Self-supervised Discovery of Human Actons from Long Kinematic Videos

Anonymous authors
Paper under double-blind review
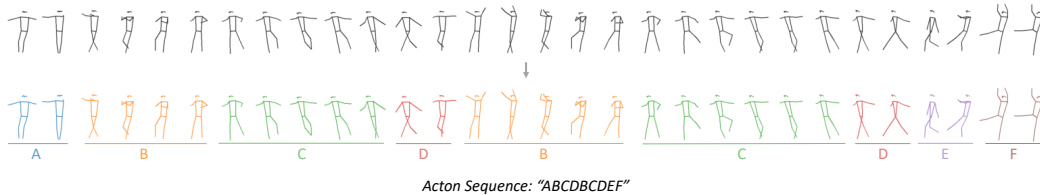
Acton Sequence: "ABCDBCDEF"

Figure 1: We present a self-supervised technique for discovering recurring temporal patterns, called *actons*, in long kinematic sequences like human dance. From a collection of such videos without any annotations, we extract a set of actons and use this lexicon to segment and model motion sequences as shown above (skeleton sequence temporally downsampled by $10\times$).

Abstract

For human action understanding, a popular research direction is to analyze short video clips with unambiguous semantic content, such as jumping and drinking. However, methods for understanding short semantic actions cannot be directly translated to long kinematic sequences such as dancing, where it becomes challenging even to semantically label the human movements. To promote analysis of long videos of complex human motions, we propose a self-supervised method for learning a representation of such motion sequences that is similar to words in a sentence, where videos are segmented and clustered into recurring temporal patterns, called actons. Our approach first obtains a frame-wise representation by contrasting two augmented views of video frames conditioned on their temporal context. The frame-wise representations across a collection of videos are then clustered by K-means. Actons are then automatically extracted by forming a continuous motion sequence from frames within the same cluster. We evaluate the self-supervised representation by temporal alignment metrics, and the clustering results by normalized mutual information and language entropy. We also study an application of this tokenization by using it to classify dance genres. On the AIST++ and PKU-MMD datasets, actons are shown to bring significant performance improvements compared to several baselines.

1 Introduction

In the past decade, computer vision research owes much of its its success to the construction of large labeled datasets. The labels in the datasets provide semantics that associate visual data with natural language descriptions. For human action understanding, the data is crowd-sourced with semantic descriptions that cover single-person movement, multi-person movement, as well as person-object interactions (Goyal et al., 2017; Kay et al., 2017; Abu-El-Haija et al., 2016). Labeling human actions by semantic descriptions, however, limits the scope of the action space, as not all human actions can be clearly and unambiguously defined by language.

Long and complex human actions in videos are difficult to depict by semantic language descriptions. We consider dance in this work, but this also applies to many other domains such as sports and opera. Though there exist a few iconic movements that are associated with a clear label, such as "the moonwalk" by Michael Jackson, a majority of dance movements are best conveyed by showing examples. Due to this complexity, a written transcription of dance uses a figure-based notation system called Labanotation (Hutchinson et al., 1977) to record human motions.

In this work, we raise the question of how to learn meaningful representations of a long human dance without any supervision. We observe that, though complex over a long time period, human dances
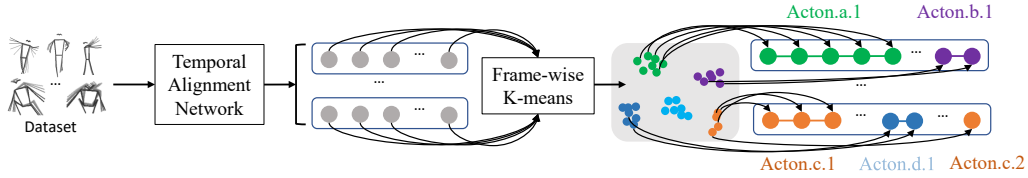
Figure 2: Overview of our acton discovery framework. First, a novel Temporal Alignment Network (TAN) through self-supervised contrastive learning is used to extract frame-wise, context-aware and temporally aligned features. Then, frame-wise K-means is used to cluster and segment long human dances into actons.

often exhibit recurring temporal patterns. We thus propose an atomic representation called *actons* as a mid-level representation for modeling long human action videos. Given this mid-level representation, human dances could be represented by a sequence of acton tokens, similar to words of a lexicon in a sentence. Building this acton dictionary may facilitate further applications such as action retrieval, dance translation, and motion stylization.

We propose a two-step self-supervised approach for clustering and segmenting long human dances into actons. An overview is illustrated in Figure 2. First, we obtain a frame-level representation by training a Temporal Alignment Network (TAN) through contrastive learning (Chen et al., 2020). Given a 3D skeleton sequence, we augment the sequence into two views by considering rotation, translation and speed augmentations. Frames in one view are mapped to corresponding frames in the other view as positives. Contrastive learning then discriminates between positive and negative pairs at the frame level using a Transformer backbone. TAN is discussed in Section 3.

With the pretrained TAN, the embeddings can be used to retrieve frames in a similar action context. We therefore propose a simple way to discover actons, by jointly clustering all frames in a video dataset by K-means. Actons are then automatically segmented from the long actions by finding continuous sequences within the same cluster assignment. Unsupervised lexicon building is discussed in Section 4.

We conduct experiments on AIST++ and PKU-MMD datasets (Tsuchida et al., 2019; Li et al., 2021; Liu et al., 2017), which provide 3D skeletons from multi-view reconstruction. AIST++ dataset contains dances of 10 genres with basic motions and advanced complex motions; PKU-MMD is an action detection dataset contains 1076 long video sequences. Alignment quality is measured by Kendall's Tau (Kendall, 1938; Dwibedi et al., 2019), and clustering performance is evaluated by normalized mutual information and language entropy. We also present two applications of the acton representation, genre classification and action detection. Extensive experimental results demonstrate the effectiveness of the method against established baselines, TCN and TCC (Dwibedi et al., 2019; Sermanet et al., 2018). We summarize the contributions of this work as follows:

- This work raises a novel question of how to discover meaningful elementary representations from a long human action sequence **without any supervision**, namely, unsupervised acton discovery. Analogous to words in a language, the discovered actons can serve as words in a body language, which then can be further used for action assessment, translation and composition. This is valuable to domain specialists like choreographers, athletes and so on. To our knowledge, this is the first time that this interesting, important and challenging task has been presented, discussed, explored and evaluated in depth.

- We believe the biggest technical challenges of this task lie in learning strong frame-wise, context-aware and temporally aligned features. To this end, we propose a novel Temporal Alignment Network (TAN) together with effective speed augmentation and negative sampling techniques to address the above issues.

- Extensive experiments, including ablation studies on the augmentation and negative sampling strategies, as well as comparison with the state-of-the-art self-supervised sequence learning methods Time-Contrastive Networks (TCN) (Dwibedi et al., 2019) and Temporal Cycle-Consistency (TCC) (Sermanet et al., 2018), demonstrate the effectiveness of the proposed method.

- We use the discovered actons as tokenized features and apply them to the application of genre classification, action detection and choreography based on random acton composition.

