Fine-Tuning Human Pose Estimations in Videos

Digvijay Singh[1] Vineeth Balasubramanian[2] C. V. Jawahar[1]
[1] KCIS, International Institute of Information Technology, Hyderabad, India
[2] Computer Science and Engineering Department, Indian Institute of Technology, Hyderabad, India
{digvijay.singh@research.,jawahar@}iiit.ac.in vineethnb@iith.ac.in

Abstract

*We propose a semi-supervised self-training method for fine-tuning human pose estimations in videos that provides accurate estimations even for complex sequences. We surpass state-of-the-art on most of the datasets used and also show a 2.33% gain over the baseline on our new dataset of unrestricted sports videos. The self-training model presented has two components: a static Pictorial Structure (*PS*) based model and a dynamic ensemble of exemplars. We present a pose quality criteria that is primarily used for batch selection and automatic parameter selection. The same criteria works as a low-level pose evaluator used in post-processing. We set a new challenge by introducing a full human body-parts annotated complex dataset,* CVIT-SPORTS*, which contains complex videos from the sports domain. The strength of our method is demonstrated by adapting to videos of complex activities such as cricket-bowling, cricket-batting, football as well as available standard datasets.*

1. Introduction

Over the past few years, we have seen inspiring advancements in different paradigms trying to solve the key problem of human pose estimation from image and video data. Two such major paradigms are based on Pictorial Structures (*e.g*. Yang and Ramanan [24]) and Deep Convolutional Networks (*e.g*. Tompson *et al*. [22]). These models perform well on generic images having considerably comprehensible human pose; however, in videos of complex activities such as sports, such methods turn out to be unreliable. Such models put a higher emphasis on pairwise connections among body-parts and regulate the configuration to follow a generic trend that may be violated in conditions such as playing sports. Videos are generally dealt with tracking strategies like optical flow, SIFT-flow because of possessing redundant temporal information, over a base model that functions at a frame level. However, the generic base models have been observed irregular to be able to produce efficient per-frame estimations that can later be harnessed by tracking strategies. The observed possible complications faced by single image models include occlusions, background cluttering, limb foreshortening, illumination variation *etc*. Moreover, the system settings become restricted when we rely on object tracking methods because of their limitation to track parts that move rapidly, change shape or reappear with a different outlook.

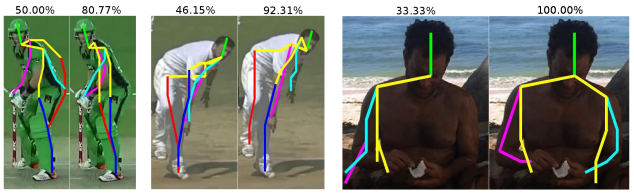


Figure 1. Three pairs of frames comparing Human-Pose estimations using, first column: Yang and Ramanan's [24] and second column: Our full model. First two images are from CVIT-Sports-Videos and third belongs to PIW [3] dataset. Percentage of correct keypoints [24] (PCK) score above each frame shows the improvement brought upon by our method over YR baseline.

This work focuses on improving individual frame estimations by using temporal information more observantly before handing it to a post-processing smoothness strategy. This results in a formulation that introduces resilience in terms of becoming more independent of part tracks, greater sequence lengths and pose complications. The strategy we are proposing utilizes semi-supervised self-training to fine-tune pose estimations in a video. The self-training model is composed of two components: i) PS-based model that is trained once initially and is used as it is for any arbitrary test video and ii) dynamic ensemble of exemplars model that is progressively augmented with newer examples in a phase-wise manner for each video. The general framework is identical to a PS-based framework in which exemplars assist in enforcing part-level appearances to generalize to newer videos. We will show that this amalgamation satiates the need of strengthening the appearance term which otherwise gets overridden by the pairwise score. Another intu-

ition behind our idea is that neighboring frames are likely to have similar poses and thus, an exemplar generated from a good estimation can help correct its temporal neighborhood. We use Yang and Ramanan's [24] (YR) model as the base (considering this as a model example of the part-based PS approach to pose estimation) because of its computational efficiency and reliability, and build on it iteratively . For its efficiency and ease of computation, we use Exemplar-SVM (E-SVM) proposed by [10] to synthesize exemplars from instances. The classifying boundary for an E-SVM is sufficiently taut to disparage far-off indexed instances but remains tolerant enough as to influence a temporal neighborhood in a more direct manner. Some corrections done by our combined model over YR baseline are shown in Figure 1. The eventual self-training framework is sensitive to the correctness of instances picked in each phase. For this purpose, we present a new pose quality ranking criteria that prunes estimations based on their geometric configurations satisfying certain criteria which are presumably valid for any frontal-human pose. Scores obtained from this criteria are also used for automatic parameter selection and in post-processing as will be shown later.

