Learning a Sequential Search for Landmarks

Saurabh Singh, Derek Hoiem, David Forsyth
University of Illinois, Urbana-Champaign
http://vision.cs.uiuc.edu/projects/lssland/
{ss1, dhoiem, daf}@illinois.edu

Abstract

We propose a general method to find landmarks in images of objects using both appearance and spatial context. This method is applied without changes 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 are scored using a learned function, which is used to expand them 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.

1. Introduction

Identifying sets of landmarks, such as joints on the human body or parts of a car, is a major problem in computer vision. Many parts, such as a wrist or car door handle, are difficult to find on their own, so the key research problem is modeling the relations among the appearances and positions of landmarks. Relations among landmarks 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. The result is a surprisingly small list of strategies for finding landmarks (§1.1).

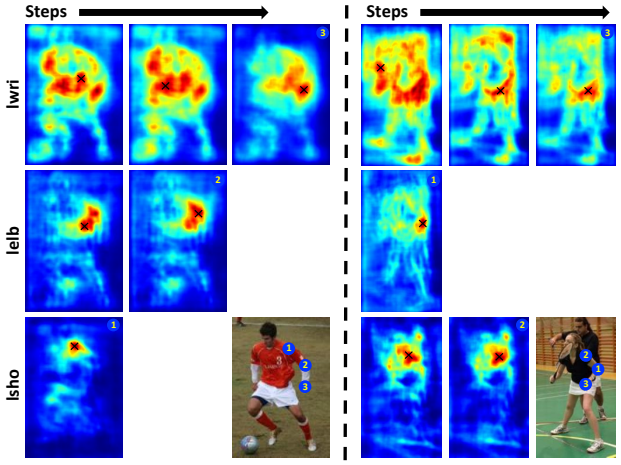

Figure 1. Our method learns to find landmarks sequentially in an image dependent order. It uses the already detected landmarks to provide the context needed for detecting the next landmark. We show a visualization of the implicit spatial model learned by our method for the case of three landmarks in 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 bottom-right we show 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* in left image shifts to the correct location in step 2 after lsho is detected in step 1. Similarly, peak for *lwri* shifts in step 3 once *lelb* is detected.

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 (Figure 1).

We assume that in each image at least one landmark can be detected with high accuracy without additional context (though which one is easy to detect might change from image to image). Our method automatically learns to find the easy landmark, and uses that landmark to provide context to support an image description and so find the next landmark; it then uses those two landmarks to find the third, and so on. In particular, the sequence of landmarks 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.

Contributions: We propose a novel approach that learns a sequential search for finding landmarks in an image-dependent order. It is: 1) simple, easy to implement and reproduce, and computationally efficient; 2) general, as it makes no assumptions about how landmarks relate to each other making it applicable to any landmark localization task (we apply it *as is* on humans as well as birds); 3) able to model appearance and locations of very high order cliques of landmarks jointly.

1.1. Background

Landmark detection is a well-studied problem, usually in the domain of finding human body joints or parts [39, 37, 34, 10, 31, 1, 7, 33]. Our method applies to humans but is designed to apply equally well to other object landmark-finding problems.

Modeling Landmark Dependencies: The modeling options currently available are:

- *Bag*: where one ignores relations (e.g. [6]).
- *Full Relational*: model all pairwise relations leading to intractable inference (e.g. [23, 15, 12, 36, 34]). One solution is to reduce the search space, i.e., by segmentation [24], local searches [34, 17] or cascades [13, 31].
- *Star*: model positions relative to a root (e.g. [22, 5, 16, 3].
- *Tree*: model a subtree of the graph of relations, possibly conditioned on the image (most active for the case of landmark location on humans, e.g. [39, 19, 29, 14, 1]; the best current human parser has this form [4]).
- *K-fan*: a variant of tree models, where the relations preserved form a junction tree of tolerable size (e.g. [5]; a variety of comparable approaches are reviewed in [18]).
- *Implicit*: model relations implicitly, for example with auto-context [28, 35].

We offer an alternative: model relations as an ordered search procedure. Our model assumes that $k-1$ landmarks and a local detector can be used together to determine which of the remaining landmarks should be located on a per-image basis. After locating the head, shoulder, and elbow, the wrist may be the best landmark to locate next in one image, while the hip is best in another. Thus, we substitute sequential inference for joint inference in order to benefit from more expressive dependency models.

Modeling Landmark Appearance: There is a rich set of options for modeling image appearance; space does not allow a comprehensive review. We use sums of rectangles offset from landmark centers, which performs reasonably well, but recent advances using feature learning [4, 32, 33] might improve results.

