Learning to Estimate Pose by Watching Videos

Prabuddha Chakraborty and Vinay P. Namboodiri
Department of Computer Science and Engineering
IIT Kanpur
{prabudc, vinaypn} @iitk.ac.in

Abstract

In this paper we propose a technique for obtaining coarse pose estimation of humans in an image that does not require any manual supervision. While a general unsupervised technique would fail to estimate human pose, we suggest that sufficient information about coarse pose can be obtained by observing human motion in multiple frames. Specifically, we consider obtaining surrogate supervision through videos as a means for obtaining motion based grouping cues. We supplement the method using a basic object detector that detects persons. With just these components we obtain a rough estimate of the human pose.

With these samples for training, we train a fully convolutional neural network (FCNN)[20] to obtain accurate dense blob based pose estimation. We show that the results obtained are close to the ground-truth and to the results obtained using a fully supervised convolutional pose estimation method [31] as evaluated on a challenging dataset [15]. This is further validated by evaluating the obtained poses using a pose based action recognition method [5]. In this setting we outperform the results as obtained using the baseline method that uses a fully supervised pose estimation algorithm and is competitive with a new baseline created using convolutional pose estimation with full supervision.

1. Introduction

Understanding human pose is a long standing requirement with interesting applications (gaming and other applications using Kinect, robotics, understanding pedestrian behavior, etc.). There has been strong progress over the years particularly using deep learning based pose estimation methods. However, progress is still required for accurate pose estimation in real world settings. One drawback faced is that the pose estimation methods require manual supervision with explicit labeling of the joint positions. This is particularly more for training state of the art deep learning systems. We address this requirement by proposing a method for obtaining automatic coarse human pose estimates. The method provides us with dense blob based pose estimates that suffices for most practical purposes (such as action recognition). Moreover it is obtained *without any manual supervision*. In fig. 1 we illustrate the dense pixel-wise estimates of body parts that are obtained from our method. We can clearly delineate separate regions such as head, neck, torso, knee area and legs as obtained by our method. The use of dense pixel-wise pose estimation allows our method to be robust to a wide variety of pose variations and problems such as occlusions and missing body parts. Further, these are obtained by using only motion cues for the various parts in videos.

The approach in this paper relies on self-supervision or surrogate supervision. Some approaches based on this rely on surrogate tasks such as re-assembling dislocated patches [7] or tracking people [30]. An interesting recent line of work that is related to this work relies on learning segmentation by using motion flows [21]. These surrogate tasks can be used for obtaining visual representations for generic tasks like classification or segmentation. Visual representations obtained through the techniques proposed so far however do not address granular tasks such as human pose estimation. Yet, we as humans can solve the problem easily. The primal cue that enables us in this task is observing the motion of the different body parts. This was evident early on and used by Gunnar Johansson in his seminal early work that analysed human body motion [16]. In this work Johansson observed that the relative motion between the body parts can be used for analysing human pose. Inspired by this insight we use the relative grouping of motion flow of humans to obtain the pose supervision required.

Our approach uses embarrassingly simple techniques that can be easily obtained in any setting for obtaining automatic supervision. These can always be improved upon. Our aim in using these techniques was to show that even the most basic grouping of human motion flow suffices to obtain the supervision required to be competitive to current state of the art techniques trained using carefully annotated supervised data. Interestingly, with enough data, the deep

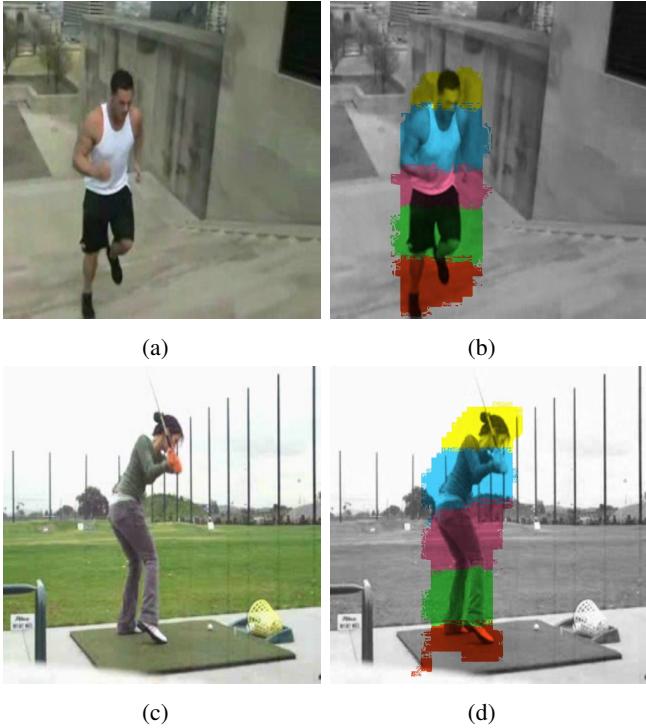|	
(a)	(b)
(c)	(d)

