https://doi.org/10.2352/ISSN.2470-1173.2018.10.IMAWM-453

Semantic Pose Machines

Ying-Kai Huang, Andreas Savakis; Rochester Institute of Technology; Rochester, New York; United States

Abstract

The objective of human pose estimation is to estimate the locations of keypoints on the human body using a single image. Convolutional pose machines is one of the most popular pose estimation techniques that is based on deep learning with convolutional features. In this paper, we propose semantic pose machines, a pose estimation technique that enhances convolutional pose machines by utilizing a semantic segmentation heatmap in addition to convolutional features. Semantic segmentation methods leverage the success of object class recog-nition networks for the segmentation of important object classes, including people. We consider the CRF as RNN semantic seg-mentation approach to obtain a heatmap that is incorporated in the pose estimation process as an additional channel. Our results on the LEEDS dataset indicate improvements over the convolutional pose machines method.

Introduction

Deep convolutional neural networks (CNNs) and their variants have become the architecture of choice for many computer vision applications. Deep networks have achieved state-of-the-art results in object class recognition [6], [13], [5], [10] face recognition [3], semantic segmentation [8], [11], pose estimation [12], and other appications. Pose estimation in images deals with the localization of body keypoints, such as head, neck, shoulders, elbows, wrists, hips and ankles. Deep learning techniques for pose estimation include DeepPose [16], Flowing ConvNets [15], Convolutional Pose Machines [12] and Stacked Hourglass Networks [1]. Due to the complexity and variety of human poses, human pose estimation stands to benefit from recent advances in semantic segmentation, such as Fully Convolutional semantic segmentation [8] and CRF as RNN [11]. The result of semantic segmentation is an image map where pixels are labeled with their corresponding class. The class that can be useful for human pose estimation is naturally the human class. In our work, for each stage of the Convolutional Pose Machine, the input is not just a RGB image and a center map as in the original design, but also the semantic segmentation feature heatmap from CRF as RNN [11].

Related Work

Pose Machine

The idea of pose machine was first introduced in [17] to deal with the articulated human pose estimation problem. By incorporating rich spatial interactions among multiple parts and information across parts of different scales, the pose machine was able to achieve state-of-the-art performance on challenging datasets. Conceptually, the pose machine is a sequential learning algorithm that outputs a confidence map for each part of human body, iteratively improving its estimates in each stage. In each stage, there are multiple classifiers trained to estimate the likelihood of a pixel belonging in the classes of interest (human parts). The output of all the classifiers becomes the input to the classifiers in the next stage. Since every classifier takes into consideration the estimation of all the classifiers in the previous stage, the spatial correlations are included into the input features of each stage. In terms of the classifier and image features used in this system, the choice of classifier was random forest since it outperformed other shallow methods on several datasets. The histogram of gradients (HOG) features, Lab color features, gradient magnitude, and context patch features were applied as the image features for the classifier. The LEEDS Sports Pose (LSP) [14] dataset and FLIC dataset were used for training and testing.

Convolutional Pose Machine

Based on the idea of the pose machine, the Convolutional Pose Machine (CPM) was developed in [12] with improvement on both accuracy and efficiency for the task. Inheriting the benefits of the pose machine architecture, the Convolutional Pose Machine combines the advantages of convolutional neural networks and pose machine. i.e. the implicit learning of long-range dependencies between image and multi-part cues and convolutional neural networks, and the ability to learn feature representations for both image and spatial context directly from data. Similarly to the pose machine, in each stage of a CPM, image features and the confidence maps produced by the previous stage are used as inputs in the current stage. Described in [12], larger receptive fields on belief maps help to learn long range spatial relationships and yield better accuracy which can be achieved either by increasing the size of the convolutional filters. Large filters come with the cost of having a larger number of parameters to learn. An alternative is increasing the depth of the network but this introduces the risk of encountering the vanishing gradient problem. A larger number of convolutional layers was chosen over larger filters in the CPM system because the sequential prediction framework of the pose machine provides a natural approach to training the deep network with intermediate supervision. This type of intermediate supervision between stages addresses the vanishing gradient problem. An example of the CPM system is shown in Figure 1. The CPM system was tested on various datasets and achieved state-of-the-art results.

