

An X3D Neural Network Analysis for Runner’s Performance Assessment in a Wild Sporting Environment

David Freire-Obregón, Javier Lorenzo-Navarro, Oliverio J. Santana,
Daniel Hernández-Sosa and Modesto Castrillón-Santana
Universidad de Las Palmas de Gran Canaria
david.freire@ulpgc.es

Abstract

We present a transfer learning analysis on a sporting environment of the expanded 3D (X3D) neural networks. Inspired by action quality assessment methods in the literature, our method uses an action recognition network to estimate athletes’ cumulative race time (CRT) during an ultra-distance competition. We evaluate the performance considering the X3D, a family of action recognition networks that expand a small 2D image classification architecture along multiple network axes, including space, time, width, and depth. We demonstrate that the resulting neural network can provide remarkable performance for short input footage, with a mean absolute error of 12 minutes and a half when estimating the CRT for runners who have been active from 8 to 20 hours. Our most significant discovery is that X3D achieves state-of-the-art performance while requiring almost seven times less memory to achieve better precision than previous work.

1 Introduction

With the progress of technology, the world of sports undergoes a tremendous transformation as competition teams seek new ways to gain an advantage. Computer vision, which employs artificial intelligence algorithms to analyze camera footage in real-time, is one of the most promising research areas on this topic. In this regard, computer vision has already been utilized in various applications, such as player position estimation, ball trajectory prediction, and technological assistance to referee decisions [1].

Recently, algorithms for action quality assessment (AQA) have emerged as a result of human action recognition research [2]. AQA aims to design a system that can automatically and objectively evaluate specific human actions based on input videos. Contrary to the traditional video action recognition problem, AQA evaluates the execution of an action.

Sports have benefited from AQA in many practical scenarios, such as athlete posture correction, coaching systems, and action evaluation. In recent years, the score to be assigned to an athlete’s performance by a panel of judges has been estimated, such as diving and gymnastics movements. Consequently, numerous AQA approaches treated this task as a regression problem to

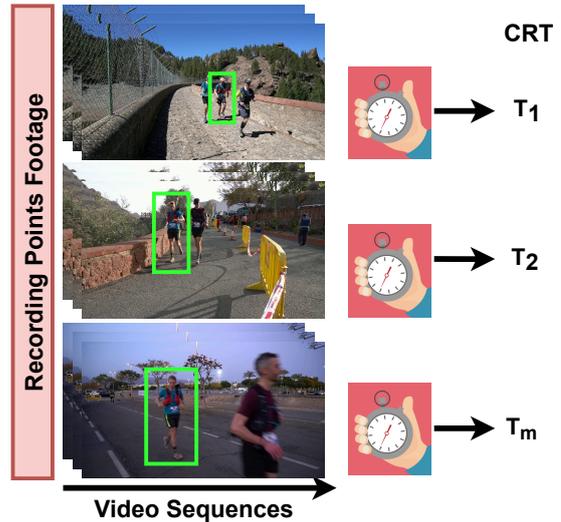


Figure 1. Samples of a runner’s footage at each recording point. The runner of interest is surrounded by a green container. We analyze different X3D instances for each footage to extract the runner’s embeddings. Then, these embeddings are fed into a model to infer the CRT at a specific recording point.

learn the direct mapping between videos and action scores [3, 4].

Lately, ultra-distance competitions have been considered for runners’ performance evaluation [5]. Contrary to previous sporting AQA works, the task is not to measure the performance of methods between the ground truth and a predicted score series but the CRT. The CRT at a particular recording point RP_i can be defined as $T_i = T_1 + \sum_{j=2}^{j=i} (T_j - T_r)$ where $r = j - 1$, see Figure 1. Moreover, the problem framed in ultra-distance races is challenging due to the highly dynamic scenes and the race span, i.e., runner’s appearance variance, multiple scenarios, occlusive elements, etc.

