Combining image-level and object-level inference for weakly supervised object recognition. Application to fisheries acoustics.

R. Lefort\textsuperscript{1,2,3}, R. Fablet\textsuperscript{2,3}, I. Karoui\textsuperscript{1,2,3}, J.-M. Boucher\textsuperscript{2,3}

\textsuperscript{1} Ifremer/STH, Technopôle Brest Iroise, 29280 PLOUZANE, France
\textsuperscript{2} Institut Telecom/Telecom Bretagne/Lab-STICC, Technopôle Brest-Iroise - CS 83818, 29238 Brest cedex, France
\textsuperscript{3} Université Européenne de Bretagne

Abstract

This paper addresses weakly supervised object recognition. We show how the combination of an image-level inference, in terms of image-level object class priors, can lead to better training of object recognition models. Stated within a probabilistic setting, the proposed approach is applied to fisheries acoustics and fish school recognition.

Figure 1. In order to assess pelagic fish stock in an area (left), the vessel acquires images of the water column throughout transversal motion (left). Images contain fish schools (right) that are assigned to their class species according to school characteristics. The training data for classification model learning is issued from the acoustic echograms corresponding to trawling (fishing with a net) sites. Trawl catches spots are shown on the left.

1 Introduction

Object recognition and classification have received a great attention in recent years, especially with the development of novel machine learning techniques. The main effort has been initially dedicated to supervised learning, that is to say training classification models from a representative dataset of labelled objects exist. In a number of real applications, such representative datasets are however rare, if available. Often, the training dataset may contain both labelled and unlabelled samples and adapted training algorithms have to be developed \cite{1}. Weakly supervised learning is among the cases of interest: the prior labelling information provided in the training refers to a subset of potential classes for each object. This statement is typical of object recognition in images. For instance, it is often the case that images comprise several different objects and that the labelling information, in terms of object classes, is given globally at the image level \cite{2}. Other examples may include remote sensing applications such as forest or field cover analysis \cite{3}. Two types of prior labelling information may be considered: the knowledge of the presence or absence of the object classes in each training image \cite{2} \cite{4}, or some information on the relative class priors \cite{5}.

Our interest in weakly supervised learning comes from fisheries acoustics. In this case, the objects of interest are fish schools in acoustic echograms, referred to hereafter as acoustic images, and object classes correspond to fish species or group of species. From the acoustic data acquired by an echosounder mounted under an oceanographic vessel (figure 1), the goal is to assess fish species biomass. At the operational level, this assessment is carried out by experts from the conversion into biomass of the acoustic energy of the fish schools observed in the acoustic images. This conversion being species-specific, a prior fish school classification is required. With a view to developing computer-assisted tools, we aim at automating this classification task. The training dataset is issued from the correspondence between the analysis of the trawl catches and the acoustic image of the trawled area, such that the training dataset is provided as a set of images, containing fish schools, with the associated relative species proportions. Hence, object class proportions are known in training images but not the individual class objects. This can be viewed as a generalization of the case considered in \cite{2} in which the training dataset is formed by object groups with the knowledge of the presence or absence of the different classes in each group.

In this work, we investigate how to combine image-level and object-level inference to improve object recognition. As sketched in Figure 2, we aim at initially propagating image-level object priors from the training images.
In our case, this image-level inference exploits both global image similarities and a spatial regularity constraint. In a second step, the knowledge of the image-level object priors constrains the object recognition step. This combination of image-level and object-level inference is shown to bring a significant improvement in terms of correct classification rate. The paper is organized as follows. We first present the outline of the proposed approach in section 2. The image-level inference stated as a variational issue and the object classification models are respectively detailed in section 3 and 4. Results and concluding constitutes the last section 5.

2 Our approach

Let us denote by \( \{ I_{tr}(k), \pi_k \}_{1 \leq k \leq N_{tr}} \) the training images for which the relative object class priors \( \pi_k = \{ \pi_{ki} \} \) are known and \( \{ I(k) \}_{1 \leq k \leq N} \) the remaining images. Index \( i \) refers to object classes. We assume that objects are detected in all images and \( n \) is the index of the different objects in an image. Let us denote by \( N_{tr}(k) \) the number of objects in training image \( k \), and \( N(k) \) in the \( k^\text{th} \) image of the other set of images. Any object is characterized by a feature vector \( x_{kn} \) and \( y_{kn} \) denotes its class.

An overview of the proposed approach is given in Figure 2. Three main steps are involved:

- We first infer class priors \( \tilde{\pi}_{ki} = p(y_{kn} = i) \) in all images from the known priors in training images (section 3). This step relies on the decomposition of the statistics of object features in a given image with respect to the feature statistics in training images. This is stated as a variational interpolation.