Learning to Search: Our approach to learn which landmark to find first in an image has similarities with Q learning [2], though, we do not learn an explicit value function. The view of inference as a search procedure has been taken in several existing works [8, 30]. Ratliff *et al.* [30] study this in the context of imitation learning to learn non-linear cost functions. Daume *et al.* [8] treat learning as a parameterized search optimization. Our method uses a computationally efficient greedy search but learns complex intermediate representations and a non-linear scoring function.

2. Learning to Find Landmarks

We want our method to be general and reasonably efficient in inference. Therefore, we avoid any prior knowledge about the relations between the landmarks such as upper/lower body, arms/legs and treat all landmarks similarly. For efficiency, we choose a greedy inference procedure but optimize our model specifically for such inference.

We make the following assumptions about the problem structure: (1) In an image, some landmarks may need more contextual information than others to be accurately detected yielding a loose ordering based on required context, e.g. an occluded wrist. Typically, this information is the location and appearance of other landmarks, (2) A subset of landmarks can be detected with fairly high accuracy without additional context. These can then be used to provide context for the next set of landmarks. Note that these assumptions are quite general and true for most practical applications.

Above assumptions lead naturally to our approach. It learns to detect landmarks one by one in the increasing order of contextual information required. It first automatically learns to find landmarks that do not need much contextual information relying on (2). Then, it uses the set of detected landmarks to provide the context for the next landmark based on (1). As more landmarks are detected

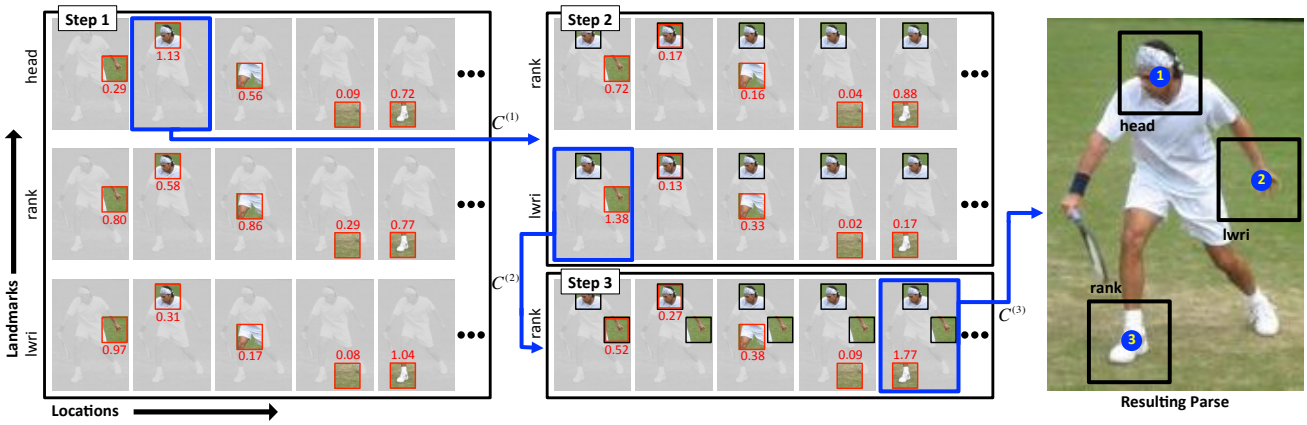

Figure 2. Visualization of the inference procedure using our learned function. In the first step, all locations in an image are evaluated as candidates for all landmarks. Highest scoring of these yields a landmark detection (blue box in left column). In the next step, all the locations are evaluated jointly with the first landmark to yield the next detection and so on until all landmarks are detected.

the available context becomes richer facilitating the detection of harder landmarks. Our approach doesn't assume that some fixed set of landmarks is always easy; it allows these to be different from image to image. 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 (§3.1) of the landmarks and it learns to use them as needed. Thus, *our approach learns not just to score correct landmark locations but also the image dependent order in which to find them.*

2.1. Model, Inference and Training

Let c be the image location of a landmark which can be either a point in the image or $\emptyset$, indicating *unknown*. The current state of our inference is an ordered set of P locations, $C = (c_1, \ldots, c_P)$, one for each target landmark. At each step our algorithm can either replace a $\emptyset$ with a location or stop.

