Multi-Domain Referee Dataset: Enabling Recognition of Referee Signals on Robotic Platforms

Anonymous Author(s)
Affiliation
Address
email

Abstract: Recognizing referee signals is crucial in human and RoboCup soccer games, where an emphasis currently lies on full robot autonomy through understanding referee signals. To advance towards this goal, we introduce the Multi-Domain Referee Dataset aimed at high-efficiency action recognition in RoboCup and examine the transfer between simulated and real domains in strongly structured settings. Our dataset includes 3,108 action sequences across four domains with over 183,000 images. Utilizing a recognition model on an Intel-Atom-based NAO robot, we demonstrate enhanced performance by merging real and synthetic data, and efficient learning of new signals with synthetic data updates, reducing acquisition efforts for future RoboCup rule modifications.

1 Introduction

The RoboCup competition, as a platform for testing autonomous systems in a real setting, requires robots to interpret human signals, particularly referee actions [1]. Despite the current emphasis on this research direction within the Standard Platform League, the performance of existing methods during recent research challenges remains low and varies considerably among teams. This can be attributed to the distinct challenges faced in the RoboCup environment as well as the lack of a common dataset that can be used for training recognition models.

In this paper, we approach this goal by providing a comprehensive referee action dataset and investigate the unique constraints and opportunities present in RoboCup. Unlike traditional human action recognition as defined in literature [2, 3, 4, 5, 6, 7, 8], RoboCup's constraints stem from the use of an affordable humanoid robotic platform, leading to issues like using low-cost cameras, latency constraints, and limited compute capacity. To address these challenges, we present a dataset that not only models all referee actions used in the tournament but also utilizes the strengths of the RoboCup environment. Our contributions include:

- A diverse dataset for referee action recognition in RoboCup, covering synthetic, hybrid, and real data in multiple environments.
- An action recognition method utilizing the data to demonstrate its use as a potential benchmark.
- Experiments showing the performance improvement compared to using single-domain data.

2 Related Work

Human Action Recognition (HAR) has been a long-standing problem to be solved in the computer vision community [9, 2, 3, 4, 6, 5, 7, 8, 10, 11, 12]. With the advent of deep learning, Simonyan and Zisserman [5] introduce a two-stream convolutional neural network for action recognition, laying the foundation of deep video action recognition. Subsequent works, such as the two-stream I3D [10], TSN [11], LRCNs [12] make progress on proposing networks to capture spatiotemporal features, with the attention mechanism [13, 14, 15] being introduced recently.

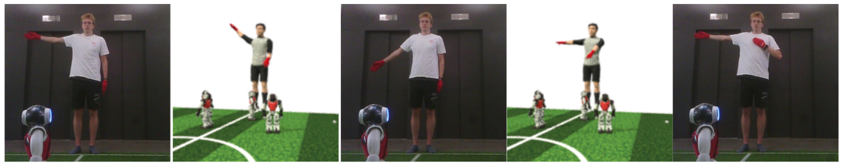


Figure 1: Static referee actions with real and synthetic data.

Human Action Recognition Datasets UCF101 [6], HMDB51 [16], and Kinetics [7] are widely-used video action recognition datasets, which cover a diverse set of human activities. However, collecting and annotating large-scale video datasets requires extensive work, and therefore, synthetic datasets are also used to train visual models for many computer vision tasks [17, 18, 19, 20, 21, 22, 23]. Though models trained with synthetic data show good performance when testing on real-world scenarios [21, 22, 23], a domain gap remains to be an issue [24].

3 Dataset Description

The work aims at enabling referee gesture detection on mobile robots. To this end, we provide a dataset that contains rendered **synthetic** videos, two sets of videos with a **chroma key** background and different acquisition protocols, and a set of **real** videos for benchmarking in a realistic setting. In this work, the real data has been solely used for the purpose of testing. Examples of the gestures contained in our dataset are depicted in Fig 1.

3.1 Real Data - Test Setting

We collect data from the robot cameras at 6 different locations that cover a variety of backgrounds and lighting conditions representative of environments present during RoboCup. To record a single session, the robots are randomly placed on the field with all robots facing the referee, who performs the 12 actions present in RoboCup.

Data Acquisition Challenges

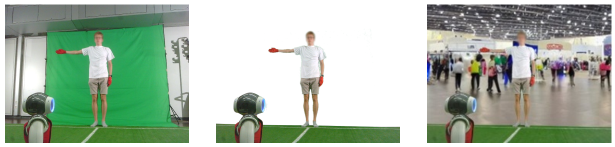
Collecting real data representative of the RoboCup environment is expensive and time-consuming due to extensive annotation, the field setup in different environments, and training individuals for performing the gestures. The real-time robotic framework used during acquisition can compromise further data quality, with issues like frame drops and camera resets causing non-consecutive frames. This disrupts synchronization and adds additional manual annotation effort, increasing the costs and potential human error. Despite these challenges, realistic data is essential for training models that generalize well. In our work, we therefore, examine two solutions: creating fully synthetic data and using chroma key sequences with synthetic backgrounds for training, while saving all fully real data for testing. Further details on these methods are provided in subsequent sections.