Figure 1: Illustration of pose estimation: Figures (a) and (c) show the original images and (b) and (d) show the respective pose estimates with the various colours depicting the different body parts estimated

network learns to generate output parts that are substantially better than the noisy supervision provided as input. The results are evaluated in terms of pose estimate comparisons as well as components of an action recognition method. The end-result is a competitive pose estimation method for free (zero supervision cost) by using easily available video data.

2. Related Work

There are two streams of work that are of relevance to the proposed work:

2.1. Pose Estimation

Human pose estimation has been solved by estimating a deformable mixture of body parts by Felzenszwalb and Huttenlocher [11]. This method provides a robust estimation of pose by using a spring and parts model allowing for deformation of human pose. The human body deformation is a significant challenge in pose estimation and this line of work allows for such deformation. This line of work has been successfully followed up by Andriluka *et al.* [3] and Eichner *et al.* [8]. Johnson and Everingham [18] consider a method that is able to learn from inaccurate annotation. In their work, the authors use clustered centroids obtained

by a larger dataset to obtain cluster specific priors for pose estimation. While, this approach is pertinent to our aim of working in the presence of noisy annotation, we are able to tolerate much larger inaccuracies than is considered in this work. Ladicky *et al.* [19] consider an interesting approach that combines pixel wise pose estimation with pictorial structures based pose estimation. In our work, we consider only pixel wise pose estimation. The advantage is that this pose estimation is more tolerant to occlusion of joints in various poses. We observe this phenomenon that pixel wise pose estimation similarly provides robustness towards occlusion and missing body parts. Ramakrishna *et al.* [24] in their work move beyond tree structured models by using inference machines that allow for richer interaction and better estimates of the parts by considering joint structured output prediction inference machines. While, similar in nature, we use recent advances in deep learning to avoid explicit structured representation learning by allowing fully convolutional networks to provide data dependent prediction. An related line of work is the seminal work by Shotton *et al.* [26] where the authors used synthetic renderings in order to estimate pose in depth images. This work however, is applicable to depth images and not to real-world color images.

Recently there have been a number of approaches that target solving the pose estimation problem in the deep learning framework [29, 31, 13]. An initial deep learning based approach was proposed by Jain *et al.* [13] where the authors considered a number of independent convolutional neural networks used for binary prediction of each body part. This binary classifier was used in a sliding window approach to generate a response map per body part. Subsequent work from Toshev and Szegedy [29] follow an interesting pose estimation approach that uses a cascade of deep regressors for pose estimation. At the first stage the architecture predicts the initial pose with the subsequent stages predicting finer pose in terms of displacement from the initial predicted pose. This approach of using sequential prediction is also adopted by Wei *et al.* [31] in their work that allows for sequential prediction in multiple stages with each stage operating on the belief map of the previous stage. In our work, we adopt the fully convolutional segmentation prediction framework [20] that is easier to train. Further, none of the methods so far considered could be trained without requiring manual supervision for training. As is well known by the community, each training set has its own bias and methods trained in one scenario would not work well in other scenarios due to a domain shift or dataset bias [28]. Our approach due to its ability to automatically generate supervision for training could always be applied in any novel scenario by just obtaining relevant data and obtaining automatic supervision through simple methods.
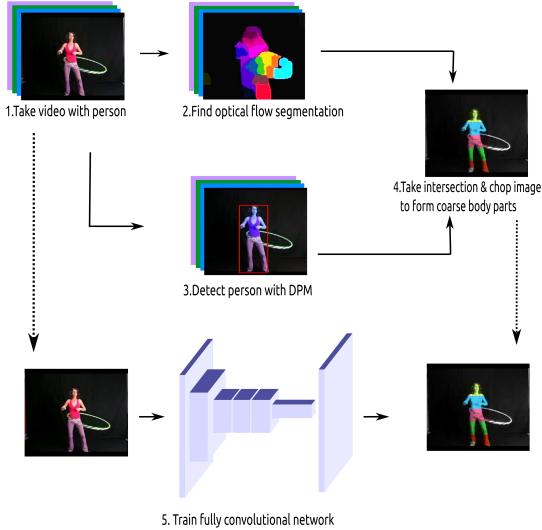
Figure 2: Illustration of the method

