Exploring the use of local descriptors for fish recognition in LifeCLEF 2015

Jorge Cabrera-Gámez, Modesto Castrillón-Santana, Antonio Domínguez-Brito, Daniel Hernández-Sosa, Josep Isern-González, and Javier Lorenzo-Navarro

SIANI
Universidad de Las Palmas de Gran Canaria
Spain
jcabrera@iusiani.ulpgc.es
http://berlioz.dis.ulpgc.es/roc-siani

Abstract. This paper summarizes the proposal made by the SIANI team for the LifeCLEF 2015 Fish task. The approach makes use of standard detection techniques, applying a multiclass SVM based classifier on large enough Regions Of Interest (ROIs) automatically extracted from the provided video frames. The selection of the detection and classification modules is based on the best performance achieved for the validation dataset consisting of 20 annotated videos. For that dataset, the best classification achieved for an ideal detection module, reaches an accuracy around 40%.

Keywords: Local descriptors, score level fusion, SVM based classification

1 Introduction

There are different scenarios of application where underwater monitoring is a required ability such as biological, fisheries, geological and physical surveys. The everyday larger availability of media captured in this environment poses the challenge to extract useful data automatically. This is indeed a hard scenario where effective techniques are needed to reduce costs and human exposition.

With this aim, CLEF presented in 2014 for the first time LifeCLEF: the Labs dedicated to multimedia life species identification [10], including FishCLEF: a video-based fish identification task. The short term goal was simply to automatically detect any fish and its species. The medium term goal is to provide researchers tools to automatically monitor species with high accuracy, in order to extract information of living species for a sustainable development and biodiversity conservation.

This year the Labs [3] and task have been reedited [11, 14]. The participants could initially access to training data, to later submit labels for the test set. The task to be accomplished was "count fish per species in video segments".

This paper describes the approach adopted by the SIANI team. The following sections detail the different elements integrated that basically perform initially a detection to later identify the fish species of the cropped image.

2 The approach

As succinctly mentioned above, the fish identification task has been decomposed into two phases: detection and classification.

2.1 Detection

The goal of the detection phase is to reduce the searching area extracting candidate ROIs from the video stream. Three different foreground detection approaches have been tested: fast, histogram backprojection and Gaussian Mixture Modeling (GMM).

Fast. This approach makes use of a simple and fast background model computed from the video frames, that is robust enough for the detection and extraction problem in some scenarios. This background modeling solution takes advantage of the static camera configuration in this particular scenario.

To define the scene background model, bg, we have used a similar method to commonly-used techniques like mean filter or median filter [13]. We compute the mode image, $\overline{I}$, that is calculated as the most frequent values in each RGB component of each pixel each pixel along the video frames.

Once the background model is available, simple and fast background subtraction techniques may be applied in each RGB component of the input image, I. The foreground is computed based on a defined threshold applied to the sum of squares of RGB components of the subtracted image (D_R, D_B, D_G) pixel value

$$S(i,j) = D_R(i,j)^2 + D_G(i,j)^2 + D_B(i,j)^2 \tag{1}$$

where $D_x = I_x - bg_x$ for every RGB component $(x = R, G, B)$.

For a pixel in a given image, $I(i,j)$, its corresponding pixel in the foreground image, fg, is computed as

$$fg(i,j) = \begin{cases} I(i,j) \text{ if } S(i,j) \geq \tau_1 \\ 0 \quad \text{otherwise} \end{cases} \tag{2}$$

The definition of the threshold (τ_1) in equation 2 has been determined by the application scenario. The extracted pixels are considered foreground, i.e. region of interest in the detection problem.

