

SVW-UCF Dataset for Video Domain Adaptation

Artjoms Gorpincenko^a and Michal Mackiewicz^b

School of Computing Sciences, University of East Anglia, Norwich, U.K.

Keywords: Dataset, Deep Learning, Domain Adaptation, Video.

Abstract: Unsupervised video domain adaptation (DA) has recently seen a lot of success, achieving almost if not perfect results on the majority of various benchmark datasets. Therefore, the next natural step for the field is to come up with new, more challenging problems that call for creative solutions. By combining two well known sets of data - SVW and UCF, we propose a large-scale video domain adaptation dataset that is not only larger in terms of samples and average video length, but also presents additional obstacles, such as orientation and intra-class variations, differences in resolution, and greater domain discrepancy, both in terms of content and capturing conditions. We perform an accuracy gap comparison which shows that both SVW→UCF and UCF→SVW are empirically more difficult to solve than existing adaptation paths. Finally, we evaluate two state of the art video DA algorithms on the dataset to present the benchmark results and provide a discussion on the properties which create the most confusion for modern video domain adaptation methods.

1 INTRODUCTION

Deep neural network architectures continue to show impressive results across a number of computer vision tasks, such as classification (Szegedy et al., 2017), image generation (Karras et al., 2020), denoising (Tian et al., 2020), and upscaling (Haris et al., 2018). However, many modern methods call for large amounts of labeled training data that might not always be available in real-life scenarios. Collecting and annotating new imagery for the task of interest is often infeasible due to associated costs and time constraints. In theory, a model could be trained on a similar dataset that shares the same task (Sun and Saenko, 2014). In practice, however, neural networks frequently fail to perform well, as they are too sensitive to differences in appearance and image capture conditions (Shimodaira, 2000). This scenario is described as the domain shift problem - where the training data distribution (the source domain) is not aligned with the test data distribution (the target domain) on the data manifold. Domain adaptation (Csurka, 2017) is a branch of transfer learning (Pan and Yang, 2010) which aims to train robust models in the presence of aforementioned misalignments and is categorized based on label availability in the target domain.


Both image and video-based DA have recently

demonstrated great advancements, scoring 95%+ accuracy on a large portion of available benchmark datasets (French et al., 2018; Shu et al., 2018; Mao et al., 2019; Chen et al., 2019; Gorpincenko et al., 2020). However, whereas image-based DA problems include various degrees of domain discrepancy, one would argue that many video data paths do not present the same level of complexity. In fact, a sophisticated neural network architecture can achieve good results even without any adaptation mechanism present in the method itself (details in Section 3 and Table 3), which leads to believe that the lack of challenging problems is one of the main reasons as to why video-based DA has not received as much attention as its counterpart.

In this paper, we introduce a large-scale video-based domain adaptation dataset called SVW-UCF¹ (Safdarnejad et al., 2015; Soomro et al., 2012), which presents a larger domain gap and consists of 25 overlapping categories. We evaluate two state of the art video DA methods (Chen et al., 2019; Gorpincenko et al., 2020) on SVW-UCF, and empirically prove that both SVW→UCF and UCF→SVW paths are more challenging than those which are currently available to public².

¹<https://github.com/ArtjomUEA/SVW-UCF>

²We are aware of the Kinetics-Gameplay dataset (Chen et al., 2019), however, it was released only in feature vector format, which significantly limits its usability in research.

^a  <https://orcid.org/0000-0001-7853-8458>


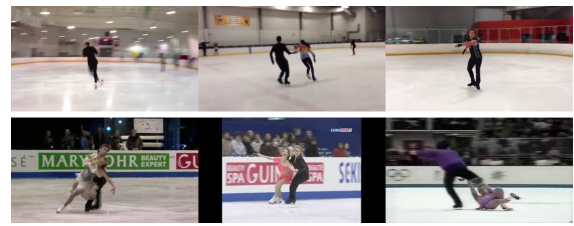
^b  <https://orcid.org/0000-0002-8777-8880>

Table 1: Collected categories for SVW-UCF.