2.2. Self-supervision

There have been a number of works that are based on self-supervision or surrogate supervision. The initial methods were aimed at obtaining unsupervised means of generating visual representations that were competitive to supervised object classification task by performing other tasks for which supervision was directly obtainable such as context prediction [7] or ego-motion [14] or by tracking objects in videos [30]. Subsequently this concept has been explored for a wide range of tasks such as learning visual representations by using robotic motions [1, 23]. Further recent works include using the task of inpainting an image [22] or predicting the odd subsequence from a set of video sub-sequences [12]. The task of self-supervision for a semantically granular task such as human pose estimation has not yet been solved by the methods proposed so far. In the next section we provide details of the proposed method for obtaining self supervision for solving the problem of pose estimation.

3. Method

Our method is a simple sequence of steps that provides the coarse supervision necessary for pose estimation as illustrated in figure 2. We obtain the dataset in terms of videos with very little assumptions on the videos. We have evaluated using videos from two action recognition datasets for obtaining training data, viz. UPenn action recognition dataset [32] and UCF 101 action dataset [27]. We obtain optical flows between consecutive pairwise frames in a video from the videos in a dataset using Farneback's optical flow technique [9]. We use two thresholds on the flow, one to ensure that there is some motion in the frame (more than

10% pixels are having optical flow values above zero) and the other to ensure that the whole frame is not moving (less than 70% of the frame has optical flow values above zero). Using this motion flow we group the optical flow values into blobs using a simple mean shift based grouping technique [6]. This step yields blobs of motion flow that are grouped. We then need to ensure that the motion flow contains motion from a person and not some extraneous source such as motion of vehicles or other moving objects such as swings or animals. This is done by using a deformable part model based person detector [10]. We observed that the root filter predictions could be used to prune non-person blobs from person-blobs. These were noisy detections (as shown in section 4.6), however, as can be seen from the experimental section, these proved sufficient for obtaining reasonable training supervision.

From the sequence of steps above, we obtain a set of frames that have person detections and motion flow blobs. The intersection of these two steps is used to obtain a set of blobs for detected persons. In our method we use only the frames with a single person detection per frame as training data. This simplifying assumption allows us to avoid the problem of forming the association of motion flow blobs to multiple persons during training. The method learned is able to estimate a set of motion flow blob segments that belongs to a single person. Note that this does not limit our method and multiple persons pose estimation can be predicted during testing as is shown in figure 8(s). Having obtained the blobs for a person, we now have to obtain the part estimates. In our method, we divide the root filter horizontally into five parts that coarsely provides pose estimates corresponding to head, torso and arms, and legs. These are obtained by uniformly dividing the root filter detection bounding box into five equal horizontal blocks. The resulting bounding boxes result in coarse pose estimation that still corresponds rather accurately to the five parts as is verified experimentally. We evaluated various number of horizontal bands (discussed in section 4.5) and observed that five parts was providing us with an appropriate number of parts that was discriminative and representative of the human pose as required for recognizing actions.

Having obtained these coarse pose supervision, we train a fully convolutional neural network for segmentation [20] that we adapt for segmenting pose estimation blobs. The whole pipeline is illustrated in fig. 2 where we show how videos are used to obtain optical flow that is segmented using mean shift to provide motion blobs. Further the DPM based detector is used to provide person detections. The intersection of the motion blobs with the person detections provides us with estimates of the parts of a moving person. These are divided into five horizontal partitions resulting in five dense pixel-wise part estimates. These are then trained using a fully convolutional neural network (FCNN)

[20] to generate pixel-wise estimates of the five part segments. Each of the steps in our pipeline (except for the final segmentation prediction step) uses basic building blocks and can be improved upon. The main aim was to ensure that our method is not contingent on an advanced building block and even the simplest of building blocks suffices to obtain automatic supervision for pose estimation. In the next section we evaluate this basic approach thoroughly and compare it competitively with state of the art pose estimation techniques.

