Real-time Fish Localization with Binarized Normed Gradients

Xiu Li [1, 2], Jing Hao [1, 2], Hongwei Qin [1, 2], Liansheng Chen [1, 2]
1. Department of Automation, Tsinghua University, Beijing 100084
2. Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055
li.xiu@sz.tsinghua.edu.cn, hao-j14@mails.tsinghua.edu.cn
qhw12@mails.tsinghua.edu.cn, cls13@mails.tsinghua.edu.cn

***Abstract*—Fast and accurate fish localization is an important step for fish detection, identification, counting and tracking. In this paper, we introduce how to localize the fish with an efficient way, which can capture almost all fish locations in an image. First, we exploit the normed gradients (NG) feature of 8×8 image windows to discriminate the fish from the background, and then we binarize the NG feature to accelerate the fish localization. As there is no existing appropriate dataset, we make a dataset of underwater imagery to achieve fish localization. The dataset contains 9,963 images of underwater videos for training, validation and testing. The details about how to label the fish of this dataset further be showed. Last, we evaluate our method on this dataset. Experiments show that our method is fast and efficient, and fish localization takes only about 0.00234 sec. per image (400 fps on an Intel i5-3540 CPU) and achieves 97.1% recall with 1000 proposals. This method satisfies computational efficiency and high detection rate simultaneously.**

Keywords—fish localization; real-time; underwater videos; Binarized Normed Gradients

I. INTRODUCTION

The advancements of the underwater video system, such as NEPTUNE and VENUS observatories [1], makes monitoring rich ocean resources a reality [4]. It is a powerful way for marine biologists to do related research such as the census and behavior observation of ocean species. However, massive video information makes it difficult to obtain the useful information efficiently. So fast, automatic and effective video analysis becomes increasingly important.

As one of important ocean species, the fish gets the marine biologists' attention. Main studies of the fish in image or video processing and computer vision areas are fish detection, recognition, tracking, counting and behavior analysis. It is clear that fish detection is the foundation of other tasks and has drawn wide attention in recent years. At the present, a lot of object detection methods [2, 3, 5] are proposed. For object detection methods, high detection rate and computational efficiency are two most important measurements. In this paper, we show a surprisingly fast and high quality way to locate the fish in an image.

As there is no existing appropriate dataset of underwater imagery, we need to make a dataset by ourselves. The dataset contains 9,963 images with 2,501 images for training, 2,510 images for validation and 4,952 images for testing. And these images all come from underwater videos. We follow the pipeline of the dataset production:

- convert the video into images, as adjacent frames are very similar, we keep one of them and delete the others of those almost same images;
- label the fish which are stand out from the background with the bounding boxes for 5,011 images for training and validation;
- save the information of bounding boxes to XML file, resulting in a total of 5,011 XML files.
- transfer XML file to YML file, because OpenCV (Open Source Computer Vision Library) can only call YML file, thus we have 5,011 YML files.

The images abstracted from underwater videos have some characteristics, such as poor image quality, blurred, crowded (a school of fish can be in an image), and camouflage fish (background and fish have similar colors or textures). Some images of underwater videos are showed in Figure 1.

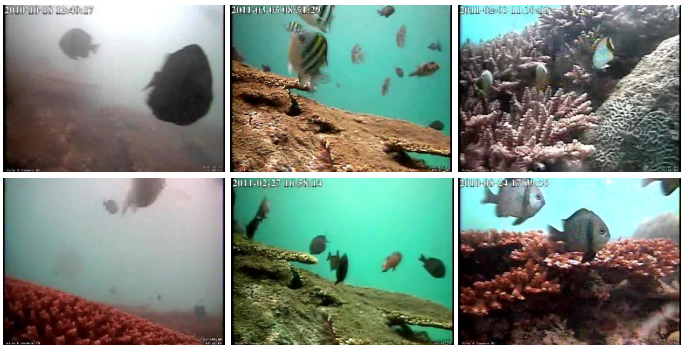

Fig. 1. Some examples of our underwater images dataset. The first column shows the low contrasted (blurred) images; the second column shows the crowded images; and the third column shows that background and fish have similar colors.