Histogram Backprojection (BackProj). The second detection method evaluated is inspired in the idea proposed in [16] that we have adapted to background segmentation. The method is based on the backprojection of temporal color histogram, and comprises the following steps:

1. Calculate for each color component the temporal histogram of every image pixel: $h_x = hist_t(I_x(i,j))$

2. Add to each histogram bin, k, the values of its neighborhood, $\pm s$, (convolution mask): $c_x(k) = \sum_{l=k-s}^{k+s}(h_x(l))$
3. Normalize the resulting histogram: $\overline{h_x}(k) = c_x(k)/max(c_x)$
4. Backproject the histogram on every image: $P_x(i,j) = \overline{h_x}(I_x(i,j))$
5. Sum the squares of values of each component of the pixels: $S(i,j) = P_R(i,j)^2 + P_G(i,j)^2 + P_B(i,j)^2$
6. Use a threshold to separate the foreground of the background:

$$fg(i,j) = \begin{cases} I(i,j) \text{ if } S(i,j) < \tau_2 \\ 0 \quad \text{otherwise} \end{cases} \tag{3}$$

Also in this case, the definition of the threshold (τ_2) in equation 3 has been determined by the application scenario.

GMM Based Background Modeling (GMM). The third background subtraction method analyzed is the one proposed by Zivkovic and van der Heijden [20]. This method performs a pixel-level background subtraction, modeling each background pixel with a GMM, extending the method proposed by Stauffer and Grimson [15]. Thus the background model is defined as:

$$p(\boldsymbol{x}|\mathcal{X}_T, bg) \approx \sum_{m=1}^{B} \hat{\pi}_m \mathcal{N}(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_m, \hat{\sigma}_m^2 Id) \tag{4}$$

where $\mathcal{X}_T = \{x^{(t)}, \ldots, x^{(t-T)}\}$ is the training set, for the time period T, $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_B$ are the estimates of the means, $\hat{\sigma}_1, \ldots, \hat{\sigma}_B$ are the estimates of the variances, and Id is the identity matrix. B is the number of components weighted by $\hat{\pi}_m$.

An optimization process was launched over the training videos to try to find a suitable configuration for the GMM foreground detection algorithm, including the number of distributions, the background ratio and the number of training frames and learning rates.

2.2 Classification

Detected ROIs are fed to the detection phase to identify the fish species. The classification phase has been designed based on local descriptors, that are currently well known techniques in different Computer Vision (CV) problems.

In texture analysis, an image is described in terms of a local descriptor codes making use of a histogram, h_i, where the bins contain the number of occurrences of the different descriptor codes present in the image. This approach follows a Bag of Words scheme [6]. For some problems, the use of a single histogram may introduce the loss of spatial information. To avoid this effect, a grid of cells is used defining the number of horizontal and vertical cells, respectively cx and cy, making a total of $cx \times cy$ cells on the analyzed pattern.

Once defined the grid setup, for a particular descriptor, d, the resulting feature vector, $\mathbf{x}_I^d$, contains the concatenation of $cx \times cy$ cell histograms, i.e. the

feature vector is defined as $\mathbf{x}_I^d = \{h_1, h_2, ..., h_{cx \times cy}\}$, where h_i is the descriptor histogram for cell i.

In this particular task, we have evaluated different descriptors and grid configurations. In this sense, we have considered the following 8 descriptors:

- Histogram of Oriented Gradients (HOG) [7].
- Local Binary Patterns (LBP) and uniform Local Binary Patterns (LBP^{u2}) [1].
- Local Gradient Patterns (LGP) [12].
- Local Ternary Patterns (LTP) [17].
- Local Phase Quantization (LPQ) [18].
- Weber Local Descriptor (WLD) [5].
- Local Oriented Statistics Information Booster (LOSIB) [8].

3 Results

This section describes the results obtained for the different fish identification task phases, highlighting those configurations that were submitted to the 2015 Lab focused on this particular problem [14].

Fig. 1. From left to right, a training sample, and two validation samples of Abudefduf vaigiensis. They are presented in similar relative scale.

3.1 Datasets

Before granting the access to the test data, the organizers provided two datasets, see Figure 1. Even though a better description of the data may be found in [14], we summarize some relevant characteristics below.

The first dataset, that we call the training dataset, is a collection of cropped images of the different fish species. The second collection contains annotated videos, including media that may present a similar scenario to the test data. We called this collection the validation dataset.