4. Experimental Evaluation

In this section we initially describe the experimental setup, followed by a quantitative evaluation of body pose estimates. We then use the pose estimate as a component for action recognition and provide a comparison. Next, we consider the effect of amount of data and number of parts followed by qualitatively considering the results for object localizatione and visualising the results for a number of samples.

4.1. Experimental Setup

Dataset For training the fully convolution neural network we have used videos from UCF-101 [27] and Penn Action Dataset [32].

Training We trained the network with a minibatch size of 10 using adam optimizer. For training the model with 40k images we used a learning rate 10^{-4}, beta1 0.9 , beta2 0.999 and no decay.

All our models are implemented with Keras having Theano backend using NVIDIA GeForce GTX TITAN X. Further details regarding the method is available in our public repository [1]

Hard Mining After we obtained the model trained with 20,000 images, we trained it further on 20,000 more images, sampled from 60,000 images, for which our model was inaccurate. We could thus reduce the number of images we needed to consider. This provided us with our final model that was trained with a total 40,000 images.

4.2. Body Pose Estimate comparison

We compare our proposed method for pose estimation against convolutional pose machine (CPM) [31] method that is the best model present trained using the Leeds dataset [17] and MPII pose dataset [2]. We obtain distance of the part locations from the ground-truth part locations in JHMDB dataset [15]. The exact part locations are obtained as centroids of the parts for our method whereas they are directly predicted using CPM [31]. As can be observed from the results presented in table 1, for various part locations the results are quite close to the ground-truth part locations

on average. The predictions are especially better as compared to CPM for part 5 that predicts the part around knees. As the distance from the torso increases it becomes harder to predict and so this part is a difficult part to reliably predict. The other parts such as the part around face and belly are also very close. The part around hips and shoulders are harder as they are not consistently obtained through our automatic annotation. The results for the automatic pose generation method is definitely much worse as compared to the output obtained after training. Note that our method is not trained on JHMDB, but only on UPenn and UCF datasets without using any pose ground-truth. The performance gap is clearly visible by considering the distances obtained in the second column against those obtained by our method in the third column. This is also evident in section 4.6 when we consider the object localisation results as the outputs obtained by the DPM detector [10] are qualitatively much worse as compared to the localisation we obtain. We further evaluate our method on a subset of MPII pose dataset with 17372 training images. For this we use the best CPM model not trained using MPII dataset as the training images are used and test it with our model trained on 40,000 images from UCF and Penn datasets. In this setting we observe that we are able to outperform CPM in most of the part estimates as shown in table 2

4.3. Pose estimation in Action Recognition

We next evaluate our method indirectly by considering its use in action recognition. We do this through an action recognition method that uses pose for recognizing action proposed by Cheron *et al.* [5]. Their method uses a supervised pose estimation method [4] that they had proposed earlier that especially handles mixed body poses. The actions are evaluated on the realistic JHMDB dataset [15]. We compare the action recognition accuracy by also considering the state-of-the-art CPM pose estimation method that is the best model present trained using the Leeds dataset and MPII pose dataset. This is not a fair comparison as our method is not trained with manual supervision. However, as can be observed from the results shown in table 3, we out-perform the supervised method of P-CNN [5] using mixed body pose estimates [4] even in this setting by around 2.2%. Small improvements can be obtained by varying the PCNN parameters improving the accuracy of our method to around 65.01%, but as this would not be the result of the pose estimation, but rather the recognition method, we do not consider such optimizations in the rest of the paper and report the original value obtained in the table 3. Thus, our method does not attain the accuracy of P-CNN with CPM features, however, we are close to their performance and the proposed method can be further improved by validating the pose estimation with P-CNN method parameters or finetuning on the JHMDB dataset. Such optimisations are not