2 RELATED WORKS

Self-Supervised Learning for Action Recognition   An effective representation for action recognition should capture the temporal order of human movement, the speed of movement, as well as the detailed motions. For self-supervised learning, pretext tasks which encode such representations include learning the arrow of time (Wei et al., 2018), shuffle and learn (Misra et al., 2016; Lee et al., 2017), and learning the temporal speed (Benaim et al., 2020). Recent advances on contrastive learning for image representation learning (Wu et al., 2018; Chen et al., 2020) also show promising results on videos. Research along this direction has achieved the state-of-the-art by contrasting on the temporal dimension (Gordon et al., 2020; Sermanet et al., 2018), distilling motion representations (Han et al., 2020), and designing better video augmentations (Qian et al., 2020).

The Kinetics human action video dataset (Kay et al., 2017) is the most popular dataset for pretraining action representations. However, the dataset is curated since each video is trimmed to temporally focus on the underlying action. The holistic representations for action recognition learned from Kinetics cannot be scaled to long kinematic videos that describe a series of complex motions.

Video Alignment   In contrast to holistic representations, a frame-level action representation may effectively capture temporal action progressions. For example, the act of pouring may consist of sub-action phases such as grabbing a bottle, tilting it to allow liquid to flow out, and putting down the bottle. The transformation which maps the frames of one video to another can be learned by CCA (Andrew et al., 2013) and Gaussian mixture (Sener & Yao, 2018; Swetha et al., 2021) losses. Self-supervised approaches to alignment include cycle consistency (Dwibedi et al., 2019) and time-contrastive networks (Sermanet et al., 2018). The video alignment task assumes that all data in the same action category are loosely aligned. Our task is much more challenging because the training data is uncurated, spanning multiple categories of actions.

Unsupervised Pattern Discovery from Sequences   Natural sensory signals, of audio, vision or text, often come in the form of long sequences. An important research topic (Oates, 2002) is to discover recurring patterns in the sequences. Early works (De Marcken, 1996; Park & Glass, 2007) show that it is possible to extract words and linguistic entities from audio data without any supervision. Further research (Myers & Rabiner, 1981) uses recurring words for connected word recognition. Recently, a zero resource speech challenge (Nguyen et al., 2020) has demonstrated that neural language models can be successfully learned from speech recordings. Pattern discovery models have also been applied to discover meaningful patterns in music (Dannenberg & Hu, 2003). Research on bioinformatics (Brazma et al., 1998; Rigoutsos & Floratos, 1998) has found structurally important gene patterns (motifs) in DNA sequences. The key method for pattern discovery is to perform sequence to sequence alignment, e.g. using dynamic time warping (Sakoe & Chiba, 1978), and aggregate aligned sequences into clusters.

For applications of unsupervised pattern discovery on videos, topic models such as latent Dirichlet allocation are applied to surveillance videos to discover recurrent activities on crossroads (Emonet et al., 2011; Hospedales et al., 2009; Emonet et al., 2013; Varadarajan et al., 2013). The alignment between short video clips is measured by the similarity of features, which often use histograms of optical flow vectors. Instead of traffic activities, the goal of our work is to analyze human motions, which are of significantly greater complexity.

Atomic Actions   In contrast to action classification, the study of atomic actions aims to provide a detailed representation of complex action sequences. With its combinatorial structure, such a representation can also be used to reduce the complexity of recognition systems. In pursuit of this, Breakfast (Kuehne et al., 2014) and MPII-Cooking 2 (Rohrbach et al., 2016) annotate 48 and 67 acton classes, respectively, over various cooking activities. FineGym (Shao et al., 2020) provides a fine-grained temporal annotation for gymnastic actions with up to 530 acton classes, which are organized into three semantic and two temporal hierarchies. Methods have been proposed for unsupervised segmentation of complex activities (Sener & Yao, 2018; Swetha et al., 2021), but they presume knowledge of acton classes and canonical orders. There has been an effort to collect actons in a semi-supervised manner (Zhu et al., 2013), which is driven by the goal of action classification, while our actons are learned in a task-agnostic manner and may be used in a broad range of applications.

Despite the expense to annotate them, acton classes are generally not shared between activities, e.g., between cooking and gymnastics. Therefore, a method for simultaneous self-supervised acton discovery and segmentation is needed to facilitate analysis of complex human actions.

3 TEMPORAL ALIGNMENT NETWORK

We observe that recurring temporal patterns, i.e. *actons*, occur in dance collections across multiple dancers and genres. Though the repeated patterns are essentially the same actions, they may vary in speed, rotation, range of motion, as well as other properties not intrinsic to the action itself. Therefore, we first seek to learn, in a self-supervised fashion, a representation that is effective for clustering actons.

Inspired from prior works that discover *motifs* in the discrete domain for bioinformatics (Brazma et al., 1998; Rigoutsos & Floratos, 1998) and in the continuous domain for audio (De Marcken, 1996; Nguyen et al., 2020; Dannenberg & Hu, 2003; Park & Glass, 2007), the task of acton discovery uniquely calls for the following properties in representation learning:

- Frame-wise features. Unlike action recognition where features in the temporal dimension are usually globally pooled for an overall representation, the features in this task should be at the local frame level since the boundaries of actons have yet to be determined.
- Context-aware. However, the feature of a frame should still represent pose status in the context of a particular motion. For example, the same pose that appears in different motions should have different feature representations.
- Temporal alignment. An acton can generally be performed at different speeds, so the learned feature is expected to help identify the same acton regardless of speed. The feature distance of the same frame at different motion speeds should be close to facilitate temporal alignment.
- Intra-compactness and Inter-separability. A key to effective clustering is to learn discriminative features that enlarge the decision margins between clusters while reducing the variations within each cluster (Liu et al., 2016). Intra-class compactness and inter-class separability between learned features should therefore be encouraged.
- Temporal continuity. The frames of an acton should be continuous in time. To facilitate acton segmentation, the noise among frames within a potential acton should be suppressed by drawing the features of adjacent frames closer together.

Formally, given a long kinematic video $\boldsymbol{x} \in \mathbb{R}^{T \times 3J}$, our goal is to learn a neural network encoder $f(\cdot)$ that extracts representation vectors $\boldsymbol{z} \in \mathbb{R}^{T \times F}$ from $\boldsymbol{x}$,

$$\boldsymbol{z} = f(\boldsymbol{x}). \tag{1}$$

Here, T and J are the number of video frames and human joints respectively and F is the dimension of latent feature vectors. For encoding the representation network $f(\cdot)$ of pose sequences in our task, we adopt the Transformer encoder network (Vaswani et al., 2017) for its superior performance on recent vision applications (Girdhar et al., 2019; Lohit et al., 2019; Plizzari et al., 2020; Carion et al., 2020; Zou et al., 2021). Please see Appendix A for more implementation details.