We present an extensive evaluation of our method on different datasets having arbitrary sequence lengths, as well as varying degrees of part motion and deformation. Poses in the Wild [3] (PIW) and VideoPose [20] (VP) are standard datasets having background clutter and self-occlusion; however, there is minimal camera motion, body deformation or rapid part movements which is present in most of the outdoor videos today. For this purpose, we introduce a new dataset called CVIT-SPORTS having pose configurations that can be called extreme. Quantitative results show that we surpass the state-of-the-art on most of the datasets and lead by a huge margin on CVIT-SPORTS-Videos dataset.

Related Work. Past attempts [11] [1] using exemplars to tackle human pose estimation rely on matching estimations with exemplar 2D configurations. Such methods fall short because of innumerable combinations of pose configurations and camera viewpoints. Relatively recent models based on deformable part models give reliable estimations in most real-life scenarios. After the initial work of Felzenszwalb and Huttenlocher [6], Yang and Ramanan [24] considered 'types' for each body part and performed efficient structured learning. Sapp *et al*. [19] presented a multimodal approach that captured a wider range of configurations. Dong *et al*. [4] proposed a unified framework for human parsing and pose estimation simultaneously. Pons-Moll *et al*. [15] estimated 3D pose by reducing information to boolean geometric relationships among body parts. Ye and Yang [25] estimated pose and shape by embedding the deformation model into a Gaussian Mixture Model for single-depth images. Initial attempts by Toshev and Szegedy [23] and Ouyang *et al*. [13] have parameterized DeepNet architectures for this problem. Later improvements by Tompson *et al*. [22] and Fan *et al*. [5] have successfully dealt with more complicated configurations.

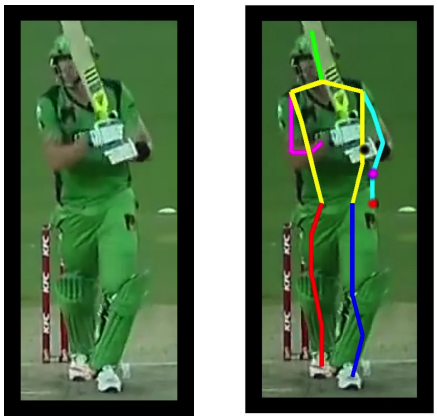
Most related to our work are methods that estimate poses in a video sequence. Ferrari *et al*. [7] use a spatio-temporal model in between consecutive frames and capture kinematic constraints within a frame. Rohrbach *et al*. [18] try to bring consistency among neighboring estimations based on SIFT-flow information. Ramakrishna *et al*. [16] have presented an occlusion-aware model that uses symmetric parts like shoulders and hands as supportive information. Another work by Yu *et al*. [26] infers 3D human poses from frames by spectral embedding and retrieving similar candidates from an exemplar database. Nie *et al*. [12] present a spatial-temporal And-Or model that simultaneously predicts pose estimation and action event. Recent work by Cherian *et al*. [3] mixes body part configurations across the sequence to find best fit detections in each frame with the underlying information provided by optical flow. Pfister *et al*. [14] used DeepNet architecture by incorporating optical flow information of each part at mid-layer followed by matching neighboring frames predictions across the sequence. Such methods rely heavily on their base model to produce good estimations and fail otherwise. Besides, using object tracking strategies impose limitations for tracking parts with deformation and high movement as can happen in long videos having considerable human and camera motion.

2. Fine-tuning Pose Estimation using Semi-Supervised Self-Training

Similar to [24], we denote I for a video frame, $p_i = (x, y)$ for the pixel location of part i and part type t_i for the mixture component of part i, where $i \in \{1, \cdots, K\}$, $p_i \in \{1, \cdots, L\}$ and $t_i \in \{1, \cdots, T\}$. $G = (V, E)$ is a K-node relational graph (in our case, a tree) The full score for a configuration of part types and positions in a video frame is then given by:

$$S(I, p, t) = S(t) + \sum_{i \in V} w_i^{t_i} \cdot \phi(I, p_i) + \sum_{ij \in E} w_{ij}^{t_i, t_j} \cdot \psi(p_i - p_j) \tag{1}$$

where $\phi(I, p_i)$ is a feature vector (such as HOG) and $\psi(p_i - p_j) = [dx \quad dx^2 \quad dy \quad dy^2]^T$. $dx = x_i - x_j$ and $dy = y_i - y_j$ are the relative positions on the pixel grid. $S(t) = \sum_{i \in V} b_i^{t_i} + \sum_{ij \in E} b_{ij}^{t_i, t_j}$ where b_is are priors on part types and b_{ij}s are parameters that weight co-occurrences of part types. Evidently, the first sum in Eqn 1 is an appearance model that computes the score of a template $w_i^{t_i}$ for part i, tuned for type t_i, at location p_i. The second sum is a spring model that influences the relative placement of p_i and p_j, parameterized by $w_{ij}^{t_i, t_j}$. Inference is carried out by maximizing $S(I, p, t)$ over p and t. Considering the graph G is a tree, this is achieved efficiently using message passing