---

Comparison of Pose estimation on JHMDB dataset [15]			
Part Name	Distance of CPM from ground truth [31]	Distance of Pose Supervision Generator-	Distance with our 40k Image Model trained on Penn [32] and UCF 101 [27]
Average Euclidean Distance 1 unit = 1 pixel			
Face - Part-1	38.93	58.11	40.46
Between Shoulders - Part-2	27.47	55.08	39.82
Belly - Part-3	55.10	68.60	55.76
Between Hips – Part-4	50.54	70.87	61.72
Between Knees – Part-5	87.11	88.45	77.38
Between Ankles – Part-5	112.09	116.54	92.0088

Table 1: In this table we provide a comparison of pose estimates with the ground-truth pose in JHMDB dataset. The CPM [31] model is trained with MPII [2] and LSP [17] datasets. Our method is trained with automatic annotation on other videos (not JHMDB) without manual supervision.

Comparison of Pose estimationon MPII dataset [2]		
Part Name	Distance of CPM [31]	Distance with our model
Average Euclidean Distance 1 unit = 1 pixel		
Part-1	214.05	209.55
Part-2	210.14	183.63
Part-3	285.14	245.16
Part-4	291.01	255.48
Part-5 (knees)	369.46	393.66
Part-5 (Ankles)	428.36	462.70

Table 2: In this table we provide a comparison of pose estimates with the ground-truth pose in MPII dataset [2]. This test is performed on 17372 training images from MPII training component. The CPM model used is trained on LSP[17]

Action recognition using P-CNN [5]	
A comparison with various pose estimation methods	
Method Name	Accuracy
Mixed body pose [4]	61.1%
CPM [31]	66.13%
Proposed Method	63.26%

Table 3: In this table we provide a comparison of various pose estimation methods evaluated through recognition of actions using the JHMDB dataset. The proposed method is still competitive even in the absence of ground truth training

currently considered in our method.

4.4. Varying amount of data

We next evaluate our method using the action recognition setting to analyse how the amount of data would affect the result. The results are illustrated in the graph shown in figure 3. As can be observed from the graph, the results consistently improved. The amount of data-samples used for training the fully convolutional neural network through automatic annotation is varied from 7000 samples to 40,000 samples. The addition of samples has aided the recognition and we were constrained only in terms of physical memory limitations in terms of the data-set with which we could train the system. Normally, any method is usually limited by amount of supervised training data available and this is not a constraint for our method. We can visualize this qualitatively in figure 4 by observing variation of the result in terms of extraction of all the parts jointly as we increase the data. As can be seen, as we increase the amount of data,

the full body extraction of the person is increasingly improved. This is reflected in the results as well as shown in the graph 3.
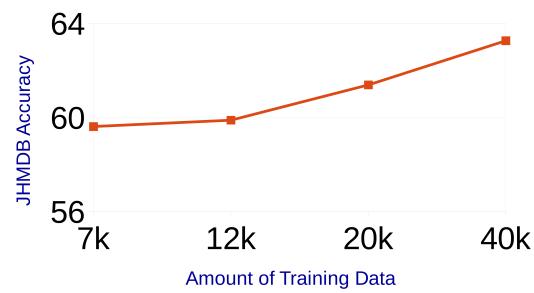