3.1 FRAME-WISE TEMPORAL CONTRASTIVE LEARNING

As illustrated in Figure 3, we follow the framework of SimCLR (Chen et al., 2020), which learns representations by maximizing agreement between augmented views of the same instance via a contrastive loss. All frame representations (for example, $\boldsymbol{z}(i)$ for the i-th frame) are transformed by an MLP projection head $h(\cdot)$ to a latent vector $\boldsymbol{v}(i)$ on a F-dim L_2 unit sphere,

$$\boldsymbol{v}(i) = \frac{h(\boldsymbol{z}(i))}{\|h(\boldsymbol{z}(i))\|}. \tag{2}$$

As we seek here to learn frame-wise representations, we consider a frame conditioned on its temporal action context as an individual instance. Given a minibatch of N video clips $\{\boldsymbol{x}_n\}$, denote the i-th frame in the video $\boldsymbol{x}_n$ as $\boldsymbol{x}_n(i)$. We augment the sequence into two views by the same family of
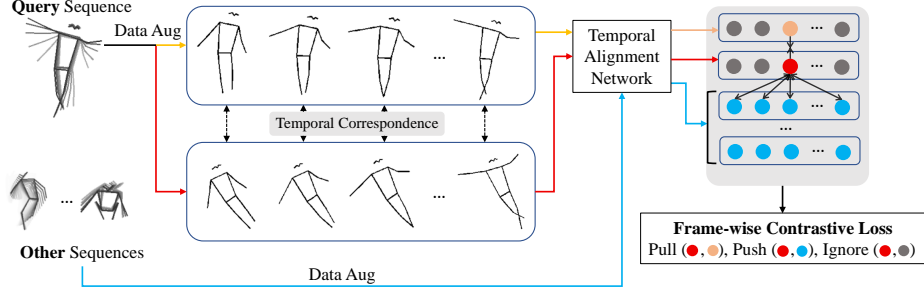
Figure 3: Illustration of Temporal Alignment Network (TAN), the first part of our system. Two separate augmentations are sampled from the an augmentation family $\mathcal{T}$ (*speed, rotation and translation*) to obtain two correlated views probably of different lengths. Ground-truth correspondence between frames of the two views is calculated. All frames are projected into a common embedding space, where a *frame-wise* contrastive loss is minimized.

augmentations, $\mathcal{T}$, composed of random translation, rotation around the gravity direction, and speed change. In order to preserve temporal smoothness, the random augmentations are applied *consistently throughout a video*. As a result, the frame $\boldsymbol{x}_n(i)$ is augmented into $\boldsymbol{x}'_n(i)$ and $\boldsymbol{x}''_n(i)$ in the two views. We use $\boldsymbol{v}'_n(i)$ and $\boldsymbol{v}''_n(i)$ to denote the latent embeddings for the two frames. Then, the contrastive loss function for a positive pair of frame examples is formally defined as:

$$\mathcal{L}_n^i = -\log \frac{\exp(\boldsymbol{v}'_n(i) \cdot \boldsymbol{v}''_n(i))/\tau)}{\exp(\boldsymbol{v}'_n(i) \cdot \boldsymbol{v}''_n(i)/\tau) + \sum\limits_{x_k(j) \in D_n^i} \exp(\boldsymbol{v}'_n(i) \cdot \boldsymbol{v}_k(j)/\tau)}, \tag{3}$$

where τ is the temperature parameter. It is worth noting that our contrastive instance ($\boldsymbol{z}_n(i) = f(\boldsymbol{x}_n)[i]$) differs from those in standard image or video based contrastive methods whose features are learned either from a single image ($\boldsymbol{z}_n(i) = f(\boldsymbol{x}_n[i])$) (Chen et al., 2020) or to represent the entire video ($\boldsymbol{z}_n = f(\boldsymbol{x}_n)$) (Feichtenhofer et al., 2021).

Due to random augmentations, some of the frames in one view may no longer have correspondences in the other view, such as when speed augmentation changes the sequence length. We simply neglect these frames during contrastive learning. The notation $\mathcal{D}_n^i$ is used to denote the set of negative samples for the frame instance $\boldsymbol{x}_n(i)$. We symmetrize the loss in Eq. 3 by swapping $\boldsymbol{x}'_n(i)$ and $\boldsymbol{x}''_n(i)$ to compute $\widetilde{\mathcal{L}}_n^i$. The overall contrastive loss is formulated as:

$$\mathcal{L}^{TAN} = \frac{1}{2NT} \sum_{n=1}^{N} \sum_{i=1}^{T} (\mathcal{L}_n^i + \widetilde{\mathcal{L}}_n^i). \tag{4}$$

Negative Samples For typical instance discrimination, all instances other than the reference are considered as negatives. However, for frame-wise representation learning, frames that are temporally close or in the same video may not be appropriate choices as negatives. We consider the following alternatives for negative pair selection.

- All frames in batch: all frames from every video in one batch other than the reference frame are considered negatives. In this case, $\mathcal{D}_n^i = \{\boldsymbol{x}_k(j) \mid k \neq n \text{ or } j \neq i\}$. This is a straightforward extension of instance discrimination to frame-level representation learning.

- Excluding close frames: ignore negative samples within the same video clip. In this case, $\mathcal{D}_n^i = \{\boldsymbol{x}_k(j) \mid k \neq n\}$. Frames within a video are usually highly correlated with one another. Ignoring them as negatives may help to learn a smooth representation over time.

3.2 EVALUATION METRICS

To evaluate the performance of the learned representations for temporal alignment, we use the Kendall's Tau (Kendall, 1938; Dwibedi et al., 2019) metric, which is a statistical measure to determine how aligned in time two sequences are in a representation space. Given a pair of sequences that perform the same action in different dance recordings with possibly different motion speed, Kendall's Tau is calculated over every pair (u_i, u_j) of frames with $i < j$ in the first video. For all these pairs,

we retrieve their nearest frames (v_p, v_q) in the second video with respect to the learned representation space. A pair is called concordant if $p < q$, and otherwise it is discordant. Kendall's Tau is defined over all pairs of frames in the first video as:

$$\tau = \frac{\text{\# concordant pairs - \# discordant pairs}}{n(n-1)/2}. \tag{5}$$

The metric ranges from -1 to 1, indicating completely reversed and aligned, respectively.

4 UNSUPERVISED LEXICON BUILDING

The second step for acton discovery is lexicon building, which involves two correlated task: segmentation and clustering of unlabeled video sequences.

Earlier methods (De Marcken, 1996; Park & Glass, 2007) mainly seek *part-coverage*, which discovers isolated segments in sequences and excludes many background frames. In another problem formulation, namely *full-coverage*, entire sequences are segmented and clustered into word-like units (Kamper et al., 2016; 2017b). In this setting, these two tasks can be treated as a joint minimization problem of the sum of distance of each segment to their corresponding cluster center. In speech processing, there is a similar problem called unsupervised word discovery (Nguyen et al., 2020) and often addressed using probabilistic Bayesian models (Kamper et al., 2017a;b). Recently, K-Means has been incorporated into these Bayesian models to ease the computational burden (Kamper et al., 2017a; Nguyen et al., 2020).

4.1 ACTON SEGMENTATION AND CLUSTERING

After representation learning, we apply K-Means clustering on the output Transformer backbone. The clustering is performed over all the output features at every time step of all the videos in a given training set, following (Nguyen et al., 2020). The number of actons in the lexicon is determined by the number of clusters in K-Means. For inference, a nearest neighbour search is conducted among the K cluster centers.

After a lexicon is built, we consider the task of segmenting long video sequences. A simplistic method is to consider a segment as a group of consecutive frames that share the same acton cluster, as shown in Figure 1. We find that our feature representation works well with this simple segmentation method. We note that ignoring negative samples within the same video clip plays a vital role in forming smooth and clean segments.

