Learning a Sequential Search for Landmarks

Saurabh Singh, Derek Hoiem, David Forsyth
University of Illinois, Urbana-Champaign.

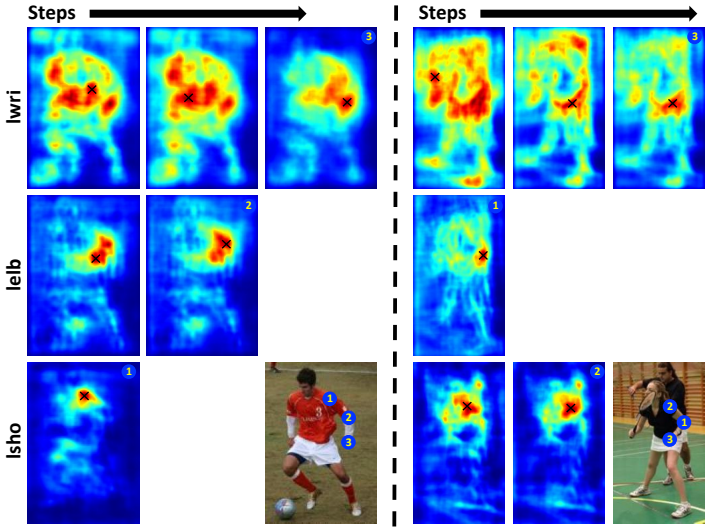

Figure 1: Visualization of the implicit spatial model learned by our method for the case of three landmarks in two different images. Each column corresponds to a *step* of our method and displays the scores for every location in the image, for each remaining landmark, as a heat map. In each step the highest scoring location-landmark pair is chosen as a detection. Bottom-right shows the inferred locations numbered by the step in which they were detected. Note that the landmarks are detected in a different order in the two images. The peaks, marked with a black cross, shift to the correct locations as steps progress; e.g., peak for *lelb* (left elbow) in left image shifts to the correct location in step 2 after lsho (left shoulder) is detected in step 1. Similarly, peak for *lwri* (left wrist) shifts in step 3 once *lelb* is detected.

We propose a general method to find landmarks in images of objects using both appearance and spatial context. This method is applied *as is* to two problems: parsing human body layouts, and finding landmarks in images of birds. Our method learns a sequential search for localizing landmarks, iteratively detecting new landmarks given the appearance and contextual information from the already detected ones. The choice of landmark to be added is opportunistic and depends on the image; for example, in one image a head-shoulder group might be expanded to a head-shoulder-hip group but in a different image to a head-shoulder-elbow group. The choice of initial landmark is similarly image dependent. Groups of landmarks are scored using a learned function, which is used to expand groups greedily. Our scoring function is learned from data labelled with landmarks but without any labeling of a detection order. Our method represents a novel spatial model for the kinematics of groups of landmarks, and displays strong performance on two different model problems.

Landmark detection is a well-studied problem, usually in the domain of finding human body joints or parts [1, 2, 3, 4, 5, 6]. Typically, this is done by modeling relations among landmarks which may be very complicated; for example, human bodies are posed and appareled in structured but complex ways, so that the position and appearance of a wrist depends on the positions of shoulders, elbows, hips, presence of long sleeves, and so on.

To make learning and inference tractable, existing approaches are forced to use at least some of a menu of assumptions: that each landmark can be identified relatively easily; that appearance and spatial terms factorize; that spatial relations fit a convenient model; that discriminative methods can satisfactorily handle relational information without expressing it explicitly; or that intractable inference problems can be dealt with approximately in a satisfactory way.

Figure 2: Qualitative results of our method on Leeds Sports Dataset (left), Fashion Pose (middle) and Caltech-UCSD Birds 200 (right).

We offer a novel, alternative strategy for finding landmarks. Instead of fixing a model structure and then dealing with an intractable inference, we treat the inference as a sequential search procedure and learn parameters such that the search remains tractable. In every step of the search a landmark is detected and our model uses the detected landmarks to capture increasingly complex appearance and spatial relations jointly to help find the next landmark. Thus, we substitute sequential inference for joint inference in order to benefit from more expressive dependency models (Figure 1).

Our approach doesn't assume that some fixed set of landmarks, e.g. *head*, is always easy; it allows these to be different from one image to another. It uses no additional supervision about their easiness or detection order. Further, it doesn't impose an explicit spatial model nor does it treat one landmark differently from other. This information is coded in the features of the landmarks and it learns to use them as needed. The paper describes the model and the training algorithm used to train it.

An interesting property of our approach is that the order in which landmarks are found may differ from image to image. Our system *learns* how each landmark depends on what has already been found, and *learns* to identify which landmark to find next. Because our method does not require any expert guidance for spatial dependencies or which landmark to detect first, it can be applied easily to any landmark detection problem. We demonstrate this by detecting landmarks on people and birds (Figure 2).

[1] Mykhaylo Andriluka, Stephan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[2] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *ICCV*, 2009.

[3] Benjamin Sapp, Alexander Toshev, and Ben Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

[4] Duan Tran and David Forsyth. Improved human parsing with a full relational model. In *ECCV*, 2010.

[5] Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011.

[6] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 2013.

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