# Hierarchical Models for 3D Visual Inference

Atul Kanaujia[‡], Cristian Sminchisescu[†] and Dimitris Metaxas[‡]
[†]TTI-C, *crismin@nagoya.uchicago.edu, ttic.uchicago.edu/~crismin*
[‡]Rutgers University, *{kanaujia,dnm}@cs.rutgers.edu,*
*www.cs.rutgers.edu/~{kanaujia,dnm}*

Recent research in visual inference from monocular images has shown that discriminatively trained image-based predictors can provide fast, automatic qualitative 3d visual reconstructions (human poses, dominant scene ground planes or facades) in real scenes. However, the stability of existing image representations tends to be perturbed by deformations and misalignments in the training set, which degrades the quality of learning and generalization. Instead, we advocate the use of hierarchical image descriptions in order to better tolerate variability at multiple levels of detail. We combine multilevel encodings with improved stability to geometric transformations, with metric learning and semi-supervised manifold regularization methods in order to further profile them for task-invariance – resistance to clutter and 'within the same human pose class' differences. We analyze the effectiveness of both descriptors and metric learning methods and show that each one can contribute, often substantially, to better 3d human pose estimates in cluttered images.

Existing methods have successfully demonstrated that bag of features or regular-grid based representations of local descriptors (*e.g.* bag of shape context features, block of SIFT features [9, 1, 10] can be surprisingly effective at predicting 3d human poses, but the representations tend to be too inflexible for reconstruction in general scenes. It is more appropriate to view them as two useful extremes of a multilevel, hierarchical representation of images – a family of descriptors that progressively relaxes block-wise, rigid local spatial image encodings to increasingly weaker spatial models of position / geometry accumulated over increasingly larger image regions. Selecting the most competitive representation for an application (a typical set of people, motions, backgrounds or scales) reduces to either directly or implicitly learning a metric in the space of image descriptors, so that both good invariance and distinctiveness is achieved, *e.g.*, for 3d reconstruction – suppress noise by maximizing correlation within the desired pose invariance class, but keep different classes separated, and turn off components that are close to being statistical random for the task of prediction, disregarding the class. Our research relies on techniques from object recognition, metric learning and semi-supervised learning, as follows:

- We analyze hierarchical, coarse to fine multilevel image encodings, recently proposed for object recognition, for the *different task* of human pose prediction. This includes HMAX [7], spatial pyramids [4] and vocabulary trees [6]. These representations offer multiple levels of selectivity / invariance that can better tolerate deformation, misalignment and clutter in the training set.

- We study algorithms based on Canonical Correlation Analysis and Relevant Component Analysis for noise suppression and metric learning [2, 8] in order to refine and further align the image descriptors within individual pose invariance classes.

- We construct models based on both labeled and unlabeled data in order to make training with diverse, real-world data possible with existing pose prediction methods. We extend semi-supervised regression models to the more general case of learning multi-valued predictors.

We follow a manifold regularization approach [3] in order to construct smoothness priors that bias the model to give similar pose predictions for inputs close in the descriptor space intrinsic geometry (as represented, say, by the graph Laplacian of a training set).

The three components are strongly dependent in practice. To make unlabeled data useful for generalization, perceptually similar descriptors have to be close in the selected input metric. Learning an appropriate one becomes a necessary intermediate step.

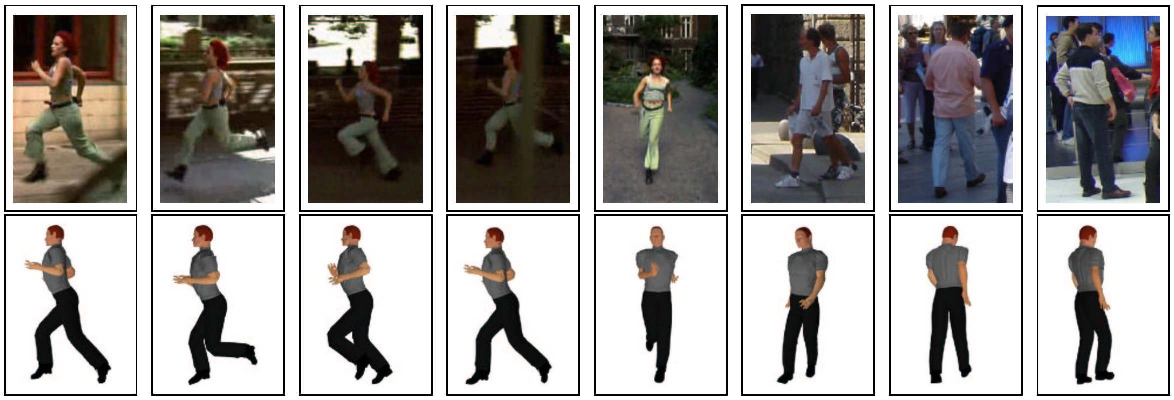

Figure 1: Qualitative 3d reconstruction results obtained on images from the movie ‘Run Lola Run’ (block of leftmost 5 images) and the INRIA Dalal pedestrian dataset (rightmost 3 images). *(a) Top row* shows the original images, *(b) Bottom row* shows automatic 3d reconstructions.

**Topic: visual processing and pattern recognition**
**Preference: oral/poster**

## References

[1] A. Agarwal and B. Triggs. A local basis representation for estimating human pose from cluttered images. In *ACCV*, 2006.

[2] A. Bar-hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning distance functions using equivalence relations. In *ICML*, 2003.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. In *AISTATS*, 2005.

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of IEEE*, 1998.

[6] D. Nistér and H. Stévenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[7] T. Serre, L Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, pages 994–1000, Washington, DC, USA, 2005.

[8] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[9] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *CVPR*, volume 1, pages 390–397, 2005.

[10] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning Joint Top-down and Bottom-up Processes for 3D Visual Inference. In *CVPR*, 2006.