(a) Original image (b) Skeleton detected using [24]

Figure 2. (Best viewed in color) The dominance of the pairwise potential over the unary causes poor performance of [24] on complex unseen poses. The red dot (C) denotes the detected wrist; the black dot denotes the ground-truth wrist (G); and the pink dot denotes the parent of the wrist (P).

[24].

While the aforementioned model showed good results in [24], generalizing this model to complex poses and long video sequences does not yield acceptable performance, often due to the dominance of the pairwise potential in the random field which is influenced by the configurations seen while training the model. An example of this issue is shown in Figure 2. This image illustrates the complex pose when there is a foreshortening of limbs, in this case of that from elbow to wrist. In [24]'s 26-joint human model, there are 3 joint positions from the elbow to the wrist, viz. elbow, mid-joint (between elbow-wrist) and wrist. The parent, P, in the figure corresponds to the mid-joint. Due to the foreshortening of the limb, the mid-joint will not be detected, and the pairwise potential forces the wrist to be placed at location C, instead of G, to ensure inter-part compatibility (observing the unary and pairwise potential values corroborates this claim too). We note that a similar observation of unary potentials being dominated by pairwise potentials at the boundaries was made by Horne *et al*. [8] recently , although their work was in a completely different context.

Our studies have shown that failure cases on complex human pose data, even when from the same domain, often occur because: (i) unary potentials may be predicted poorly due to the fact that the appearance terms defining a part may not strongly resemble the appearances of the same part as seen earlier, or (ii) the unary potential is subdued by the pairwise potential (of already seen configurations), when a new configuration of parts is encountered. Hence, in this work, we propose to fine-tune pose estimation for complex poses using a semi-supervised self-training approach that strengthens the unary score of parts with respect to the newer video without additional labeling effort. We achieve this objective using three steps: (i) a pose quality ranking SVM that identifies the quality of the pose detected by a base model on a given image; (ii) a coarse-to-fine strategy that uses the pose quality scores to identify suitable exemplars for self-training; and (iii) an ensemble of exemplar SVMs for semi-supervised self-training to improve the base model from [24]. While the proposed method has been studied with [24], the ideas behind each of the steps are independent of the method, and can easily be integrated with other part-based models for human pose estimation.

Pose Quality Ranking SVM: Given a trained base model, self-training refers to the process where the given model is trained iteratively on its own output on newer data. In a typical self-training setting, the model trained on a given dataset is applied on newer data, and the data with the most confident labeling from the new data is chosen with respective labels to retrain the original model. Identifying data with the most confident labels is, however, not straight-forward when dealing with human pose labels. Our initial experiments also showed that using the highest output scores of the base model does not directly translate to pose quality, especially when complex poses are present in the new video sequence (similar to the illustration in Figure 2).

Hence, in this work, we use the characteristics of the geometric configuration of the pose detected on a newer video by the base model to obtain a pose quality ranking, which can then be used to select suitable data instances for self-training. Table 1 shows the configuration characteristics of the detected pose that are used for this purpose. Items 1-18 are binary variables that indicate the presence of certain configurations that indicate errors in the pose (for instance, the "shoulder swap" binary variable indicates that the detected position of the right shoulder is to the left of the detected position of the left shoulder - which is typically an error, barring exceptional settings). Items 19-20 are real-valued variables that provide the pose scale, and the average half-angle between the parts. Characteristics which are suffixed with N in Table 1 indicate that these items are compared with a local temporal neighborhood of the video sequence for outlierness. If the values significantly differ from the local temporal neighborhood, the corresponding binary variable assumes value 1. More details are provided in Section 3.3 Figure 3 illustrates scenarios of improbable pose configurations, which motivate the use of the aforementioned characteristics to rank the detected pose quality.

Given the aforementioned characteristics, a linear Support Vector Regressor is trained to finally provide a pose quality ranking score. This score is used for selecting instances in each phase of the self-training pipeline, as described further below.

1. L-R shoulder swap	2. L-R hip swap
3. L-R torso parts intersctn.	4. Unlike L-R torso length
5/6. Converging torso width from top to bottom	7/8. L and R torso lengths (N)
9/10. L and R legs length (N)	11/12. L and R arms lengths (N)
13/14. L and R hip location (N)	15/16. L and R shoulder location (N)
17. L and R shoulder distance (N)	18. L-sho to neck to R-sho traversal distance(N)
19. Pose scale (N)	20. Half parts angle

Table 1. Our Pose Quality Ranking-SVM is trained with features designed using above criteria. Entries with N tag are compared with mean neighborhood entries of the same type. l-r denotes left-right.

