Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 00 (2012) 1–17

Abstract

Keywords:

1. Introduction

During the last twenty years, we have assisted to a exponential progress in digital cameras, in network bandwidth and in information storage capacities, which has led to a proliferation of visual data content in the form of images and videos. At the same time, powerful object detection and recognition approaches have been proposed [], [], [] which, however, are not able to scale up to many scenarios and objects due mainly to the lack of large labeled training sets. Indeed, as demonstrated in other research fields [**?**], [**?**] the performance of classifiers increases dramatically when a conspicuous set of labeled training data is available. Moreover, ground truth data plays a central role not only for the development of algorithms but also for quantitative performance evaluation, which has also received significant attention by the vision community with the aim to establish a valid reference for a systematic evaluation. Therefore, large scale annotated datasets, covering as much scenarios and objects as possible, are needed in order to train and evaluate the existing approaches. The main limitation to achieve this goal is the daunting amount of time needed to generate high quality ground truth data which requires a lot of human concentration, in fact it has been estimated that labeling an image may take from to two to thirty minutes, depending on the operation, and it is, obviously, even worse in the case of videos.

There exist, in the literature, a few attempts [], [] (which will be reviewed in the next section) performed by some vision groups that have collected consistent annotated datasets which however are too task-oriented and can not be generalised for any object category and scenario. To reach the objective of creating more diverse and larger annotated datasets, collaborative methods, exploiting large population of expert and motivated users, have been proposed [], []. One of the most relevant examples is LabelMe [], a web-based platform to collect user annotations with which a large set of annotations, for training and testing of

detection, segmentation and classification algorithms in still images, has been collected. However, the main shortcoming of LabelMe is the lack of in intelligent mechanisms to combine and integrate user annotations, in fact usually an user can delete the annotations provided by other users and create his/her own annotations. Moreover, LabelMe is thought only for image annotation, although a video based version has been proposed that, however, is not as successful and flexible as the image based version. Together with the web-based collaborative efforts, approaches for crowd sourcing the annotation effort to non-experts have been proposed [], []. However, these approaches lack mainly in mechanisms for testing the reliability of annotators, thence their annotations can not be fully trusted. Therefore, it appears rather evident which are limits of the current approach in this field. In this paper we propose a web-based approach which has two main objectives: 1) to propose an approach able to guide and to speed up the annotation phase and 2) to build up a large scale database of labeled visual data (image and video) to be used for a variety of tasks ranging from image enhancement to object detection and tracking to image classification, etc.

Two fold: a collaborative effort to collect ground truth data on videos and a new controlled dataset focused on an emerging topic in video-surveillance applications: environmental monitoring and animal behavior studied

Limiti generici delle soluzioni esistenti

Obiettivi della nostra ricerca

Variety of fish species makes it a multi class classification problem. Aims:

The proposed system supports multiple levels of labeling: 1) Scene Level, Object Level and Pixel Level.

Descrizione del dataset F4K e quale impatto ha sulla ricerca sia per la collection di ground truth che per gli algoritmi di detection e classificazione Challenging datasets are catalyst for progress in computer vision.

We have available a large dataset of videos taken in underwater environment and we wish to collect a large dataset of ground truth labels for detection, tracking and recognition purposes

Underwater environment:

- Not many detection, tracking and recognition algorithms have been developed to deal with unconstrained environments;
- Underwater scenario is a challenging one for developing/testing detection, tracking and recognition algorithms. Add Features

The remainder of the paper is as follows. Sect. **??** reviews the existing approaches for collecting ground truth data together with the most used datasets of visual data. Sect. **??** describes the proposed web based approach to collect large scale ground truth data for different visual tasks. Sect. **??**, instead, shows the application of the propsed system to the aforementioned underwater scenario, providing some statistics on the collected data. Finally, in Sect. **??** concluding remarks and ideas for future developments are given.

2. Related Works

Many available datasets contain only a small number of objects and classes and the ones that have large number of objects/classes lack in quality since they are hand-labeled from users on the web and the mechanisms for quality assessment are not effective

Main features of the existing approaches (PETS, LabelMe, CalTech, Berkeley Segmentation Dataset, PASCAL collection, CAVIAR, etc...)

