Active Learning for Human Pose Estimation

Buyu Liu
University of Edinburgh
buyu.liu@ed.ac.uk

Vittorio Ferrari
University of Edinburgh
vittorio.ferrari@ed.ac.uk

Abstract

Annotating human poses in realistic scenes is very time consuming, yet necessary for training human pose estimators. We propose to address this problem in an active learning framework, which alternates between requesting the most useful annotations among a large set of unlabelled images, and re-training the pose estimator. To this end, (1) we propose an uncertainty estimator specific for body joint predictions, which takes into account the spatial distribution of the responses of the current pose estimator on the unlabelled images; (2) we propose a dynamic combination of influence and uncertainty cues, where their weights vary during the active learning process according to the reliability of the current pose estimator; (3) we introduce a computer assisted annotation interface, which reduces the time necessary for a human annotator to click on a joint by discretizing the image into regions generated by the current pose estimator. Experiments using the MPII and LSP datasets with both simulated and real annotators show that (1) the proposed active selection scheme outperforms several baselines; (2) our computer-assisted interface can further reduce annotation effort; and (3) our technique can further improve the performance of a pose estimator even when starting from an already strong one.

1. Introduction

Human pose estimation, the localization of human body joints, has enjoyed substantial attention. Starting from classical pictorial structures [2, 12, 14], recent state-of-the-art approaches employ convolutional networks [36, 55, 47, 6]. These methods aim to learn discriminative patterns that enable to distinguish patches around body joints from the rest of the image. This requires good training data, but data collection is particularly time-intensive for human pose estimation, as annotators are typically asked to click on 14 joints per person [1]. The reference analysis paper [1] suggests a reasonable annotation rate of one pose per minute.

Weakly supervised learning [9] and active learning [38, 8] have been proposed to address data collection problem for several tasks, such as image classification [22, 20, 35], object detection [51, 57] , object recognition [21, 13] and semantic segmentation [49, 27, 44, 16]. However, the problem remains largely unaddressed for human pose estimation.

In this paper, we propose the first active learning approach for human pose estimation (Fig. 1). We follow the general scheme of active learning: an active learner automatically selects a subset of unlabelled data. After that, human annotators label the selected data. Finally, the learner updates the pose estimator with the labelled data, and the process iterates. Since the goal of active learning is to maximize performance while minimizing annotation effort, we focus on two main elements in the scheme: active selection and human annotation procedures.

For active selection, we first explore various individual cues to measure the informativeness of as-yet unlabelled images. We first adapt classical active learning cues to the human pose estimation task, such as highest model probability [25, 26], best v.s second best [37, 20], and influence [17, 42]. In addition, we propose an uncertainty measure which takes into account the spatial distribution of the model's response on an image, coined *multiple peak entropy*. This cue provides a better estimation of the images where the model is uncertain on. Moreover, we propose a dynamic way to combine the influence and uncertainty cues, where their weights vary during the active learning process. During the early selection iterations when the pose estimator only sees little annotated data, the influence cues play a more important role. Later, as the model gets better, our scheme gradually switches to rely more on uncertainty. Our weighting term approximates the expected reliability of the current pose estimator on unlabelled images, and its value increases as the estimator gets better.

To further reduce annotation effort, we propose a computer assisted interface to help the annotator to rapidly click on a body joint (Fig. 5). The main idea is to discretize the image space into large regions, each associated with a single candidate point for the true location of the joint. Thanks to this, the annotator no longer needs to click exactly on the joint, but just anywhere inside the associated region. To find the set of candidate points, we use the full 2D distribu-