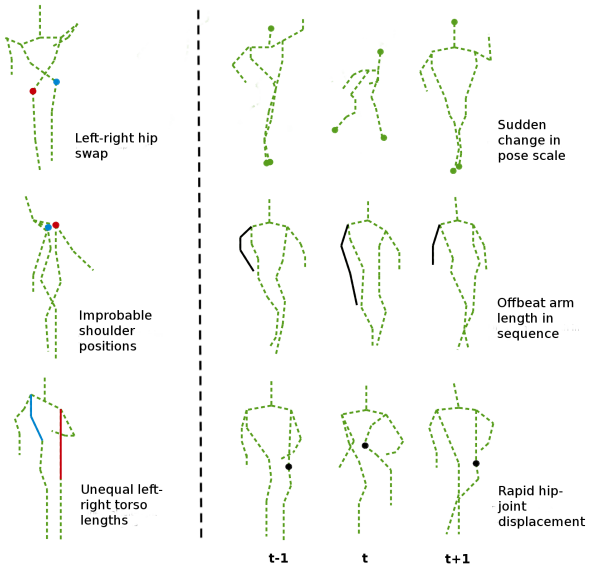


Figure 3. Examples of some characteristics of configuration of pose detected by base model used to rank quality of pose estimations in a video sequence. Left column shows improbable estimations at a single image level and right column shows eccentric behavior in a temporal locality of estimations.

Coarse-to-Fine Strategy for Exemplar Selection: Instead of directly using the obtained pose quality scores for selecting exemplars (video frames), we employ a coarse-to-fine strategy over the temporal resolution of the new video sequence to ensure representative data instances are picked from different temporal segments of the video sequence for self-training. This is achieved using a two-step process:

- In the first step, temporal neighborhoods with average pose quality scores that are greater than the mean pose quality scores are chosen. Among these high-ranked neighborhoods, the neighborhoods that have the highest mean output scores from the base method are selected for further processing. This provides a coarse selection of potential exemplars.

- In the second step, in each of the selected neighborhoods, the specific exemplars with pose quality scores greater than the mean pose quality scores of the neighborhood are selected, with a suitable threshold limiting the maximum number of exemplars that can be picked from one temporal neighborhood.

This coarse-to-fine strategy was primarily employed to address scenarios when the highest pose quality scores are all in the same temporal neighborhood of the entire video sequence, which if selected can lead to poor and biased self-training results. This simple multi-resolution approach significantly improved the self-training performance. We also note that where there are ties in the pose quality score, we use the output score of the base model to break the tie.

Ensemble of Exemplar SVMs for Semi-Supervised Self-Training: Given a set of exemplars that capture the best performance of the base model, we now describe how self-training is performed using this set. Self-training has been used with Conditional Random Fields (CRFs) successfully earlier [2][21] in natural language processing tasks, either by using the most confidently classified data from the new set in a SoftMax SVM formulation, or by simply using the probabilistic outputs. However, to the best of our knowledge, self-training with CRFs has not been attempted earlier where structural SVMs are used, which poses unique challenges as in the case of human pose estimation. Besides, many existing self-training methods tend to reinforce the knowledge of the base supervised model. Hence, in this work, to overcome these issues and to strengthen the unary scores in the base model, we use the identified exemplars to train an ensemble of exemplar SVMs [10] for each part individually. Hence, each selected exemplar-frame leads to the training of 26 exemplar SVMs corresponding to each of the parts.

As in [10], exemplar SVMs are based on the simple idea to train a classifier for each exemplar in the dataset, and then calibrating the output scores of these exemplar SVMs to obtain a final classifier. Each exemplar SVM, $(\mathbf{w}_E, b_E)$, tries to separate an exemplar $\mathbf{x}_E$ from a set of negative examples, N_E, by the largest possible margin. This is achieved by learning the parameters $(\mathbf{w}_E, b_E)$ that optimize the following convex objective:

$$\min_{(\mathbf{w}_E, b_E)} ||w||^2 + C_1 h(\mathbf{w}^T \mathbf{x}_E + b_E) + C_2 \sum_{x \in N_E} h(-\mathbf{w}^T \mathbf{x} - b)$$

where h stands for a loss function such as the hinge loss $h(x) = \max(0, 1 - x)$. This optimization is solved as in

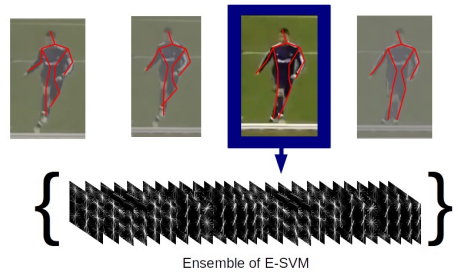
[10], for each of the 26 parts from the selected exemplar-frames in our case. Negative examples are randomly chosen from non-part windows from CVIT-SPORTS-Images set.