1. Instance recognition
2. Low quality labeling
3. Copyrighted-images/videos
4. Static scenario: it is useful to vary the scene type, distances, degree of clutter, etc..
5. Portability of the annotation tool
6. interoperability (XML Schema of the GT)

Categorization of the existing approaches:

- Primi approcci non collaborativi
  CAVIAR: resolution half of PAL (384x288) : ground truth consists of hand-labeled bounding boxes surrounding objects together with a label indicating object activity (e.g. walking, running, etc..) Limitations of existing approaches: OpenMind Initiative: description and limitations

  Compatibility with existing labeling tools: ViPer and ODViS but not user friendly

  ATTENZIONE: non confondere i metodi per collect GT con i dataset

  i-LIDS: event related ground truth.

  AVSS: This page provides publicly available benchmark datasets for testing and evaluating detection and tracking algorithms. The datasets are free for research and educational purposes only and can be used in scientific publications at the condition of respecting the requested citation acknowledgment. Caltech 101 and 256: They are not designed to learn objects in cluttered scenes, i.e. they consist of small cropped images with limited viewpoints. They work fine for training patch-based object detections while they are not suitable for detections which use contextual information. The Berkely Segmentation Database: limited in regards of scale and content

- Web based approaches LabelMe: rough annotation of object boundary, it does not support or better it does not integrate in an intelligent way the annotations of different users

- Approcci per il crowdsourcing

- Metodi per la generazione automatica del GT
  Semi-automatic approaches still require the user to manually label the obtained results.

3. Detection and Tracking Ground Truth

3.1. Goals of our project

Here a description of our platform for generating ground truth (hand-labeling, crowdsourcing), testing algorithms performance and validating results.

3.2. Hand-Labeled GT on Underwater Videos

3.2.1. Web-based Platform for Hand-Labeling Ground Truth

One of the most widely used tool for generating ground truth on videosequences is ViPer. More specifically, it allows frame-by-frame markup of video stored in a specific XML format.
Improvements with respect to the ViPer
Semantic Information on detected blobs

- Functionalities of the web-annotation tool:
    1. Description on how to use it:
        (a) new blob creation;
        (b) blob deletion
        (c) contour extraction (both manually and automatically)
        (d) semantic label association
    2. Interoperability with existing tools
    3. XML Schema for GT
- Combing GT of different users
    1. Overlap Score: PASCAL VOC
    2. Euclidean Distance Score
- Ground Truth Quality Ranking. Annotation Scoring Functions
    1. Control Points
    2. Annotation size
    3. Edge Detection
    4. Bayesian Matting
    5. Active/Statistical Shape Models

3.2.2. Content of the collected GT

3.3. Performance Evaluation (PE)

VIVID evaluation website: not web-platform

- PE of Detection Algorithms on :
    1. Hand-Labeled GT Data;
    2. Crowsourced GT Data
- PE of Tracking Algorithms

4. Recognition Ground Truth

The goal of this section is to obtain a groundtruth annotation for fish recognition. More specifically, we want to know which fish images belong to the same species, along with the species names. This allows us to train and evaluate our fish recognition methods in the F4K project. In order to support the manual labelling of images, we propose to use an automatic clustering method to groups and retrieve similar images, which allows us to label large dataset in an efficient manner.

This section is organised as follows: We start with a discussion of the fish clustering method that is used to support the labelling task. We then explain how we combine the clustering method with the annotation interfaces to support manual annotation work. Two types of annotation interfaces are defined: i) the interface for normal annotators to label fish images; and ii) the interface for marine biologists to verify the obtained labels. We close this section with a discussion of the quality of the labels obtained.

I changed the subsection titles here, the original one was “First method to determine the similarity between fish”. It’s confusing: why is it a “first” method? as there doesn’t seem to have a “second” method ....

4.1. Fish Clustering

4.1.1. Measuring similarity between fish images

The fish clustering method starts with the assumption that the segmentation of the fish is correctly performed. In the case that we do not have groundtruth segmentation data, we can use visual inspection to manual remove failures in the segmentation from the dataset, this migt however be very time consuming.Do you mean manually remove the wrong segmentations? What’s the relation of this to the similarity measure?

It is confusing here whether we are talking about similarity measures/fish representations or clustering method.I thought we are talking about similarity measures/fish representations... In order to compare fish, we need a method which is also able to compare unseen fish species with already know species in the dataset. The similarity method needs to be invariant against a lot of variantions because of the uncontrolled nature

of the video recording. The last requirement on the method is that it must be able to deal with objects which are quite similar, as opposed to most methods in image retrieval where the classes with are quite dissimilar (car, building, people, etc). To compute the similarity between images, we use a method that is very similar to the method described in [**?**].