Let $x^T = \Phi(C)^T = [\phi(c_1)^T \ldots \phi(c_P)^T]$ be a feature vector corresponding to C formed by concatenating individual feature vectors for c_i. We define $\phi(\emptyset) = \mathbf{0}$ and use the same ϕ for each landmark. Our algorithm chooses whether to replace $\emptyset$ or stop at current state by evaluating a learned scoring function $\mathbf{F}$ for each possible next state as follows:

$$\mathbf{F}(x) = \mathbf{F}(\Phi(C)) \tag{1}$$

Inference: We use a greedy search strategy (Figure 2). Let $C^{(s)}$ be the state at some step $s \in \{0, \ldots, P\}$. Let $C^{(s)} \oplus c_{ij}$ be a candidate next state obtained by replacing the i-th unknown landmark with the j-th image location. Then we have:

$$C^{(s+1)} = \arg\max_{i,j} \quad \mathbf{F}(C^{(s)} \oplus c_{ij}) \tag{2}$$

Note that at step s, exactly $P - s$ elements of C are $\emptyset$. If the image contains N locations then $\mathbf{F}$ would evaluate $(P - s) \times N$ possible next states.

This clearly places significant demands on $\mathbf{F}$, which much take a large value for good groups of locations and a small value for the bad groups. For example, we require that $\mathbf{F}$ be large for good elbow groups, and good shoulder-elbow groups, and good shoulder-elbow-wrist groups. We use a 5 layer fully connected neural network as the learner to model $\mathbf{F}$ (Figure 3). All the activation functions are rectified linearities except the output which is linear.

There are several interesting properties of this inference scheme. Firstly, it doesn't impose an ordering over landmarks. Thus, they can be detected in one order for one image and in a different order for another image. Figure 7 visualizes the frequencies of detections of a subset of landmarks in each step. While elbows and wrists have a preference to be detected later, shoulders and ankles tend to be detected in earlier steps. Figure 9 shows several images in which the inference followed different order. Secondly, the next landmark is scored in conjunction with already detected landmarks exploiting the context provided by them.

Training: We train our model in an online fashion, so we consider the loss due to a single image $\mathbf{I}$. Let $C^{(s)} = C^{(s-1)} \oplus c_{ij}$ be the state in step s reached by selecting the image location j for landmark i. Similarly, let $C^{(s)}_* = C^{(s-1)} \oplus c^*_{kl}$ be the target ground truth state reachable by selecting the image location l for landmark k. We explain ground truth selection in the next section. Our training loss $\mathbf{J}$ for predicting P landmarks is an average of individual losses J_s for each landmark as follows:

$$\mathbf{J}(\mathbf{I}; W) = \frac{1}{P} \sum_{s=1}^{P} J_s(C^{(s)}, C^{(s)}_*) \tag{3}$$

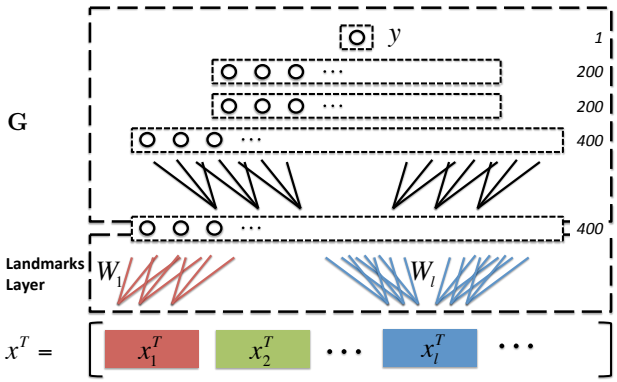

Figure 3. Structure of the neural network learner used to model the scoring function. We refer to the first hidden layer as *landmarks layer* (see text) with colored blocks corresponding to different target landmarks. The rest of the model acts as a scoring function that learns to score landmark groups of various sizes. Our model is fully connected; numbers on the right show the layer sizes.

We would like to train our model $\mathbf{F}$ such that it scores $C_*^{(s)}$ higher than any other $C^{(s)}$. Additionally, it should penalize predictions which are further away from ground truth more than those which are closer. We use a margin based structured loss J_s which uses a candidate dependent margin function $\Delta_{ij,kl}$ to achieve this. Let $x = \Phi(C^{(s)})$ and $x_* = \Phi(C_*^{(s)})$ be the features for $C^{(s)}$ and $C_*^{(s)}$ respectively. Further, let d_{ij}^* be the distance of image location j from the ground truth location of landmark i in image space. Then the loss J_s is defined as follows:

$$J_s(C^{(s)}, C_*^{(s)}) = \max(0, \Delta_{ij,kl} + \mathbf{F}(x) - \mathbf{F}(x_*)) \quad (4)$$

$$\Delta_{ij,kl} = \min(\alpha \min(d_{ij}^*, d_{kj}^*), 1) \quad (5)$$

We use a scaling constant, $\alpha < 1$, to control the steepness of margin function. Note how the margin depends on both the target ground truth landmark for the current step as well as the ground truth landmark of the candidate that was selected. Consider a candidate for a landmark i which lies exactly on its ground truth and the ground truth landmark for the current step k where $i \neq k$. $\Delta_{ij,kl}$ as defined above ensures that such a candidate has a margin of zero w.r.t. k.