Instead of calibrating the resulting ensemble of exemplar SVMs (obtained from the parts of all the selected exemplar-frames) using the procedure in [10], we instead propose a different approach to integrate these exemplar SVMs into the CRF model of the base method [24]. In Equation 1 of the base model, we redefine the unary potential term using the exemplar SVMs as follows:

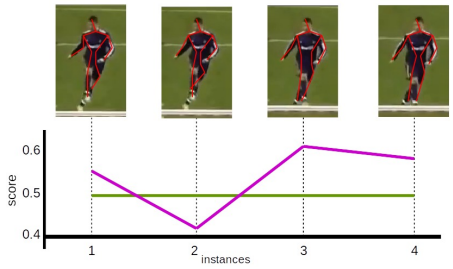
$$\sum_{i \in V} (\eta w_i^{t_i} + (1-\eta)\hat{w}_i^{t_i}) \cdot \phi(I, p_i) \qquad (2)$$

where $\hat{w}_i^{t_i}$ are the (normalized) weights learnt from the exemplar SVM for the i^{th} part tuned for type t_i, and η is a parameter that controls the weight given to the exemplars' contribution to the unary score. This unary score is then integrated into the CRF inference in [24] to obtain a new updated model that can be used for detecting poses on the (rest of the) newer video. This self-training approach provides us a seamless way not only to integrate the ensemble of exemplar SVMs into the base model, but also automatically addresses the issue of calibration in the ensemble. We note additionally that unlike other self-training methods which tend to reinforce the knowledge of the base model, the freedom in the choice of negative examples in this ensemble-of-exemplar-SVMs approach allows us to ensure newer knowledge is added to the previously learned model, rather than just reinforce old knowledge.

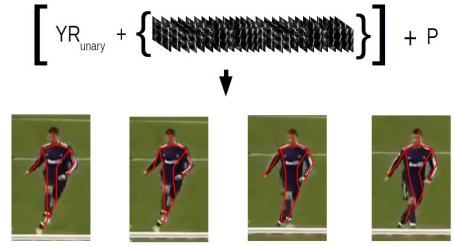
Given the aforementioned three steps of the proposed methodology to fine-tune pose estimation on newer video sequences, this entire process can be iterated in phases until a pre-defined performance level is achieved in terms of the pose quality scores or as meets subjective assessment. The overall idea of our methodology is summarized in Figure 4.



(a) Using pose quality score for batch selection, instance 3 is picked and is converted into an ensemble of E-SVM of length equal to number of body-joints.



(b) (Best viewed in color). Base model on consecutive frames with pose quality score (green) and base method's output score (red). Where there are ties among the pose quality score, the base method's output score is considered to break the tie.



(c) The procured E-SVMs are incorporated in our method equation 2 and used to infer on all frames which leads to correction in previously wrong localization.

Figure 4. Illustrative overview of proposed methodology

3. Experiments and Results

3.1. Datasets

The datasets used to train various models and for running different set of experiments in the paper.

Image-PARSE [17] This dataset has been used to train and test full-human parsing models. It has a total of 305 images containing people and is annotated with full-body pose. We use the dataset to train the base model YR-PARSE.

FLIC [19] For comparison experiments, we have used this human upper-body labeled dataset of around 4.5K samples to train another base model YR-FLIC.

VideoPose [20] This dataset consists of 17 test videos from TV shows with an average sequence length of 30 frames. The dataset is used to assess upper-body part estimations from different part localization algorithms.

Poses in the Wild [3] This recently introduced dataset consists of 30 video sequences of approximately 30 average sequence length trimmed from different Hollywood movies. The dataset has real-life scenes having some amount of camera and human motion and all the frames have been annotated with human upper-body joints. The main purpose of this dataset is to determine the localization accuracy of upper-body limb parts like shoulders, elbows and wrists.