The feature used for fish recognition are the color of the fish, the texture of the fish and the fish contour. In the case of color, we transform all the pixel values of the segmented fish to HSV (Hue, Saturtion, Value), for the Hue channel two values are used, namely the sine and cosine of the Hue channel. This removes the big difference between the different red values because of the cylindrical color definition. Using all the 4 dimensional pixel values $X = \{x_1, \ldots, x_N\}$ in the segmented image, we fit a Gaussian Mixture Model (GMM) $f(x)$, using the Expectation Maximalization algorithm described in [**?**], which uses Minimum Description Lenght to automatically determine the number of Guassian density functions. For the texture, a GMM is fitted to the magnitude and the orientation of the Canny filter at each pixel in the segmented fish. We also fitted a GMM on the curvature scale space representation of the fish contour, where we use the points on lines given by the curvature scale space of the fish contour as input vector to fit a GMM. The advantage of using the curvature scale space is that it gives an invariant representation for most affine image transformations on the contour. We use Gaussian Mixture Model (GMM) because they give us an invariant representation which is able to make a model of the presented data, allowing the method to discover new species without needing models of these species by forehand. In [**?**], the similarity between two GMMs (f_1, f_2) is determined by using a Monte-Carol simulation to approximate the Kullback-Liebler divergence:

$$D(f_1 \| f_2) = \int f_1 \log \frac{f_1}{f_2} \approx \frac{1}{N} \sum_{t=1}^{N} \log \frac{f_1(x_t)}{f_2(x_t)} \tag{1}$$

$$S(f_1, f_2) = D(f_1 \| f_2) + D(f_2 \| f_1) \tag{2}$$

We use the symmetric version of the KL-divergence (Equation 2) to measure the similarity between different fish images. Given the three different kind of features (colour,texture and contour), we obtain three different KL-divergences which we sum together to get the final similarity measure.

4.1.2. ***Clustering based on similarity between objects***

In order to compute clusters based on the similarity between objects, we use Affinity Propagation [**?**]. Although in [**?**], a method for clustering is described to combine the GMMs, the Affinity Propagation method in [**?**] scales better on large databases. The Affinity Propagation is a graph based clustering method. This clustering method passes messages around to determine the responsibility and availability. The message of responsibility between point i and k reflects how well suited point k represents point i. The availability message between point i and k indicates how appropriate it is for point i to choose point k as

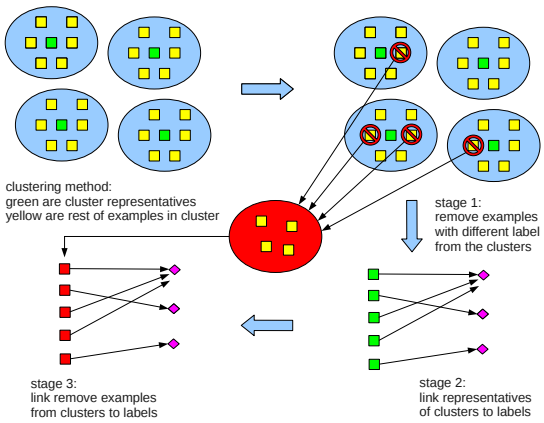

Figure 1. A schematic representation of the method we use to the annotate images with the support of a clustering method

representative. The final outcomes of the clustering method are clusters of fish images which have one image as a representative image.

The representative image is important in our interface for normal annotators because we use this image to link the clusters together.What does "link the clusters together" mean? See below. However, although [**?**] gives us representative image for each cluster, this does not mean that other clustering methods can not be used. For instance by using K-means clustering, it is easy to select the images closest to the mean of the cluster as representative; in addition, using a random images in the cluster as representative should also work.

4.2. Manual Groundtruth Annotation using Automatic Clustering

In our project, we would like to obtain the names of different fish species that we observe in the underwater cameras. However, labelling thousands of images by hand is a time consuming process. For clarity, in this paper we use the definition "cluster" for a group of images which are similar as determined by an automatic algorithm. The definition for "label" is a group of images which belong to the same category from the perspective of the human annotator and this group contains all the images in this category. Hence we aim to improve this procedure by using a clustering method. Instead of giving a label for every fish, it is much faster and easier to check if a fish image is similar to another fish image. Note that the task of the annotator changes from typing in species name for each image to judging images clusters. Although the latter task can still be difficult, the images are processed in a batch mode and it does not require as much domain knowledge as the former task.

