Object detection, classification, tracking and individual recognition for sea images and videos

Dávid Papp, Dániel Lovas, Gábor Szűcs

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Magyar Tudósok krt. 2., H-1117, Budapest, Hungary,

pappd@tmit.bme.hu, lovas.daniel@simonyi.bme.hu, szucs@tmit.bme.hu

Abstract. Manually monitoring the population displacement of fish species and the whale individuals is a painful and definitely unscalable process. Video data about fishes often require laborious visual analysis, moreover biologists often use photos of whale caudal for further analysis as it is the most discriminant pattern for distinguishing an individual whale from another. Therefore two challenges were announced in the SeaCLEF of LifeCLEF campaign, one for automatic fish categorization and enumeration, and another for automatic whale individual recognition based on visual contents. We elaborated a complex system to detect, classify and track objects (fishes) in underwater video by examining each image frame of it. We used Kalman filter to track the moving objects, and Hungarian method was used to match the pair of the objects in consecutive time periods because of many fishes. We categorized the detected fishes with C-SVC classifier, as an advanced SVM (Support Vector Machine) classifier. As further improvement we used color histograms and discriminant training method for filtering out false detections. For whale individual recognition we elaborated another system to compare the individuals by applying BoW model, during which Harris-Laplace detector and dense SIFT for creating low-level features. After that GMM based Fisher vectors were calculated and compared to each other with RBF kernel function. In addition to this we tried background segmentation as preprocessing.

Keywords: fish classification, tracking, Kalman filter, whale recognition, SVM method, RBF kernel function

1 Introduction

The need of automated methods for sea-related visual data is more important in imaging systems (both underwater and not) for marine ecosystem analysis and biodiversity monitoring. Analysis of video data usually requires very time-consuming and expensive input by human observers, and this is true for underwater videos as well, although the statistics of data collection would be very useful for exploratory applications, in particular for fisheries and biological areas. This analytical "bottleneck"

greatly restricts the use of the powerful video technologies and demands effective methods for automatic content analysis to enable proactive provision of analytical information; and in order to solve this problem a challenge is announced in SeaCLEF [20] of the LifeCLEF [7] campaign of ImageCLEF.

In this challenge there were two subtasks: (1) Coral Reef Species Recognition, where the aim was to automatically identify and recognize coral reef species, and (2) Whale Individual Recognition, where the goal was to find the images that correspond to the same individual whale.

2 Coral Reef Species Recognition

2.1 Object detection, classification

For object detection and classification we have improved our previous work [22], where the bounding boxes of consecutive images with common fish identifiers can be classified into different species; therefore the final decision of classification in our solution was based on majority voting. For object detection we have used background subtraction [8] in order to separate the foreground from background. Contours of objects have been detected using by an algorithm evolved by Suzuki and Abe [21], and based on these contours the bounding boxes and the object centers were calculated.

For the classification of the content of the bounding boxes we have used Fast-Hessian Detector to determine the "key points" in each image, and SURF (Speed Up Robust Features) [1] descriptor (both of them are low-level features) to extract local information at each key point. For creating high-level representation of each image we clustered the SURF descriptors with K-means [16] algorithm, and the resulting cluster centers were considered as codewords, since a centroid represents similar feature descriptors. For calculating a high-level descriptor for an unknown image the low-level features are extracted from it, and based on the statistics (histograms) of the nearest codewords the high-level descriptor is calculated.

For classification the high-level descriptors we used a variation of SVM (Support Vector Machine), the C-SVC (C-support vector classification) [2][3] with linear kernel function. The SVM is basically a binary linear classifier, thus in order to extend it to a number of classified categories, the one-against-all technique was used.

2.2 Tracking system

After the object detection Kalman filter [9][25] was used to track objects in three steps: (i) initialization, and after that there is a cycle process with (ii) prediction and (iii) correction. At initialization step an identity number and a confidence value were attached to every detected fish. In the next step a prediction was calculated by Kalman filter on each detected object (using the calculated object center) to forecast the future position of the investigated object. In the correction step the new detections (in next frame of the video) give the measurements (which are used in the comparison of the measurements with predictions). These measurements were used for correcting the

Kalman filter objects. In order to reach the best tracked-measured coupling we applied the Hungarian method [12][5], completed with a restriction that we removed those objects that not belong to a new measurement.