Ground Truth Selection: For each training image we have the ground truth landmark locations but we do not have the ground truth ordering $C_*^{(s)}$ in which they should be detected. Our training algorithm dynamically selects $C_*^{(s)}$ by considering the highest scoring candidate for each of the remaining landmarks in step s and picking the one which is closest to its ground truth. This scheme favors the detection of those landmarks in early steps which can be learned easily, i.e., whose predictions tend to fall closer to their ground truths. Further, it enables the learning of an image dependent ordering. We also tried using the ground truth landmark whose candidate was scored highest in the current step. This strategy is very sensitive to initialization and prone to getting stuck in bad local minima.

Practical Considerations: We train our model through back-propagation using stochastic gradient descent with momentum. Although the updates are straightforward for all the layers, we observed an optimization instability. Consider the *landmarks layer* (Figure 3) and let $\{W_L, b_L\}$ with $W_L = [W_1 \dots W_P]$ be the parameters of this layer. We have $x^T = [x_1^T \dots x_P^T]^T = \Phi(C^{(s)})$ with $x_l = \phi(c_l)$ as the features of the individual elements in $C^{(s)}$. We re-write our scoring function making the parameters of this layer explicit.

$$\mathbf{F}(x) = \mathbf{G}(W_L x + b_L) = \mathbf{G}(\sum_{l=1}^{P} W_l x_l + b_L) \quad (6)$$

Thus, W_l is the block of W_L corresponding to the landmark l. If s was the step in which landmark l was detected then the gradient of the objective w.r.t. W_l is

$$\frac{\partial \mathbf{J}}{\partial W_l} = \frac{1}{P}\left(\frac{\partial J_s}{\partial W_l} + \lambda \sum_{i=s+1}^{P} \frac{\partial J_i}{\partial W_l}\right) \quad (7)$$

Earlier terms for $s \in \{1, \dots, s-1\}$ are zeros since x_l is zero in those terms. We introduce a multiplier λ for the later terms. Setting $\lambda = 1$ yields the original gradient. In its original form, different W_l receive gradients of different magnitudes depending on the step in which landmark l was detected. This introduces instability during optimization and hurts the performance of landmarks that are detected earlier. We use λ as a normalizer to counter this effect and set its value to $\frac{0.5}{(P-s)}$ in experiments. We experimented with setting $\lambda = 0$ and found that it is better behaved when compared to using the original gradients but yields slightly inferior final performance when compared to the suggested setting. We set α (eq. 5) in a dataset dependent way such that on an average the margin for a landmark is close to one near other landmarks.

Features are centered and scaled using the mean and range computed over the training set for each dataset. For each image, we consider locations in a grid with a stride of 5 pixels during training and 2 pixels during testing for computational efficiency. We augment the data by adding left-right flips, random crops and small scalings of the images. Models are trained on an NVIDIA K40 to further speed up the training. The training and inference code, along with additional material, is available on the project webpage [1].

3. Experiments

We evaluate our approach on two different landmark prediction tasks: 1) Humans and, 2) Birds. We apply our model

[1] http://vision.cs.uiuc.edu/projects/lssland/

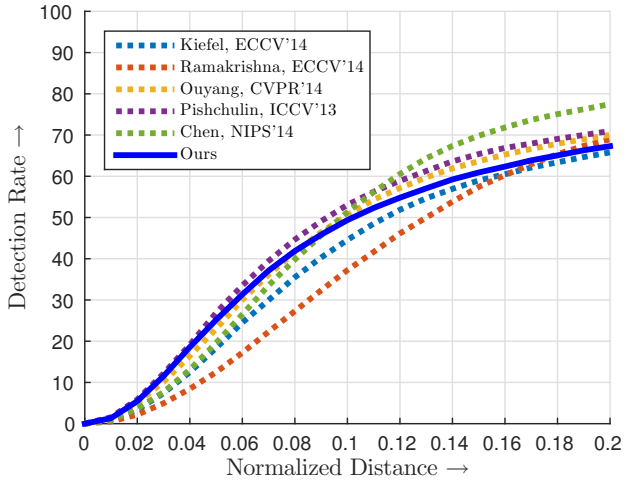

Figure 4. Comparison of our method with state of the art for human landmark detection. We are comparable with the leading approaches in the high precision area of small distances despite using a very simple set of features and no explicit imposed structure.