4.2 EVALUATION METRICS

Normalized Mutual Information (NMI) is a widely used metric for estimating the quality of clustering (Witten & Frank, 2002). Let Y be a random variable that describes the event of a testing data sample being one of the ground truth actons, while C is another random variable that describes the event of a sample belonging to one of the clusters. Mutual Information measures how much knowing one of the variables reduces uncertainty about the other. For example, if knowing that a sample belongs to a cluster determines its ground truth label ($\text{H}(Y|C) = 0$, where $\text{H}(\cdot)$ denotes the entropy of a variable), then all information conveyed by C is shared with Y; the Mutual Information is the same as the uncertainty contained in Y. On the other hand, if C and Y are independent, namely knowing that C does not give any information about Y ($\text{H}(Y|C) = \text{H}(Y)$), then Mutual Information is zero. Formally,

$$\text{NMI}(Y, C) = \frac{2(\text{H}(Y) - \text{H}(Y|C))}{\text{H}(Y) + \text{H}(C)}. \tag{6}$$

The NMI is normalized by the sum of entropy in Y and C, ranging from 0 to 1. A higher NMI value indicates better clustering results.

Language Entropy is a statistical measurement on the average uncertainty (conditional entropy) of the letter (in our case, acton), when the preceding N-1 letters are known. Specifically, let W_N represent a block of contiguous actons $(w_1, w_2, ..., w_N)$ and $\mathcal{W}_N$ represent all possible progressions

Table 1: Comparison of different augmentation approaches (left) and negative samples (right) on AIST++.

Aug. Method	$\tau \uparrow$	NMI $\uparrow$	$F_2 \downarrow$
full aug.	**0.80**	**0.79**	**0.81**
w/o speed aug.	0.77	0.74	0.90
w/o rotation aug.	0.72	0.58	1.66
w/o translation aug.	0.76	0.76	0.82

Negative Samples	NMI $\uparrow$	$F_2 \downarrow$
All frames in batch	0.35	2.29
Excluding close frames	**0.79**	**0.81**

of W_N in the given sequences. The entropy K_N and the conditional entropy F_N in N contiguous actons are defined as:

$$K_N = - \sum_{W_N \in \mathcal{W}_N} p(W_N) \log p(W_N) \tag{7}$$

$$F_N = K_N - K_{N-1} = - \sum_{W_N \in \mathcal{W}_N} p(W_N) \log p(w_N | W_{N-1}). \tag{8}$$

Finally, the language entropy in (Shannon, 1951) is defined as $H = \lim_{N \to \infty} F_N$.

We evaluate language entropy on testing sequences of basic choreography in the dataset, where each sequence forms only a single handcrafted labanotation token but recurs several times. Hence, the language entropy for these sequences is expected to be low. In the experiment, lower language entropy indicates better results under this setting. We observe consistent results between methods when $N \to \infty$, so we use 2-gram evaluation F_2 instead of H in the experiments to simplify computation.

5 EXPERIMENTS

5.1 DATASETS

AIST++ For our experiments, we use AIST++ Dance Motion Dataset (Li et al., 2021)[1], which contains 3D human keypoint annotations estimated for the AIST Dance Video Database (Tsuchida et al., 2019). AIST++ contains 1,362 sequences with 3D skeletons[2], evenly distributed across 10 dance genres. For each genre, ~85% of the sequences are of *basic* choreography and ~15% of them are of *advanced* choreography. Basic choreographies are repetitive while advanced choreographies are longer and more complicated dances improvised by the professional dancers. Advanced videos range from 27.4 seconds (1644 frames) to 46.7 seconds (2802 frames).

PKU-MMD PKU-MMD is a large-scale action detection dataset with 1,076 videos, which last 3 to 4 minutes and contain about 20 action instances each. There are 51 action classes, e.g. drinking, waving hand. Among other modalities, only 3D skeletons are used for our experiment. We adopted the Cross-Subject Evaluation split and report mean average precision of different actions (mAP$_a$) on the testing set with IOU threshold $\theta = 0.3$.

5.2 ABLATION STUDIES

At the stage of **representation learning**, we evaluate by Kendall's Tau. For each of some randomly chosen basic choreographies, we sample two videos of different tempos and crop the first complete actions that are found in common. When using 3D skeleton joint coordinates to retrieve the nearest frame by L_2 distance, we obtain a $\tau = 0.44$, significantly lower than using trained representations from Temporal Alignment Network. We start from our final setting and ablate different augmentation components. Results are shown in Table 1, from which the necessity of all three augmentations can be seen.

At the stage of **lexicon building**, we evaluate by NMI and language entropy averaged across 10 genres and 15 different K values from 10 to 150 with an interval of 10. We use the choreography label of each video as the ground-truth frame label for NMI calculation. After this inference, we get a reorganized text corpus, on which language entropy (approximated by F_2) is calculated.

---

[1]Annotations licensed by Google LLC under CC BY 4.0 license.
[2]Counted after removing sequences deemed to be poorly reconstructed by the AIST++ authors.

In Table 1, we observe again that all three augmentation methods are indispensable, corroborating our conclusion at the representation learning stage based on Kendall's Tau. To the right, We can see that excluding a specific part of the negative samples is beneficial to TAN.
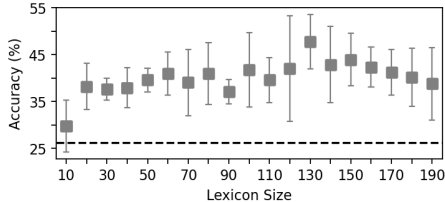
5.3 Applications

Figure 4: Mean and standard deviation over 6 trials of genre classification, plotted against raw skeleton method.

Table 2: Comparison with baseline method BLSTM (Liu et al., 2017) on PKU-MMD on mAP_a with varying θ.

θ	0.1	0.3	0.5	0.7
BLSTM	0.479	0.325	0.130	0.014
TAN (ours)	**0.490**	**0.328**	**0.132**	**0.015**

Acton discovery provides us with a discrete intermediate representation to support higher-level tasks on long kinematic videos. To demonstrate the utility of our tokenization for long kinematic video recognition, we conduct experiments to classify advanced choreography videos in AIST++ into 10 genres and do action detection on PKU-MMD. More experimental details are in the Appendix.

Genre Classification   For acton sequence input, we use the LSTM (Hochreiter & Schmidhuber, 1997) text classification model with learnable and randomly-initialized word embeddings. Test set accuracy across different lexicon sizes (K value) are presented in Figure 4. We observe that the tokenization consistently aids classification for a reasonably large K, likely because the combinatorial structure reduces requirements on recognition system capability.

Action Detection   After TAN pre-training and lexicon building, we learn a segmentation model. Our performance is compared with the best-performing method BLSTM (Liu et al., 2017) in Table 2. It shows that our method is superior to BLSTM, especially with a less strict localization requirement (smaller θ).

Action Composition   In order to use the built lexicon to guide choreography (the art of designing novel motions), we present a two-stage method. First, we generate a random list of words by thresholding the L_2 distance between the last frame of the preceding word and the first frame of the successive word. Then we randomly instantiate each word by choosing one sequence from its cluster and splice these sequences with linear interpolation. **Please check the demonstration video in Supplementary Material**.

5.4 Comparison

We compare our model with three baseline methods, two of which are prototypical methods for self-supervised sequence representation learning. We first briefly introduce them for completeness.

