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

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

Abstract

Human pose analysis is presently dominated by deep convolutional networks trained with extensive manual annotations of joint locations and beyond. To avoid the need for expensive labeling, we exploit spatiotemporal relations in training videos for self-supervised learning of pose embeddings. The key idea is to combine temporal ordering and spatial placement estimation as auxiliary tasks for learning pose similarities in a Siamese convolutional network. Since the self-supervised sampling of both tasks from natural videos can result in ambiguous and incorrect training labels, our method employs a curriculum learning idea that starts training with the most reliable data samples and gradually increases the difficulty. To further refine the training process we mine repetitive poses in individual videos which provide reliable labels while removing inconsistencies. Our pose embeddings capture visual characteristics of human pose that can boost existing supervised representations in human pose estimation and retrieval. We report quantitative and qualitative results on these tasks in Olympic Sports, Leeds Pose Sports and MPII Human Pose datasets.

1. Introduction

The ability to recognize human posture is essential for describing actions and comes natural to a human being. Different poses in a video form a visual vocabulary similar to words in text. An important objective of computer vision is to bring this ability to the computer. Finding similar postures across different videos automatically enables a lot of different applications like action recognition [3, 4] or video content retrieval.

So what makes two postures look similar? More formally, a similarity function, which is entailed by a pose embedding, needs to capture the characteristics of different postures, while exhibiting the necessary invariance to strong intra-class variations. In particular, it should be sensitive to articulation of body parts while being invariant to illumination, background, clutter, deformations (e.g. facial expressions) or occlusions. Often human joints are used as a surrogate for describing similarity, but there are several issues: First, measuring distances in pose space accurately and coming up with a non-ambiguous Euclidean embedding is already a challenging problem. Second, the manual annotation of human joints in larger datasets is expensive and time-consuming.

Convolutional networks have recently been immensely helpful to computer vision. The feature hierarchy of such a network is effectively defined by a cascade of filter banks that are recursively applied to extract discriminative features for the given task. In this work we take advantage of convolutional networks to learn pose embeddings from videos.

In supervised similarity learning we assume that we are given positive and negative pairs of postures for training. In this supervised setting convolutional networks excel and have recently surpassed human performance in some basic tasks. In contrast unsupervised training of convolutional networks is still an open problem and currently the focus of the research community. In this paper we investigate how to learn a pose representation without labels.

A solution for the problem of missing supervision is to switch to a related auxiliary task for which label information is available. For this self-supervised strategy several well-known sources of weak supervision have been recently re-visited: among them spatial configuration of natural scenes, inpainting, super-resolution, image colorization, tracking, ego-motion and even audio. Although there are many sources available, not all of them are suitable for the application in pose analysis. We exploit human motion in videos to make pose similarity apparent and learnable without labels. With an almost infinite supply of video data online exploiting this idea is very attractive.

We propose learning spatiotemporal relations in videos

*Both authors contributed equally to this work.

by means of two complementary auxiliary tasks: a *temporal ordering* task which learns whether two given person images are temporally close (similar) and a *spatial placement* task which discovers randomly extracted crops from the spatial neighborhood of persons, and learns whether given patches are a person or not. Learning spatial and temporal relations of human movement simultaneously provides us information of "what" we are looking at (person/ not person) and "how" the instances differ (similar/dissimilar poses). *Curriculum-based* learning and *repetition mining* arrange the training set by starting from only the easy samples and then iteratively extend to harder ones, while also eliminating inactive video parts. Then our spatiotemporal embeddings successfully learn representative features of human pose in a self-supervised manner.

2. Related Work

Human pose analysis deals with problems such as pose retrieval, similarity learning and pose estimation. Most approaches in pose analysis rely on supervised data and there exists only a few unsupervised approaches. Here, we summarize significant examples of pose analysis and related unsupervised learning approaches:

Pose estimation Pose estimation aims at finding locations of body joints, whereas pose retrieval or embedding finds a metric that can retrieve the most similar poses and discriminates samples according to their pose information, without localizing joints directly. With the advancements in convolutional neural networks [21], pose estimation is also dominated by deep learning-based methods. Toshev and Szegedy [31] estimated joint locations directly regressing in a CNN architecture. Instead of simply regressing joint locations, Chen and Yuille [10] learned pairwise part relations combining CNN with graphical models. Tompson *et al.* [30] exploited CNNs for relationship between body parts with a cascade refinement. A recent work by Newell *et al.* [25] used fully convolutional networks in a bottom-up top-down manner to predict heatmaps for joint locations.

Similarity learning The first *Siamese*-type architecture [8] was proposed to learn a similarity metric for signature verification. Similarity learning was also applied in human pose analysis. In [24] and [22], body joint locations are used to create similar and dissimilar pairs of instances from annotated human pose datasets. [22] also transferred a learned pose embedding to action recognition.