Recent advances in ultra-distance races CRT estimation provide high generality regarding the action recognition networks considered [5, 6]. We seek to bridge these approaches by gradually increasing the network’s complexity to resolve this task. Our work explores X3D expanded instances to produce accurate

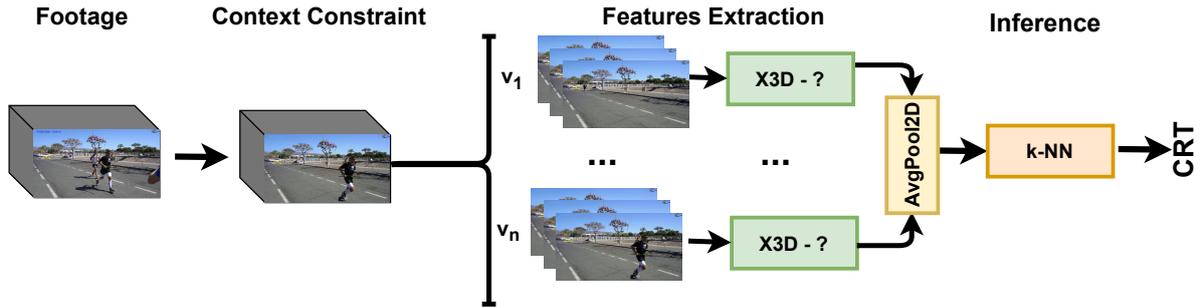


Figure 2. **The proposed pipeline for regressing runner’s CRT.** The designed process consists of two primary components: the footage pre-processing block, and the regression block. In the first scenario, the tracker aids by neutralizing the runner’s background activity. The latter entails the division into n small clips by down-sampling. Then the clips are sent into X3D instances for extracting features. An average pooling synthesizes the final features. The resulting tensor is the input to the regressor.

CRT predictions, as shown in Figure 1. The used architecture is X3D for expanding from the 2D space into the 3D space-time domain [7]. The 2D base architecture is the MobileNet. The considered expansion then progressively increases the computation by expanding only one axis at a time, i.e., frame rate, sampling rate, footage resolution, network depth, number of layers, and number of units.

We have analyzed the X3D instances in a dataset collected to evaluate runner re-identification methods in real-world scenarios. The achieved results are remarkable (up to 12 minutes and a half of MAE), and they have also provided interesting insights. Contrary to other action recognition networks, X3D instances generate shorter embeddings. As a consequence, k-NN instance-based approach turns to be enough to tackle the problem efficiently. Second, a few network expansions are enough to achieve the best performance. Finally, our proposal outperforms other approaches in literature.

2 Related Work

AQA is inherently an action recognition problem facing challenges like automatically and objectively evaluating specific actions people complete through input footage. A common approach to handling action recognition in supervised training has been to limit input data to skeleton-based approaches, e.g., detecting human body joints [8, 9]. This encourages the supervised network to infer knowledge from a metric scale that may be globally inconsistent in scenarios where the human body shifts rapidly, such as violence detection [10] and sports AQA [11]. In addition, the estimated skeleton data can frequently be noisy in realistic scenes due to occlusions or changing lighting conditions [12], particularly in ultra-distance races held in uncontrolled environments.

Regarding this issue, appearance-based approaches

have been used in the past to tackle AQA. Pioneer research conducted by Parmar et al. already uses C3D neural networks at a clip level for feature computations [13, 14]. More recently, several works have used the I3D network on clips not to predict a score but a score distribution [11, 15]. I3D ConvNets have also been used to tackle the runner’s performance in the past [5, 6]. Contrary to these works, we aim to analyze the athletes performance by progressively expanding an X3D architecture to achieve a lightweight architecture that preserves robustness.

3 Method

We develop a modular multi-stage pipeline for runners’ CRT estimation in ultra-distance races. Its structure is illustrated in Figure 2.

Context constrain. According to Freire et al. [6], action recognition networks require clean footage input for CRT inference. Therefore, objects (athletes, race personnel, cars) that are not interesting must be removed from the scene. Specifically, the initial block pre-processes the raw input data to focus on the runner of interest. We have applied ByteTrack [16], a multi-object tracking network, to track the runner of interest in each footage. Then, a context-constrain pre-processing yielded the scenario considered in our experiments. Given a runner i bounding box area $BB_i(t, RP)$ at a given time $t \in [0, T]$ and in a recording point $RP \in [0, P]$, the new pre-processed footage $F'_i[RP]$ can be formally denoted as follows:

$$F'_i[RP] = BB_i(t, RP) \cup \tau(RP) \quad (1)$$

Where $\tau(RP)$ is the average number of frames to generate the clean footage where the runner appears with a still background.

