Learning to recognize human actions: from hand-crafted to deep-learning based visual representations

PhD Thesis defense

Author Albert Clapés — **Advisor** Dr. Sergio Escalera

UNIVERSITAT DE BARCELONA

February 4, 2019

Action recognition: theory and practice

To provide the computers with the ability to visually recognize human-related actions in a possibly meaningful context and potentially involving interaction with objects.

In practice, the recognition abilities are provided by computer vision algorithms addressing the following issues:

- Which action? (classification)
- When is the action ocurring? (temporal localization)
- Where is the action taking place? (spatial localization)

The problem has been studied for nearly 30 years[1,2].

---

[1] *Yasuo Kuniyoshi, Hirochika Inoue, and Masayuki Inaba. "Design and implementation of a system that generates assembly programs from visual recognition of human action sequences". In: Intelligent Robots and Systems' 90.'Towards a New Frontier of Applications', Proceedings. IROS'90. IEEE International Workshop on. IEEE. 1990, pp. 567–574.*

[2] *Junji Yamato, Jun Ohya, and Kenichiro Ishii. "Recognizing human action in time-sequential images using hidden markov model". In: Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on. IEEE. 1992, pp. 379–385.*

KTH dataset[3] (boxing examples)


Weizmann dataset[4] (handwaving examples)


[3] *Ivan Laptev and Tony Lindeberg. "Interest point detection and scale selection in space-time". In: International Conference on Scale-Space Theories in Computer Vision. Springer. 2003, pp. 372–387.*

[4] *Moshe Blank et al. "Actions as space-time shapes". In: null. IEEE. 2005, pp. 1395–1402.*

HMDB-51 dataset[5] (smoking examples)


UCF-101 dataset[6] (pizza-tossing examples)


[5] *Hildegard Kuehne et al. "HMDB: a large video database for human motion recognition". In: Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE. 2011, pp. 2556–2563.*

[6] *Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild". In: CoRR abs/1212.0402 (2012).*

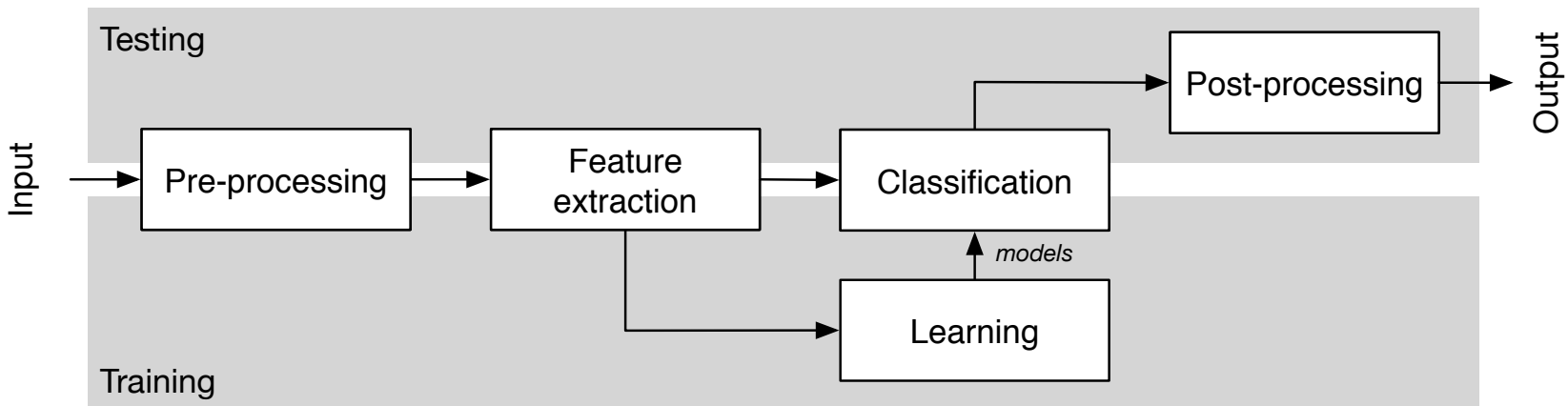
Name	Year	#Classes	#Examples
KTH	2003	6	2,391
Weizmann	2005	10	90
Hollywood2	2009	12	3,669
HighFive	2010	4	300
HMDB-51	2011	51	6,766
UCF Sports Action	2012	10	150
UCF-101	2012	101	13,320
Sports-1M	2014	487	1,000,000
ActivityNet	2015	203	28,000
Youtube-8M	2016	4,716	8,000,000
Kinetics	2017	600	500,000
AVA	2018	80	1,580,000*

* Actions are concurrent in time and space.

Table: Action classification datasets

Action recognition: pipeline and feature extraction approaches

Action recognition can be posed as a fully-supervised pattern recognition problem.



Two different approaches to **feature extraction**:

- **Hand-crafted**: manual design of the feature representation
- **Feature learning**: an "optimal" feature set is learned guided by the recognition performance, e.g. end-to-end trained *deep neural networks*

Action recognition: not all about the nets

Many deep-based approaches are combined with hand-crafted approaches, e.g. *Improved Dense Trajectories* (IDTs)[7].

Besides their combination with IDTs, deep-learning methods for action recognition also benefit from:

- Optical flow or other visual modalities – e.g. depth
- Ensemble learning
- Strategies to explicitly model longer term temporal dynamics

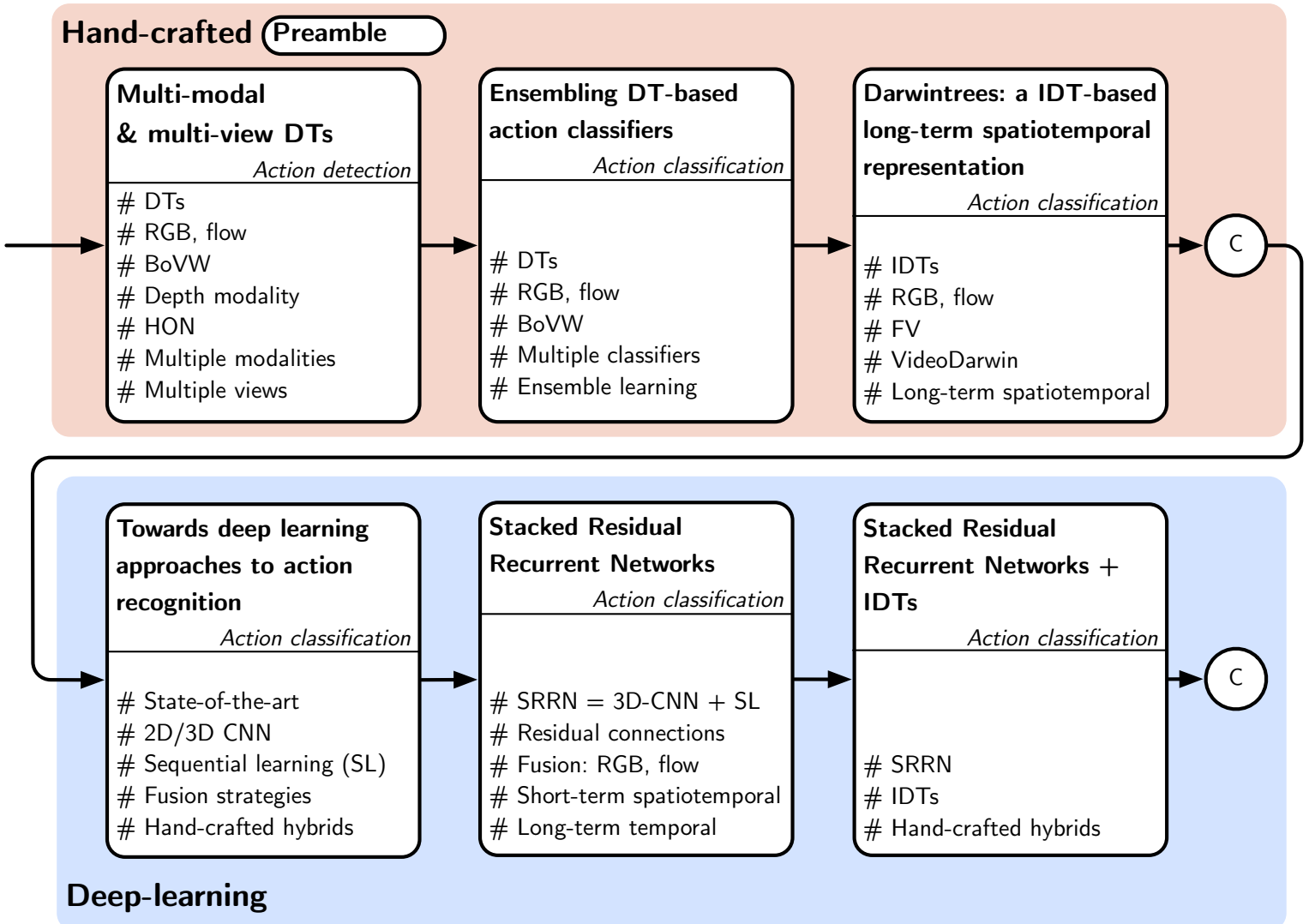
[7] *Heng Wang and Cordelia Schmid. "Action recognition with improved trajectories". In: Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE. 2013, pp. 3551–3558.*

Contributions of this thesis

The Thesis concurs in time with a change of paradigm in feature extraction approaches, as reflected in its two distinguished but complementary parts:

- In the hand-crafted scenario, we propose:
  - Multi-Modal and multi-view DTs
  - Ensembling DT-based action classifiers
  - *Darwintrees*: an IDT-based representation to model long-term temporal dynamics
- Moving towards deep-based approaches, we present:
  - A comprehensive study of current deep-learning methods
  - *Stacked Residual Recurrent Networks* for action recognition

Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
C
Deep-learning

Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
C
Deep-learning

Preamble

The design of hand-crafted features for action recognition requires addressing two key issues:

1. Which visual *context* to be represented by the features[8,9] – either global or local?
2. How to effectively model the *temporal* or *spatiotemporal* information from videos?

[8] *Ronald Poppe. "A survey on vision-based human action recognition". In: Image and vision computing 28.6 (2010), pp. 976–990.*

[9] *Thomas B Moeslund, Adrian Hilton, and Volker Krüger. "A survey of advances in vision-based human motion capture and analysis". In: Computer vision and image understanding 104.2-3 (2006), pp. 90–126.*

Which visual *context* to be represented by the features[10,11] – either global or local?

The latest hand-crafted works experimentally demonstrated local representations (LR) can better deal with real-world action videos:

- Body deformations
- Heavy occlusions
- Changes in the point of view
- Camera motion
- Avoid relying on error-prone preprocessing steps, e.g. background subtraction or human detection.

[10] *Poppe, "A survey on vision-based human action recognition".*

[11] *Moeslund, Hilton, and Krüger, "A survey of advances in vision-based human motion capture and analysis".*

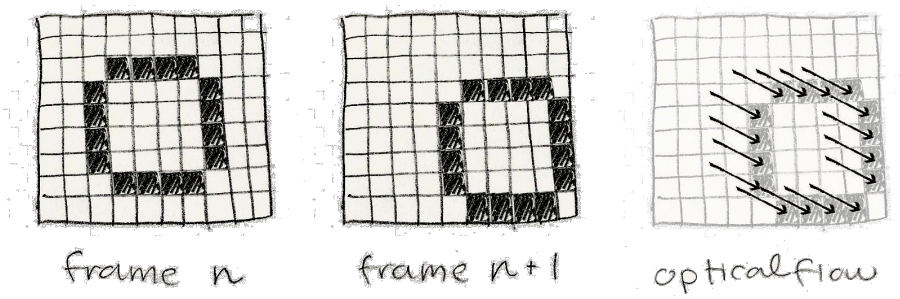
How to effectively model the *temporal* or *spatiotemporal* information from videos?

1. Finding **interest locations** based on spatial and motion characteristics
2. Defining a **local context**, e.g. cuboids or space-time tubes
3. **Extracting the features** by temporally combining 2D image descriptors or, directly, 3D feature descriptors

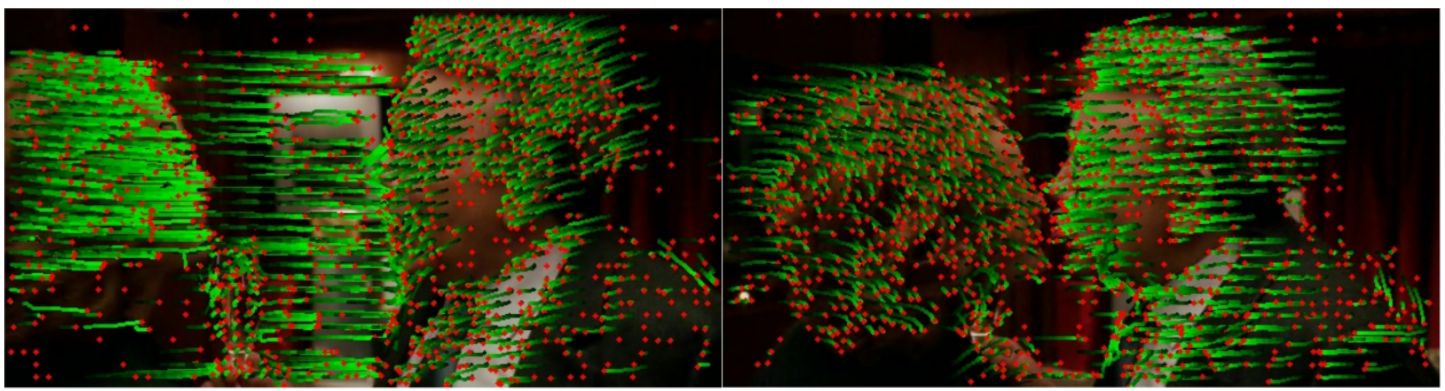
The most successful hand-crafted methods to action recognition: *Dense Trajectories* (DTs)[12] and *Improved Dense Trajectories* (IDTs)[13].

[12] *Heng Wang et al. "Action recognition by dense trajectories". In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE. 2011, pp. 3169–3176.*

[13] *Wang and Schmid, "Action recognition with improved trajectories".*



(a) Optical flow (OF)[14]

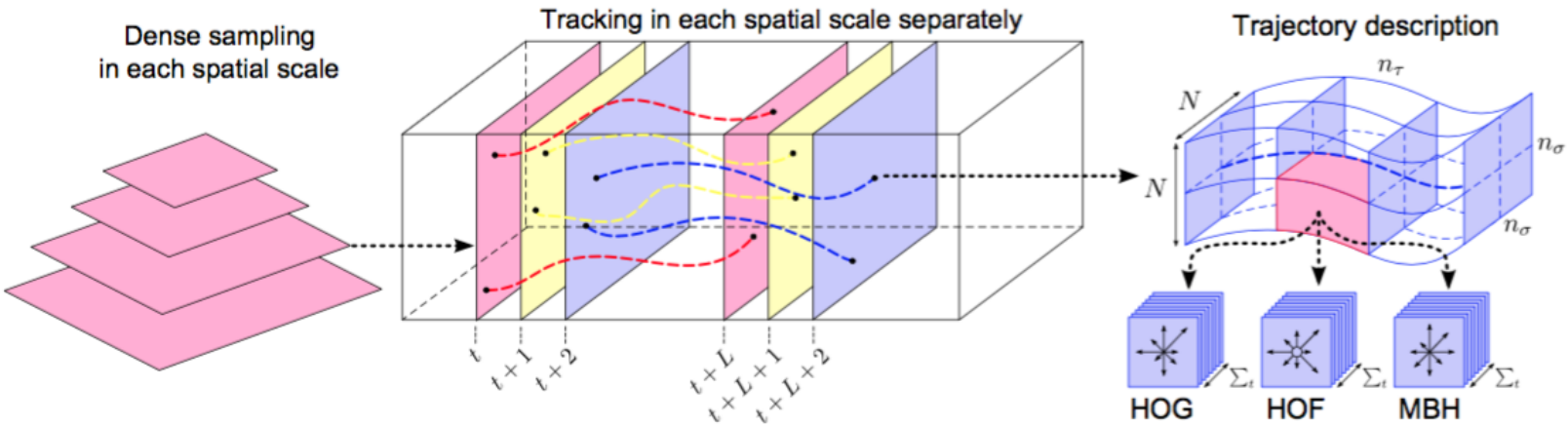


(b) Dense trajectories (DTs)

Figure: A dense trajectory (DT) is constructed from tracking a pixel's displacement through space and time

[14] *Gunnar Farnebäck. "Two-frame motion estimation based on polynomial expansion". In: Image analysis. Springer, 2003, pp. 363–370.*

Points are sampled at different spatial scales and over regular space intervals and tracked during L frames using OF vectors[15].



