Few-shot keypose detection for learning of psychomotor skills

Benjamin Paaßen[1], Tobias Baumgartner[2], Mai Geisen[2], Nina Riedl[2] and Miloš Kravčík[1]

Abstract

Some psychomotor tasks require students to perform a specific sequence of poses and motions. A natural teaching scheme for such tasks is to contrast a student's execution to a teacher demonstration. However, this requires strategies to match the teacher demonstration of each motion to the student's attempts and to identify differences between demonstration and attempt. In this paper, we investigate methods to automatically detect student attempts for poses with only a single correct teacher demonstration. We investigate relevance learning, prototype networks, and attention mechanisms to achieve a robust few-shot approach which generalizes across students. In an experiment with one teacher and 27 students performing a sequence of motion elements from the field of fitness and dance, we show that prototype networks combined with an attention mechanism perform best.s

Keywords

psychomotor training, few-shot learning, metric learning, prototype networks, convolutional neural networks

1. Introduction

Some psychomotor skills require us to execute a specific sequence of poses, such as in dance choreographies, as well as during the repetitive execution of fitness moves, e.g. squats. While research has emphasized the need for holistic teaching beyond mere imitation [1], at least part of beginner's training is concerned with getting the basic set of poses correctly executed [2]. Our aim is to automate some of this basic teaching by contrasting a student's current motion with a teacher demonstration and highlighting differences [3]. However, to correctly compare student and teacher motion, we need to establish a matching between both sequences of poses. Typically, matchings between sequences are performed via alignment distances such as dynamic time warping [4, 2]. However, such techniques are ill-suited to our scenario, where correct execution only depends on poses, not on the transition motion between poses. Our goal is to recognize the few points in time where the student's attempted a certain pose and contrast the student's execution to the teacher's demonstration of the same pose.

An additional challenge is posed by the scarcity of training data. For any specific pose,

we expect only a single (correct) demonstration by a teacher and only few annotated student sequences. In other words, we are confronted with a few-shot learning setting, where we need to share as much information across poses as possible to learn a viable model. In particular, we draw upon prior work on prototype networks [5] and metric learning [6]. Prototype networks would map each frame of student motion x_t at time t and each teacher demonstration w_k for pose k to auxiliary representations $f(x_t)$ and $f(w_k)$, and detect pose k at time t if the distance between $f(x_t)$ and $f(w_k)$ is low.

By contrast, metric learning is concerned with finding a distance function d such that the distance $d(x, y)$ is low if x and y should be close and large if x and y should not be close [6]. In our setting, we wish to learn a distance such that $d(x_t, w_k)$ is small if the student attempted pose k in frame t and large, otherwise. A special kind of metric learning is to learn weights α_l for each dimension l. If these weights are non-negative, such a weighting becomes equivalent to an attention mechanism, where we interpret a large α_l as paying attention to dimension l, whereas small α_l means that dimension l is unimportant for the current decision [7].

Our contribution in this work is that we combine prototypical networks with an attention mechanism for the purpose of keypose recognition, where such techniques have not yet been applied, to the best of our knowledge.

2. Method

Our task is to recognize times t when a student tried to execute a certain pose k. More specifically, we record the motion of a student with a Kinect camera, yielding a time series $x_1, x_2, \ldots, x_T$, where each x_t is a 26×3 matrix, storing the 3D position of 26 joints of the human body at frame t. We wish to compare the student's frames to expert demonstrations $w_1, \ldots, w_K$ for each of the K poses and recognize frames where the student attempted a pose. Our basic approach is to compute some distance $d(x_t, w_k)$ between student frames and teacher demonstrations and recognize pose k at time t, whenever $d(x_t, w_k) < \theta$ for some threshold θ. However, the Euclidean distance on the raw 3D positions is not suitable because it would be disturbed by differences in body size and orientation, as well as deviances in joints that are irrelevant for the specific pose (Fig. 1, a).

To account for body size and orientation differences, we first translate the time series to angular space, in particular the azimuth and elevation of a joint compared to its parent joint in the human skeleton (Fig. 1, b). Further, we apply learnable weights $\alpha_{k,l}$ for each pose k and each joint l, resulting in the distance $d(x_t, w_k)^2 = \sum_{l=1}^{26} \alpha_{k,l} \cdot \|x_{t,l} - w_{k,l}\|^2$ (Fig. 1, c).

