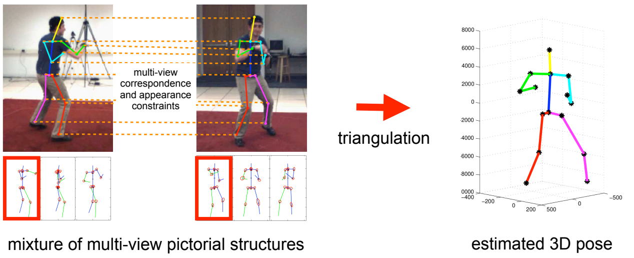
Multi-view Pictorial Structures for 3D Human Pose Estimation

Sikandar Amin[1] sikandar.amin@in.tum.de
Mykhaylo Andriluka[2] andriluka@mpi-inf.mpg.de
Marcus Rohrbach[2] rohrbach@mpi-inf.mpg.de
Bernt Schiele[2] schiele@mpi-inf.mpg.de

[1] Intelligent Autonomous Systems Group
Technische Universität München, Germany

[2] Computer Vision and Multimodal Computing
Max Planck Institute of Informatics
Saarbrücken, Germany



mixture of multi-view pictorial structures   estimated 3D pose

Figure 1: Overview of our approach to articulated 3D human pose estimation. Red boxes specify the selected component.

Pictorial structure models are the de facto standard for 2D human pose estimation. Numerous refinements and improvements have been proposed such as discriminatively trained body part detectors, flexible body models and local and global mixtures. While these techniques allow to achieve the state-of-the-art performance for 2D pose estimation, they have not yet been extended to enable pose estimation in 3D, instead this problem is traditionally addressed using 3D body models and involves complex inference in a high-dimensional space of 3D body configurations.

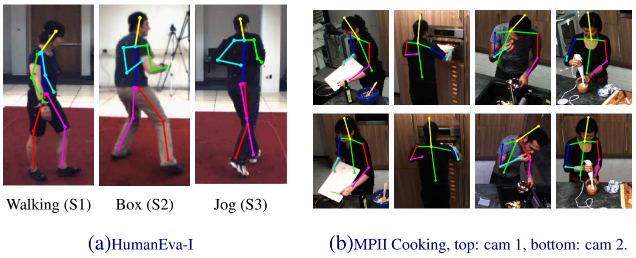
We formulate the articulated 3D human pose estimation problem as a joint inference over the set of 2D projections of the pose in each of the camera views. As a first contribution of this paper, we propose a 2D pose estimation approach that extends the state-of-the-art 2D pictorial structures model [6] with *flexible parts*, *color features*, *multi-modal pairwise terms*, and *mixtures of pictorial structures*. The second and main contribution is to extend this 2D pose estimation model to a multi-view model that performs joint reasoning over people poses seen from multiple viewpoints. The output of this novel model is then used to recover 3D pose.

We evaluate our multi-view pictorial structures model on HumanEva-I [8] and MPII Cooking [7] dataset. In comparison to related work for 3D pose estimation our approach achieves similar or better results while operating on single-frames only and not relying on activity specific motion models or tracking. Notably, our approach outperforms state-of-the-art for activities with more complex motions.

Single-view model: The pictorial structures model, originally introduced in [2, 3], represents the human body as a configuration $L = \{l_1, \ldots, l_N\}$ of N rigid parts and a set of pairwise part relationships E. The image position and absolute orientation of each part is given by $l_i = (x_i, y_i, \theta_i)$. We formulate the model as a conditional random field, and assume that the probability of the part configuration L given the image evidence I factorizes into a product of unary and pairwise terms:

$$p(L|I) = \frac{1}{Z} \prod_{n=1}^{N} f_n(l_n; I) \cdot \prod_{(i,j) \in E} f_{ij}(l_i, l_j), \qquad (1)$$