Raw Skeleton Coordinates (N/A)   For this method, we use 3D skeletons directly as input for clustering. To simulate the L_2 distance in the frame embedding space, we calculate the L_2 skeleton distance after spatially normalizing the body center to the origin.

Time-Constrastive Networks (TCN) (Sermanet et al., 2018)   TCN works by first sampling a certain number of anchor frames. For each of them, one positive frame within a preset time interval and one negative frame outside a larger time interval are sampled. For each motion, two augmented views are used to simulate the multi-view TC loss, where the anchor and positive frames come from the same view while negatives come from the other. After feature extraction, a triplet margin loss (Balntas et al., 2016) is used to push positive features closer to the anchor feature while pushing away negative features.

Temporal Cycle-Consistency (TCC) (Dwibedi et al., 2019)   TCC is a self-supervised representation learning method based on the task of temporal alignment between videos. Compared to TCC,

Table 3: Comparison with baseline methods. Accuracy refers to genre classification accuracy on AIST++ in Section 5.3 with K=100, averaged over 6 trials. mAP$_a$ refers to action detection score on PKU-MMD.

	AIST++			PKU-MMD		
	NMI ↑	F_2 ↓	Acc. ↑	NMI ↑	F_2 ↓	mAP$_a$ ↑
N/A	0.48	1.84	0.36	0.36	3.61	0.302
TCN	0.41	1.92	0.42	0.42	3.60	0.317
TCC	0.62	1.34	0.39	0.39	3.59	0.303
TAN	**0.79**	**0.81**	**0.46**	**0.46**	**3.50**	**0.328**

our method does not assume that the same event is occurring in all the videos in a dataset. We thus modify the loss by cycling between two augmented views of one sequence. For a frame a in the first view, we first find its (soft) nearest neighbour in the second view b, then force the nearest neighbour of b in the first view to be a via a cross entropy loss.

To compare different representation learning methods, we perform experiments by applying the same lexicon building algorithm on representations learnt from TCN, TCC, as well as using raw skeleton coordinates. For both metrics on both datasets, learnt features give better results than using raw skeleton coordinates, which justifies the use of temporal information in the representation for lexicon building. It can be seen in Table 3 that the lexicon built from our method shows better results on both clustering evaluation metrics and genre classification accuracy than the baselines. This might be attributed to the fact that these other methods do not utilize ground-truth frame correspondence between two augmented views.

5.5 QUALITATIVE RESULTS

Under the real-world setting in Section 5.3, we visualize the acton instances (short skeleton sequences) clustered into actons in Figure 5. More qualitative results can be found in Appendix F and Supplementary Material.
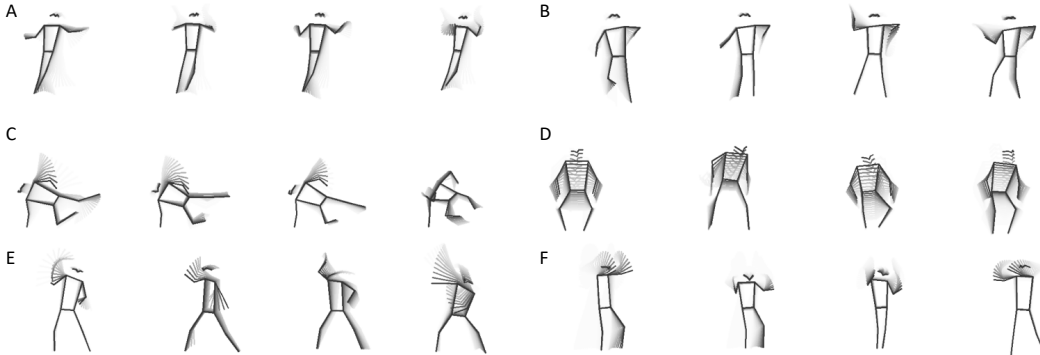
Figure 5: Visualization of discovered actons from advanced videos. For each acton, four instances are shown side to side, with temporally earlier poses fading away. At the bottom left, an acton **E** with large intra-cluster variance is shown; to its right are four acton instances facing different directions but underpinned by the same motion.

6 CONCLUSION

In this work, we assert the importance of self-supervised acton discovery for long kinematic videos. We then proposed a non-trivial pipeline composed of Temporal Alignment Network (TAN) and a Lexicon Building algorithm, and discuss some of the design choices via proposed evaluation metrics, NMI and language entropy. In the end, we successfully extract actons on completely unannotated action datasets, AIST++ and PKU-MMD, and reorganize them into a text corpus. Among many, two directions may be especially worth pursuing in the future: discovering textual motifs in a corpus to build a hierarchical lexicon (Shao et al., 2020) and building a multi-channel lexicon by decomposing the human body into parts (Hutchinson et al., 1977; Li et al., 2020).

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pp. 1247–1255. PMLR, 2013.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, pp. 3, 2016.

Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9922–9931, 2020.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Alvis Brazma, Inge Jonassen, Ingvar Eidhammer, and David Gilbert. Approaches to the automatic discovery of patterns in biosequences. *Journal of computational biology*, 5(2):279–305, 1998.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, 2008.

Roger B Dannenberg and Ning Hu. Pattern discovery techniques for music audio. *Journal of New Music Research*, 32(2):153–163, 2003.

Carl De Marcken. Unsupervised language acquisition. *arXiv preprint cmp-lg/9611002*, 1996.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1801–1810, 2019.

Rémi Emonet, Jagannadan Varadarajan, and Jean-Marc Odobez. Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model. In *CVPR 2011*, pp. 3233–3240. IEEE, 2011.

Rémi Emonet, Jagannadan Varadarajan, and Jean-Marc Odobez. Temporal analysis of motif mixtures using dirichlet processes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1): 140–156, 2013.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3299–3309, 2021.

Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 244–253, 2019.

Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5842–5850, 2017.

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Timothy Hospedales, Shaogang Gong, and Tao Xiang. A markov clustering topic model for mining behaviour in video. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 1165–1172. IEEE, 2009.

Ann Hutchinson, Ann Hutchinson Guest, and William Ambrose Hutchinson. *Labanotation: or, kinetography Laban: the system of analyzing and recording movement*. Number 27. Taylor & Francis, 1977.

Herman Kamper, Aren Jansen, and Sharon Goldwater. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):669–679, 2016.

Herman Kamper, Aren Jansen, and Sharon Goldwater. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174, 2017a.

Herman Kamper, Karen Livescu, and Sharon Goldwater. An embedded segmental k-means model for unsupervised segmentation and clustering of speech. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 719–726. IEEE, 2017b.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 780–787, 2014.

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 667–676, 2017.

Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5226–5234, 2018.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021.

Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 382–391, 2020.

Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, pp. 7, 2016.

Suhas Lohit, Qiao Wang, and Pavan Turaga. Temporal transformer networks: Joint learning of invariant and discriminative time warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12426–12435, 2019.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pp. 527–544. Springer, 2016.

C Myers and L Rabiner. A level building dynamic time warping algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):284–297, 1981.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv preprint arXiv:2011.11588*, 2020.

Tim Oates. Peruse: An unsupervised algorithm for finding recurring patterns in time series. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pp. 330–337. IEEE, 2002.

Alex S Park and James R Glass. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. *arXiv preprint arXiv:2012.06399*, 2020.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *arXiv preprint arXiv:2008.03800*, 2020.

