Disentangling Motion, Foreground and Background Features in Videos

Xunyu Lin*
Beihang University
Beijing, China
xunyulin2017@outlook.com

Víctor Campos
Barcelona Supercomputing Center
Barcelona, Catalonia/Spain
victor.campos@bsc.es

Xavier Giro-i-Nieto
Universitat Politecnica de Catalunya
Barcelona, Catalonia/Spain
xavier.giro@upc.edu

Jordi Torres
Barcelona Supercomputing Center
Barcelona, Catalonia/Spain
jordi.torres@bsc.es

Cristian Canton Ferrer
Facebook
Seattle (WA), USA
ccanton@fb.com

Abstract

This paper instroduces an unsupervised framework to extract semantically rich features for video representation. Inspired by how the human visual system groups objects based on motion cues, we propose a deep convolutional neural network that disentangles motion, foreground and background information. The proposed architecture consists of a 3D convolutional feature encoder for blocks of 16 frames, which is trained for reconstruction tasks over the first and last frames of the sequence. The model is trained with a fraction of videos from the UCF-101 dataset taking as ground truth the bounding boxes around the activity regions. Qualitative results indicate that the network can successfully update the foreground appearance based on pure-motion features. The benefits of these learned features are shown in a discriminative classification task when compared with a random initialization of the network weights, providing a gain of accuracy above the 10%.

1. Introduction

Unsupervised learning has long been an intriguing field in artificial intelligence. Human and animal learning is largely unsupervised: we discover the structure of the world by observing it, not by being told the name of every object, which is the typical method in supervised learning [12]. A system capable of predicting what is going to happen by just watching large collections of unlabeled video data needs to build an internal representation of the world and its dynamics [3]. When considering the vast amount of unlabeled data which is generated every day in contrast to its annotated counterpart, unsupervised learning becomes one of the key challenges to solve in the road towards general artificial intelligence.

Based on how a human would provide a high level summary of a video, we hypothesize that there are three key components to understand such content, namely *foreground*, *motion* and *background*. These three elements would tell us, respectively, what are the main objects in the video, what are the doing and where is it happening. We propose a framework that explicitly disentangles these three components in order to build strong features for action recognition, where the supervision signals can be generated without requiring from expensive and time consuming human annotations. The proposal is inspired in how infants who have no prior visual knowledge tend to group things that move as connected wholes and also move separately from one another [1]. Based on this intuition, we can build a similar unsupervised pipeline to segment foreground and background with global motion, i.e. the rough moving directions of objects. Such segmented foregrounds across the video can be used to model both the global motion (e.g. transition or stretch) and local motion (i.e. transformation of detailed appearance) from a pair of foregrounds at different time steps. Since background motion is mostly given by camera movements, we restrict the use of motion to the foreground and rely on appearance to model the former.

The contributions of this work are two-fold: (1) disentangling motion, foreground and background features in videos by human alike motion aware mechanism and (2) learning strong video features that improve the performance of action recognition task.

2. Related Work

Leveraging large collections of unlabeled videos has proven beneficial for unsupervised training of image models

*This work was developed during an exchange period of Xunyu Lin at the Universitat Politecnica de Catalunya.

thanks to the implicit properties they exhibit in the temporal domain, e.g. visual similarity between patches in consecutive frames [10] and temporal coherence and order [4]. Since learning to predict future frames forces the model to construct an internal representation of the world dynamics, several works have addressed such task by predicting global features of future frames with Recurrent Neural Networks (RNN) [7] or pixel level predictions by means of multi-scale Convolutional Neural Networks (CNN) trained with an adversarial loss [3]. The key role played by motion has been exploited for future frame prediction tasks by explicitly decomposing content and motion [8] and for unsupervised training of video-level models [2]. Similarly in spirit, separate foreground and background streams have been found to increase the quality of generative video models [9].

Techniques exploiting explicit foreground and background segmentations in video generally require from expensive annotation methods, limiting their application to labeled data. However, the findings by Pathak et al. [5] show how models trained on noisy annotations learn to generalize and perform well when finetuned for other tasks. Such noisy annotations can be generated by unsupervised methods, thus alleviating the cost of annotating data for the target task. In this work we study our proposed method by using manual annotations, whereas evaluating the performance drop when replacing such annotations with segmentations generated in an unsupervised manner remains as future work.

3. Methodology

We adopt an autoencoder-styled architecture to learn features in an unsupervised manner. The encoder maps input clips to feature tensors by applying a series of 3D convolutions and max-pooling operations. Unlike traditional autoencoder architectures, the bottleneck features are partitioned into three splits which are then used as input for three different reconstruction tasks, as depicted in Figure 1.