In order to label an entire dataset of images using a clustering method, we have developed a strategy which consists of three steps:

1. Cleaning the cluster, where we remove images which are not similar to the representative image

Figure 2. This is the first interface for annotation that allows annotators to click on the images in the cluster that do not below to the same label as the representative image on top of the page

2. Merging the clusters, using the representative image of the cleaned clusters to link them to labels
3. Linking removed images from the cleaning stage to the labels.

For fish, this means that a label includes all fish of a certain species in the dataset. Terms "cluster" and "label" alrady occured in the begining of the section. Are they refer to the same concepts? If so, it's better to explain them ealier; if not, it's better to change the names in previous text.

In Figure 1, we show a schematic representation of the proposed method to annotate images. The blue oval describe the clusters, where the green images are the representative images in the cluster. Our method consists of three stages, in the first stage we clean the clusters by removing the images that do not have the same label as the representative image. In the second stage, we link the represenative images to a label (shown as purple diamonds). In the third stage, we link the images which were removed from the cluster (shown as red squares) to a label. However, whether one needs to perform the last stage depends if you want to label all the images in your dataset or if a large subset of all the images is sufficiet. For comparison reasons we perform this stage, in order to make our method comparable with labeling every image in the set individually. In the next sections, we will discussed the steps in more details together with the interface design.

4.3. Annotation interface for non-expert annotators

For the first stage, we use the cleaning interface shown in Figure 2. In this case, the representative cluster image is the image on top in the interface and the rest of the images in that cluster are shown under this images. The annotator only has to select the images which are not correct and continue to the next window. In our interface, we assume that people can deal with around 30 gallery images in a screen at one time (more images often requires extra work such as scrolling down). Therefore our clustering method tries to obtain cluster of around 30 images. After cleaning all the clusters, there are basically three kinds of images in the dataset: the representative cluster images, the images that belong to a cluster and the images that are not part of a cluster.

In the second stage, we are going to link the clusters to labels using the representative image . Notice that by linking these images, we also immediately link the images that belong to the underlying clusters. In many cases, some initial examples of labels are already in the database, which can be used in the interface shown in Figure 3. In this case, we simply pick the first representative image as a label. The annotator

Figure 3. This is the second interface for annotation that allows annotators to link a image to a label by clicking on one of the gallery image which belongs to the same label as the image on top of the page or add a new label by pressing the green plus button

will then use the linking interface shown in Figure 3 to link the next representative image to this first label image. However, it is also possible that there are no initial examples. In this case, the annotator will create another label by pressing the green plus button. In the first case, the clusters are under the same label and linked. In the second case, a new label and representative label image are created. As mentioned in the previous step, ideally we would show 30 images per screen, so that the annotators do not have to perform additional actions such as scrolling down. Similarly, in this step, when there exist more than 30 labels, it becomes hard to show everything in one screen. Here, we rank the labels in the descending order of their similarity to the top image.

In the third step, we link the set of images that are not part of a cluster to a label?. In this case, we use the same interface as in the previous step to link also these images to a label. Note that it is possible that there are images in this set that do not belong to any label yet. (A speed-up can be achieved for large dataset if we recluster the images and use both the first and second interface, however from a user-perspective switching these different interfaces can be confusing so in this work we only use the second interface to link them to clusters.)

4.4. Annotation interface for expert annotators

Since the final goal of the annotation is to assign a species name to each of the fish images, beside non-expert annotators' annotation over the fish clusters, we also need expert annotators to assign species names to each of the clusters. As said before, experts are expensive and rare resources. Hence we use expert annotators to annotate only a subset of our data. The images annotated by the experts can be used as a validation set for the non-expert annotation.

The interface for expert annotators is similar to the first stage interface for the non-expert annotators. See Figure 4. Using this interface, the expert annotator first enters the species name that applies to the majority of the images in a cluster in the top-right text box. Once the name is entered, all images within the cluster are automatically assigned with the same species name. Then, the annotator is asked to select those images that do not belong to the cluster. By selecting these images, he/she can input the correct species names for them in the text box under each image. In this manner, in the worst case, the annotator will have to manually assign a species name to each of the images, i.e., when the clustering is so bad that each image within a cluster contains a different fish species. In the best case, i.e., when the cluster is pure, the annotator only needs to enter the species name once.

Figure 4. The interface for expert annotators. The expert enters the species name for the majority of the images within a cluster, then select images that should not belong to this cluster and input the correct species names for these images.

Figure 5. In each cluster, the percentage of the images where biologists disagree on their species names. Here "disagree" refers to the situation where each biologist assigns a different name to the image, or all the biologists indicate that they cannot identify the species.