2.3 Further improvements

We used different additions at machine learning phase for improving our predictions. One of them was the discrimination learning based on images extracted from training video set. We collected false positive detections according to the ground truth to define a new category, so-called 'Trash' class. This 'Trash' class was used for filtering out some particular objects from the classification procedure, which cannot be identified with high probability.

The other way that we tried to improve our prediction was that we used color histogram in addition to SURF descriptors at training phase. We created three color histograms, one for each color channel with 256 intensity levels. After that we used the same methods as in 2.1 and 2.2.

2.4 Official evaluation of Coral Reef Species Recognition

In the official evaluation the normalized counting score is measured (instead of accuracy as in our preliminary testing). The counting score (CS) is defined as can be seen in Equation (1), where d is the difference between the number of occurrences in the run (per species) and the number of occurrences in the ground truth (Ngt).

$$CS = e^{-\frac{d}{Ngt}} \tag{1}$$

The precision (Pr) is defined as Pr= TP/(TP+FP) with TP and FP being, respectively, the true positives and the false positives. The normalized counting score (NCS) is defined as NCS = CS x Pr.

Our final official results can be seen in two tables; the Table 1 presents the average (per video and species) normalized counting score (NCS), precision and counting score; and NCS values for each fish species can be seen in Table 2.

Table 1. Occurrences and NCS (Normalized Counting Score) results at fish species

	normalized counting score	**precision**	**counting score**
BME TMIT RUN1	0.28	0.35	0.54

Table 2. Counting score, precision and NCS (Normalized Counting Score) results at fish species

Fish species	Counting	Precision	Normalized
abudefduf vaigiensis	0.8273	0.6712	0.5553
acanthurus nigrofuscus	0.8257	0.3425	0.2828
amphiprion clarkii	0.1713	0.0621	0.0106
chaetodon lunulatus	0.4612	0.2748	0.1267
chaetodon speculum	1.0000	1.0000	1.0000
chaetodon trifascialis	0.6844	0.5068	0.3469
chromis chrysura	0.0456	0.0275	0.0013
dascyllus aruanus	0.8095	0.5078	0.4111
dascyllus reticulatus	0.5122	0.1988	0.1018
hemigymnus melapterus	0.7632	0.2877	0.2196
myripristis kuntee	0.2554	0.0411	0.0105
neoglyphidodon nigroris	0.6876	0.3545	0.2437
pempheris vanicolensis	1.0000	1.0000	1.0000
plectrogly-phidodon dickii	0.0698	0.0353	0.0025
zebrasoma scopas	0.1201	0.0321	0.0039

3 Whale Individual Recognition

The aim was to find the images that correspond to the same individual whale; and the basic idea was to create the representation of each image based on the visual content, and measure the similarity between the images by using their representatives. This approach consists of four steps: (i) feature detection, (ii) feature description, (iii) image description, (iv) similarity measurement as usual phases in computer vision and we solved these steps likewise to our previous work [23].

3.1 Feature detection and description

Lots of different feature types can be detected in an image, e.g. corners, edges, ridges, as "interesting" part of an image, furthermore many possible feature extraction methods are available for images. We chose SIFT (Scale-Invariant Feature Transform) algorithm [15][14], using dense sampling method (briefly dense SIFT). This sampling method can be considered as a two-dimensional grid upon the image, where SIFT descriptors were calculated at each grid point. We also used the Harris-Laplace corner detector [6][17] for feature detection. Each of these descriptor vectors belongs to only one "interesting" point of an image.