Isidore Rigoutsos and Aris Floratos. Combinatorial pattern discovery in biological sequences: The teiresias algorithm. *Bioinformatics (Oxford, England)*, 14(1):55–67, 1998.

Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119(3):346–373, 2016.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Fadime Sener and Angela Yao. Unsupervised learning and segmentation of complex activities from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8368–8376, 2018.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1): 50–64, 1951.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2616–2625, 2020.

Sirnam Swetha, Hilde Kuehne, Yogesh S Rawat, and Mubarak Shah. Unsupervised discriminative embedding for sub-action learning in complex activities. *arXiv preprint arXiv:2105.00067*, 2021.

Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pp. 501–510, Delft, Netherlands, November 2019.

Jagannadan Varadarajan, Rémi Emonet, and Jean-Marc Odobez. A sequential topic model for mining recurrent activities from long term video logs. *International journal of computer vision*, 103(1): 100–126, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8052–8060, 2018.

Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *European Conference on Computer Vision*, pp. 507–523. Springer, 2020.

Jun Zhu, Baoyuan Wang, Xiaokang Yang, Wenjun Zhang, and Zhuowen Tu. Action recognition with actons. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3559–3566, 2013.

Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. *arXiv preprint arXiv:2103.04503*, 2021.

APPENDIX

A TRANSFORMER BACKBONE DETAILS

For encoding the representation network $f(\cdot)$ of pose sequences in our task, we adopt the Transformer encoder network (Vaswani et al., 2017) for its superior performance on recent vision applications (Girdhar et al., 2019; Lohit et al., 2019; Plizzari et al., 2020; Carion et al., 2020; Zou et al., 2021). Transformer builds layers of a representation by leveraging a global attention mechanism, and thus it is capable of learning frame-wise representations while encoding rich contextual information. Particularly, the self-attention mechanism tends to identify the context and aggregates related information in the entire sequence to represent the current frame. This long-range property perfectly meets our need for context modeling.

An encoder layer of the standard transformer architecture consists of a multi-head self-attention module and a feed-forward network (FFN) (Vaswani et al., 2017). Our transformer backbone network is built by stacking three encoder layers. Similar to word embedding in NLP tasks (Collobert & Weston, 2008; Pennington et al., 2014; Bojanowski et al., 2017), a 3D skeleton sequence is simply flattened in the spatial dimension and embedded into a 512-dimensional hidden space using a two-layer MLP network. Due to the permutation invariance of the transformer architecture, standard positional encodings (Vaswani et al., 2017) using sine and cosine functions of various frequencies are added to the input embeddings to make them sensitive to relative position in the sequence.

Note that there exists more sophisticated and potentially more effective spatial modeling architectures like GCN (Yan et al., 2018), SRNet (Zeng et al., 2020) and Spatial Transformer (Plizzari et al., 2020) to model the highly structured skeleton data together with rich prior information. However, examining them and comparing their performance is not the focus of this work.

B EXPERIMENT DETAILS

For AIST++, note that for each sequence of basic choreography, dancers are asked to dance six times with different BPMs (Beats Per Minute). Each basic choreography thus naturally forms a class. This is leveraged in part of our experiments. For each genre, we extract 20 motion sequences to form a validation set. Unless otherwise stated, representation learning is conducted on all advanced videos to prevent the system from abusing repetitive patterns in basic videos; the lexicon is built on the training split of basic videos; inference and metric calculation are done on the validation split of basic videos.

We present the values of the hyper-parameters in Table 4. The representation learning experiments are conducted on 4 Nvidia V100 GPUs.

C AUGMENTATION DETAILS

In Table 5, we show the hyper-parameters of different data augmentations. For translation augmentation and rotation augmentation, the probability of different values used are uniform, while for speed augmentation, we first uniformly sample a number between 1 and the highest speed, then give a 50% chance of using its reciprocal.

D IMPLEMENTATION DETAILS OF APPLICATIONS

Genre Classification   For our genre classification method, we create a new split of advanced choreographies with a larger test set in relation to the development set. Representation learning and lexicon building is carried out on the development set. After lexicon building, we do nearest neighbour inference on the test set to obtain tokenizations of the test videos.

We use the LSTM (Hochreiter & Schmidhuber, 1997) text classification model with learnable and randomly-initialized word embeddings. We use a 2-layer unidirectional LSTM (Hochreiter & Schmidhuber, 1997) for sequence modeling. For tokenized input (ours), we use a learnable word embedding, while for the baseline, we use a linear layer for mapping 3D human skeleton input into a

Table 4: Different hyper-parameters.

	Hyper-parameter	Value
Shared	Batch Size	32
	Number of frames	64
	Optimizer	Adam
	Peak Learning Rate	2.5×10^{-5}
	Weight Decay	1.0×10^{-6}
	Gradient Clip by Norm	0.5
	Number of epochs	500
	Number of warmup epochs	50
	Learning rate Scheduler	Linear Warmup Cosine Annealing
	Frames per second	60/30 (default to AIST++/PKU-MMD)
Specific to TCN	Anchor Number	16
	Positive Window Size	2
	Negative Multiplier	4
	Triplet Loss Margin	2

Table 5: Hyper-parameters of the standard augmentations.

Hyper-parameter	Value
Translation range	± 0.2m
Rotation range	$\pm 18°$
Speed range	$\frac{1}{2} \sim 2$

hidden space of the same dimension (Li et al., 2018). Hyperparameters can be found in Table 6. Note the significant discrepancy in input sequence lengths due to tokenization.

Action Segmentation After TAN pre-training and lexicon building, we learn a classifier by assigning each acton class to one of the 51 action classes defined in the PKU-MMD dataset according to maximum agreement in the training set. Using this classifier, we densely evaluate the sliding windows in different time scales across the test video, then select high confidence local windows using the Non-Maximum Suppression (NMS) algorithm as our final action detection results.

Model Selection For both variants, we further split the AIST++ development split into a training split and a validation split by a ratio of $4 : 1$. Sizes of sets can be found in Table 6. We use the best-performing model on this validation split for testing. Results on the test split are reported.

E MORE ON PROPOSED METRICS

Here, we provide a proof that the conditional entropy sequence $\{F_N\}$ defined in Section 4.2 converges. Intuitively, the more characters that are conditioned on, the less uncertainty there is in predicting the next character.

Proposition 1. *The sequence F_N converges.*

Proof. Since F_N is a conditional entropy sequence, we know $F_N \geq 0$. We only need to prove $F_{N+1} \geq F_N$ for any $N \in \mathbb{N}^+$.