Semantic Segmentation

As a type of pixel-level prediction task, semantic segmentation plays an important role in image understanding. The significant success of CNNs in solving high-level computer vision problems such as image recognition and object detection, has inspired new approaches for adapting CNNs to pixel-level labelling tasks such as semantic segmentation. In 2015, Fully Convolutional Networks (FCNs) [8] were proposed to deal with the semantic segmentation problem. FCNs adapted contemporary deep classification networks such as AlexNet [6], VGGnet [13], and GoogLeNet [2] to the segmentation task and achieved state-of-the-art performance in several benchmarks.

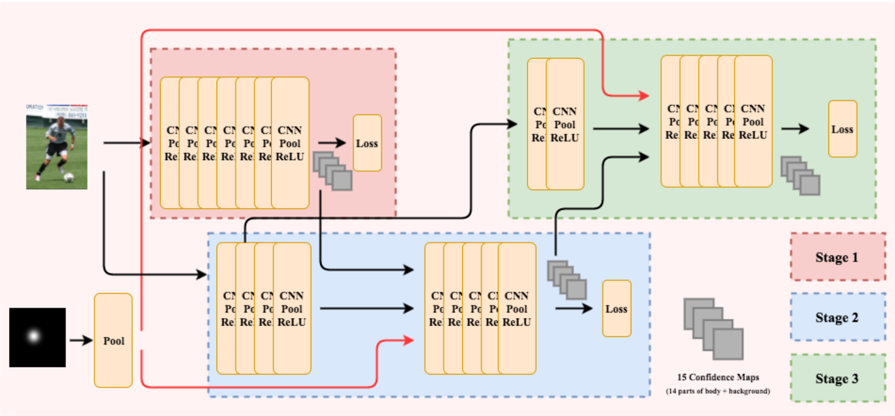

Figure 1: Convolutional Pose Machine architecture

However, one of the problems of using CNNs for low-level tasks is that successive convolution and pooling leads to non-sharp boundaries and blob-like shapes in the output feature map due to the lack of smoothness constraints. In other word, CNNs do not have a mechanism to encourage label agreement between pixels and features. In order to cope with this problem, Conditional Random Fields (CRFs) are widely used as post-processing procedures to smoothen and refine the final predictions. DeepLab [7] and Deconvolution Network [4] are some of the examples that attach CRFs at the end of CNNs to boost performance. DeepLab uses deep convolutional neural networks with their proposed atrous spatial pyramid pooling to robustly segment objects at multiple scales while the second one applies incorporates their proposed deconvolution layers and unpooling layers with the convolutional layers of VGG16 [13] and trained the network end-to-end.

Conditional Random Fields as Recurrent Neural Networks (CRF as RNN)

Unlike convolutional neural networks, recurrent neural networks (RNNs) have not gained as much popularity as CNNs due to their main disadvantage, which is short-term memory loss. However, the concept and framework of RNNs can be represented as mean-field iterations, an approximation for solving CRFs. Therefore, this allows them to be trained end-to-end with convolutional layers through backpropagation. The idea behind the probabilistic model in CRFs is to try to estimate the likelihood of a pixel being in a certain class given the surrounding predictions. the model takes the other predictions into consideration and finds the agreement between the predictions. Usually CRFs are used as a post-processing operation for fine-tuning the final output. In CRFs, in order to make a prediction on a pixel, the predictions of its neighborhood are needed. Likewise, recurrent neural networks are fed with the previous predictions. This type of similarity that is shared between CRF and RNN makes the transition possible from to the other. The main contribution of [11] is proposing a representation of the mean-field approximation in the convolutional neural network hierarchy. In [11], the convolutional neural network architecture is used to derive unary energy components for CRFs. The results show that the end-to-end trained FCN [8] with CRF as RNN model is better than simply using CRFs as a post-processing operation.

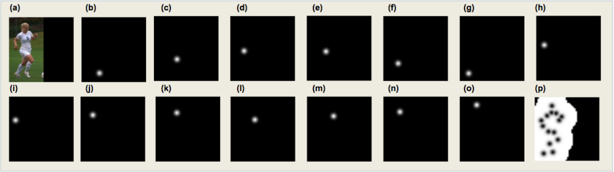

Figure 2: Groundtruth of every stage. (a) Input image, (b) Right ankle, (c) Right knee, (d) Right hip, (e) Left hip, (f) Left knee, (g) Left ankle, (h) Right wrist, (i) Right elbow, (j) Right shoulder, (k) Left shoulder, (l) Left elbow, (m) Left wrist, (n) Neck, (o) Head top, (p) Background.