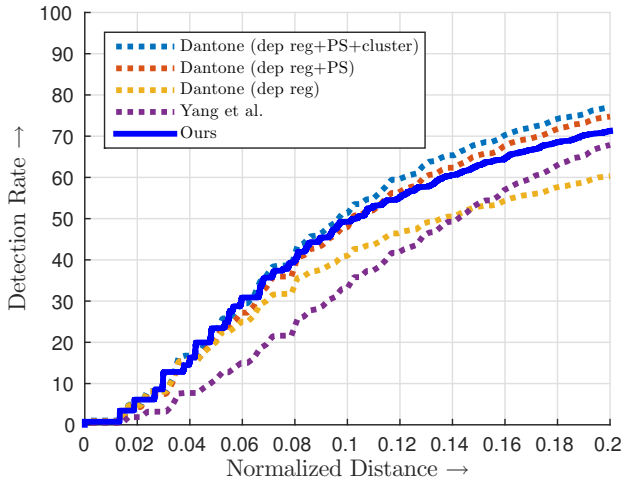

Figure 5. Performance of our approach on Fashion Pose dataset. Our performance is similar to the approach of Dantone *et al.* [7] and better than Yang *et al.* [39]

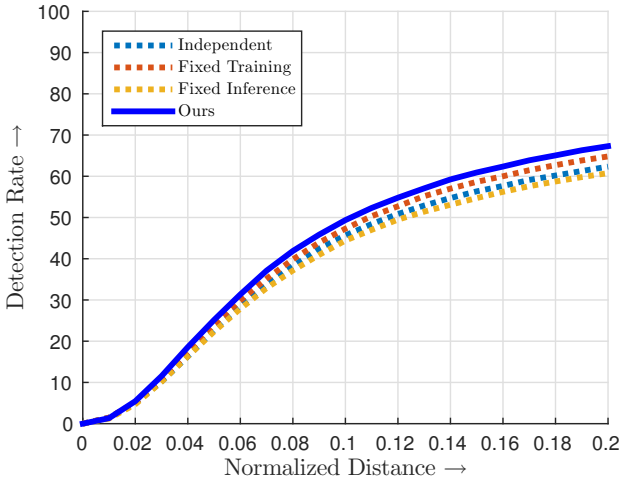

Figure 6. Our method perform better than several variants. Independent trains individual landmark detectors with no interaction. Fixed Training follows a predetermined landmark detection order and thus uses more context than Independent, performing better. Fixed Inference uses our model but follows a fixed order of detection. Our approach utilizes the flexibility of image dependent ordering to outperform all.

Method	Torso	Up. Leg	Lo. Leg	Up. Arm	Forearm	Head	Total
Yang [39]	84.1	69.5	65.6	52.5	35.9	77.1	60.8
Eichner [11]	86.2	74.3	69.3	56.5	37.4	80.1	64.3
Kiefel [21]	84.3	74.5	67.6	54.1	28.3	78.3	61.2
Pishchu- [26]	87.4	75.7	68.0	54.4	33.7	77.4	62.8
Ramakri- [28]	88.1	79.0	73.6	62.8	39.5	80.4	67.8
Ouyang [25]	88.6	77.8	71.9	61.9	45.4	84.3	68.7
Pishchu- [27]	88.7	78.9	73.2	61.8	45.0	85.1	69.2
Chen [4]	92.7	82.9	77.0	69.2	55.4	87.8	75.0
Ours	88.0	77.2	72.7	58.2	34.0	79.9	65.2

Table 1. Comparison of our approach with state of the art on PCP@0.5. We perform similar to several recent approaches as is evident in Figure 6 where our method is close to state of the art in the high precision area.

as is for both the tasks and use the same set of generic features for both. We first describe the features used and then evaluate our approach on several established datasets.

3.1. Features

We use a simple set of features which allow the model to capture appearance and spatial relations between landmarks.

Appearance Features: Each appearance feature for a given location is the average of values in a box of random size, at a random offset, from a randomly chosen image channel. This allows the model to capture relations in appearance between nearby landmarks. We convert the image into 10 channels consisting of 3 Luv channels, 6 gradient orientation channels and 1 gradient magnitude channel [9]. Boxes are constructed by sampling the square root of their areas from the range of $[\sqrt{5}, \sqrt{1000}]$ pixels, their log aspect ratio from the range of $[\log(1/5), \log 5]$ and then solving for an integer width and height. This sampling of areas is biased towards generating smaller boxes which we found to perform better. Offsets are sampled randomly within a circle of radius 50. Biasing the sampling towards smaller offsets performed better than uniform sampling.

