Supporting ground-truth annotation of image datasets using clustering

Boom, B. J., Huang, P. X., He, J., & Fisher, R. B. (2012). Supporting ground-truth annotation of image datasets using clustering. In Pattern Recognition (ICPR), 2012 21st International Conference on. (pp. 1542-1545). IEEE.

Supporting Ground-Truth annotation of image datasets using clustering

Bastiaan J. Boom[1], Phoenix X. Huang[1], Jiyin He[2], Robert B. Fisher[1]
[1]*School of Informatics, University of Edinburgh*
[2]*Interactive Information Access, CWI, Amsterdam*

Abstract

As more subject-specific image datasets (medical images, birds, etc) become available, high quality labels associated with these datasets are essential for building statistical models and method evaluation. Obtaining these annotations is a time-comsuming and thus a costly business. We propose a clustering method to support this annotation task, making the task easier and more efficient to perform for users. In this paper, we provide a framework to illustrate how a clustering method can support the annotation task. A large reduction in both the time to annotate images and number of mouse clicks needed for the annotation is achieved. By investigating the quality of the annotation, we show that this framework is affected by the particular clustering method used. This, however, does not have a large influence on the overall accuracy and disappears if the data is annotated by multiple persons.

1. Introduction

One of the most common problems given a newly acquired dataset is to attach labels to this dataset and often (especially in medical imaging) experts are needed to determine these labels. A similar problem is how to obtain groundtruth classifications (e.g. fish species) for a large dataset of images (e.g. underwater images of fish), where to guarantee the quality we would like a certain number of users to annotate the images. By supporting this task with a clustering method we solved two problems at the same time: Firstly, by translating the task from recognizing fish species to cluster validation, the expert knowledge needed is greatly reduced. Secondly, by clustering the images using computer vision features, the annotation process is more efficient.
Previous approaches for the annotation of a set of images are, for instance, the ESP Game [7] and LabelMe [6]. Most of this work is focused on a large variety of internet images where often multiple tags can be given to these images. These tools are useful in the case of random internet images, but are not efficient for solving annotation problems where we want to obtain a single specific label for an image. Recent work more suitable to this problem involves the annotation of a bird database, where users label certain properties of a bird like the color of tail, wings, beak [9]. This focuses on subject-specific image datasets, however it might not be a very efficient way of annotating images as multiple properties have to be assigned to each image.
Alternatively, there are approaches which combine user annotations and machine learning to obtain the groundtruth labels, for example, [5], but this does not speed up the annotation task. Another approach that relies less on the automatic methods and allows users to search and annotate images at the same time is [8]. These approaches are developed for internet images on the web and need all labels to be defined apriori.
In our approach, we want to annotate all the images in the dataset, where we focus on subject-specific datasets. This allows the use of specific domain dependent features to cluster these datasets, which can support the annotation by users or experts. This has not been attempted before to our knowledge. Our approach explores both the Kullback-Liebler divergence [3] and Pyramid histogram [1] for the clustering of automatically segmented fish images (Section 2.1) after which individual annotators refine and group clusters using two specialized interfaces (Section 2.2) and then the result of multiple users are combined in (Section 2.3). Experiments show a reduction of up to 77% of annotation time and 93% of mouse clicks while maintain accuracy(Section 3).

2. Ground-truth annotation using automatic clustering

2.1. Fish Clustering Methods

To cluster the fish images, two methods for obtaining features and calculating a distance measure between