Semantic Pose Machine

System Architecture

The architecture of the Semantic Pose Machine is based on extending the Convolutional Pose Machine features with sematnic segmentation maps. Figure 1 shows a 3-stage architecture of the Convolutional Pose Machine. The input to the first stage is just an RGB image followed by seven stacks of convolutional layers, max pooling layers, ReLU nonlinearity. This stage outputs 15 confidence maps of 46 x 46 in dimension, which correspond to 14 parts of the human body and a background. The second stage is similar to the first one, although instead of just taking the RGB image as input, it is also fed with the 15 confidence maps from the first stage and the max-pooled center map. The third and following stages have exactly the same structure. Each of these stages takes 32 feature maps extracted from the second stage, the max-pooled center map, and the output of the previous stage as input, and then outputs 15 confidence maps just like in the first and second stages. At the end of each stage a Euclidean loss layer is used for intermediate supervision to deal with the vanishing gradient problem. The groundtruth for each stage is the same, as illustrated in Figure 2. Since the output of the Convolutional Pose Machine includes 15 confidence maps, the final predictions for each part of human body are made by selecting the highest probability location in the corresponding belief map.

The main difference between the Semantic Pose Machine and the Convolutional Pose Machine is the input. Instead of feed-

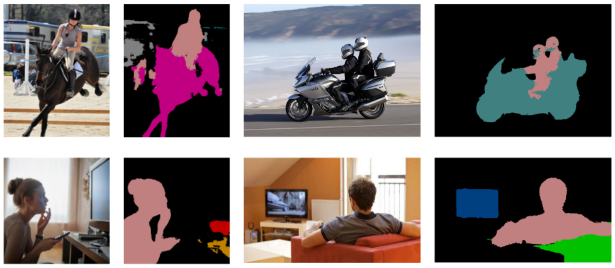

Figure 3: Semantic segmentation output of CRFasRNN.

ing the network the RGB image and a center map, the input of the Semantic Pose Machine also includes the output from the semantic segmentation heatmap based on CRFasRNN. The heatmap of CRFasRNN is a pixel-wise prediction in which each pixel is assigned a class as shown in Figure 3. However, by taking one step back and only focusing on the human class in the CRFasRNN output, we can retrieve the human belief map of the input image as shown in Figure 5. A human belief map shows the probability distribution of every pixel being human class which is the extra channel for the input of semantic pose machine. The design of the Semantic Pose Machine is shown in Figure 6. The center map is stacked on the input image as an extra channel, instead of sending it directly to the pooling layer, and the human belief heatmap is input as a separate channel, where the center map was placed in the CPM architecture. In addition, the number of stages for the Semantic Pose Machine is 3 because in the CMP paper [12], it was shown that the improvement on accuracy is not significant after the third stage and the additional system complexity would not be justified. We decided to add an extra channel in the semantic pose machine and not to simply replace the center map of the input with the human belief heatmapmap, to avoid confusing the classifier in case there are multiple people in the image.

Training Procedure

To assess the performance of the semantic pose machine in comparisonw ith teh CPM, We trained the Semantic Pose Machine and the Convolutional Pose Machine from scratch using the same data and without data augmentation or any fine tuning. Our experiments were designed to determine whether using the extra channel is helpful for estimating human pose. For training and testing, the Leeds Sports Dataset was used [14] which contains 2000 annotated images. Among the dataset, 1500 were used for training, 200 for validation and the last 300 for testing. The network hyper-parameters were set the same as in the convolutional pose machine. Both networks were trained for 36000 iterations with a batch-size of 4 and then a validation set was used to find the best model for final testing. As for the input center maps during the training phase required by the Convolutional Pose Machine and the Semantic Pose Machine, they are derived from taking the average of the neck, right hips, and left hips locations of the groundtruth. In terms of the input size, the network takes in 368x368 size of image and any images larger or smaller are resized according to the ratio of the larger side of the image to 368 and the rest the image is zero padded as shown in Figure 2(a) so as to preserve the ratio of the human shape.