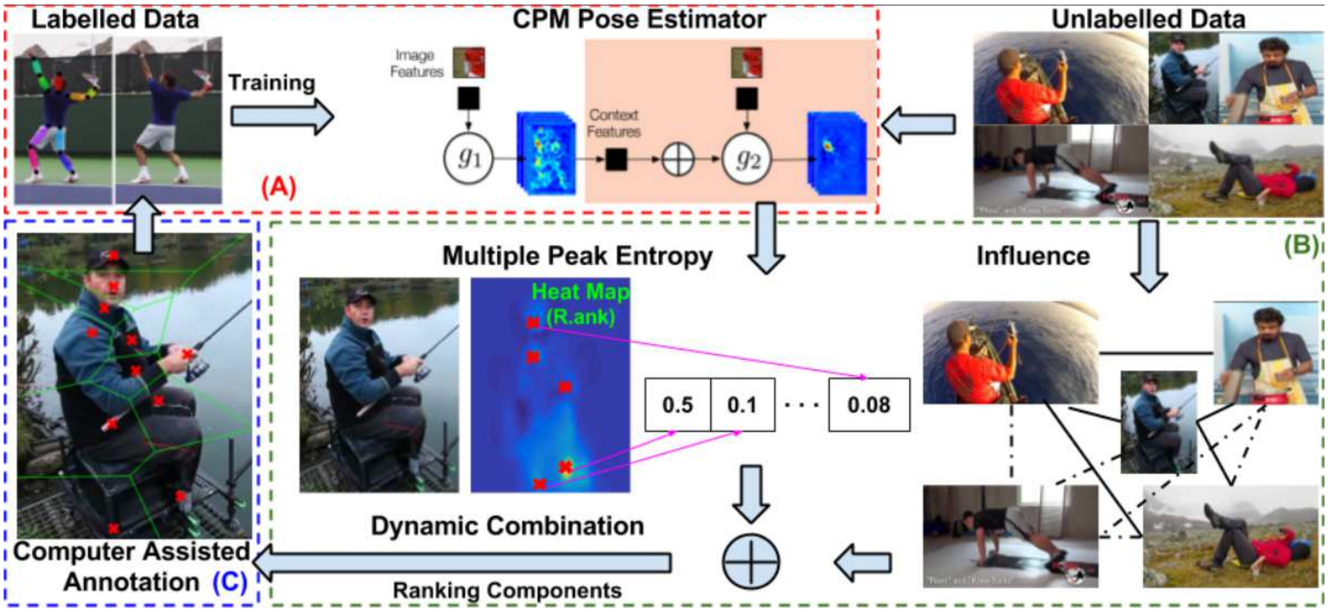


Figure 1. Overview of our approach. We begin with CPM estimator pre-trained on a small set of labelled images; then search the large unlabelled pool for informative components/images to annotate. Our novel active learning strategy dynamically combines influence and uncertainty cues, where the uncertainty is measured with our proposed multiple peak entropy that favours images with multiple weak local peaks in their predicted heat maps. Then our proposed annotation system requests human annotations. Our computer assisted annotation interface further saves the annotation time of clicking on joint locations by reducing localisation space for human labeller.

tion of responses of the current pose estimator on an image (heat maps): each local maxima of the heat map becomes a candidate. We then divide the image into non-overlapping regions consisting of all pixels closer to a candidate than to any other. Users then right click anywhere in a region to select the corresponding candidate point as the true joint location. If the true location is not among the candidates, users can left click on any other point (which takes the same time as the standard annotation interface).

We perform extensive active learning experiments using the challenging MPII [1] and LSP datasets [19]. A first series of experiments using simulated annotators demonstrates that: (1) our proposed multiple peak entropy cue outperforms previous uncertainty-based cues; (2) our proposed dynamic combination of influence and uncertainty cues further improves active selection over individual cues and outperforms a static combination strategy. (3) our method can further improve the performance of a pose estimator even when starting from an already strong one, initialized from a large training set. Moreover, we carry out experiments with real human annotators. These lead to comparable results to what achieved by the simulations, showing that (4) our method is robust to the noise naturally introduced by real annotators. Finally, we validate our proposed computer assisted interface to reduce the time to click on a joint. We found that (5) it saves about 33% annotation time on average, without reduction in performance. Overall, combing all elements we propose produces 80% of the performance of a model trained from the full MPII training set, in just 23% of the total annotation time (i.e. using multiple peak entropy, dynamic combination, and assisted interface).

2. Related Work

Human Pose Estimation. Pictorial structures are one of the classical approaches to articulated pose estimation [2, 12, 14, 40, 33]. In these methods, spatial correlations between parts of the body are expressed as a tree-structured graphical model with kinematic priors that couple connected limbs. To extend the model representation power, more flexible methods, such as non-tree models [54, 45, 24], propose to investigate different structures to model the spatial constraints among body joints on score maps. Recently, CNN-based methods [36, 55, 47, 48, 7, 29] have enjoyed considerable success. DeepPose [48] takes the first step towards adopting CNN [23] for human pose estimation, where CNN is used to directly regress joint locations in Cartesian coordinates repeatedly. Subsequently, graphical models have been introduced to incorporate spatial relationships between joints either as a post-processing [6] or in an end-to-end manner [47]. More recent work [5, 55] proposed to build up dependency among input and output spaces, where predictions at previous steps are concatenated with the image as input of the current step to iteratively refine predictions. Our work is built on the state-of-the-art method [55].

Active Learning The core problem of active learning is to quantify the informativeness of an as-yet unlabelled example [38, 8, 22]. Selection strategies include uncertainty sampling [25, 41], reducing the classifier's expected error [52, 38], maximizing the diversity among the selected images [15, 17], or maximizing the expected labelling change [49, 13]. In computer vision, active learning has been used for scene classification [20, 35] and annotating large image and video datasets [56]. In addition to these unstructured prediction tasks, researchers also explored active learning for structured prediction, e.g. semantic/geometric segmentation [44, 27, 49, 16]. These methods either annotate the most marginally uncertain single variable/pixel by estimating the local entropy of the marginal distribution [27] or request labels of the most uncertain image by estimating the entropy of the joint distribution [44]. In contrast, Maji et al. [28] propose a novel uncertainty measurement for structured models, which estimates upper-bounds of the true entropy of the Gibbs distribution via MAP perturbations [31]. In this paper, we tackle active learning for human pose estimation for the first time. We propose an uncertainty measure specific to this task, and a dynamic combination strategy that outperforms several active selection alternatives proposed in other domains.

User Interaction. Interactive techniques provide another way to minimise manual effort, e.g. tools for efficient video annotation [53] and object labelling [34]. Methods that intelligently design the query space [39, 32, 30] also share the spirit of reducing annotation effort. Other works have looked into active learning schemes that query for multiple types of annotator feedback [50, 4, 43]. In this paper, we propose a new computer assisted annotation interface for human pose estimation. It leverages the predictions of the current pose estimator to guide the annotator while it clicks on a joint, reducing annotation time by one third without damaging accuracy.

3. Approach

We are given a small set of images with full human pose annotations F_s, which we use to train an initial human pose estimator, and a large set of unlabelled images F. The goal is to obtain body joint locations in the unlabelled set and to train a strong human pose estimator while minimizing human annotation effort.