To learn the weights $\alpha_{k,l}$, we require training data where the correct matching between student and teacher is known. In particular, assume that we have one time series $x_1^i, \ldots, x_{T_i}^i$ per student i. For each of these time series, assume that experts provided a label sequence $y_{1,k}^i, \ldots, y_{T_i,k}^i$ for each pose k, where $y_{t,k}^i = 1$ if pose k should be recognized at time t, $y_{t,k}^i = 0$ if pose k should *not* be recognized at time t, and $y_{t,k}^i = -1$ if we consider it irrelevant whether pose k is recognized at time t or not.

a) Euclidean distance   b) angular distance   c) weighted ang. distance   d) embedding distance
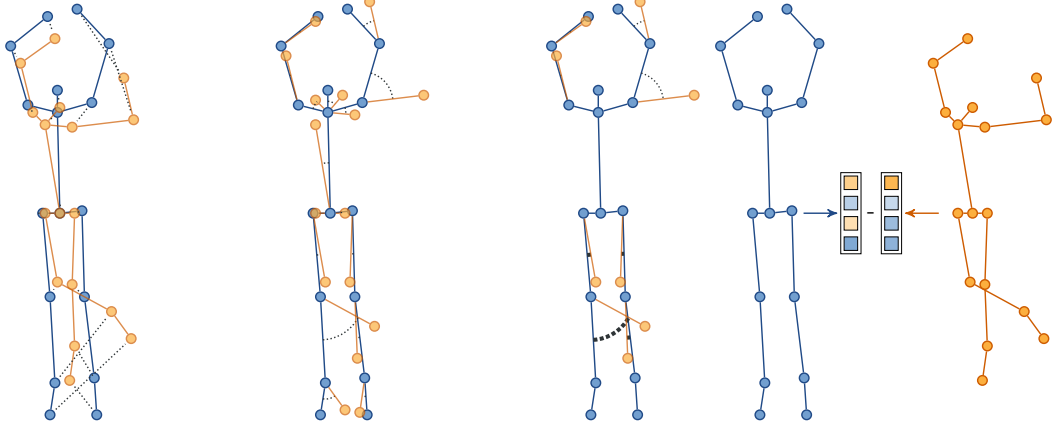
Figure 1: An illustration of different ways of computing distances between student poses x_t (blue) and teacher poses w_k (orange). a) Euclidean distance, b) angular distance, c) weighted angular distance (relevance learning / attention net), d) embedding distance (prototype net, prototype attention net).
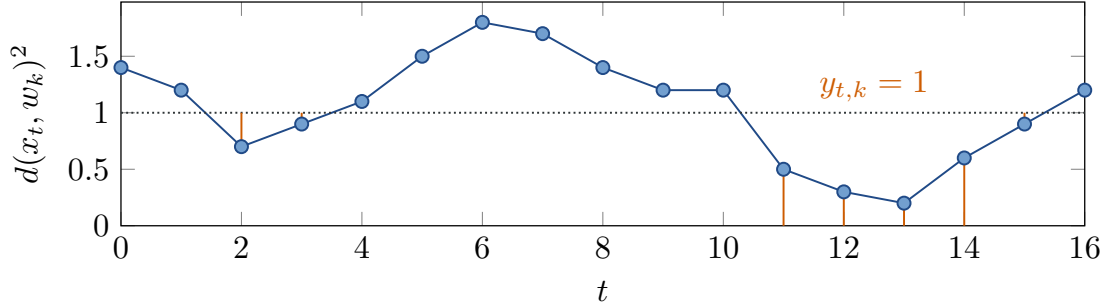

Figure 2: An illustration of the contrastive loss (1). If $y_{t,k} = 1$, we punish any distance larger zero (orange lines at 11, 12, 13, 14). Otherwise, we punish distances smaller than 1 (orange lines at 2, 3, and 15).