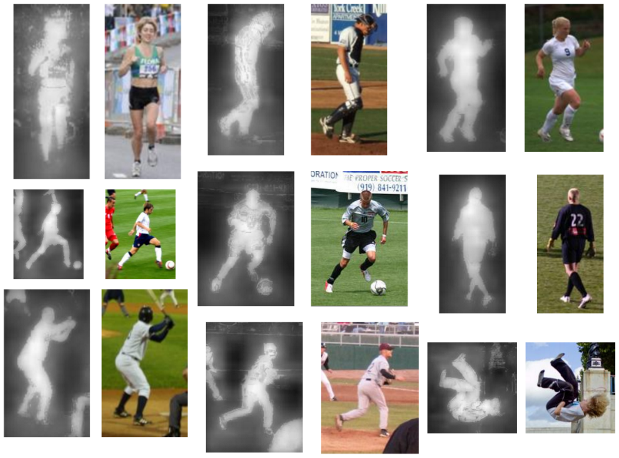

Figure 4: Human confidence maps derived from the output of CRFasRNN.

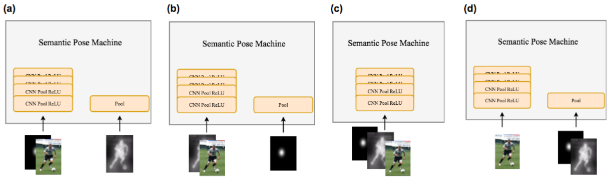

Figure 5: Human confidence maps derived from the output of CRFasRNN.

Experimental Results

The architecture of the Convolutional Pose Machine, illustated in Figure 1, shows that there are two inputs to the network, a center map and an RGB image. In the semantic pose machine architecture, the question arises of where the human confidence map should be. The original idea was to simply stack the human map with RGB image as the fourth channel, however, it could also go with the center map and then combined with the feature maps through concatenation. In order to find out which way of incorporating the segmentation heatmap results in the best performance, combinations of the placement of RGB image, center map, and human confidence map are tested. The experimental input variations are shown in Fig. 7. In the experiment, the two best performing placements of the human belief map are (a) and (b) in Figure 7 which are used to compare to the Convolutional Pose Machine in the following sections.

Evaluation

When it comes to evaluating the performance of human pose estimation, there are several different metrics, such as probability of correct pose (PCP), average precision of keypoint (APK), probability of correct keypoint (PCK), and PCKh [9] [18]. Among them, PCK is conventionally used for the Leeds Sports Dataset used in this paper. In PCK metric, a prediction is considered correct if it is within a radius from the groundtruth. The radius differs from image to image since it is derived from formula 1 which takes the size of the input image into account. The alpha from formula 1 acts as a threshold, the larger the value, the lower the

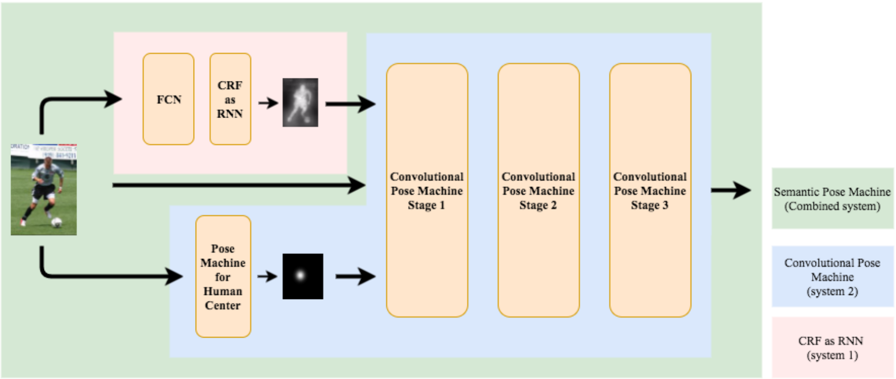
FCN
CRF as RNN
Pose Machine for Human Center
Convolutional Pose Machine Stage 1
Convolutional Pose Machine Stage 2
Convolutional Pose Machine Stage 3
Semantic Pose Machine (Combined system)
Convolutional Pose Machine (system 2)
CRF as RNN (system 1)

Figure 6: Semantic Pose Machine architecture.

