Supplementary Material: Aligning Moments in Time using Video Queries

Yogesh Kumar[1*] Uday Agarwal[1*] Manish Gupta[2] Anand Mishra[1]
[1]Indian Institute of Technology Jodhpur [2]Microsoft
{kumar.204, agarwaluday, mishra}@iitj.ac.in, gmanish@microsoft.com

1. Background: Dynamic Time Warping

Dynamic Time Warping (DTW) [3] is a widely used method for aligning and comparing two sequences, such as time series, that may vary in length or exhibit local accelerations and decelerations. Unlike the Euclidean distance, which requires sequences to be of the same length and aligned in time, DTW is robust to temporal shifts and distortions, making it a popular choice in speech, gesture, and time-series analysis [1].

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{m \times d}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{n \times d}$ denote two sequences of possibly different lengths m and n, where d is the feature dimension. The pairwise dissimilarity between sequence elements is encoded in a cost matrix $\mathbf{C} \in \mathbb{R}^{m \times n}$, with entries $C_{i,j} = \delta(\mathbf{x}_i, \mathbf{y}_j)$ for a chosen distance function δ (e.g., squared Euclidean).

An alignment between $\mathbf{X}$ and $\mathbf{Y}$ is represented by a binary matrix $\mathbf{A} \in \{0,1\}^{m \times n}$, where $\mathbf{A}_{i,j} = 1$ indicates that $\mathbf{x}_i$ is matched to $\mathbf{y}_j$. The set of all valid alignments, $\mathcal{A}_{m,n}$, enforces monotonicity and continuity constraints so that alignments are contiguous and order-preserving.

DTW computes the minimum total alignment cost over all valid alignments:

$$\mathrm{DTW}(\mathbf{X}, \mathbf{Y}) = \min_{\mathbf{A} \in \mathcal{A}_{m,n}} \langle \mathbf{A}, \mathbf{C} \rangle,$$

where $\langle \mathbf{A}, \mathbf{C} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{A}_{i,j} C_{i,j}$ is the Frobenius inner product. This optimization can be efficiently solved by dynamic programming in $\mathcal{O}(mn)$ time and space.

Limitations DTW's flexibility enables robust sequence comparison and alignment, even when sequences are stretched or compressed in time. However, a key limitation of DTW is that it is not differentiable with respect to its inputs, as the minimum over alignments introduces discontinuities. This non-differentiability makes DTW unsuitable as a loss function in gradient-based learning frameworks.

Soft-DTW Differentiable Relaxation To address this, *soft-DTW* [1] replaces the hard minimum in DTW with a differentiable soft-minimum (log-sum-exp) operator, parameterized by a smoothing parameter $\gamma > 0$. The soft-minimum operator is defined as:

$$\min{}^{\gamma}\{a_1, \ldots, a_K\} = -\gamma \log \sum_{k=1}^{K} \exp\left(-\frac{a_k}{\gamma}\right).$$

The soft-DTW cost is then given by:

$$\text{soft-DTW}_{\gamma}(\mathbf{X}, \mathbf{Y}) = \min{}^{\gamma}\left\{\langle \mathbf{A}, \mathbf{C} \rangle : \mathbf{A} \in \mathcal{A}_{m,n}\right\}.$$

As $\gamma \to 0$, soft-DTW recovers the original DTW; for $\gamma > 0$, it aggregates the costs of all possible alignments, weighting them exponentially by their cost. This smoothing ensures that soft-DTW is differentiable everywhere with respect to its inputs, enabling its use as a loss function for end-to-end training of models that output sequences.

2. Additional Result Discussion

2.1. Impact of Query Length on Performance

Fig. 1 shows the performance of MATR across different query lengths on ActivityNet-VRL. We observe that mIoU remains relatively stable across varying query lengths, indicating that MATR effectively captures relevant moments regardless of the duration of the query. In contrast, R@1 demonstrates a slight downward trend as query length increases. This suggests that while MATR can maintain alignment with longer queries in terms of overlap (mIoU), accurately retrieving the exact match (R@1) becomes slightly more challenging for longer queries.

2.2. MATR Scalability and Efficiency

Our model with 4 layers on both encoders was sufficient to achieve strong performance. We perform additional experiments by varying number of layers. Increasing the number of layers from 4 to 6 led to only a 0.3% improvement in R@1 (54.8 to 55.1), despite a 25.4% increase in parameters (82.7M to 103.7M).

[*]Equal Contribution

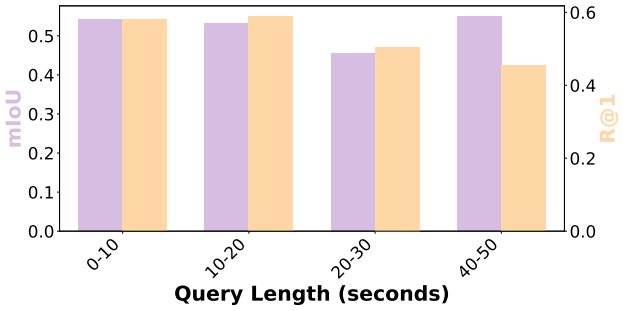

Figure 1. Performance comparison of MATR on varied query length using ActivityNet-VRL. The mIoU remains relatively stable across different query lengths, while R@1 shows a slight downward trend as query length increases.