Feature extraction and regression. The input footage consisting of n frames is down-sampled

and split into n video clips (v_1, \dots, v_n), each containing q consecutive frames representing an activity snapshot (see Figure 2). Next, each video clip v_i is passed through a pre-trained X3D network, resulting in a 192-dimensional feature vector. These X3D instances have been pre-trained with the Kinetics dataset [17], which comprises 400 action categories. Once all the feature vectors of the n video clips are obtained, an average pooling layer is applied to ensure that the information from each clip is given equal consideration. Finally, the extracted features are used to train a k-NN regressor, which is used to infer the CRT.

X3D instances. We have used four X3D expanded instances that are named according to their size; extra small (X3D-XS), small (X3D-S), medium (X3D-M), and large (X3D-L). Each considered expansion is used for sequentially expanding X2D from a tiny spatial network to a spatiotemporal X3D network by performing the following operations on temporal (frame rate and sampling rate), spatial (footage resolution), width (network depth), and depth dimensions (number of layers and number of units) [7]. X3D-XS is the output after five expansion steps. The following larger model is X3D-S, defined by one backward contraction step following the seventh expansion step. The contraction step reduces the frame rate and, therefore, temporal resolution while holding the clip duration constant. The eighth and the tenth expansions generate the X3D-M and X3D-L, respectively. As seen in Table 1, X3D-M expands the spatial resolution by increasing the spatial sampling resolution of the input video. In contrast, X3D-L expands not only the spatial resolution but also the depth of the network by increasing the number of layers per residual stage.

4 Dataset and Experiments

A key challenge in performing AQA is the lack of publicly available sporting datasets. Our pipeline needs athlete’s data to regress CRT properly. While many works in this sporting domain rely on statistical data, manually gathering sufficient multimedia data is expensive. We employ a dataset derived from TGC20ReId dataset [18] provided by the authors, that contains seven-second clips at 25 fps at each recording point for each participant.

The initial dataset includes annotations for nearly 600 participants across six recording points. Given the varying performances of the runners, the gap between the leaders and the last runners increases along the course as the number of active participants decreases. Consequently, a subset of 214 runners is eligible for estimating the CRT, that is, those runners that have covered the last three recording points during the dataset recording time.

Metric. An athlete i observation $o_i[RP]$ at a recording point $RP \in [0, P]$ consists of a pre-processed footage $F_i[RP]$ and a CRT $\phi_i[RP]$. In addition, the

CRT of the runners has been normalized between [0,1] using Equation 2.

$$\phi'_i[RP] = \frac{\phi_i[RP] - \min(\phi_i[0])}{\max(\phi_i[RP])} \quad (2)$$

Our task is to identify an end-to-end regression technique that minimizes the following objective:

$$\min L(\phi'_i[RP], \psi_i[RP]) = \frac{1}{N} \sum_{j=0}^N |\phi'_i[RP]_j - \psi_i[RP]_j| \quad (3)$$

where $\psi_i[RP]$ represents for the runner i predicted value at a recording point RP based on seven seconds of movement observation, and N is the batch size.

The following section presents the average Mean Absolute Error (MAE) across 20 repetitions of 10-fold cross-validation. On average, 410 samples are chosen for training, leaving 46 for testing.

4.1 X3D Instances Evaluation

As Section 3 points out, we have considered several X3D instances, namely X3D-XS, X3D-S, X3D-M, and X3D-L. Additionally, inspired by [6], we have combined these instances averaging them (192 embeddings) or concatenating ($192 \times \#I$ embeddings, where $\#I$ is the number of combined instances) the last ResNet block output.

Table 1 shows the achieved results by each configuration. The table is divided into three blocks, with four entries in the first and three in the rest. The first block is related to the basic X3D instances, the second is the average of different X3D instances embeddings, and the last is the average of different X3D instances concatenation. Average and concatenation experiments combine models sequentially by both, model size -first XS and S, then XS, S and M, and so on- and individual performance. From the temporal dimension perspective, each X3D-XS and X3D-S input clip is composed by four frames and a large sample rate (12 frames), whereas X3D-M and X3D-L increase the number of frames per clip (13 and 16) but reduce the sample rate by a half.