3.2 Synthetic Data

We create synthetic data by modeling the 3D environment in the procedural 3D animation framework *Side FX Houdini* that closely resembles the setup during RoboCup. Subsequently, photo-realistic referee action sequences are rendered from diverse camera views, utilizing the flexibility to adjust camera positions, referee poses, models, and textures. This facilitates the efficient creation of a diverse, large-scale dataset with precise and easy annotation of the generated video sequences.

Simulation environment setup The simulation environment is defined by the official field definition and a model of the NAO robot with differently colored jerseys. To represent the referee, we a 3D human models that encompass different body shapes and textures is used.

Robot positions In our simulated setup, robots and cameras are randomly distributed across the field, remaining stationary during a single data session to enable data fusion from multiple robots,



(a) Green Screen. (b) Masked Image. (c) Augmented.

Figure 2: Chrome key data collection and augmentation.

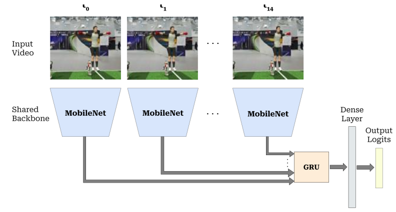


Figure 3: Action recognition pipeline.

with positions randomized between sessions for real data variation. As cameras are distributed over the whole field, certain viewpoints are not suitable for observing the referee's actions. A detailed analysis of the impact of the relative position on action recognition is conducted by categorizing camera positions. Positions with robots' cameras one-quarter field away from the referee and with a view angle below $45°$ are labeled as *easy positions*, others as *hard positions*.

Backgrounds We augment the simulated images with a set of 65 synthetic backgrounds. The backgrounds are generated using Stable Diffusion [25] with prompts representative of the environments encountered during RoboCup such as crowded exhibition centers.

3.3 Real Data - Chroma Key

A high-quality animation framework and raytracing renderer is utilized for generating synthetic data. However, a domain gap still remains, which we approach by collecting additional data from NAO robots. Using a single robot for recording yields one video per location, whichr equired multiple sessions at varied locations to model diverse backgrounds. Thus, we record in front of chroma key backgrounds (Greenscreen), where post-processing allows the insertion of different backgrounds, as shown in Fig 2.

Two data collection methods for chroma key images are employed, differing in the number of robot per session and location. In Chroma Key Front (CK Front), a single robot, positioned directly in front of the referee according to the RoboCup 2022 rules, is used with 9 individuals participating. Extending this, Chroma Key Game (CK Game) follows the RoboCup 2023 rules, utilizing multiple robots in varied field positions, with 5 referees participating.

Chroma Key Front This setting comprises videos from a single robot placed in front of the referee, allowing for an easy background extraction and action recognition. In each session, the chroma key background is manually removed. Adhering to RoboCup 2022 rules, class 12 is absent in this dataset, enabling the study of our approach's learning capabilities with a synthetic data-exclusive class. This can indicate, how much new real data needs to be collected for future rule changes.

Chroma Key Game In this setting, robots are randomly placed on the field, with the layout being changed for each session to provide sufficient variability. Fig 2a shows the view from one of the robots Adobe Premiere has been used to generate a mask of the greenscreen. The same methodology as for annotation of real data has been used, which helps to synchronize annotations between robots.

4 Action Recognition

To gain deep insight into our dataset and to provide a public benchmarking model to all RoboCup teams, we develop an approach for human action recognition designed for low-resource contexts. The method employs a MobileNet [26] architecture for image feature extraction as a backbone. After resizing each image from a window of 15 frames to 90 x 120 px, the corresponding deep feature is extracted. To further capture the temporal relationships among the images, the sequence of 15 deep features is further processed by a GRU [27]. The GRU's 64-dimensional output is directed through 2 subsequent dense layers, each with a preceding Dropout layer [28] and ReLU[29] activation functions. Finally, the class is predicted directly from the logits. Our approach is further depicted in Fig 3 for clarity.

SYN easy	SYN hard	CK front	CK game	Test full	Test easy	Test hard
✓				27.9	33.3	16.1
✓	✓			21.9	25.2	14.6
		✓		30.8	28.0	37.1
			✓	60.4	65.0	50.2
		✓	✓	69.3	74.3	<u>58.4</u>

Table 1: Test accuracy, single-domain training.

SYN easy	SYN hard	CK front	CK game	Test full	Test easy	Test hard
✓		✓		52.8	54.7	48.7
✓			✓	74.4	<u>82.5</u>	56.6
✓		✓	✓	**76.1**	**85.6**	55.4
✓	✓	✓		46.8	52.5	34.5
✓	✓		✓	72.5	81.3	53.2
✓	✓	✓	✓	<u>73.3</u>	79.4	**59.9**

Table 2: Test accuracy, multi-domain training.

