Merging Pose Estimates Across Space and Time

Xavier P. Burgos-Artizzu[1]
xpburgos@caltech.edu
David Hall[1]
dhall@caltech.edu
Pietro Perona[1]
perona@caltech.edu
Piotr Dollár[2]
pdollar@microsoft.com

[1] California Institute of Technology
Pasadena, CA, USA
[2] Microsoft Research
Redmond, WA, USA

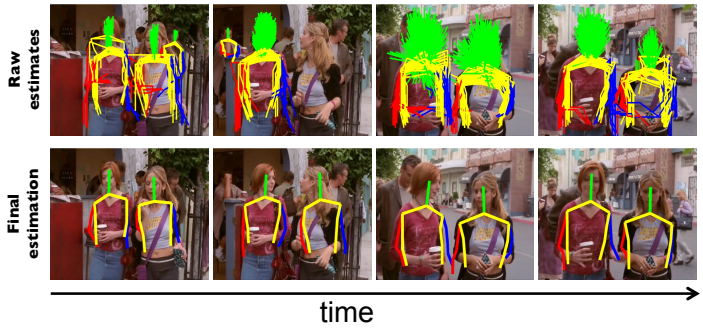

Figure 1: Our pose-NMS approach merges together pose estimates in space and time, outputting a final set of estimates more accurate than any of the individual ones. This is achieved by performing a robust average while simultaneously solving the correspondence problem.

Data driven approaches for pose estimation naturally output a set of pose hypothesis and rely on 'non-maximum suppression' (NMS) techniques to merge detections that are associated with the same objects. NMS is well developed for the case of object detection where the goal is to merge rigid object locations (bounding boxes) [5, 6]. However, it is still unclear how to extend it to flexible pose estimates. Applying standard NMS independently to each part location as in [7] fails to explicitly leverage the higher dimensionality of pose parameterization.

Our first contribution is a principled framework for merging multiple pose estimates in a single frame. This can be viewed as a generalization of NMS beyond bounding boxes. Our proposed approach makes minimal assumptions about the underlying method for pose estimation and generates a final set of pose estimates that are more accurate than any of the individual ones. We achieve this by performing a robust average while simultaneously solving the correspondence problem between pose estimates generated by multiple objects.

Our second contribution is to extend our approach to the multi-frame setting using the same mathematical framework, resulting in pose estimates that are further improved. While our approach is inspired by the recent success of 'tracking by detection' approaches [1, 2, 3, 4], we sidestep many of the inherent challenges associated with full tracking (e.g. objects entering and leaving a scene, extended periods of occlusion, etc.). Instead we present a principled, simple approach for merging multiple independent pose estimates across space and time and outputting both the number and pose of the objects present in a scene, see Fig. 1.

Pose-NMS is a versatile approach. It can be used to merge pose estimates in a single image or in an entire video, controlling the desired amount of temporal consistency. In scenarios where number of objects is fixed for long periods of time, (*e.g.* animals in a cage) it can be used to perform joint optimization over $K > 1$, improving joint reasoning. Pose-NMS can also be used to find all relevant trajectories when number of objects is variable.

We evaluate our approach on three different tasks: 1) Human body pose estimation, 2) Human face landmarks and 3) Animals. We collected 1,000 clips and manually annotated pose on the last frame of each clip, to measure how much pose estimation gets improved on frame T given previous $1 \leq t \leq T-1$ frames. Clip lengths vary from 1-10s.

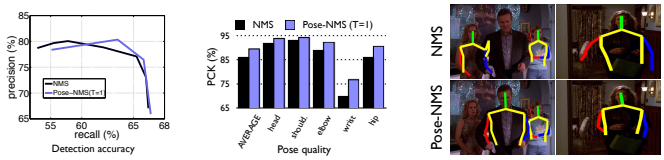

Figure 2: Results on (static) Buffy Stickmen dataset. Pose-NMS performs slightly better for detection but consistently improves the quality of pose estimates around 5%.

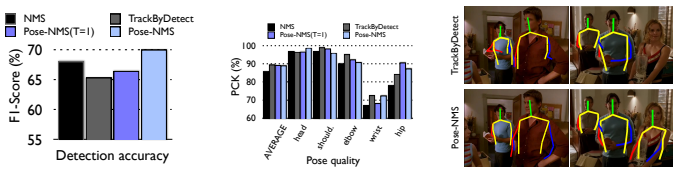

Figure 3: Results on the Video Buffy dataset. Single-frame Pose-NMS improves pose quality around 5% at similar detection accuracy compared with standard NMS. Full (multi-frame) Pose-NMS improves detection accuracy 6% while maintaining similar pose quality compared to running NMS prior to our tracking phase.

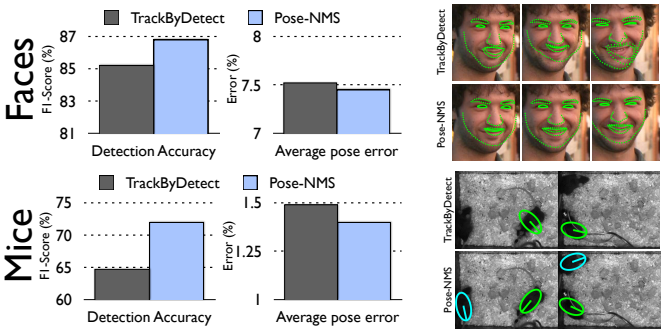

Figure 4: Results on Faces and Mice. Our approach improves detection by 3-7% and pose estimation quality between 2-7%. See text for details.

Single-frame Pose-NMS reaches 3% higher precision at similar recall rates compared to standard NMS, see Figure 2. More importantly, it consistently improves the quality of pose estimates by more than 5% on all body parts. This shows that Pose-NMS, unlike standard NMS, is capable of generating a final set of pose estimates that are more accurate than any of the original ones.

Full Multi-frame Pose-NMS improves detection accuracy between 3-7% and pose estimation quality between 1-7% when compared with running NMS prior to our multi-frame optimization (related to standard tracking-by-detection schemes), see Figures 3 and 4. Code can be downloaded from the authors' websites.

[1] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[2] K. Schindler B. Leibe and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, 2007.

[3] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *PAMI*, 33(8):1619–1632, 2011.

[4] D. A. Forsyth D. Ramanan, and and A. Zisserman. Tracking people by learning their appearance. *PAMI*, 29(1):65–91, 2007.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] C. Desai, D. Ramanan, and C.C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.

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