After finishing annotation, we also include a questionnaire for the expert in order to collect information such as which features the expert used to identify certain species. This type of information may be a useful hint for selecting useful features when developing automatic methods for fish recognition.

In our experiment, we invited 3 marine biologists, who have over 10 years research experience in the area where the underwater video cameras are located. In order to obtain relatively high quality clusters, we manually constructed clusters over a small sub set of our dataset: 27 clusters over 524 images. Since the sizes of clusters are very imbalanced, for each cluster, we randomly sampled 30 images to be shown to the biologists. In total 190 images were annotated by the biologists.

For 82.6% of the images, at least two biologists agreed on a species name; for 56.3% of the images, all biologists agreed on a name (including the cases where two biologists agreed on a species name while the third biologist indicated that he/she cannot identify the fish). Further, if we look at the biolgoists' annotation per cluster, we see that for different clusters, or in other word, different potential species of fish, biologists have different levels of agreement on their species names. From Figure 5 we see that for 9 out of 27 clusters the biologists cannot agree on a species name for all images. A closer check-up on the annotations shows that for 7 out of the 9 clusters, although the biologists do not agree on the species names, they do agree on their family or genus names. This suggests that for some fish images, it may be too difficult to recognize the species of the fish, while a family or genus level recognition may be more practical.

Futher, we compare the manual clustering results to the biologists' annotation in order to check the performance of non-expert in grouping fishes into species based on their visual similarity. In Figure 6 we show the percentage of "errors" the non-experts made according to biologists' judgement, where an "error" refers to the case when at least two biologists indicate that a image should be removed from a cluster. We

Figure 6. In each cluster, the percentage of the images that at least 2 biologists suggest that these images should be removed from the cluster.

see that in most of the cases, 21 out of 27 clusters, the groups of fishes created by the non-experts are approved by the biologists, which suggests that it is promising to use non-experts to identify fish species (without actually knowning the names of the species) based on visual similarity. Another observation we have here is that, if we compare Figure 6 to Figure 5, we see that in many cases, while the non-experts have low error rate in identify a cluster, the biologist have difficulty to agree on a species name for the cluster. For instance, in the case of cluster 3, 4, 5, 6, etc,. This observation suggests that although it is promising to cluster fish images into visually similar groups, it is not necessarily easy to associate this cluster to a specific label, in particular at the species level.

@Bas: why in my judgements 190 imgages are labeled by the biologists, but you only have 159? Did you do some sort of filtering? If we count the images that have at least two biologists agree on a name, then it's 157.

In case you are confused when checking the database: in the database the cluster version 1 has 28 clusters, however, cluster 16 is empty. In the paper, I removed the empty cluster, so from cluster 17 on, all the cluster id corresponds to the cluster id-1 in the database

4.5. Obtained annotations

I think we need to at least briefly discuss how many images we annotated, how much time it takes, how we arrange people to do this (e.g., at least 3 non-expert to annotate), disagreements, evaluation results on the experts' subset, etc.

At the moment, we have a first dataset of 3678 fish images, where 159 fish image have been given a species name by the marine biologists. We found 6 users willing to annotate the entire database of fishes for us using the clustering method to support them in their efforts. Based on the labeling of the biologists, we found out that the average user performance is 87.6% correctly labelled fish. There is however a large difference between people who saw the fish images for the first time and people who are part of this project having observed some of the fish before. The lowest user performance is 68.8%, where this person basically annotated different species that look similar to the same category. For users, it is often very difficult to determine if fish belong to a different species or not, because appearances of the fish like the colour can change due to illumination conditions. Another difference between users is their decision to ignore the fish image because you can not clearly identify fish basd on the images or the images contains multiple fish. Some users used this ignore options very frequently, while other users used it rarely.

We combined the different labels given by all the users using the probabilities that the user correcty labelled the fish measured using the biologists annotations. If user agreed on the label the probabilities become very high while with disagreements results become much lower. In Figure 7, we show the distributions on the user's disagreements, in most cases however user do agree which can be observer in the last bar. In more than 90% of the images, the probability of being correctly labelled is greater than 99.9%.