Table 6: Hyper-parameters used for long kinematic video recognition experiment

Hyper-parameter	Value
Batch Size	50
Optimizer	Adam
Peak Learning Rate	1.0×10^{-4}
Weight Decay	1.0×10^{-4}
Gradient Clip by Norm	0.1
Number of epochs	200
Hidden Space Dimension	256
Train set size	109
Validation set size	28
Test set size	61
Class number	10 (default to AIST++)

$$F_N - F_{N+1} = -\sum_{\mathcal{W}_N} p(W_N) \log p(w_N|W_{N-1}) + \sum_{\mathcal{W}_{N+1}} p(W_{N+1}) \log p(w_{N+1}|W_N)$$

$$= \sum_{\mathcal{W}_{N-1}} \left(\sum_{w_N, w_{N+1}} p(W_{N+1}) \log p(w_{N+1}|W_{N-1}, w_N) - \sum_{w_N} p(W_N) \log p(w_N|W_{N-1}) \right)$$

$$\geq \sum_{\mathcal{W}_{N-1}} \left(\sum_{w_N, w_{N+1}} p(W_{N+1}) \log p(w_{N+1}|W_{N-1}) - \sum_{w_N} p(W_N) \log p(w_N|W_{N-1}) \right)$$

$$= \sum_{\mathcal{W}_{N-1}} \left(\sum_{w_{N+1}} p(W_{N-1}, w_{N+1}) \log p(w_{N+1}|W_{N-1}) - \sum_{w_N} p(W_N) \log p(w_N|W_{N-1}) \right)$$

$$= 0.$$

$\square$

The value it converges to is then defined as language entropy H by Shannon (1951).

We measure the correlation strengths among three proposed metrics, namely, Kendall's Tau, NMI, and language entropy on 1000 runs of experiments differing only in augmentation parameters with Pearson's r, Spearman's ρ, and Kendall's τ (Pedregosa et al., 2011). From Table 7, we can interpret the results as indicating that NMI and language entropy have strong correlation while Kendall's Tau is linked to them weakly.

Table 7: Correlation among proposed metrics

	$|r| \uparrow$	$\rho \uparrow$	$\tau \uparrow$
Kendall's Tau & NMI	0.45	0.45	0.31
Kendall's Tau & language entropy	0.35	0.36	0.24
NMI & language entropy	0.82	0.81	0.62

F   QUALITATIVE RESULTS

More acton instances (short skeleton sequences) clustered into actons can be found in in Figure 6. An animated version can be found in the demo video in Supplementary Material.
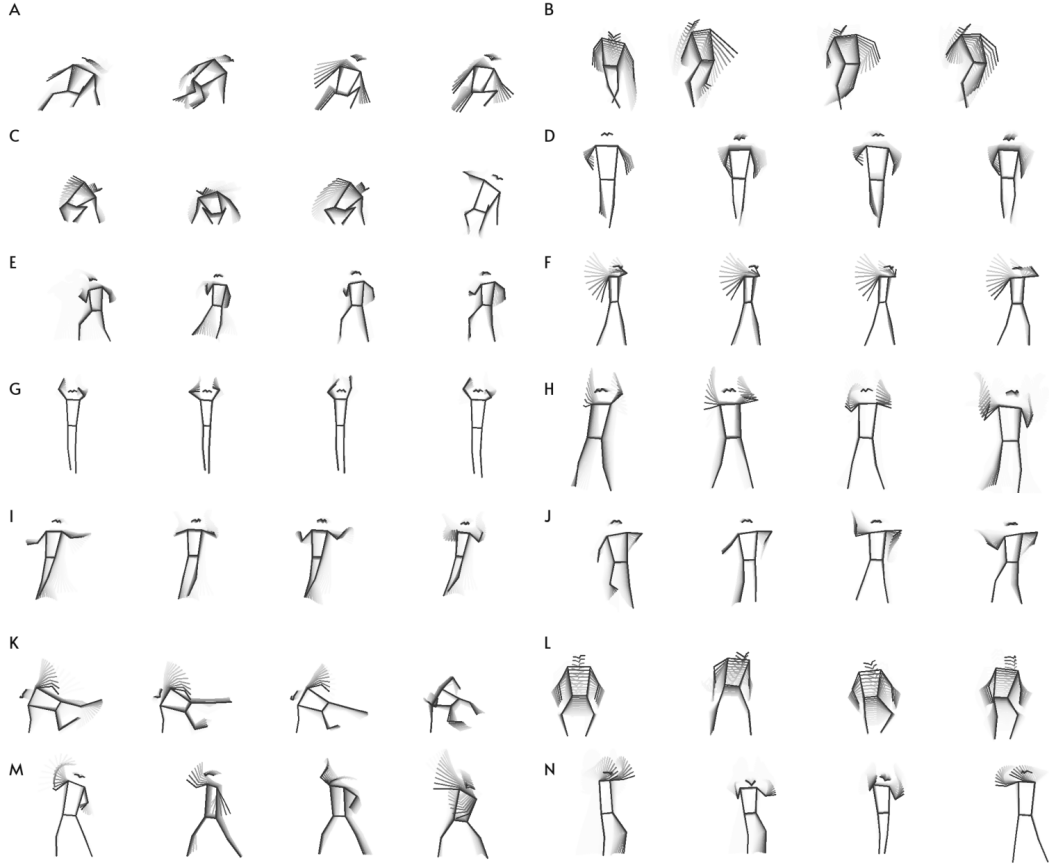
Figure 6: Continued Figure 5, visualization of discovered actons from advanced videos. For each acton, four instances are shown side to side, with temporally earlier poses fading away. At the bottom left, an acton **M** with large intra-cluster variance is shown; to its right are four acton instances facing different directions but underpinned by the same motion.

To check the overall quality of a built lexicon, the length distribution of all segments and the distribution of the number of instances for all actons are shown in Figure 7, with $K = 450$. We can see that the discovered actons have a median duration of about 0.5 second and have reasonable numbers of repetition.
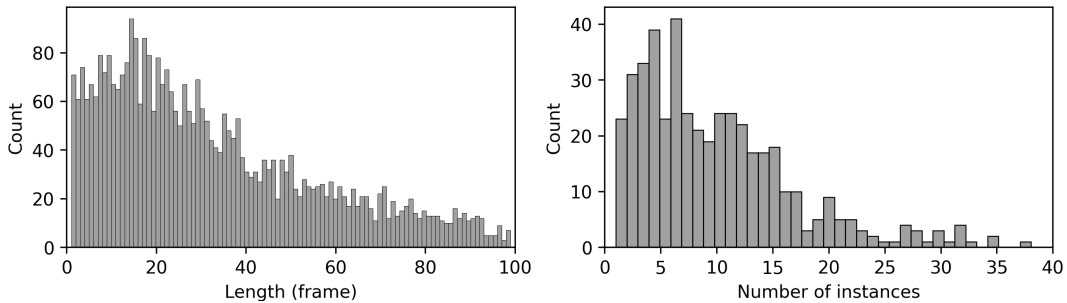

Figure 7: The left histogram shows the distribution of segment length across all segments. The right histogram shows the distribution of repetition number, or number of instances, for all discovered actons.

To qualitatively assess the quality of frame embeddings produced by Temporal Alignment Network (TAN), we use t-SNE to reduce the embedding space to two dimensions and plot frames from four basic videos with different tempos. Each of these four videos is a repetition of one basic motion. From Figure 8, we can learn the following:

- Embeddings of frames in a video form ordered *lines*, which show that our embeddings possess the quality of temporal continuity.

- Despite different video lengths due to different tempos, different lines of videos have roughly the same starting, turning and ending points. This shows that our embeddings are temporally aligned.

- In the left part, different lines are largely overlapping, in concordance with the fact that they are the same sequence of motions. This shows that the frame embedding is capturing skeleton features. Non-overlapping parts can be attributed to the dancers rendering the same motion differently each time.

- Four segments in each line are overlapped, which reflects the fact that there are repeated motions in each video for four times.

