Metric learning by active crowd-sourcing

Graham W. Taylor, Ian Spiro, Chris Bregler, and Rob Fergus
Courant Institute of Mathematical Sciences, New York University
{gwtaylor,spiro,bregler,fergus}@cs.nyu.edu

Introduction

Effective systems for visual reasoning must cope with input variability caused by many different factors. In response, a variety of approaches have attempted to learn image representations that are invariant to photometric and geometric distortions. Methods like Boost-SSC/PSH [1], variants of Neighborhood Components Analysis (NCA) [2, 3, 4] and Dimensionality Reduction by Learning an Invariant Mapping (DrLIM) [5] use supervised learning to map high-dimensional images to a low-dimensional space in which nearest neighbors are easily computable, and observations that are perceptually similar have high measurable similarity. However, these methods employ a binary notion of pairwise similarity, either through predefined classes or by thresholding real-valued labels. Such labelings are expensive to obtain, often difficult to define and cannot represent graded similarity which may benefit learning.

In recent years, Amazon Mechanical Turk and other crowd-sourcing platforms have emerged as a way of accelerating vision and other tasks, often for rapid labeling of massive image datasets. Most of these techniques ask the participants to explicitly provide desired segmentations, feature points, configurations, pose, or class labels. For continuous domains, especially for similarity measures, people have more difficulty supplying consistent labels. We propose a new paradigm for learning invariant mappings: *active* crowd sourcing through imitation. Consider the problem of learning an embedding in which people in similar pose lie close-by. The first step in this task is to obtain many images of people in similar pose but with different clothing, backgrounds, lighting and other appearance changes. Obtaining this data is time-consuming. Moreover, judging the degree of similarity between observations is non-trivial and inconsistent across observers. Other works (e.g. [1, 6, 7]) have used synthetic renderings to a modest degree of success, but we believe there is a better source of real data that exhibits the same amount of variability a model would observe at test time. Given an image of a person in pose, people have a remarkable ability to mimic its content. Therefore, we can exploit the abundance of webcams to quickly crowd-source a massive dataset of people in similar pose by asking people to imitate images (Fig. 1a).

A key aspect of our approach is the use of temporal coherence in video to both increase the number of similar examples and introduce graded similarity, which we demonstrate improves the quality of the embedding. Temporal coherence has been used to learn invariant features [9] but in a very different context. These methods directly learn from frames of video while we use video only as a source of seed images presented to users. Our model learns only from user-contributed imitations, many of which could correspond to a single frame of the original video. We use scene cuts and correspondence by frame number to determine the graded similarity of the imitations.

Method

Our online learning algorithm is based on DrLIM [5] and adapted to graded similarity. It is trained by stochastic gradient descent, and can be used with arbitrary nonlinear mappings. Our mapping, shown in Fig. 1b, is a siamese network that processes pairs of images through two identical pathways, each of which is a standard convnet, similar to ones used for object recognition. The loss is computed on the output of each image, and its gradient is backpropagated through each pathway. We examine several ways of mapping discrete distances obtained from frame numbers to real-valued similarity scores.

Experiments

We consider the problem of matching people in similar pose but in severely different settings. Rather collect data ourselves, we leverage an existing project in an unintended way. *One Frame of Fame* [8] is a music video created by the Dutch band C-Mon & Kypski. It contains members of the band performing a variety of poses choreographed to music. However, there is a twist: the band aims to replace selected frames of their video with imitations captured from the webcam of an anonymous visitor. The band has created a web application to solicit frames. A visitor to the website is presented with a frame of the original video and they are asked to imitate that pose using their own webcam. At the time of writing, the band had collected 25,998 images. The images are of sufficient resolution to impede linear or vector-based nonlinear mappings. The number of images is too large to work suitably with the batch methods like NCA. Our model is trained on a subset of 24,065 (uncleaned) images and tested on a subset of 1,269 (uncleaned) images. Sample query results are shown in Fig. 2.