3.2 Image description

The final step of creating the representation is the completion of a high-level descriptor for each image. Following the general trend, we applied BoW (bag-of-words) model [4][13] for this purpose, where images are treated as documents. We have used GMM (Gaussian Mixture Model) for determining the codebook [19][24] (whole set of codewords gives the codebook), which is a parametric probability density function represented as a weighted sum of (in our case 256) Gaussian component densities. GMM parameters were estimated based on the training set by using the iterative EM (Expectation Maximization) algorithm [24], but an initial model was needed for EM. In our training procedure the k-means clustering was performed over all the vectors with 256 clusters, which resulted the initial model for EM. As a result of the algorithm described above, a codebook with 256 codewords was available for further calculations, which can be considered as a concise representation of the training image set. According to the codebook the next step was to create a descriptor that specifies the distribution of the visual codewords in any image, called high-level descriptor. To represent an image with high-level descriptor, the GMM based Fisher vector [19][18] was calculated. These vectors were the final representations (image descriptor) of the images.

3.3 Similarity measurement

Firstly we tried Euclidean and other distance measurement techniques (e.g. City Block, Mahalanobis) for determining the relation between two fisher vectors. As far

as we were able to check the performance (by manually checking the most confident matches), these methods resulted very poor accuracy. Thus, we built a kernel matrix using RBF (Radial Basis Function) kernel function, because an entry of kernel matrix *K(i,j)* describes how similar the i^{th} Fisher vector to the j^{th} Fisher vector (i.e. the i^{th} image to the j^{th} image). According to this matrix, the list of our discovered matches in a descending confidence order were the elements of our submitted run file (i.e. 'BME TMIT Whalerun1'). Note that we included the first approximately 1 million pairs in our run file, although with a shorter list we could have achieved a probably higher score in the official competition.

3.4 Segmentation

We used the segmentation propagation technique introduced in [10][11] for separating the background (the water) from the whale's caudal fin. After that, we performed the same methods as in 3.1-3.3 on the 'masked images' for matching the individuals. As it can be seen in Figure 1, the algorithm was not able to perfectly separate the fins in all cases. In the first column of the figure, there are a few examples for a perfect segmentation; the second column includes acceptably good segmentations, where a portion of water still present in the masked version of the images; finally, some wrong segmentations can be seen in the last column, but this occurred rarely. Thus we created the 'BME TMIT Whalerun2' and the 'BME TMIT Whalerun3' on this basis. The only difference between them was that during the feature extraction phase, for the 'BME TMIT Whalerun3' we restricted the dense sampling of SIFT descriptors to sample exclusively from the masked portion of the image (i.e. from the caudal fin).

Fig. 1. Some examples for the segmentation of whales

3.5 Official evaluation of Whale Individual Recognition

The metric used for evaluating the submitted run files was the Average Precision (i.e. the precision averaged across all good matches of the ground truth). The following table provides the AP values of our submitted runs.

Table 3. Results of Whale Individual Recognition

Run name	Average Precision
bmetmit_whalerun_1	0.25
bmetmit_whalerun_3	0.10
bmetmit_whalerun_2	0.03

4 Conclusion

Two challenges were announced in the SeaCLEF of LifeCLEF campaign, one for automatic fish categorization and enumeration, and another for automatic whale individual recognition based on visual contents. For the first task we elaborated a complex system to detect, classify and track objects (fishes) in underwater video by examining each image frame of it. We used Kalman filter to track the moving objects, and Hungarian method was used to match the pair of the objects in consecutive time periods because of many fishes. We categorized the detected fishes with C-SVC classifier, as an advanced SVM (Support Vector Machine) classifier. As further improvement we used color histograms and discriminant training method for filtering out false detections. In the official evaluation the normalized counting score is measured, our final official result was 0.28.

For whale individual recognition we elaborated another system to compare the individuals by applying BoW model, during which Harris-Laplace detector and dense SIFT for creating low-level features. After that we calculated the GMM based Fisher vectors and then we compared them to each other with RBF kernel function, but firstly we tried several distance measurements. In addition to this we tried background segmentation as preprocessing. Average Precision (AP) metric was used for evaluating, our final official results for submitted run files were 0.25, 0.10, 0.03. Based on these we can conclude that the segmentation had no positive effect on the recognition.

