Recognition of Underwater Creatures Using SIFT and Bag-of-words Model

HAN Yechen, Ritsumeikan University, gr0119xr@ed.ritsumei.ac.jp
Shinichi HIRAI, Ritsumeikan University, hirai@se.ritsumei.ac.jp

With the rapid improvement of the computation performance recently, pattern recognition has entered the generation of big data. Many different kinds of features and machine learning model have already been applied on the implementation based on big data. This paper proposed an approach of underwater creature recognition using SIFT feature and Bag-of-words model. SIFT feature, which considered to be a robust local feature, is introduced into deal with the deformation of underwater creatures. For machine learning and classification, Bag-of-words model is used in this paper to create the classifier, in a form of histogram. Evaluation experiments are executed and proved the effect of the system implemented in this paper.

Key Words**:** computer vision, machine learning, big data, SIFT feature, Bag-of-words model

1. Introduction

Object recognition is a long lasting topic for computer vision, or pattern recognition. In the past, because of the limitation of computation performance, the former methods for this task used a designed target or a given template. The features detected from the template lead us to recognize them from other images or videos. The disadvantage is quite obvious that we could recognize only a single class of objects, or only several classes.

Recently, with the explosion of information sets, rapid improvement of computation performance such as GPGPU, parallel computation, cloud computing, big data method has been widely used in all kinds of applications: data mining and searching, artificial intelligence, air traffic control, city planning, etc. In computer vision, the applications of big data are also commonly used.

Andre, et al. proposed an approach using big data to improve the content-based image retrieval (CBIR) technology and made some extension by introducing visual similarity [1]. In their approach, Bag-of-words (BoW) model for the visual significance.

Pavlov, et al. developed a video analysis system to deal with face detection, face tracking, and gender recognition. They utilized AdaBoost Classifier for the face detection and support vector machine (SVM) to deal with the tracking [2].

Ukwatta et al. developed a vision based metal spectral system using multi-label classification. They applied both SVM and Artificial Neural Network (ANN) and yield a correct identification of metals to an accuracy of 99% [3].

In this paper, we proposed an approach to recognize the underwater creatures. Which would probably be used for the ecological survey or the aid of manipulation of underwater robot.

2. Construction of system

Basic flow of recognition based on big data has two phases, training phase and recognition phase. In this paper, we used SIFT feature to do this training and used BoW model to create the classifier for the recognition as shown in Fig.1.

Firstly, a lot of training data should be prepared. These data could be images and videos. Then we detect SIFT features from all these training data. All the SIFT features would be recorded into files. After applied to BoW model, we will create the histogram for each class of underwater creature. (Section 2.1) These histograms would be used in the recognition phase as classifier. When it goes to the recognition phase, again we detect SIFT features from the input data (images or videos), and apply BoW model to the input data to create the histogram for them. Finally, we compare the histogram of input data with the classifier histograms to recognize the targets in the input data.

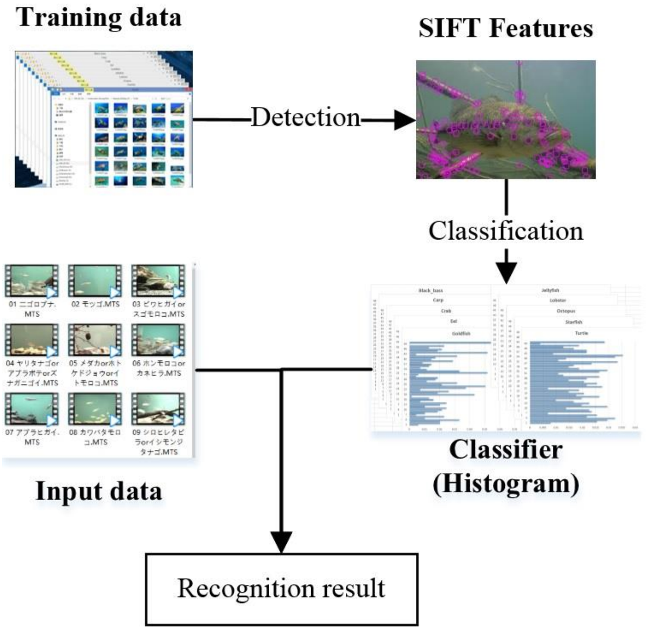
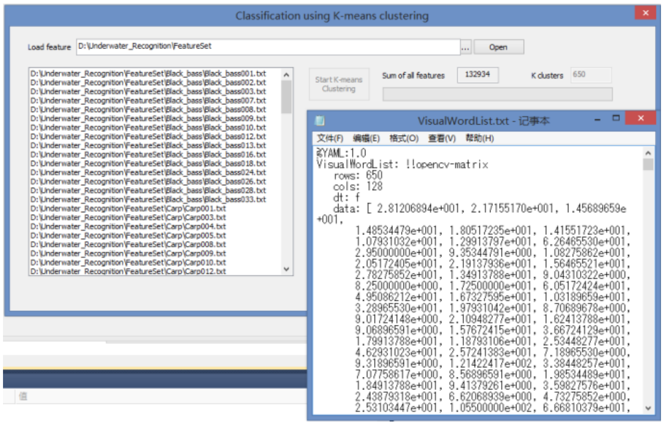