5 Experiments and Results

In our experiments, we assessed baseline performance using single-domain training with synthetic, chroma key data. Synthetic data was divided into easy and hard sets (3.2), and evaluation used real test data, also split into easy and hard (3.1). Considering the application of RoboCup, the test easy class is of major interest, as it best represents the current tournament scenario where only the robot locations that are known to have good viewing angles need to be considered for making a decision. Results for single- and multi-domain training are presented in Tables 1 and 2 respectively. In this section, the domains synthetic and chroma key will be indicated by their abbreviations SYN and CK.

Single Domain Performance Investigating the set of single domain experiments in Table 1, the performance improves for an increasing overlap between the training and testing domains. On the full test set, this corresponds to the sequence of SYN, CK Front and CK Game. CK Game has the strongest performance by a large margin, with 60.4% and 65.0% accuracy on test full and easy respectively. SYN and CK front both exhibit a considerably lower performance, which can be attributed to the two different domain gaps. The former has a considerably different image appearance, while the latter covers a much smaller domain of viewing angles. Using SYN hard for training deteriorates results, likely because recognizing actions from hard positions backpropagates incorrect signals, reducing model performance.

Multi-Domain Performance We tested various combinations of SYN, CK Front, and CK Game for multi-domain training to determine optimal data collection and augmentation strategies. This can help in making decisions to extend the dataset when new rules or actions are introduced at RoboCup. These results are provided in Table 2.

Combining SYN and CK Front considerably improves performance despite their individual domain gaps with the test real data. Performance jumps by 24.8% and 22.0% on test full when using them jointly. This improvement can be attributed to the complementary domain gaps which allows the training to cover the full domain when using them together. Further adding the CK Game data allows us to raise the model's accuracy to 76.1%. For our task, this supports the use of a multi-domain dataset, that contains large portions of data that are cheap to generate on a large scale.

Amount of Data As the data collection and annotation require a large amount of resources, we provide an analysis of the model performance on a lower amount of data. The results indicate that even combining SYN data with a single referee from a CK dataset can improve the performance considerably from 27.9% to 70%, which is a promising perspective for data collection.

6 Conclusion

In this work, we presented a new multi-domain referee action dataset that aims at providing the basis for bringing more autonomy to the RoboCup competition. Comprehensive experiments demonstrate that combining different domains improves the performance considerably and allows easy adaptability of the dataset to future rule changes. Finally, the implemented action recognition method is able to run in real-time on low-performance robot hardware and can serve as a baseline to benchmark future approaches.

References

[1] H. Kitano and M. Asada. The robocup humanoid challenge as the millennium challenge for advanced robotics. *Advanced Robotics*, 13(8):723–736, 1998.

[2] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3):257–267, 2001.

[3] Efros, Berg, Mori, and Malik. Recognizing action at a distance. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 726–733. IEEE, 2003.

[4] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding*, 104(2-3):249–257, 2006.

[5] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[6] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[9] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3192–3199, 2013.

[10] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[12] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[13] R. Girdhar and D. Ramanan. Attentional pooling for action recognition. *Advances in neural information processing systems*, 30, 2017.

[14] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487, 2022.

[15] C. Plizzari, M. Cannici, and M. Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, pages 694–701. Springer, 2021.

[16] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[17] C. De Souza, A. Gaidon, Y. Cabon, and A. López. Procedural generation of videos to train deep action recognition networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[18] O. Matthews, K. Ryu, and T. Srivastava. Creating a large-scale synthetic dataset for human activity recognition. *arXiv preprint arXiv:2007.11118*, 2020.

[19] J.-N. Zaech, C. Gao, B. Bier, R. Taylor, A. Maier, N. Navab, and M. Unberath. Learning to Avoid Poor Images: Towards Task-aware C-arm Cone-beam CT Trajectories. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 11–19, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32254-0. doi: 10.1007/978-3-030-32254-0_2.

[20] M. Unberath, J.-N. Zaech, S. C. Lee, B. Bier, J. Fotouhi, M. Armand, and N. Navab. DeepDRR – A Catalyst for Machine Learning in Fluoroscopy-guided Procedures. *arXiv:1803.08606 [physics]*, Mar. 2018.

[21] J. Marin, D. Vázquez, D. Gerónimo, and A. M. López. Learning appearance in virtual scenarios for pedestrian detection. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 137–144. IEEE, 2010.

[22] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[23] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla. Understanding real world indoor scenes with synthetic data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4077–4085, 2016.

[24] X. Peng, B. Usman, K. Saito, N. Kaushik, J. Hoffman, and K. Saenko. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *arXiv preprint arXiv:1806.09755*, 2018.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[26] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017. URL `https://api.semanticscholar.org/CorpusID:12670695`.

[27] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

[29] A. F. Agarap. Deep learning using rectified linear units (relu). *ArXiv*, abs/1803.08375, 2018. URL `https://api.semanticscholar.org/CorpusID:4090379`.