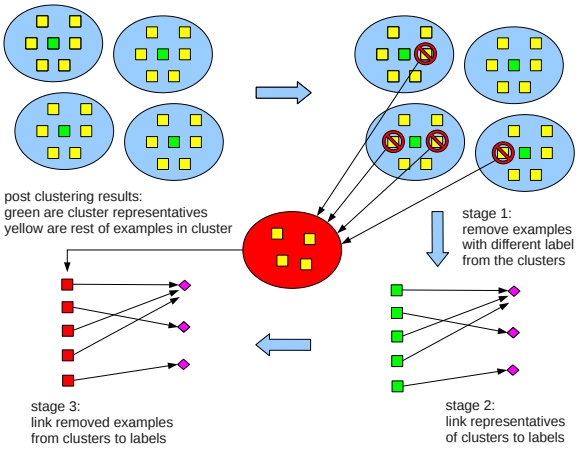


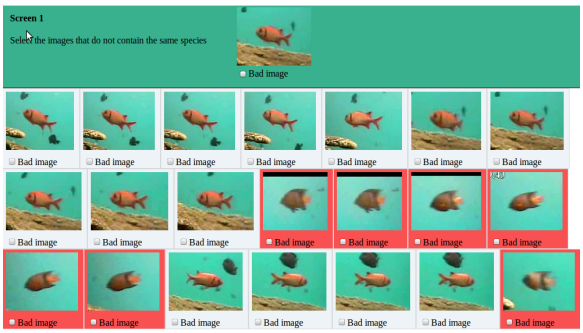
Figure 1: A schematic representation of the framework for annotating images with the support of a clustering method

fish images were used: The first method [3] computes the Kullback-Liebler divergence (*KL divergence*) between feature sets. For the fish, we create sets of color, texture and contour features (where we respectively used the Hue/Saturation/Value, the Canny edge detector and the Curvature Scale Space representation). A Gaussian Mixture Model (GMM) is estimated from the set of features of a fish (for instance the color) and KL divergence is computed between two GMMs. The second method [1] uses a pyramid histogram of visual words (dense SIFT features with color information). These histograms describing each fish are normalized and the Euclidean distance between the histograms are computed. To compute clusters based on these distance measures, we use Affinity Propagation [2] which also provides a representative image for each cluster. The representative image is important because we can represent a cluster by a single image.

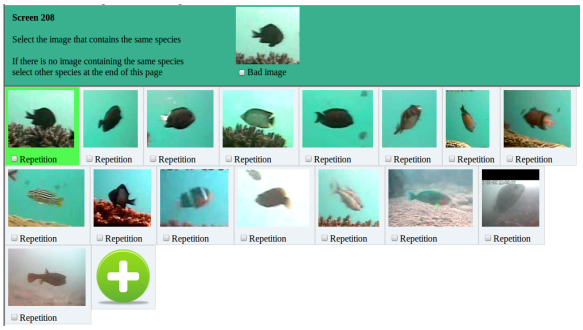
2.2. A Cluster-based Annotation Framework

Manual annotation of thousands of images for the task of recognition can be time consuming. Efficiency is improved by using a clustering method. Instead of giving a label for every fish, the user verifies that a fish image is similar to another fish image. Thus the task of the user changes from entering fish names to judging the estimated similarity between images. Although this task can still be difficult, it does not require as much domain knowledge as the previous task.

The framework to label an entire dataset of images using a clustering method consists of three stages (Figure 1 shows a schematic of this framework):



(a) The first interface to remove images from the cluster by clicking on the image that does not belong to the same label as the representative image in the top row



(b) The second interface to link the image in the top row to a label by clicking on one of the gallery images which belonging to the same label or add a new label by pressing the green plus button

Figure 2: Interfaces

1. Cleaning the clusters (blue ovals in Figure 1), where we remove images which are not similar to the representative image (green square).
2. Merging the clusters, using the representative images of the cleaned clusters to link them to labels (shown as purple diamonds)
3. Linking removed images (shown as red squares) from the cleaning stage to the labels.

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. In the case of fish, this means that a label includes all fish of a certain species in the dataset. Note that the particular species name is not necessary at this stage and can be added afterwards to the labels by a domain expert.