Given this kind of training data, we learn the weights $\alpha_{k,l}$ by minimizing the contrastive loss

$$\ell = \sum_i \sum_{k=1}^{K} \sum_{t:y_{t,k}^i=1} d(x_t^i, w_k)^2 + \sum_{t:y_{t,k}^i=0} \left[1 - d(x_t^i, w_k)^2\right]_+, \tag{1}$$

where $[1 - d]_+ = \max\{0, 1 - d\}$. This loss punishes large distances $d(x_t^i, w_k)$ if $y_{t,k}^i = 1$ and distances below 1 if $y_{t,k}^i = 0$ (Fig. 2). In other words, this loss tries to ensure that we can recognize pose k correctly by checking if $d(x_t, w_k)$ is smaller than $\theta = 1$. We can optimize this loss by standard, gradient-based non-linear techniques, such as L-BFGS. We call this approach *relevance learning*, in line with [8].

Note that this scheme is, essentially, a simple metric learning scheme [6, 8]. It is also nicely intepretable because we can inspect the learned weights $\alpha_{k,l}$ and check whether they make

sense to a domain expert. Further, we can provide feedback by highlighting joints k to students where the weighted deviation $\alpha_{k,l} \cdot \|x_{t,l} - w_{k,l}\|^2$ is large. For example, imagine a virtual mirror with an avatar of the student where the avatar's joint k is color-coded as red, similar to the scheme of [2].

Unfortunately, relevance learning is limited to situations where the set of poses is fully known. For every new pose k, we need to record new training data to train new weights $\alpha_{k,l}$, which may be infeasible. Instead, we would prefer an approach which can be applied to new poses without any re-training. To that end, we apply a two-layer feedforward neural network g which receives the expert demonstration w_k of a pose as input and maps it to weights $\alpha_{k,l} = g_l(w_k)$. By applying a sigmoid nonlinearity, we ensure that the weights remain in the range $[0, 1]$. This is, essentially, an attention mechanism, which have become popular for speech processing [7]. We call this approach *attention net* and train it with the same contrastive loss (1) as before.

In addition to the attention mechanism, we also test the effect of a refined motion representation. In particular, we use a 1D convolution with kernel length 31 and 32 filters, followed by a sigmoid nonlinearity and a linear layer which reduces the 32 filter dimensions to a single number[1]. Finally, we apply another sigmoid nonlinearity and a linear layer from the 26 joints to n latent dimensions which integrates information across joints. Overall, we obtain an n-dimensional representation $f(x_t)$ of each motion frame x_t. We compare to the correct demonstration $f(w_k)$ via the distance $d(x_t, w_k)^2 = \sum_{l=1}^{n} g_l(w_k) \cdot \left(f_l(x_t) - f_l(w_k)\right)^2$ (Fig. 1, d). We learn all neural network parameters by minimizing loss (1), but we add a regularization term $\lambda \cdot \sum_{t=1}^{T_i-1} \|f(x_t^i) - f(x_{t-1}^i)\|^2$ to ensure that the learned motion representation is smooth over time. We call this approach *prototype attention net* because it integrates the concept of the attention net with the representation approach of prototypical networks [5].

As a final model, we also consider a *prototype net* where we omit the attention net and set the weights $g_l(w_k)$ to 1, instead.

3. Experiments

We compare relevance learning, attention net, prototype net, and prototype attention net on a dataset of 27 students and one teacher, all executing a series of 25 fitness and dance motion elements (namely squat, raise arms (2x), tree squat, "pirate" (6x), elbow to knee (3x), crossover (2x), airplane, airplane squat, standing, lunge (4x), T-pose, clock (2x)) while being recorded with a Kinect camera. We had 15 female and 12 male participants with mean age 27 (std. 4.49 years). Table 1 displays the self-reported prior experience of the participants with sports in general and video tutorials in particular.

We note that the annotators only annotated the first frame where the students attempted a certain pose. To arrive at complete labels $y_{t,k}^i$, we used a heuristic scheme where we automatically labeled 30 frames after the actual annotation with $y_{t,k}^i = 1$ and extended the annotation further as long as the 3D marker positions did not change beyond 5 times the average distance between adjacent frames. We further set $y_{t,k}^i = -1$ for the 30 frames before and after the annotated

---

[1]32 was chosen as the next power of 2 above the number of keyposes, which was 25. However, future work could investigate more hyperparameter combinations.

Table 1
The number of participants with a specific level of self-reported prior experience with sports (top row) and video tutorials (bottom row).