- The second step (section 4) involves the training of probabilistic object classification models of the form \( p(y_{kn} = i | x_{kn}, \Theta) \), where \( \Theta \) stands for model parameters. The proposed training criterion exploits the knowledge of class priors in training images.

- The third step consists in combining image-level class priors \( \tilde{\pi}_{ki} \) to object-level classification likelihoods \( p(y_{kn} = i | x_{kn}, \Theta) \) to carry out object classification:

\[
p(y_{kn} = i | x_{kn}, \Theta, \tilde{\pi}) \propto \tilde{\pi}_{ki} p(y_{kn} = i | x_{kn}, \Theta) \quad (1)
\]

The classification resorts to selecting the most likely class according to posterior likelihood (1).

3 Image-level inference

In this section we present the image-level inference of class priors. The key principle is to match the statistics of object features in a given acoustic image to the feature statistics in training images. Further assuming that images are observations sampled on a spatial grid, as in the considered application where each image refers to a site along the survey track of the oceanographic vessel, spatially neighbouring images are expected to display homogeneous variations of class priors.

Formally, we infer assignment weights \( \{ \tau_l(k) \} \) corresponding to the likelihood of the assignment of image \( k \) to training images \( l \), such that \( \sum_l \tau_l(k) = 1 \). Given known class priors in training images \( \pi_{l,i} \), class priors \( \tilde{\pi}_{ki} \) in image \( k \) can be estimated from assignment weights as a weighted average:

\[
\tilde{\pi}_k = \sum_{l=1}^{N_{tr}} \frac{N_{tr}}{n_l} \pi_l \tau_l(k) \quad (2)
\]

Image content is characterized by the statistics of object features. Here we consider a quantile-based representation of the marginal histograms of object features. Denoting by \( H(k) \) the set of these marginal distributions for image \( k \), we assume that the latter can be decomposed with respect to the marginal distributions in training images, i.e.

\[
H^f(k) = \sum_{l=1}^{N_{tr}} \tau_l(k) H^f_{tr}(l) + B^f(k) \quad (3)
\]

where \( H^f(k) \) is the marginal distribution for feature \( f \) in image \( k \), \( H(k) = \{ H^f(k) \} \), \( H_{tr}(l) \) the analogous marginals for training image \( l \) and \( B^f(k) \) a zero-mean noise.

Using a Bayesian formulation, assignment weights are
estimated according to the MAP criterion:
\[ \hat{\pi} = \arg \max_{\pi} \left\{ p(\tau|H) \right\} = \arg \min_{\pi} \left\{ -\log p(H|\tau) - \log p(\pi) \right\} \] (4)

From observation equation (3), data-driven energy 
\[ E_1(\tau) = \int_{k} D \left( H^f(k), \sum_{l=1}^{N_\tau} \tau_l(k) H_{il}^f(l) \right) \] (5)

where \( D \) is a distance between marginal distributions. Here, the Bhattacharrya distance [6], known to be a relevant similarity measure between histograms in a number of applications, between histograms \( h_1 \) and \( h_2 \):
\[ D(h_1, h_2) = 1 \frac{1}{L} \sum_{l=1}^{L} \sqrt{h_1 \cdot h_2} \] (6)

for L-dimensional histograms. In addition to this data-driven term, prior energy 
\[ -\log p(\tau) \] is stated as a regularization term:
\[ E_2(\tau) = \int_{\{l(k)\}} |\nabla \tau_l(k)| \] (7)

The image-level estimation of class priors then resorts to the following variational interpolation:
\[ \hat{\pi} = \arg \max_{\pi} \left\{ E_1(\tau, H) + E_2(\tau) \right\} \] (8)

subject to the constraint:
\[ \sum_{l=1}^{N_\tau} \tau_l(k) = 1 \] (9)

The above maximization is achieved using a gradient descent where the gradient of the variational criterion is issued from Euler-Lagrange equation.

4 Training discriminative object classification model

Object classification models are stated according to the framework proposed in [5]. Given previous results [5], we here focus on discriminative classification models (DCMs). In the linear case, these models are stated as an explicit probabilistic parameterization of the linear discriminant classifier:
\[ p(y_{kn} = i|x_{kn}, \Theta) \propto \exp \left\{ \beta \left[ \langle \omega_i, x_{kn} \rangle + b_i \right] \right\} \] (10)

where \( \langle \omega_i, x_{kn} \rangle + b_i = 0 \) is the equation of the decision hyperplan parameterized by normal vector \( \omega_i \) and intercept \( b_i \) for class \( i \) and parameter \( \beta \) sets the dynamic of the probabilistic decision. Let us denote \( \Theta = \{ \omega_i, b_i \} \). Using a kernel approach, this linear model is extended to a non-linear one [7]. More precisely, we exploit a non-linear mapping issued from a kernel PCA (principal component analysis) [5], which leads to a parameterization of classification likelihood \( p(y_{kn}|x_{kn}) \) similar to Eq.(10) except that the original feature vector \( x_{kn} \) is replaced by its mapped version in the kernel space spanned by the kPCA. Here, we consider a Gaussian kernel.