References

[1] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, pages 750–759, 2003.
[2] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, 2004.
[3] R. Salakhutdinov and G. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *AISTATS*, volume 11, 2007.
[4] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *CVPR*, 2008.
[5] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742, 2006.
[6] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. Boostmap: A method for efficient approximate similarity rankings. *CVPR*, 2004.
[7] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, pages 641–648, 2003.
[8] C-Mon and Kypski. One frame of fame. http://oneframeoffame.com, 2010.
[9] H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. In *ICML*, pages 737–744, 2009.

Category: vision Preference: oral

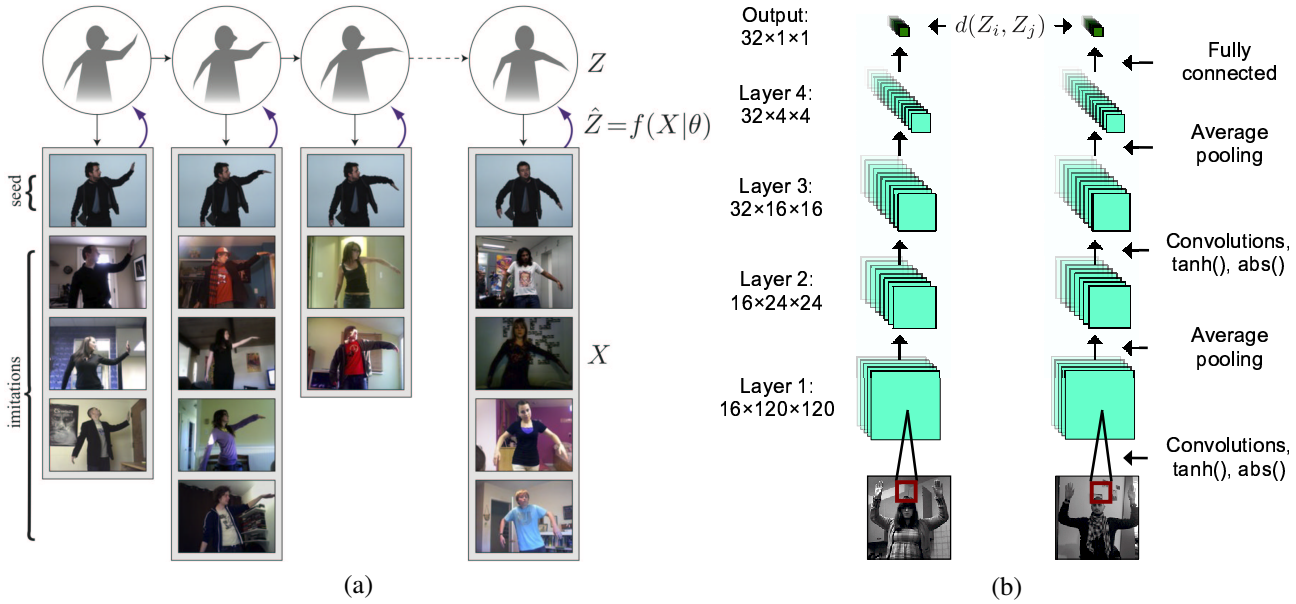

Figure 1: (a) Schematic of our approach. We assume for each frame of video, there exists an unobserved low-dimensional representation of pose, Z. A seed image is generated by mapping from pose space to pixels, X, through an interpretation function. Our method learns a nonlinear embedding, $f(X|\theta)$ which approximates Z with a low-dimensional vector. In the example above, users are asked to imitate seed images taken from a music video [8]. (b) Convnet architecture for learning a nonlinear mapping.

Figure 2: Sample retrieval results. Each row is a query. We select a test image (column 1) and find its 10 nearest neighbors using our learned embedding: Ord-Conv (simple). Text indicates seed id (left) and distance from the query (right).

Category: vision Preference: oral