Self-supervised Learning of Pose Embeddings from Spatiotemporal Relations in Videos Supplementary Material

Ömer Sümer* Tobias Dencker* Björn Ommer
Heidelberg Collaboratory for Image Processing
IWR, Heidelberg University, Germany
firstname.lastname@iwr.uni-heidelberg.de

1. Introduction

In the following we present material that supplements the approach and the experiments from our paper. Different configurations of the temporal ordering task are investigated followed by a visualization of the curriculum that we use for training. Moreover, we show qualitative results of our repetition mining procedure and of pose retrievals on the MPII Human Pose dataset.

2. Temporal Ordering Task

Positive τ^+	Negative $[\tau^-_{min}, \tau^-_{max}]$	Avg. AUC
3	[8, 16]	0.768
4	[8, 16]	**0.781**
5	[8, 16]	0.769
4	[6, 12]	0.776
4	[10,18]	0.778

Table 1: Positive and negative sampling ranges in temporal ordering task.

Complementary to ablation studies in Section 4.2, we study the sensitivity of the temporal ordering task to different configurations. In particular, we vary the ranges τ^+, τ^- for sampling positive and negative pairs as describe in Section 3.1 and train a network using only the modified temporal ordering task. Table 1 shows that the best setting for the two sampling ranges is $\tau^+ = 4$ and $\tau^- = [8, 16]$.

Our understanding is that the difficulty of the tasks is influenced by τ^+. The larger this parameter the more variation in the sampled positive class has to be accounted for. The selection is therefore a trade-off: A small τ^+ results in little variation and potentially overfitting, a large τ^+ results in more variation and training might show little or no convergence. The range τ^- behaves similarly, but has a smaller impact on the performance.

3. Curriculum

The curriculum, that we employ for training the spatial and temporal tasks on the Olympic Sports dataset, is visualized as a histogram in Figure 1. The histogram bins correspond to the training data which is used at different stages of the curriculum. As described in Section 3.2, the difficulty of training data increases from stage to stage. It is measured by the fg/bg ratio of estimated optical flow. The figure visualizes the amount of data samples as well as the absolute frequencies of the sixteen Olympic Sports categories in each stage.

The gradual augmentation of training data according to the curriculum changes the distribution of the categories across stages. This is due to the varying difficulty of the categories: 1) Videos in categories like hammer throw usually contain large clips with clearly visible motion which are already included in the early stages of the curriculum. 2) Videos in categories like clean-and-jerk as well as shot-put largely consist of slow motion parts with less obvious differences between nearby frames and thus are more prominent in the later stages of the curriculum.

4. Repetitions

In Figures 2, 3 and 4 we illustrate the kind of repetitive poses that our method is able to locate inside individual videos as described in Section 3.3. Our self-supervised embeddings detect groups of repetitions in spite of changes in lightening, camera angle and background.

5. Pose Retrieval

In Figure 5 we show qualitative results for the task of pose retrieval on the MPII Human Pose dataset. We compare Alexnet with our best performing model. Even though

*Both authors contributed equally to this work.