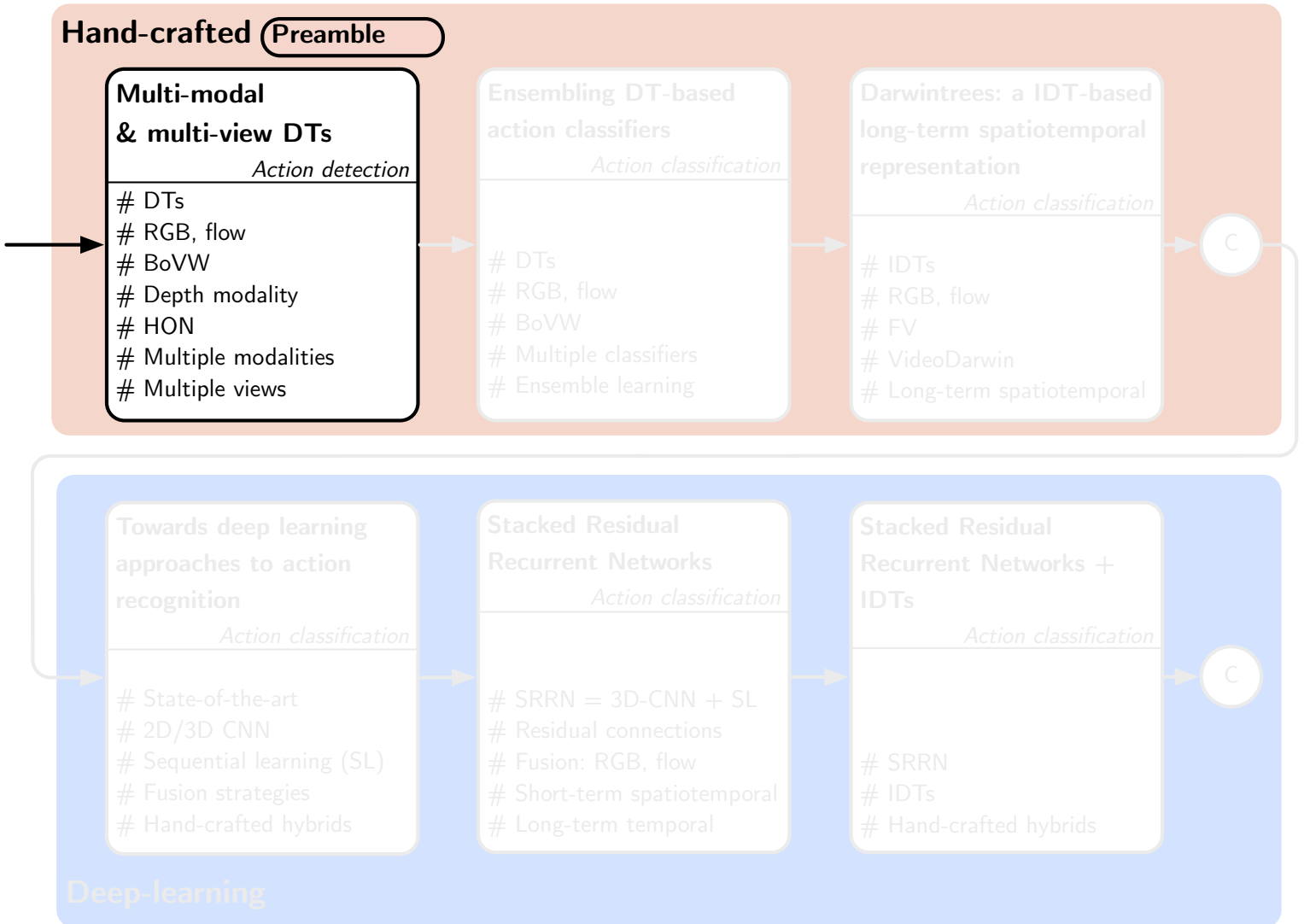
Each trajectory is described with:

- The trajectory shape: $\mathcal{T} = (\Delta P_t, \ldots, \Delta P_{t+L-1})$, where $\Delta P_t = (P_{t+1} - P_t)$
- The cells' descriptors: HOG, HOF, and MBH. The $n_x \times n_y \times n_t$ cells are obtained from uniformly subdividing the $N \times N \times L$ spatiotemporal tubelet surrouding the trajectory.

A cell descriptor (HOG, HOF, or MBH) is the accumulation of descriptors from the image patches contained in that cell.

[15] *Farnebäck, "Two-frame motion estimation based on polynomial expansion".*

Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
C
Deep-learning

Multi-modal and multi-view DTs

Multi-modal Dense Trajectories (MmDT) We enrich the set of RGB descriptors (HOG, HOF, and MBH) and TS, with a Histogram of Oriented Normals (HON).

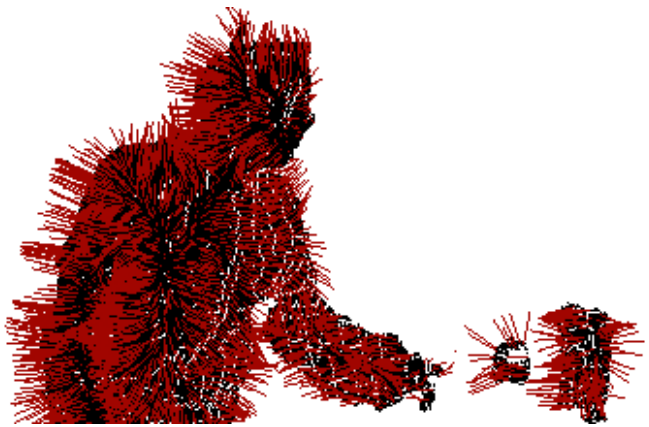


Figure: A 3D normal can be expressed in terms of its inclination ψ and azimuth φ (polar coordinates)

Our HON representation consists of a two-dimensional $\delta_\psi \times \delta_\varphi$ histogram, with each bin counting occurrences of pairs (ψ, φ).

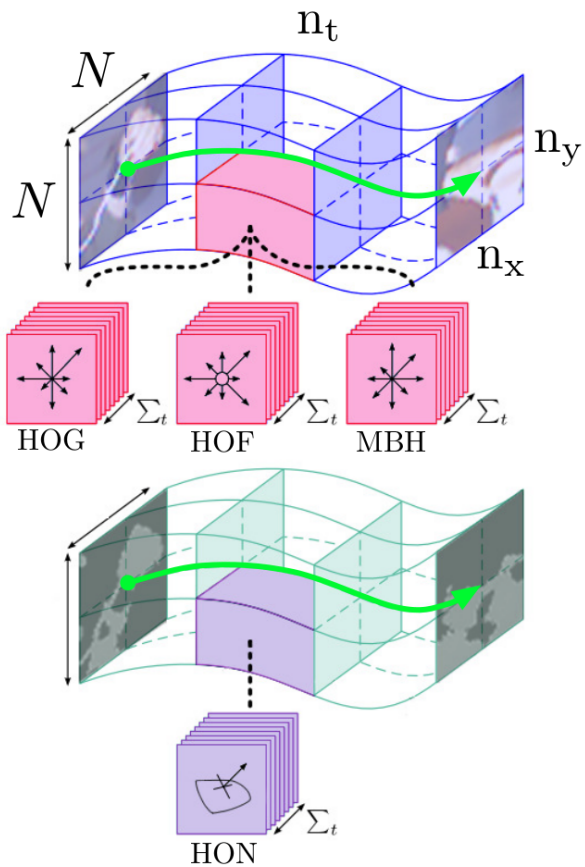


Figure: Mulit-modal dense trajectories

Multi-modal and multi-view DTs

MmDTs extracted from **multiple views** are used to learn a codebook (vocabulary) of trajectories. Then, MmDTs from all views of a video can be encoded as a matrix of per-frame BOVW features (words). A temporal sliding window is convolved to efficiently compute window-level descriptors and serve as input to learning/classification stages.

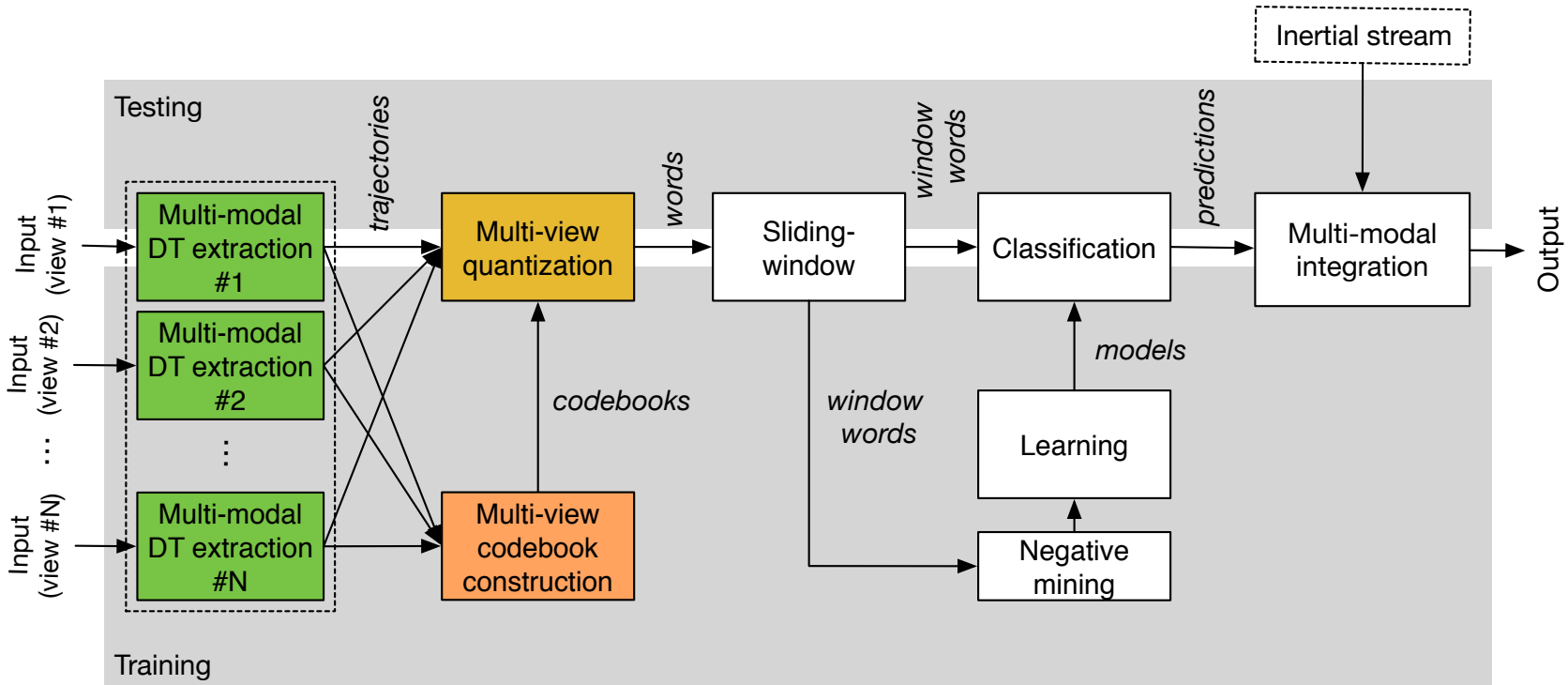


Figure: Sliding-window based action localization by using multi-modal and multi-view DTs.

Results on SARQuavitae dataset consisting of 31 sequences and 14 elderly people in a real-life elder home monitoring scenario.

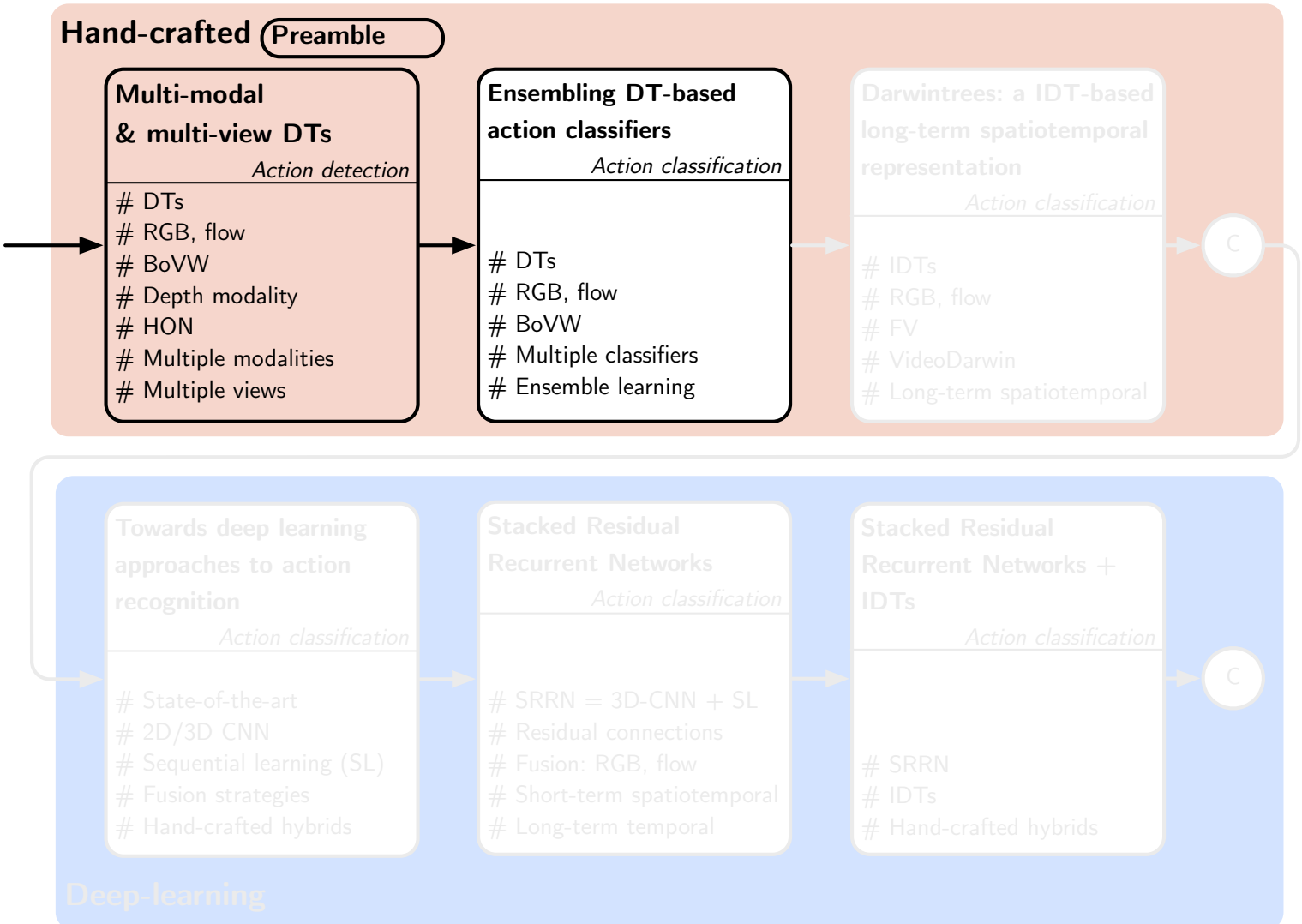
	Trajectory descriptors					
Action name	TS	HOG	HOF	MBH	HON	Accuracy
Drinking	0.5	0	0	0.3	0.2	91.54%
Eating	0.1	0.1	0	0.4	0.4	86.42%
Reading	0.2	0.1	0.1	0.1	0.5	91.57%
Taking-pill	0	0	0.4	0.1	0.5	83.80%