In this paper, we extend a simple and efficient method [1] to achieve the fish localization. The structure of this paper is organized as follows:

In Section Ⅱ, we introduce the fish localization method based on binarized normed gradients and show the details of the dataset making rules. We evaluate our method on the dataset, yielding high fish detection rate, and compare the

[1] *http://www.oceannetworks.ca/*

computational time with other popular methods [2,3] in Section Ⅲ. Finally, in Section Ⅳ, we conclude our work.

Our work has two major contributions:

First, we make a dataset including 9,963 images of the underwater videos, labeling 5,011 images with 2,501 images for training and 2,510 images for validation. Some examples of our labeling images is showed in Figure 2.

Second, we extend a simple and efficient model [1] to achieve fish localization, achieving the high computational efficiency and detection rate.

II. METHODOLOGY

Inspired by [1], we use the binarized normed gradients (NG) features to discriminate the fish, capturing all possible fish locations in an image. In this section we detail our fish localization methods in two parts:

- the pipeline of fish location method;
- the making rules and the strategies of the dataset production process.

A. Fish Localization

This paper aims to capture all possible fish locations efficiently, yielding real-time fish localization. However, the classical sliding window fish detection paradigm has a high time and space complexity. A good object detection model should be accelerate the evaluation processes without decreasing the accuracy and detection rate [5]. Based on this, we extent a surprisingly powerful way [1] to accelerate the classical sliding window paradigm, resulting in the time of fish localization only 0.00234 sec. per image, and achieve high detection rate simultaneously, yielding 97.1% fish detection rate with 1000 proposals.

Moreover, the recall can be improved with the increasing of the number of proposals and color spaces. When we increase the proposals to 5,000 and use 3 different color spaces [1], the detection rate can be improved to 99.8%.

The main idea of this method is as follows :

Training. We resize the source image to 9 different quantized sizes according to the characteristics of the dataset, and then calculate respectively normed gradients for each resized image. In order to reduce the computational time, we binarize the NG feature to help search for the fish proposals. By this way, we greatly reduce the time and space complexity. And then we can learn a model $w \in R^{8\times8}$ by cascaded SVM [1].

In our experiments, the number of quantized sizes is 9, because the average computational time is lowest without decreasing detection rate with 9. The details of experiments' results with different quantized sizes are showed in Tab. I.

There are 2,501 images for training procedure. Figure 3 shows the pipeline of fish localization, and the training stage is from a) to c).

Validation. In this stage, we fine-tune the model obtained in training process using 2,510 validation images. Then we obtain the modified model, which is used in testing stage to predict the fish locations.

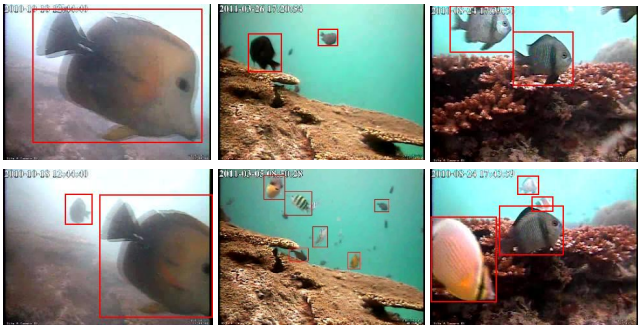

Fig. 2. Some examples of our labeling images used in training and validation stage. The first column shows the shape of fish has two kinds: complete or incomplete; the second column shows the number of fish is uncertain, and some fishes have hazy outline; and in the last column, the fish and the background have a big overlap.

Testing. We use the obtained model w to generate a small set of proposals with the same method. In this process, the candidate windows are ordered by scores. Then we choose the top 1,000 proposals as the effective proposals. We can calculate average detection rate by 4,952 testing images with 1000 proposals per image. The details is showed in Section Ⅲ.

The flowchart of our fish localization procedure is shown in Fig. 3.