3. Dataset Details

Table 1 presents a comparative analysis of the SportsMoments dataset and ActivityNet-VRL dataset [2]. The table highlights the number of query-target pairs and action classes in each dataset's training, validation, and testing splits. The SportsMoments dataset consists of 750,393 pairs spanning 16 classes in the training split, making it significantly larger and more focused than the ActivityNet-VRL dataset, which has 462,872 pairs distributed across 160 classes. In the validation and testing splits, SportsMoments contains 10,000 pairs each, with 4 classes and 9 classes, respectively. In contrast, ActivityNet-VRL includes only 829 pairs and 978 pairs spanning across 20 classes each for validation and testing respectively. These statistics underscore the tailored structure and scalability of the SportsMoments dataset for diverse training and evaluation scenarios.

Dataset	Split	# Pairs	# Classes
SportsMoments	Train	750,393	16
	Val	10,000	4
	Test	10,000	9
ActivityNet-VRL [2]	Train	462,872	160
	Val	829	20
	Test	978	20

Table 1. Comparison between ActivityNet-VRL and our proposed dataset SportsMoments.

Dataset	Type	Mean	Median	Max.	Min.
SportsMoment	Target Video	98.3	98.2	208.8	35.4
	Query Video	6.0	5.0	66.0	1.0
ActivityNet-VRL	Target Video	38.2	29.3	191.5	1.3
	Query Video	8.2	7.2	57.2	0.5

Table 2. The mean, median, max, and min lengths (in seconds) of the unique target and query videos in the SportsMoments and ActivityNet-VRL datasets.

Table 2 provides insights into the temporal characteristics of the unique target and query videos in the SportsMoments dataset. The mean duration of target videos is 98.3 seconds, with a median duration of 98.2 seconds. These videos have a maximum duration of 208.8 seconds and a minimum of 35.4 seconds, showcasing their relatively long and consistent lengths. On the other hand, query videos are significantly shorter, with a mean duration of 6 seconds, a median duration of 5 seconds, a maximum length of 66 seconds, and a minimum of 1 second. These contrasting temporal scales reflect the dataset's emphasis on connecting succinct queries with comprehensive target contexts. For the ActivityNet-VRL dataset, target videos are shorter on average compared to those in SportsMoment, with a mean duration of 38.2 seconds and a median of 29.3 seconds. The maximum duration of these videos is 191.5 seconds, while the minimum is 1.3 seconds, indicating a wider range of temporal lengths. Similarly, query videos in ActivityNet-VRL are slightly longer than those in SportsMoment, with

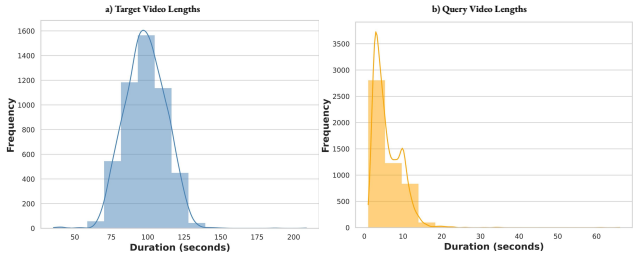

Figure 2. We visualize the distribution of **a)** target video durations and **b)** query video durations in our proposed dataset, SportsMoments.

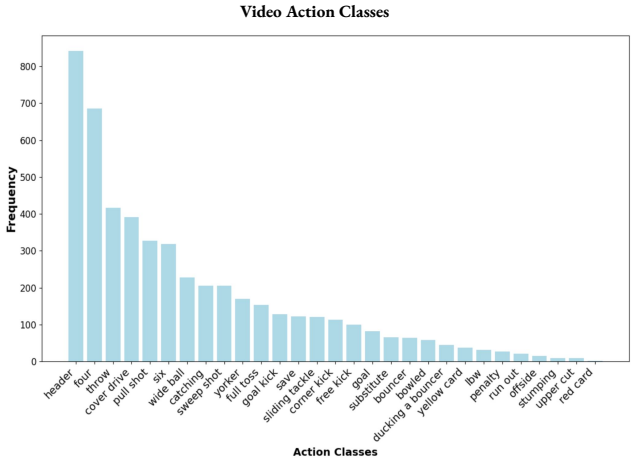

Figure 3. Frequency distribution of action classes in the SportsMoments dataset. The dataset comprises a wide range of actions across the two sports, including frequent activities such as "header," "four," and "throw," as well as less common ones like "red card" and "uppercut." The figure highlights the variation in class occurrences, reflecting the diversity and imbalance in real-world sports scenarios.