References

1. Bay, H. and Tuytelaars, T. and Van Gool, L.: SURF: Speeded Up Robust Features, *9th European Conference on Computer Vision*, (2006)

2. Boser, B., Guyon, I., Vapnik, V.: A Training Algorithm for Optimal Margin Classifier, Proc. of the *5th Annual ACM Workshop on Computational Learning Theory*, pp. 144-152 (1992)
3. Cortes, C., Vapnik, V.: Support-vector networks, *Machine Learning*, Vol. 20, No. 3, pp. 273-297 (1995)
4. Fei-Fei, L., Fergus, R., & A. Torralba, A.: Recognizing and Learning Object Categories, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, (2007)
5. Frank, A.: On Kuhn's Hungarian method—a tribute from Hungary. *Naval Research Logistics (NRL)*, 52(1), 2-5. (2005)
6. Harris, C, Stephens, M.: A combined corner and edge detector. In C. J. Taylor, editors, *Proceedings of the Alvey Vision Conference*, pages 23.1-23.6. Alvey Vision Club, September 1988. doi:10.5244/C.2.23.
7. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W. P., Champ, J., Planqué, R., Palazzo, S., Müller, H.: LifeCLEF 2016: multimedia life species identification challenges, *Proceedings of CLEF 2016* (2016)
8. KaewTraKulPong P. and Bowden, R.: An Improved Adaptive Background Mixture Model for Real- time Tracking with Shadow Detection, In Proc. *2nd European Workshop on Advanced Video Based Surveillance Systems*, AVBS01. Sept (2001)
9. Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1), 35-45. (1960)
10. Kuettel, D., Guillaumin M., Ferrari V.: ImageNet Auto-annotation with Segmentation Propagation, Technical Report, *International Journal of Computer Vision*, 2013.
11. Kuettel, D., Guillaumin, M., Ferrari, V.: Segmentation Propagation in ImageNet *European Conference on Computer Vision (ECCV)*, Firenze, Italy, October 2012.
12. Kuhn, H. W.: The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2), 83-97. (1955)
13. Lazebnik, S., Schmid, C. and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, New York, Vol. 2, pp. 2169-2178 (2006)
14. Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, 60, 2, pp. 91-110, (2004)
15. Lowe, D.: Object recognition from local scale-invariant features. In: *ICCV* (1999)
16. MacQueen, J.: Some methods for classification and analysis of multivariate observations, Proceedings of the *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281-297 (1967)
17. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors, International *Journal on Computer Vision* 60(1), 2004, pp. 63-86.
18. Perronnin, F., Dance, C.: Fisher kernel on visual vocabularies for image categorization, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, (2007)
19. Reynolds D. A.: Gaussian Mixture Models, *Encyclopedia of Biometric Recognition*, Springer, February, pp. 659-663 (2009)
20. SeaCLEF 2016 http://www.imageclef.org/lifeclef/2016/sea (LifeCLEF 2016), *CLEF working notes 2016*
21. Suzuki, S. and Abe, K.,: Topological Structural Analysis of Digitized Binary Images by Border Following. *Computer Vision, Graphics, and Image Processing*, 30(1), 32-46. (1985)

22. Szűcs, G., Papp, D., Lovas, D., SVM classification of moving objects tracked by Kalman filter and Hungarian method, In: L Cappellato, N Ferro, G J F Jones, E S Juan (eds.) *Working Notes of CLEF 2015 Conference*, Toulouse, France, September 8-11, 2015, Paper 40. 10 p. Vol. 1391.
23. Szűcs, G., Papp, D., Lovas, D., Viewpoints Combined Classification Method in Image-based Plant Identification Task In: Cappellato L., Ferro N., Halvey M., Kraaij W. (eds) *Working Notes for CLEF 2014 Conference*. Sheffield, Great Britain, September 15-18, 2014., pp. 763-770. Vol.1180.
24. Tomasi C.: Estimating gaussian mixture densities with EM: A tutorial, *(Tech. rep., Duke University); Chinese Journal of Electron Devices*, pp, 15-18 (2004)
25. Welch, G. F.: Kalman Filter. *Computer Vision: A Reference Guide*, 435-437. (2014)