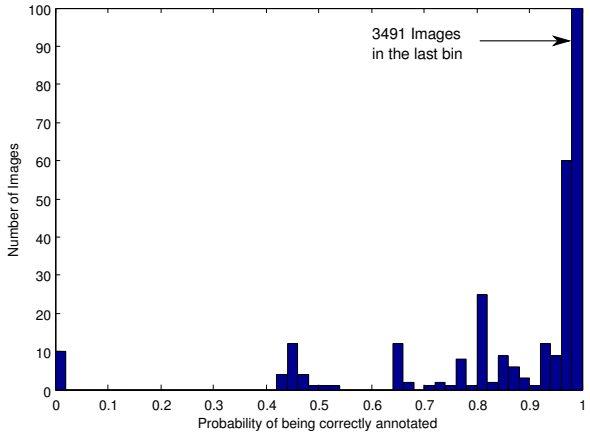

Figure 7. We show the distribution on the probabilities that a images labelled by all users is correctly, there are in this database however still a lot of disagreements between user. This images can be communicate to the marine biologists, for now we exclude them from the dataset for training model and evaluation of our methods.

We also look at the profit that we obtained by labelling using a clustering methods. Let us assume that a user labels everything correctly. For the first interface, we need $\frac{M}{30}$ screens, assuming M is the number of image and that we show around the 30 images. For the second interface, we need screens to link all the clusters, where C is the number of clusters and we need to label all the images that do not have the same label as the representative images, let us denote this by the number R, where for our clustering methods this was around $R = 300$. In order to measure the time it takes a user to label, we asked one user to label images non-stop for us, where the average time in which he took annotating a screen in stage 1 is $T_1 = 19.7$ seconds and for stage 2 this was $T_1 = 7.3$ seconds. Given the Equation 3 below we can easily estimate how much time it will take users, notice that labelling the entire dataset using the second interface will take MT_2.

$$time = \frac{1}{30}MT_1 + CT_2 + RT_2 \tag{3}$$

$$clicks = \frac{1}{30}M + 2C + 3R \tag{4}$$

Equation 4 estimates the number of click necessary to label the entire database, In the first interface user need to click on very incorrect image R and to go the next screen users also need to click. We have add this to the number of mouse clicks required for the second interface which is two mouse clicks, one to select the images and one to go to the next screen. If we have to annotate all screen using the second interface we end up with $2M$ clicks, based on these equations it can be shown that with a good performing clustering method (determining by the value R) large gains in time and number of clicks can be obtained

for annotating a dataset.

4.6. Clustering Method

The fish clustering method start from the assumption that the detection and segmentation of the fish is correctly performed. This is not neccessary, however, in previous section we already obtained groundtruth data for detection and segmentation, which we use for the fish clustering methods. In the case that we do not have groundtruth data we can use visual inspection to remove failures in the segmentation from the dataset, this is however more time consuming.

In order to cluster fish, we need a method which allows us to discover new species in the dataset. The clustering method also needs to be invariant against a lot of variantions because of the uncontrolled nature of the video recording. The last property of the clustering methods is that it will be able to deal with objects which are quiet similar, as suppose to most methods in image retrieval with cluster classes with are quiet dissimilar (car, building, people, etc). Our clustering method is very similar to the method describe in [**?**], especially in the way that this method measures the similarity between images. In order to cluster the images, we use Affinity Propagation [**?**] instead of [**?**], because this is faster and more scalable with larger databases.

In order to represent every image, we fit a Guassian Mixture Model (GMM) to each of the seperate fish feature. The GMM are fitted using the Expectation Maximalization algorithm described in [**?**], which uses Minimum Description Lenght to automatically determine the number of Guassian density functions. We use GMM because they allow to compare unseen fish feature. In new cases, a new feature (like a certain color) will be described using its own Guassian density functions. In order to model fish, the most important features of fish are the color, texture (spot,strips,etc) and shape. In the case of color, we transform all the pixel value of the segmented fish to HSV (Hue, Saturtion, Value), for the Hue channel two values are used, namely the sine and cosine of the Hue channel. This removes the big difference between the different red values because of the cylindrical color definition. From these 4 dimensional data points, we then obtain a GMM using [**?**]. For the texture, a GMM is fitted to the magnitude and the orientation of the Canny filter at each pixel in the segmented fish. We also fitted a GMM on the curvature scale space representation of the fish contour, where we use the points on lines given by the curvature scale space of the fish contour input data points. The advantage of using the curvature scale space is that it gives an invariant representation for most affine image transformations on the contour.