Our framework iteratively alternates between (A) retraining the pose estimator using all currently available annotations; (B) actively select a subset of the unlabelled images; (C) human annotators label the selected images. We focus on steps (B) and (C) since they are the two main factors for reducing annotation effort. For step (B), we propose a new uncertainty measurement and a strategy to combine multiple cues in a dynamical manner. For step (C), we further reduce the annotation effort by introducing a computer assisted interface, which reduces human localisation space by leveraging the current estimator predictions. We discuss the three steps in detail in the following sections.

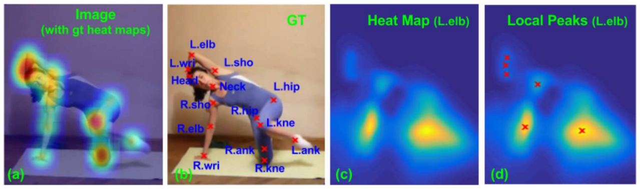


Figure 2. (a) shows input image (with ground-truth heat maps). (b) illustrates pixel-level ground-truth annotations for all 14 joints. (c) shows heat map prediction for $L.elb$ (left elbow for short). (d) visualizes local peak predictions of $L.elb$.

Notation. We denote by U_t the set of unlabelled image and joint pairs at training iteration t and denote by L_t the set of labelled pairs. We denote the dataset that our active learner works on as F, where $F = U_t \cup L_t$ and $U_t \cap L_t = \emptyset$. F_s denotes a separate, small set of fully labelled images that we use to initialize our pose estimator. Each image I_i has a person with $p = \{1, \ldots, P\}$ body joints (Fig. 2(b)). We associate a binary variable $I_i^p \in \{0, 1\}$ with p-th joint in image I_i. $I_i^p = 1$ if and only if this joint is labelled. U_0 is F. The goal of pose estimation is to predict the joint locations $Y = \{Y_p\}$, where $Y_p = z$ defines the location of the p-th joint and $z = (u, v) \in \mathcal{Z} \subset \mathbb{R}^2$ is the 2D coordinate.

3.1. Step (A): Model Training

In this step, we re-train the human pose estimator. We use Convolutional Pose Machine (CPM) [55], but other approaches [29, 7, 5] that predict 2D heat maps could be used.

CPM is a CNN-based sequential prediction framework. It consists of several stages $n \in \{1, \ldots, N\}$, each of which encodes both appearance cues and context information as its features. Specifically, the contextual cues are incorporated in the form of predictions from previous stage. Each stage of the pose machine is trained to produce the belief maps for the locations of the joints. A typical 2D heat map generated by the CPM model is shown in Fig. 2(c). The CPM model encourages the network to iteratively approach the correct location by defining a loss function at the output of each stage that minimizes the L_2 distance between the predicted and ground-truth heat maps for each joint.

At iteration t of active learning, we obtain a set of labelled pairs L_t. For each non-zero I_i^p, the annotated ground-truth location of joint p is denoted as $\hat{Y}_p$. Following [55, 47], we can generate a ground-truth belief map $b_p^t(\hat{Y}_p = z)$ for $\hat{Y}_p$ by putting Gaussian peaks at location z of each body joint p (Fig. 2(a) for example). Then the loss function of the n-th stage of CPM that we aim to minimise is defined by:

$$f_n^t = \sum_{I_i,p} \sum_{z \in \mathcal{Z}} I_i^p \| b_p^t(\hat{Y}_p = z) - b_p^t(Y_p = z) \|^2 \qquad (1)$$

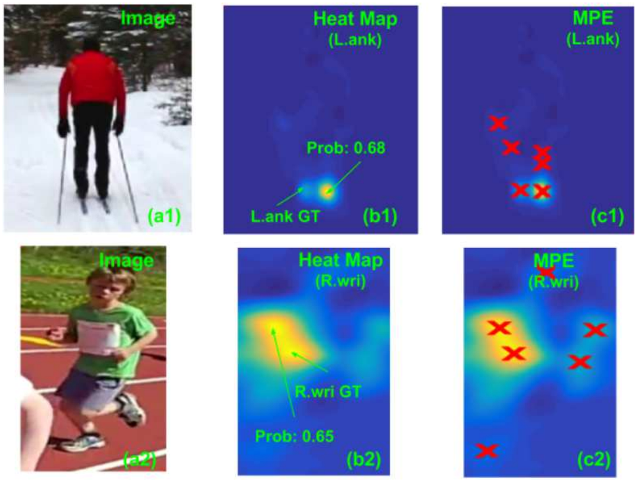


Figure 3. (a$*$), (b$*$) and (c$*$) shows an example image, heat map prediction of $L.ank$ or $R.wri$, and our MPE measurement, respectively. In (b1) and (b2), 'Prob' indicates the highest probability point in each heat map, while 'GT' indicates the ground-truth position of the joint. In (c1), (c2) red crosses shows all local maxima of the heap map, which we use to compute MPE.

The overall loss function of CPM is obtained by adding the losses at each stage and defined as follows:

$$\mathcal{F}^t = \sum_{n=1}^{N} f_n^t \tag{2}$$

where N is the number of stages in CPM.

Applying CPM to any image I_i would generate a set of heat maps. We denote $\mathbf{b}_p^t \in \mathcal{R}^{w \times h}$ as all the beliefs of joint p evaluated at every location z in the image with CPM trained at t-th active learning iteration, where w and h are the width and height of the I_i, respectively. Then the generated set of belief maps is denoted as $\mathbf{b}^t = \{\mathbf{b}_p^t\}_P \in \mathcal{R}^{w \times h \times (P+1)}$ (P joints plus one for background).