This validation dataset is used in the following subsections to analyze the different detection and classification alternatives, providing a cue to decide the final system setup chosen for the Fish task submission. In fact, we used both training and validation datasets to select the classification approach, and the validation dataset, to select the detection technique and tune its parameters.

Briefly, the training set contains samples of the 15 different fish species, i.e. classes. The number of samples per species is indicated in Table 1. The reader will observe that the different species are not equally represented through the dataset, circumstance that also is present in the validation and test sets. The average dimension in the training samples is $88 \pm 38 \times 102 \pm 49$ pixels.

Table 1. Number of samples per class in the training and validation datasets.

	Number of instances per dataset	
Fish type	Training	Validation
Abudefduf vaigiensis	305	132
Acanthurus nigrofuscus	2511	294
Amphiprion clarkii	2985	363
Chaetodon lunulatus	2494	1217
Chaetodon speculum	24	138
Chaetodon trifascialis	375	335
Chromis chrysura	3593	275
Dascyllus aruanus	904	894
Dascyllus reticulatus	3196	3165
Hemigymnus melapterus	147	214
Myripristis kuntee	3004	242
Neoglyphidodon nigroris	129	85
Pempheris Vanicolensis	49	999
Plectrogly-Phidodon dickii	2456	737
Zebrasoma scopas	271	72
Total	22443	9162

The validation dataset contains 9162 samples distributed per class according to the last column of Table 1. The average dimension of those samples is $52 \pm 37 \times 56 \pm 39$ pixels.

3.2 Detection Results

As mentioned above, the annotated validation dataset videos were used to analyze the performance of different detection algorithms. The detection rates for the three implementations are shown in Table 2, being computed as the total number of correct or positive detections divided by the number of annotations. The false detection rate presented is also the ratio between the number of unmatched or false detections and the number of annotations. This was done to have a clear evidence of the number of false detections in relation to the number of annotations. False detections do not necessarily mean a failure in the detection module, but that there is not annotation for that particular frame and ROI. Indeed, the annotations were done only when the fish species was clearly identifiable [14]. In this sense, we have made use of the minimal size for annotation, and applied a dimension filter to remove small detected ROIs.

A positive detection is considered when there is a significant intersection between a given detection container, B and an annotation container, A. As confidence measure, we employed the Jaccard Index, JI. This index relates the intersection of both containers with their union, $JI = \frac{A \cap B}{A \cup B}$, providing a value between 0 and 1, larger values meaning better matching. For the analysis summarized in Table 2, we have considered 0.4 and 0.5 threshold values.

Table 2. Detection rates considering different detection techniques and JI thresholds.

Detection	Detection rate (false detection rate)	
algorithm	JI=0.5	JI=0.4
Fast I	0.80(4.60)	0.85(4.55)
Fast II	0.74(3.74)	0.80(3.67)
BackProj	0.82(2.77)	0.88(2.49)
GMM	-	0.40(2.49)

The high variability of the video segments made extremely difficult to obtain a good tuning of the algorithm parameters. As a consequence, simple approaches yielded better results both in execution time and detection. Indeed, among the techniques analyzed, both **Fast** and **BackProj** algorithms provided not brilliant but acceptable detection rates. **Fast** was chosen with different tuning parameters to setup *run1*, while **BackProj** was used for the other two submitted runs (*run2* and *run3*). The detection approach is later combined with the classifier providing the best performance in the validation dataset classification.

3.3 Classification Results

The detection rates achieved, described in the previous section, allowed us to explore our model based approach on the dataset. Certainly, a model approach is not a priori the best solution for the unbalanced classification task, but being newcomers, we were interested in applying our experience in other CV problems to evaluate local descriptors in this scenario.

The analysis described in this section presents results in two steps. Firstly, the study evaluates different descriptors with the training set, i.e. the collection of cropped images, see Table 1. Secondly, the best descriptors are later evaluated with the validation dataset, to adopt the most promising configuration for the test set.