Disentangling of foreground and background: depending on the nature of the training data, reconstruction of frames may become dominated either by the foreground or background. We explicitly split the reconstruction task to guarantee that none of the parts dominates over the other. Partitioned foreground and background features will be passed into two different decoders for reconstruction. While segmentation masks are often obtained by manual labeling, it is worth noting they can be obtained without supervision as well, e.g. by using methods based on motion perceptual grouping such as uNLC [5]. The latter approach has proven beneficial for unsupervised pre-training of CNNs [5].

Disentangling of foreground motion: leveraging motion information can provide a boost in action recognition performance when paired with appearance models [6]. We encourage the model to learn motion-related representations by solving a predictive learning task where the foreground in the last frame needs to be reconstructed from the foreground in the first frame. Given a pair of foregrounds at timesteps t_1 and t_2, namely (f_{t_1}, f_{t_2}), we aim to estimate a function M from motion features $m_{t_1 \to t_2}$ throughout t_1 and t_2 that maps f_{t_1} to f_{t_2} in deep feature space G:

$$G\left(f_{t_2}\right) = M\left(G(f_{t_1}), m_{t_1 \to t_2}\right) \tag{1}$$

Throughout this work, the space of encoded features is used for G and M is parametrized by a deterministic version of cross convolution [11]. The foreground decoder weights are shared among all foreground reconstruction task. Gradients coming from the reconstruction of f_{t_2} are blocked from backpropagating through $G(f_{t_1})$ during training to prevent $G(f_{t_1})$ from storing information about f_{t_2}.

Frame selection: assuming that the background semantics stay close throughout the short clips, only the background in the first frame is reconstructed. First and last frames are chosen to perform foreground reconstruction, since they represent the most challenging pair in the clip.

Loss function: the model is optimized to minimize the L1 loss between the original frames and their reconstruction:

$$L_{rec} = L_{ffg} + L_{fbg} + L_{lfg} \tag{2}$$

where L_{ffg}, L_{fbg} and L_{lfg} represent the reconstruction loss for the first foreground, first background and last foreground, respectively. During preliminary experiments, we observed that the reconstruction of first foreground always outperformed the reconstruction of the last one by a large margin, given the increased difficulty of the latter task. In order to get finer reconstruction of the last foreground, we introduce an L2 loss L_{feat} on $G(ft_2)$. The pseudo ground truth for this task is obtained by getting first foreground features from the encoder fed with the temporally reversed clip. The final loss to optimize is the following:

$$L_{total} = L_{rec} + L_{feat} \tag{3}$$

4. Experimental setup

Please note again we are showing results trained with ground truth masks to check the feasibility of our proposal and the pure unsupervised framework generating masks from uNLC [5] remains as future work.

Dataset: there are 24 classes out of 101 in UCF-101 with localization annotations. Following [5], we first evaluate the proposed framework with supervised annotations and use the bounding boxes in the subset of UCF-101 for such purpose. Evaluating the proposal in weak annotations collected by means of unsupervised methods remains as future work. We follow the original splits of training and test set and also split 10% videos out of the training set as validation set in order to perform early stopping and prevent the network from overfitting the training data.