3.2. Step (B): Active Selection

We now describe how we actively select the most informative images for annotation. In each active learning iteration $t \in \{0, \ldots, T\}$, we solicit annotations for the actively chosen batch S_t, and augment L_t with the newly labelled data: $L_t = S_t \cup L_{t-1}$. Our active selection algorithm considers both influence and uncertainty cues. The influence accounts for influential property among images, where images that are similar to other unlabelled images are more valuable as they are likely to propagate information [17]. Conversely, the uncertainty is measured on individual body joints inside each image. Our contributions lie in a new uncertainty measurement, that we call multiple peak entropy, and a dynamic combination of multiple cues.

Uncertainty Uncertainty aims to find unlabelled images where the current pose estimator is not confident to have localized the joints correctly. In those images, it is more likely to have made mistakes. For each image $I_i \in U_t$, we can obtain the heat maps $\mathbf{b}^t$ at the t-th active learning iteration. Typically, uncertainty is measured by the Highest Probability (HP) [25, 26] among all possible outputs for a variable. In our case, a variable is a joint and the possible outputs are all pixels in an image. So the HP criterion for selecting the p-th joint in image I_i can be written as:

$$C_{HP}(I_i, p) = (1 - \max_z b_p^t(Y_p = z|I_i)) * (1 - I_i^p) \tag{3}$$

However, this criterion considers only the highest probability in the heat map, ignoring the information about the remaining distribution. To address this, [37] proposes margin sampling (aka. Best vs Second Best (BSB)), and [42] uses entropy-based methods.

None of these methods is ideal for human pose estimation. Fig. 3(b1) and (b2) show example heat maps for *L.ank* or *R.wri*, respectively. Note how there are typically multiple modes in a heat map. Hence, despite the presence of a high probability peak (Prob), the location predictions are actually wrong (i.e. not on the GT position), and the Highest Probability criterion is not able to identify these examples as uncertain. Moreover, the modes are widely spread and these heat maps are spatially diffuse. The BSB criteria would return scores near 0 in these cases, as the second best pixel in the heat map is just next to the top scoring pixel, with nearly identical value. Similarly, plain entropy would not be able to differentiate between a single wide mode (likely to be a correct case) and multiple tighter modes (an uncertain case).

To improve on this, we propose a **Multiple Peak Entropy** (MPE) criterion. MPE considers the above mentioned properties and accounts for inherent spatial relations between pixels in the heat map. These are not independent possible output values, but they form a spatial structure instead. Specifically, we find all locally-optimal predictions for the p-th joint by applying a local maximum filter on $\mathbf{b}_p^t$. We denote this set of peaks by $\mathcal{M}$, where each peak $m \in \mathcal{M}$ has coordinates $z_m = (u_m, v_m)$, and prediction confidence $b_p^t(Y_p = z_m|I_i)$. We define the normalised prediction as:

$$\text{Prob}(I_i, m, p) = \frac{\exp b_p^t(Y_p = z_m|I_i)}{\sum_m \exp b_p^t(Y_p = z_m|I_i)} \tag{4}$$

Finally, the MPE uncertainty of joint p in image I_i is quantified as:

$$C_{MPE}(I_i, p) = \sum_m -\text{Prob}(I_i, m, p) \log \text{Prob}(I_i, m, p) \tag{5}$$

C_{MPE} favours joints that have multiple weak peaks in their heat maps. The reason why MPE works in human pose estimation task is that it provides a compact but multi-mode aware representation for heat map predictions. On the one hand, MPE ignores the information around the local peaks,

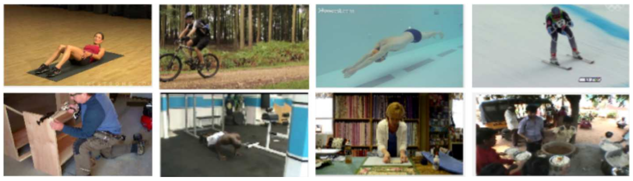


Figure 4. The top and bottom column shows typical high and low influence images, respectively. High influence images are typically uncluttered and more prototypical. This kind of images occurs more frequently in the dataset we consider.

which enables us to tackle the over-smooth property in heat maps. On the other hand, it handles the multi-mode property by collecting predictions of these local peaks and measuring their entropy. As illustrated in Fig. 3(c1) and (c2), our MPE handles both cases quite well.

Influence. Unlabelled images that are similar to many others are good candidates for annotation, because they would then effectively propagate their labels to others. We denote the influential property [17] of an unlabelled image I_i at iteration t as:

$$C_{INF}(I_i) = \frac{1}{|U_t| - 1} \sum_{I_j \in U_t \setminus I_i} d(I_i, I_j) \tag{6}$$

where $|U_t|$ denotes the number of unlabelled images. $d(*, *)$ is an appearance distance function. We denote $I_j \in U_t$ when $\sum_p I_j^p < P$. Examples of high and low influence images can be seen in Fig. 4. We see that high influence images are typically uncluttered and more prototypical. This kind of images occurs more frequently in the dataset we consider.