These works in pose estimation and similarity learning exploited large amounts of annotations (body joints or labeling of similar/dissimilar postures). However, unsupervised learning methods without using labels showed promising performance in various learning tasks in the last decade. Self-supervised learning is very popular similar to classical unsupervised methods such as clustering, autoencoders [29], restricted Boltzman machines [17].

Self-supervised learning The availability of big data motivated the community to investigate alternative sources of supervision such as ego-motion [1, 33], colorization [34], image generation [28], spatial [12, 27] or temporal clues [32, 23]. As our approach belongs to the class of self-supervised methods using spatial and temporal information, we describe these methods in detail.

Doersch *et al.* [12] trained convolutional neural networks to take image patches from a 3×3 grid and classify the relative location of 8 patches with respect to a center patch. Norozzi and Favaro [27] argued that solving locations of relative patches could introduce ambiguities and proposed a localization problem given all 9 patches at once. Also, they used 100 relative locations as class labels out of $9!$ permutations using a Hamming distance-based selection.

Wang and Gupta [32] exploited videos by detecting interesting regions with SURF keypoints and tracking them. Then, they used a Siamese-triplet architecture with a ranking loss together with random negative selection and hard negative mining. However, tracking is not the best solution in the challenging context of pose analysis due to the non-rigid deformations of person patches which are in low resolution and contain too few keypoints to detect parts and track them precisely.

Misra *et al.* [23] defined a temporal order verification task, which classifies whether given 3-frame sequences are temporally ordered or not by altering the middle frame. In action/pose benchmarks or internet videos, there are a lot of cyclic human actions (*e.g.* running based sports, dancing), which often produce confusing samples and interfere with representation learning.

In order to learn a better representation, we argue that temporal cues which aim to learn whether given inputs are from temporally close windows or not will be a more effective approach. The use of temporal cues to learn whether given inputs are from temporally close windows or not is an effective approach for representation learning. Local proximity in data (slow feature analysis, SFA) has first been proposed by Becker and Hinton [6]. The most recent spatial and temporal self-supervised learning methods are inspired from SFA. Goroshin *et al.* [16] created a connection between slowness and metric learning by temporal coherence. Motivated by temporal smoothness in feature space, Jayaraman and Grauman [18] exploited higher order coherence, which they referred to as steadiness, in various tasks. Slowness or steadiness criterion can introduce significant drawbacks mostly because of limited motion and the repetitive nature of human actions. Thus, we learn auxiliary tasks in relatively small temporal windows which do not contain more than a single cycle of action. Moreover, the use of curriculum learning [7] and repetition mining refine and guide our self-supervised tasks to learn stronger temporal features.

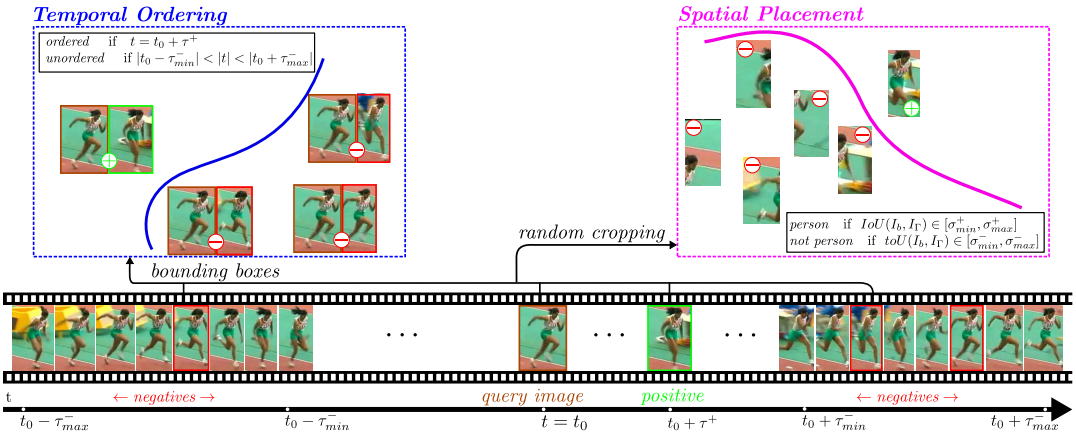

Figure 1: Sampling procedure for training self-supervised pose embeddings. For each query image in a video, positive and negative pairs of temporal ordering are collected from specific temporal ranges *(left)*. In spatial placement, samples are cropped using the IoU criterion *(right)*.

Curriculum learning has been proposed by Bengio *et al.* [7] and it speeds up training and improves test performance by using samples whose difficulties are gradually increasing in shape recognition and language modeling. To the best of our knowledge, the potential of a curriculum has not been studied in the self-supervised setting, where we associate the difficulty of training samples with their inherent motion.