For the first stage, we use the cleaning interface shown in Figure 2(a). In this case, the representative cluster image is shown at the top of the screen and the rest of the images in that cluster are put under this image. The user only has to select the images which are not cor-

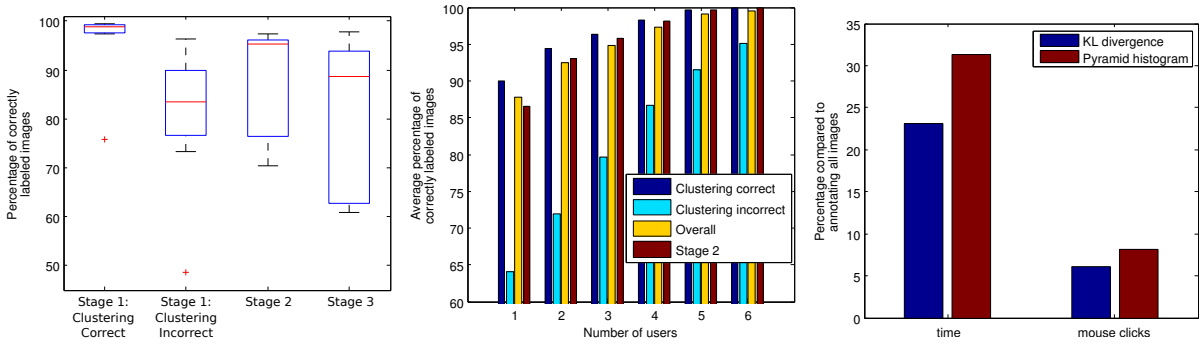
rectly clustered and continue to the next window. After cleaning all the clusters, there are basically three kinds of images in the dataset: 1) The representative cluster images, 2) the images that belong to clusters and 3) images that are not part of a cluster.
In the second stage, users link the clusters to labels using the representative images. This is because, for improved cluster coherence, we overcluster (e.g. 156 clusters for 32 labels) and therefore need to merge clusters. Notice that by linking these images, we also immediately link the images that belong to the underlining clusters. The second interface shown in Figure 2(b) is used to link the representative image either to one of the previous representative images of a label or the user will create a new label by pressing the green plus button. In the first case, the cluster is categorized under the same label. In the second case, a new label and representative label image are created.
In the third stage, we link the set of images that are not part of a cluster, using the same interface as in the previous step to also link these images to a label. In this work, the final goal is to label an entire dataset which is comparable with the normal labeling task of annotating each image individually. It is however possible to skip stage 3 or in case of a very large datasets, it is possible to recluster the images which are removed from the clusters in stage 1, which may speedup the annotation even more.

2.3. Combining Multiple Annotators

The problem of combining the annotations from multiple users is discussed in [10] and [4]. In the framework describe by [10], we have an observed label L_{ij} for each image j of the M images given by each user i of the N users. The expertise (accuracy in annotation) of user i is modeled by the parameter α_i and the difficulty of the image j is given by the parameter β_j. The groundtruth image label is denoted by Z_j. In [10], Expectation-Maximization is used to infer both α_i, β_j and Z_j given the observed labels L_{ij}. In [4], the labels from an expert are used to estimate α_i and β_j on a small number of images that this expert also annotated, from which we can compute β_j and Z_j on the remaining images. We extended the work of [4] from two classes to support multiple classes, which allows us to find groundtruth labels for all images.