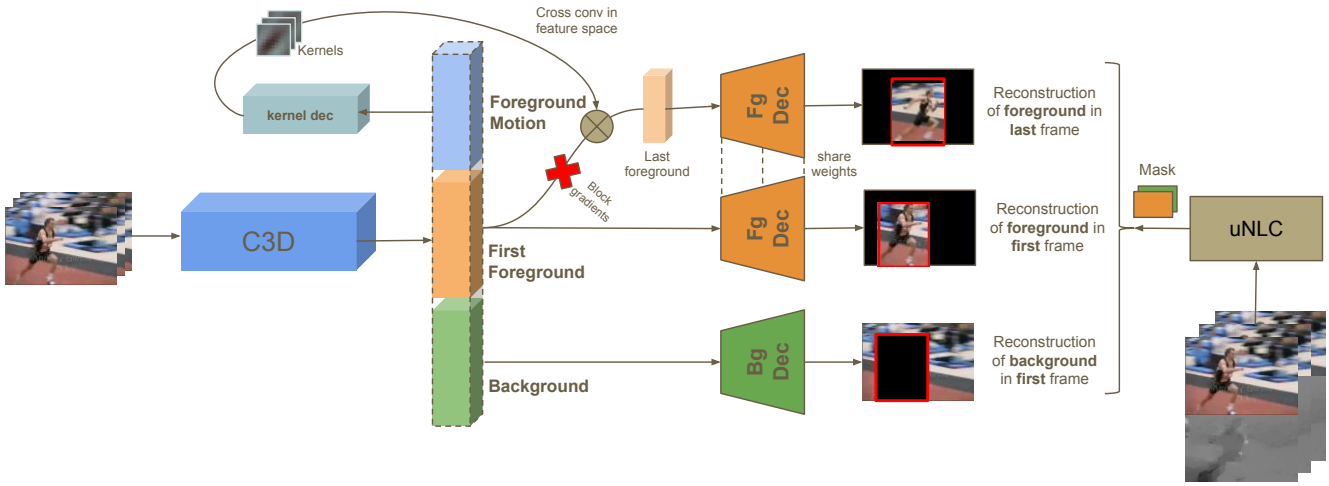

Figure 1: System architecture. Please note that in this work, the masks used to generate ground truth are from manual annotations while uNLC will be utilized in our future work.

Training details: videos are split into clips of 16 frames each. These clips are then resized to 128×128 and their pixel values are scaled and shift to $[-1, 1]$. The clips are randomly temporally or horizontally flipped for data augmentation. Weight decay with rate of 10^{-3} is added as regularization. The network is trained for 125 epochs with Adam optimizer and a learning rate of 10^{-4} on batches of 40 clips.

5. Results

We tested our model on test set for reconstruction task. For better demonstrating the efficiency of our proposed pretraining pipeline, we also trained the network to do action recognition with pretrained features.

Reconstruction task: reconstruction results on test set are shown in Figure 2. From these results, we can clearly see that the network already can predict similar foreground segmentation as ground truth. However, the image reconstructions are still blurry. We argue that this is due to the properties of the L1 loss we are adopting [3]. One interesting fact is that the network has learned to generalize foreground to some other moving objects in the scene even though they are not included in the annotations. For example, the result shown in the top-right corner: instead of only segmenting the person, the dog walking beside the person is also included. This fact suggests that the network has successfully learned to identify foreground from motion cues.

Besides from foreground and background features, these results also demonstrate a good extraction of motion features. The learned motion features contain both global motions, e.g. transition of foreground, and local motions, e.g. change of human pose. In the bottom-center result, the generated kernels from motion feature successfully shift the object from right to the left and change its gesture.

Action recognition: a good pretraining pipeline should show better performance on some typical discriminative tasks than random initialization, especially when training data is scarce [4, 2, 5, 9, 10]. We also conducted comparative experiments on the task of action recognition. By discarding the decoders in our framework and training a linear softmax layer on top of the disentangled features, we can obtain a simple network for action recognition. For the first experiment, we first pretrain our encoder on the subset of UCF-101 with the settings discussed above and then fine-tune the whole action recognition network with added softmax layer on the same dataset. As a baseline, we trained another action recognition network with all weights initialized randomly.

During training, we observed that the pretrained model reached 90% accuracy on training set immediately after one epoch while the randomly initialized network took 130 epochs to achieve it. Both of the network reached around 96% accuracy in later training phase and encountered severe overfitting problems. The accuracy of different methods on the validation set during training time is shown in Figure 3. The best accuracy obtained on the test set with the pretrained model is 62.5%, while it drops to 52.2% when using a random initialization as depicted in Table 1. We observe a margin of more than 10% on accuracy between our proposed method and random initialization on both validation set and test set. This further demonstrates that with our proposal, the network can learn better generalized features. These results are specially promising given the reduced size of the dataset used during pretraining, which is just a fraction of UCF-101. While this demonstrates the efficiency of the approach, using a larger dataset for pretraining should provide additional gains and better generalization capabilities.