Dynamic Combination. A good active learning scheme should considers both uncertainty and influence cues simultaneously. To this end, we further introduce $W(U_t)$ to dynamically combines these two cues. $W(U_t)$ measures how reliable our pose estimator is on unlabelled images. Ideally, the reliability would be defined as the estimator's expected error. However, exact estimation of CPM's reliability is computationally intractable. Thus we approximate the expected error with the following:

$$W(U_t) = \frac{1}{|U_t|} \sum_{I_i \in U_t} \text{Prob}(I_i, m^*, p^*) \tag{7}$$

where for the p-th joint in image I_i, m^*, p^* is defined as:

$$(m^*, p^*) = \underset{m,p}{\arg\min}\ \text{Prob}(I_i, m, p) \tag{8}$$

we define the value of selecting image $I_i \in U_t$ at t-th iteration as:

$$\begin{aligned} C_{DC}(I_i) = &(1 - W(U_t)) * C_{INF}(I_i) + \\ &W(U_t) * \frac{1}{P} \sum_p C_{MPE}(I_i, p) * (1 - I_i^p) \end{aligned} \tag{9}$$

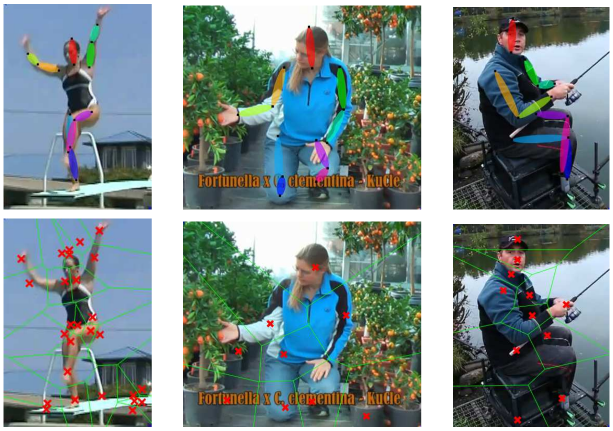


Figure 5. Three examples for VD-based annotation. The top row are the ground-truth annotation. The second row are the VD figures for left knee. The green lines define the boundaries of VD regions while red cross are the local peaks obtained from predicted heat maps.

Given a selection criterion, an active learner would rank components and select top k_t joints at t-th iteration. These joints are then fed to human annotation interface (Sec. 3.3).

Note that our proposed active learning scheme differs from previous methods [17, 38, 8] in several ways. Firstly, none of the existing methods focus on human pose estimation task. Secondly, we propose a novel MPE uncertainty measurement. Finally, our scheme dynamically combines both uncertainty and influence cues. Both the MPE and dynamic fusion prove to be effective for human pose estimation (see Fig. 6 for results).

3.3. Step (C): Human Annotation

The conventional way to annotate body joints is to ask the annotators to click on the exact pixel where the required joint is located. However, such annotation is very time consuming, as annotators are required to label 14-16 [1, 19] joints per person. The annotation time of labeling a joint is shown in Fig. 8(a). These timings are obtain from one annotator in our university.

To further reduce the annotation effort, we propose a computer assisted interface. Given an image I_i and the required joint p, instead of providing the raw image to users, we generate candidates for where joint p may be located. To achieve this, we first compute the predicted heat map for the joint with the current pose estimator. Then we take the local peaks of this heat map as the candidates for the joint location. Given these location candidates, we divide the image into non-overlapping regions consisting of all pixels closer to that candidate than to any other. Specifically, we generate a Voronoi Diagram (VD) based on the candidates. If the true joint location is included in the candidate pool, the annotator can right-click anywhere on the corresponding region.

This takes less time than clicking on the exact location. If the true joint location is not among the candidates, the annotator can left-click on the true location. This takes the same time as the unassisted annotation setting. We refer to our interface as **VD-based annotation** and show some examples in Fig. 5.

4. Experiments

4.1. Experimental Settings

We report extensive experiments to evaluate various aspects of our work. Sec. 4.2 compares several active selection cues, including cues previously used for other tasks, and our proposed multiple peak entropy. Sec. 4.3 combines multiple cues and compares our proposed dynamic combination to a simpler static combination baseline. Sec. 4.4 studies the effect of using real human annotators, instead of simulations. In sec. 4.5 we carry out active learning starting from an already strong pose estimator. Finally, in sec. 4.6 we explore the benefits brought by our proposed computer-assisted annotation interface.

Datasets. We use two datasets: MPII Human Pose [1] and the Extended Leeds Sports Pose [19] (LSP). For MPII, we use the training set (25K person samples) and the validation set (3K samples). For LSP, we use the training set (11K person samples) and the test set (1K samples). In both MPII and LSP a person is represented by $P = 14$ body joints.

Protocol. All experiments in Sec. 4.2-4.4 and 4.6 follow the same protocol: (1) we train an initial pose estimator on 100 fully labelled images randomly sampled from the LSP training set; (2) we perform active learning on the MPII training set, iteratively adding samples and re-training the pose estimator with all samples labelled so far; (3) at each iteration, we evaluate the current pose estimator on the MPII validation set and report its performance on it.

The experiments in Sec 4.5 differ in that we use a much larger initial training set in step (1) and evaluate on a different set in step (3) (see Sec. 4.5 for details).

As common in the active learning literature [50, 27, 44], we simulate annotations in step (2) by using the ground-truth annotations provided in the MPII training set in Sec. 4.2, 4.3 and 4.5. In Sec. 4.4 instead, we use real human annotators.

Implementation Details. We use the publicly available Caffe [18] framework as well as the CPM code provided by [55] to train our model. We set the number of CPM stages N (Eq. (2)) to 6. We define our $d(I_i, I_j)$ as the cosine similarity between the appearance features on I_i and I_j (Eq. (6)), computed by applying AlexNet [23] pre-trained on ImageNet [11] on the target image and extracting the $fc6$ layer output. We also tried the diversity cues as suggested in [17] but they led to slightly worse results. To obtain local peaks, we apply a 5×5 local maximum filter on the heat maps. The number of active learning iterations T is set to 5 and the number of joints P is 14. The number of joints selected at each active learning iteration t is referred to as k_t, and progresses as follows during iterations: $[5\%, 5\%, 20\%, 20\%, 20\%, 20\%] * J_F$ and J_F denotes the number of joints on MPII training set F. Definitions of T and k_t can be found in Sec. 3.2. Note that we select fewer joints at early iterations because we especially care about the active learning performance with a small amount of training data.

