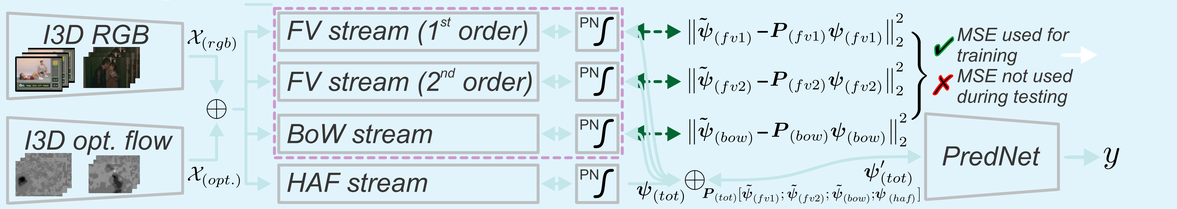
Hallucinating IDT Descriptors and I3D Optical Flow Features for Action Recognition with CNNs

Lei.Wang@data61.csiro.au*,1,2 Piotr.Koniusz@data61.csiro.au*,1,2 Du.Huynh@uwa.edu.au3
*equal contribution [1]Data61/CSIRO [2]Australian National University [3]University of Western Australia

Motivation, contributions and key ideas

- We revive the use of Improved Dense Trajectory descriptors via CNN hallucination of their Bag-of-Words (BoW) and Fisher Vector (FV) for action recognition.
- We show that features of optical flow passed via I3D network can be hallucinated by the I3D RGB stream.
- An external network or descriptors co-regularize a target network by the alignment of feature statistic from different streams, which resembles Domain Adaptation.
- Our pipeline uses self-supervision e.g., IDT BoW/FV and Optical Flow Features (OFF) represent easy to obtain (but computationally costly) self-information about videos.



- Our pipeline consists of: (i) the BoW and FV hallucinating streams, (ii) Optical Flow Features (OFF), (iii) the High Abstraction Features stream (HAF), and (iv) the Prediction Network (PredNet).
- The proposed model saves 20–55h of computations: BoW/FV/OFF are computed only at the training stage and they are hallucinated at the stestinfg stage.
- Descriptor Encoding Schemes which provide self-supervision:
  - **BoW** assigns each local descriptor $\boldsymbol{x}$ to the closest visual word from $\boldsymbol{M}=[\boldsymbol{m}_1,...,\boldsymbol{m}_K]$ built via k-means. In order to obtain mid-level features $\boldsymbol{\phi}$, we solve:

$$\boldsymbol{\phi}=\arg\min_{\boldsymbol{\phi}'}\left\|\boldsymbol{x}-\boldsymbol{M}\boldsymbol{\phi}'\right\|_2^2, s.t.\ \ \boldsymbol{\phi}'\in\{0,1\}, \mathbb{1}^T\boldsymbol{\phi}'=1. \quad (1)$$

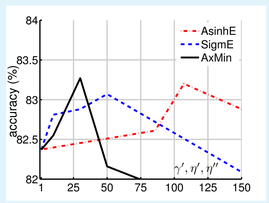
  - **FV** uses a Mixture of K Gaussians from a GMM used as a dictionary. It performs descriptor coding w.r.t. to Gaussian components $G(w_k,\boldsymbol{m}_k,\boldsymbol{\sigma}_k)$ which are parametrized by mixing probability, mean, and on-diagonal standard deviation. The 1^{st}- and 2^{nd}-order features are $\boldsymbol{\phi}_k=(\boldsymbol{x}-\boldsymbol{m}_k)/\boldsymbol{\sigma}_k$, and $\boldsymbol{\phi}'_k=\boldsymbol{\phi}_k^2-1$. Concatenation of per-cluster features $\boldsymbol{\phi}_k^*\in\mathbb{R}^{2D}$ forms the mid-level features $\boldsymbol{\phi}\in\mathbb{R}^{2KD}$:

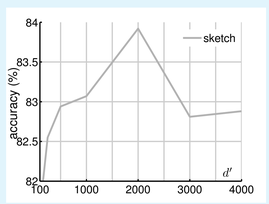
$$\boldsymbol{\phi}=[\boldsymbol{\phi}_1^*;...;\boldsymbol{\phi}_K^*],\ \ \boldsymbol{\phi}_k^*=\frac{p\left(\boldsymbol{m}_k|\boldsymbol{x},\theta\right)}{\sqrt{w_k}}\left[\boldsymbol{\phi}_k;\boldsymbol{\phi}'_k/\sqrt{2}\right], \quad (2)$$