CVIT-SPORTS Taking motivation from [9], a dataset containing still images from sports, we present an extremely challenging dataset of humans playing different sports. The

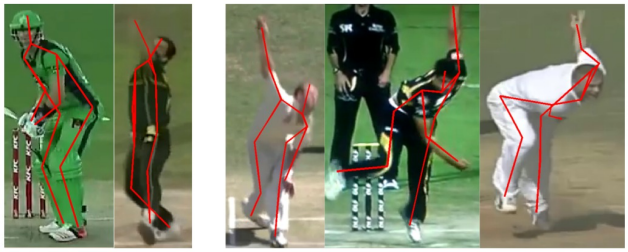


Figure 5. Sample frames with ground-truth annotations from CVIT-Sports-Videos dataset with complications such as (a) Half body self occlusion and (b) Extreme body deformation.

dataset comprises of two parts: CVIT-SPORTS-Videos and CVIT-SPORTS-Images. All the frames in the complete dataset have been annotated with 14 keypoints *i.e* full human pose. From this set, we generate a set of 26 keypoints which we have used all through in our experiments.

CVIT-SPORTS-Videos This set has a total of 11 videos of a human playing sports retrieved from YouTube. We have included intricate activities like cricket-bowling, cricket-batting and football-kicking. In total, this set has a total of 1446 frames averaging out to 131 frames per video.

CVIT-SPORTS-Images This set has 698 images, all belonging to cricket-frontal-bowling class. It has 150 random cricket-bowling-action images retrieved from Google. Rest of the images are random frames of cricket-bowling-runup and cricket-bowling-action from different videos. These images are used to get negative examples for training E-SVM in all the experiments. For both sets, in cases where most of the body parts are not obviously visible, we have annotated the frames to the best of our comprehensibility by looking at previous and future frames in the sequence. Some of these tangled cases are shown in Figure 5.

3.2. Evaluation Metric

We have used the Percentage of Correct Keypoints (PCK) metric proposed by [24] for evaluation of the proposed method. PCK considers only those estimated keypoints correct which lie within a pixel threshold distance of D_T from their counterpart in ground truth. The threshold distance is determined by the formula : $D_T = \beta.max(h, w)$, where h and w are the height and width of the tight bounding box created around the ground truth pose and β is a threshold controlling the relative correctness magnitude. β is set to 0.1 when dealing with full-body poses and to 0.2 when only half-body poses are considered.

3.3. Implementation Details

E-SVM as Exemplars. Each instance picked using our batch selection criteria is converted into a set of E-SVM filters. Since a full body is divided into 26 parts, we procure a set of 26 E-SVMs for every picked frame and for upper body setting, a set of 18 E-SVMs is obtained. E-SVM for a part is trained as a linear SVM with the part features as the only positive and 2000 negatives belonging to either background or non-part classes. Negative 2000 random samples for each part are picked out of many more possible candidates from CVIT-SPORTS-Images set.

Figure 6. YR baseline estimations on a set of frames from sequence V1. These detections are obtained in phase 1 of our pipeline.

Method	PCK(%)
YR-PARSE [24]	62.67
FT @ phase 2	65.97
FT-Full	67.83
FT-Full + PP	67.95

Table 2. Performance of our approach with different settings on example video V1 having 99 frames. Last row shows the post-processor (PP) improvement over the final phase of our full fine-tuning (FT-Full) model.

Iteration parameters and Stoppage criteria. In our coarse-to-fine exemplar selection strategy, the number of neighborhoods to be processed in each phase is dependent on a combination of pose quality score and method score where each neighborhood is restricted to a maximum size of 10 frames. Threshold value for both scores is taken to be the mean score of neighborhoods in consideration.

From chosen neighborhoods, a maximum of 3 instances are picked using pose quality score and are converted into E-SVM. η, a scalar defining the unary contribution from YR and E-SVM in Equation 2, is automatically picked from 11 possible values (ranging from 1 to 0 with 0.1 step size) using summation of pose quality scores for the complete video. From the observation that higher η value in phases ensures long-run consistency, we pick η which gives us the first local maxima. A video is processed iteratively until a threshold percentage of neighborhoods have been exhausted which by validation has been set to 60%.

Pose Quality Ranking SVM training. The SVM is trained similar to a linear incremental-SVM with the weights initialized from the domain knowledge of frontal-human-pose. Features as mentioned in Table 1 are obtained from a sam-

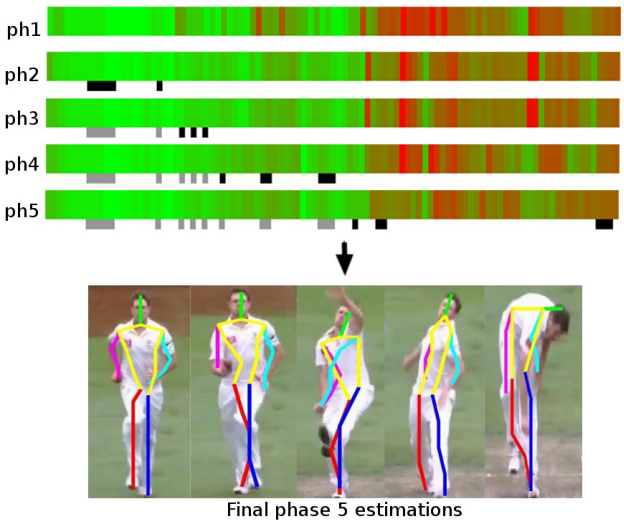


Figure 7. (Best viewed in color) Phase-wise PCK performance on sequence V1 showing the improvement trend as we iterate. Transition of color from green to red represents transition from correctness to false estimations. Indexes marked with black are instances converted into E-SVM and used for testing in the same phase, whereas gray-marked indexes are exemplar instances from previous phases.