Location Features: Location is encoded in two ways.

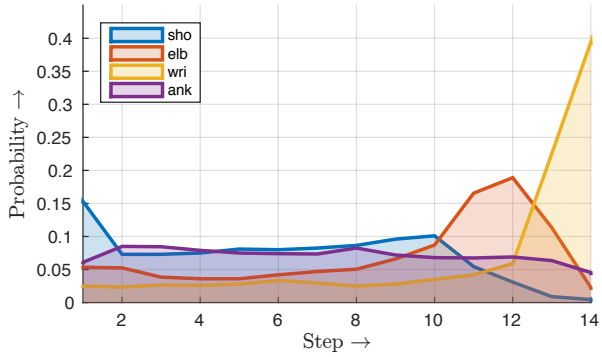

Figure 7. We visualize the frequency of detection of a few landmarks in all the steps. Elbows tend to be detected later while shoulders tend to be detected earlier. However, all have a non-zero frequency at each step.

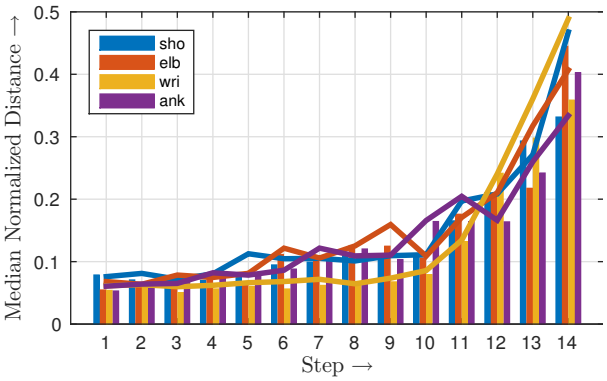

Figure 8. Median errors at various steps. The bars correspond to errors from our model while the lines correspond to errors from the *Independent* baseline. Every landmark had a minimum of 33 detections at any step. From the bars it is clear that, independent of landmarks, earlier steps tend to have lower error indicating that our model has learned to detect them in a meaningful image dependent order. For each step, error for *Independent* for a certain landmark is computed by first selecting those images in which our method detected that landmark in that step and reporting the median error of *Independent* over those. There is high correlation between cases that our method finds hard (detects in later steps) and cases that the independent detector finds hard. Thus, **our method is able to pick out landmarks which are more likely to be correct and is detecting them early yielding reliable context for the later steps**. Also note that there is larger gain in later steps where the context would play a larger role.

First, directly normalized (x, y) yield two features. Normalization is done by assuming the origin as the center of the image and dividing by the maximum of width and height of images, m, across the whole dataset. Next, location is encoded in terms of some fixed points. For any image, we set down 20×20 fixed points at equal intervals in range $[-m/2, m/2]$ from the center along both axis. Posi-

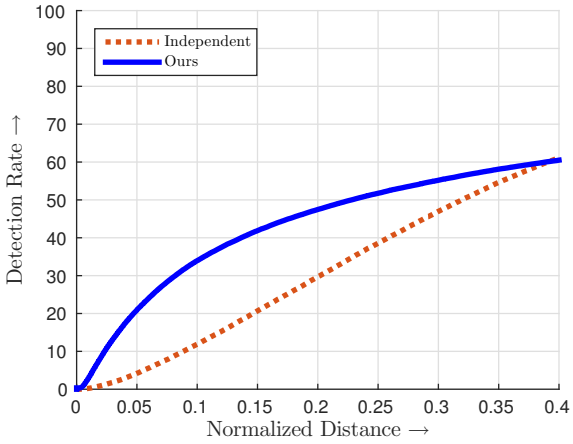

Figure 11. Comparison of our approach on CUBS 200 dataset with a baseline of independent detectors. Our method is able to exploit the context yielding significant improvement. Normalized distance is computed by dividing the pixel distance by the height of the bird.

tion is then encoded as a 400 dimensional vector capturing proximity to these points. The proximity is computed by $1 - \frac{min(d,r)}{r}$, where d is the distance to the point and r is a dataset dependent constant computed as $0.15 \times m$ to ensure some points are always in proximity. We found that these features improve the recall for larger error thresholds.

3.2. Datasets

We demonstrate the performance of our approach for human landmarks on the Leeds Sports Dataset (LSP) [20] and the Fashion Pose dataset [7]. LSP contains 1000 training and 1000 testing images of humans in difficult articulated poses with 14 landmarks. We use the Observer Centric (OC) annotations [11]. Fashion Pose Dataset contains 6530 training and 775 testing images of human models with 13 landmarks. This dataset has less pose variation but significant appearance variation due to clothing. Further, to demonstrate the general applicability of our approach we apply it to the Caltech-UCSD Birds 200 (2011) dataset (CUBS 200) [38]. It contains 5994 training and 5794 testing images with 15 landmarks for birds. For all the datasets we work with provided train and test splits.