4a5Oou5u3io	GrIOmtbyHNc	5j2tQfj2Osk	mo3H_qusvxI	RpmSeHG8SJ0	cpU0GSxliSU
EOvpUWZC6ng	Txq8RoclyKM	-vYVBUyEuAA	IvjBlpb8GTA	ukuNlWxcR2k	LmqBbPwCy8o
zuPuDzEWI-A	OFbyNU6UQQs	iPXz5XWfj1k	IA5BNsNu6fw	9sRb61qsoLs	X0we8220k74
MbdWVX2jUhA	04vWaGaqw-4	92s9HCGG4vQ	BMXfphiHj1s	1yHWGw8DH4A	shyBGceNASI
YqKYpgZ9FWU	GxO1tlZKa8E	qElsVosGiXA	zUTUA0TKvfQ	G4vxbWF79f4	v9Akttif8gY
Mtnmvuhs2RA	3K90O0Y9NZc	9gyv2xh7qQw	1-SCubr2yc4	ezQUPi74IIQ	cTqY53zWypk
qQfuOCns4P4	iGGElV40TcU	X6VNTk-Vl5w	iFYm_ogzmdA	6j1Oj5N7kwE	WkEvdVImkec
wTN82hUkhWk	rpPCjyVfQOo	L7olqH38dzw	farmGHpVCek	S825iLaZQhg	T-c3B0lJJvE
ku0Y8TA8L44	bSRB7ysDRQw	90yyJvfbwAs	9oRmmeKJ6WM	6f9v2z5HYX4	eeaHNmIuP3c
5OGkvFUl3G8	ePmiyHWxMDk	CuEC8Twrcgo	nc9nA8fG87Q	9dy7kpALSAE	EedISWx_fCo
9lz5VV-8XEc	qEgOfHT_MnI	snM7xGdWHSc	Us_6Nzj3Fic	H7q7q1IwEY4	xYbklmu4ZsE
stWSFacouPM	RRk3S82oH9M	XJgOiFkOgzo	RKP_Mg4piUA	2rLIjqEM3Nw	rfl6elqYCKA
cdkiRWd15RA	FXF7c2LyGDo	hUBLgzSEaS4	X9ny1p0D61I	qWlfNQRd0uY	tDeoIdYvh6g
W_j6PHiin_0	fWCU_oJyD38				

Table 3. YouTube IDs used for SportsMoments dataset creation. We used 80 videos for dataset curation.

a mean duration of 8.2 seconds, a median of 7.2 seconds, a maximum length of 57.2 seconds, and a minimum of 0.5 seconds. These statistics highlight the diverse temporal characteristics of the two datasets, emphasizing their complementary nature for tasks involving video-to-video grounding and retrieval.

Figure 2 visualizes the duration distribution of the target and query videos, highlighting the dataset's temporal diversity. Figure 3 illustrates the frequency of annotated segments across action classes, providing an overview of the class distribution in the SportsMoments dataset.

The 29 classes in our proposed benchmark dataset SportsMoments are as follows: yorker, bouncer, full toss, wide ball, cover drive, six, four, pull shot, sweep shot, ducking a bouncer, uppercut, catching, run-out, stumping, LBW, run, goal, sliding tackle, off side, free kick, corner, goal kick, header, throw, yellow card, red card, substitute, penalty, and save. These action classes included in the SportsMoments dataset span a diverse range of activities from cricket (e.g., yorker, cover drive, LBW) and soccer (e.g., goal, sliding tackle, yellow card). The dataset covers a good breadth of actions, highlighting its applicability to multi-sport analysis and video understanding tasks.

4. SportsMoments Dataset Curation

The construction of query and target videos follows a systematic process. Each annotated segment in the dataset serves as a potential query video associated with its corresponding action class. To create a target video, the annotated segment is offset by a duration ranging between 30 to 60 seconds on either side, i.e., both towards the start and end of the segment. Additionally, a scaling factor of 25% of the annotated segment's duration is applied to this offset to account for variability in segment lengths.

During the annotation phase, to maintain data quality, we ensured that immediately overlapping segments (representing the same action occurring within 2 seconds) are merged into a single segment. Furthermore, since the dataset uses full-length matches for annotation, the number of overlapping segments across different classes is negligible, ensuring that the dataset remains clean and avoids ambiguities.

Each pair in the dataset consists of a short clip acting as a query video coupled with a longer target video belonging to the same action class. For the creation of training pairs, we adopt a strategy similar to that used in [2]. A query video can be paired with all target videos of the same class. Unlike [2], which generates such exhaustive pairings on the fly, we precompute and store all possible pairs during the dataset creation process.

For the test and validation splits, we curate a fixed set of 10,000 pairs for each split. These splits are designed to ensure fair representation of all classes, thereby enabling balanced evaluation across the dataset.

5. YouTube IDs used in SportsMoments

We also release the YouTube IDs of videos used in our dataset, 80 in total as shown in Table 3. For cricket, we have a total of 64 videos, 26 of which are full length twenty over games, 9 are fifty over games, 26 of which are highlights of different matches, one of which is a part of a full length match and 3 videos are that of a single day Test-match innings. The remaining 16 matches are that of full length football matches from various leagues.

References

[1] M. Cuturi and M. Blondel. Soft-DTW: a differentiable loss function for time-series. In *ICML*, 2017. 1

[2] Y. Feng, L. Ma, W. Liu, T. Zhang, and J. Luo. Video re-localization. In *ECCV*, 2018. 2, 3

[3] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23:67–72, 1975. 1