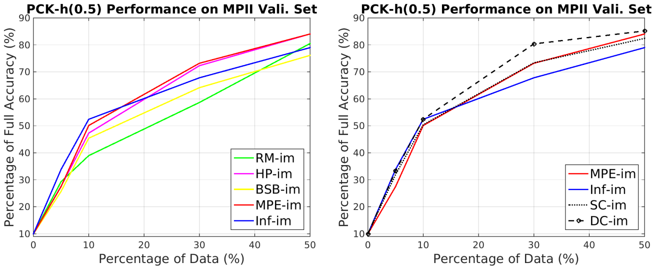
Evaluation Metric. To compare with published results, we use the widely accepted PCK-h [47, 1] metric, where a joint is considered correctly localized if the distance between the predicted and the true location is within a certain fraction of the head diameter.

4.2. Individual Active Selection Cues

Since no prior work does active learning for human pose estimation, we explore several informative individual cues and adapt these cues proposed in other active learning domains [37, 27] to our task. The following section describes how various cues measure the informativeness of images and joints at the t-th active learning iteration. After this, we can rank all components and select $\frac{k_t}{P}$ images or k_t joints from unlabelled set U_t.

- *Random (RM):* We randomly select images or joints.
- *Highest Probability (HP):* See Eq. (3) for joint-level measurement. We use the averaged score $\frac{1}{P}\sum_p C_{HP}(I_i, p)$ for image-level measurement.
- *Best v.s Second Best (BSB):* Due to the spatially smooth nature of predicted heat maps, directly comparing the highest and the second-highest value is meaningless (Fig. 3). Instead, we compute the difference between the highest and the second-highest peak in each joint's heat-map as the BSB score. We average joint-level BSB over all joints and take it as the image-level BSB score.
- *Influence (Inf):* We estimate the influential property of unlabelled images by Eq. (6). Note that *Inf* can only be applied at the image-level.
- *Multiple Peak Entropy (MPE):* See Eq. (5) for joint-level measurement. The image-level MPE measurement on I_i is defined as $\frac{1}{P}\sum_p C_{MPE}(I_i, p)$.

Fig. 6 shows the percentage of full accuracy as a function of percentage of annotation data, where $*$-im denotes the image-level active learning with $*$ cues. The full accuracy on the MPII validation set is obtained by applying CPM trained with all labelled images in MPII training set. We achieve 87.8% average accuracy on MPII validation set, which is comparable to 86.3% reported in [3]. Fig. 6(a) compares the performances of all individual cues. We can



(a) (b)

Figure 6. (a) shows the effectiveness of various individual cues. We can see that the proposed *MPE* is almost always the best among all uncertainty measurements. (b) compares the performances static and dynamic combination of top two individual cues on MPII validation set.

see that our *MPE* is almost always above other existing uncertainty measurements in this figure, which demonstrates the effectiveness of our proposed method. In comparison, *HP* performs better than *BSB* and can gradually achieve comparable performance to our *MPE* when labelling larger fractions of the data. *RM* is almost always the worst cue. Interestingly, *Inf* seems to be the best option at the early stage of active learning and becomes less competitive than uncertainty cues later on. Such phenomenon confirms our hypothesis that uncertainty and influence play different roles at different times in the active learning process.

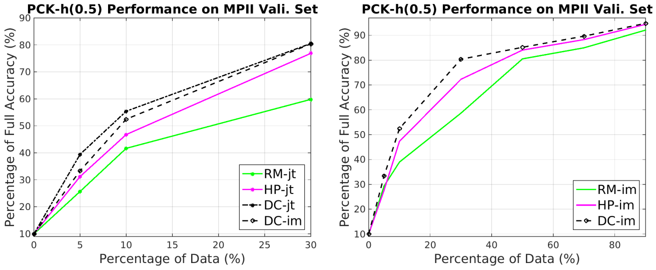
4.3. Active Selection with Multiple Cues

We also explore multiple cues in active learning for pose estimation task. In addition to our **Dynamic Combination** (*DC*) method, we also compare to a simpler **Static Combination** (*SC*). In *SC*, we simply fuse multiple cues by fixing $W(U_t)$ to 0.5 in Eq. (9) for all t. In these experiments, we combine *Inf* and *MPE* as these are the two best individual cues, and they are intuitively complementary.

We compare the performance of *DC* and *SC* on the MPII validation set and show the results on Fig. 6(b). This figure shows that *SC* is only as good as *MPE*, whereas *DC* is better than either of the two cues alone. This demonstrates the effectiveness of our proposed dynamic combination *DC* (with time-varying $W(U_t)$). *DC* effectively embeds the observation that influence cues (*Inf*) are more effective at early active learning iterations, where little data has been labelled, whereas uncertainty cues (*MPE*) become more reliable later on, when a good amount of labelled data is available.

