Does Human Action Recognition Benefit from Pose Estimation?

Angela Yao[1]
yaoa@vision.ee.ethz.ch
Juergen Gall[1]
gall@vision.ee.ethz.ch
Gabriele Fanelli[1]
fanelli@vision.ee.ethz.ch
Luc Van Gool[1,2]
vangool@vision.ee.ethz.ch

[1] Computer Vision Laboratory,
ETH Zurich, Switzerland
[2] IBBT, ESAT-PSI
K.U. Leuven, Belgium

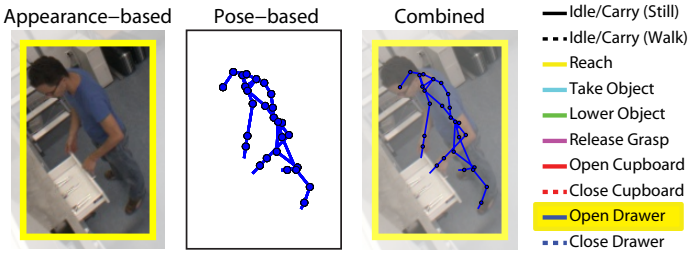

Figure 1: We address the question of whether it is useful to perform pose estimation for the task of action recognition by comparing the use of appearance-based features, pose-based features and combined appearance- and pose-based features.

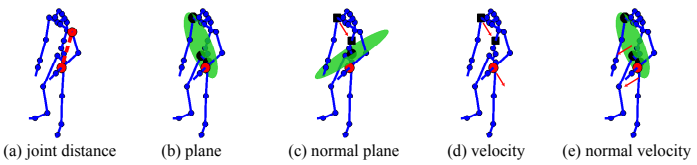

Figure 2: Pose-based features. *(a)*Euclidean distance between two joints. *(b)*Plane: distance between a joint and a plane spanned by three other joints. *(c)*Normal plane: same as plane feature, but the plane has a normal vector in the direction of two joints and is fixed at a third joint. *(d)*Velocity: velocity component of a joint in the direction of two other joints. *(e)*Normal velocity: velocity component of a joint in the direction of the normal vector of the oriented plane spanned by three other joints.

Introduction The earliest works in action recognition focused on tracking body parts and classifying the joint movements. These *pose-based approaches*, while straight-forward, require accurate tracking of body parts, which is a challenging task in its own right. As recent trends in action recognition have shifted towards natural and unconstrained videos (*e.g.* films, broadcast sports, Youtube videos), efforts have shifted from high-level modelling of the human body to directly classifying actions with abstract and low-level appearance features in *appearance-based approaches*. But despite requiring more initial processing, pose representations have several advantages. First, they have fewer intra-class variances; in particular, 3D skeleton poses are viewpoint and appearance invariant, such that actions vary less from actor to actor. Secondly, using pose representations simplifies learning for action recognition, since relevant high-level information has already been extracted. Given the great progress in pose estimation over the past few years [1], we feel that pose-based action recognition systems warrant a second look.

In this work, we compare pose-based and appearance-based features for action recognition as depicted in Fig. 1. Our pose-based features are derived from articulated 3D joint information; we label as appearance-based any feature which can be extracted from video data without explicit articulated modelling of the human body. For fair comparison, we apply the same action recognition system [4] to the two different sets of features. Finally, we combine the two feature types into a single system.

Methods For classifying the actions, we use the Hough-transform voting method of [4], which has been shown to provide state-of-the-art results and can easily be adapted to use different feature sets. In the original work, Yao *et al*. trained a random forest to learn a mapping between appearance-based feature patches (colour, optical flow, spatial and temporal gradients) and a corresponding vote in an action Hough space. We use the publicly available source code[1] and apply it directly as our appearance-based system. While more sophisticated spatio-temporal (appearance) features exist in the literature, we omit them from our experimentation as [4] showed that the above-mentioned low-level features achieve comparable results. We then modified the code to accept pose-based features and combined features.

One of the biggest challenges of using posed-based features is that semantically similar motions which can be grouped into a single action may not necessarily be numerically similar [2]. As such, we do not directly compare 3D skeleton joints in space and time. Instead, we use relational pose features, which describe geometric relations between specific joints in a single pose or a short sequence of poses. Relational pose features, introduced in [2], have been used previously for indexing and retrieval of motion capture data; we modify a subset of them for use in the random forest framework (see Fig. 2 for a qualitative description). For combining appearance and pose-based features, we pass both appearance and pose information to the Hough forest and allow the classifier to automatically select the relevant features.

We focus our comparison on a home-monitoring scenario and use the TUM kitchen dataset [3]. Even though we use the 3D joint positions from the dataset directly, these values were determined by a markerless motion capture system [1] and exemplify state-of-the-art pose estimation results, *i.e.* are not measured from markers.

Results & Conclusion Using the same classifier on the same dataset, our results showed that pose-based features (81.5%) outperform appearance features (69.8%). Combining the two types of features showed no improvement (80.1%), since the action classes which involve interaction with the environment (cupboard or drawer) are already accurately estimated by the pose features. While pose-based action recognition is often criticised for requiring extensive preprocessing for accurate limb tracking, we also tested robustness of the pose-based features by corrupting the test joint data. Even with high levels of noise (up to 100mm of additive Gaussian noise), the pose-based features either matched or outperformed appearance-based features, indicating that perfect pose estimates are not necessary. On the other hand, appearance features are applicable in many cases in which poses cannot be extracted and are also capable of encoding contextual information. We believe that a combination of appearance and pose features would be most ideal when actions cannot be classified by the pose alone, even though this was not the case in our experiments.

[1] J. Bandouch and M. Beetz. Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models. In *Int. Workshop on Human-Computer Interaction*, 2009.

[2] M. Müller, T. Röder, and M. Clausen. Efficient content-based retrieval of motion capture data. *ACM Trans. Graph.*, 24:677–685, 2005.

[3] M. Tenorth, J. Bandouch, and M. Beetz. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE Workshop on THEMIS*, 2009.

[4] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *CVPR*, 2010.

[1] http://www.vision.ee.ethz.ch/~yaoa