Fig. 1 Flow of basic recognition machine learning

2.1 BoW model

Bow model is a statistical framework which is originally used in the document categorization. Suppose there are N words in all documents represented by $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$. Then we count the frequency of each word that appears in each document and get their frequency as another vector with N elements. Thus each document could be represented by an N-dimensional vector $\boldsymbol{p} = \{p_1, p_2, \ldots, p_N\}$. The distance between vector $\boldsymbol{p}$ of different documents represent their similarity.

When this model is applied into computer vision, we use the definition of visual words instead of the text words in BoW model. Because the visual words are not exact things as the text words in documents, a clustering should be implemented first to create the visual word list. In this paper, we used the widely used K-means method to do this clustering. Considering the volume of the database, we set the K value about 1/200 of the number of features, which works well at present.


Classification using K-means clustering
...
Open
Start K-means Clustering
VisualWordList.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

Load feature D:\Underwater_Recognition\FeatureSet

Sum of all features 132934

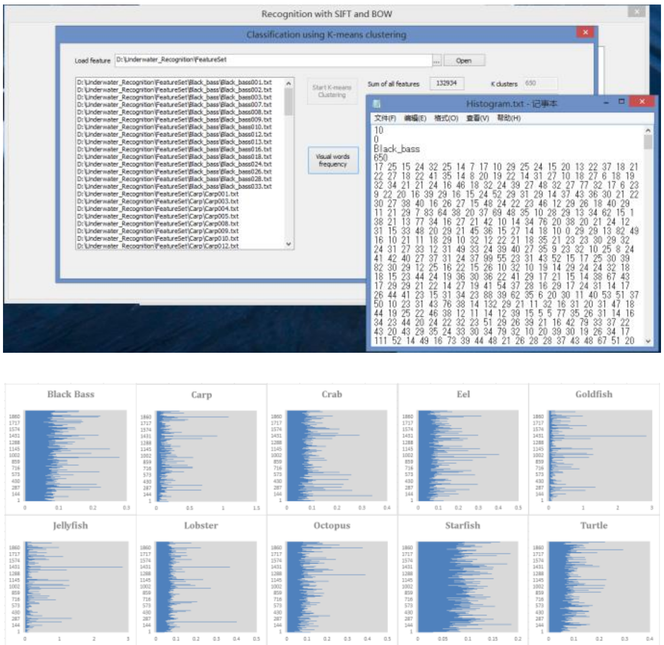
K clusters 650

```
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass001.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass002.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass003.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass007.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass008.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass009.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass010.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass012.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass013.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass016.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass018.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass024.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass026.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass028.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass033.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp001.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp003.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp004.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp005.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp008.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp009.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp010.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp012.txt
```

```
%YAML:1.0
VisualWordList: !!opencv-matrix
   rows: 650
   cols: 128
   dt: f
   data: [ 2.81206894e+001, 2.17155170e+001, 1.45689659e
+001,
       1.48534479e+001, 1.80517235e+001, 1.41551723e+001,
       1.07931032e+001, 1.29913797e+001, 6.26465530e+001,
       2.95000000e+001, 9.35344791e+000, 1.08275862e+001,
       2.05172405e+001, 2.19137936e+001, 1.56465521e+001,
       2.78275852e+001, 1.34913788e+001, 9.04310322e+000,
       8.25000000e+000, 1.72500000e+001, 6.05172424e+001,
       4.95086212e+001, 1.67327595e+001, 1.03189659e+001,
       3.28965530e+001, 1.97931042e+001, 8.70689678e+000,
       9.01724148e+000, 2.10948277e+001, 1.62413788e+001,
       9.06896591e+000, 1.57672415e+001, 3.66724129e+001,
       1.79913788e+001, 1.18793106e+001, 2.53448277e+001,
       4.62931023e+001, 2.57241383e+001, 7.18965530e+000,
       9.31896591e+000, 1.21422417e+002, 3.38448257e+001,
       7.07758617e+000, 8.56896591e+000, 1.98534489e+001,
       1.84913788e+001, 9.41379261e+000, 3.59827576e+001,
       2.43879318e+001, 6.62068939e+000, 4.73275852e+000,
       2.53103447e+001, 1.05500000e+002, 6.66810379e+001,
```

Fig. 2 Visual word list created by clustering, where *K*=650

Therefore there is a set of visual words ***VW*** = {***vw***$_1$, ***vw***$_2$, ***vw***$_3$, …, ***vw***$_K$}, in which each element is a 128-dimensional vector, having same length with SIFT features. Then we scan all the SIFT features again to find the most close visual word and count the number. After fitting all the SIFT features of a single class of underwater creature to corresponding visual words, each class could be represented by a histogram with *K* 128-dimensional vectors. All the histograms are recorded into a file.