3. Approach

Our motivation is to learn pose embeddings from videos without labels. We follow the insight that spatiotemporal relations in videos provide sufficient information for learning. For this purpose, we propose a self-supervised pipeline that creates training data for two auxiliary tasks: 1) temporal ordering and 2) spatial placement. Since the raw self-supervised output needs refinement, we introduce curriculum learning and repetition mining as key ingredients for successful learning. The two auxiliary tasks are trained in a Siamese CNN architecture and the learned features are eventually used as pose embeddings in order to retrieve similar postures and estimate pose.

3.1. Self-supervised Pose Embeddings: Temporal Ordering and Spatial Placement

We consider a temporal and a spatial auxiliary task which are automatically sampled from videos as described in Fig. 1. Both tasks capture complementary information from inside videos essential for learning a pose embedding. The temporal task teaches the pose embedding to become more sensitive to body movements and more invariant to camera motion (*i.e.* panning, zoom in/out, jittering), while the spatial task relies on the spatial configuration of a single frame and focuses on learning a human appearance model which strengthens the ability to separate posture from background.

For the temporal ordering task, a tuple of two frames is sampled from the same video together with a binary label which indicates whether the first frame (anchor) is closely followed in time by the second frame (candidate). In order to focus on learning human posture, we do not sample the full frames, but instead crop bounding box estimates of the person of interest. Thus, the training input for the temporal ordering task consists of two cropped boxes and a binary label indicating whether the two boxes are temporally ordered.

For a frame I_{t_0} sampled at time point t_0, we sample a candidate frame I_t with a temporal offset of $\Delta_t = t - t_0$. In order to sample a positive candidate the offset needs to be $\Delta_t = \tau^+$, while a negative candidate is sampled if

$$\Delta_t \in \tau^- = [\tau^-_{min}, \tau^-_{max}] \cup [-\tau^-_{max}, -\tau^-_{min}]$$

holds. $\tau^-_{min}, \tau^-_{max}$ are the range limits of the negative candidates. In other words, a positive candidate comes exactly from τ^+ frames in the future, while negative candidates come from ranges before or after the anchor frame.

The temporal ordering task relies on the assumption of temporal coherence that frames in a small temporal neighborhood are more similar than distant frames. We add the constraint that positive candidates can only come from the future. Since the self-supervised sampling from videos already introduces large amounts of variation, we want the positive class to be as homogeneous as possible in order to facilitate training. In contrast the negative class is sampled from a larger range that allows more variation, but is still close enough to the positive class to provide challenging similarities for discriminative learning.

For the spatial placement task, a box is randomly cropped from a single frame together with a binary label

that indicates whether the cropped box overlaps with the estimated bounding box of a person in this frame. The overlap is measured with the Intersection-over-Union (IoU) criterion [13]. For the estimated bounding box I_b and a randomly cropped box I_r, the binary label y_S is defined as

$$y_S(I_b, I_r) = \begin{cases} 1, & \text{if } IoU(I_b, I_r) \in [\sigma^+_{min}, \sigma^+_{max}] \\ 0, & \text{if } IoU(I_b, I_r) \in [\sigma^-_{min}, \sigma^-_{max}] \end{cases} \quad (1)$$

where $IoU(\cdot, \cdot)$ computes the IoU and $[\sigma^+_{min}, \sigma^+_{max}]$ defines the positive range of overlap while $[\sigma^-_{min}, \sigma^-_{max}]$ defines the negative. Since the estimated bounding boxes are not completely reliable, the positive and negative IoU ranges are usually selected with a gap between them to help the separation of the classes.

In both auxiliary tasks, three negative samples are used for each positive posture, because sampling of negatives (what it is not) from larger ranges helps with learning positive similarities (what it is) precisely. The intuition is that the pose embedding learns to discriminate between a homogeneous positive and a more heterogeneous negative class in both tasks. Since both tasks focus on different aspects of human posture, the best pose embedding is obtained by joint training. We investigate the contribution of different configurations in Sect. 4.2.

3.2. Creating a Curriculum for Training

In supervised training with human annotations, it is often beneficial to avoid difficult samples with ambiguous or even incorrect labels, because this kind of data can inhibit convergence and lead to inferior results. In the self-supervised case, we find that data quality fluctuates even more and needs to be taken into account. On the other hand, skipping too many difficult training samples can result in overfitting on a small subset of easy samples and hurts generalization to unseen datasets. We propose to strike a balance by using a curriculum of training data that gradually increases in difficulty over the course of training. We create the curriculum with regard to the temporal ordering task which produces far more inconsistent samples than spatial placement.

In order to determine the difficulty of temporal ordering for a particular training sample, we look into the motion characteristics of the respective video. For instance, a clean-and-jerk video mainly consists of inactive parts with little motion, whereas a long-jump video is dominated by a highly repetitive structure with fast moving, deforming postures. Training samples from video sequences with clear foreground motion (e.g. a long-jump video) are preferable for learning temporal ordering, because their negative candidates, which are sampled from the range of τ^-, are easier to distinguish from the positive ones from τ^+. Therefore, we determine the difficulty of a training sample by estimating the motion in videos and sample training frames with sufficient action.