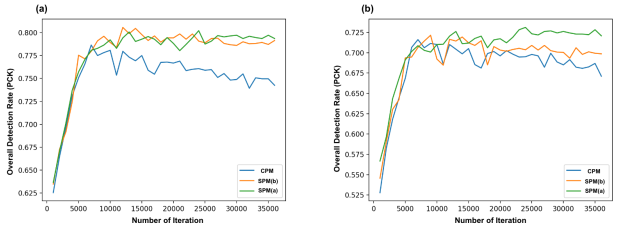

Figure 7: (a) Validation using groundtruth center maps. (b) Validation using predicated center maps

threshold. The alpha parameter is set to 0.1 for our evaluation.

$$\text{radius} = \alpha \times \text{Max}(\text{width}, \text{height}) \qquad (1)$$

Results and Comparison

The comparison between the Convolutional Pose Machine and the Semantic Pose Machine was done in two ways. The first one directly uses the groundtruth center map as part of the input, while the second way uses a model provided by the Convolutional Pose Machine to estimate the center of the human in an image. During the training, we assume that the center maps are given. In this paper, we decided to evaluate by using both the groundtruth and the prediction from the pre-trained model as input respectively.

The accuracies of the testing are shown in Tables 1 and 2. There is no surprise that the accuracies when using the groundtruth center maps are overall higher than those using predictions from the pre-trained model, because during the training the network is fed with the groundtruth center maps. As a result, the network has learned the parameters for that pattern and naturally has better performance if the input is consistent. Expectedly, the predictions derived from the model are not always accurate which will result in a lower accuracy. In either evaluation, the Semantic Pose Machine has higher overall accuracy than the Convolutional Pose Machine. The Semantic Pose Machine makes better predictions in most parts of human body. Furthermore, it is worth noticing that the Semantic Pose Machine does especially better at estimating the location of extremities, such as ankles and wrists, compared to the Convolutional Pose Machine. This is attributed to incorporating the results of CRFasRNN, which are quite sophisticated even at the limbs of human body. By including the human heatmaps from CRFasRNN as input features, the information of where to focus on the image is taken into account by the semantic pose machine. Such spatial features in the human heatmaps are extracted and learned through backpropagation and therefore produce a better performance than simple postprocessing.

The difference between the accuracies of the two implementations of the Semantic Pose Machine is not significant, although there are some differences. If the the groundtruth center maps are used as input during testing, stacking the center map with the RGB image while leaving the human confidence map for the pooling results in higher accuracy. On the other hand, if the center map inputs are derived from the pre-trained model, stacking the human heatmap with the RGB image gives out a better result. In the design of the Convolutional Pose Machine, the human center maps are used as an aid, but because it is not always accurate, it is placed separately from the RGB image to have no influence on the first stage feature extraction of the RGB image. However, if the input center map is guaranteed to be correct, it can be more than just a support and therefore it is more beneficial to be stacked on the RGB image, while using the human confidence heatmap as an assistance. This is shown in the implementation (a) in Figure 7. On the contrary, if the correctness of the center map is in doubt, then it should be used as an aid while stacking the human belief map with the RGB image just like the (b) implementation in Figure 7.

During the validation, it is observed that by including the extra channel, the semantic pose machine is less affected by the overfitting problem than it is in the Convolutional Pose Machine as shown in Figure 7. After around 7000 thousand iterations, the validation accuracy of the Convolutional Pose Machine starts to ramp down while the validation performance of the Semantic Pose Machine remains robust. Examples of final predictions on the test set from the Semantic Pose Machine and the Convolutional Pose Machine are shown in Figure 8. The performance improvements obtained by the semantic pose machine are illustrated in several examples.