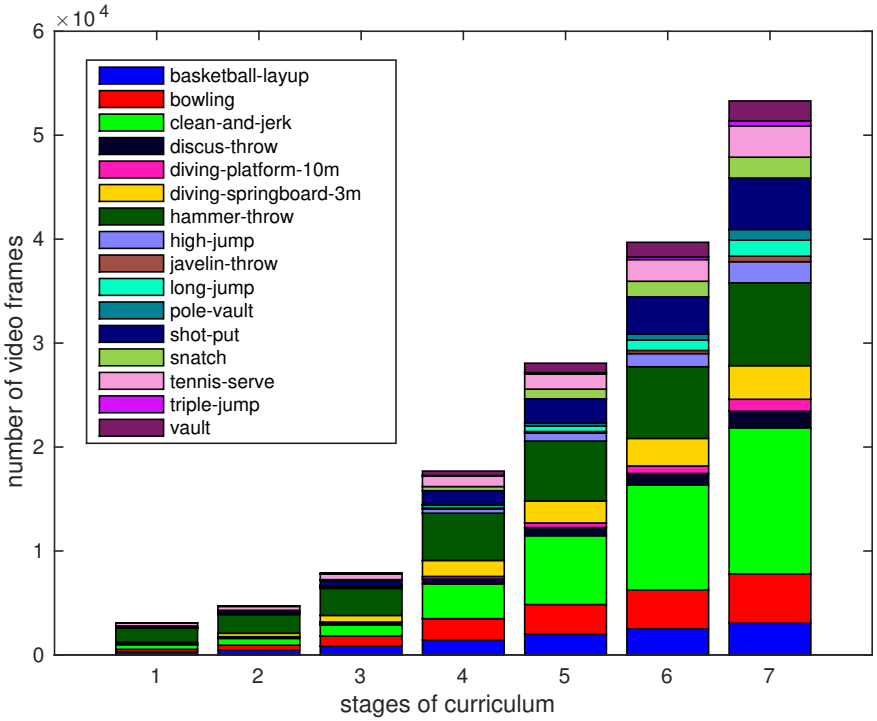

Figure 1: A histogram describing the curriculum used for training on Olmpic Sports dataset. The curriculum contains seven stages with training samples of increasing difficulty.

our method is only trained on Olympic Sports videos (not MPII images) and without any pose labels, our method shows favorable retrievals. The last row shows a typical failure cases due to occlusions.

Figure 2: Groups of repetitive poses mined from a single video in the Olympic Sports dataset. Each row shows a query frame and its repetitions, which are retrieved by our method. The column of query frames and each row of repetitive poses are sorted according to frame numbers that are shown inside the yellow boxes.