3. Experiment

We empirically investigate our proposed framework with a dataset of 3678 automatically segmented fish images obtained from underwater surveillance cameras with 32 different fish species in the dataset. The dataset is annotated by 6 users using the KL divergence and by 2 of the 6 users again using the Pyramid histogram. A part of the dataset (159 images) is also labeled by marine biologists, allowing us to obtain the groundtruth (using [4]) by combining all the annotations.
Figure 3(a) shows the accuracy at each of the stages. Because the annotation in the first stage depends on the clustering performance, this stage is divided into two boxplots. It is clear from these boxplots that users make more mistakes with removing the incorrectly clustered images than with correctly clustered images. We assume that this has two causes: The first cause is that users do not scan the images very comprehensively, which leads to labeling mistakes which could be avoided. The second cause is that some images are hard to recognize and users might not be able to separate them correctly. The performance in stage 2 (see Figure 3(a)) is a good indication of the labeling performance without using clustering, because stage 2 has the user select a pictorial "label" for each presented image. In our case, we only present the representative images rather than the full set, but we argue that accuracy would be similar if all images were presented. From the performance of stage 3, we observe that it is also more difficult to link the images excluded from stage 1 (which were incorrectly clustered), than linking the representative images.
Overall quality: In order to measure the performance of multiple users annotating the dataset, we calculate all subsets of users and combine their annotations for the six users who labeled with KL divergence. Figure 3(b) gives the average performance in annotation for combining a certain number of users. By comparing the "Overall" results, which shows the accuracy of annotation with clustering, to the "Stage 2" results, which estimates the accuracy of annotating all images without clustering, there is in most cases a small decrease in accuracy due to the clustering. The first 2 bins in Figure 3(b) show the difference between the user performance on the correctly clustered images and incorrectly clustered images as discussed before. The incorrectly clustered images have only a small influence on the overall performance, because the percentage of incorrectly clustered images for KL divergence and Pyramid histograms is respectively 9.8% and 16.9%.
Gain in time and mouse clicks: To estimate the time it takes to annotate the images, one of our users performed the labeling non-stop, allowing us to measure the average time it takes to finish one screen. In the case of the first interface, it took an average time of 19.7 seconds to complete one screen and for the second interface it took an average time of 7.3 seconds. Figure 3(c)



(a) Box plot of the performance in labeling of all users for the individual stages

(b) Histogram of the average combined performance of a certain number of users

(c) The improvement in both time and mouse clicks over annotating all images

Figure 3: Evaluation

shows the improvement in time based on extrapolation of these values for all users and both clustering methods in comparison to labeling all images using the second interface. The number of mouse clicks is important in crowdsourcing, because users often get paid per click. If one labeled all the images using the second interface, we would need $2M$ clicks to select and confirm the correct species (stage 2 interface). In the first interface, we only click on images that have to be removed from the cluster and need an extra click to confirm our annotation for each cluster. In the second stage we click twice to select and confirm the label for only the representative images and all the images excluded from stage 1. The net results is about a 77% reduction in label time and 93% reduction in mouse clicks when using KL divergence.

4. Conclusion

An efficient framework to annotate images is presented in this paper. The quality of this annotation framework is affected by the clustering method (5.1% error by combining 3 users), however it does not seem to affect the quality of the annotations too much compared to the estimated quality of labeling all the images in the dataset without clustering (4.2% error by combining 3 users). This difference in quality gets smaller if more users are annotating. With the clusters based framework, we can label the dataset three times in the time it takes to label all images without clustering, which also gives a better quality of labels. This framework has also been used without stage 3 to label a dataset of around 23000 fish images, which took about 8 hours for each of the three users annotating.

Acknowledgements: This work is supported by the Fish4Knowledge project, which is funded by the European Union 7th Framework Programme [FP7/2007-2013].

References

[1] A. Bosch, A. Zisserman, and X. Muñoz. Image classification using random forests and ferns. In *IEEE 11th Int Conf on Computer Vision, 2007. ICCV 2007*, pages 1–8, Oct. 2007.

[2] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972 – 976, Feb. 2007.

[3] J. Goldberger, S. Gordon, and H. Greenspan. Unsupervised image-set clustering using an information theoretic framework. *IEEE Trans. on Image Processing*, 15(2):449–458, Feb. 2006.

[4] F. K. Khattak and A. Salleb-Aouissi. Quality control of crowd labeling through expert evaluation. In *Second Workshop on Computational Social Science and the Wisdom of Crowds (NIPS 2011)*, pages 1–5, Dec 2011.

[5] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *J. Mach. Learn. Res.*, 99:1297–1322, 2010.

[6] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *Int J of Computer Vision*, 77:157–173, 2008.

[7] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc of the SIGCHI conf on Human factors in computing systems*, CHI '04, pages 319–326, 2004.

[8] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *CVPR 2006*, volume 2, pages 1483 – 1490, 2006.

[9] P. Welinder and P. Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *CVPR Workshops 2010*, pages 25 –32, june 2010.

[10] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, pages 2035–2043, 2009.