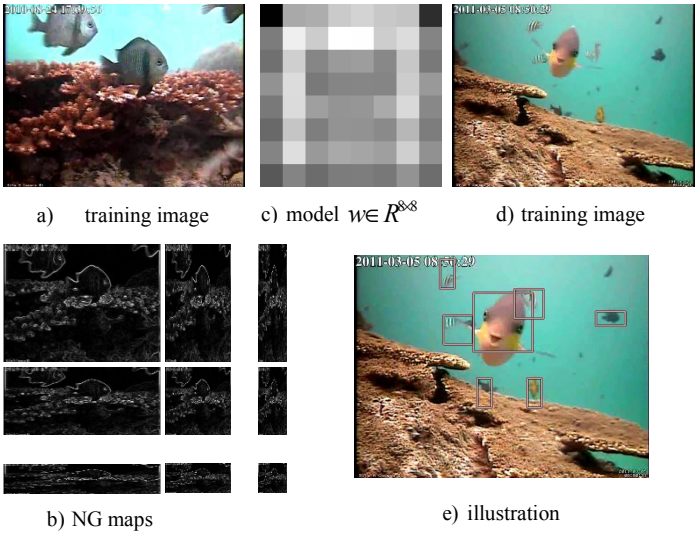

Fig. 3. The procedure of fish localization is from a) to e): calculating NG maps of training images a), and resizing the NG map to 9 different sizes, resulting in b), then training the model c) based on the NG features, finally, we can predict the fish locations e) for testing images d).

B. Dataset

We created a new dataset of underwater imagery for our experiments. This dataset consists of 9,963 images with 2,501 images for training, 2,510 images for validation and 4,952 images for testing.

Part of this dataset are from underwater videos of Fish4Knowledge[2], and the other is from ImageCLEF[3]. The images' spatial resolution has two sizes, 320×240 and 640×480. We labeled the ground truth of fish locations for training and validation images in order to learn the model w and calculate

[2] *http://f4k.dieei.unict.it/datasets/bkg_modeling/.*
[3] *http://www.imageclef.org/lifeclef/2015/fish.*

the detection rate. The total of labeling images are 5,011 and we detail the dataset production rules in the follow.

In this paper, we can training a model $w \in R^{8\times8}$, a 64D NG feature, by ground truth of the fish locations; and in the testing stage, we use the ground truth to calculate the recall and accuracy, which can evaluate the performance of the model.

Because of the limit of underwater imaging equipment, luminosity variations and water properties for light, such as absorption, reflection and scattering, the quality of images in our dataset is poor, it is a challenge for fish labeling. In this way, we divide those images into four parts: fish integrity (whether the shape of fish is complete), fish quantity, the outline of fish, fish and background.

The rules of labeling the fish are as follows:

- Fish integrity

  The shape of fish on the edge of image may be not complete (see Fig. 4). In this case, if the texture or outline of fish is clear, such as the top line in Fig. 4, we label the fish, otherwise ignore them.

Fig. 4. The first line shows that the shape of fish is not complete but the texture or outline is obvious, in this way, we label them; and in second line, the fish is unclear in both shape and texture labeled in green oval, in this case we ignore them.

- Fish quantity

  In an image, the number of fish is uncertain (see Fig. 5). We label all the obvious fish which can test whether this method is good enough to capture all possible fish locations.

Fig. 5. The left image shows there are one fish and the right image have a school of fish.

- The outline of fish

  Because the quality of underwater videos is poor and the distance between fish and camera is far, the contour of the fish may not be clear. In this case, we ignore them unless the texture of fish is obvious (see Fig. 6).

Fig. 6. The outline and texture of fish in green oval of left image is unclear, in this way, we ignore it; and in right image, the texture of fish in green oval is obvious though the shape is blurred, we label it.

- Fish and background

  Sometimes, there is a big overlap between fish and background, or the fish and the background have similar colors or textures. Under the circumstance, we label all the fish that can be perceived (see in Fig. 7).

Fig. 7. The fish labeled in green oval is clear although the background and fish is similar in colors.

III. Experiments and Analysis

We evaluate the performance of our method based on binarized normed gradients with the dataset. In this section we shows the experimental details, results and analysis, and we compare our method with other popular methods [2, 3] in computational time by the dataset (see Tab. Ⅱ). We use the implementation in [1].