Table: Best performing combinations of descriptor weights for each of the classes in a classification experiment on pre-cut action clips. Accuracy reported

	Streams		Integration strategies		
	Vision	Inertial	Intersection	Union	Learning-based (RF)
Drinking	**0.46**	0.32	0.30	0.37	0.42
Eating	0.48	0.26	0.06	0.22	**0.49**
Reading	0.05	0.10	0.04	0.05	**0.20**
Taking-pill	0.22	0.04	0.00	0.08	**0.34**
TOTAL (mean)	0.30	0.18	0.10	0.18	**0.36**

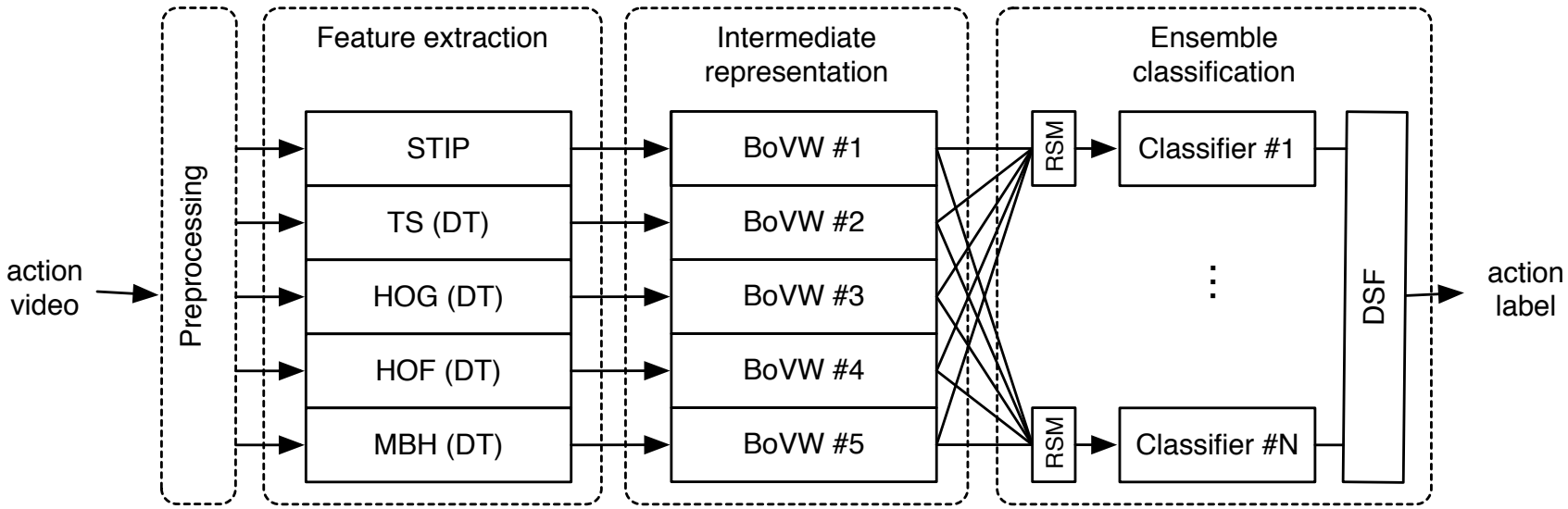
Table: Temporal detection results. The learning-based integration is done via a Random Forest (RF) classifier. F1-score reported

Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
C
Deep-learning

Ensembling DT-based action classifiers

A codebook is generated for each LR – i.e. STIP, TS, HOG, HOF, and MBH. For an input video, each BoVW generates a $4,000$-dimensional word.



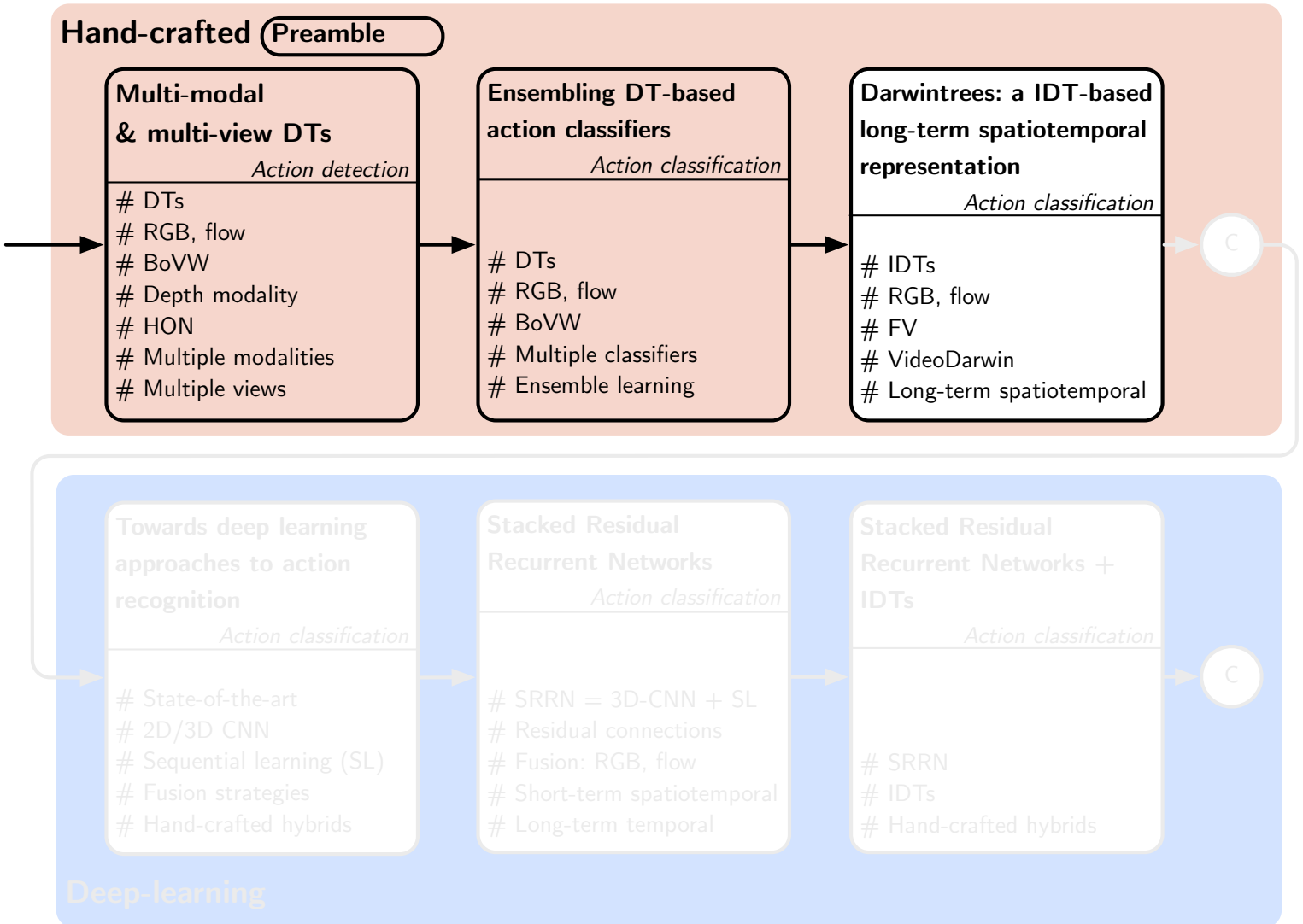
The classifiers' outputs are combined via Dempster-Shafer Fusion (DSF). It computes a *degree of belief* for each classifier on each class (from training data), and uses these to weight the classifiers' decisions during prediction.

A Random-subspace Method (RSM) samples M features among $5 * 4,000 = 20,000$ to be fed to each of the N classifiers.

Method	Acc	#1	#2	#3
Karpathy et al. (2014)	65.40	-	-	-
Phan et al. (2013)	73.39	71.10	73.67	**75.39**
Murthy et al. (2013)	73.10	-	-	-
Rostamzadeh et al (2013)	70.50	70.45	69.80	71.27
Nga et al (2014)	66.26	65.16	66.73	66.90
Cho et al (2013)	65.95	65.22	65.39	67.24
Paez et al (2013)	65.68	65.31	65.48	66.23
Chen et al (2013)	64.30	63.41	65.37	64.12
Nga et al (2014b)	60.10	-	-	-
Wang et al (2013)	54.74	54.76	55.16	54.29
Soomro et al (2012)	43.90	-	-	-
STIP + BoVW	42.56	42.12	41.89	43.67
DT (TS) + BoVW	49.88	49.76	50.05	49.83
DT (HOG) + BoVW	51.10	50.19	51.76	51.35
DT (HOF) + BoVW	46.59	46.47	46.69	46.60
DT (MBH) + BoVW	62.93	62.54	62.78	63.46
Baseline: Holistic	60.73	61.13	60.11	60.95
Ensemble I: DSF	69.10	69.43	68.09	69.79
Ensemble II: RSM+DSF	**75.05**	**75.11**	**74.80**	75.23

Table: Comparison of our proposal to accuracies (%) from state-of-the-art recognition methods on UCF-101 (global overall accuracy and on the 3 different train/test splits). M was fixed to $20,000/N$ and, experimentally, $N = 85$

Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
C
Deep-learning

Darwintrees: an IDT-based representation to model long-term temporal dynamics

Key aspects

- We model the evolution of IDTs[16] throughout the entire video sequence, but also on its subparts
- Subparts: a spectral divisive clustering[17] yields to an unordered binary tree decomposing the cloud of trajectories
- Videodarwin[18] provides a meta-representation for tree nodes and branches
- Our Darwintree kernel classifies videos based on node-videodarwin and branch-videodarwin representations
- State-of-the-art performance on UCF-Sports and competitive results on HighFive and Olympic Sports

---

[16] *Wang and Schmid, "Action recognition with improved trajectories".*

[17] *Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. "Activity representation with motion hierarchies". In: International Journal of Computer Vision 107.3 (2014), pp. 219–238.*

[18] *Basura Fernando et al. "Modeling video evolution for action recognition". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, pp. 5378–5387.*

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T), \boldsymbol{x}_i \in \mathbb{R}^D$ be a set of local of d-dimensional feature vectors (e.g. IDT descriptors). Let $\boldsymbol{\Theta}_{\mathrm{GMM}} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}_k | k = 1, \ldots, K\}$ the parameters of a GMM that fit the distribution of descriptors.

Each $\boldsymbol{x}_i$ is associated to the k-th mode in the mixture by a strength:

$$\gamma_{ik} = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right]}{\sum_{c=1}^{K} \exp\left[-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_c)\right]}$$

Then, for each mode k, we consider mean and covariance deviation:

$$\boldsymbol{m}_k = \frac{1}{T\sqrt{\pi_k}} \sum_{i=1}^{T} \gamma_{ik} \frac{\boldsymbol{x}_i - \boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k}$$

$$\boldsymbol{z}_k = \frac{1}{T\sqrt{2\pi_k}} \sum_{i=1}^{T} \gamma_{ik} \left[\left(\frac{\boldsymbol{x}_i - \boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k}\right)^2 - 1\right]$$

The **fisher vector** representation of $\boldsymbol{X}$:

$$FV(\boldsymbol{X}) = [\boldsymbol{m}_1; \ldots; \boldsymbol{m}_K; \boldsymbol{z}_1; \ldots; \boldsymbol{z}_K] \in \mathbb{R}^{2DK}$$

Videodarwin: in-a-nutshell

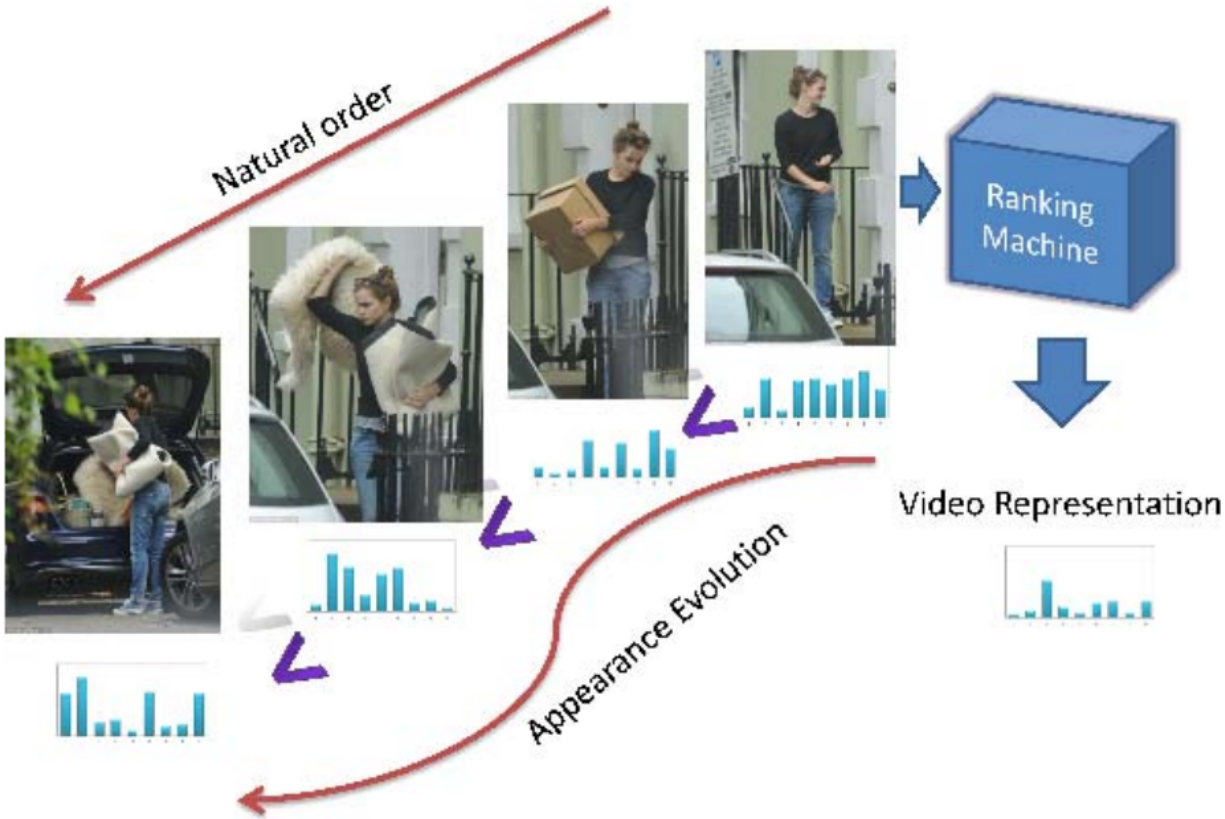


Figure: Modeling the evolution of features by learning to order frames in a video. Figure obtained from Fernando et al., "Modeling video evolution for action recognition"

