. Codebook size is set to 256, which is produced by k-means. Finally, Lib-linear SVM is adopted for event classification. The classification results are shown in Table 3. As seen, our FRSC based method achieves good performance in such simple video events, which shows potential application in intelligent elevator video surveillance, such as overload detection, violence detection, video summarization.



(a) Empty

(b) Full loading

(c) Violence

Fig.8 Some example video frames.

Table 3 Classification accuracy of elevator video event classification.

Event	Accuracy
Empty	100%
Full loading	99%
Violence	98.6%

5 Conclusion

To alleviate the ambiguity and non-robustness problem in saliency based coding, an improved salient coding named FRSC was proposed in this

paper. First, we introduce efficient saliency detection method and use the generated saliency maps to measure the saliency degree of each image, then divide each image into salient and non-salient region to get a structured dictionary. Then, FRSC was proposed to improve the stability by introducing fuzzy reasoning rules. Finally, a saliency detection based spatial pooling scheme was proposed to incorporate spatial information to obtain a more compact image representation. Experiment on several common used datasets demonstrated the effectiveness of the proposed coding approach. At the same time, our method is more efficient than the reconstruction based coding schemes. We further apply it into elevator video event classification and achieved good performance, which demonstrate the potential application in intelligent elevator video surveillance, such as overload detection, violence detection, video summarization. Our future works will focus on conducting extensive experiment on more complicated elevator video events classification.

Acknowledgement

The authors would like to express their sincere thanks to the anonymous reviewers for their invaluable suggestions and comments to improve this paper. This work is supported by science and technology planning project of chongqing bureau of quality and technology supervision.

References:

[1] Z. Wu, Q. Ke, M. Isard, J. Sun, Bundling features for large scale partial-duplicate web image search, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp.25–32.

[2] Jiang Yu-Gang, Ngo Chong-Wah, Yang Jun, Towards optimal bag-of-features for object categorization and semantic video retrieval, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, 2007, pp.494-501.

[3] R. Collins, A. Lipton, T. Kanade, H. Fujuyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, A system for video surveillance and monitoring, technical report: CMU-RI-TR-00-12, Pittsburgh, PA, 2000.

[4] V. I. Pavlovic, R. Sharma, T. S. Huang, Visual interpretation of hand gestures for human-computer interaction: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, 1997, pp. 677–695.

[5] R. Kosala, H. Blockeel, Web mining research: A survey, ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, 2000, pp. 1–15.

[6] A. K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 1, 2004, pp.4–20.

[7] G. Csurka, C. Bray, C. Dance, L. Fan, Visual categorization with bags of keypoints, in: European Conference on Computer Vision, 2004.

[8] S. Lazebnik, C. Schmid, J. Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[9] Http://www.vision.caltech.edu/Image_Datasets /Caltech101/.

[10] Http://www.cs.unc.edu/-lazebnik/research/scenecategories.zip/, 2006.

[11] Http://www.vision.caltech.edu/Image-Datasets/Caltech256/.

[12] Http://pascallin.ecs.soton.ac.uk/challenges/voc, 2005-2012.

[13] C. Harris, M. Stephens, A combined corner and edge detector, in: Proceedings of the Fourth Alvey Vision Conference, 1988, pp.147–151.

[14] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, International Journal of Computer Vision, vol. 60, no. 1, 2004, pp.63–86.

[15] D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, vol. 60, no. 2, 2004, pp.91-110.

[16] B. T. Navneet Dalal, Histograms of Oriented Gradients for Human Detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp.886-893.

[17] S. P. Lloyd, Least squares quantization in PCM, IEEE Transactions on Information Theory, vol. 28, no. 2, 1982, pp.129–137.

[18] C. Cortes, V. Vapnik, Support-vector network, Machine Learning, 1995, pp.273–297.

[19] Http://www.csie.ntu.edu.tw/~cjlin/liblinear/.

[20] Y. Huang, Z. Wu, L. Wang, T. Tan, Feature Coding in Image Classification: A Comprehensive Study, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, 2014, pp.493-506.

[21] J. C. Gemert, J. Geusebroek, C. J. Veenman, A. W. M. Smeulders, Kernel Codebooks for Scene

Categorization, in: European Conference on Computer Vision, 2008, pp.696-709.
[22] L. Liu, L. Wang, X. Liu, In Defense of Soft-assignment Coding, in: International Conference on Computer Vision, 2011, pp.2486-2493.
[23] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1794-1801.
[24] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp.3360-3367.
[25] S. Gao, I. W. Tsang, L. Chia, Laplacian Sparse Coding, Hypergraph Laplacian Sparse Coding, and Applications, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.35, no.1, 2013, pp.92-104.
[26] F. Perronnin, J. Sanchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: European Conference on Computer Vision, 2010, pp.143-156.
[27] X. Zhou, K. Yu, T. Zhang, T. S. Huang, Image classification using super-vector coding of local image descriptors, in: European Conference on Computer Vision, 2010, pp.141-154.
[28] Y. Huang, K. Huang, Y. Yu, T. Tan, Salient coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp.1753-1760.
[29] Z. Wu, Y. Huang, L. Wang, T. Tan, Group Encoding of Local Features in Image Classification, in: International Conference on Pattern Recognition, 2012, pp.1505-1508.
[30] A. Shaban, H. R. Rabiee, M. Farajtabar, M. Ghazvininejad, From Local Similarity to Global Coding; An Application to Image Classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp.2794-2801.
[31] H. Lee, B. Alexis, R. Rajat, Ng. Andrew Y, Efficient sparse coding algorithms, in: Conference on Neural Information Processing Systems, 2006, pp. 801-808.
[32] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, S. Ma, Image Classification by Non-Negative Sparse Coding, Low-Rank and Sparse Decomposition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp.1673-1680.
[33] L. Zhang, C. Ma, Low-rank decomposition and Laplacian group sparse coding for image classification, Neurocomputing, 2014.
[34] S. Chen, W. Shi, W. Zhang, Visual saliency detection via multiple background estimation and spatial distribution, Optik, vol. 125, no. 1, 2014, pp.569-574.
[35] L.X. Wang, Adaptive Fuzzy Systems and Control: Design and Stability Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1994.
[36] Reza Tavoli, Classification and Evaluation of Document Image Retrieval System, WSEAS Transactions on Computers, vol. 11, no. 10, 2012, pp.329-338.
[37] Mahmoud Al-Ayyoub, Duaa Alawad, Khaldun Al-Darabsah, Inad Aljarrah, Automatic Detection and Classification of Brain Hemorrhages, WSEAS Transactions on Computers, vol. 12, no. 10, 2013, pp.395-405.