Video Modeling with Correlation Networks

Heng Wang, Du Tran, Lorenzo Torresani, Matt Feiszli

The paper proposes a novel correlation operator to learn a frame-to-frame spatiotemporal matching, where for each pixel in frame A, we compute its correlation within the K x K neighborhood in the corresponding pixel in frame B.

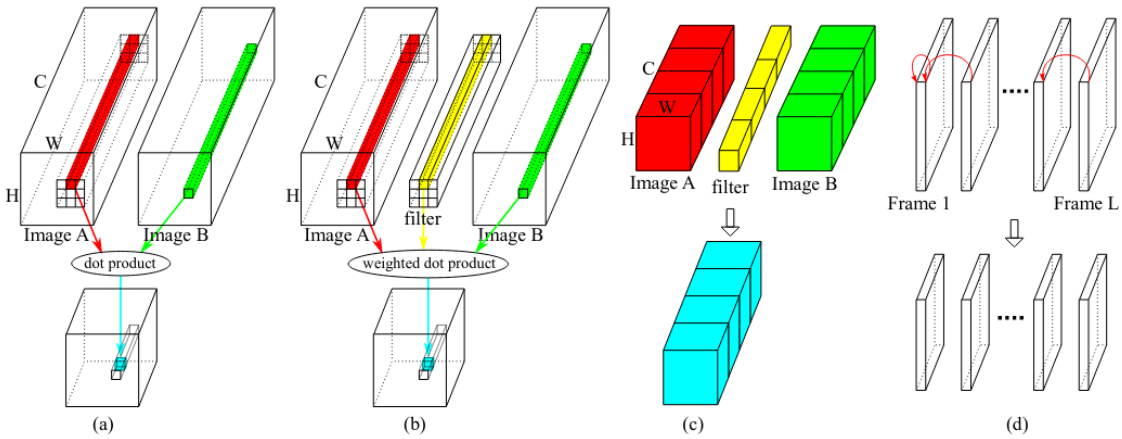

Figure 1: An illustration of the proposed correlation operator. (a) Correlation operator used for optical flow and geometric matching. (b) The introduction of filters renders the operator "learnable." (c) Groupwise correlation increases the number of output channels without adding computational cost. (d) Extending the correlation operator to work on a sequence of video frames.

More precisely, for two feature maps of two images A and B of size C x H x W, first we define a K x K correlation region; then for each spatial location in A, we compute the correlation of the C-d vector at that location and the K x K region centered at the same location in B. So for each location in A, we compute K x K correlations, and by doing this over the whole input; the final volume is of size (K x K) x H x W, where in this case (K x K) plays the role of the channel dimension. To make this operation learnable, the K x K is first convolved with a filter of the same size before computing the correction. To consider larger K x K regions, the paper proposes a dilated version, where the correlation is computed over a dilated region to a larger receptive field. Additionally, to maintain the case number of channels (C = K x K), they propose a group separated region, where the inputs and filters are grouped into G groups, and the correlation is computed for each group separately where G x K x K = C. Finally, the operation is also applied temporally where the operation is computed for every adjacent pair in the input sequence. The operation can be integrated into resblocks by replacing the 3x3x3 conv with 7x7 correlation.