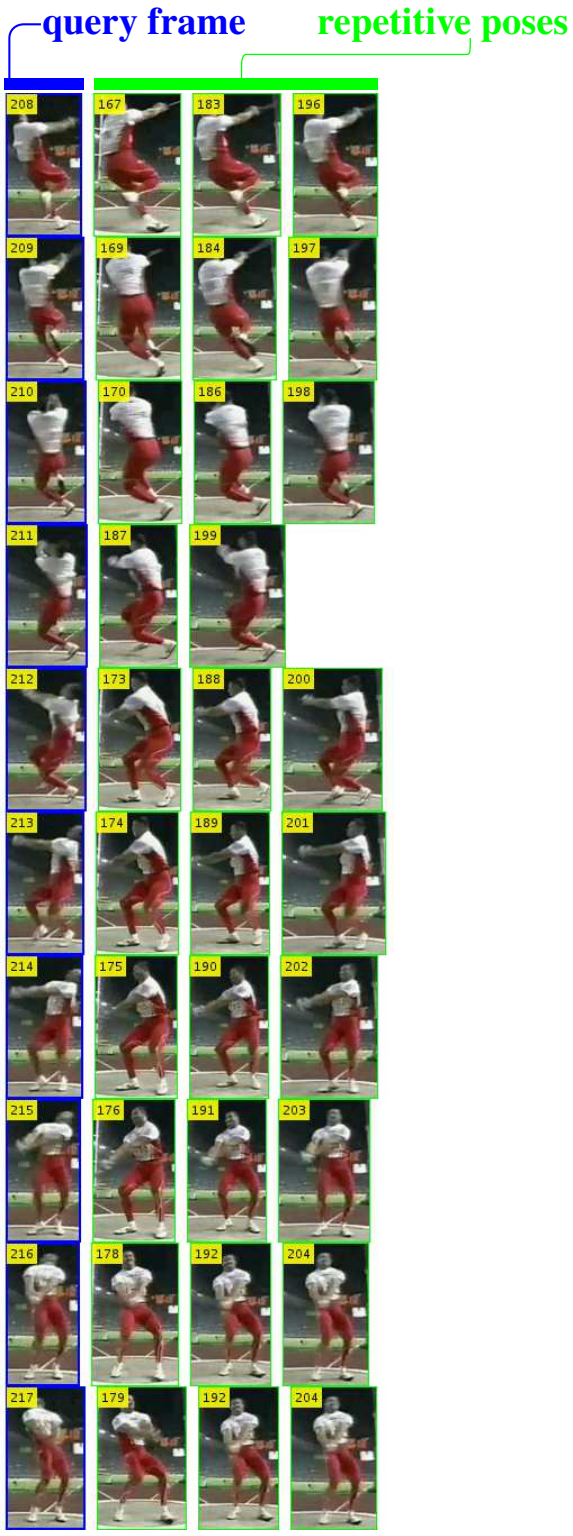

Figure 3: Groups of repetitive poses mined from a single video in the Olympic Sports dataset. Each row shows a query frame and its repetitions, which are retrieved by our method. The column of query frames and each row of repetitive poses are sorted according to frame numbers that are shown inside the yellow boxes.

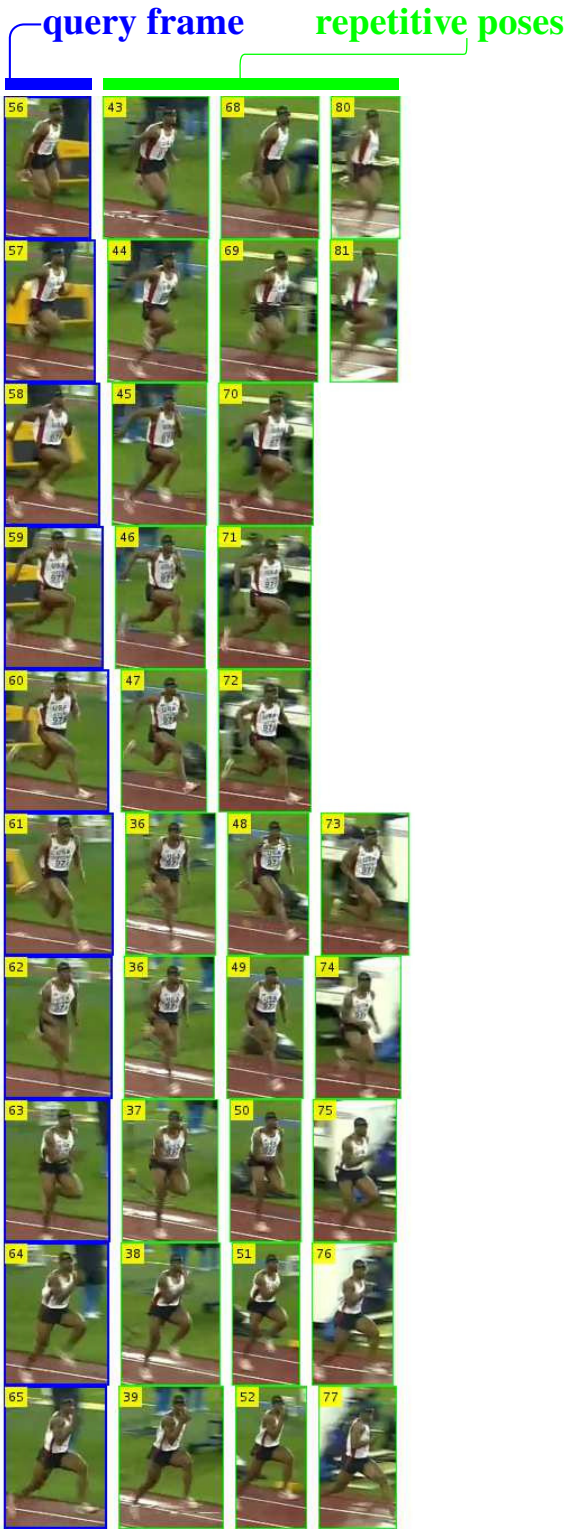

Figure 4: Groups of repetitive poses mined from a single video in the Olympic Sports dataset. Each row shows a query frame and its repetitions, which are retrieved by our method. The column of query frames and each row of repetitive poses are sorted according to frame numbers that are shown inside the yellow boxes.