Conclusions

In this work, we have demonsrtated that we can obtain improved pose estimation results by combining aspects of two state-of-the-art systems based on deep CNNs, namely convolutional pose machines and CRF as RNN semantic segmentation. Our

	Convolutional Pose Machine	Semantic Pose Machine (a)	Semantic Pose Machine (b)
Right ankle	35.33%	39.33%	40.33%
Right knee	46.66%	47.33%	49.66%
Right hip	68.66%	70.66%	71.00%
Left hip	65.66%	63.66%	65.33%
Left knee	47.00%	47.66%	48.66%
Left ankle	31.33%	39.33%	36.00%
Right wrist	50.66%	57.00%	55.33%
Right elbow	55.66%	59.00%	61.33%
Right shoulder	75.00%	79.00%	78.33%
Left shoulder	68.66%	69.66%	70.00%
Left elbow	52.33%	54.66%	53.00%
Left wrist	46.33%	50.00%	41.33%
Neck	84.33%	85.00%	83.00%
Head top	82.33%	84.00%	85.66%
Overall	57.88%	60.45%	59.92%

Table 1: Accuracies of the Convolutional Pose Machine, implementation (a) of Semantic Pose Machine, and implementation (b) of Semantic Pose Machine. Groundtruth center maps as input.

	Convolutional Pose Machine	Semantic Pose Machine (a)	Semantic Pose Machine (b)
Right ankle	32.66%	40.66%	38.33%
Right knee	42.66%	43%	44.66%
Right hip	53.66%	51.33%	51%
Left hip	47.33%	49.66%	50%
Left knee	40%	38.66%	44.66%
Left ankle	29.66%	36%	35%
Right wrist	48.33%	44.66%	52.33%
Right elbow	51.33%	55.33%	56%
Right shoulder	68.33%	70%	73%
Left shoulder	63%	67.66%	65%
Left elbow	49.66%	50.66%	48%
Left wrist	42.33%	44.33%	39.33%
Neck	80.33%	79%	81%
Head top	81.33%	82.66%	84.66%
Overall	52.19%	53.83%	54.5%

Table 2: Accuracies of the Convolutional Pose Machine, implementation (a) of Semantic Pose Machine, and implementation (b) of Semantic Pose Machine. Predicted center maps as input.

architecture incorporates the human segmentation heatmap as an extra channel for the pose estimation task. The human confidence maps help to more accurately estimate the joints of human body, especially at the extremities. Future work will further validate this approach by using a larger datasets and provide comparisons with other state-of-the-art systems in this field.

References

[1] K. Yang A. Newell and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499, 2016.

[2] Y. Jia-P. Sermanet S. Reed D. Anguelov D. Erhan V. Vanhoucke C. Szegedy, W. Liu and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[3] D. Kalenichenko F. Schroff and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[4] S. Hong H. Noh and B. Han. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[7] I. Kokkinos-K. Murphy L.C. Chen, G. Papandreou and A.L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In *https://arxiv.org/abs/1606.00915*, 2016.

[8] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[9] P. Gehler M. Andriluka, L. Pishchulin and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014.

[10] R. Girshick S. Ren, K. He and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[11] B. Romera-Paredes V. Vineet Z. Su D. Du C. Huang S. Zheng, S. Jayasumana and P. Torr. Conditional random fields as recurrent neural networks. In *IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.

[12] T. Kanade S.E. Wei, V. Ramakrishna and Y. Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[14] S.Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[15] J. Charles T. Pfister and A. Zisserman. Flowing convnets for human pose estimation in videos. In *CVPR*, 2015.

[16] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2016.

[17] M. Hebert-Martial V. Ramakrishna, D. Munoz, J.A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *European Conference on Computer Vision*, pages 33–47, 2014.

[18] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013.

Author Biography

Ying-Kai Huang received his BS in computer science from National Taiwan University of Science and Technology and MS in computer engineering from Rochester Institute of Technology (2017). His research focus has been in the field of computer vision and deep learning.

Andreas Savakis is Professor of Computer Engineering at Rochester Institute of Technology (RIT). He received the BS and MS degrees in Electrical Engineering from Old Dominion University and the PhD in Electrical and Computer Engineering with Mathematics Minor from North Carolina State University. His research interests include adaptive learning, object tracking, activity recognition, change detection, deep learning and computer vision applications. He has co-authored over 100 publications and holds 12 U.S. patents. He serves as Associate Editor of the Journal for Electronic Imaging and the IEEE Transactions on Circuits and Systems for Video Technology.