| SVW-UCF | SVW | UCF101 |
|----------------|----------------|--|
| Archery | Archery | Archery |
| Baseball | Baseball | Baseball Pitch |
| Basketball | Basketball | Basketball, Basketball Dunk |
| Biking | BMX | Biking |
| Boating | Rowing | Rowing, Rafting, Kayaking, Skijet |
| Bowling | Bowling | Bowling |
| Discus throw | Discusthrow | Throw Discus |
| Diving | Diving | Diving |
| Figure skating | Skating | Ice Dancing |
| Golf | Golf | Golf Swing |
| Gymnastics | Gymnastics | Floor Gymnastics |
| Hammer throw | Hammer throw | Hammer Throw |
| High jump | High jump | High Jump |
| Javelin throw | Javelin | Javelin Throw |
| Long jump | Long jump | Long Jump |
| Pole vault | Polevault | Pole Vault |
| Punching | Boxing | Boxing-Punching Bag, Boxing-Speed Bag, Punch |
| Shot put | Shotput | Shotput |
| Skiing | Skiing | Skiing |
| Soccer | Soccer | Soccer Juggling, Soccer Penalty |
| Swimming | Swimming | Breaststroke, Front Crawl |
| Tennis | Tennis | Tennis Swing |
| Volleyball | Volleyball | Volleyball Spiking |
| Weight lifting | Weight lifting | Bench Press, Clean and Jerk, Lunges |
| Wrestling | Wrestling | Sumo Wrestling |

2 DATASET DESCRIPTION

To build the dataset, we collected 25 distinct classes that are present in both SVW (Safdarnejad et al.,



(a) figure skating



(b) hammer throw



(c) punch

Figure 1: Snapshots of different categories from the SVW-UCF dataset. For each class, the samples from SVW and UCF101 are in the top and bottom row, respectively.

2015) and UCF101 (Soomro et al., 2012), listed in Table 1. For UCF101, we coupled some of the categories that are similar to each other into one, such as Basketball and Basketball Dunk to form Basketball, or Breaststroke and Front Crawl to form Swimming. In total, SVW-UCF consists of 5878 training and 2410 testing videos, shared between two domains (Table 2). For consistency, we followed the suggested SVW and UCF evaluation protocols^{3,4}, and chose train/test splits №1 for both.

The visual domain gap between SVW and UCF101 mostly comes in the differences between skill levels of sportsmen and camera operators in the videos. Whereas UCF101 mainly consists of clips that were filmed in their natural category environments that might provide additional cues to a classifier, e.g., basketball court or baseball pitch with teams and viewers, it is not the case with SVW. The latter dataset was captured solely with smartphones by amateurs, and contains more casual scenes, such as playing football in a backyard or practicing field penalties on an empty pitch. The gap is further expanded

³<http://cvlab.cse.msu.edu/project-svw.html>

⁴<https://www.crcv.ucf.edu/data/UCF101.php>

Table 2: Comparison of video domain adaptation datasets.

| | UCF-Olympic | UCF-HMDB _{small} | UCF-HMDB _{full} | SVW-UCF |
|-----------------------|-------------|---------------------------|--------------------------|-----------|
| Average length (sec.) | 5.4-10.6 | 5.8-3.3 | 7.2-3.3 | 15.5-6.6 |
| Classes | 6 | 5 | 12 | 25 |
| Training samples | 601-250 | 482-350 | 1438-840 | 2466-3412 |
| Testing samples | 240-54 | 189-150 | 571-360 | 1057-1353 |

by variations in camera angles and vibrations, lighting, and different orientations, which are not present in previously proposed datasets.

From the domain adaptation perspective, it is important to learn semantic information about actions with minimum bias to their visual appearance. For example, both bench press and lunges with dumbbells are considered to be weight lifting, although these exercises look very different from each other. Consequently, activities that represent same actions in a different manner introduce an additional yet positive challenge for this field of research. Hence, we grouped some of the classes that meet these criteria where it was possible, for example, wrestling and sumo wrestling, gymnastics and floor gymnastics. We also note that SVW-UCF has the largest amount of overlapping categories and unique samples, as well as greater average video length when comparing to the existing datasets (Table 2).

3 COMPARISON WITH OTHER WORK

In this section, we first present a set of potential reasons as to why existing video DA datasets do not present enough challenges for modern algorithms. Then, to test our hypotheses, we perform an evaluation of accuracy gaps between adaptation paths.

UCF-Olympic (Jamal et al., 2018) and UCF-HMDB_{small} (Sultani and Saleemi, 2014) were initially designed to have visually similar categories, while the main goal of domain adaptation is to learn semantics of an object or an action, regardless of its appearance (Ganin and Lempitsky, 2015). Indeed, the absence of labels alone without the visual gap present turns this into a semi-supervised learning problem (Zhu and Goldberg, 2009), where the target domain can simply be treated as a portion of unlabelled samples. To ensure that there is a sufficient domain discrepancy in SVW-UCF, we selected datasets that are different in both visual and semantic sense.

