Transfer Learning of Object Classes: From Cartoons to Photographs

Geremy Heitz, Gal Elidan, Daphne Koller
Stanford University, Stanford, CA
{gaheitz,galel,koller}@cs.stanford.edu

Abstract

We consider the important challenge of recognizing a variety of deformable objects in images. Of fundamental importance and particular difficulty in this setting is the problem of "outlining" an object, rather than simply deciding on its presence or absence. A major obstacle in learning a model that will allow us to address this task is the need for hand-segmented training images. In this paper we present a transfer learning approach that circumvents this problem by transferring the "essence" of an object from cartoon images to natural images, using a landmark-based model. The use of transfer to create an automatic model-learning pipeline greatly increases our efficiency and flexibility in learning novel objects with minimal user supervision. We show that our method is able to automatically learn, detect and localize a variety of classes.

1 Introduction

Recognition and localization of instances of object classes is an important open problem in the field of computer vision. Many papers address this problem using a variety of methods. The constellation method [3] attempts to recognize and localize objects of interest using a generative model of object "parts", which appear in the image as interest operator patches. Berg *et al.* [1] use a landmark-based model and perform recognition using a nearest neighbor search relative to exemplar images. They solved the correspondence problem using a quadratic integer program.Coughlan and Ferreira [2], solve the correspondence problem using loopy belief propagation (LBP), for simple objects including handwritten letters and stick figures.

All of these suffer from the fact that fully-supervised data is hard to obtain. They solve this by either using Expectation-Maximization [3], using many image points and hoping that most of them are inside the object [1], or concentrating on simple objects with trivial segmentations [2].

In this work, we present a new transfer learning approach that circumvents this major obstacle, enabling us to "self-learn" using minimal user supervision. In the first phase of our algorithm we automatically learn a simple landmark model from cartoon images of the object. We then correspond our model to candidate natural training images using MRF inference in order to automatically identify useful candidates. In the second phase of the algorithm, we use these candidates to construct a more elaborate landmark based model (i.e. one that also takes into account appearance). We then use this model to identify and predict the location of objects in unseen test instances.

We show that our method effectively bootstraps the simple cartoon images and is able to localize objects in natural images surprisingly well. Interestingly, we demonstrate that learning from cartoon images is often superior to learning from hand-segmented examples, as the "essence" of the object is captured in the first phase of our algorithm without human bias.

2 Landmark-Based Object Model

We model an object as a set of landmark points lying on the object boundary. Each landmark has local information about the image appearance in its neighborhood as well as local edge information. In addition, pairs of landmarks have information about their relative locations, rotations, and scales.

2.1 Localization using Inference

To localize an object in an image, we define a Markov Random Field (MRF) whose variables correspond to the landmarks of the model. An assignment to these variables is a correspondence between the model and image pixels and the potentials of the MRF take into account both the local and inter-landmark features. Thus, the problem of object localization is translated into the problem of finding the most likely assignment in a probabilistic graphical model. We search for this assignment using loopy belief propagation.

3 Transfer Learning

Our learning procedure is composed of two principal stages, one that involves only simple cartoon images and one that involves natural images. By using this two-phase approach, we can automatically determine which landmarks to use and avoid some of the pitfalls described above. We are thus able to provide a more stable model than those produced using previous methods.

In phase 1 of the model learning (see Figure 1), we automatically extract outlines from a set of cartoon images of the object to be learned. This gives us a high-resolution contour of the outline. We then determine the correspondences between all of the training instances using a slight modification of our correspondence algorithm described above. This step produces a common parameterization for the training contours. By using MDL-like considerations, we can then select a set of landmarks that most accurately represents the shape. From the chosen landmarks, we construct our phase 1 model, which captures local edge features and pairwise interactions between the landmarks. Such a model will contain little or no appearance information, due to the artificial nature of the training images.

In phase 2 (see Figure 2), we correspond the phase 1 model to each instance in our natural image training set. From this we select the correspondences that score most highly (i.e. the ones we are most confident about), and use these as the phase 2 training instances. By doing this, we have bootstrapped the creation of a training set using the information automatically extracted from the cartoon images. This model will now contain the full appearance model derived from the natural images.

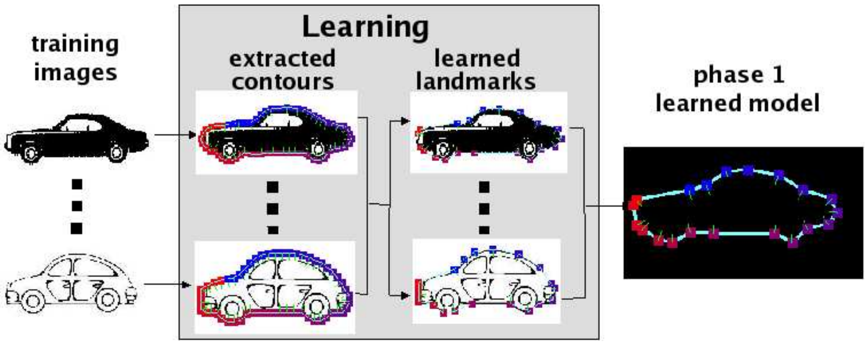

Figure 1: Phase 1 of model learning. We begin with a set of cartoon images and extract a high resolution outline contour. We then correspond these contours and select a set of landmarks that best represents the shape across all of the training instances. A model is then learned from these cartoon instances.