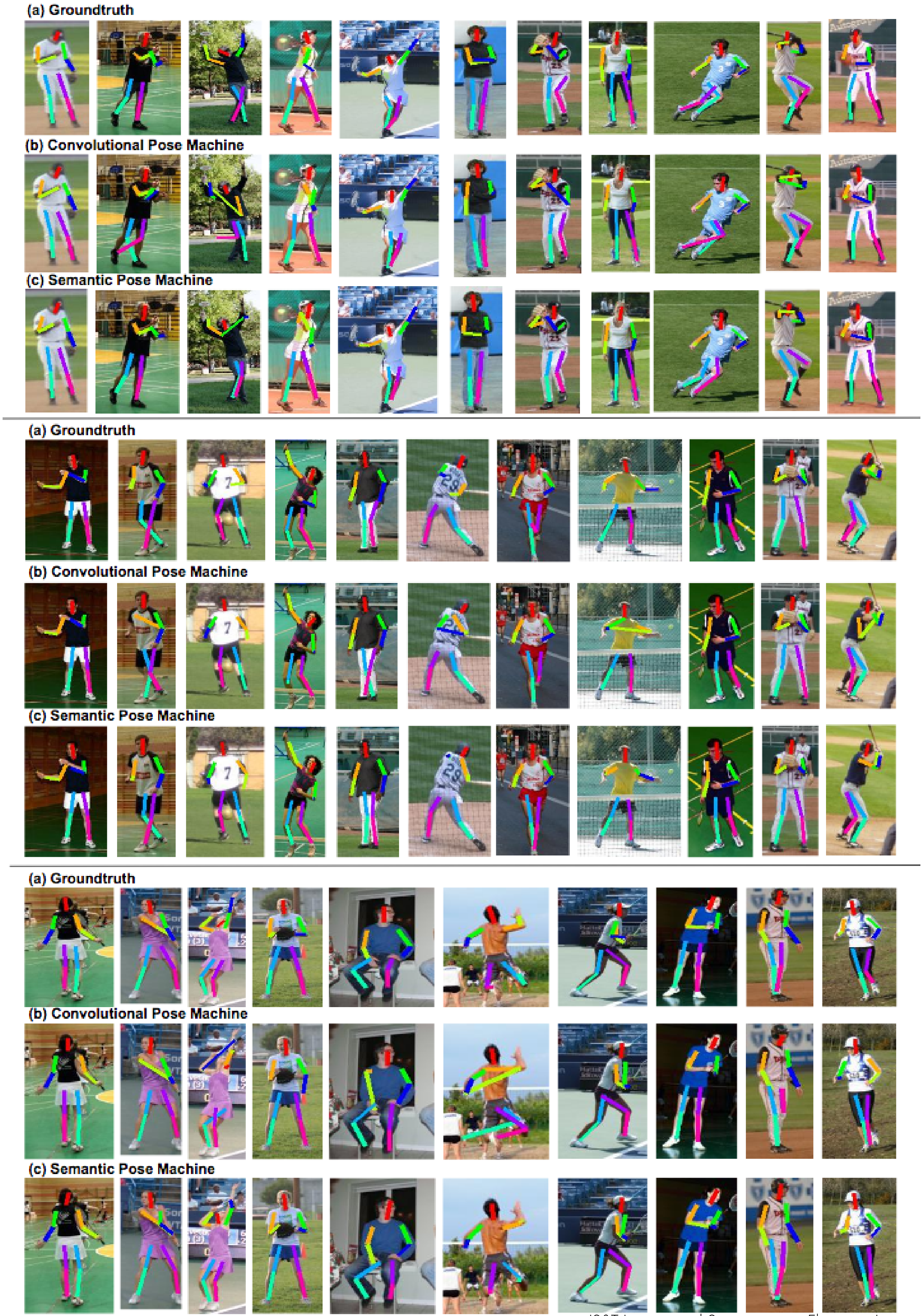
(a) Groundtruth
(b) Convolutional Pose Machine
(c) Semantic Pose Machine
(a) Groundtruth
(b) Convolutional Pose Machine
(c) Semantic Pose Machine
(a) Groundtruth
(b) Convolutional Pose Machine
(c) Semantic Pose Machine

Figure 8: (a) Groundtruth and predictions from the Convolutional Pose Machine and the Semantic Pose Machine.