Load feature D:\Underwater_Recognition\FeatureSet

Sum of all features 132934

K clusters 650

```
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass001.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass002.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass003.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass007.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass008.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass009.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass010.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass012.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass013.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass016.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass018.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass024.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass026.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass028.txt
D:\Underwater_Recognition\FeatureSet\Black_bass\Black_bass033.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp001.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp003.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp004.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp005.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp008.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp009.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp010.txt
D:\Underwater_Recognition\FeatureSet\Carp\Carp012.txt
```

```
10
0
Black_bass
650
17 25 15 24 32 25 14 7 17 10 29 25 24 15 20 13 22 37 18 21
22 27 18 22 41 35 14 8 20 19 22 14 31 27 10 18 27 6 18 19
32 34 21 21 24 16 46 18 32 24 39 27 48 32 27 77 32 17 6 23
9 22 20 16 39 29 16 15 24 52 29 31 29 14 37 43 36 30 21 22
30 27 38 40 16 26 27 15 48 24 22 23 46 12 29 26 18 40 29
11 21 29 7 83 64 38 20 37 69 48 35 10 28 29 13 34 62 15 1
38 21 13 77 34 16 27 21 42 10 14 34 76 20 38 20 21 24 12
31 15 33 48 20 29 21 45 36 15 27 14 18 10 0 29 29 13 82 49
16 10 21 11 18 29 10 32 12 22 21 18 35 21 23 23 30 29 32
24 31 27 33 12 31 49 33 24 39 40 27 35 9 23 32 10 25 8 24
41 42 40 27 37 31 24 37 99 55 23 31 43 52 15 17 25 30 39
82 30 29 12 25 16 22 15 26 10 32 10 19 14 29 24 24 32 18
18 15 23 44 24 19 36 30 36 22 41 29 17 21 15 14 38 67 43
17 29 29 21 22 14 27 19 41 54 37 28 16 29 17 24 31 14 17
26 44 41 23 15 31 34 23 88 39 62 35 6 20 30 11 40 53 51 37
50 10 23 31 43 76 38 14 132 29 21 11 32 16 31 20 31 47 18
44 19 25 22 46 38 12 11 14 12 39 15 5 5 77 35 26 31 14 16
34 23 44 20 24 22 32 23 51 29 26 39 21 16 42 79 33 37 22
43 20 43 29 35 24 33 30 34 79 32 10 20 39 30 19 26 34 17
111 52 14 49 16 73 39 44 48 21 26 28 28 37 43 48 67 51 20
```

Fig. 3 Histogram (classifier) of underwater creatures

2.2 Recognition

This phase starts quite similar with the training phase. For a single image or a frame of an input video, we detect all the SIFT features. Then every feature would be compared with visual word list to find the most fit one and count the frequency. Thus we could create the histogram for each input image. Then we calculate the distance between the histogram of input images and that used as classifier. One thing that should be pay attention to is that until now the histogram is the count of visual words, so normalization here is required to turn the histograms from a vector of count to that of frequency.

Another important thing is that, with the increase of features, the number of visual words also increase because we set the number of visual word 1/200 of that of features. This makes the histogram quite long and the search and recognition cost much time. A hash table is applied here trying to save some time. A rapid matching approach would be one of our future works.

3. Evaluation experiments

The evaluation experiments in this paper focused chiefly on the accuracy of recognition. In this paper we collected 289 images of 10 classes of underwater creatures from internet. These 10 classes of underwater creatures are: black bass, carp, crab, eel, goldfish, jellyfish, lobster, octopus, starfish, and turtle. By modifying the resolution of images, the number of features detected could be from tens of thousands to nearly one million. All the experiments were executed on a computer with Intel I5-4460 @ 3.2GHz and 32GB memory. The algorithm is implemented under Windows 7 operating system and Visual Studio 2010, combined with OpenCV 2.4.9 and Rob Hess's free library OpenSIFT.

3.1 Basic accuracy experiment

For this basic accuracy experiment, we modify the resolution of images to control the number of SIFT features. We then observe the accuracy and time consuming. Detail is shown in Table 1.