In [**?**], the similarity between two GMM is determined by using a Monte-Carol simulation to approximate the Kullback-Liebler divergence:

$$D(f\|g) = \int f \log \frac{f}{g} \approx \frac{1}{N} \sum_{t=1}^{N} \log \frac{f(x_t)}{g(x_t)} \tag{5}$$

$$S(f,g) = D(f\|g) + D(g\|f) \tag{6}$$

We use the symmetric version of the KL-divergence (Equation 2) to measure the similarity between different fish images. Given the three different kind of features, we obtain three different KL-divergence which we sum together to get the final similarity measure. Instead of using the clustering approach proposed in [**?**], we use Affinity Propagation [**?**] which scales better on large image databases. This is a graph based clustering method, which can automatically find clusters given a matrix of similarity measurements. This clustering method passes messages around to determine the responsibility and availability. The message of responsibility between point i and k reflects how well suited point k to represent point i. The availability message between point i and k indicates how appropriate it is for point i to choose point k as representative. The final outcome of the clustering method is are clusters of fish images which have one images as a representative image. The representative image is important in our interface because we use this image to link the clusters together. Although the clustering method should not matter, speed up in the manual annotation can be achieved if the clusters and the similarity between fish are of better quality. Because [**?**] gives us representative image for each cluster, this does not mean that other clustering methods can not be used. For instance by using K-means clustering, it is easy to select the images closes to the mean as representative, but as using a random images in the cluster as representative should also work.

4.7. Manual Annotation using Automatic Clustering

Manual annotation of image for the task of recognition can be time consuming work. In our project, we would like to obtain the different species which we observe in the underwater cameras. However, labelling thousands of images by hand is a time consuming operation, so we improve this task by using a clustering method. Instead of giving a label for every fish, we agrue that it is much faster to check if a fish image is similar to another fish image. Notice that the task of the user change from typing fish names to judging images. Although this task can still be difficult, it does not require as much domain knowledge as the previous task.

In order to label an entire dataset of image using a clustering method, we have developed a strategy which consists of three steps:

- Cleaning the cluster, where we remove images which are not similar to the represenative image
- Merging the clusters, using the representative image of the cleaned clusters to link them to labels
- Linking removed images from the cleaning stage to the labels.

In this paper, we use the definition cluster for a group of images which are similar determine by a automatic algorithm. The definition for label is a group of images which are similar to the human annotator and this group contains all the similar image in the entire set. For fish, this means that in a label we obtain all fish

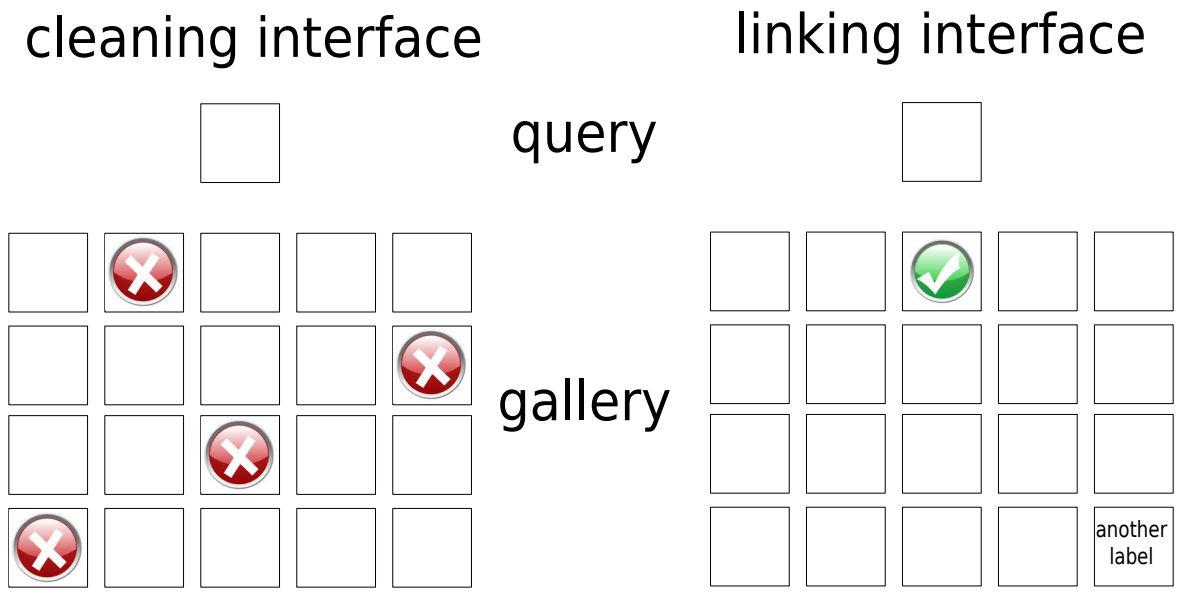

Figure 8. This are the two interface which we use to annotated the images, the first interface allows us to clean the clusters by clicking on gallery images that do not below to the same label as the query image, the second interface is to link the clusters together allowing a person to click one gallery image with the same label. In the second interface, we assume that there are no images with the same label in the gallery and that it is possible to indicate that there is no image with the same label in the gallery using the Another Label button (right-bottom)

of a certain species in the set.