Table 3 summarizes the results for different local descriptors in a 5-fold cross validation experiment defined on the provided training dataset, considering a single multi-class SVM based classifier. This kind of approach has already been applied for the task [2]. Each descriptor is evaluated for different grid configurations in the ranges $cx \in [1, 4]$ and $cy \in [1, 4]$. Unfortunately, for the given deadline (its extension was not evident), we could not manage to evaluate all the grid configurations with dimensions larger than 3×3.

Table 3. Mean accuracy achieved in a 5-fold cross validation experiment on the training dataset. The cell color serves to cluster accuracies.

Descriptor	Grid setup			
	1×1	2×2	3×3	4×4
HOG	54.69	87.71	95.66	97.50
LBPu2	77.79	94.11	96.72	96.88
LBP	88.20	91.62	85.98	65.58
LGP	61.57	84.00	92.06	95.13
LTP	89.90	86.18	34.91	26.60
LPQ	90.74	95.70	88.73	53.87
WLD	16.01	16.01	15.92	13.30
LOSIB	40.32	73.13	87.63	92.28

With the exception of WLD, the whole collection of descriptors reported a high accuracy at least for a particular grid setup. However, this was not the case in the following analysis on the annotated ROIs extracted from the validation dataset, as summarized in Table 4. It seems, that the grid configuration is important for some descriptors. A larger number of cells is preferred for HOG, LBP^{u2}, LGP and LOSIB, while other descriptors such as LBP, LTP and LPQ provide better results with a lower number of cells. Again, WLD is not providing useful classification results.

Table 4. Accuracy achieved with a single descriptor training with the training dataset and testing with the extracted annotated samples on validation dataset.

Descriptor	Grid setup			
	1×1	2×2	3×3	4×4
HOG	19.25	28.24	33.55	36.76
LBPu2	19.09	25.92	33.08	31.18
LBP	20.91	18.94	12.88	4.99
LGP	14.85	22.36	28.17	30.84
LTP	18.55	16.88	4.63	3.00
LPQ	22.15	20.85	15.72	6.48
WLD	3.00	3.00	3.00	3.96
LOSIB	12.55	19.21	27.42	28.47

Considering the vital importance of combining several descriptors [19], a further evaluation of a fusion approach was considered. According to the score level (SL) fusion literature and previous results in the context of facial processing [4, 9], we adopted a score level fusion approach where the first layer is composed by a set of classifiers designed according to the chosen descriptors, while the second layer classifier takes the first layer scores as input. In summary, the fusion alternatives analyzed below follow the approach outlined in Figure 2.

Table 5 summarizes the results achieved for different fusion alternatives. The selection of descriptors and grids are based on the single descriptors results

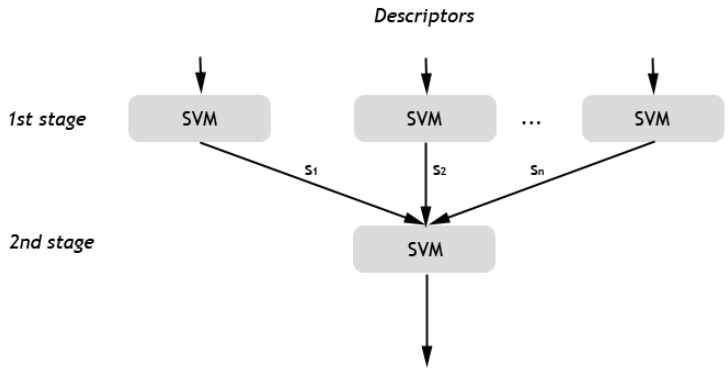

Fig. 2. Illustration of two stage classification fusion architecture, with n classifiers in the first stage whose scores are fed into a second stage *meta* classifier.

achieved for the validation dataset, see again Table 4. This table does not include results with 4×4 grids, as they were not available in time for the deadline.