When creating a curriculum, we use an optical flow based criterion that computes the ratio of the optical flow in the foreground and background of the frame. To compute the *fg/bg ratio* the mean magnitude of optical flow in the foreground bounding box is divided by mean magnitude of optical flow of the background. We use the method from [9] to estimate the optical flow between two frames. The fg/bg ratio acts as a proxy of a *signal-to-noise* ratio, as examples with higher values are more easily separated from the background.

The curriculum is assembled by sorting the training samples according to their flow ratio and splitting them in discrete blocks, curriculum updates, with increasing difficulty (decreasing flow ratio). We analyze the impact of the curriculum in an ablation experiment in Table 1 where we train the network with and without a curriculum using the same subset of self-supervised training data. Details of ablation experiments and the effect of curriculum will be explained in Sect. 4.1 and Sect. 4.2.

3.3. Mining Repetitive Poses

There are two reasons why we pay special attention to repetitive poses in video sequences: First, they impair the training of the temporal ordering task. Second, if the location of repetitions were known, they could be extracted and used as valuable training data, which we refer to as *repetition mining*. The mined repetitions augment temporal ordering by providing a new similarity learning task.

While the proposed curriculum avoids difficult samples in the early stages of self-supervised training, repetitive poses in videos are not filtered by the motion-based curriculum. The training of the temporal ordering task suffers from repetitions which can cause incorrect labeled image pairs by violating the assumption of temporal coherence. For instance, if a negative frame is sampled from a video with a repetitive action like running or walking, it might be more similar to the anchor frame than the positive candidate.

After an initial training of the temporal ordering task, we use the learned pose embeddings to detect repetitive poses in the training data. For each video, we obtain a self-similarity matrices by computing all the pairwise distances between frames. As distance measure, we use the Euclidean norm of the normalized `pool5` features. In order to extract reliable and strong repetitions, we convolve the self-similarity matrix with a 5x5 circulant filter matrix to suppress potential outliers that are not aligned with the off-diagonals by thresholding. The maxima of each row indicate the fine-scaled repetitions of the respective query frame. Fig. 2 shows an example self-similarity matrix and repetitions which are mined using this approach.

Repetitive poses form groups of very similar but not identical images due to small variations over time caused by the persons movement, changes in the camera viewpoint,

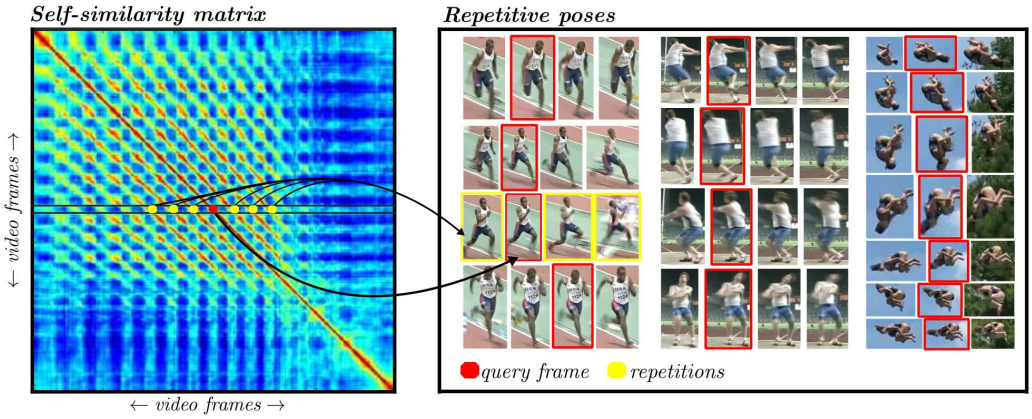

Figure 2: Mining repetitive poses. Off-diagonal structures of the self-similarity matrix on the left indicate repetitions in a video. For each row, repetitions are mined using a query frame. Repetitive poses from three videos are shown on the right.

or even the frame rate of the video camera. These groups of highly similar images help to learn the more fine-grained details of human posture. They can be used to create a new type of similar-dissimilar problem. Similar pairs are chosen among repetition groups, negative candidates are picked from regions between the repetitions.

As repetitions occur only in a subset of the available video data, they are combined with samples from non-repetitive videos and added to the first stages of the curriculum. The mined repetitive poses are in quality close to human annotated similarities and provide a stabilizing effect on the whole training procedure.

Our method can be employed in a bootstrapping fashion, by repeatedly training the temporal ordering task and mining repetitions which provide better training samples without additional supervision.

3.4. Network Architecture

For the two self-supervised tasks we train two convolutional neural networks which differ in the number of images they process as shown in Fig. 3. The temporal ordering task is trained using a Siamese architecture [8] that takes a pair of images as input while the spatial placement task is trained on single images using a common single stream architecture.