Table 1 also highlights the relative importance of the model size. As can be appreciated, smaller models consistently outperform bigger models, i.e., X3D-XS is 18% better than X3D-L. Averaging model embeddings partially outperform individual approaches, but the rates are inconclusive since there is no correlation between the size and the performance. For instance, the middle combination configuration (XS+S+M) is worse than any other, bigger or smaller, configuration. Freire et al. have recently achieved their best results when concatenating I3D ConvNet embeddings [6]. Similarly, in our study, we consistently observed a substantial and consistent reduction in loss as the model size increased, highlighting the effectiveness of

Table 1. **Mean average error (MAE) achieved by each configuration.** The first column displays the model configuration (+ stands for average and \cup for concatenation, respectively). The second column shows the number of frames per video clip, the third column shows the sampling rate (SR), and the last column shows the achieved MAE. Lower is better.

| Instance | #Frames | SR | MAE |
|-----------------|---------|-------|--------------|
| X3D-XS | 4 | 12 | 0.010 |
| X3D-S | 4 | 12 | 0.011 |
| X3D-M | 13 | 6 | 0.011 |
| X3D-L | 16 | 5 | 0.012 |
| XS+S | Mixed | Mixed | 0.012 |
| XS+S+M | Mixed | Mixed | 0.013 |
| XS+S+M+L | Mixed | Mixed | 0.019 |
| XSUS | Mixed | Mixed | 0.011 |
| XSUSUM | Mixed | Mixed | 0.011 |
| XSUSUM \cup L | Mixed | Mixed | 0.011 |

concatenating embeddings. Despite these findings, it is worth noting that the X3D-XS model, albeit with a slight margin, still maintains the highest success rate among all tested models.

After considering various classifiers such as linear regression, random forest, gradient boosting, SVM, and a multi-layer perceptron, we have found that k-NN outperforms all of them. To ensure the optimal performance, we conducted a grid search to identify the most suitable regressor. It turns out that k-NN reported the best result. Furthermore, due to the small number of dimensions and the moderate number of observations, we have reported rates using a k-NN regression method. Consequently, an instance-based classifier is good enough to select the best embeddings for inference. In terms of minutes, a 0.010 MAE is roughly 12 minutes and a half. Since the fastest runner was recorded after 8 hours of CRT, and the last one after 20 hours of CRT, the achieved MAE is a really positive outcome.

To better compare the proposed pipeline with the related work, we have included our best result in Table 2. This table summarizes the performance reported in recent literature on the mentioned dataset but also the size of the model. The table includes three major architectures, C3D, 3D ResNets considering different depths, and the I3D ConvNet. Overall, the X3D-XS model outperforms other considered prior architectures on this task. Moreover, Table 2 shows that the X3D-XS model is more than six and a half times smaller than the model with the second best result. Note that the ranking in this table shows no correlation between the model size and the model performance.

Table 2. **Comparison of different architectures on the dataset used in the present work.** The first column shows the considered pre-trained architectures, whereas the second and the third columns show the number of parameters and the MAE, respectively. Lower is better.

| Architecture | #Params | MAE |
|----------------------|---------|--------------|
| C3D [19] | 34.8M | 0.038 |
| 3D ResNets-D30 [20] | 60.5M | 0.036 |
| 3D ResNets-D50 [20] | 45.8M | 0.033 |
| 3D ResNets-D101 [20] | 84.8M | 0.032 |
| 3D ResNets-D200 [20] | 146.4M | 0.031 |
| I3D-800SB [6] | 25M | 0.019 |
| I3D-2048SB [6] | 25M | 0.015 |
| X3D-XS (Ours) | 3.7M | 0.010 |

5 Conclusions

Combining metric accuracy and lightweight models is a key challenge in AQA. We propose an X3D analysis by progressively expanding the architecture on temporal, spatial, width, and depth dimensions. Then, an instance-based classifier (k-NN) provides good performance on the generated embeddings. We show improved error reduction with each basic X3D instance alone and demonstrate successful results when concatenating instance signals. Our best result was achieved by a model almost seven times smaller and a 34% better than the best proposal in the literature. Several applications can benefit from our proposal, not only monitoring a runner’s performance, but also relieving the race staff from paying exhausting continuous attention to health concerns. In addition, we hope it will assist in deploying robust and general CRT estimation models.