Figure 3: Difference in accuracy as the number of data samples is increased from 7000 to 40,000 samples. Increasing amount of data continuously increases the performance of the method.
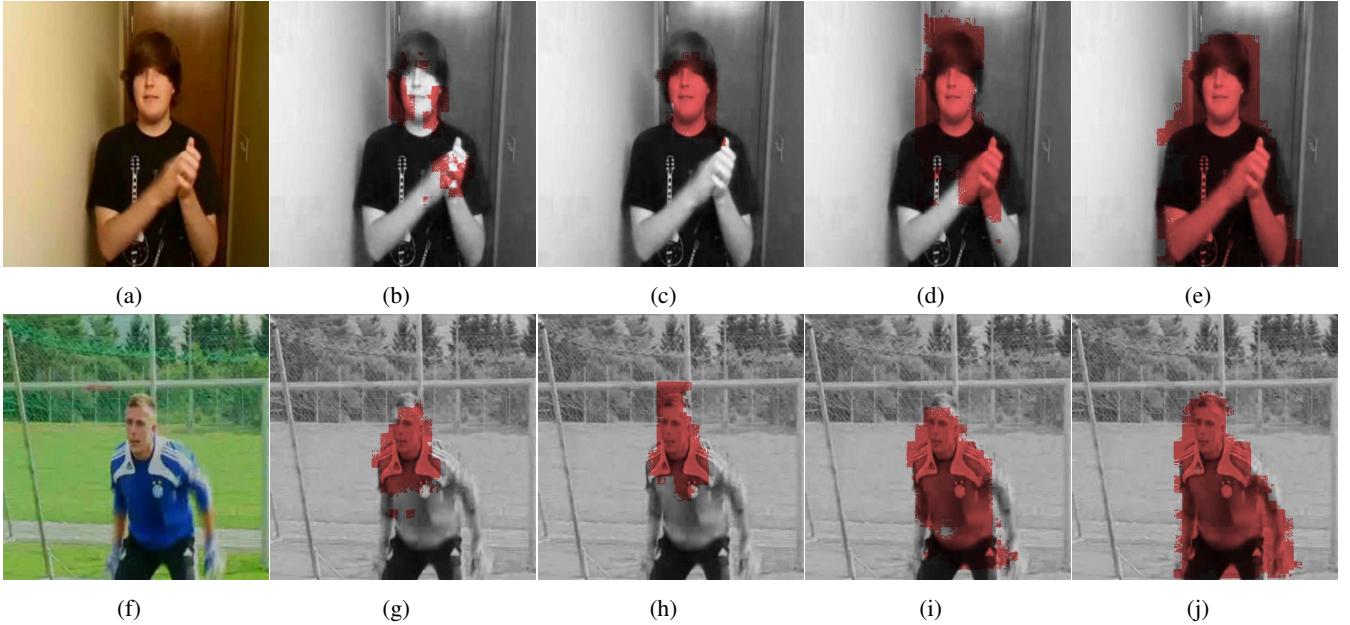
Figure 4: Difference in accuracy in full body estimation against the amount of data. The amount of data used for training is from left to right 7000 samples, 12000 samples, 20000 samples and 40000 samples.

4.5. Varying number of parts

We next analyse the effect of number of parts in our proposed method. We evaluate the effect of varying the number of parts for the task of action recognition on the JHMDB dataset. As can be observed from the graph 5 we obtained maximum accuracy using 5 parts. This experiment was carried out by fixing the number of samples to around 12000 samples and varying the number of parts. We can also observe this phenomenon visually in figure 6. Using a single part we observe in figure 6 that the pose estimation is attracted towards the golf club as a single part and does not detect the man or woman. With three parts, the pose estimation improves and we obtain three gross parts. This is further improved and tightly obtained when we use five parts. With seven parts, the individual part samples are not discriminative enough and are not reliably estimated. In figure 6(f) - (j) we consider the whole body being estimated by considering different number of parts as a slight mismatch in the individual parts may be tolerated. As can be seen the figure 6(i) the model with 5 parts provides us the best estimate of the person as a whole as compared to other varying number of parts. We therefore use five parts in our proposed method for all the remaining experiments.

4.6. Qualitative results and comparison

We now obtain the comparison of the proposed method qualitatively with the Faster RCNN [25] that was trained on Pascal VOC using ground-truth data and analyse the results from the proposed method qualitatively.
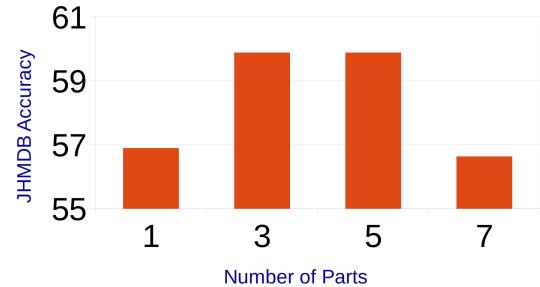

Figure 5: Difference in accuracy in action recognition task against number of parts

