Learning Localized Perceptual Similarity Metrics for Interactive Categorization

Catherine Wah[1], Subhransu Maji[2], and Serge Belongie[3]

[1]UC-San Diego [2]UMass Amherst [3]Cornell Tech

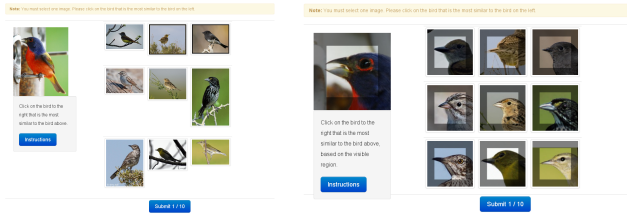
Abstract. Current similarity-based approaches to interactive categorization rely on learning metrics from holistic perceptual measurements of similarity between objects or images. However, making a single judgment of similarity at the object level can be a difficult or overwhelming task for the human user to perform. Secondly, a single general metric of similarity may not be able to adequately capture the minute differences that discriminate fine-grained categories. In this work, we propose a novel approach to interactive fine-grained categorization that leverages multiple perceptual similarity metrics learned from localized and roughly aligned regions across images, reporting state-of-the-art results and outperforming methods that use a single nonlocalized similarity metric.

1 Introduction

Fine-grained visual categorization (FGVC) is an area of computer vision that has experienced an increased amount of attention in recent years across various visual domains [4, 1–3]. The goal is to distinguish between *fine-grained categories* or subcategories (*e.g.*, a Cardinal vs. a Lazuli Bunting) that belong to the same *basic-level category* (*e.g.*, Bird). Some work has focused on interactive methods for FGVC [1, 7, 2], in particular using perceptual similarity judgments from human users [8]. Perceptual similarity-based categorization systems do not require part and attribute vocabularies, reducing both the burden on the non-expert human users as well as the reliance on experts.

While similarity can be holistic in nature (*e.g.*, object utility or function, or overall shape), it can also be highly localized, for instance, when specific corresponding regions or parts of the object differ from one other. Especially at the fine-grained category level in which classes tend to be visually coherent, it is likely that the small yet important characteristics that distinguish subcategories are localizable. In these scenarios, a single metric of perceptual similarity that is observed at the object level can be overly general, and asking a user to make holistic nonlocalized similarity comparisons can be difficult.

By using localized similarity comparisons and constraining the user's view to a portion of the image, we are able to highlight certain aspects of similarity; these localized judgments tend to be easier for humans to perform than holistic similarity judgments (see Figure 1). Moreover, we can potentially reduce the effect of nuisance factors such as background noise and differing object poses. For each common region or part, we learn a separate perceptual space that captures local visual information.



(a) Nonloc. comparison (b) Loc. comparison

Fig. 1. We use perceptual similarity metrics learned from localized comparisons (1(b)) to perform interactive categorization, aiming to reduce both overall human effort required as well as improve performance over using nonlocalized comparisons (1(a)).

2 Methods

Given a reference image x, our goal is to predict as quickly as possible the true object class c from C possible classes that fall within the same basic-level category. We do so by using both computer vision and user responses to similarity-based questions posed by the system at test-time. Our system supports two types of similarity comparisons: nonlocalized and localized (see Figure 1). We define a region as a visually discriminative and recurring object part that does not have to be semantically defined or meaningful. In practice, it is a spatially localized and roughly aligned template derived from an associated descriptor (see Figure 2).

We describe as follows an extension to the system proposed by Wah *et al.* [8], which supports the use of multiple similarity metrics but does not adequately handle instance-level variations, specifically the presence of certain pose-aligned parts in the image. We therefore require a set of regions in order to localize similarity comparisons, as well as a methodology for choosing which images and regions to show in the display.

In order to highlight the same localized region across images for performing localized similarity comparisons, we require instance-level region correspondences. We discover a set of localized and roughly aligned mid-level discriminative visual representations in an unsupervised manner [5]. This method has multiple advantages: first, we can determine spatial correspondences between images by using the discovered patches as detectors; second, the regions are by nature common in gradient appearance; and last, the discovered regions may provide implicit (albeit noisy) pose alignment. At test time, we can use these templates as part detectors that are evaluated on input images in a sliding window manner. In generating the set of discriminative regions, we keep discovered patches that have sufficient overlap (50%) with the ground truth object bounding box. From this resulting set, we select 5 diverse and representative regions to use in our experiments (see Figure 2).