Let $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_F] \in \mathbb{R}^{2DK \times F}$ be column-wise per-frame feature vectors – e.g. IDT-based FVs.
Let $\tilde{\mathbf{V}}^+ = [\tilde{\mathbf{v}}_1^+, \ldots, \tilde{\mathbf{v}}_F^+]$ and $\tilde{\mathbf{V}}^- = [\tilde{\mathbf{v}}_1^-, \ldots, \tilde{\mathbf{v}}_F^-]$, i.e. $\tilde{\mathbf{V}}^+, \tilde{\mathbf{V}}^- \in \mathbb{R}^{2DK \times F}$, be the smoothed versions of $\mathbf{V}$:

$$\tilde{\mathbf{v}}_i^+ = \varrho_1 \left(\frac{1}{i} \sum_{j=1}^{i} \mathbf{v}_j \right) \ , \ \tilde{\mathbf{v}}_i^- = \varrho_1 \left(\frac{1}{i} \sum_{j=0}^{i-1} \mathbf{v}_{i-j} \right) \tag{1}$$

where $\varrho_1(\cdot) = \frac{\cdot}{||\cdot||_1}$ is the L1-normalization function.
The videodarwin representation of $\mathbf{V}$:

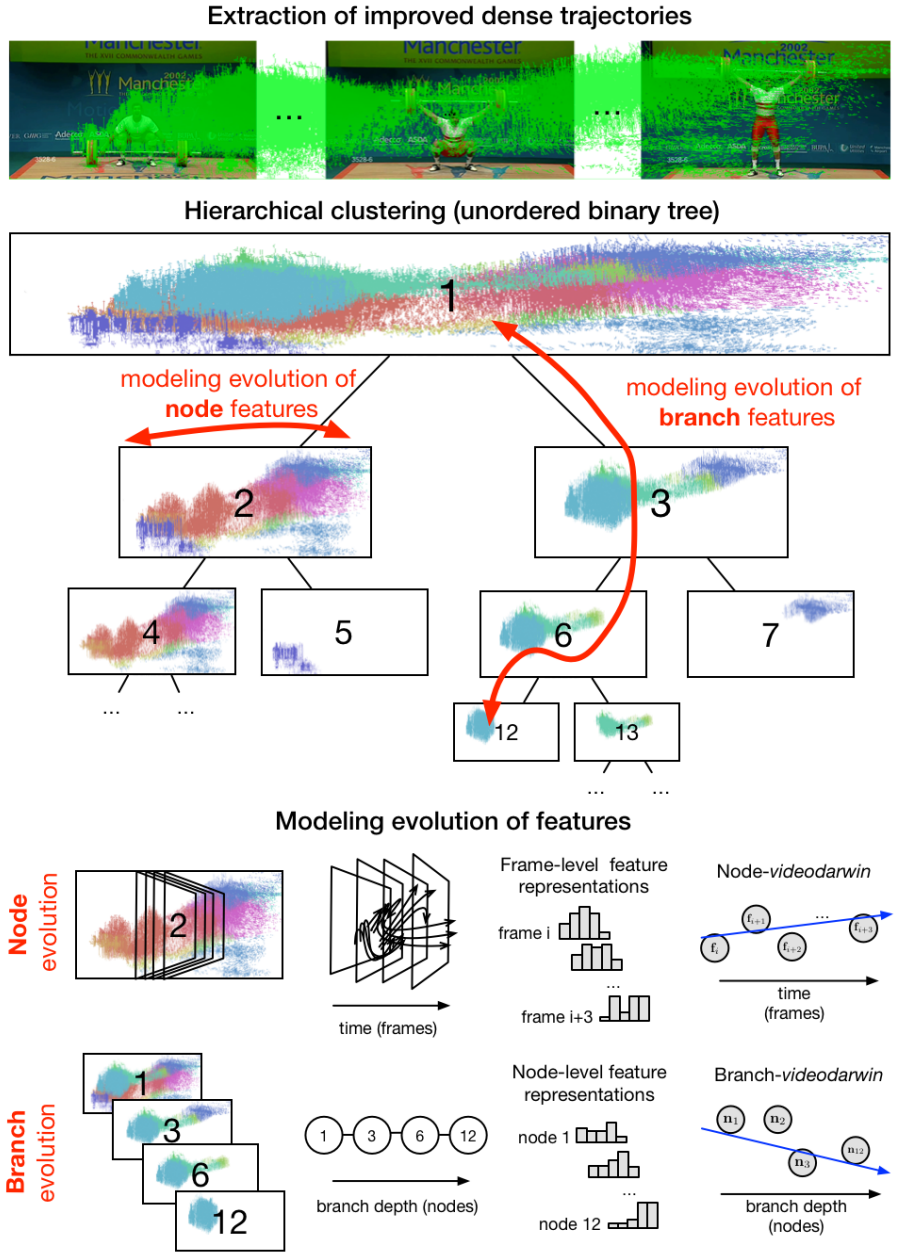
$$\mathbf{w} = \left[\boldsymbol{\theta}^+ ; \boldsymbol{\theta}^- \right] \in \mathbb{R}^{4DK}$$

where $\boldsymbol{\theta}^+, \boldsymbol{\theta}^- \in \mathbb{R}^{2DK}$ are parameters of two regressors trained on $\tilde{\mathbf{V}}^+$ and $\tilde{\mathbf{V}}^-$:

$$\{(\tilde{\mathbf{v}}_i^+, i) | 1 \leq i \leq F\} \rightarrow \boldsymbol{\theta}^+ \ , \ \ \{(\tilde{\mathbf{v}}_i^-, i) | 1 \leq i \leq F\} \rightarrow \boldsymbol{\theta}^-$$

The meta-descriptor $\mathbf{w}$ can be input to a discriminative classifier. However, the smoothing from Eq. 1 can be problematic for longer sequences.

Darwintrees: an IDT-based representation to model long-term temporal dynamics



The **similarity between trajectories** in the video is defined for each feature $f \in \{x, y, z, v_x, v_y\}$ using a RBF-gaussian kernel:

$$\boldsymbol{A}_f(i,j) = (\mathbf{t}_{i,f}, \mathbf{t}_{j,f}) = exp\left(-\frac{||\mathbf{t}_{i,f} - \mathbf{t}_{j,f}||_2}{2\hat{d}}\right),$$

where d_f is the median of the distances between the corresponding tracklet features.
The similarity matrices are aggregated via element-wise product:

$$\mathbf{A} = \mathbf{A}_x \odot \mathbf{A}_y \odot \mathbf{A}_t \odot \mathbf{A}_{v_x} \odot \mathbf{A}_{v_y}.$$

Given **A**, we can perform **spectral clustering**:

- Nyström approximation method[19] instead, allows to use a small portion of the trajectories to extrapolate the results and obtain the approximate leading eigenvectors.
- A divisive hierarchical clustering algorithm[20] recursively thresholds on the leading eigenvectors' values and build the corresponding unordered binary tree.

[19] *Charless Fowlkes et al. "Spectral grouping using the Nystrom method". In: Pattern Analysis and Machine Intelligence, IEEE Transactions on 26.2 (2004), pp. 214–225.*

[20] *Gaidon, Harchaoui, and Schmid, "Activity representation with motion hierarchies".*

Node videodarwin

Let $\tilde{\boldsymbol{V}}_q^+ = [\boldsymbol{v}_{q,1}^+, \ldots, \boldsymbol{v}_{q,F_q}^+]$ and $\tilde{\boldsymbol{V}}_q^- = [\boldsymbol{v}_{q,1}^-, \ldots, \boldsymbol{v}_{q,F_q}^-]$ be the smoothed per-frame IDT-FVs of the q-th node, where $F_q = b - a + 1$ s.t. $1 \leq a < b \leq F$.

The **node videodarwin** representation is:

$$\mathbf{n}_q = [\boldsymbol{\theta}_q^{\text{node}+}; \boldsymbol{\theta}_q^{\text{node}-}] \in \mathbb{R}^{4DK},$$

where $\boldsymbol{\theta}_q^{\text{node},+}, \boldsymbol{\theta}_q^{\text{node}-} \in \mathbb{R}^{2DK}$ are the two (forward and reverse) regressors' parameters trained, respectively, on $\{(\tilde{\mathbf{v}}_{q,i}^+, i) \mid 1 \leq i \leq F_q\}$ and $\{(\tilde{\mathbf{v}}_{q,i}^-, i) \mid 1 \leq i \leq F_q\}$.

This provides an additional solution to the whole-video smoothing degradation. Nodes with smaller groups of trajectories are likely to span shorter time intervals.

Branch videodarwin

Let $\mathbf{u}_j \in \mathbb{R}^{2DK}$ be the j-th node global IDT-FV representation and the stack of per-node representations from the node itself (j) to the tree root node (1):

$$\mathbf{U}_j = [\mathbf{u}_{\lfloor j/2^0 \rfloor}, \mathbf{u}_{\lfloor j/2^1 \rfloor}, \ldots, \mathbf{u}_{\lfloor j/2^{\log_2(j)} \rfloor}]$$

where $\lfloor \cdot \rfloor$ refers to the floor operation.

The **branch videodarwin** representation of the j-th is computed:

$$\mathbf{b}_j = [\boldsymbol{\theta}_j^{\text{branch}^+}; \boldsymbol{\theta}_j^{\text{branch}^-}] \in \mathbb{R}^{4DK},$$

where $\boldsymbol{\theta}_q^{\text{branch}+}, \boldsymbol{\theta}_q^{\text{branch}-} \in \mathbb{R}^{2DK}$ are the two (forward and reverse) regressors' parameters trained, respectively, on $\{(\tilde{\mathbf{u}}^+_{\lfloor j/2^i \rfloor}, i) \mid 1 \leq i \leq log_2(j)\}$ and $\{(\tilde{\mathbf{u}}^-_{\lfloor j/2^i \rfloor}, i) \mid 1 \leq i \leq log_2(j)\}$.

Note branch videodarwin for the root note $\mathbf{b}_1$ is not defined.

Darwintree kernel

Let define a tree structure $\mathcal{E} = \{\mathbf{n}_1, \mathcal{S}\}$. More precisely,

$$\mathcal{S} = \left\{ \mathbf{s_i} = [\mathbf{n_i}; \mathbf{b_i}] \ : \ 1 < i < |\mathcal{B}|, \ \mathbf{s}_i \in \mathbb{R}^{8DK} \right\}.$$

We compute the darwintree kernel based on the pairwise similarity of those node-branch representations:

$$K_{\mathrm{DT}}(\mathcal{S}, \mathcal{S}') = \frac{1}{|\mathcal{S}||\mathcal{S}'|} \sum_{\mathbf{s}_i \in \mathcal{S}} \sum_{\mathbf{s}_j \in \mathcal{S}'} \phi(\mathbf{s}_i, \mathbf{s}_j) \, , \ \forall i, j > 1 \tag{2}$$

where $\phi(\cdot, \cdot)$ can be any valid mapping function, e.g. dot product for linear mapping. The normalization factor $\frac{1}{|\mathcal{S}||\mathcal{S}'|}$ is a normalization factor.

Results

Datasets

- UCF Sports Action[21]: 150 examples, 10 classes. Accuracy on 147/53 train/test holdout.
- HighFive[22]: 300 examples, 4+1 classes. mAP on 2-fold CV.
- Olympic Sports[23]: 783 examples, 16 classes. mAP on 640/143 train/test holdout.

Parameters and settings

- Trajectory features: MBH only ("RootSIFT" and PCA=0.5 reduced dimensionality from 192 to 96 dimensions)
- GMMs: $K = 256$ and $N = 1,000,000$
- FVs: $2 * 96 * 256 = 49,152$ (and $2 * 49,152 = 98,304$ after pos-neg kernel mapping and $l2$-normalization prior to VD)
- Spectral clustering: same parameters from[24], but no tree levels = 4
- VideoDarwin: "RootSIFT" and $l2$-normalization
- Classification: linear SVM with $C = 1$

[21] *Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. "Action mach a spatio-temporal maximum average correlation height filter for action recognition". In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE. 2008, pp. 1–8.*

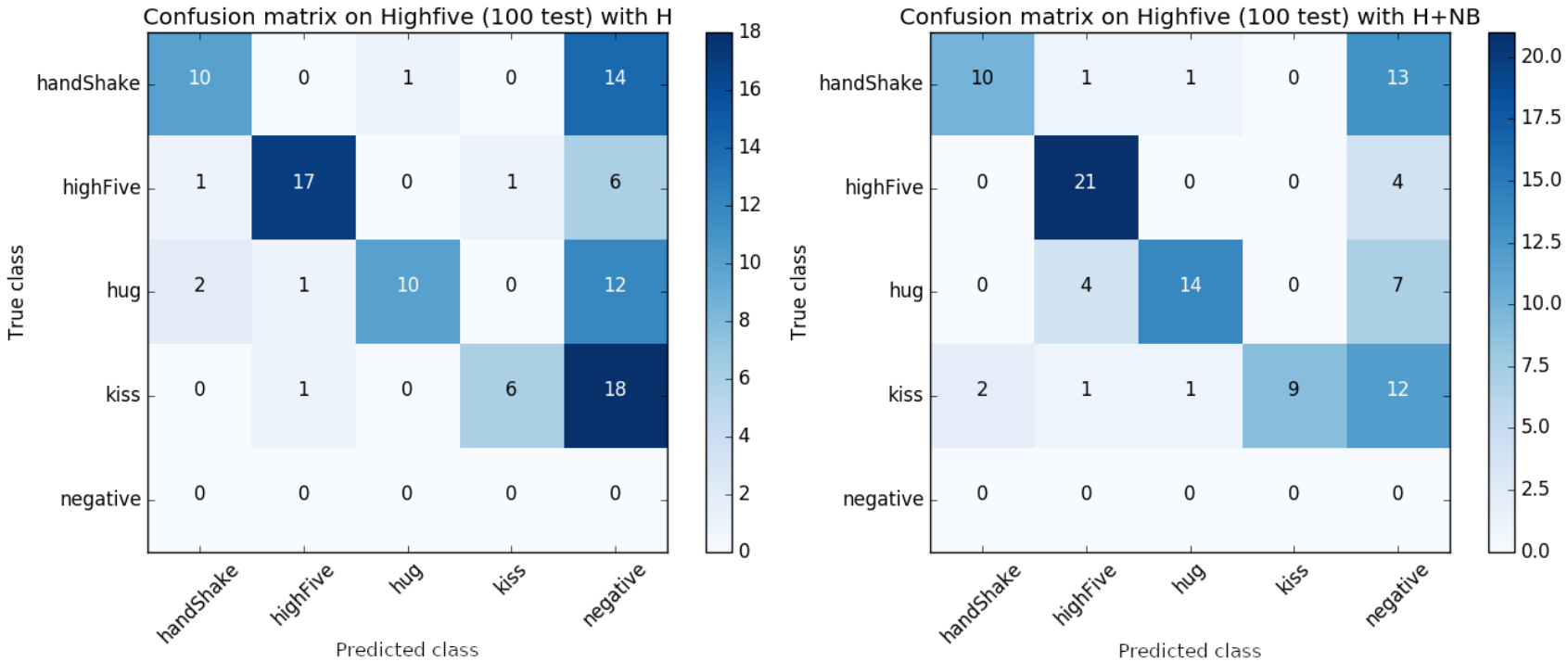
[22] *Alonso Patron-Perez et al. "Structured learning of human interactions in TV shows". In: IEEE Transactions on Pattern Analysis and Machine Intelligence 34.12 (2012), pp. 2441–2453.*