By breaking the process into two phases, we have reduced the number of parameters that must be learned at any one time. In phase 1 the only parameters to learn are which points along the full-resolution contour to use as landmarks. In phase 2, we begin with fully-supervised training instances, so maximum-likelihood learning of the model parameters is straightforward.

4 Preliminary Results

We have applied our procedure to several object categories with promising results. Below we show a few representative results for the "car side" class in the Caltech 101 dataset. Figure 3 shows a few successful localizations of cars as well as a failure.

We evaluate these results using an overlap metric that measures the extent to which the object is correctly localized. Figure 4 compares the performance of our approach to learning from hand-segmented images as a function of the number of training samples used in the second phase of the learning algorithm. Surprisingly, learning from cartoon images is significantly better than learning from models where the landmark were chosen by hand. This is the result of a human bias in labeling that tends to use both too few and the wrong landmarks. For instance, humans often exclude landmark points along gently sloping contour points that can aid recognition but often don't appear to help describe the shape.

References

[1] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[2] J. M. Coughlan and S. J. Ferreira. Finding deformable shapes using loopy belief propagation. In *European Conference on Computer Vision (ECCV)*, pages 453–468, 2002.

[3] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 264–271, 2003.

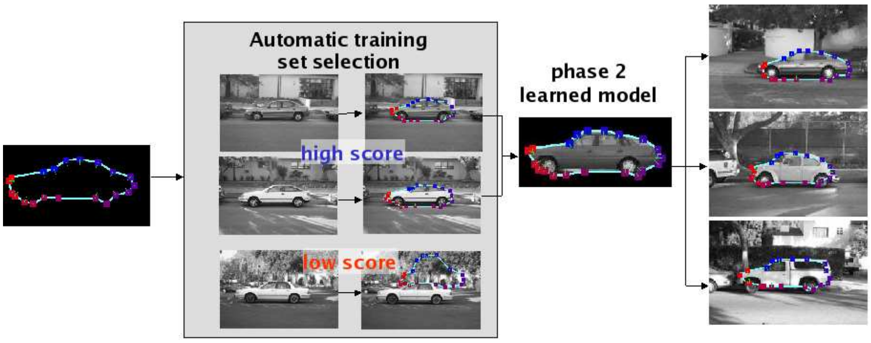

Figure 2: Phase 2 of model learning. We apply our phase 1 model to infer the landmarks in the images of our training set. Selecting the best such correspondences as our new training set, we learn a new model that contains both the appearance and geometry information.

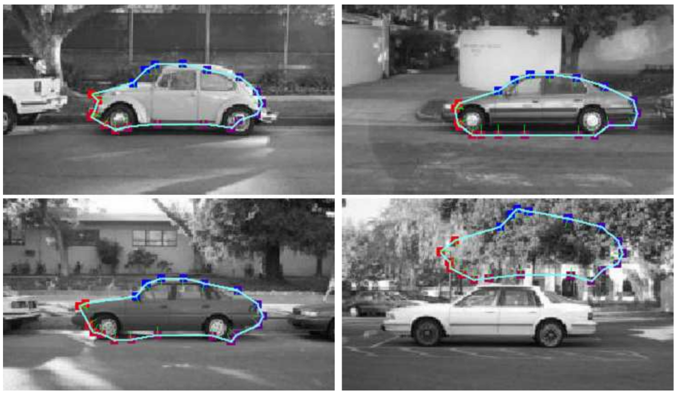

Figure 3: Localization results for the car side object class in the Caltech 101 dataset. Shown are three successful localizations and a single failure. Reasonable localization was achieved in over 85 percent of the test instances.

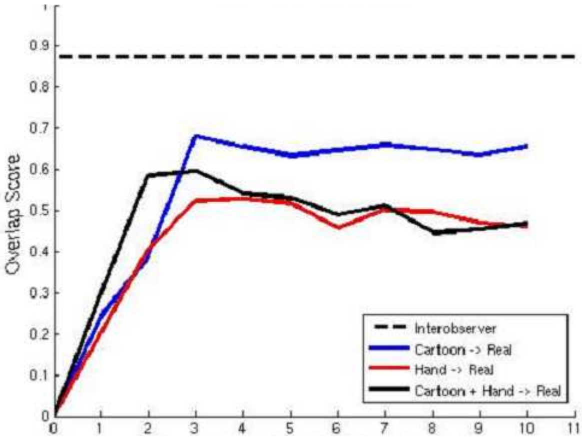

Figure 4: Average overlap score as a function of the number of real-life training instances used in phase 2 of out learning algorithm. We compare three scenarios: only cartoon images were used in phase 1 of the algorithm; only hand segmented images were used; both hand and cartoon images were used.