We adopt the well-known Alexnet architecture [21] for both tasks. In the temporal task the two Siamese streams consist of convolutional layers. After the last pooling layer the output from the two streams is concatenated. The fully-connected layers compute a binary output probability for testing. The convolutional networks are trained by minimizing binary cross-entropy loss functions. For joint training of both tasks the weights in the convolutional layers are not only shared between the Siamese parts but are also shared with the convolutional layers of the spatial placement task. Moreover, the joint loss of the two auxiliary tasks is computed in a weighted sum.

After training the network, we use the feature representation from the last shared layer `Pool5` as pose embeddings. Features of this layer provide good localization which is important for pose retrieval and estimation.

We make several modifications to the Alexnet architecture: 1) Because we want to avoid overfitting and our binary tasks do not require a large number of parameters, both networks have a reduced number of neurons in the fully connected layers compared to the original Alexnet (namely 2048/1024 vs 4096/4096). 2) To improve training of the temporal task we replace the regular rectified linear unit in the last convolutional layer with a non-linearity that has a negative slope. We find that this modification is critical for

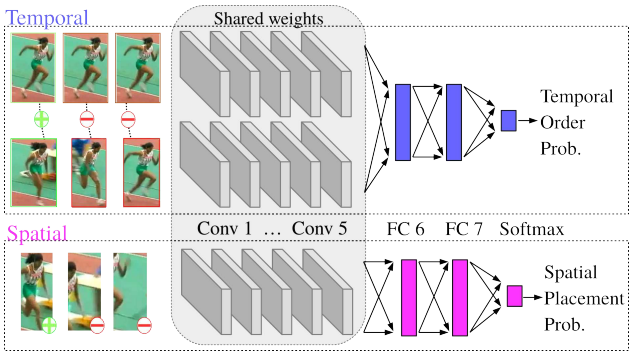

Figure 3: Network architecture for temporal ordering and spatial placement.

performance. 3) The use of batch normalization in the fully connected layers is an important regularizer in our training that helps with generalization to other datasets.

4. Experiments

We present experiments on posture analysis, pose estimation and pose retrieval. The training of our method is demonstrated on the Olympic Sports dataset (OSD) [26]. In different ablation experiments we highlight the design decisions in our proposed method. To study the ability of our approach to generalize to unseen datasets we include experiments on Leeds Sports Pose (LSP) [20] and the challenging and unconstrained MPII Human Pose [2]. Additionally in a supervised pose estimation setting [31], we report performance of our method in comparison with other initialization approaches.

4.1. Training and Testing Details

From 680 videos in Olympic Sports dataset, we extract approximately 140,000 frames for which we obtain bounding box estimates using the method in [14]. Our training curriculum uses about 80,000 frames which are ordered using the flow ratio criterion described in Sect. 3.2. It starts out with about five percent of the easiest training samples and increases the amount of training data in seven steps every 2.5K iterations. The amount of training data grows exponentially during the first few curriculum updates, but does not surpass 25 percent of training data for a single update.

For training of the convolutional networks we use the *Caffe* framework [19]. We optimize our model using the Adam solver for stochastic batch gradient descent with batch size of 48 and a fixed learning rate of 10^{-4}. In the convolutional layers, we use a reduced learning rate of 10^{-5}. The training is stopped after 40K iterations. For joint training we reduce the loss weight for the spatial task by a factor of 0.1. In the auxiliary tasks we use $\tau^+ = 4$ and $\tau^- = [8, 16]$ as well as $\sigma^+ = [0.65, 0.95]$ and $\sigma^- = [0.25, 0.55]$. For mining repetitions we follow the procedure described in Sect. 3.3 and iterate it two times to collect about 15000 frames with repetitive poses. We find that two iterations are sufficient, since our method has found most of the repetitions by this time. For testing, we use the pairwise Euclidean distance of `Pool5` features as a similarity measure between images.

4.2. Ablation Experiments on Posture Analysis in Olympic Sports Dataset

We demonstrate on the Olympic Sports dataset, how different configurations of our method affect the performance of the learned pose embeddings. For the evaluation on the the Olympic Sports dataset, we adopt the posture analysis benchmark proposed in [5]. It consists of 1200 exemplar postures for each of which ten positive (similar) and ten negative (dissimilar) poses are defined. The performance is determined by the ability of the pose embeddings to separate positives from negatives and measured in terms of the area under the curve (AuC) of a ROC.

Task	without curriculum	with curriculum
temporal(T)	0.592	0.630
temporal& spatial	0.664	0.679
temporal(T)$*$	0.762	0.781
temporal& spatial$*$	0.767	0.784

Table 1: Average AUC in Olympic Sports benchmark shows effect of curriculum training. Methods with $(*)$ are initialized with Imagenet pre-trained weights.