In figure 7 we provide a comparison of the proposed method qualitatively as a localisation method against fully supervised method of Faster RCNN [25] that is a benchmark method for object localisation and deformable part model (DPM) approach [10] that we use in our method as a means of person identification for various images. As can be seen from the figure, both the supervised object localisation methods fail to localise the person. This can be explained as the JHMDB dataset for action recognition [15] has a different distribution of objects and the persons in figures 7(e) and (i) are not in a usual upright pose. However, the proposed method succeeds in estimating the pose of the persons accurately though the proposed method has not seen a single image from the JHMDB dataset during training. This shows the efficacy of the method in being able to
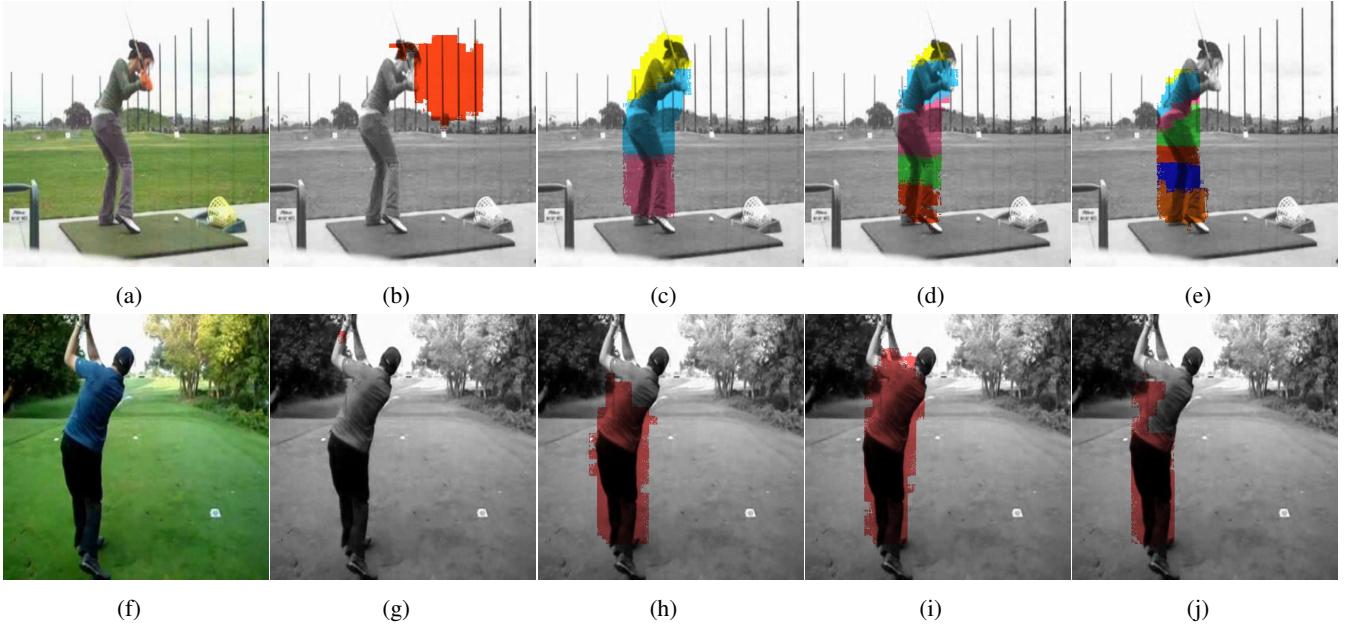
6

Figure 6: Difference in pose estimation using proposed method as obtained by varying the number of parts

localise persons accurately and even performing much better than the base method it was trained on.

As can be seen in figure 8 the method performs very well on varying kinds of data ranging from complex pose of a child pushing a table (figure 8(a)and(f)) and a baby sitting ((figure 8(e)and(j)) to that of persons playing in the field (figures 8(b),(c),(d) and(g)(h)(i)) to persons climbing stairs (figure 8(l)and(q)) or ladder of a ship in adverse lighting (in this result figure 8(k) was the original image and the result figure 8(p) is enhanced for visualisation). Similarly figure 8(o) shows a person walking in the street at night and we show the result in figure 8(t) with enhanced brightness to visualise the result. Interestingly figure 8(m) shows a person sitting that is also accurately estimated as shown in figure 8r. Further figure 8(n) shows the generalization of the method towards estimating the pose of two people that are quite accurately estimated as shown in figure 8(s). Thus as can be seen the proposed method is applicable for a variety of images and provides us with a rather good estimate of pixel-wise dense pose estimates, albeit with fewer detail in terms of the exact joint locations.

5. Conclusion

We have obtained through this paper a method that can be automatically trained using basic techniques to obtain pose estimation from a single image without requiring any manual supervision. This is possible by harvesting data regarding coarse pose through the relative motion of people in videos. This method can be easily applied in various scenarios and shows robust dense pixel-wise estimates of human body pose in challenging situations.