It is likely that the localized regions may not be present in certain images; this corresponds to a low detection score for a particular region detector. As such, we modify the display model of [8] to take part presence into account. We

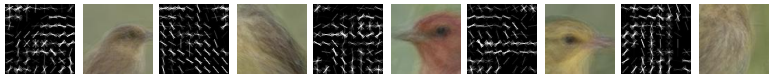
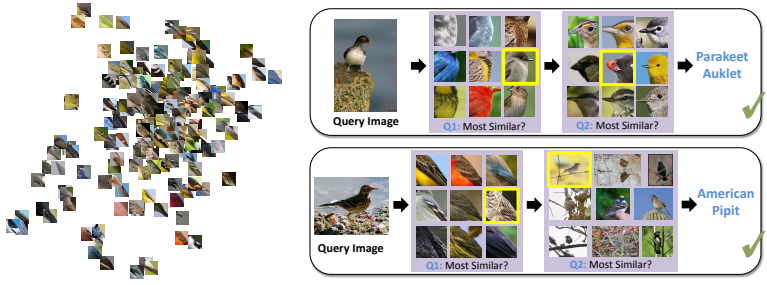


Fig. 2. The selected discovered regions, each visualized as a HOG template alongside the averaged image of the corresponding highest confidence positive detections.



(a) Region 13 (b) Qualitative results

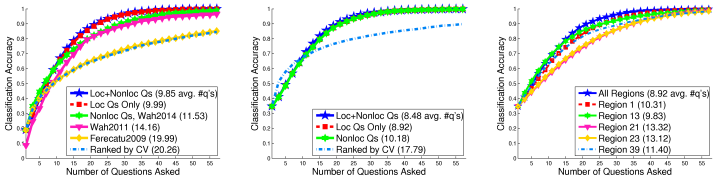
Fig. 3. 3(a): We visualize the first two dimensions of a learned local embedding. 3(b): Qualitative examples using only the 5 localized similarity metrics (top) and using the localized metrics along with a nonlocalized metric (bottom).

employ an approximate solution that groups the images into clusters to ensure that each image in the display is equally likely to be selected, maximizing the expected information gain. For the display, we thus pick the image within the cluster with the highest mass as weighted by the region presence probability.

We perform experiments on the fine-grained CUB-200-2011 dataset [6] of 200 bird species. For learning the perceptual metrics, we collect similarity comparisons using the crowdsourcing workplace Amazon Mechanical Turk, sampling images based on region detection scores. For each localized region, we generate triplets from similarity comparisons to learn an independent localized embedding (*e.g.*, Figure 3(a)) [8]. This is then used directly in our interactive categorization system. In order to compare to previous work, we initialize our computer vision estimate of class probabilities using the same setup as [8], with multiclass 1-vs-all SVMs trained on color/grayscale SIFT features and color histograms. We also compare to a method that uses Fisher vector encodings (FVs).

3 Experiments

We present our interactive classification results in Figure 4; qualitative examples are shown in Figure 3(b). For testing, we use an interface similar to that used in training (Figure 1(b)). We use simulated user responses at test time; we refer the reader to [8] for details on the user model. Our experimental setup and performance metrics are the same as [7, 8], in which the user can verify perfectly



(a) SIFT/color hist. (b) Fisher vectors (c) Local (FVs)

Fig. 4. *Interactive categorization.* 4(a): We compare to prior baselines from [8]. 4(b): We observe performance when the initial class probability estimates are improved with Fisher vectors. 4(c): We compare performance using each localized metric separately.

the highest probability class, and we evaluate our system based on the average number of questions a user must answer per test image to classify it correctly.

We can draw several observations from our results: (1) It is advantageous to use localized and nonlocalized metrics together (see Figure 4(a)). (2) Localized comparisons are more informative than nonlocalized comparisons. In general, our interactive categorization system will tend to ask users to make localized comparisons in the beginning, as these questions provide the most expected information gain. (3) Some localized regions are more useful for categorization than others (see Figure 4(c)). (4) Localized similarity comparisons require less human effort. On average, it takes a human user 11.35 ± 10.17 sec to answer a localized comparison, compared to 16.36 ± 14.31 sec for a nonlocalized comparison.

To conclude, we have presented an approach to interactive fine-grained categorization that leverages localized similarity comparisons and does not rely on part or attribute vocabularies; we discover a set of discriminative, localized and roughly aligned regions for this categorization task. We demonstrate that localized similarity comparisons are more intuitive for users to perform, and that by using independent localized metrics we can improve categorization accuracy over using a single nonlocalized metric.

References

1. Branson, S., et al.: Visual recognition with humans in the loop. In: ECCV (2010)
2. Kumar, N., et al.: Leafsnap: A computer vision system for automatic plant species identification. In: ECCV (2012)
3. Maji, S., et al.: Fine-grained visual classification of aircraft. Tech. rep. (2013)
4. Nilsback, M., Zisserman, A.: Automated flower classification over a large number of classes. In: ICCVGIP (2008)
5. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: ECCV (2012)
6. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Caltech-UCSD Birds-200-2011. Tech. Rep. CNS-TR-2011-001, Caltech (2011)
7. Wah, C., et al.: Multiclass rec. and loc. with humans in the loop. In: ICCV (2011)
8. Wah, C., Van Horn, G., Branson, S., Maji, S., Perona, P., Belongie, S.: Similarity comparisons for interactive fine-grained categorization. In: CVPR (2014)