First, we study the impact of curriculum learning. We train our temporal (T) and temporal & spatial (ST) tasks once with and once without a curriculum, but using the same amount of training data. The experiments in Table 1 show that the curriculum as proposed in Sect. 3.2 improves the performance of our method by 5% in mean AUC in random initialized temporal task. When the temporal task is initialized with Imagenet pre-trained weights, it improves by 2%, and this improvement is preserved even in joint learning of temporal ordering and spatial placement. Temporal ordering itself seems less powerful than spatial placement and cannot be learned without curriculum learning. However, our *fg/bg ratio* based curriculum significantly increases its performance in posture analysis.

Second, we analyze the contributions of repetition mining and the two individual auxiliary tasks in Table 2. Temporal ordering underperforms with respect to spatial placement when initialized with random weights. We argue that temporal ordering is a more challenging task, since the temporal nature of actions has to be learned by the network. When the network is initialized from Imagenet, temporal ordering performs well. It has already learned to filter relevant visual information and improves with additional temporal cues from videos. On the other hand, spatial placement does not improve on Alexnet by such a large margin, because pre-trained Alexnet already comes with a good localization ability. In both settings (initialized randomly or from Imagenet pre-trained weights), repetition mining further boosts performance. This improvement highlights the benefit of the usage of repetitions.

Additionally in Table 2, we compare our best performing method with related work. When randomly initialized, our method performs better than several different self-supervised methods [12, 32, 27] and surpasses the best competitor by nearly 5 points. It even approaches the performance of the Imagenet pre-trained Alexnet, which is impressive considering that our training leveraged 680 sport videos (approx. 80K frames used) without labels, whereas

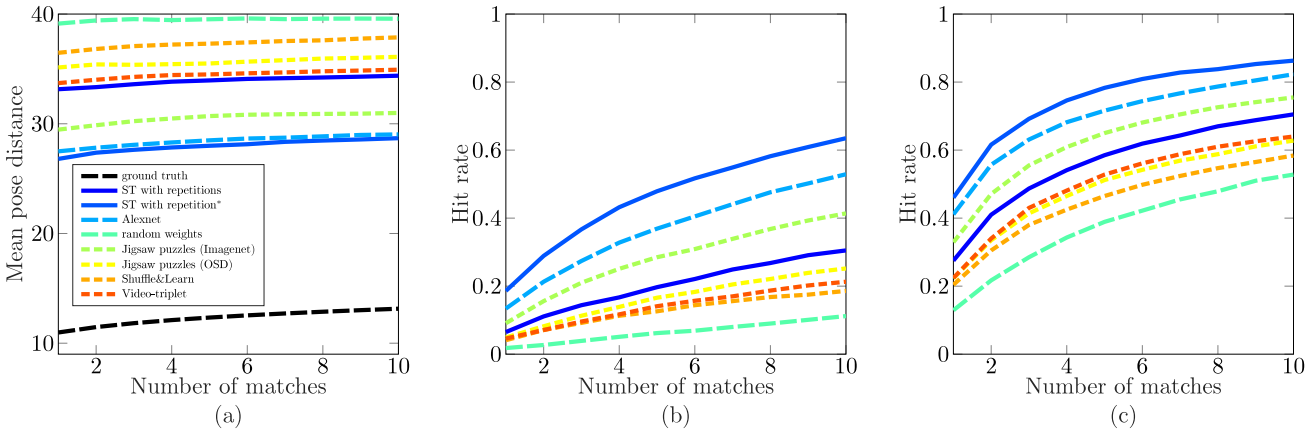

Figure 4: Pose retrieval results on MPII validation set: (a) Mean pose distance, (b) Hit rate@K using nearest neighbor criterion, (c) Hit rate@K using relative distance criterion. Model with $(*)$ initialized with Imagenet pre-trained weights.

Method	Avg. AUC
temporal(T)	0.630
spatial(S)	0.668
temporal & spatial	0.679
T with repetitions	0.658
S&T with repetitions	**0.701**
HOG-LDA	0.580
Doersch *et al.* [12]	0.580
Jigsaw puzzles [27] (Imagenet)	0.653
Jigsaw puzzles [27] (OSD)	0.646
Shuffle&Learn [23]	0.646
Video triplet [32]	0.598
Alexnet [21]	0.722
temporal*	0.781
spatial*	0.756
temporal & spatial*	0.784
T with repetitions*	0.794
S&T with repetitions*	**0.804**
CliqueCNN* [5]	0.790

Table 2: Comparative posture analysis performance of auxiliary tasks in Olympic Sports dataset. Methods with $(*)$ are initialized with Imagenet pre-trained weights of Alexnet.

Imagenet contains 1.2M labeled images. In the case of finetuning our model improves about 8 points on ImageNet pre-trained Alexnet and surpasses CliqueCNN [5] which is trained in the same setting.