Experiments show that the number of quantified sizes can influence the performance of the model (see Tab. Ⅰ). The notation of the number of quantified sizes is *Nm*, and we use *Nm*=9 in our experiments. The average computational time and the recall are different with the different m, some details are showed in Table Ⅰ.

TABLE I. COMPUTIONAL TIME ON THE DATASET

Nm	***4***	***9***	***16***	***25***
Time (sec.)	0.00239	0.00234	0.00242	0.00814
Recall (%)	93.84	97.8	97.76	97.54

A. Computational time

We show the average computational time and make the comparison with other methods [2,3] by the dataset in Tab. Ⅱ. It is obvious that our method is much faster than others.

TABLE II. COMPUTIONAL TIME ON THE DATASET

Method	***Measure[2]***	***Selective[3]***	***Our method***
Time (sec.)	3.19	13.66	0.00234

As shown in Tab. Ⅱ, our method takes only about 0.00234 sec. per image with the dataset, that is, we can capture almost fish locations in one image at 400 fps, which is more than 1,200 times faster than [2, 3].

B. *Detection rate*

The recall means:

$$\text{Recall} = \frac{True\ possitive}{true\ possitive + false\ negtive}$$

The average recall is the ratio of true positive proposals to 1,000 proposals by 4,952 testing images with 1,000 proposals per image.

As demonstrated in Fig. 8, we show the trade-off between #proposal (the number of proposals) and Recall for the testing images, the recall is improved to 97.1% with 1000 proposals (see in Fig. 8). That is to say, we capture nearly all fish locations of the dataset with 1000 proposals per image. The recall can be approved into 99.8% when the number of proposals is 5,000 and in 3 different color spaces.

The illustration of the true positive fish proposals is shown in Fig. 9.

IV. Conclusion

We extend a surprisingly fast and efficient method, binarized normed gradients, to capture the fish locations with our dataset. Our method has a better performance and is more than 1000 times faster than [2,3], and the fish localization takes only

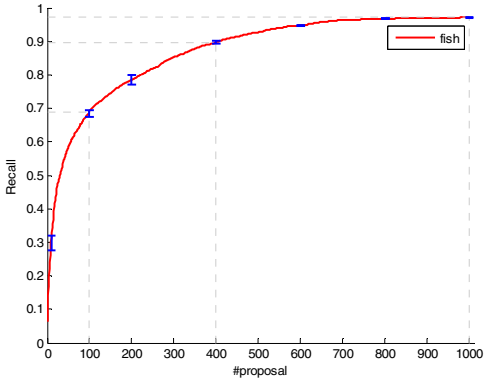

Fig. 8. Trade-off between #proposal and the fish recall. We can capture nearly all fish locations of this dataset.

0.00234 sec. per image. In this speed-up way, we can achieve fish localizations in real time. This method can also be applied to other real-time object estimation for underwater imagery.

Acknowledgment

The work in this paper is supported by National Natural Science Foundation of China (Grant No. 71171121/61033005) and National 863 High Technology Research and Development Program of China (Grant No. 2012AA09A408).

Fig. 9. Qualitative results for the fish localization in our testing images. We show the accuracy and integrity of the fish localization method.

References

[1] Cheng, Ming-Ming, et al. "BING: Binarized normed gradients for objectness estimation at 300fps." Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014.

[2] Alexe, Bogdan, Thomas Deselaers, and Vittorio Ferrari. "Measuring the objectness of image windows." Pattern Analysis and Machine Intelligence, IEEE Transactions on 34.11 (2012): 2189-2202.

[3] Uijlings, Jasper RR, et al. "Selective search for object recognition."International journal of computer vision 104.2 (2013): 154-171.

[4] Qin, Hongwei, Yigang Peng, and Xiu Li. "Foreground Extraction of Underwater Videos via Sparse and Low-Rank Matrix Decomposition."

Computer Vision for Analysis of Underwater Imagery (CVAUI), 2014 ICPR Workshop on. IEEE, 2014.

[5] Zhang, Ziming, Jonathan Warrell, and Philip HS Torr. "Proposal generation for object detection using cascaded ranking SVMs." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.