For the first step, we use the cleaning interface shown in Figure 8. In this case, the representative cluster image is the query image in the interface and the rest of the images in that cluster are put in the gallery. The user only has to select the images which are not correct and can continue to the next window after finishing this task. In our interface, we assume that people can deal with around the 30 gallery images in a screen at once (more images usually requires actions like scrolling down which is also extra work). So our clustering method trieds to obtain cluster of around the 30 images and people can select the images which are not part of the clusters. After cleaning all the clusters, there are basically three kind of images in the dataset. The representative cluster images, the images that belong to cluster and the images that are not part of a cluster.

In the second step, we are going to link the clusters using the representative cluster images to labels. Notice that by linking the clusters, we also immediately link the images that belongs to that cluster. In order to link the representative cluster images, we are going to link all the representative cluster images to representative label images. In many cases, some initial examples of labels are already in the database which can be used to link to the representative cluster images. However, it is also possible that there are no initial examples.

In this case, you can just pick the first representative cluster image as first representative label image. The user will then use the linking interface shown in Figure 8 to link the next representative cluster image to either first representative label image or to another label. In the first case, the clusters are under the same label and will be linked. In the second case, a new label and representative label image are created. In the previous step, we already mentioned that we let users deal with around the 30 gallery images, however if there are more than 30 labels, it becomes hard to show everything. In our case, we are able to compute the similarity between images, allowing us to compute which representative label image are most likely to be the representative cluster image.

In the thrid step, we link the set of images that are not part of a cluster. In this case, we use the same interface as in the previous step to link also these images to a label. Notice, that it is possible that there are images in this set that do not belong to any label yet. (A speedup can be achieve in this step if we use a combination of the first and second interface. There are probably a lot of images that match almost equally well to two label and are just in the first stage cluster with the wrong label. Using the first interface to link them to the second most probable label can achieve a speed up, however from a user perspective switching these different interfaces can be confusing so in our initial test we only use the second interface to link them to clusters.)

4.8. Simulation Program

We create a simulation program to see how much windows and clicks one needs to label a set of 3678 fish images, from which we already obtain the labelling. In a lot of labelling programs, we label each image individually so we need to view 3678 windows, type in the correct name or select the name from a dropdown box. We usual need to confirm our finding with a click which takes us to the next window. If we improve this interface slightly for our problem, we can use basically the linking interface in Figure 8 to label each image seperately. This will take 3678 windows but only 3678×2 mouse clicks assuming that you have to click on the correct gallery image and a confirm button to go to a next window. Here, we did not take into account that you have more than 30 labels you need more screens anyway.

In our new program, we want to minimize the number of screens and clicks. For the cleaning interface, this means that the users normally click all incorrect image and gives a final click to conform and go to a next screen. For the linking interface, we already explained the procedure needs only 2 clicks for each window. In the case, that we combine our new proposed interface with random selection of cluster and similarity score between images, we need a total of 6535 clicks and 1358 screens to annotate the entire database. In the case, that we use the clustering method describe in Section , we only need 1558 clicks and 621 screens. This is a large improvement of the previous proposed interface and the method without using clustering. Since a lot of the crowdsource website let you pay for the number of clicks a user have to perform these kind of strategies might be an interesting tools to get good results in a cheaper way.

4.9. Possible Todo List

- combining labeling ... very easy with simulation program at the moment
- majority voting ... disagreement solutions .... can use simulation programs
- different kind of users mistake .... random and systematic
- how to use experts? for instance to check final labels or solving user disagreements or testing user accuracy
- creating a real interface for labelling (Jiyin on holiday, until end this month)
- different database .... (flower database)
- bayesian network to describe label certain (or other model to describe certainty and quality of normal label and labeling with our new approach)
- write paper for Concetto and CVPR

majority voting ... disagreement solutions .... can use simulation program keeping expert in the

5. Concluding Remarks

Here some discussion on what we have achieved and on our future plans.