4.3. Pose Retrieval

In this set of experiments we want to study the ability of our trained pose embeddings to generalize to unseen datasets. For this purpose, we evaluate our methods in the task of pose retrieval on the challenging MPII Human Pose dataset. We adopt the same procedure as described in [22], and split the fully annotated MPII training set into train and validation set. The validation set is further split in 1919 images for query and 8000 images for test purposes. The input images and pose annotations are normalized with respect to smallest square patch tightly enclosing all body part locations, and normalized into the input size of our network.

According to [22] three performance metrics are used: *mean pose distance*, *hit rate using nearest neighbor* and *relative distance criterion*. The pose distance is the mean of Euclidean distances between normalized pose vectors. The mean pose distance is computed across the first K nearest neighbors. The hit rate measures the correctness of retrieval and is defined in two different ways: 1) *nearest neighbor criterion* determines whether at least one retrieval among the K nearest neighbors belongs to the first fifty nearest neighbors in the pose space. 2) *relative distance criterion* uses a $+10$ margin of minimum pose distance between query and test set.

The pose retrieval results evaluated on the three performance metrics on MPII are shown in Fig. 4. Here, we trained our method on the spatial&temporal (ST) tasks with repetition mining using OSD only. It successfully generalizes to the challenging MPII dataset. When randomly initialized, it shows better mean pose distance and hit rate performance than previous methods, which are also trained on videos [32, 23].

When the jigsaw puzzles method [27] is trained on the larger Imagenet dataset, they clearly outperform our method. We argue that this performance gap is due to different training data. To support this assumption, we re-train their method on OSD person boxes using their official im-

plementation [1], and find it to perform worse than our self-supervised method across all measures.

When initialized from Imagenet pre-trained weights, our method outperforms Alexnet across all measures particularly in hit rates.

4.4. Pose Estimation

For pose estimation we evaluate on the Leeds Sports Pose dataset [20]. We follow the procedure described in [5] and use the 1000 training images and 3938 (fully annotated) images from the extended training set as test set for retrieval while the original 1000 test images are used as query. In both query and test images, joint locations are normalized into our networks input size.

Method	*Head*	*Torso*	*U.arms*	*L.arms*	*U.legs*	*L.legs*	*Mean*
random weights	19.3	45.2	9.6	4.1	21.1	20.3	19.9
ground truth	72.4	93.7	58.7	36.4	78.8	74.9	69.2
Chu *et al.*[11]	89.6	95.4	76.9	65.2	87.6	83.2	81.1
Shuffle&Learn [23]	36.7	66.6	20.1	8.3	37.5	35.0	34.0
Video triplet [32]	40.5	76.6	23.9	10.0	46.1	39.6	39.4
Jigsaw puzzles [27] (Imagenet)	49.3	80.1	27.5	11.9	50.5	47.4	**44.4**
Jigsaw puzzles [27] (OSD)	41.0	72.8	23.8	12.2	43.0	39.8	38.7
S&T with repetitions	40.3	74.7	23.8	11.5	45.8	42.8	39.8
Alexnet [21]	42.4	76.9	47.8	41.8	26.7	11.2	41.1
CliqueCNN [5] *	45.5	80.1	27.2	12.6	50.1	45.7	43.5
S&T with repetitions*	55.8	86.5	35.0	18.9	58.7	53.8	**51.5**

Table 3: Pose estimation results in Leeds Sports Pose dataset with PCP measures for each method. Methods with $(*)$ are initialized with Imagenet pre-trained weights.

We report the Percentage of Correct Parts (PCP) measure [15] on 14 body joints for different methods. According to PCP a part is considered correct, if its endpoints are within 50% part length of the corresponding ground truth endpoints.

Unsupervised pose estimation results of LSP in Table 3 show that our method, when initialized randomly, performs better than other self-supervised methods except for jigsaw puzzles trained on Imagenet. As in the case of pose retrieval, we argue that it is due to the size of Imagenet. When initialized from pre-trained weights, our method clearly outperforms [21, 5].

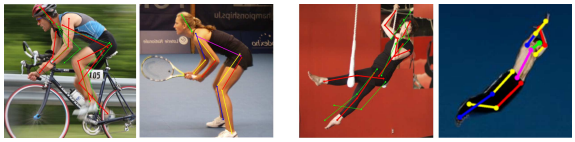

Figure 5: Pose estimation results in Leeds Sports Pose dataset. First images are from test set with the superimposed ground truth skeleton depicted in red and the predicted skeleton in green. Second images are corresponding nearest neighbors.

[1] https://github.com/MehdiNoroozi/JigsawPuzzleSolver

Some qualitative samples from the query set together with their nearest neighbors are shown in Fig. 5. Our method is able to retrieve similar poses even if the query is very different from our training data.

In addition to our unsupervised experiments, we use our pose embeddings as an initialization of the supervised DeepPose [31] method. In total, we evaluate four different initializations of [31] on the MPII dataset: (i) our randomly initialized spatial&temporal (ST) with repetitions model, (ii) Shuffle&Learn [23], (iii) random initialization, and (iv) Imagenet pre-trained Alexnet [21].