Acknowledgments: This work is partially funded by the the Spanish Ministry of Science and Innovation under project PID2021-122402OB-C22, and by the ACIISI-Gobierno de Canarias and European FEDER funds under project, ProID2021010012, ULPGC Facilities Net, and Grant EIS 2021 04

References

- [1] B. T. Naik, M. F. Hashmi, and N. D. Bokde, “A comprehensive review of computer vision in sports: Open issues, future trends and research directions,” *Applied Sciences*, vol. 12, no. 9, 2022.
- [2] S. Wang, D. Yang, P. Zhai, Q. Yu, T. Suo, Z. Sun, K. Li, and L. Zhang, “A survey of video-based action quality assessment,” in *Int. Conf. on Networking Systems of AI (INSAI)*, 2021, pp. 1–9.

- [3] P. Parmar and B. T. Morris, “What and how well you performed? a multitask learning approach to action quality assessment,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 304–313, 2019.
- [4] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, “Group-aware Contrastive Regression for Action Quality Assessment,” *2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pp. 7899–7908, 2021.
- [5] D. Freire-Obregón, J. Lorenzo-Navarro, and M. Castrillón-Santana, “Decontextualized I3D ConvNet for ultra-distance runners performance analysis at a glance,” in *Int. Conf. on Image Analysis and Processing (ICIAP)*, 2022, pp. 242–253.
- [6] D. Freire-Obregón, J. Lorenzo-Navarro, O. J. Santana, D. Hernández-Sosa, and M. Castrillón-Santana, “Towards cumulative race time regression in sports: I3D ConvNet transfer learning in ultra-distance running events,” in *Int. Conf. on Pattern Recognition (ICPR)*, 2022, pp. 805–811.
- [7] C. Feichtenhofer, “X3D: Expanding Architectures for Efficient Video Recognition,” *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 200–210, 2020.
- [8] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *AAAI Conf. on Artificial Intelligence*, 2018.
- [9] H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, and B. Dai, “Revisiting skeleton-based action recognition,” *2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2959–2968, 2021.
- [10] D. Freire-Obregón, P. Barra, M. Castrillón-Santana, and M. de Marsico, “Inflated 3D ConvNet context analysis for violence detection,” *Machine Vision and Applications*, vol. 33, no. 15, 2022.
- [11] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, “Uncertainty-aware score distribution learning for action quality assessment,” *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 9836–9845, 2020.
- [12] N. Carissimi, P. Rota, C. Beyan, and V. Murino, “Filling the gaps: Predicting missing joints of human poses using denoising autoencoders,” in *ECCV Workshops*, 2018.
- [13] P. Parmar and B. T. Morris, “Learning to score olympic events,” 2017.
- [14] P. Parmar and B. Morris, “Action quality assessment across multiple actions,” in *IEEE Winter Conf. on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*. IEEE, 2019, pp. 1468–1476.
- [15] B. Zhang, J. Chen, Y. Xu, H. Zhang, X. Yang, and X. Geng, “Auto-encoding score distribution regression for action quality assessment,” *ArXiv*, 2021.
- [16] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang, “Bytetrack: Multi-object tracking by associating every detection box,” in *European Conference on Computer Vision*, 2021.
- [17] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” *CoRR*, 2017.
- [18] A. Penate-Sanchez, D. Freire-Obregón, A. Lorenzo-Melián, J. Lorenzo-Navarro, and M. Castrillón-Santana, “TGC20ReId: A dataset for sport event re-identification in the wild,” *Pattern Recognition Letters*, vol. 138, pp. 355–361, 2020.
- [19] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “C3D: generic features for video analysis,” *CoRR*, vol. abs/1412.0767, 2014.
- [20] K. Hara, H. Kataoka, and Y. Satoh, “Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet,” in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.