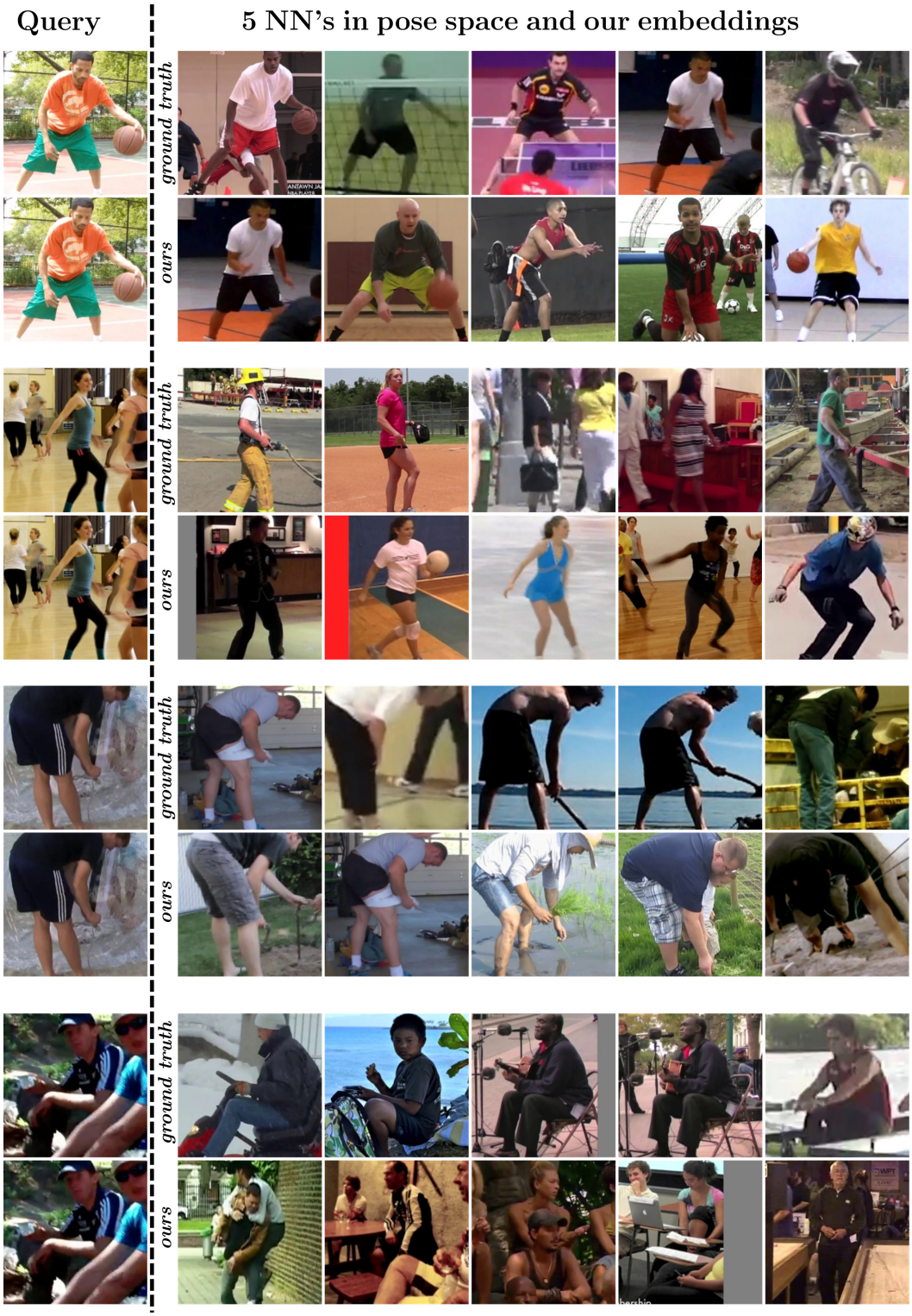

Figure 5: Pose retrievals for four query images on the MPII Human Pose dataset. We retrieve five nearest neighbors using either the ground truth by computing Euclidean distances in pose space, or using our self-supervised pose embeddings. The last row illustrates a failure case related to occlusions.