The part likelihood terms $f_n(l_n; I)$ are represented with boosted part detectors that rely on the encoding of the image using a densely computed grid of shape context descriptors [1]. We concatenate these shape context features with color features and learn a boosted part detector on top of this combined representation. Note that augumenting shape information with the color allows us to automatically learn the relative importance of both features at the part detection stage. The pairwise terms $f_{ij}(l_i, l_j)$ which encode the spatial constraints between parts are traditionally modeled with Gaussian distribution in the transformed space of the joint between two parts. We extend our model by introducing mixture models at the level of these pairwise part dependencies. To that end we replace the unimodal Gaussian with the term that maximizes over multiple modes and represent each mode with a Gaussian. Following [4, 5] we extend our approach to a mixture of pictorial structures models. We obtain the mixture components by clustering the training data with k-means and learning a separate model for each cluster. The components typically correspond to major modes in the data, such as various viewpoints of the person with respect to the camera. The index of the component is treated as a latent variable to be inferred at test time. We select the best component with the minimal uncertainty in the marginal posterior distributions of the body parts. In our experiments this approach worked slightly better compared to a trained holistic classifier that distinguishes the mixture component based on the contents of the person bounding box.



(a)HumanEva-I   (b)MPII Cooking, top: cam 1, bottom: cam 2.

Figure 2: Example 3D pose estimation results from our approach (projected to 2D).

Multi-view model: To exploit the multi-view information we augument the model with appearance and spatial correspondence constraints across views. In order to estimate the 3D pose we proceed in two steps. In the first step we jointly infer the 2D projections of the 3D body joints across views exploiting multi-view constraints. In the second step, we recover the 3D pose by triangulation of the estimated 2D projections. For simplicity, we describe our multi-view model for the case of two views. For view m, let us denote the 2D body configuration as L_m and image evidence as I_m. According to Eq. 1 the single-view factors $f(L_1; I_1)$ and $f(L_2; I_2)$ representing the conditional posterior over body configurations decomposes into a product of unary and pairwise terms that define appearance and spatial constraints between parts independently for each view. The joint posterior over configurations in both views is given by

$$p(L_1, L_2 | I_1, I_2) = \frac{1}{Z} f(L_1; I_1) f(L_2; I_2) \prod_n f_n^{app}(\mathbf{l}_n^1, \mathbf{l}_n^2; I_1, I_2) f_n^{cor}(\mathbf{l}_n^1, \mathbf{l}_n^2), \qquad (2)$$

The multi-view appearance factor f_n^{app} encodes the color and shape of the body part seen from multiple viewpoints. We define the joint appearance feature vector by concatenating the features from multiple views and train a boosted part detector using this representation. The multi-view correspondence factor f_n^{cor} encodes the constraint that part locations in each view should agree on the same 3D position. Given a pair of corresponding part locations $\mathbf{l}_n^1$ and $\mathbf{l}_n^2$ in each view and the projections of their reconstructed 3D point $\hat{\mathbf{l}}_n^1$ and $\hat{\mathbf{l}}_n^2$, we represent multi-view correspondence factor by $f_n^{cor}(\mathbf{l}_n^1, \mathbf{l}_n^2) = \exp(-(\|\mathbf{l}_n^1 - \hat{\mathbf{l}}_n^1\|^2 + \|\mathbf{l}_n^2 - \hat{\mathbf{l}}_n^2\|^2))$. When more than two views are available we connect the corresponding 2D body parts in all pairs of views. The posterior in Eq. 2 then includes multi-view appearance and correspondence factors for each pair of connected parts in all views as well as within-view spatial and appearance factors.

As for the single-view case, our multi-view model employs mixtures of pictorial structures, however, in this case the mixture components correspond to groups of poses consistent across views because they are learned from the projections of the same 3D poses.

[1] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *NIPS*, 2000.

[2] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[3] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67–92, 1973.

[4] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[5] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[6] Andriluka Mykhaylo, Roth Stefan, and Schiele Bernt. Discriminative appearance models for pictorial structures. *IJCV*, 2011.

[7] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[8] L. Sigal, A. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2), 2010.