Supplemental material for Flip-Invariant Motion Representation

Takumi Kobayashi
National Institute of Advanced Industrial Science and Technology
Umezono 1-1-1, Tsukuba, Japan
takumi.kobayashi@aist.go.jp

In this supplemental material[1], we detail the procedure of the proposed method for achieving flip-invariant descriptors. As shown in Table 2 of the main manuscript, there are four types of methods: {patch-level invariance (Sec.3.1), descriptor-level invariance (Sec.3.2&3.3)} $\times$ {hand-crafted, ConvNet}-descriptors. The methods of the patch-level invariance are graphically depicted in Fig. A&B, while the descriptor-level invariance for the hand-crafted and ConvNet descriptors is shown in Fig. C. Each component in those methods is implemented as follows.

Improved dense trajectory [36]: The dense trajectories are first extracted from an input video. We use the toolbox[2] provided by the authors [36] to compute *improved* dense trajectories with the default parameter values suggested in the toolbox; *e.g.*, spatial sampling with 5 pixel interval, 8 image scales and 15 frames for trajectory length.

Hand-crafted descriptors: As described in Sec.2, there are five types of the hand-crafted local descriptor, HOG, HOF, MBH_x/MBH_y and LMS. Except for LMS, we compute the descriptors by following the implementation in the dense-trajectory toolbox[2]; on $2_{(X)} \times 2_{(Y)} \times 3_{(T)}$ volume grids in a trajectory of 32×32 spatial extent, 8 orientation bins for HOG and MBH_x/MBH_y, and 9 orientation bins including one null direction for HOF. The LMS is computed on $3_{(X)} \times 3_{(Y)} \times 3_{(T)}$ volume grids in a trajectory of 32×32 spatial extent.

ConvNet descriptors [37]: For a learning-based descriptor which contrasts with the hand-crafted ones, we compute the trajectory-pooled dense ConvNet descriptor [37] by using the toolbox[3] provided by the authors [37]. While basically following the default procedure suggested in the toolbox, as described in Sec.4, we pool the ConvNet activations around the dense trajectories computed as above, for fair comparison to the hand-crafted descriptors.

PCA: As described in Sec.4, the descriptors are projected into 64-dimensional subspace via PCA which is trained on the training samples.

iFK [27]: As described in Sec.4, we apply improved Fisher kernel encoding (iFK) [27] with 256 GMMs on the 64-dimensional space of descriptors.

linear SVM [34]: The linear SVM [34] is learned by libSVM toolbox[4].

[1] We use the same reference numbers as in the main manuscript, such as for the sections and the bibliographies shown in the main manuscript.
[2] http://lear.inrialpes.fr/~wang/download/improved_trajectory_release.tar.gz
[3] https://github.com/wanglimin/TDD
[4] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

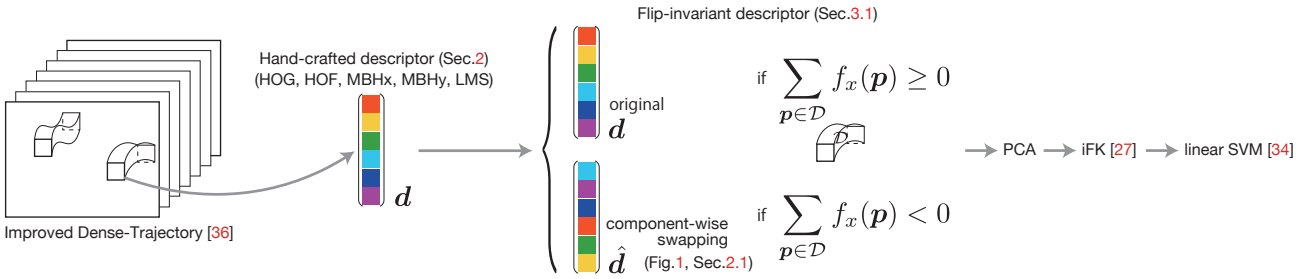

Figure A. Patch-level invariance (Sec.3.1) for hand-crafted descriptors.

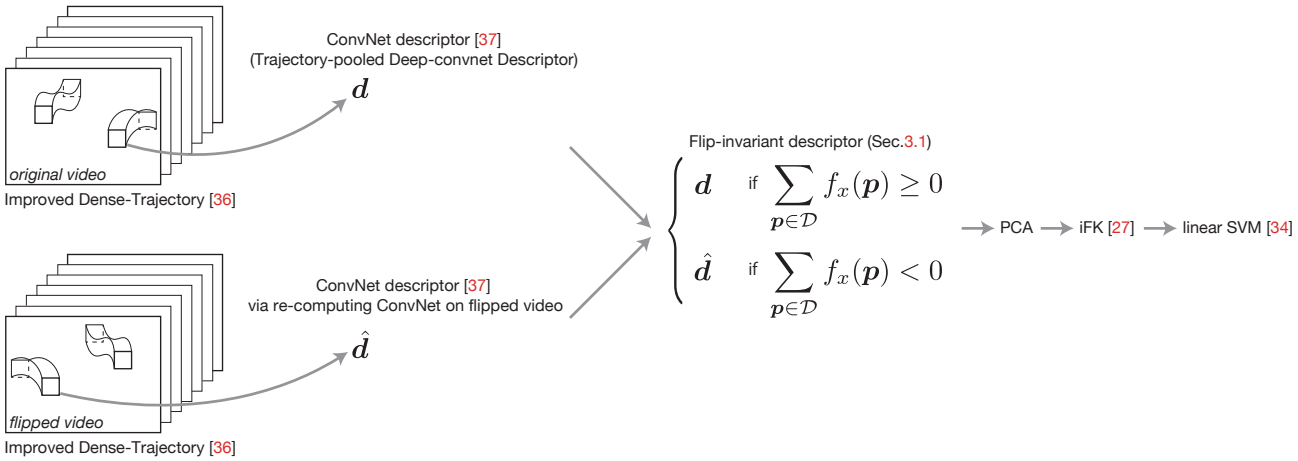

Figure B. Patch-level invariance (Sec.3.1) for ConvNet descriptors [37].

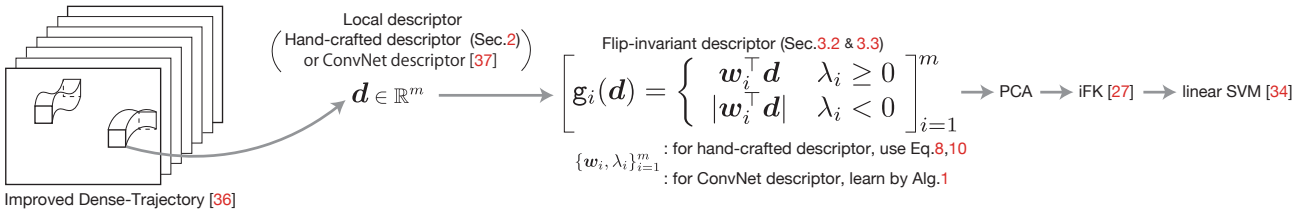

Figure C. Descriptor-level invariance (Sec.3.2&3.3) for hand-crafted/ConvNet descriptors.