- We have also investigated **Power Normalization** (PN) to prevent so-called burstiness in BoW, FV and CNNs, and the **count sketch** of features to avoid overfitting when fusing several streams:
  - Three PN operators were investigated: *AsinhE*, *Sigmoid* and *AxMin*. Despite the similar role of these three pooling operators, we investigate each of them as their interplay with end-to-end learning differs.
  - Sketching vectors by the count sketch is used for the dimensionality reduction. Let d and d' denote the dimensionality of the input and sketched output vectors, respectively. Let vector $\boldsymbol{h}\in\mathcal{I}_{d'}^d$ contain d uniformly drawn integer numbers from $\{1,...,d'\}$ and vector $\boldsymbol{s}\in\{-1,1\}^d$ contain d uniformly drawn values from $\{-1,1\}$. Then, the sketch projection matrix $\boldsymbol{P}\in\{-1,0,1\}^{d'\times d}$ becomes:

$$P_{ij}=\begin{cases}s_i & \text{if } h_i=j,\\ 0 & \text{otherwise,}\end{cases} \quad (3)$$

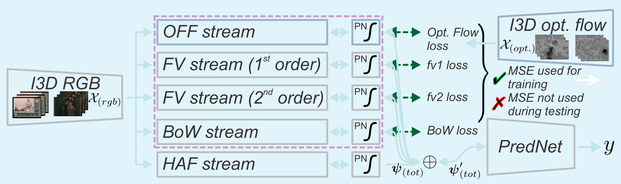
  and the sketch projection $p:\mathbb{R}^d\rightarrow\mathbb{R}^{d'}$ is a linear operation given as $p(\boldsymbol{\psi})=\boldsymbol{P}\boldsymbol{\psi}$ (or $p(\boldsymbol{\psi};\boldsymbol{P})=\boldsymbol{P}\boldsymbol{\psi}$ to highlight $\boldsymbol{P}$).

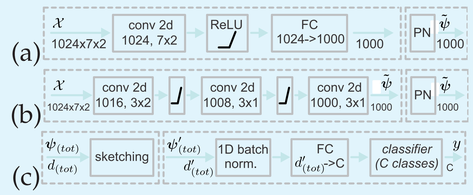




The pipeline: further details

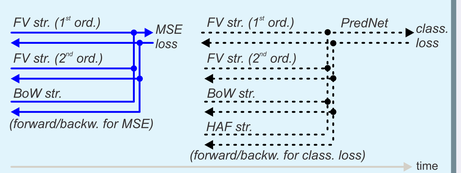
- **BoW/FV/HAF stream** take the I3D intermediate representation $\mathcal{X}_{(rgb)}$ and $\mathcal{X}_{(opt.)}$ of size 1024×7 which were obtained by stripping the classifier and the last 1D conv. layer of I3D pre-trained on Kinetics-400.
- **OFF stream** uses the I3D intermediate representation $\mathcal{X}_{(rgb)}$ only, and it is fed to hallucination/HAF streams. I3D Optical Flow Features $\mathcal{X}_{(opt.)}$ are pre-computed as the training ground-truth for the OFF layer (the MSE loss is used). Fig. (a) and (b) show *Fully Connected* and *Convolutional* variants used for the FV/BoW/OFF and HAF streams.





- **Combining BoW/FV/OFF/HAF**:
  - Fig. (a) and (b): arch. of halluc. streams. Fig. (c) is PredNet.
  - During training, we combine MSE loss functions responsible for training hallucination streams with the class. loss:

$$\ell^*(\boldsymbol{\mathcal{X}},\mathbf{y};\bar{\boldsymbol{\Theta}})=\frac{\alpha}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\tilde{\boldsymbol{\psi}}_i-\boldsymbol{\psi}'_i\right\|_2^2+\ell\Big(f(\boldsymbol{\psi}'_{(tot)};\boldsymbol{\Theta}_{(pr)}),\mathbf{y};\boldsymbol{\Theta}_{(\ell)}\Big),$$



  where: $\forall i\in\mathcal{H},\tilde{\boldsymbol{\psi}}_i=g(\hbar(\boldsymbol{\mathcal{X}},\boldsymbol{\Theta}_i),\eta)\,,\boldsymbol{\psi}'_i=\boldsymbol{P}_i\boldsymbol{\psi}_i,$

$$\boldsymbol{\psi}_{(haf)}=g\big(\hbar(\boldsymbol{\mathcal{X}},\boldsymbol{\Theta}_{(haf)}),\eta\big)\,,\boldsymbol{\psi}'_{(tot)}=\boldsymbol{P}_{(tot)}\left[\oplus_{i\in\mathcal{H}}\tilde{\boldsymbol{\psi}}_i;\boldsymbol{\psi}_{(haf)}\right]. \quad (4)$$