- In the right part, we can see how lexicon building is working. It generally over-segments one *motion* from a human perspective into a sequence of actons. For a conceptual example, if a labanotation describes the dance as $AAAA = (A)^4$, as it is a repetition of the same motion A, our system represents it as $(abcd)^4$.

- In the bottom part, we can see that for all four videos, the same patterns are repeated two times in the middle. It is desired that an acton level pattern recognition system can identify this and group over-segmented actons into one larger acton. This inspired us to use language entropy as a metric.

Using the same setting as in Section 5.3, we show the reorganized corpus in Figure 9. Note that due to the specific segmentation approach used in our method, the same acton cannot appear continuously in the tokenization results.
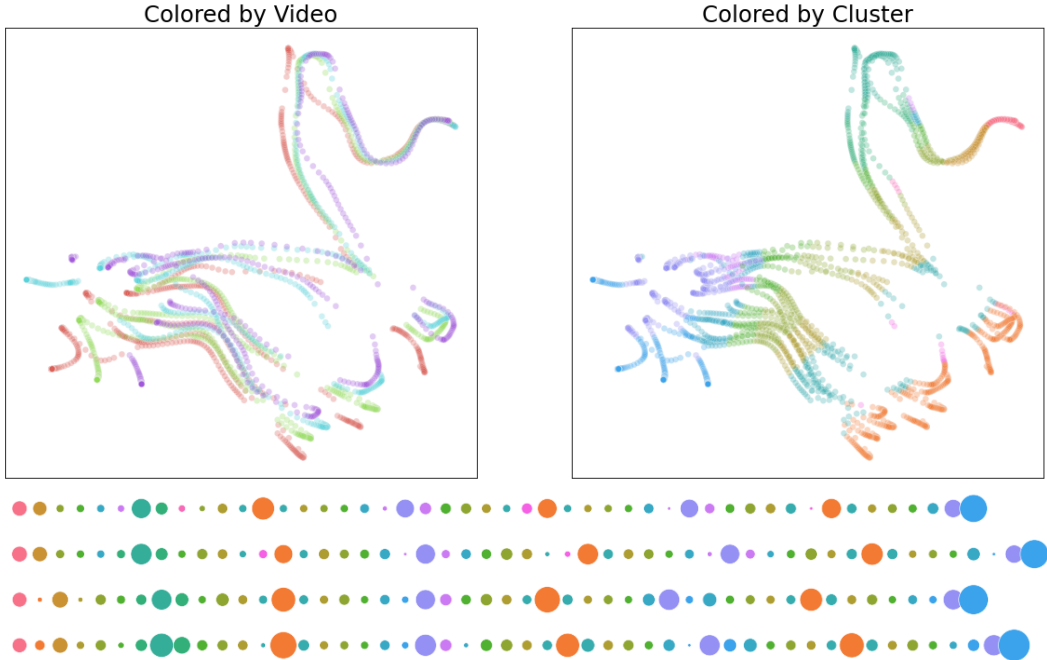

Figure 8: t-SNE visualization of frame embeddings of videos with different tempos. In the left, frames are colored by the video they belong to; on the right, they are colored by the cluster they belong to. At the bottom is a visualization of tokenization results on the same four videos. Each line is one advanced video. Dot size corresponds to acton instance duration. Consistent colors are used for the t-SNE plot in the right and the tokenization visualization.
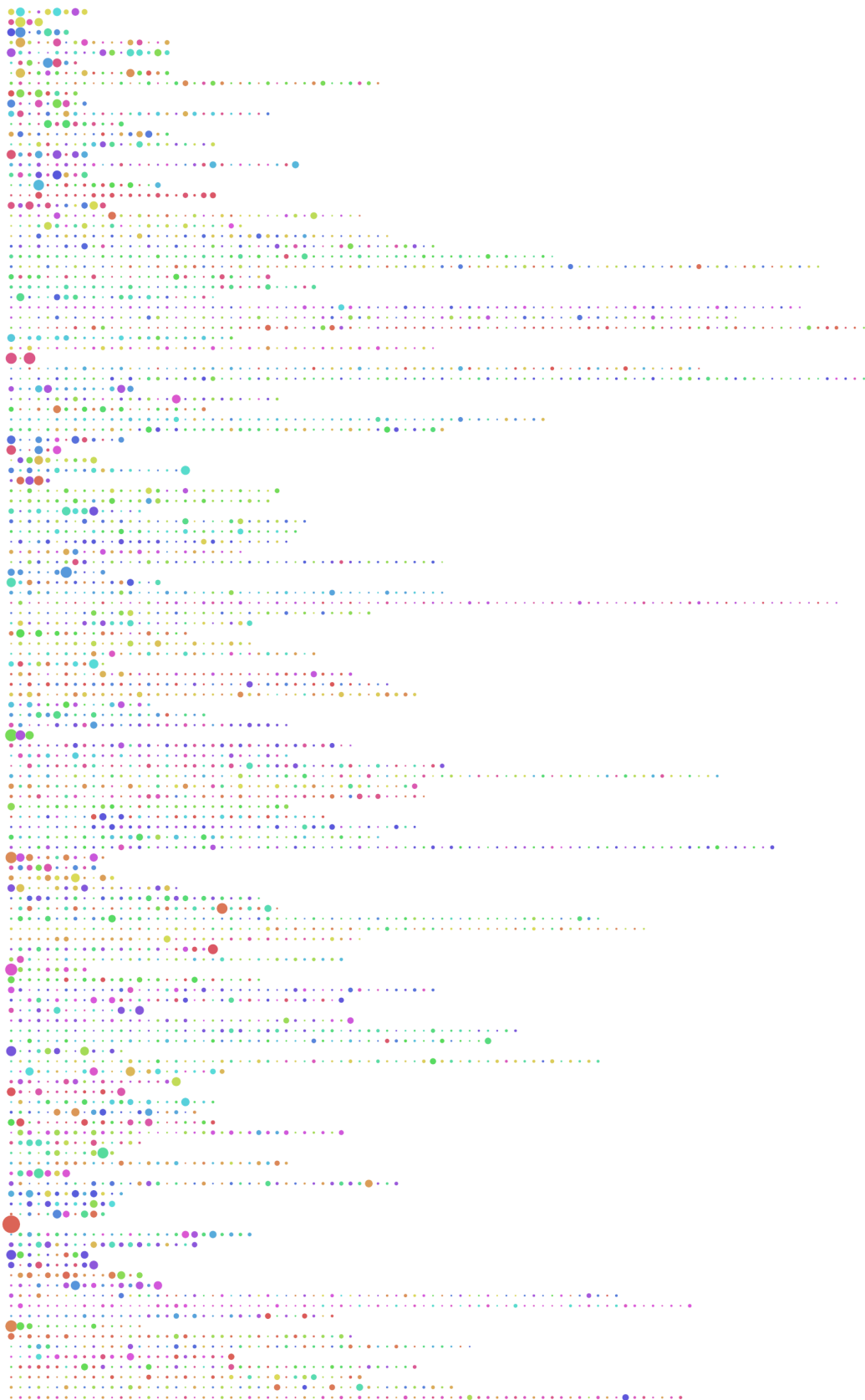
Figure 9: Illustration of segmentation on advanced videos in AIST++. Each line is one advanced video. Dot size corresponds to acton instance duration. Different colors represent different actons.