The limitation of the proposed method is in terms of being limited to only coarse blob based pose estimation. In future we would like to consider further advanced models such as hierarchical estimation of parts in order to obtain a more fine-grained pose for humans. To conclude, the performance of the proposed method without manual supervision is definitely encouraging and motivates the use of such self supervision for more tasks.
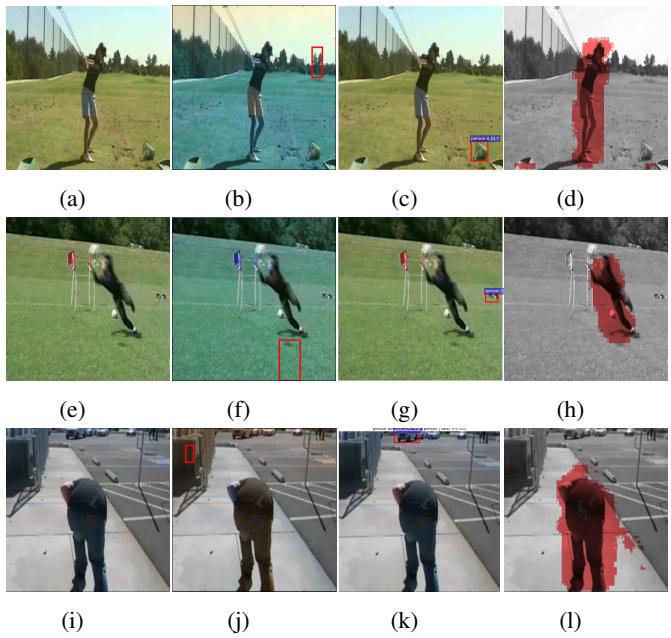
Figure 7: Figure provides comparison of results with respect to supervised object detectors for person localisation. Figures (a),(e) and (i) are the original images, (b),(f) and (j) are results using the DPM [10], (c),(g) and (k) are results from Faster-RCNN and (d),(h) and (l) are from the proposed method. As can be seen, the automatically supervised method provides much better results even in hard examples not detected by the supervised object detectors.

Figure 8: Illustration of results: Figures (a) - (e) and (k) - (o) show the original images and (f) - (j) and (p) - (t) show the respective pose estimates with the various colours depicting the different body parts estimated

References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[4] A. Cherian, J. Mairal, K. Alahari, and C. Schmid. Mixing body-part sequences for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] G. Chéron, I. Laptev, and C. Schmid. P-CNN: Pose-based CNN Features for Action Recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, May 2002.

[7] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*, 99:190–214, 2012.

[9] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, pages 363–370, 2003.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, Sept. 2010.

[11] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[12] B. Fernando, H. Bilen, E. Gavves, and S. Gould. Self-supervised video representation learning with odd-one-out networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler. Learning human pose estimation features with convolutional networks. In *International Conference on Learning Representations (ICLR)*, April 2014.

[14] D. Jayaraman and K. Grauman. Learning image representations tied to egomotion. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[15] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3192–3199, Dec. 2013.

[16] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973.

[17] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010.

[18] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[19] L. Ladicky, P. H. S. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR (to appear)*, Nov. 2015.

[21] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] L. Pinto, D. Gandhi, Y. Han, Y.-L. Park, and A. Gupta. The curious robot: Learning visual representations via physical interactions. In *European Conference on Computer Vision (ECCV)*, 2016.

[24] V. Ramakrishna, D. Munoz, M. Hebert , J. A. D. Bagnell, and Y. A. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *European Conference on Computer Vision (ECCV)*, July 2014.

[25] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[26] J. Shotton, A. Fitzgibbon, , A. Blake, A. Kipman, M. Finocchio, R. Moore, and T. Sharp. Real-time human pose recognition in parts from a single depth

image. In *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011.

[27] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. Technical report, CRCV, University of Central Florida, CRCV-TR-12-01, November., 2012.

[28] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011.

[29] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2014.

[30] X. Wang and A. Gupta. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[31] S.-E. Wei, V. Ramakrishna, and T. K. andYaser Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] W. Zhang, M. Zhu, and K. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.