ple video sequence and are used to train the desired SVM. For upper-body experiments, a separate SVM is trained by ignoring features that correspond to lower body-parts.

3.4. Baseline

We use the PS-based model proposed by [24] as our base model and the iterative procedure thrives on it. For full human-body experiments, the base model YR-PARSE is trained on 305 images from Image-PARSE [17] dataset. Number of human body parts considered are 26 and number of types for each part is taken to be 6. Similarly, for human upper-body experiments, YR-FLIC base model is trained on FLIC [19] dataset of around 4.5K labeled images. Types for each part is again 6, whereas total number of body parts are only 18.

We consider video sequence V1, having 99 total frames belonging to cricket-batting domain, as our primary example to show different experiment settings and their eventual impact. Table 2 shows YR baseline gives average PCK score of 62.67% on V1. Figure 6 shows YR estimations on frames picked from different time indexes and clearly stipulates its ineffectiveness to track body-parts when faced with occlusion, part foreshortening and other complications.

3.5. Fine-tuning using Self-training

After running the first phase using base model, we pick instances and synthesize E-SVM from them. Video length is sub-divided into N_t number of neighborhoods and using our coarse-to-fine selection strategy, we pick 3 instances from each neighborhood finally chosen. Generated E-SVMs are augmented in the ensemble and with the updated ensemble and YR base model, we run equation 2 across the whole sequence. Same procedure is followed for subsequent phases. The impact of iterations are more explicitly shown in Figure 7 and table 2 shows our 5.16% gain over the baseline on sequence V1 by our full fine-tuning (FT-Full) model.

Post-processor: Neighborhood Interpolation. After getting estimations from a phase decided by the stoppage criteria, we run a simple post-processor (PP) that uses information from pose quality scores. For an index, if its pose quality score is less than both of its neighboring indexes we choose the instance for post processing. Estimation of the picked instance is replaced by the mean interpolation of its neighboring estimations *i.e.* $est_i = mean(est_{i-1}, est_{i+1})$. Table 2 shows an improvement of 0.12% PCK on our sample video sequence V1 over our FT-Full model and a cumulative 5.28% PCK gain over the baseline.

Method	PCK(%)
YR-PARSE [24]	72.41
FT @ phase 2	74.31
FT-Full	74.19
FT-Full + PP	**74.74**

Table 3. Full Performance on CVIT-SPORTS-Videos dataset with variants. The pose setting considered is full 26 human body-joints.

Method	CVIT-SPORTS-Videos	PIW	VP
YR [24]	64.87	70.74	63.35
[3]	42.12	71.43	**71.95**
YR + FS [18]	31.75	48.04	60.01
FT @ phase 2	**65.20**	72.78	69.11
FT-Full	64.90	72.44	68.65
FT-Full + PP	64.50	**73.26**	68.96

Table 4. Comparison results: Performance comparison using PCK measure on CVIT-SPORTS-Videos, PIW [3] and VP [20]. The pose setting used includes 6 upper-body joints.

3.6. Quantitative and Qualitative Results

Diagnosing the Model. Table 3 shows PCK evaluation of our complete iterative method along-with variants on our CVIT-SPORTS-Videos set. The model being used estimates full 26 human body-joints in each frame. A performance improvement of 1.90% is achieved after running the first phase of combined model: YR and E-SVM. Full model shows net improvement of 1.78% and our post-processor (PP) stretches the lead to 2.33% over the YR baseline.

Comparisons. The method presented in this work is compared with recent well-performing methods tackling the problem of human pose estimation in videos on three

Figure 8. Top row shows the estimations with YR [24] and second row shows the corrections being done by our FT-Full model. First two columns are sequence of frames from CVIT-SPORTS-Videos dataset, whereas last column is from PIW [3] dataset.