Evaluation Metric: We adopt the generally accepted metric of plotting detection rate against normalized distance from ground truth. Normalization is done using the torso height. PCP hides improvements in high precision region, but we report it on LSP to compare with existing work.

3.3. Comparison with the SOA Parsing

Figure 4 and 5 show that our method compares favorably with the state of the art on both the LSP and the Fashion Pose datasets. Note that our performance is close to

Figure 9. **Qualitative Results for different error percentiles of Leeds Sports dataset**: We sort the test images by sum of squared errors of our predictions from the ground truth. Each rows shows results on atypical poses from different quarters of this ordering with first row being most accurate. Numbers in blue circles show the step in which each landmark was detected. We also display the limbs as a visual aid. Different images exhibit different detection order. From top to bottom, as poses get harder, wrists quickly lose precision. Other failure modes include confusion with clutter, confusion with other people and under-represented poses such as inverted people.

Figure 10. Qualitative examples of landmark detections on CUBS 200 birds dataset. Although we detect all the landmarks, we visualize a subset, {*beak, belly, leftleg, tail, back, nape*}, to avoid clutter. Asterisks show their ground truth locations, blue circles with colored borders show the corresponding detections and the numbers indicate the step in which they were detected. Lines connecting detections are shown as a visual aid and are color coded to denote same pair of landmarks. Our approach does quite well in comparison to a baseline of independent detectors (See Figure 11). Note the difference in order in which the landmarks are detected from image to image.

the best performing approaches in the high precision area. Figure 9 shows qualitative results on LSP from various error quartiles. Table 1 compares our method with a few more approaches on LSP using PCP@0.5. Our performance is close to several recent approaches, though, other methods such as Chen *et al.* [4] outperform with a looser criterion (Figure 4). However, note that in contrast to other state of the art body parsers, our method can be applied without modification to parse bird landmarks too (§3.6).

3.4. Ordering and Relations

Figure 1 visualizes the implicit spatial model learned by our method in two images. It is clear that the landmarks are detected in different orders and the ones detected earlier help localize the harder ones as shown by the shift in peaks. Our opportunistic ordering strategy is able to use appearance and spatial relations to improve performance.

In Figure 6, we compare our method with several variants and show that it improves upon them. *Independent* trains all the landmark detectors independently. *Fixed Training* trains the model using a fixed ordering of $head \rightarrow shoulders \rightarrow elbows \rightarrow wrists \rightarrow hips \rightarrow knees \rightarrow ankles$. Note that this ordering deliberately puts harder landmarks like $wrists$ and $ankles$ after $shoulders$ and $hips$ as they may be needed as context. *Fixed Inference* uses our trained model but follows the aforementioned fixed ordering during inference. *Fixed Training* does better than *Independent* by exploiting the context provided by earlier landmarks. Our method improves upon it by allowing a flexible choice of ordering and learning a richer dependency model. *Fixed Inference* does the worst as a fixed ordering forces it into a part of search space for which it has not learned.

Figure 7 shows that the learned model is indeed using an image dependent ordering. It plots the frequency of a few landmarks being detected in various steps. Although harder landmarks such as wrists and elbows tend to be detected later, there is a significant spread over the steps. Other landmarks (not visualized) show a trend similar to that of ankle.

Figure 8 shows that landmarks which are detected in earlier steps tend to have less error. We plot median normalized distance of the predictions for various landmarks at each of the steps. Irrespective of the landmark type, landmarks which are detected earlier tend to have lower error. This is expected since landmarks that are hard for an image will be detected in later steps.

3.5. Uniqueness of Detection Orderings

We counted the number of unique orderings of landmark detections in test images and found that our method follows a unique ordering for each image in both the Leeds Sports Dataset (1000 test images) and the Fashion Pose Dataset (775 test images). This, in conjunction with Figure 8, indicates that an image dependent ordering is indeed useful.

Ordering	% Occurrences
{*left-shoulder, left-elbow, left-wrist*}	62.3
{*left-shoulder, left-wrist, left-elbow*}	11.6
{*left-elbow, left-shoulder, left-wrist*}	9.6
{*left-wrist, left-shoulder, left-elbow*}	7.9
{*left-wrist, left-elbow, left-shoulder*}	5.3
{*left-elbow, left-wrist, left-shoulder*}	3.3

Table 2. Percent occurrences of detection orderings in LSP for a restricted set of landmarks, namely {*left-shoulder, left-elbow, left-wrist*}. Clearly first order is preferred but other orderings are common as well.