[23] *Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. "Modeling temporal structure of decomposable motion segments for activity classification". In: European conference on computer vision. Springer. 2010, pp. 392–405.*

[24] *Gaidon, Harchaoui, and Schmid, "Activity representation with motion hierarchies".*

Method	UCF Sports Action (Acc)	HighFive (mAP)	Olympic Sports (mAP)
H = VD	87.23	**78.30**	88.34
N	85.11	73.48	83.17
B	80.85	74.39	87.70
NB = DT	**91.49**	73.21	84.38
H+NB	**91.49**	75.78	**88.84**

Table: Results of the different methods in the benchmarking datasets for holistic-videodarwin (H), node-videodarwin (N), branch-videodarwin (B), darwintree (NB or DT), and the combination of NB with H



(a) H on HighFive (b) H+NB on HighFive

Figure: Confusion matrices on HighFive illustrating the improvement of H+NB over H in terms of TP, FP, and FN

(a) Golf-Swing-Back 005. GT=1, H=2, N=2, B=9, NB=1, H+NB=1

(b) Run-Side 001. GT=5, H=9, N=9, B=5, NB=5, H+NB=9

Figure: Visual data and trajectory clusters on 5 frames evenly spaced in time on UCF Sports Action's examples. See in the captions the groundtruth label (GT) and outputs of H, N, B, NB, and H+NB. Classes are (1) "Diving-Side", (2) "Golf Swing", (3) "Kicking", (4) "Lifting", (5) "Riding Horse", (6) "Running", (7) "Skateboarding", (8) "Swing-Bench", (9) "Swing-Side", and (10) "Walking"

(a) hug 0017. GT=3, H=1, N=3, B=1, NB=3, H+NB=3

(b) kiss 0040. GT=4, H=4, N=5, B=5, NB=5, H+NB=5

Figure: Visual data and trajectory clusters on 5 frames evenly spaced in time for 5 different examples on the HighFive dataset. See in the captions the groundtruth label (GT) and outputs of H, N, B, NB, and H+NB. Classes are: (1) "handShake", (2) "highFive", (3) "hug", (4) "kiss", and (5) negative

Method	Accuracy (%)
Karaman et al. (2014)	90.8
Ma et al. (2015)	89.4
Wang et al. (2013)	85.2
Ma et al. (2013)	81.7
Raptis et al. (2012)	79.3
VD	87.2
Ours (VD+DT)	**91.5**

Table: UCF Sports Action dataset

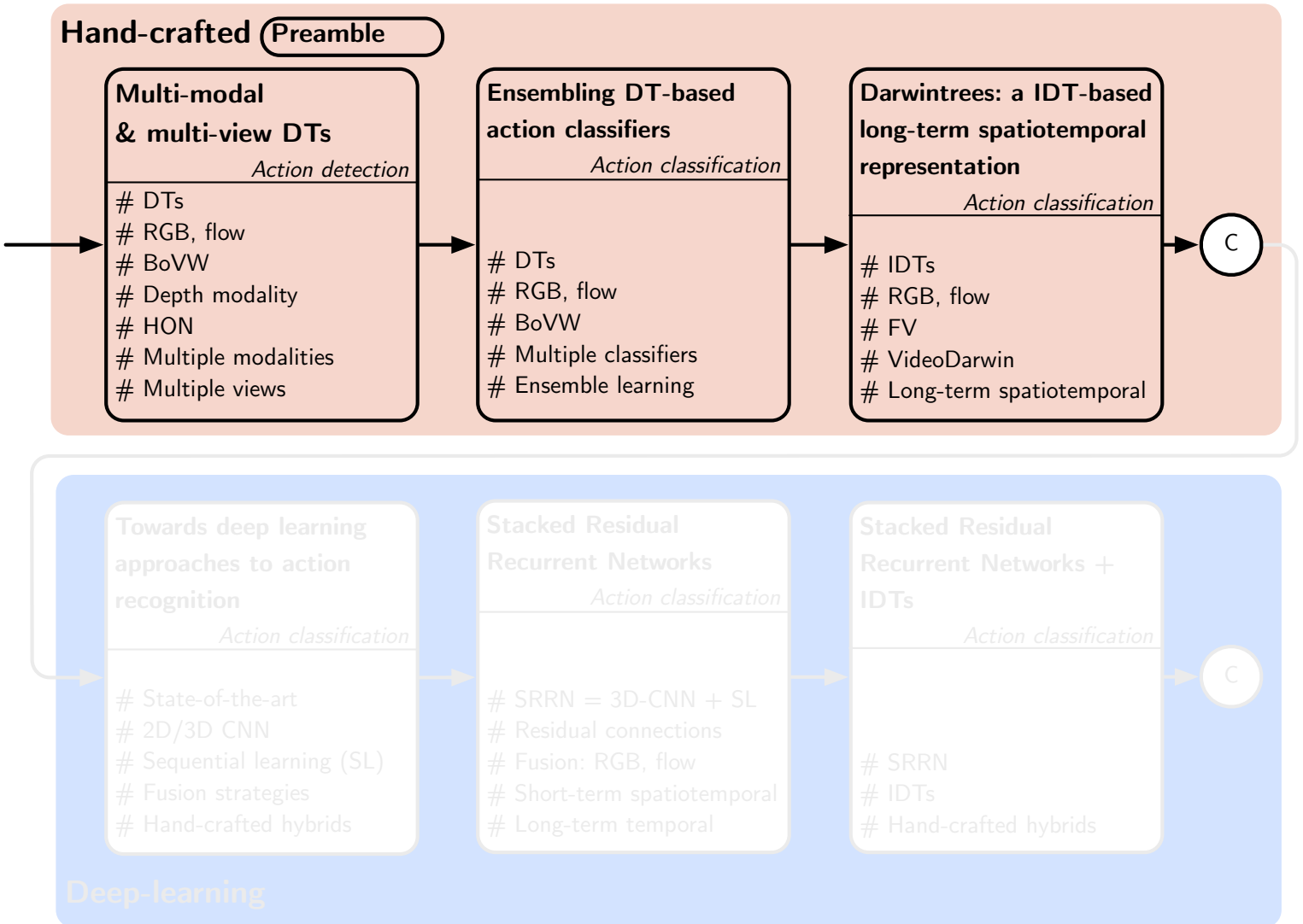
Method	mAP
Wang et al. (2015)	69.4
Karaman et al. (2014)	65.4
Ma et al. (2015)	64.4
Gaidon et al. (2014)	62.4
Ma et al. (2013)	36.9
Patron-Pérez et al. (2012)	42.4
VD	**78.3**
Ours (VD+DT)	75.8

Table: HighFive dataset

Method	mAP
Peng et al. (2014)	**93.8**
Ni et al. (2015)	92.3
Lan et al. (2015)	91.4
Wang et al. (2013)	91.1
Gaidon et al. (2014)	85.5
VD	88.3
Ours (VD+DT)	88.8

Table: Olympic sports dataset

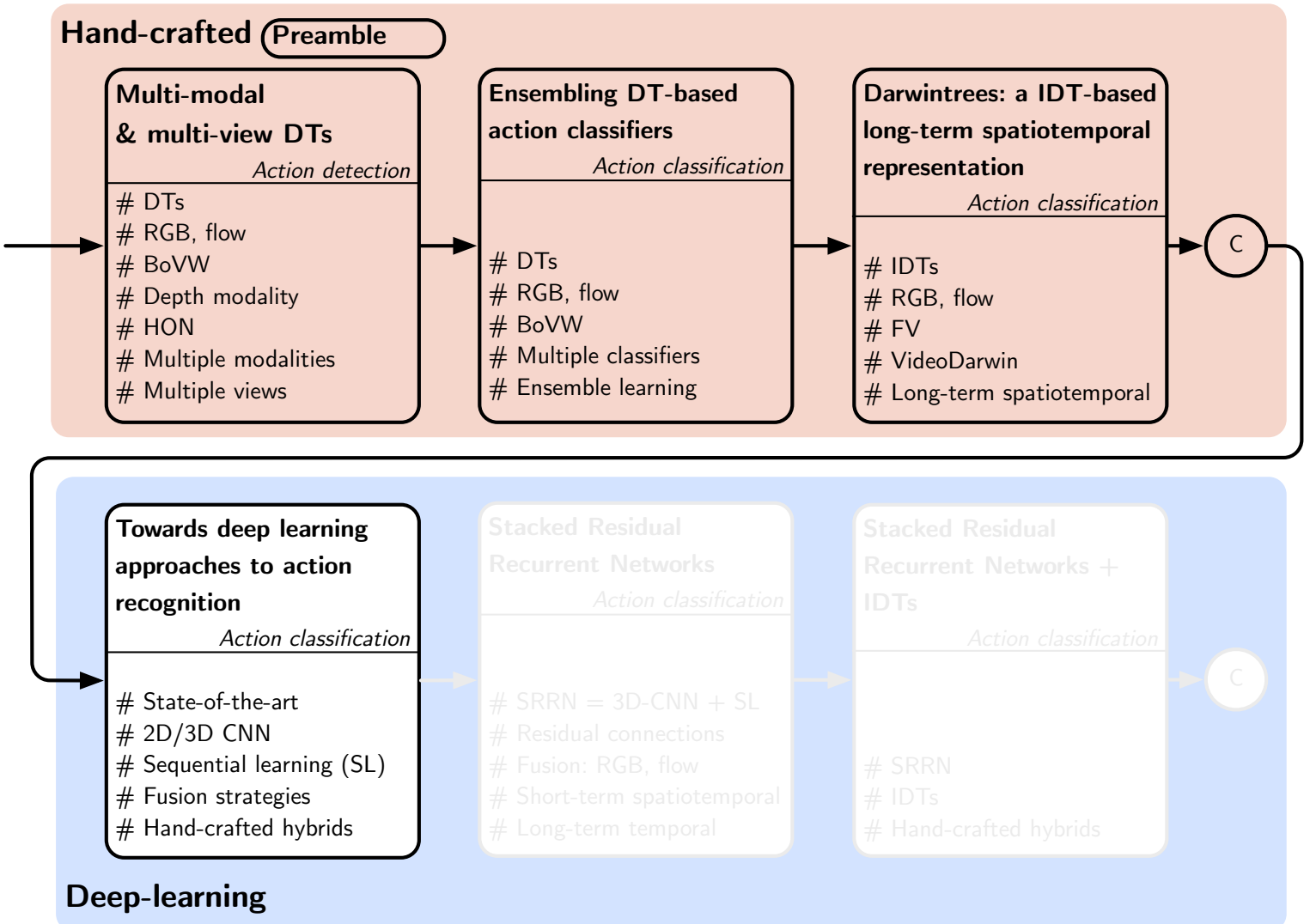
Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
C
Deep-learning

Conclusions: hand-crafted approaches

- We contributed to the main stages of the PR pipeline
- Improved on DTs/IDTs, which dominated the state-of-the-art, and over other base methods
- Combining modalities, ensembling classifiers, and effectively modeling spatio-temporal information are different and complementary ways to enhance action recognition
- The design of reliable representations often require going through a painful loop of re-design, re-adjustment of parameters, and re-evaluation
- Hand-crafted methods have been superseded by deep-based ones

Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
C
Deep-learning

Moving towards deep learning

A large number of architecture variations appeared in the recent years.

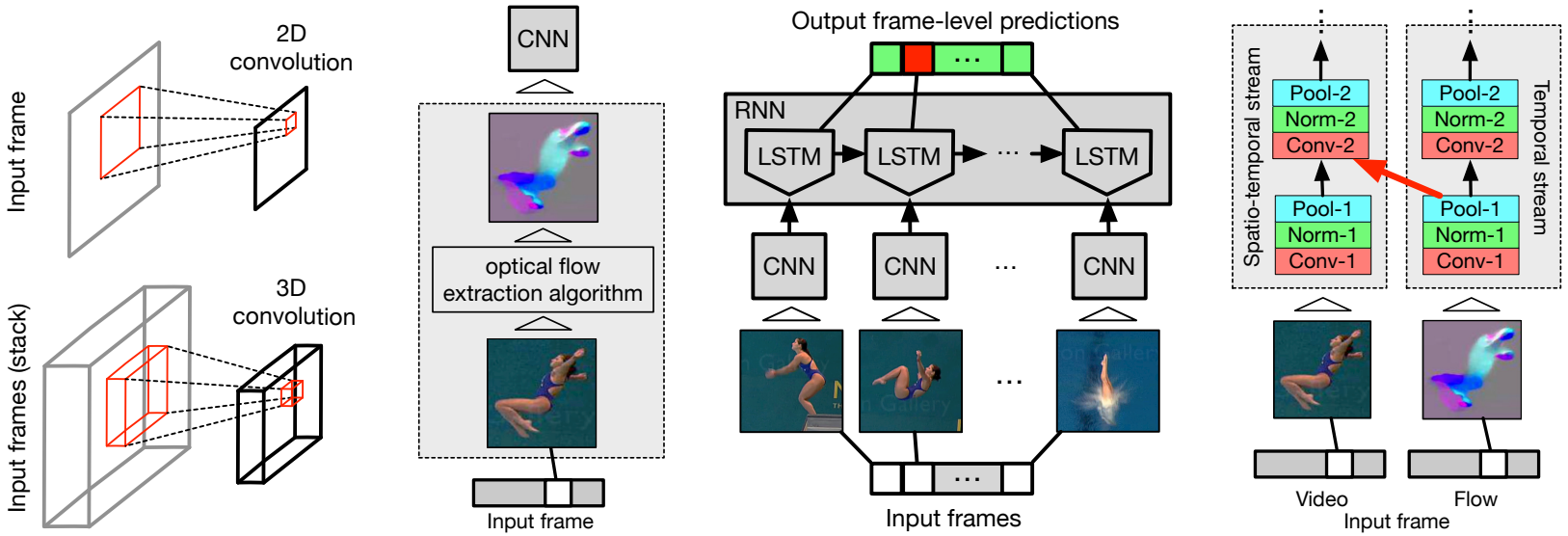
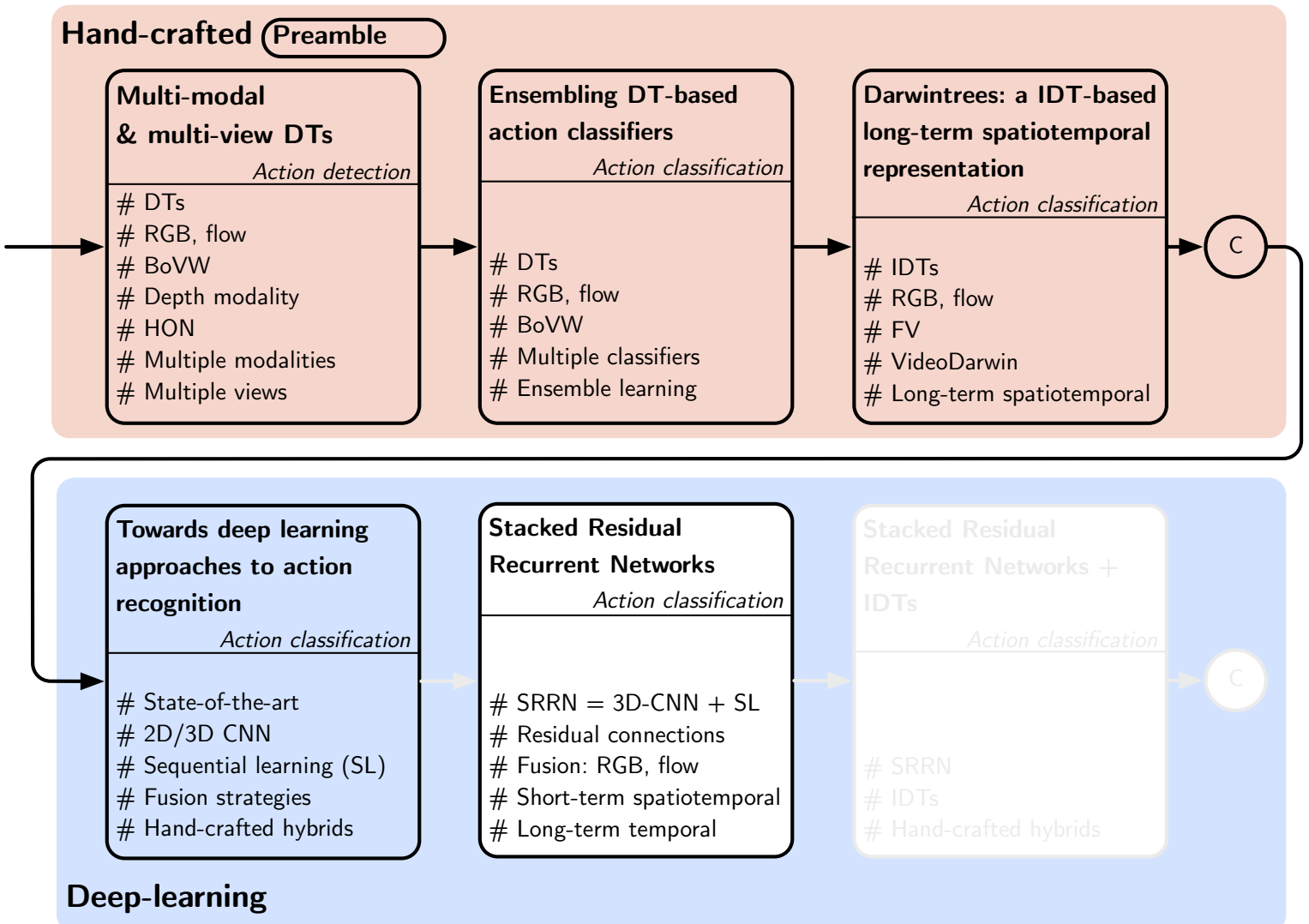