Neural network models are very sensitive to inputs, and require large amounts of training points to generalise and develop strong, reliable decision boundaries on the data manifold. All the previ-

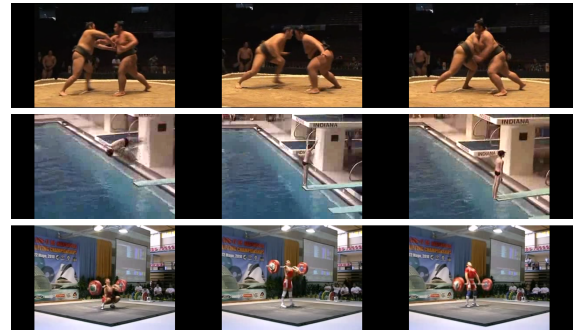


Figure 2: Snapshots of videos from the UCF101 dataset. Each frame belongs to a different sample, although it is clear that they are just small parts of a single, larger video.

ous datasets underdeliver in that regard, on average having less than 100 videos per class, per domain. This problem is further amplified by samples from UCF101, where many groups of clips are comprised of single large videos, and arguably cannot be considered as fully individual, distinct samples, as they share a lot of similar visual properties (Figure 2). We extend the numbers by providing 331.5 clips per class on average (versus second largest 267.4 for UCF-HMDB_{full} (Chen et al., 2019)), as well as greater video length to ensure a significant number of frames in SVW-UCF (Table 2). Clearly, there are methods that were specifically created to aid the requirements of neural networks in cases where the amount of training data is limited (Miyato et al., 2019; Yun et al., 2019), however, that is rather the focus of other research fields, such as unsupervised and semi-supervised learning. As for domain adaptation, it was created to address the problem of the lack of data in the target domain, however, on condition that the source domain has sufficient task-related information (Ganin and Lempitsky, 2015).

To test whether aforementioned properties are important in domain adaptation datasets and deep learning algorithms, we trained the TA³N model in one domain setting, i.e., in a standard, supervised manner, with adaptation mechanisms disabled. Obtained results (Table 3) suggest that the field of video DA needs to shift towards problems with greater visual and semantic discrepancies, as well as more classes and samples, since strong CNN backbone model alone is already enough to achieve good performance on

Table 3: Accuracy gap comparison between different video domain adaptation paths, using TA³N model architecture. U→O stands for UCF→Olympic, U→H_s for UCF→HMDB_{small}, U→H_f for UCF→HMDB_{full}, S→U for SVW→UCF, and vice versa.

| | U→O | O→U | U→H _s | H→U _s | U→H _f | H→U _f | S→U | U→S |
|-------------|--------|--------|------------------|------------------|------------------|------------------|--------|--------|
| Source only | 93.82% | 84.58% | 94.40% | 92.61% | 71.67% | 73.91% | 60.55% | 54.41% |
| Target only | 100.0% | 99.41% | 98.67% | 98.94% | 82.78% | 94.92% | 89.10% | 91.45% |
| Gap | 6.18% | 14.83% | 4.27% | 6.33% | 11.11% | 21.01% | 28.55% | 37.04% |

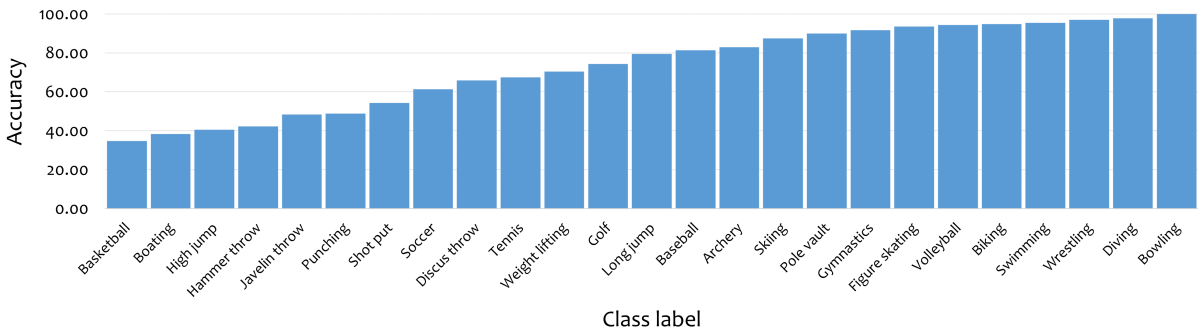


Figure 3: Individual category results of TA³N+VAT on the UCF→SVW adaptation path, sorted by accuracy.