However, a random model may also yield unique orderings. To study this, we consider only the {*left-shoulder, left-elbow, left-wrist*} landmarks and show the percentage occurrences of various orderings for LSP in Table 2. Clearly, the intuitive ordering of {*left-shoulder, left-elbow, left-wrist*} is preferred, but other orderings are common as well. A χ^2 test rejects the null hypothesis of a uniform distribution with a significance value of zero, clearly showing that the orderings are not uniformly random.

3.6. Finding Landmarks on Birds

Landmark finding is a general problem, and our approach is especially well suited for a general application as it doesn't make any strong assumptions about the landmarks and their relations. We verify this by applying our method *as is* to the demanding problem of finding landmarks in birds on CUBS 200 dataset. Unlike human pose datasets, there is no clear intuitive ordering of landmarks in this dataset making our approach all the more appealing. Figure 10 shows qualitative results while Figure 11 compares our method with an *Independent* baseline which learns a set of detectors independently. Our method shows significant gains over this baseline by using the learned ordering to better propagate context.

4. Conclusion

We described a general method to find landmarks in images by greedily expanding groups, exploiting appearance and contextual information within the group to identify the next best landmark using a learned scoring function. Our method learns to detect landmarks in an image dependent order. Using a very simple set of features, our method is able to achieve good performance. Further, we have shown that our method performs very well on two distinct landmark finding problems underlining its general applicability. Learning the low level features and stronger encoding of context are some of the promising future directions.

Acknowledgments: This material is based upon the work

supported in part by the National Science Foundation under Grants No. IIS 09-16014, IIS-1421521, and IIS-1029035, ONR MURI Award N00014-10-10934, and a Sloan Fellowship. We would also like to thank NVIDIA for donating some of the GPUs used in this work.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 2

[2] A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998. 2

[3] L. Bourdev and J. Malik. Poselets: Body part detectors training using 3d human pose annotations. In *ICCV*, 2009. 2

[4] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. *arXiv preprint arXiv:1407.3399*, 2014. 2, 5, 8

[5] D. Crandall, P. Felzenswalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR*, 2005. 2

[6] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision*, 2004. 2

[7] M. Dantone, J. Gall, C. Leistner, and L. van Gool. Human pose estimation from still images using body parts dependent joint regressors. In *CVPR*. IEEE, 2013. to appear. 2, 5, 6

[8] H. Daumé III and D. Marcu. Learning as search optimization: Approximate large margin methods for structured prediction. In *ICML*. ACM, 2005. 2

[9] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 5

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

[11] M. Eichner and V. Ferrari. Appearance sharing for collective human pose estimation. In *ACCV*, 2012. 5, 6

[12] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. 28(4):594–611, 2006. 2

[13] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *CVPR*, 2010. 2

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, January 2005. 2

[15] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. *CVPR*, 2003. 2

[16] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *CVPR*, 2005. 2

[17] V. Ferrari, M. Marín-Jiménez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 2

[18] K. Grauman and B. Leibe. Part-based category models. In *Visual Recognition*. 2

[19] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. In *ICCV*, 2001. 2

[20] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. doi:10.5244/C.24.12. 6

[21] M. Kiefel and P. V. Gehler. Human pose estimation with fields of parts. In *ECCV*. Springer, 2014. 5

[22] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, pages I: 878–885, 2005. 2

[23] T. Leung, M. Burl, and P. Perona. Finding faces in cluttered scenes using random labelled graph matching. In *ICCV*, 1995. 2

[24] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configuration: Combining segmentation and recognition. In *CVPR*, volume 2, pages 326–333, 2004. 2

[25] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *CVPR*. IEEE, 2014. 5

[26] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 588–595. IEEE, 2013. 5

[27] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3487–3494. IEEE, 2013. 5

[28] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*, 2014. 2, 5

[29] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, volume 19, 2006. 2

[30] N. D. Ratliff, D. Silver, and J. A. Bagnell. Learning to search: Functional gradient techniques for imitation learning. *Autonomous Robots*, 2009. 2

[31] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010. 2

[32] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *CoRR*, abs/1406.2984, 2014. 2

[33] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. *arXiv preprint arXiv:1312.4659*, 2013. 2

[34] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*, 2010. 2

[35] Z. Tu. Auto-context and its application to high-level vision tasks. In *CVPR*, 2008. 2

[36] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, 2008. 2

[37] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011. 2

[38] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 6

[39] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 2013. 2, 5