Figure: Deep-learning based approaches to action recognition: (a) 2D/3D CNNs, (b) motion-based, (c) sequential models, and (d) fusion strategies

Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
C
Deep-learning

The two-stage pipeline

$V \in \mathbb{R}^{m_x \times m_y \times l_v}$: a video of length l_v and frame size $m_x \times m_y$ pixels.
$S_i \in \mathbb{R}^{m_x \times m_y \times s}$: a clip of length s.
$K_v = \lfloor \frac{l_v}{r} \rfloor - 1$: the number of clips in V, r the stride between clips.
$A = \{a_i = E(S_i) | i = 1 \dots K_v; a_i \in \mathbb{R}^{d_f}\}$: spatiotemporal features extracted.

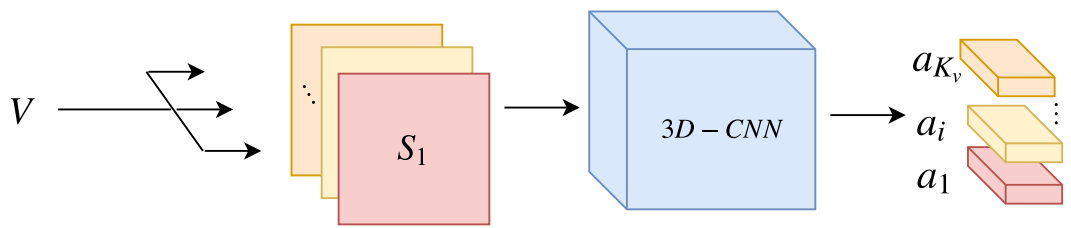


Figure: 3D-CNN for feature extraction

$E(\cdot)$ is not tied to any particular CNN architecture.

The input to the stacked residual recurrent network:

$$x^0 = \{x_t^0 = a_{\sigma(t)} | t = 1, \ldots, T; a_{\sigma(t)} \in A\},$$

where $\sigma(t) = 1 + \lfloor (t-1)\frac{K_v - 1}{T-1} \rfloor$.

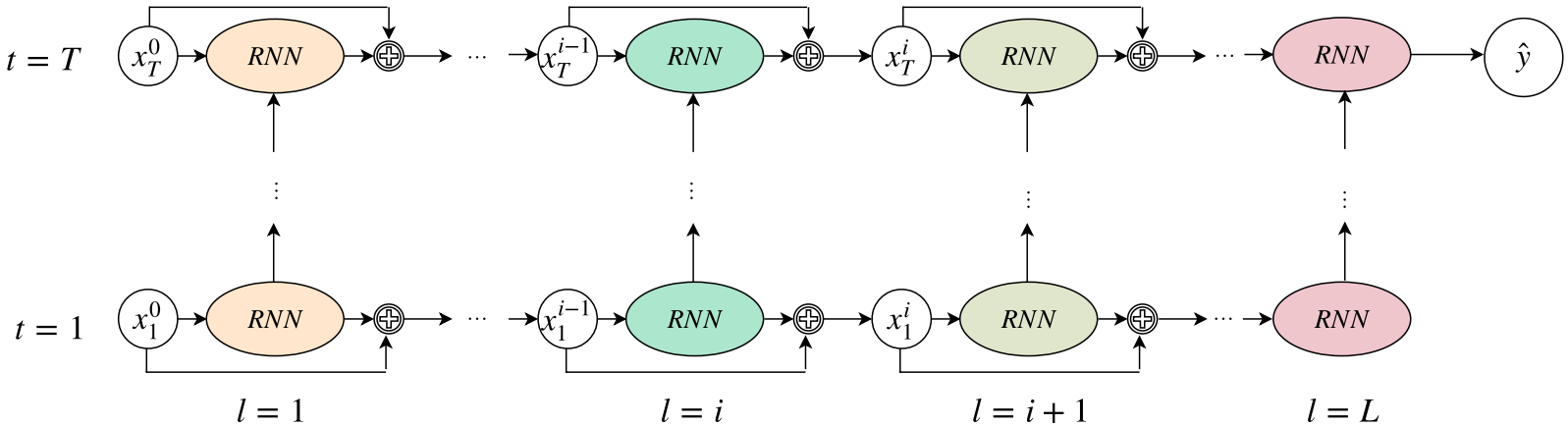


Figure: Stacked residual recurrent neural network of L layers and T timesteps

The model is updated as follows:

$$\begin{aligned} c_t^l, m_t^l &= LSTM_l(c_{t-1}^l, m_{t-1}^l, x_t^{l-1}; \Theta^l) \\ x_t^l &= m_t^l + x_t^{l-1} \end{aligned} \tag{3}$$

Training

Backpropagation adjusts

$$\Theta^l \in \{W_i^l, W_o^l, W_f^l, W_c^l, U_i^l, U_o^l, U_f^l, U_c^l, b_i^l, b_o^l, b_f^l, b_c^l\}, l = 1 \dots L,$$

where $W_.^l \in \mathbb{R}^{h \times n}$, $U_.^l \in \mathbb{R}^{h \times h}$, $b_.^l \in \mathbb{R}^h$ the parameters of the recurrent model at layer $l \in \{1 \dots L\}$.

In our case, because of the recurrent skip connections in Eq. 3, $h = n = d_f$.

Testing

Once the model is trained, the **video level prediction**:

$$\hat{y} = \mathrm{softmax}(W_y m_T^L),$$

where $W_y \in \mathbb{R}^{P \times h}$ and P is the no. classes.

Fusion

We propose several strategies for fusing RGB (c) and flow (f) information.

- Mid-level fusion
  $x^0 = \{x_t^0 = x_t^c \circ x_t^f | t = 1, \ldots, T\}$
- Late fusion
  $\hat{y} = \hat{y}^c \circ \hat{y}^f$

where $\circ \in \{\oplus, \odot\}$ are the element-wise sum and product:

- $u \oplus v = (u_1 + v_1, \ldots, u_m + v_m)$
- $u \odot v = (u_1 \cdot v_1, \ldots, u_m \cdot v_m)$

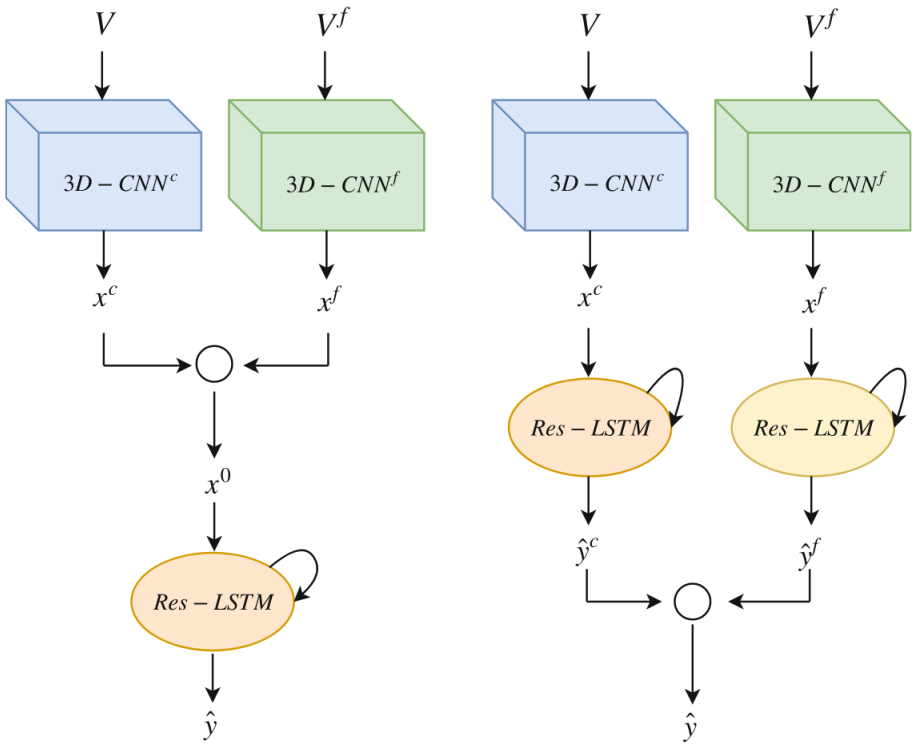


Figure: Mid-level fusion (left) versus late fusion (right). V and V^f, the video and the "flow video"

Results

Datasets

- UCF-101: 13,320 examples, 101 classes. Acc on 1st split (3-fold CV)
- HMDB-51: 6,766 examples, 51 classes. Acc on 1st split (3-fold CV)

Parameters and settings We use a C3D pre-trained on 1-M Sports in order to chose the following parameters:

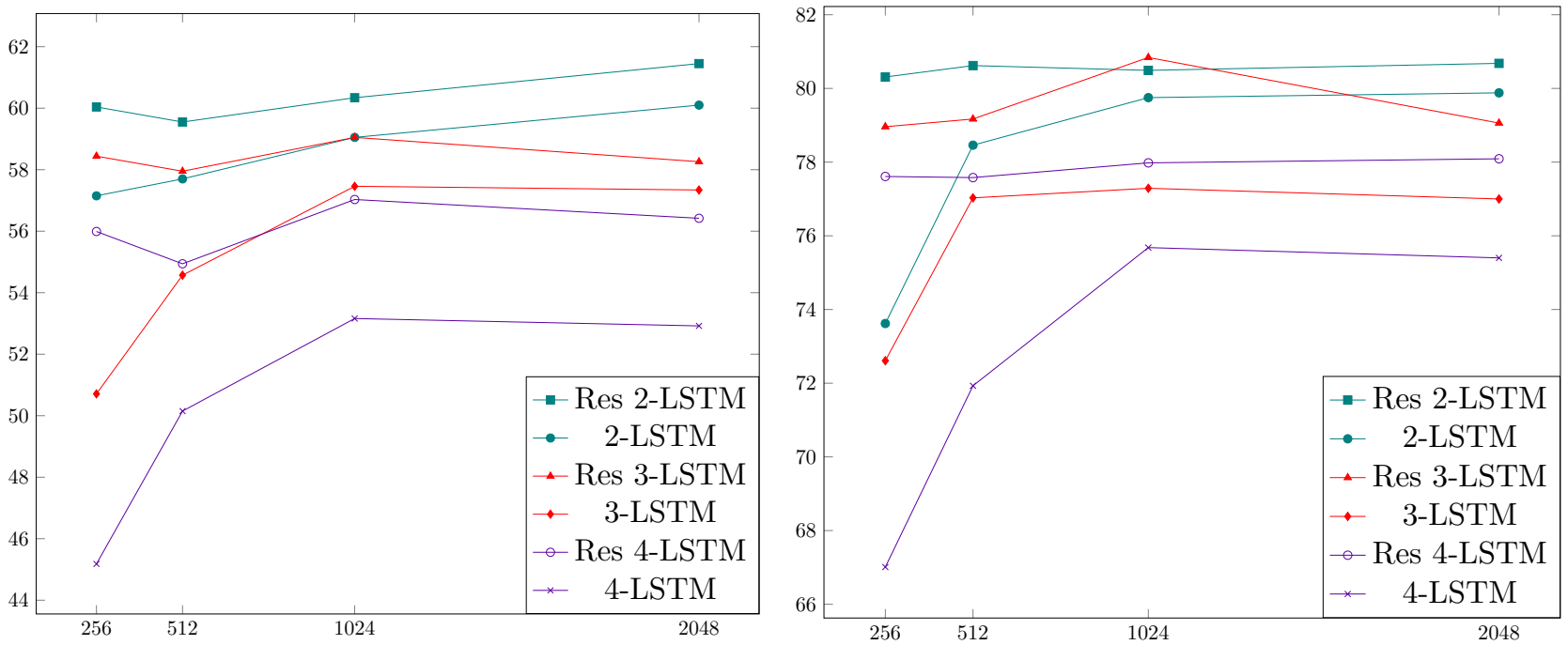
- Hidden layer size
- Depth (no layers in the stack)
- Duration (no timesteps)
- Fusion strategy

Final model evaluation We use TSN as feature extractor in:

- Comparison to other RNN-like architectures
- Qualitative evaluation and confusion matrices
- Combination with IDTs
- Comparison to SOTA methods

Results: parameters and settings

We tested $h = \{256, 512, 1024, 2048\}$ and $L = \{2, 3, 4\}$. PCA is applied to match the output dimensionality of the C3D to the desired h.