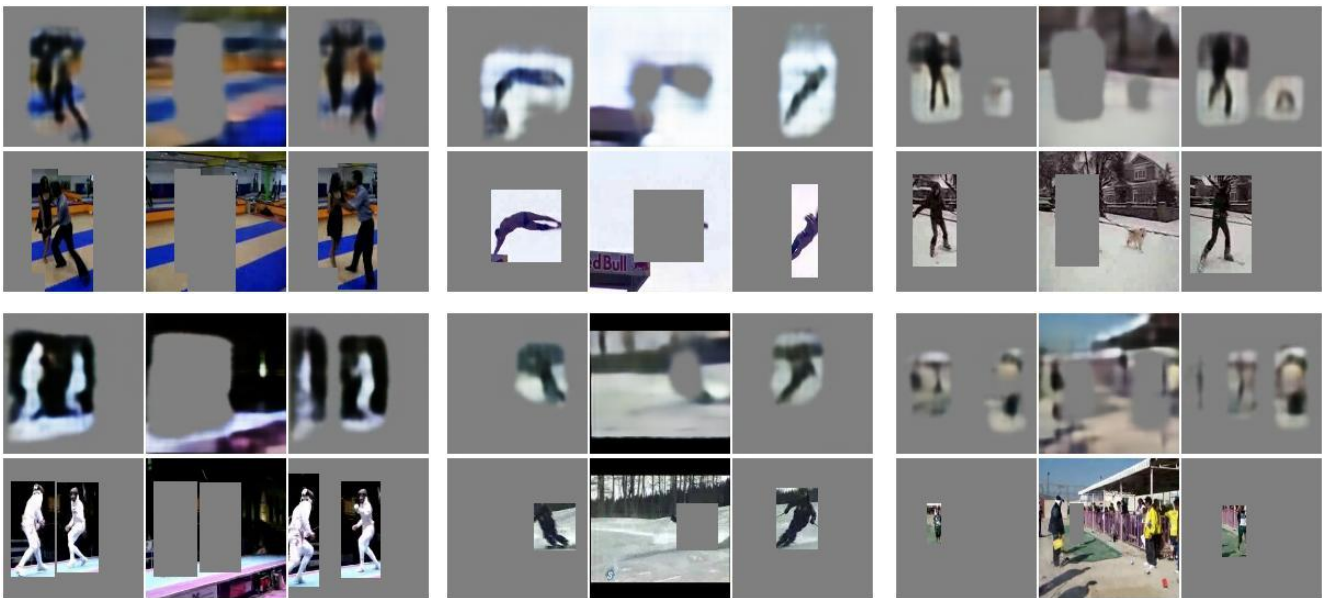

Figure 2: Reconstruction results on the test set. For each example, the top row shows the reconstruction while the bottom one contains the ground truth. Each column shows the segmentation of foreground in first frame, background in first frame and foreground in last frame, respectively.

Method	Accuracy
Random initialization	52.2%
Pretrained (ours)	**62.5%**

Table 1: Test action recognition accuracy of different methods on the subset of UCF-101.

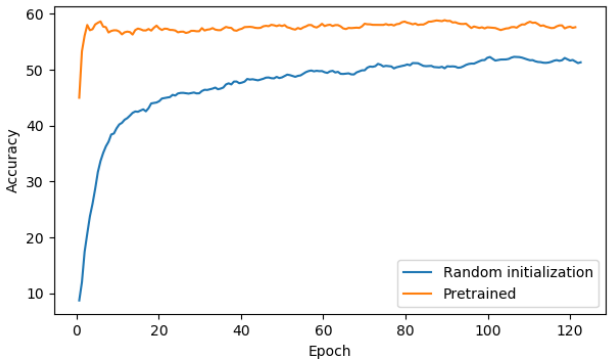

Figure 3: Action recognition results on validation set. This figure shows the accuracy of each method on validation set during the training time.

6. Conclusions

This work has proposed a novel unsupervised framework to disentangle motion, foreground and background features in videos. Our method mostly exploits motion in videos and is inspired by human perceptual grouping with motion cues. Our experiments using ground truth boxes render convincing results on both frame reconstruction and action recognition, showing the potential of the proposed architecture.

However, multiple aspects still need to be explored in our work. As our plans for the future work, we decide to (1) introduce unsupervised learning for foreground segmentation as well, as proposed in uNLC [5]; (2) train with a larger amount of unlabeled data; (3) introduce adversarial loss to improve the sharpness of the reconstructed frames [3]; and (4) fill the gap of absent motion features between the first frame and the last frame by reconstructing any random frame in the clip.

References

[1] S. S. Elizabeth. Principles of object perception. In *Cognitive Science*, 1990.

[2] Z. Luo, B. Peng, D.-A. Huang, A. A. Alahi, and L. Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *CVPR*, 2017.

[3] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016.

[4] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, 2016.

[5] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. In *CVPR*, 2017.

[6] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[7] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015.

[8] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017.

[9] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *NIPS*, 2016.

[10] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015.

[11] T. Xue, J. Wu, K. L. Bouman, and W. T. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*, 2016.

[12] L. Yann, B. Yoshua, and H. Geoffrey. Deep learning. In *Nature*, 2015.