Image-level v.s Joint-level Active Selection. Here we compare the results of labelling k_t joints and $\frac{k_t}{P}$ images in the active learning process. $*$-jt denotes the joint-level active learning with $*$ cues. We define the value of selecting the p-th joint in I_i as:

$$C_{DCJ}(I_i, p) = ((1 - W(U_t)) * C_{INF}(I_i) \\ + W(U_t) * C_{MPE}(I_i, p)) * (1 - I_i^p) \quad (10)$$



(a) (b)

Figure 7. (a) compares the performances of joint-level and image-level active learning. The proposed *DC* is always more effective than other baselines in both image and joint level. (b) illustrates the full performance on MPII validation set.

The results are illustrated in Fig. 7(a). We see that the proposed *DC* is always more effective than the baselines in both image and joint level. *DC*-jt saves half of the annotation effort in comparison to *RM*-im while achieving 40% of overall performance. More interestingly, *DC*-jt seems to perform better than *DC*-im when the labelling budget is low.

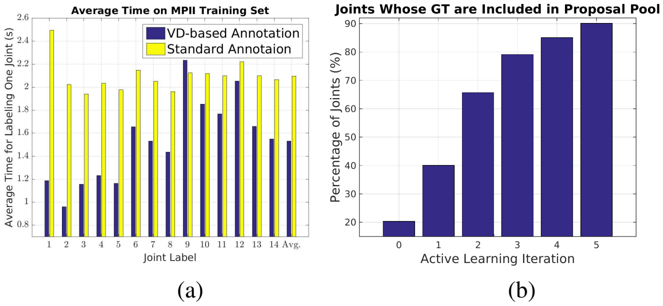
Using up to 90% of all training data. We show the full results of the active learning process in Fig. 7(b). Here we continue the curves until 90% of all initially unlabelled data has been labelled. We compare our proposed *DC* to *HP* and *RM* and show that it always outperforms these two active learning criteria. With 30% annotation budget, the performance of *DC* is 8% and 19% better than that of *HP* and *RM*, respectively. Our *DC* is still marginally better than *HP* and *RM* with 90% annotation data.

4.4. Using Real Annotators

We investigate here how robust our method is to noise introduced by using real human annotators. To this end, we use the active learning criterion that performed best in our image-level simulations (*DC*-im, Fig. 6 and 7) and run it for two iterations (5% and 10% of MPII training images). We ask 7 real human annotators to click on joint locations in images selected by our active learning process and use their responses to re-train the pose estimator. This leads to 25.4% and 46.2% of the full PCK-h accuracy with 5% and 10% annotations, respectively. This is comparable to what achieved by the simulations.

4.5. Starting from a Strong Initialization

To explore the model performance when initialized with a strong pose estimator, we conduct here an experiment starting from a CPM model trained on the full LSP [19] training set (11000 images, step (1) of the protocol in Sec. 4.1). We then apply our proposed *DC*-im active learning method on the MPII training set to gradually add more annotations (step (2) of the protocol). We evaluate pose es-



(a) (b)

Figure 8. (a) average time consumption with standard and VD-based annotation interface, for each joint and their average; (b) percentage of joints whose local peaks include the correct joint location, at each active learning iteration.

timator performance on the LSP test set (step (3) of the protocol).

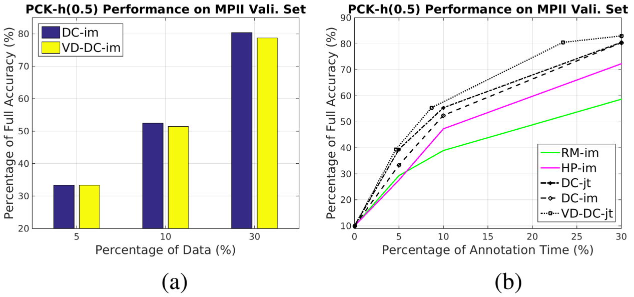
The initial model trained on 100% LSP training images gives 84.3% accuracy. Our active learner achieves 86.7% and 88.9% accuracy, when adding 10% and 30% of the MPII training set, respectively. This shows that our method can further improve the performance of a pose estimator even starting from a strong one. Moreover, using 100% LSP and 100% MPII training images would yield an upper bound of 90.5% accuracy [55]. This shows that our active learning criterion is cost-efficient: 74% of the performance improvement that could be gained by adding the full MPII training set is recovered by annotating only 30% of it.

4.6. Computer Assisted Annotation Interface

We compare here the annotation time between standard annotation and our proposed VD-based annotation interface. In both cases, we ask one annotator to label 400 images with the standard interface, and 400 images with the VD-based interface. In the standard interface, the annotator is asked to click on the exact pixel where the joint is located. Instead, our VD-based interface enables the annotator to click anywhere in a large region containing the joint (Sec. 3.3). Note how all experiments in this subsection use done with the standard protocol of Sec. 4.1, i.e. initializing the pose estimator with 100 LSP images (step (1)) and evaluating on the MPII validation set (step (3)).

Annotation Times. Annotation is performed by one student from our university. For both standard and VD-based interface, we created a full-screen interface. Fig. 8(a) reports the average time for annotating each joint with the two interfaces. We see that our VD-based interface can save about 33% annotation time compared with the standard method.

Candidates Quality Analysis. We measure the quality of VD-based annotation interface by measuring the percentage of requested joints whose ground-truth locations are among the candidates (Sec. 3.3). For each location candidate, we use the PCK-h(0.5) metric to determine whether it is sufficiently close to the ground-truth location to count. Fig. 8(b) shows that the percentage of joints whose ground-truth are included among the candidates grows with the active learning iteration. Hence, the more training data the pose estimator sees, the more our VD-based annotation interface can reduce the labelling time.