As proposed in [5], a weakly supervised training based on class prior information can be derived to estimate model parameters \( \Theta \) from the minimization of an image-level criterion, defined as the error of the estimation of image-level class priors from object classification:
\[ \hat{\Theta} = \arg \min_{\Theta} \sum_{k} D(\hat{\pi}_k(\Theta), \pi_k) \] (11)

where \( D(\cdot, \cdot) \) stands for the Bottacharrya distance as in Eq.(6), and \( \hat{\pi}_k(\Theta) \) class priors estimated from object classification likelihood as follows:
\[ \hat{\pi}_k(\Theta) = \sum_{n=1}^{N_\tau(k)} E_{kn} p(y_{kn} = i|x_{kn}, \Theta) \] (12)

Where \( E_{kn} \) is the physical feature of object \( n \) in training image \( k \) (e.g., \( E_{kn} \) equals one if these proportions are computed as relative object occurrences).

5 Results

<table>
<thead>
<tr>
<th>Approach developed in [8]</th>
<th>mean error</th>
<th>mean error per class</th>
<th>variance per class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14.66%</td>
<td>9.07%</td>
<td>0.0142</td>
</tr>
<tr>
<td></td>
<td>21.46%</td>
<td>13.90%</td>
<td>0.0437</td>
</tr>
<tr>
<td></td>
<td>5.31%</td>
<td>3.90%</td>
<td>0.0045</td>
</tr>
<tr>
<td>Our approach</td>
<td>9.93%</td>
<td>6.94%</td>
<td>0.0028</td>
</tr>
<tr>
<td></td>
<td>12.08%</td>
<td>8.80%</td>
<td>0.0034</td>
</tr>
<tr>
<td></td>
<td>13.90%</td>
<td>10.90%</td>
<td>0.0070</td>
</tr>
<tr>
<td></td>
<td>5.33%</td>
<td>4.30%</td>
<td>0.0028</td>
</tr>
</tbody>
</table>

Table 1. Comparison between our approach for assessing proportion in test images and the approach developed in [8]. The mean error, the mean error per species and the variance of the error are shown.

In order to evaluate the classification performance of the proposed approach, the objects class membership must be known. To this end, a labelled fish school data set is considered. Training and test images are generated by picking out randomly schools from the fish school data set. Images with one, two, three or four species are built allowing a reliable evaluation of the model behaviour. Species proportion is generated randomly too. Note that if there is one species per image, it leads to a supervised learning case.
A procedure was developed in [8] to infer the proportion in test images. Table 1 shows the improvement brought by considering our inference issue detailed in section 3 instead of [8]. School classification performance is shown in figure 3. In figure 3-(a) we first compare the classification rate as a function of $\beta$ defined in Eq.(10) for both the proportion case (triangle marker) and the binary case (square marker) and considering two-class mixture (two classes per images). The binary case, widely studied in semi supervised learning literature [2], considers a binary vector label indicating whether classes are present in images or not. We show in figure 3-(a) the improvement brought by considering proportion label instead. The best results are obtained when the inferred image-level prior is considered which proves the relevance of fusing image-level and object-level information for object classification. In the binary case, gains in excess of 5% are obtained when $\beta$ is going from $10^{-4}$ to $3.10^{-2}$. In the proportion case, gains in excess of 5% are obtained when $\beta$ is going from $1.5.10^{-3}$ to $4.10^{-3}$. Other results in figure 3-(b) and 3-(c) show the correct classification rate as a function of the mixture complexity when $\beta$ is fixed ($\beta = 4 \cdot 10^{-3}$). The mixture complexity refers to the number of class per image. There can be one (supervised learning), two, three or four classes per image deteriorating the performance when the class number per image increases. The improvement brought by considering weighted classification probability is obvious for both the proportion and the binary case. The proportion case outperforms the binary case for multi-species mixtures.

As a conclusion, we have developed a method for assessing a proportion prior in test images and a fusion has been carried out between the prior and a non linear DCM for objects classification. A comparison between a binary label and a proportion label has been established revealing the improvement brought by considering proportion labelled data. Then, we have shown that correct object classification rate has been considerably improved when mixing prior and classification models.

References