Table 1 Accuracy experiment using various numbers of features

	Test1	Test2	Test3	Test4
Feature number	58387	217071	217071	448624
Visual word number	300	500	1000	2000
Training time	10 min	27 min	62 min	10 hours
Accuracy	45.7%	46.1%	54.1%	65.41%

This table showed that with the increase of features and visual words, higher accuracy could be achieved, with a cost of higher time consuming. For test1, test3 and test4, the number of visual words is set to about 1/200 of the number of features. The number of features and that of visual words increased simultaneously brought us an increase of accuracy. For test2, we reduce the number of visual word to 1/400 of the number of the feature. Although the number of features is four times than that of test1, but the accuracy does not have an obvious increase. Thus we supposed that the rate of visual words number to feature number set to 1/200 is suitable.

3.2 Details for test4

This analysis is trying to find some further information from test4. First of all, the detail accuracy information is listed in Table 2.

Table 2 Detailed accuracy information of test4

Class	Black bass	Carp	Crab	Eel	Goldfish
Accuracy	75.00%	71.43%	68.97%	22.73%	29.41%
Class	Jellyfish	Lobster	Octopus	Starfish	Turtle
Accuracy	45.45%	88.89%	86.96%	84.62%	80.65%

It is obvious that the accuracy of the recognition of eel, goldfish and jellyfish is lower than the other classes. For detailed investigation, we create the confusion matrix for the recognition of all the ten classes as shown in Table 3.

This confusion matrix is not symmetrical as we expected, which means that even a class A is recognized as class B by mistake, the reverse process does not establish. For example, class Eel is quite often recognized as Goldfish or Turtle, with an accuracy of 22.73% to 18.18%. But Goldfish was NOT recognized as Eel by mistake. Turtle has a probability of 3.23% been recognized as Eel but this error is quite small compared with the 80.65% accuracy.

4. Conclusion

In this paper, we implemented an approach of recognition of underwater creatures using big data method, combined with SIFT feature and BoW machine learning model. The evaluation experiments showed that for our training data, some of the classes could be recognized with a high accuracy.

Reference

[1] Andre B., Vercauteren T., Buchner A. M., Wallace M.B., Ayache N., “Learning Semantic and Visual Similarity for Endomicroscopy Video Retrieval”, IEEE Transactions on Medical Imaging, volume: 31, pp.1276-1288, 2012.

[2] Pavlov V., Khryashchev, V., Shmaglit, L., Pavlov E., “Application for video analysis based on machine learning and computer vision algorithms”, 14th Conference of Open Innovations Association (FRUCT), pp.90-100, 2013.

[3] Ukwatta E., Samarabandu J., “Vision Based Metal Spectral Analysis using Multi-label Classification”, Canadian Conference on Computer and Robot Vision (CRV), pp.132-139, 2009.

[4] D.G. Lowe, “Distinctive Image Features from Scale Invariant Keypoints,” Int’l J. Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[5] http://robwhess.github.com/opensift/

[6] D.FILLIAT, “A visual bag of words method for interactive localization and mapping”, IEEE International Conference on Robotics and Automation, pp.3921-3926. 10-14 April, 2007.

[7] Rusinol, M., Llados, J. , “Logo Spotting by a Bag-of-words Approach for Document Categorization”, 10th International Conference on Document Analysis and Recognition, pp.111-115, 2009.

[8] Zisheng Li, Imai J., Kaneko, M., “Robust Face recognition Using Block-based Bag of Words”, 20th International Conference on Pattern Recognition (ICPR), pp.1285-1288, 2010.

[9] http://docs.opencv.org/modules/refman.html

Table 3 Confusion matrix

	Black bass	Carp	Crab	Eel	Goldfish	Jellyfish	Lobster	Octopus	Starfish	Turtle
Black bass	75.00%	0.00%	7.14%	10.71%	0.00%	0.00%	0.00%	0.00%	0.00%	7.14%
Carp	0.00%	71.43%	0.00%	4.76%	9.52%	0.00%	4.76%	9.52%	0.00%	0.00%
Carb	0.00%	0.00%	68.97%	3.45%	0.00%	10.34%	3.45%	0.00%	13.79%	0.00%
Eel	0.00%	4.55%	4.55%	22.73%	18.18%	13.64%	9.09%	9.09%	0.00%	18.18%
Goldfish	0.00%	0.00%	5.88%	0.00%	29.41%	11.76%	17.65%	11.76%	23.53%	0.00%
Jellyfish	0.00%	9.09%	9.09%	18.18%	18.18%	45.45%	0.00%	0.00%	0.00%	0.00%
Lobster	0.00%%	0.00%	5.56%	5.56%	0.00%	0.00%	88.89%	0.00%	0.00%	0000%
Octopus	4.35%%	4.35%	0.00%	0.00%	0.00%	0.00%	4.35%	86.96%	0.00%	0.00%
Starfish	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	15.38%	0.00%	84.62%	0.00%
Turtle	3.23%	3.23%	0.00%	3.23%	0.00%	0.00%	0.00%	6.45%	3.23%	80.65%