Procedural Generation of Videos to Train Deep Action Recognition Networks

César Roberto de Souza[12], Adrien Gaidon[3], Yohann Cabon[1], Antonio Manuel López[2]

[1] NAVER LABS Europe, France; [2] Centre de Visió per Computador & UAB, Spain, [3] Toyota Research Institute, USA

CVPR July 21-26 2017 HONOLULU

http://adas.cvc.uab.es/phav

Introduction

- **Problem:** labeled video data bottleneck
  - Cost: manual video annotation is expensive
  - Bias: lack of diversity and variations
- **Solution:** synthetic video generation
  - Modern computer graphics and game engines
  - Realistic, efficient, and scriptable simulation

- Contributions
  - A **procedural generative model** of human action videos
  - **Physically plausible variations** of scenes and actions
  - **Large synthetic dataset** with real and procedural actions
  - Experimental validation as a data supplement for training deep action recognition models

Procedural Human Action Videos (PHAV)

Action clips:	39,982
Categories:	35
Clips / class:	1,142.34
Clip duration:	1-10s
Total duration:	2d 07h 31m
Frames / modality:	5,996,286
Data modalities:	6

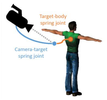
Generating a Large Synthetic Dataset for Action Recognition

- Scene composition
  - *"A human action video contains a **protagonist** performing an **action** in an **environment**, under particular **weather conditions** at a specific **period of the day**. There can be one or more **background actors** in the scene, as well as one or more **supporting characters**. The scene is filmed using a **camera behaviour**."*

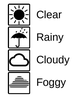
- Protagonist

- Environment (regions of the virtual world)

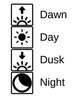
- Camera



- Weather



- Period

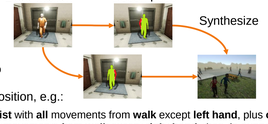


- Actions
  - **sub-HMDB (21):** brush hair, catch, clap, climb stairs, golf, jump, kick ball, push, pick, pour, pull up, run, shoot ball, shoot bow, shoot gun, sit, stand, swing baseball, throw, walk, wave
  - **One person synthetic (10):** car hit, crawl, dive floor, flee, hop, leg split, limp, moonwalk, stagger, surrender
  - **Two people synthetic (4):** walking hug, walk hold hands, walk the line, bump into each other

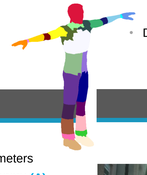
- Adding variations
  - Limited-time physics simulations for human body (**ragdoll physics**)
  - Random perturbations
  - Muscle weakening
  - Interaction with objects

Human ragdoll with 15 parts

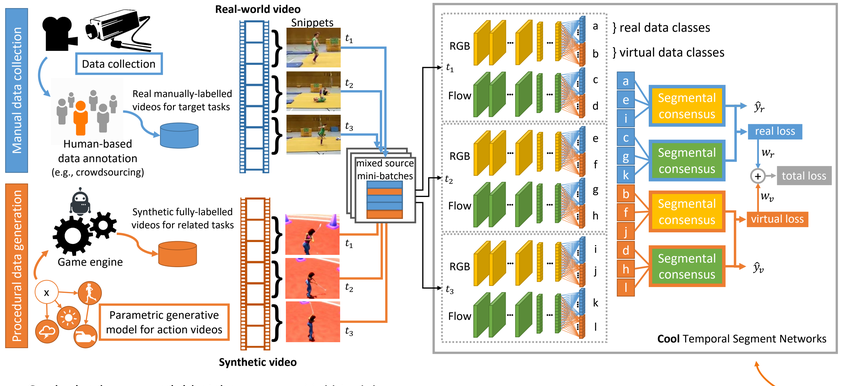
- Synthetic action generation
  - Library of atomic actions, e.g.:
    - Left arm movement during walk
    - Upper right leg movement during jump



  - New actions by programmatic composition, e.g.:
    - Hold hands → Place a **protagonist** with **all** movements from **walk** except **left hand**, plus **one supporting character** with **all** movements from **walk** except **right hand,** then tie puppet hands together
    - Flee → Place a **protagonist** with **leg** movements from **run**, raise puppet hands
    - Limp → Place a **protagonist** with **leg** movements from **walk**, increase physical weight of left leg



Cool Temporal Segment Networks



- Cool mixed-source mini-batches
  - Virtual-world and real-world frames mixed together in the same minibatch
  - Two paths after last feature layer:
    - Prediction of **virtual** dataset class labels
    - Prediction of **real** dataset class labels
  - Different losses for each data path

- Multi-task loss

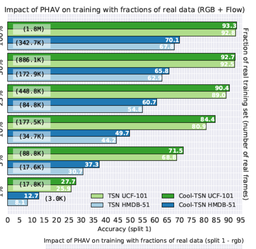
$$\mathcal{L}(y, \boldsymbol{G}) = \sum_{z \in \{real, virtual\}} \delta_{\{y \in C_z\}} w_z \mathcal{L}_z(y, \boldsymbol{G})$$

$$\mathcal{L}_z(y, \boldsymbol{G}) = -\sum_{i \in C_z} y_i \left(G_i - \log \sum_{j \in C_z} \exp G_j \right)$$

  - where:
    - z indexes the source dataset of the video
    - w_z is a loss weight (e.g. relative proportion of z in the mini-batch)
    - C_z is the set of action categories for dataset z
    - $\delta_{\{y \in C_z\}}$ is the indicator function that returns one when label y belongs to C_z, and zero otherwise