(a) HMDB-51 (b) UCF-101

Figure: Effect of the hidden layer size (x-axis) and depth (line colors)

Results: parameters and settings

We evaluated $T = \{5, 15, 25, 35\}$ and different fusion strategies of $\{\text{mid}, \text{late}\} \times \{\oplus, \odot\}$.

T	5	15	25	35
HMDB-51	59.5	60.2	**61.5**	61.4
UCF-101	77.9	79.9	79.5	**80.9**

Table: Impact of the value of no. timesteps T on accuracy

Strategy	Mid fusion	Late fusion
$\oplus$ (element-wise sum)	59.3	65.2
$\odot$ (element-wise product)	56.5	**68.0**
$w_1.Flow + w_2.RGB$	–	63.3

Table: Different mid and late fusion strategies for the 2-layer Res-LSTM on the HMDB-51 dataset. The weighted sum $w_1 = 1.5$ and $w_2 = 1.0$[26] performed poorly

[25] *Limin Wang et al. "Temporal segment networks: towards good practices for deep action recognition". In: European Conference on Computer Vision. Springer. 2016, pp. 20–36.*

[26] *Wang et al., "Temporal segment networks: towards good practices for deep action recognition".*

Results: final model evaluation

Comparison to L^{2}STM[27]

Data Types	L^{2}STM	Res-LSTM	Gain
Human-Object Interaction	86.7	88.2	$\uparrow$ **1.5**
Human-Human Interaction	95.4	96.9	$\uparrow$ **1.5**
Body-Motion Only	88.6	90.7	$\uparrow$ **2.1**
Playing Instrument	-	97.3	-
Sports	-	93.2	-
Pizza Tossing	72.7	66.7	$\downarrow$ 6.0
Mixing Batter	86.7	91.1	$\uparrow$ **4.4**
Playing Dhol	100	100	$\equiv$
Salsa Spins	100	97.7	$\downarrow$ 2.3
Ice Dancing	100	100	$\equiv$

Table: Comparison to L^{2}STM on Split-1 of UCF-101 for coarse categories and fine-grained classes

[27] *Lin Sun et al. "Lattice Long Short-Term Memory for Human Action Recognition". In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE. 2017, pp. 2166–2175.*

Results: final model evaluation

Qualitative evaluation

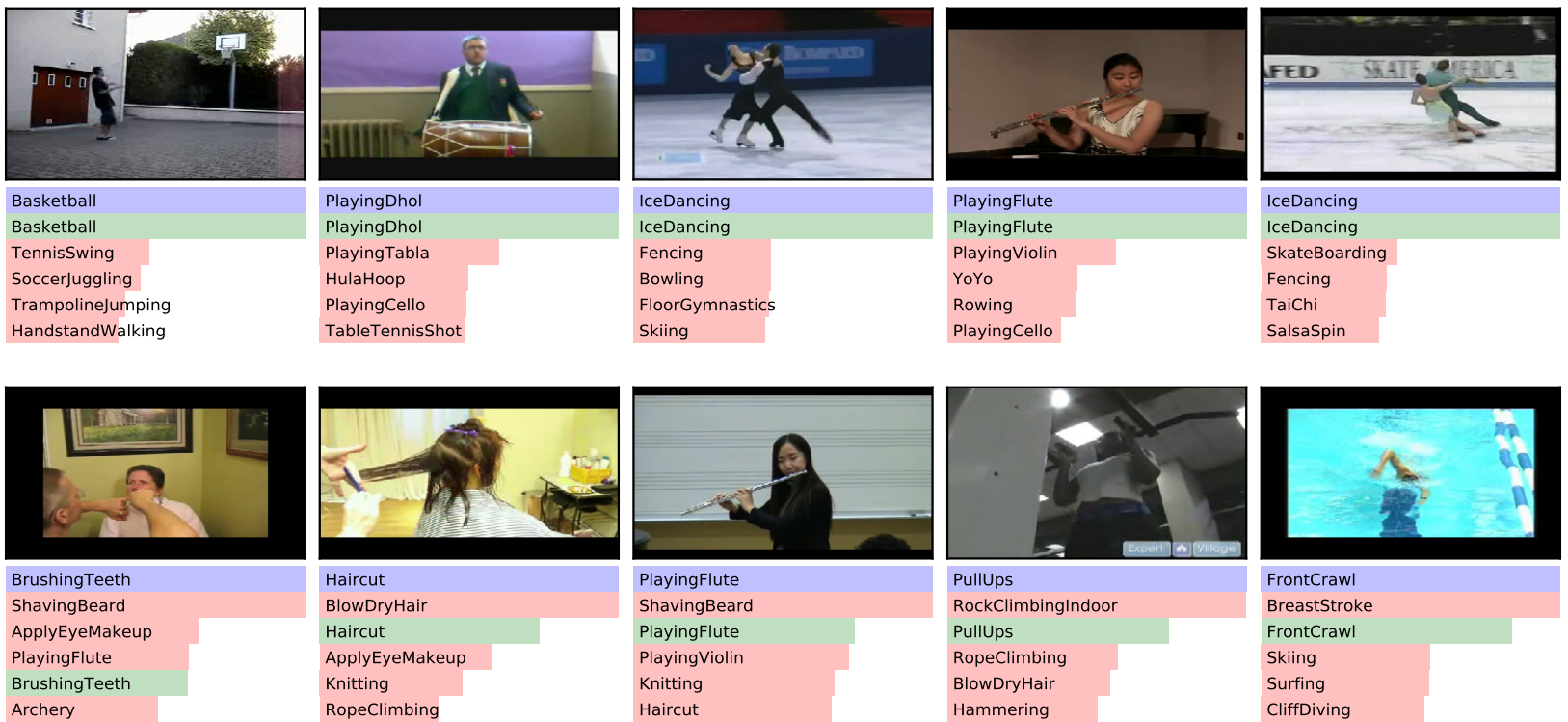
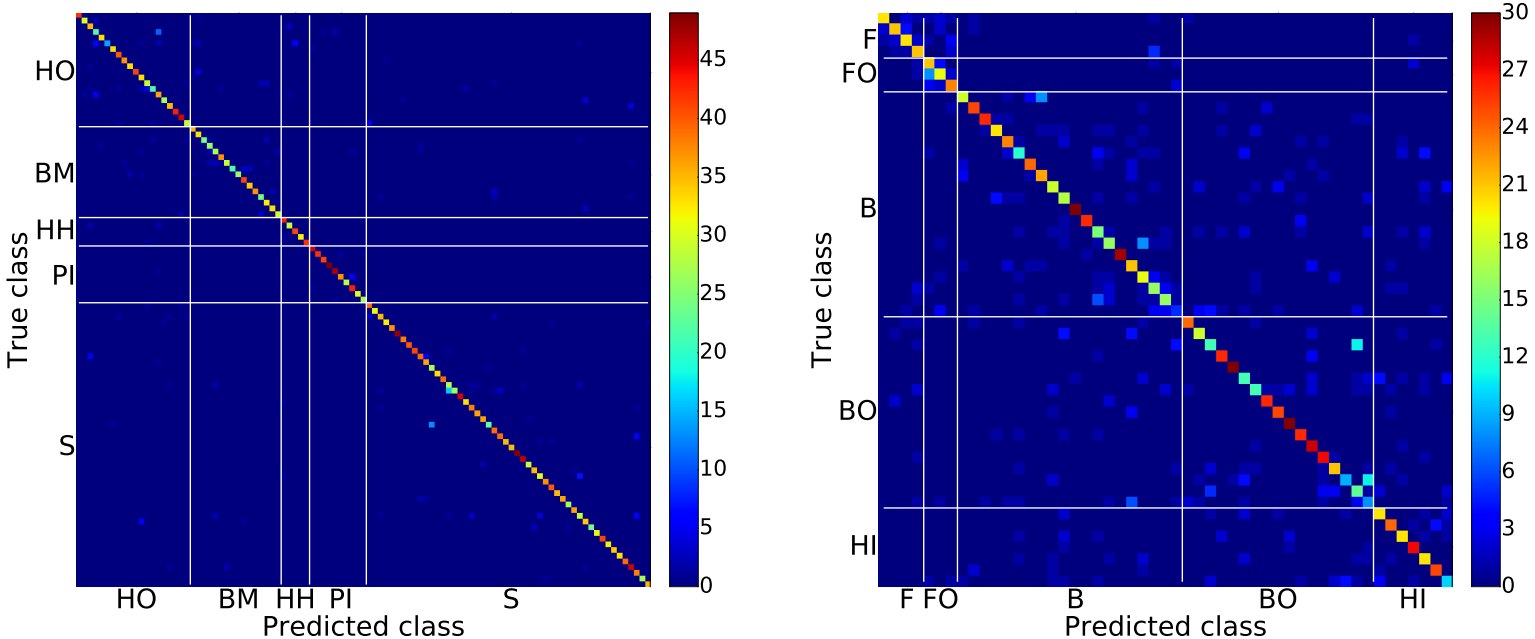


Figure: Sample video classification results showing the top-5 class predictions based on confidence. First row: correctly classified videos; second row: miss-classified videos. Key – blue bar: groundtruth class; green bar: correct class prediction; red bar: incorrect class prediction

Results: final model evaluation

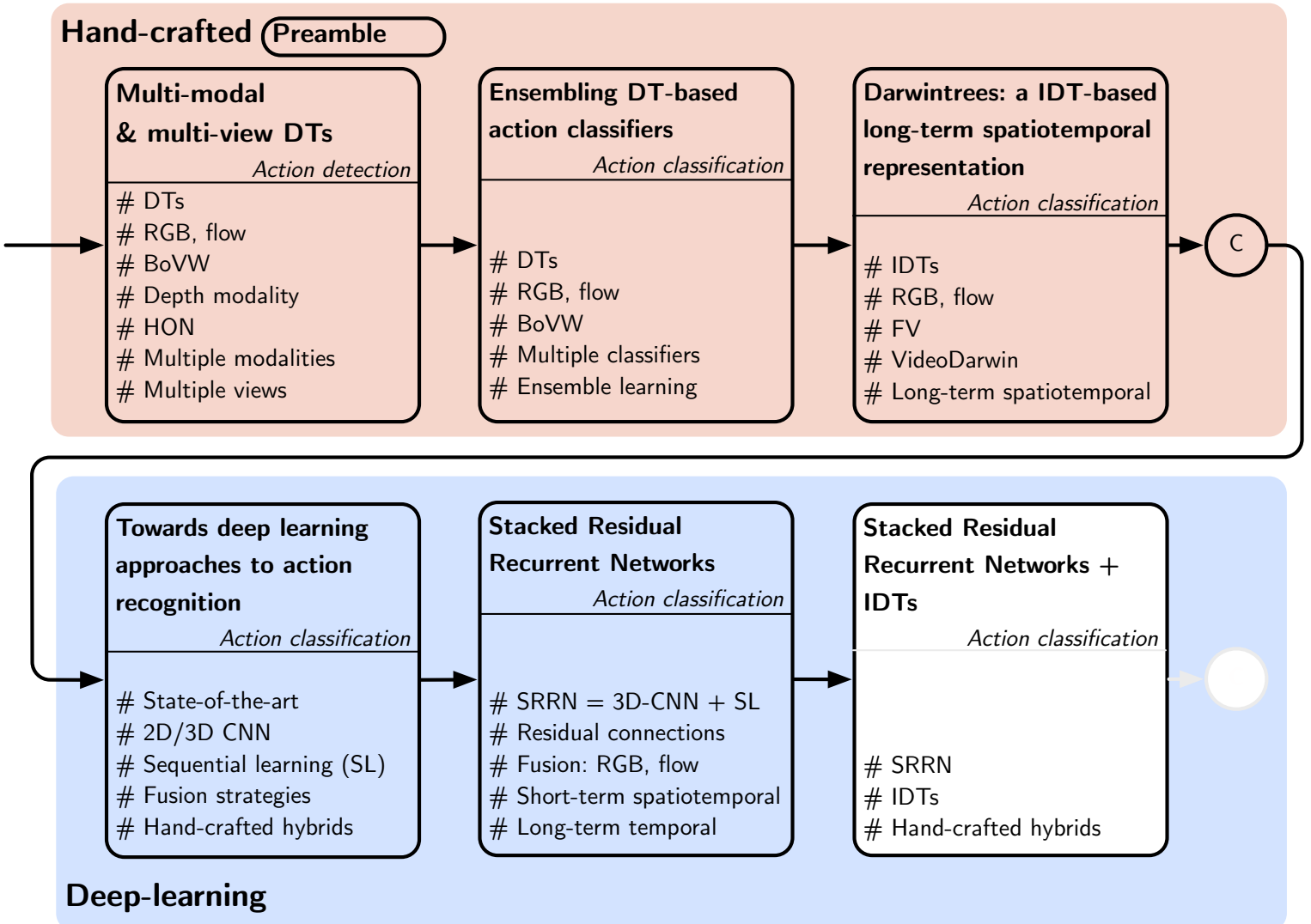
Confusion matrices



(a) Res-LSTM on UCF-101 (b) Res-LSTM on HMDB-51

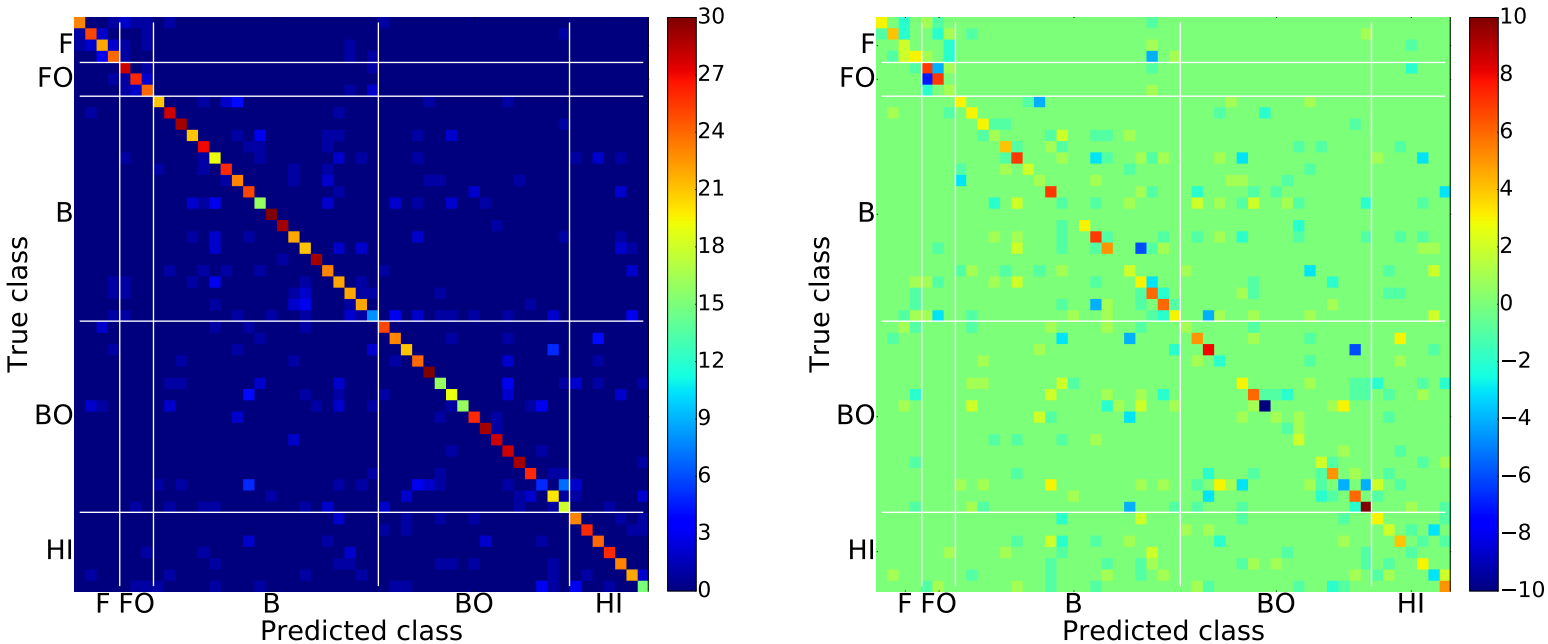
Figure: Re-arranged classes to group coarse categories. UCF-101: human-object interaction (HO), body-motion only (BM), human-human interaction (HH), playing musical instrument (PI), and sports (S). For HMDB-51: facial actions (F), facial actions w/ object manipulation (FO), body movements (BM), body movements w/ object interaction (BO), and body movements for human interaction (HH)

Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
Deep-learning

Results: final model evaluation

Combination with IDTs (I)



(a) Res-LSTM ⊙ IDT on HMDB-51 (b) Res-LSTM ⊙ IDT - Res-LSTM on HMDB-51

Figure: Re-arranged confusion matrices after combining our Res-LSTM with IDT on HMDB-51

Results: final model evaluation

Combination with IDTs (II)

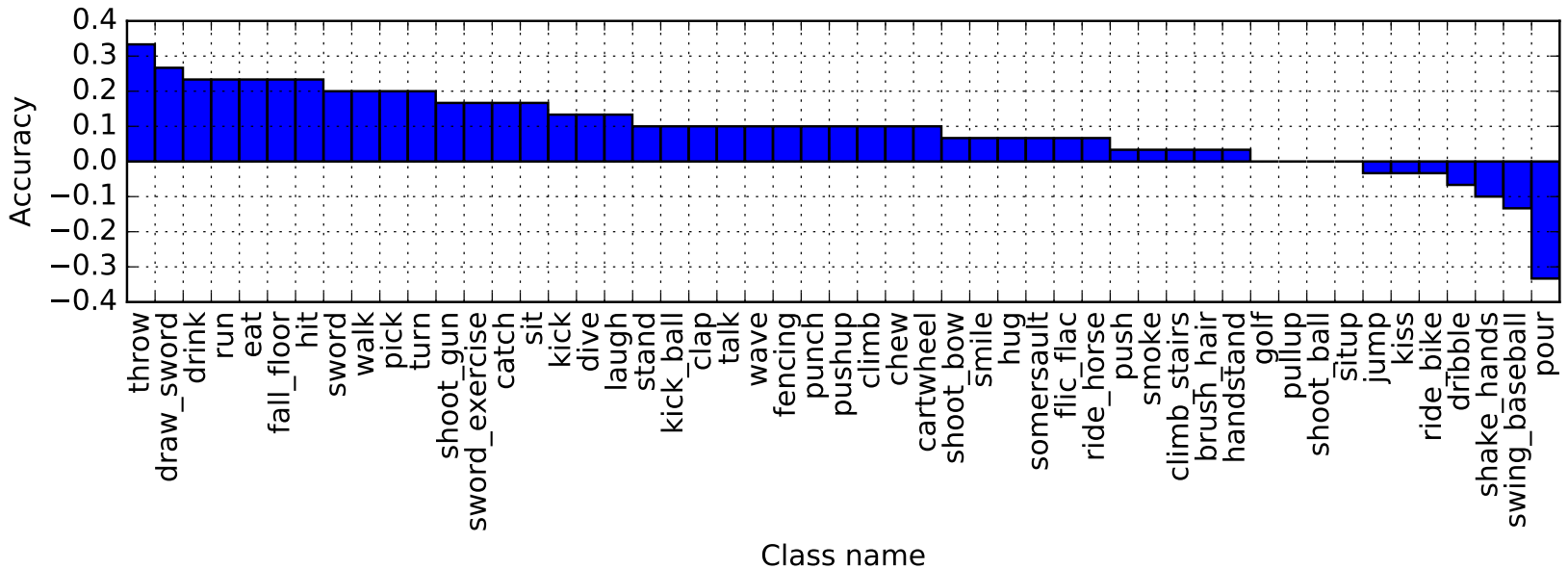


Figure: Per-class accuracy improvement after combining our Res-LSTM with IDT on HMDB-51

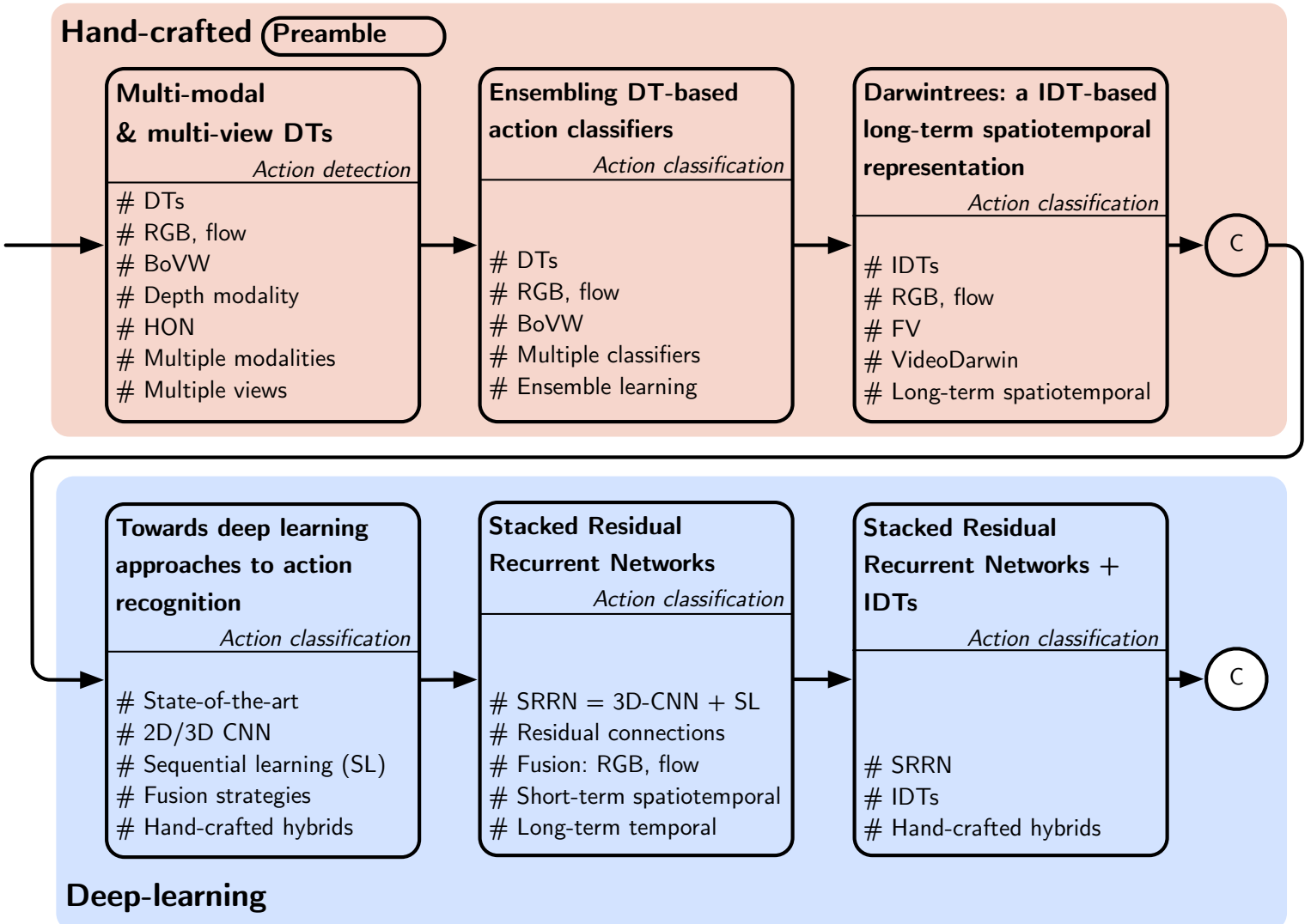
Results: final model evaluation

Comparison to state-of-the-art

Model	Method	UCF-101	HMDB-51
Static models	FST-CNN (2015)	88.1	59.1
	TDD (2015)	90.3	63.2
	KV-CNN (2016)	93.1	63.3
	LTC (2018)	91.7	64.8
	TDD + IDT (2015)	91.5	65.9
	ST-ResNet (2016)	93.4	66.4
	STM-ResNet (2017)	94.2	68.2
	LTC + IDT (2018)	92.7	67.2
	TSN (2016)	94.2	69.4
	ST-ResNet + IDT (2016)	94.6	70.3
	STM-ResNet + IDT (2017)	94.9	72.2
	STC-ResNext (2018)	95.8‡	72.6‡
	I3D (2017)	**98.0**	**80.7**
Sequential models	LRCN (2015)	82.9	-
	AttLSTM (2015)	77.0*	41.3
	UnsuperLSTM (2015)	84.3*	44.0‡
	RLSTM-g3 (2016)	86.9	55.3
	TwoLSTM (2015)	88.3*	-
	VideoLSTM (2018)	89.2*	56.4
	VideoLSTM + IDT (2018)	91.5*	63.0
	L^2STM (2017)	**93.6**	66.2
	PreRNN (2018)	**93.7***	–
	Res-LSTM (ours)	92.5*	68.0*
	Res-LSTM (ours) $\odot$ IDT	93.0*	**76.9***

‡ Only RGB modality is used
* Evaluation on split-1

Outline


Hand-crafted
Preamble
Multi-modal & multi-view DTs
Action detection
DTs
RGB, flow
BoVW
Depth modality
HON
Multiple modalities
Multiple views
Ensembling DT-based action classifiers
Action classification
DTs
RGB, flow
BoVW
Multiple classifiers
Ensemble learning
Darwintrees: a IDT-based long-term spatiotemporal representation
Action classification
IDTs
RGB, flow
FV
VideoDarwin
Long-term spatiotemporal
C
Towards deep learning approaches to action recognition
Action classification
State-of-the-art
2D/3D CNN
Sequential learning (SL)
Fusion strategies
Hand-crafted hybrids
Stacked Residual Recurrent Networks
Action classification
SRRN = 3D-CNN + SL
Residual connections
Fusion: RGB, flow
Short-term spatiotemporal
Long-term temporal
Stacked Residual Recurrent Networks + IDTs
Action classification
SRRN
IDTs
Hand-crafted hybrids
C
Deep-learning

Conclusions: deep-based approaches

- The two stages: modeling of local spatiotemporal features (3D-CNN) and long-term temporal dynamics (SRRN)
- Residual connections are effective in multi-layer RNNs
- Shallower SRRN perfomed better
- Late fusion $>$ middle fusion. Element-wise product $>$ element-wise sum
- Combination with hand-crafted features – in particular, IDTs – further boosts performance
- Sequential models obtain $\sim 5-10\%$ less than static models

Conclusions

- **The problem** is far from being solved, but we are already dealing with realistic videos
- In **larger datasets**: larger intra-class and lower inter-class variance
- **Performance saturation** on some benchmarking action classification datasets? Transfer learning!
- **"Optimality" of networks** for action recognition. The convenience of pre-computed optical flow
- **Longer videos** with more complex temporal semantics will challenge current techniques
- **Work is required on** temporal and spatiotemporal detection (and segmentation)
- **Non-fully supervised approaches** are becoming more trendy and worth to look at: unsupervised, semi-supervised, weakly supervised, and so on

Future work

- Experiments on deeper SRRNs
- To build on top of I3D (feature extractor) pre-trained on Kinetics dataset[28]
- End-to-end training of 3D-CNN and SRRN
- Design a dataset with richer temporal semantics
- Weakly-supervised temporal action segmentation

[28] *Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. IEEE. 2017, pp. 4724–4733.*

Publications

5 Journals	**11** conferences	**248** citations	**9** h-index	**8** i10-index

Indexed journals

- Clapés, A., Reyes, M., & Escalera, S. (2013). Multi-modal user identification and object recognition surveillance system. Pattern Recognition Letters, 34(7), 799-808.
- Reyes, M., Clapés, A., Ramrez, J., Revilla, J. R., & Escalera, S. (2013). Automatic digital biometry analysis based on depth maps. Computers in Industry, 64(9), 1316-1325.
- Cepero, A., Clapés, A., & Escalera, S. (2015). Automatic non-verbal communication skills analysis: A quantitative evaluation. AI Communications, 28(1), 87-101.
- Palmero, C., Clapés, A., Bahnsen, C., Mgelmose, A., Moeslund, T. B., & Escalera, S. (2016). Multi-modal rgbdepththermal human body segmentation. International Journal of Computer Vision, 118(2), 217-239.
- Clapés, A., Pardo, ., Vila, O. P., & Escalera, S. (2018). Action detection fusing multiple Kinects and a WIMU: an application to in-home assistive technology for the elderly. Machine Vision and Applications, 1-24.

International conferences and workshops

- Clapés, A., Reyes, M., & Escalera, S. (2012). User identification and object recognition in clutter scenes based on rgb-depth analysis. In International Conference on Articulated Motion and Deformable Objects (pp. 1-11). Springer, Berlin, Heidelberg.
- Pardo, ., Clapés, A., Escalera, S., & Pujol, O. (2014). Actions in context: system for people with dementia. In Citizen in Sensor Networks (pp. 3-14). Springer, Cham.
- Konovalov, V., Clapés, A., & Escalera, S. (2013). Automatic Hand Detection in RGB-Depth Data Sequences. In CCIA (pp. 91-100).
- Cepero, A., Clapés, A., & Escalera, S. (2013). Quantitative Analysis of Non-Verbal Communication for Competence Analysis. In CCIA (pp. 105-114).
- Mogelmose, A., Bahnsen, C., Moeslund, T., Clapés, A., & Escalera, S. (2013). Tri-modal person re-identification with rgb, depth and thermal features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 301-307).
- Bagheri, M., Gao, Q., Escalera, S., Clapes, A., Nasrollahi, K., Holte, M. B., & Moeslund, T. B. (2015). Keep it accurate and diverse: Enhancing action recognition performance by ensemble learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 22-29).

- Escalante, H. J., Ponce-Lpez, V., Wan, J., Riegler, M. A., Chen, B., Clapés, A., ... & Mller, H. (2016). ChaLearn Joint Contest on Multimedia Challenges Beyond Visual Analysis: An overview. In ICPR (pp. 67-73).
- Ponce-Lpez, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., ... & Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions-dataset and results. In European Conference on Computer Vision (pp. 400-418). Springer, Cham.
- Asadi-Aghbolaghi, M., Clapes, A., Bellantonio, M., Escalante, H. J., Ponce-Lpez, V., Bar, X., ... & Escalera, S. (2017). A survey on deep learning based approaches for action and gesture recognition in image sequences. In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on (pp. 476-483). IEEE.
- Clapés, A., Tuytelaars, T., & Escalera, S. (2017, October). Darwintrees for action recognition. In The IEEE Int. Conf. on Computer Vision (ICCV) (pp. 3169-3178).
- Asadi-Aghbolaghi, M., Clapés, A., Bellantonio, M., Escalante, H. J., Ponce-Lpez, V., Bar, X., ... & Escalera, S. (2017). Deep learning for action and gesture recognition in image sequences: A survey. In Gesture Recognition (pp. 539-578). Springer, Cham.

Questions?