Discovering States and Transformations in Image Collections

Phillip Isola*, Joseph J. Lim*, Edward H. Adelson
Massachusetts Institute of Technology

Objects in visual scenes come in a rich variety of transformed states. A few classes of transformation have been heavily studied in computer vision: mostly simple, parametric changes in color and geometry. However, transformations in the physical world occur in many more flavors, and they come with semantic meaning: e.g., bending, folding, aging, etc. The transformations an object can undergo tell us about its physical and functional properties.

Much work in computer vision has focused on the problem of invariant object recognition [2], scene recognition [8], and material recognition [7]. The goal in each of these cases is to build a system that is invariant to all within-class variation. Nonetheless, the variation in a class is quite meaningful to a human observer. Consider Figure 1. The collection of photos on the left only shows tomatoes. An object recognition system should just see "tomato". However, we can see much more: we can see peeled, sliced, and cooked tomatoes. We can notice that some of the tomatoes are riper than others, and some are fresh while others are moldy. Being able to infer such properties is essential to interacting with objects.

Given a collection of images of an object, what can a computer infer? Given 1000 images of tomatoes, can we learn how tomatoes work? In this paper we take a small step toward that goal. From a collection of photos, we infer the states and transformations depicted in that collection. For example, given a collection of photos like that on the left of Figure 1, we infer that tomatoes can be undergo the following transformations, among others: ripening, wilting, molding, cooking, slicing, and caramelizing. Our system does this without having ever seen a photo of a "tomato" during training (although overlapping classes, such as "fruit", may be included in the training set). Instead we transfer knowledge from other related object classes.

The problem of detecting image state has received some prior attention. For example, researchers have worked on recognizing image "attributes" (e.g., [5], [6]), which sometimes include object and scene states. However, most of this work has dealt with one image at a time and has not extensively catalogued the state variations that occur in an entire image class. Unlike most previous work, we focus on understanding variation in image *collections*. In addition, we go beyond previous attributes work by linking up states into pairs that define a transformation: e.g., *raw↔cooked, rough↔smooth, defalted↔inflated*. We explain image collections both in terms of their states (unary states) and transformations (antonymic state pairs). In addition, we show how state pairs can be used to extract a continuum of images depicting the full range of the transformation (Figure 1 bottom-left).

To demonstrate our understanding of states and transformations, we test on three tasks. As input we take a set of images depicting a noun class our system has never seen before (e.g., *tomato*; Figure 1). We then parse the collection:

- Task 1 – Discovering relevant transformations: What are the transformations that the new noun can undergo in (e.g., a *tomato* can undergo *slicing*, *cooking*, *ripening*, etc).
- Task 2 – Parsing states: We assign a state to each image in the collection (e.g., *sliced*, *raw*, *ripe*).
- Task 3 – Finding smooth transitions: We recover a smooth chain of images linking each pair of antonymic states.

Similarly to previous works on transfer learning [3] , our underlying assumption is the transferability of knowledge between adjectives (states and transformations). To solve these problems, we train classifiers for each state using a convolutional neural net [4]. By applying these classifiers to each image in a novel image set, we can discover the states and transformations in the collection. We globally parse the collection by integrating the per-image inferences with a conditional random field.

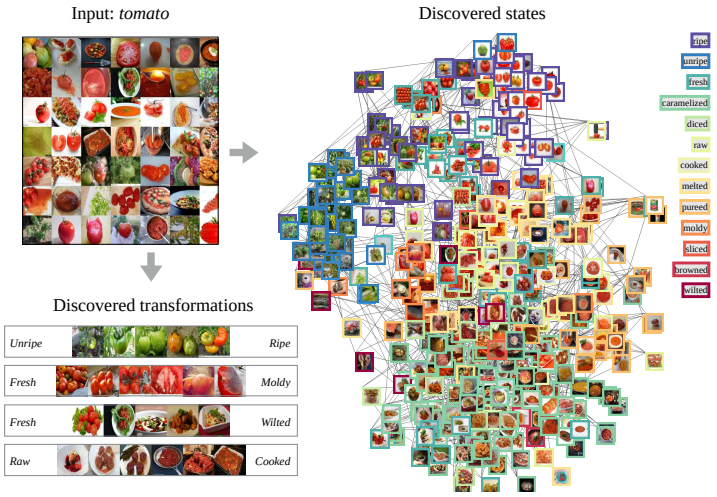

Figure 1: **Example input and automatic output of our system:** Given a collection of images from one category (top-left, subset of collection shown), we are able to parse the collection into a set of states (right). In addition, we discover how the images transform between antonymic pairs of states (bottom-left).

Our contribution in this paper is threefold: (1) introducing the novel problem of parsing an image collection into a set of physical states and transformations it contains (2) showing that states and transformations can be learned with basic yet powerful techniques, and (3) building a dataset of objects, scenes, and materials in a variety of transformed states.

[1] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. NEIL: Extracting Visual Knowledge from Web Data. In *ICCV*, 2013.

[2] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV*, 2014.

[4] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[5] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[6] Devi Parikh and Kristen Grauman. Relative attributes. In *ICCV*, 2011.

[7] Lavanya Sharan, Ce Liu, Ruth Rosenholtz, and Edward H Adelson. Recognizing materials using perceptually inspired features. *IJCV*, 103 (3):348–371, 2013.

[8] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

* - indicates equal contribution