Table 4: Video domain adaptation algorithms evaluation on SVW-UCF.

| Method | SVW→UCF | UCF→SVW |
|-----------------------|---------|---------|
| Source only | 60.55% | 54.41% |
| TA ³ N | 67.85% | 59.98% |
| TA ³ N+VAT | 74.65% | 65.94% |
| Target only | 89.10% | 91.45% |

a large portion of available paths. In addition to that, we also trained and tested the two currently best performing video DA methods: TA³N (Chen et al., 2019) and TA³N+VAT (Gorpincenko et al., 2020) on SVW-UCF (Table 4). Even with adaptation in place, the difference between the highest accuracy and ‘Target only’ is still significant (14.45% and 25.51% for SVW→UCF and UCF→SVW, respectively), which opens up opportunities for future research. When it comes to individual category performance (Figure 3), we found that the lowest scores were obtained in cases where the domain gap mainly consists of visual discrepancies. On UCF frames, classes such as Basketball, High jump, Hammer throw, Javelin throw, Punching, and Shot put mostly have professional pitch/stadium setups and crowd in the background, while many SVW videos of these sports do not share the same properties. On the other hand, activities that do not possess the same degree of freedom when it comes to the place where they can be performed in (e.g., skating always requires an ice rink), generally have higher accuracy: Skiing, Gymnastics, Figure skating, Swimming, Diving, and Bowling. This leads to a conclusion that video DA al-

gorithms still heavily rely on visual cues, while the actions themselves are rather complementary. The above is further supported by a very strong performance on the Wrestling class - even though the domains have slightly different sports constructing the category (Table 1), the presence of the arena makes it easy for the classifier to assign the correct prediction.

4 CONCLUSIONS

In this paper, we presented SVW-UCF - a large-scale dataset for video domain adaptation, which contains more categories and delivers a greater domain discrepancy, when compared to existing datasets. We also evaluated the accuracy gap between all available adaptation paths and found that most of them do not present enough challenge for modern techniques, as good performance can be achieved by simply using a robust neural network model. Finally, we tested two state of the art video DA algorithms on SVW-UCF and looked at the results of the TA³N+VAT method on UCF→SVW in detail. The latter suggests that the field of unsupervised video DA needs to shift towards problems with greater visual gaps, as that is the area where current methods struggle the most.

ACKNOWLEDGEMENTS

The project was jointly funded by Innovate UK (grant #102072), Cefas, Cefas Technology Limited and EDF

Energy, and has also been supported by the Natural Environment Research Council; and Engineering and Physical Sciences Research Council through the NEXUSS Centre for Doctoral Training (grant #NE/RO12156/1).

REFERENCES

- Chen, M.-H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., and Zheng, J. (2019). Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Csurka, G. (2017). *A Comprehensive Survey on Domain Adaptation for Visual Applications*, pages 1–35. Springer International Publishing, Cham.
- French, G., Mackiewicz, M., and Fisher, M. (2018). Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*.
- Ganin, Y. and Lempitsky, V. S. (2015). Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, pages 1180–1189.
- Gorpincenko, A., French, G., and Mackiewicz, M. (2020). Virtual adversarial training in feature space to improve unsupervised video domain adaptation.
- Haris, M., Shakhnarovich, G., and Ukita, N. (2018). Deep back-projection networks for super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1664–1673.
- Jamal, A., Namboodiri, V. P., Deodhare, D., and Venkatesh, K. (2018). Deep domain adaptation in action space. In *BMVC*.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan.
- Mao, X., Ma, Y., Yang, Z., Chen, Y., and Li, Q. (2019). Virtual mixup training for unsupervised domain adaptation.
- Miyato, T., Maeda, S., Koyama, M., and Ishii, S. (2019). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993.
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.
- Safdarnejad, S. M., Liu, X., Udpa, L., Andrus, B., Wood, J., and Craven, D. (2015). Sports videos in the wild (svw): A video dataset for sports analysis. In *Proc. International Conference on Automatic Face and Gesture Recognition*, Ljubljana, Slovenia.
- Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227 – 244.
- Shu, R., Bui, H., Narui, H., and Ermon, S. (2018). A DIRT approach to unsupervised domain adaptation. In *International Conference on Learning Representations*.
- Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild.
- Sultani, W. and Saleemi, I. (2014). Human action recognition across datasets by foreground-weighted histogram decomposition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–771.
- Sun, B. and Saenko, K. (2014). From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *Proceedings of the British Machine Vision Conference*. BMVA Press.
- Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI’17*, page 4278–4284. AAAI Press.
- Tian, C., Xu, Y., Li, Z., Zuo, W., Fei, L., and Liu, H. (2020). Attention-guided cnn for image denoising. *Neural Networks*, 124:117 – 129.
- Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., and Choe, J. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031.
- Zhu, X. and Goldberg, A. (2009). *Introduction to Semi-Supervised Learning*.