	Ours	Shuffle&Learn [23]	Random init.	Alexnet[21]
Head	82.6	75.8	79.4	87.2
Neck	90.3	86.3	87.1	93.2
LR Shoulder	79.5	75.0	71.6	85.2
LR Elbow	62.8	59.2	52.1	69.6
LR Wrist	47.1	42.2	34.6	52.0
LR Hip	75.5	73.3	64.1	81.3
LR Knee	65.3	63.1	58.3	69.7
LR Ankle	59.5	51.7	51.2	62.0
Thorax	90.1	87.1	85.5	93.4
Pelvis	80.3	79.5	70.1	86.6
Total	73.3	69.3	65.4	78.0

Table 4: PCKh@0.5 measure for DeepPose method [31] on MPII Pose benchmark dataset comparing different initialization approaches.

For all initializations, we train the DeepPose method using the same setup and evaluate using PCKh@0.5 metric as shown in Table 4. Our method shows an improvement of 7.9% and 4% compared with random initialization and Shuffle&Learn, respectively. It is only 4.7% below Alexnet, which is learned using the labels of 1.2 million images.

5. Conclusion

In this paper, we have proposed two complementary self-supervised tasks, temporal ordering and spatial placement which are trained jointly on unlabeled video data. To boost self-supervised training, we have introduced a motion-based curriculum and a procedure for mining repetitive poses and using them as valuable training data. Our pose embeddings capture the characteristics of human posture, which we have demonstrated in experiments on pose analysis. In the Olympics Sports dataset, the learned representation decreases the gap between self-supervised methods and Imagenet supervision, and fine-tuning with our self-supervised approach significantly improves the performance of models pre-trained on Imagenet. Finally, we have shown that the trained embeddings are able to generalize to unseen datasets in pose analysis without fine-tuning.

Acknowledgments: This work has been supported in part by the Heidelberg Academy for the Sciences, DFG, and by an NVIDIA hardware grant.

References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 37–45, Dec 2015.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, June 2014.

[3] B. Antic and B. Ommer. *Learning Latent Constituents for Recognition of Group Activities in Video*, pages 33–47. Springer, Cham, 2014.

[4] B. Antic and B. Ommer. Per-sample kernel adaptation for visual recognition and grouping. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1251–1259, Dec 2015.

[5] M. A. Bautista, A. Sanakoyeu, E. Tikhoncheva, and B. Ommer. Cliquecnn: Deep unsupervised exemplar learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3846–3854. Curran Associates, Inc., 2016.

[6] S. Becker and G. E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163, 1992.

[7] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA, 2009. ACM.

[8] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. In *NIPS*, pages 737–744. 1994.

[9] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, volume 3024 of *Lecture Notes in Computer Science*, pages 25–36, 2004.

[10] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014.

[11] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4715–4723, June 2016.

[12] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, Dec 2015.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[15] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[16] R. Goroshin, J. Bruna, J. Tompson, D. Eigen, and Y. LeCun. Unsupervised learning of spatiotemporally coherent metrics. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4086–4093, Dec 2015.

[17] G. E. Hinton and T. J. Sejnowski. Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1983.

[18] D. Jayaraman and K. Grauman. Slow and steady feature analysis: Higher order temporal coherence in video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3852–3861, June 2016.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[20] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[22] S. Kwak, M. Cho, and I. Laptev. Thin-slicing for pose: Learning to understand pose without explicit pose estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4938–4947, June 2016.

[23] I. Misra, C. L. Zitnick, and M. Hebert. *Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification*, pages 527–544. Springer, Cham, 2016.

[24] G. Mori, C. Pantofaru, N. Kothari, T. Leung, G. Toderici, A. Toshev, and W. Yang. Pose embeddings: A deep architecture for learning to match human poses. *arXiv preprint arXiv:1507.00302*, 2015.

[25] A. Newell, K. Yang, and J. Deng. *Stacked Hourglass Networks for Human Pose Estimation*, pages 483–499. Springer, Cham, 2016.

[26] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. *Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification*, pages 392–405. Springer, 2010.

[27] M. Noroozi and P. Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*, pages 69–84. Springer, Cham, 2016.

[28] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[29] D. E. Rumelhart and J. L. McClelland, editors. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, MA, USA, 1986.

[30] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling,

C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1799–1807. Curran Associates, Inc., 2014.

[31] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, June 2014.

[32] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, pages 2794–2802, 2015.

[33] A. R. Zamir, T. Wekel, P. Agrawal, C. Wei, J. Malik, and S. Savarese. *Generic 3D Representation via Pose Estimation and Matching*, pages 535–553. Springer, Cham, 2016.

[34] R. Zhang, P. Isola, and A. A. Efros. *Colorful Image Colorization*, pages 649–666. Springer, Cham, 2016.