Experiments

Target	Model	Spatial	Temporal	Full
PHAV	TSN	65.9	81.5	82.3
UCF-101	[75]	85.1	89.7	94.0
UCF-101	TSN	84.2	89.3	93.6
UCF-101	TSN-FT	86.1	89.7	94.1
UCF-101	Cool-TSN	86.3	89.9	**94.2**
HMDB-51	[75]	51.0	64.2	68.5
HMDB-51	TSN	50.4	61.2	66.6
HMDB-51	TSN-FT	51.0	63.0	68.9
HMDB-51	Cool-TSN	53.0	63.9	**69.5**

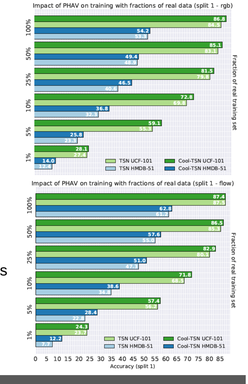


- Procedural generation can act as
  - Strong prior during model initialization: **fine-tuning**
  - Regularizer during network training: **multi-task**
  - Surrogate for missing data samples: **fractioning**

- **Fine-tuning** from PHAV to the real world
  - Improvements for all modalities (RGB and flow)
  - Does not require storing virtual training data

- **Multi-task** loss mixing virtual and real data
  - Improvements for all modalities (RGB and flow)
  - Better results for smaller datasets (HMDB-51)

- **Fractioning** real world data sets into smaller sets
  - What performance improvements we could expect if we had only a fraction of the real world training data?
  - Improvements of up to 6.6 percent points on small data

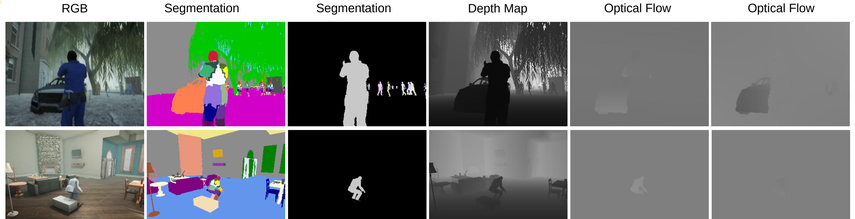


Procedural Generation with an Interpretable Parametric Generative Model

Generative model

- Interpretable parameters
  - Action category (A)
  - Human model (H)
  - Environment (E)
  - Weather (W)
  - Motion variation (V)
  - Period of the day (D)
  - Camera behavior (C)
  - Base motion (B)
  - Video length (L)
    - Random perturbations
    - Animation motion blending
    - Ragdoll muscle weakening
    - Objects / inv. kinematics

Data modalities and annotations



- Annotations
  - Extrinsic and intrinsic camera parameters
  - Actor location in camera coordinates
  - 2D bounding boxes in screen coordinates
  - 3D bounding boxes in world coordinates
  - Body joint locations in screen coordinates (pose)
  - Physical properties of body parts (weight, strength, and limits)
  - Detailed information about the generation parameters: environment, weather, day phase, location, virtual objects

Conclusion

	Method	UCF-101 %mAcc	HMDB-51 %mAcc
One source	iDT+FV [73]	84.8	57.2
	iDT+StackFV [43]	-	66.8
	iDT+SFV+STP [72]	86.0	60.1
	iDT+MIFS [30]	89.1	65.1
	VideoDarwin [13]	-	63.7
Multiple sources	2S-CNN [53]	88.0	59.4
	TDD [74]	90.3	63.2
	TDD+iDT [74]	91.5	65.9
	C3D+iDT[62]	90.4	-
	Actions~Trans [76]	92.0	62.0
	2S-Fusion [12]	93.5	69.2
	Hybrid-iDT [9]	92.5	70.4
	TSN-3M [75]	94.2	69.4
	VGAN [70]	52.1	-
	Cool-TSN	94.2	69.5

- Procedural generation is useful for action recognition
- Strong physical priors present in game and physics engines can be leveraged to train deep neural networks
- Procedural generation can be controlled by a generative probabilistic graphical model of human action videos
- Quantitative evidence that virtual samples can act as drop-in complement for small datasets
- Not necessary to generate particular action categories to obtain performance improvements
- New perspectives for video modelling and understanding

References

9. C. R. de Souza, A. Gaidon, E. Vig, and A. M. López. Sympathy for the Details: Dense Trajectories and Hybrid Classification Architectures for Action Recognition. In ECCV, pages 1–27, 2016.
12. C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition. In CVPR, 2016.
13. B. Fernando, E. Gavves, M. Oramas, A. Ghodrati, T. Tuytelaars, and K. U. Leuven. Modeling video evolution for action recognition. In CVPR, 2015.
30. Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In CVPR, 2015.
43. X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In ECCV, 2014
53. K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.
62. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In CVPR, 2014.
70. C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In NIPS, 2016.
72. H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. IJCV, pages 1–20, July 2015.
73. H. Wang and C. Schmid. Action recognition with improved trajectories. In ICCV, 2013.
74. L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In CVPR, 2015.
75. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In ECCV, 2016.
76. X. Wang, A. Farhadi, and A. Gupta. Actions ~ Transformations. In CVPR, 2015.