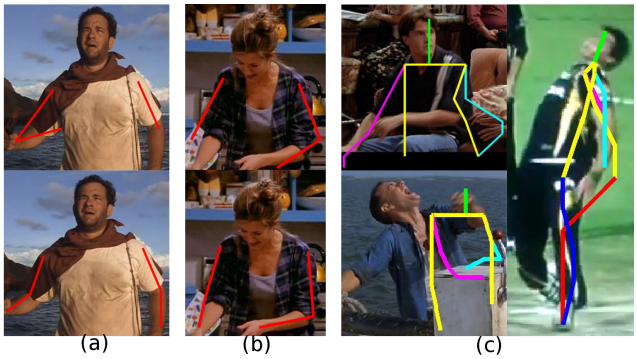


Figure 9. Comparison and Failure cases: Top showing estimations by [3] and bottom shows our FT-Full results on (a) PIW and (b) VP dataset. (c) Failure cases from three datasets used.

datasets: i. CVIT-SPORTS-Videos ii. PIW[3] and iii. VP[20]. Cherian *et al*. [3] uses optical flow to process single frame estimations and Rohrbach *et al*. [18] work on the same line but using SIFT-flow. We use the codes provided by the authors of both. The comparison evaluation is done on the upper-body joints as previous methods use this setting as shown in Figure 9 (a) and (b).

For comparisons on CVIT-SPORTS-Vidoes dataset, we have used the upper body-parts detections from our full model trained on PARSE [17] dataset. Table 4, first column demonstrates our lead over previous approaches on this extremely complicated dataset using PCK evaluation.

For standard datasets PIW [3] and VP [20], we have trained a new base model, YR-FLIC trained on FLIC [19] dataset similar to [3]. Table 4, second and third columns shows 2.02% and 5.76% improvements over the baseline by our method at phase 2 (FT @ phase 2) of the iterative procedure. Performance deteriorates when we run all the phases, although surpasses previous baseline for PIW dataset.

3.7. Discussion

Our self-training method iteratively strengthens positive unary responses across temporal sequence without any external manual intervention, which the baseline loses because of greater false pairwise influence. The whole iterative procedure is sensitive towards the quality of exemplars synthesized at each iteration making the task of determining good quality exemplars of prime significance.

Failure Cases. For shorter videos like in PIW and VP datasets, very few exemplars prominently influence the complete sequence. Failure in grabbing correct instances results in influencing neighborhoods inadequately as shown in Figure 9 (c).

Parameter Selection issues. Choosing optimal parameters in each phase is extremely relevant, otherwise the improvement across the iterations reduces to a minimal amount. We automatically determine the parameters using pose quality score in a manner determined by validation.

4. Conclusion

We have presented a self-training approach tackling the problem of Human Pose Estimation in videos which instead of directly relying on the baseline's predictions, empowers unary response enabling us to capture intricate pose configurations. We have also presented a pose quality criteria to pick instances in each iteration which also assists in automatic parameter selection and functions as a low-level pose evaluator. The setting thus obtained surpasses the previous state-of-the-art on most of the datasets used and leads by a huge margin on our introduced CVIT-SPORTS-Videos dataset.

References

[1] S. Carlsson and J. Sullivan. Action recognition by shape matching to key frames. In *IEEE Computer Society Workshop on Models versus Exemplars in Computer Vision*, 2001.

[2] M. Chen, J.-T. Sun, X. Ni, and Y. Chen. Improving context-aware query classification via adaptive self-training. In *CIKM*, 2011.

[3] A. Cherian, J. Mairal, K. Alahari, and C. Schmid. Mixing body-part sequences for human pose estimation. In *CVPR*, 2014.

[4] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014.

[5] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR*, 2015.

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

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

[8] L. Horne, J. Alvarez, M. Salzmann, and N. Barnes. Efficient scene parsing by sampling unary potentials in a fully-connected crf. In *IV*, 2015.

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

[10] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[11] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV*, 2002.

[12] B. X. Nie, C. Xiong, and S. Zhu. Joint action recognition and pose estimation from video. In *CVPR*, 2015.

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

[14] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. *CoRR*, abs/1506.02897, 2015.

[15] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for monocular human pose estimation. In *CVPR*, 2014.

[16] V. Ramakrishna, T. Kanade, and Y. Sheikh. Tracking human pose by tracking symmetric parts. In *CVPR*, 2013.

[17] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[18] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[19] B. Sapp and B. Taskar. MODEC: multimodal decomposable models for human pose estimation. In *CVPR*, 2013.

[20] B. Sapp, D. J. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011.

[21] A. Subramanya, S. Petrov, and F. Pereira. Efficient graph-based semi-supervised learning of structured tagging models. In *EMNLP*, 2010.

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

[23] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

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

[25] M. Ye and R. Yang. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *CVPR*, 2014.

[26] J. Yu, Y. Guo, D. Tao, and J. Wan. Human pose recovery by supervised spectral embedding. *Neurocomputing*, 2015.