The first three alternatives combine the best descriptors for a given grid resolution; the higher the resolution, the better the accuracy. However, there is no real restriction to make use of a unique grid resolution. For that reason we also evaluated the fusion of the best descriptors with different grid resolutions, achieving the best overall accuracy with a RBF kernel using the first 90 PCA components.

For each combination, the selected descriptors and grid setups are indicated, reporting the results using SVM based classifiers, including linear and RBF kernels, with and without a previous dimensionality reduction by means of a Principal Component Analysis (PCA).

The best performing classifier, the fourth combination using RBF kernel with a PCA based features, is used in combination with the selected detection approaches.

3.4 Discussion

As mentioned above, our team submitted three runs. They all made use of an identical second phase based on a two stages classifier. The selected descriptors combination is the one highlighted in Table 5. This score fusion selection contains six single descriptor classifiers in the first stage: $\text{LTP}_{1\times1}$, $\text{LPQ}_{1\times1}$, $\text{HOG}_{3\times3}$, $\text{LBP}^{u2}_{3\times3}$, $\text{LGP}_{3\times3}$ and $\text{LOSIB}_{3\times3}$. The second stage makes use of the classifiers scores, that are projected into a PCA space.

Each run differs in its detection phase. Our first run made use of the **Fast** detection algorithm, while the other two integrate the **BackProj** detector with different parameters setup.

Table 5. Accuracy achieved with fusion approaches training with the image folder and testing with the extracted annotated samples on the folder of validation videos.

Descriptors	SVM Approach	Accuracy
$HOG_{1\times1}$+$LBP_{1\times1}$+$LBP^{u2}_{1\times1}$+$LTP_{1\times1}$+$LPQ_{1\times1}$	RBF	24.56
	RBF+PCA	25.45
	Linear	27.42
	Linear+PCA	27.21
$HOG_{2\times2}$+$LBP_{2\times2}$+$LBP^{u2}_{2\times2}$+$LTP_{2\times2}$+$LPQ_{2\times2}$	RBF	20.44
	RBF+PCA	29.54
	Linear	21.06
	Linear+PCA	29.69
$HOG_{3\times3}$+$LBP^{u2}_{3\times3}$+$LGP_{3\times3}$+$LOSIB_{3\times3}$	RBF	35.63
	RBF+PCA	38.65
	Linear	36.96
	Linear+PCA	38.75
$LTP_{1\times1}$+$LPQ_{1\times1}$+$HOG_{3\times3}$+$LBP^{u2}_{3\times3}$+$LGP_{3\times3}$+$LOSIB_{3\times3}$	RBF	34.99
	RBF+PCA	40.41
	Linear	36.43
	Linear+PCA	39.74

The normalized counting scores of the referred runs in the overall Lab analysis are reported in Figure 3. Two teams are over 50%, followed at a remarkable distance by the best runs of other two teams, including our **run3**, achieving over 30%. Our main focus was on the classification phase, that has provided unbalanced results for different classes, likely due to the non homogeneous number of training samples per class. A focus based exclusively on the fusion of local descriptors seems not to be reliable enough for the problem. However, the detection phase requires further attention as a larger number of proper detections would improve the overall score.

4 Conclusions

This document describes the model based approach submitted to the LifeCLEF 2015 Fish task by the SIANI team. The proposal explores the use of local descriptors for this problem. We employed standard detection techniques to later apply an ensemble of SVM multiclass classifiers.

Three runs were submitted with identical classification stage. One is based on the **Fast** detection algorithm, while the other two are based on the **BackProj** algorithm.