- For **optimization**, in each step, we have (i) forward/backward passes via BoW/FV/OFF streams for MSE loss followed by (ii) forward/backward passes via all sub-streams and PredNet for the classification loss.

Results

- Evaluations (*FC* vs. *Conv* halluc. streams) on the **HMDB51** dataset:

	sp1	*sp2*	*sp3*	mean
HAF(Conv)+BoW/FV(FC) hal.	82.0%	80.4%	80.5%	81.0%
HAF(FC)+BoW/FV(Conv) hal.	82.4%	81.3%	81.5%	81.7%
HAF(FC)+BoW/FV(FC) hal.	82.9%	82.7%	81.5%	**82.4%**

- Evaluations of (*top*) (*HAF+BoW/FV halluc.*) and (*bottom*) comparisons to the state of art on **HMDB51**:

	sp1	*sp2*	*sp3*	mean
HAF only	81.8%	80.8%	80.5%	81.0%
HAF+BoW/FV halluc.	83.5%	82.6%	81.4%	**82.5%**

ADL+ResNet+IDT 74.3%	STM Network+IDT 72.2%
ADL+I3D 81.5%	Full-FT I3D 81.3%

- Evaluations of (*top*) (*HAF+BoW/FV halluc.*) and (*bottom*) comparisons to the state of art on **YUP++**:

	static	*dynamic*	*mixed*	mean stat/dyn	mean all
HAF only	92.0%	81.7%	89.1%	86.8%	87.6%
HAF+BoW/FV halluc.	94.8%	**89.6%**	**93.3%**	**92.2%**	**92.6%**
T-ResNet	92.4%	81.5%	89.0%	87.0%	87.6%
ADL I3D	**95.1%**	88.3%	-	91.7%	-

- Evaluations of (*top*) (*HAF+BoW/FV halluc.*) pipeline w/o sketching/PN (*SK/PN*). (*) denotes human-centric pre-processing while (*MSK/PN*)in (*HAF*+BoW hal.+MSK/PN*) denotes multiple sketches per BoW followed by *PN*. (*bottom*) Other methods on the **MPII**:

	sp1	*sp2*	*sp3*	*sp4*	*sp5*	*sp6*	*sp7*	mAP
HAF+BoW hal.	73.9	71.6	76.2	70.7	76.3	71.9	63.4	71.9%
HAF+BoW hal.(SK/PN)	73.9	75.8	72.2	73.9	77.0	73.6	68.8	**73.6%**
HAF* only	74.6	73.2	77.0	75.1	76.1	75.6	71.9	74.8%
HAF*+BoW hal.	78.8	75.0	84.1	76.0	77.0	78.3	75.2	**77.8%**
HAF*+BoW hal.(MSK/PN)	80.1	79.2	84.8	83.9	80.9	78.5	75.5	**80.4%**
HAF•+BoW hal.(MSK/PN)	80.8	80.9	85.0	83.9	82.0	79.8	79.6	**81.7%**
ditto+OFF hal.	81.2	81.2	84.9	83.4	84.2	78.9	79.1	**81.8%**
I3D+BoW MTL•	79.1	78.1	83.6	78.7	79.1	78.6	76.5	79.1%

KRP-FS 70.0%	KRP-FS+IDT 76.1%	GRP 68.4%	GRP+IDT 75.5%

- Evaluations of our methods on the **Charades** dataset. Our best pipeline yielded 43.1% (much more complex feature banks yield 43.4%):

HAF only	HAF+BoW/ FV exact	HAF+BoW/FV/OFF halluc. +MSK×2/PN	HAF+BoW/FV/OFF halluc. +MSK×4/PN	HAF+BoW/FV/OFF halluc. +MSK×8/PN
37.2	41.9	42.0	42.2	**43.1**