(a) (b)

Figure 9. (a) simulated results with VD-based annotation. (b) compares the performance of various active learning schemes as a function of percentage of annotation time. Here we show that combining our best scheme *DC-jt* with VD-based annotation can further reduce the annotation time.

Quality of Models Trained from VD-based Annotations. We also explore the performance of model trained with our VD-based annotation interface. We use *DC-im* as our active learning criterion and refer to the corresponding VD-based annotation interface as *VD-DC-im*. In this setting, we simulate the full VD-based annotation by replacing ground-truth annotation of requested joints with peak locations selected by users (Sec. 3.3). Our VD-trained model can achieve comparable performance (within 3% gap) to the original model trained from exact ground-truth joint locations, when using *DC-im* as our active criterion (Fig. 9(a)).

Overall Performance. We report the percentage of accuracy as a function of annotation time on Fig. 9(b). Combining our dynamic active selection scheme and VD-based annotation interface further improves efficiency, e.g., we can get 80% performance with 23% annotation time.

4.7. Conclusions

We took the first steps towards active learning for human pose estimation. Our method reduces the human annotation time both through an active selection scheme and through improvements in the annotation interface. We proposed an uncertainty measurement, Multiple Peak Entropy, which outperforms standard uncertainty baselines used in other active learning tasks. Moreover, we proposed an effective dynamic combination of influence and uncertainty cues. Finally, we introduced an efficient computer assisted annotation interface which reduces labelling time by one third without significant loss in accuracy.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1, 2, 5, 6

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

[3] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. *arXiv preprint arXiv:1605.02914*, 2016. 6

[4] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*, 2013. 3

[5] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 2, 3

[6] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, pages 1736–1744, 2014. 1, 2

[7] X. Chu, W. Ouyang, X. Wang, et al. Crf-cnn: Modeling structured information in human pose estimation. In *NIPS*, 2016. 2, 3

[8] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 1996. 1, 3, 5

[9] D. J. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV*, 2006. 1

[10] N. Dalal and B. Triggs. Histogram of Oriented Gradients for human detection. In *CVPR*, 2005.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Feifei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

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

[13] A. Freytag, E. Rodner, and J. Denzler. Selecting influential examples: Active learning with expected model output changes. In *ECCV*. Springer, 2014. 1, 3

[14] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *CVPR*, pages 3582–3589, 2014. 1, 2

[15] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Semisupervised svm batch mode active learning with applications to image retrieval. *ACM Transactions on Information Systems*, 27(3):16, 2009. 3

[16] J. E. Iglesias, E. Konukoglu, A. Montillo, Z. Tu, and A. Criminisi. Combining generative and discriminative models for semantic segmentation of ct scans via active learning. In *Biennial International Conference on Information Processing in Medical Imaging*, 2011. 1, 3

[17] S. D. Jain and K. Grauman. Active image segmentation propagation. In *CVPR*, 2016. 1, 3, 4, 5, 6

[18] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013. 6

[19] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2, 5, 6, 7

[20] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009. 1, 3

[21] C. Kading, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *CVPR*, 2015. 1

[22] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, 2007. 1, 3

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 6

[24] X. Lan and D. P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *ICCV*. IEEE, 2005. 2

[25] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *ICML*, 1994. 1, 3, 4

[26] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 1994. 1, 4

[27] W. Luo, A. Schwing, and R. Urtasun. Latent structured active learning. In *NIPS*, 2013. 1, 3, 6

[28] S. Maji, T. Hazan, and T. Jaakkola. Active boundary annotation using random map perturbations. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014. 3

[29] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*. Springer, 2016. 2, 3

[30] D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. In *CVPR*, 2016. 3

[31] G. Papandreou and A. L. Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *ICCV*. IEEE, 2011. 3

[32] A. Parkash and D. Parikh. Attributes for classifier feedback. In *ECCV*, 2012. 3

[33] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, 2013. 2

[34] B. L. Price, B. S. Morse, and S. Cohen. Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. In *ICCV*, 2009. 3

[35] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional active learning for image classification. In *CVPR*, 2008. 1, 3

[36] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*. Springer, 2014. 1, 2

[37] D. Roth and K. Small. Margin-based active learning for structured output spaces. In *ECML*. Springer, 2006. 1, 4, 6

[38] N. Roy and A. Mccallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001. 1, 3, 5

[39] O. Russakovsky, L.-J. Li, and L. Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *CVPR*, 2015. 3

[40] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *CVPR*, 2010. 2

[41] T. Scheffer, C. Decomain, and S. Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*. Springer, 2001. 3

[42] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *EMNLP*. Association for Computational Linguistics, 2008. 1, 4

[43] B. Siddiquie and A. Gupta. Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In *CVPR*, 2010. 3

[44] Q. Sun, A. Laddha, and D. Batra. Active learning for structured probabilistic models with histogram approximation. In *CVPR*, 2015. 1, 3, 6

[45] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *ECCV*. Springer, 2012. 2

[46] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015.

[47] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, pages 1799–1807, 2014. 1, 2, 3, 6

[48] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 2

[49] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012. 1, 3

[50] S. Vijayanarasimhan and K. Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *CVPR*, 2009. 3, 6

[51] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *IJCV*, 108(1-2):97–114, 2014. 1

[52] S. Vijayanarasimhan and A. Kapoor. Visual recognition and detection under bounded computational resources. In *CVPR*, 2010. 3

[53] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 2013. 3

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

[55] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 1, 2, 3, 6, 8

[56] J. Yang et al. Automatically labeling video data using multi-class active learning. In *ICCV*. IEEE, 2003. 3

[57] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pages 702–709, 2012. 1