The best accuracy achieved for the ideal annotated containers reaches 40%, suggesting that the approach is still far from being reliable in this scenario. In the close future, our aim is on the one side at improving detection, that might be combined with tracking. On the other side, once we have observed the problems originated in the multiclass classification of an unbalanced dataset, and apart

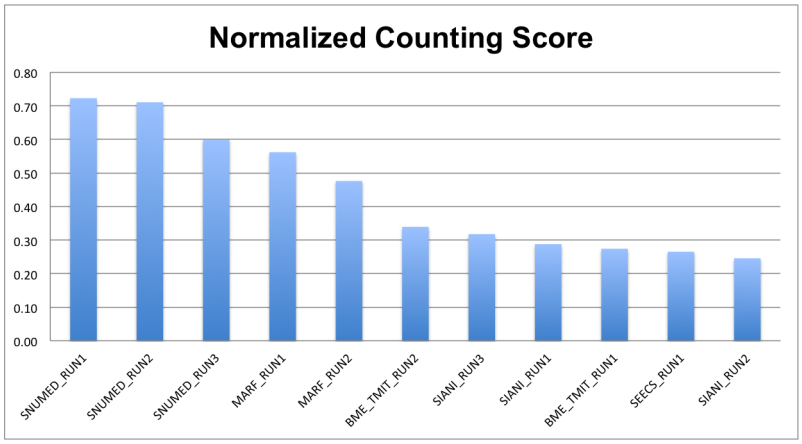

Fig. 3. Normalized counting score of the participant runs. Extracted from [14].

from computing more dense grids, we should explore the combination with other techniques to leverage the classification stage.

Acknowledgments. Work partially funded by the Institute of Intelligent Systems and Numerical Applications in Engineering and the Computer Science Department at ULPGC.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12) (December 2006)
2. Blanc, K., Lingrand, D., Precioso, F.: Fish species recognition from video using svm classifier. In: CLEF (Working Notes). pp. 778–784 (2014)
3. Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.): CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org) (2015), ISSN 1613-0073, http://ceur-ws.org/ Vol-1391/.
4. Castrillón, M., Lorenzo, J., Ramón, E.: Improving gender classification accuracy in the wild. In: 18th Iberoamerican Congress on Pattern Recognition (CIARP) (2013)
5. Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: WLD: A robust local image descriptor. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(9), 1705–1720 (September 2010)
6. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray., C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. pp. 1–22 (2004)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C. (eds.) International Conference on Computer Vision & Pattern Recognition. vol. 2, pp. 886–893 (June 2005)

8. García-Olalla, O., Alegre, E., Fernández-Robles, L., González-Castro, V.: Local oriented statistics information booster (LOSIB) for texture classification. In: International Conference in Pattern Recognition (ICPR) (2014)
9. Heisele, B., Serre, T., Poggio, T.: A component-based framework for face detection and identification. International Journal of Computer Vision Research 74(2) (August 2007)
10. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Planque, R., Rauber, A., Fisher, R., Müller, H.: Lifeclef 2014: Multimedia life species identification challenges. In: nformation Access Evaluation. Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science Volume, vol. 8685, pp. 229–249. Springer (2014)
11. Joly, A., Müller, H., Goëau, H., Glotin, H., Spampinato, C., Rauber, A., Bonnet, P., Vellinga, W.P., Fisher, B.: Lifeclef 2015: multimedia life species identification challenges. In: Proceedings of CLEF 2015 (2015)
12. Jun, B., Kim., D.: Robust face detection using local gradient patterns and evidence accumulation. Pattern Recognition 45(9), 3304–3316 (2012)
13. Lo, B., Velastin, S.: Automatic congestion detection system for underground platforms. In: Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. pp. 158–161 (2001)
14. Spampinato, C., Fisher, B., Boom, B.: Lifeclef fish identification task 2015. In: CLEF working notes 2015 (2015)
15. Stauffer, Grimson: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 246–252 (1999)
16. Swain, M.J., Ballard, D.H.: Color indexing. International Journal on Computer Vision 7(1), 11–32 (1991), `http://www.springerlink.com/content/n231141541p1211g/`
17. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. Image Processing, IEEE Transactions on 19(6), 1635 – 1650 (2010)
18. V, O., J., H.: Blur insensitive texture classification using local phase quantization. In: Proc. Image and Signal Processing (ICISP). pp. 236–243 (2008)
19. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision 73(2), 213–238 (2007)
20. Zivkovic, Z., der Heijden, F.: Effcient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters 27, 773–780 (2006)