Recognize Actions by Disentangling Components of Dynamics Supplementary Materials

Yue Zhao[1], Yuanjun Xiong[2], and Dahua Lin[1]

[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong [2]Amazon Rekognition

1. Network Details

In this section, we provide the architectural details of the proposed model. We plot the network architecture in Figure 1. Input to the network is a sequence of 8 consecutive RGB frames. The architecture is based upon the BN-Inception model [1]. The first spatial convolutional layer of the BN-Inception is used as the shared lower-level feature extractor for all three branches. We set the stride to 1 to have a displacement map with the same size with the original input.

1.1. Static Appearance Branch

The static appearance branch mostly follows the BN-Inception, except that (1) the `pool_2` layer with spatial Max-pooling is replaced with a 3D version with kernel size $3 \times 3 \times 2$ and stride $2 \times 2 \times 2$; (2) An extra temporal Max pooling layer with kernel $1 \times 1 \times 2$ and stride $1 \times 1 \times 2$ is inserted after `inception_3a` and `inception_4e`. This results in a gradual decrease of temporal resolution from 8, 4, 2, to 1.

1.2. Apparent Motion Branch

The apparent motion branch receives the low-level feature maps and calculates a 4-D cost-volume of size $224 \times 224 \times 11 \times 11$ for each neighboring image pair. The cost volume is then transformed into a displacement map. 7 displacement maps calculated from 8 consecutive frames are stacked into a 14-channel motion representation, which is processed by a subnet similar to BN-Inception [1].

1.3. Appearance Change Branch

The appearance change branch starts with a warping module, which takes in a feature map and reference displacement map calculated from the apparent motion branch. Then we calculate the warped difference by subtracting the feature map at time t from the feature map at time $t-1$ warped with the displacement map at time $t-1$. The warped differences from consecutive frames are stacked as well and processed by a subnet similar to BN-Inception [1].

2. Visualization

In this section, we present a qualitative study by visualizing the intermediate results. As shown in Figure 2, the displacement map can preserve the apparent motion information in most cases. However, the displacement map contains much more noise than the optical flow calculated by TV-L1 [2]. This is because in TV-L1[2] the optimization objective includes a regularization term which favors smoothness and penalizes abrupt motion. In the cost-volume formulation, however, we do not enforce any smoothness constraint. The deep subnet that follows is expected to filter out the distraction of noise and focus on the real motion. In experiments, we did observe that there is just a slight decrease of performance in these cases. It is important to note that the displacement map fails in those regions that either have a large area of similar background color or have wavy texture which evolves with time. In the former case, the small receptive field of the convolutional layer, on which the cost volume is built upon, makes it difficult to track large displacement. In the latter case, the wavy texture, for example, grass, water, or snow makes pixel-level matching between frames impossible without introducing smoothness assumption.

We also compare the method between the naive RGB difference and the motion-warped RGB difference. Compared with RGB difference, motion-warped RGB difference focuses more on the change of appearance, reflected by the more distinctive edges of moving objects. We can also observe that RGB difference warped by the displacement map is visually similar to the RGB difference warped by TV-L1.

References

[1] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift.

In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015. 1

[2] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. *Pattern Recognition*, pages 214–223, 2007. 1

