Beyond Frontal Faces: Improving Person Recognition Using Multiple Cues

Ning Zhang[1], Manohar Paluri[2], Yaniv Taigman[2], Rob Fergus[2], Lubomir Bourdev[2]
[1]UC Berkeley [2]Facebook AI Research

Recognizing people we know from unusual poses is easy for us. However, in the absence of a clear, high-resolution frontal face, we rely on a variety of subtle cues from other body parts, such as hair style, clothes, glasses, pose and other context. While a lot of progress has been made recently in recognition from a frontal face, non-frontal views are a lot more common in photo albums than people might suspect. For example, in our dataset which exhibits personal photo album bias, we see that only 52% of the people have high resolution frontal faces suitable for recognition. Thus the problem of recognizing people from any viewpoint and without the presence of a frontal face or canonical pedestrian pose is important, and yet it has received much less attention than it deserves. We believe this is due to two reasons: first, there is no high quality large-scale dataset for unconstrained recognition, and second, it is not clear how to go beyond a frontal face and leverage these subtle cues. In this paper we address both of these problems.

We introduce the *People In Photo Albums (PIPA)* dataset, a large-scale recognition dataset collected from Flickr photos with creative commons licenses. It consists of 37,107 photos containing 63,188 instances of 2,356 identities and examples are shown in Figure 1. The dataset is publicly available. We tried carefully to preserve the bias of people in real photo albums by instructing annotators to mark every instance of the same identity regardless of pose and resolution. Our dataset is challenging due to occlusion with other people, viewpoint, pose and variations in clothes. While clothes are a good cue, they are not always reliable, especially when the same person appears in multiple albums, or for albums where many people wear similar clothes (sports, military events). As an indication of the difficulty of our dataset, the DeepFace system [2], which is one of the state-of-the-art recognizers on LFW, was able to register only 52% of the instances in our test set and, because of that, its overall accuracy on our test set is 46.66%.

We propose the Pose Invariant PErson Recognition (PIPER) method, which accumulates the cues of poselet-level person recognizers trained by deep convolutional networks to discount for the pose variations, combined with a face recognizer and a global recognizer. We propose a Pose Invariant PErson Recognition (PIPER) method, which uses part-level person recognizers to account for pose variations. We use poselets [1] as our part models and train identity classifiers for each poselet. Poselets are classifiers that detect common pose patterns. A frontal face detector is a special case of a poselet. Other examples are a hand next to a hip or head-and-shoulders in a back-facing view, or legs of a person walking sideways. A small and complementary subset of such salient patterns is automatically selected as described in [1]. While each poselet is not as powerful as a custom designed face recognizer, it leverages weak signals from specific pose pattern that is hard to capture otherwise. By combining their predictions we accumulate the subtle discriminative information coming from each part into a robust pose-independent person recognition system.

We are inspired by the work of Zhang et al. [3], which uses deep convolutional networks trained on poselet detected patches for attribute classification. However our problem is significantly harder than attribute classification since we have many more classes with significantly fewer training examples per class. We found that combining parts by concatenating their feature in the manner of [3] is not effective for our task. It results in feature vectors that are very large and overfit easily when the number of classes is large and training examples are few. Instead, we found training each part to do identity recognition and combining their predictions by using sparsity filling, as shown in Figure 2, can boost the performance by a large margin.

We demonstrate the effectiveness of PIPER by using three different experimental settings on our dataset: 1) Person recognition 2) One-shot learning and 3) Unsupervised identity retrieval. For person recognition task, our method can achieve 83.05% accuracy over 581 identities on the test set.

Figure 1: **Example photos from our dataset.** These are taken from a single album and show the associated identities. Each person is annotated with a ground truth bounding box around the hear, with each color representing one identity. If the head is occluded, the expected position is annotated.

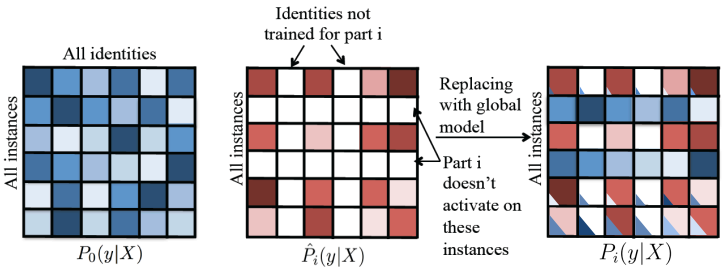

Figure 2: Example of sparsity filling. We show the predictions of the global model on the left and poselet-level classifiers in the middle (white cells means missing predictions). On the right we show how we fill in the rows from global model and how we fill in the columns by linearly interpolating based on the global model in the normalized probability. More details can be found in the full paper.

Moreover when a frontal face is available, it improves the accuracy over DeepFace from 89.3% to 93.4%.

[1] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[2] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *CVPR*, 2014.

[3] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir D. Bourdev. PANDA: pose aligned networks for deep attribute modeling. *CVPR*, 2014.

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