	Expert	Advanced	Some	None
Sports	6	5	11	5
Video	7	13	6	1

Table 2
The average evaluation measures $\pm$ standard deviation across poses

model	recall	precision	F1	AUC
relevance learning	$\mathbf{0.81 \pm 0.13}$	0.18 ± 0.09	0.28 ± 0.12	0.53 ± 0.23
attention net	0.60 ± 0.21	0.38 ± 0.15	0.43 ± 0.17	0.49 ± 0.22
prototype net	$\mathbf{0.81 \pm 0.16}$	0.53 ± 0.14	0.61 ± 0.15	0.72 ± 0.18
prototype attention net	$\mathbf{0.81 \pm 0.14}$	$\mathbf{0.56 \pm 0.12}$	$\mathbf{0.64 \pm 0.14}$	$\mathbf{0.74 \pm 0.16}$

region to ignore cases where the students was already/still close to the target pose but not close enough to count as training data.

Even after this preprocessing, though, our dataset is highly imbalanced: $y_{t,k}^i = 1$ is relatively rare, whereas $y_{t,k}^i = 0$ is common. Accordingly, we do not report accuracy but recall, precision, and F1 score, as well as the area under the precision-recall curve (AUC).

Table 2 shows the results. As we can see, the prototype attention net performs best, according to all measures. A Wilcoxon signed-rank test revealed that the AUC for relevance learning and the attention net were both significantly lower ($p < 10^{-3}$) but the AUC of prototype net and prototype attention net was statistically indistinguishable. This finding indicates that representation learning is more crucial than joint weighting in achieving good keypose detection results, at least on this particular dataset.

4. Conclusion

We considered the problem of keypose detection for a sequence fitness and dance motion elements in a few-shot setting, where only a single teacher demonstration and few student demonstrations per pose exist. To detect keyposes, we evaluated methods which compute a distance between teacher demonstrations and student frames and detect a keypose if the distance is below 1. We compared several schemes to arrive at a distance, namely (1) relevance learning, which optimized joint weights for each keypose, (2) an attention neural net which inferred the joint weights from the respective teacher demonstration, (3) a prototype network which represented both teacher and student motion in a latent space before computing distance, and (4) a combination of prototype and attention network. As expected, the prototype attention net (4) performed best but we found that the prototype net (3) performed nearly as well. Therefore,

we conclude that representation learning is more crucial compared to attention, at least for our example.

In future work, it should be investigated how well a prototype network generalizes to new keyposes it was not trained on and how far performance can be improved with refined architectures. Beyond keypose detection, future work should investigate the ability to recognize entire motion, in addition to purely static poses. For all these future research opportunities, we believe that our proposed loss function and training scheme can contribute to robust detection approaches, which in turn can become a crucial component for new feedback methods in psychomotor learning.

Acknowledgments

References

[1] S. Anu, V. Ele, Teaching dance in the 21st century: A literature review, The European Journal of Social & Behavioural Sciences 7 (2013) 624–640. doi:`10.15405/ejsbs.2013.7.issue-4`.

[2] F. Hülsmann, C. Frank, I. Senna, M. O. Ernst, T. Schack, M. Botsch, Superimposed skilled performance in a virtual mirror improves motor performance and cognitive representation of a full body motor action, Frontiers in Robotics and AI 6 (2019) 43. doi:`10.3389/frobt.2019.00043`.

[3] B. Paaßen, M. Kravčík, Teaching psychomotor skills using machine learning for error detection, in: R. Klemke, K. Asyraaf Mat Sanusi, et al. (Eds.), Proceedings of the 1st International Workshop on Multimodal Immersive Learning Systems (MILeS 2021), 2021, p. 8–14. URL: http://ceur-ws.org/Vol-2979/paper1.pdf.

[4] T. K. Vintsyuk, Speech discrimination by dynamic programming, Cybernetics 4 (1968) 52–57. doi:`10.1007/BF01074755`.

[5] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: I. Guyon, U. V. Luxburg, S. Bengio, et al. (Eds.), Proc. NeurIPS, 2017, pp. 4077–4087.

[6] A. Bellet, A. Habrard, M. Sebban, A survey on metric learning for feature vectors and structured data, arXiv 1306.6709 (2014).

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, et al. (Eds.), Proc. NeurIPS, 2017, pp. 5998–6008.

[8] B. Hammer, T. Villmann, Generalized relevance learning vector quantization, Neural Networks 15 (2002) 1059–1068. doi:`10.1016/S0893-6080(02)00079-5`.