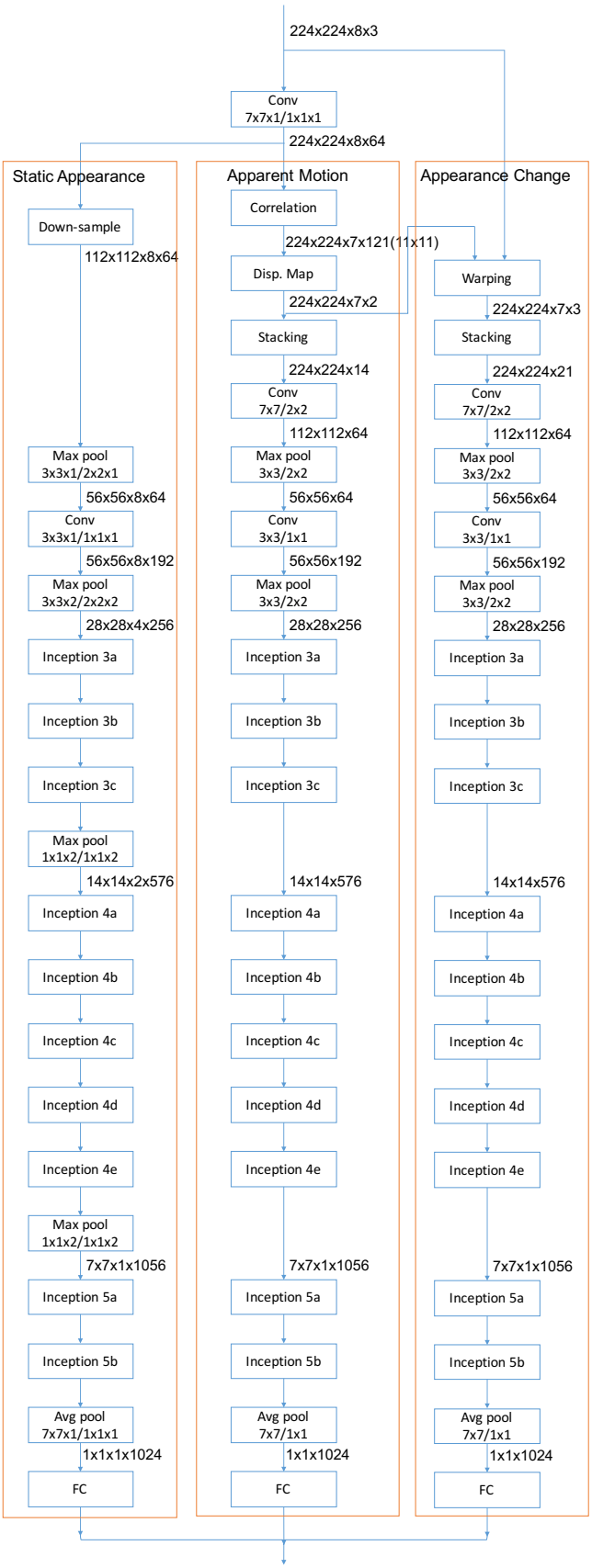

Figure 1. The illustration of the network structure.

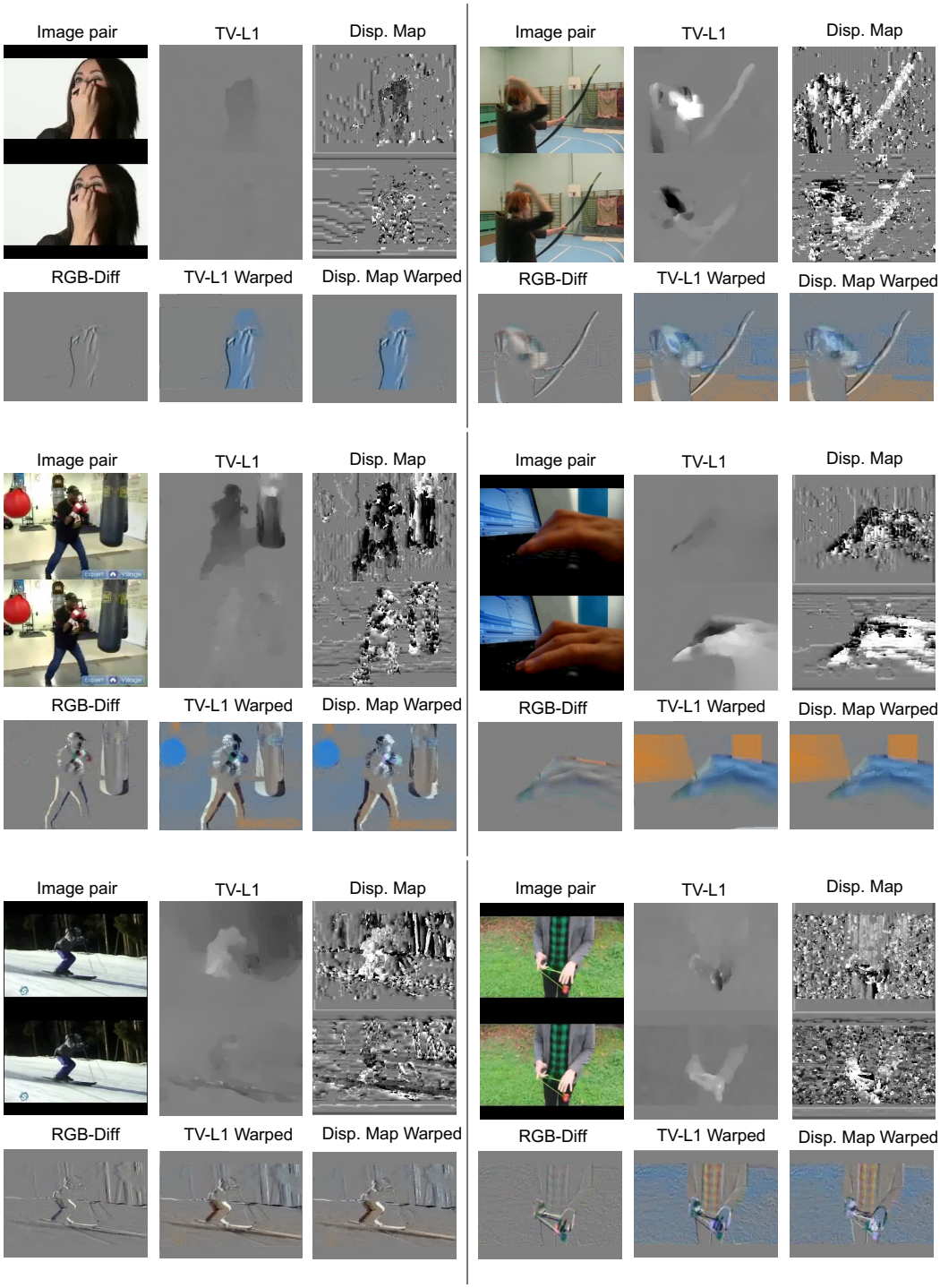

Figure 2. Visualization of the displacement map and warped RGB-difference. For each image pair (upper-left), the TV-L1 optical flow (upper-middle), the cost-volume-based displacement map (upper-right) is compared. In the lower part, RGB-difference without warping (lower-left), warped by TV-L1 flow (lower-middle) and displacement map (lower-right) are compared.