# Comprendre les données visuelles à grande échelle

**ENSIMAG**
**2019-2020**

Karteek Alahari & Diane Larlus

https://project.inria.fr/bigvisdata/

# Au programme

- Organisation du cours
- Introduction
  - Contexte et applications
  - Aperçus des taches
  - Evaluation
- Représentation des données visuelles
  - Descripteurs locaux et globaux, réseaux de neurones
  - Application à la fouille de donnée
- Problème de la reconnaissance
  - Classification d'images et de vidéo
  - Séparateurs à Vaste marge (SVM)
  - Pour aller plus loin

Crédits pour les transparents: C. Schmid, H. Wang

# Why automatic video understanding?

- Query for videos in professional Archives and YouTube
- Analyze and describe content of videos

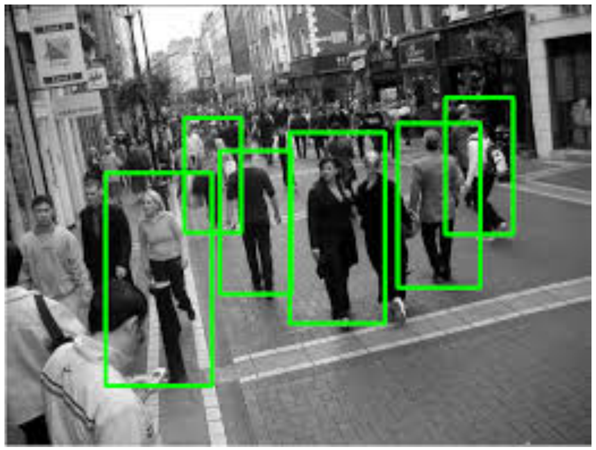
Education: How do I make a pizza?

Sociology research: Influence of character smoking in movies

# Why automatic video understanding?

- Car safety & self-driving and video surveillance
  - Detection of humans (pedestrians) and their motion, detection of unusual behavior

Courtesy Volvo



Courtesy Embedded Vision Alliance

# Machine visual perception - applications

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

# Machine visual perception - applications

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and  the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

# Machine visual perception - applications

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

# Machine visual perception - applications

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

# Action recognition: Difficulties

- Large variations in appearance
  - Viewpoint changes
  - Intra-class variation
  - Camera motion

# Difficulties: Viewpoint change

# Difficulties: within-class variations

# Action recognition: Difficulties

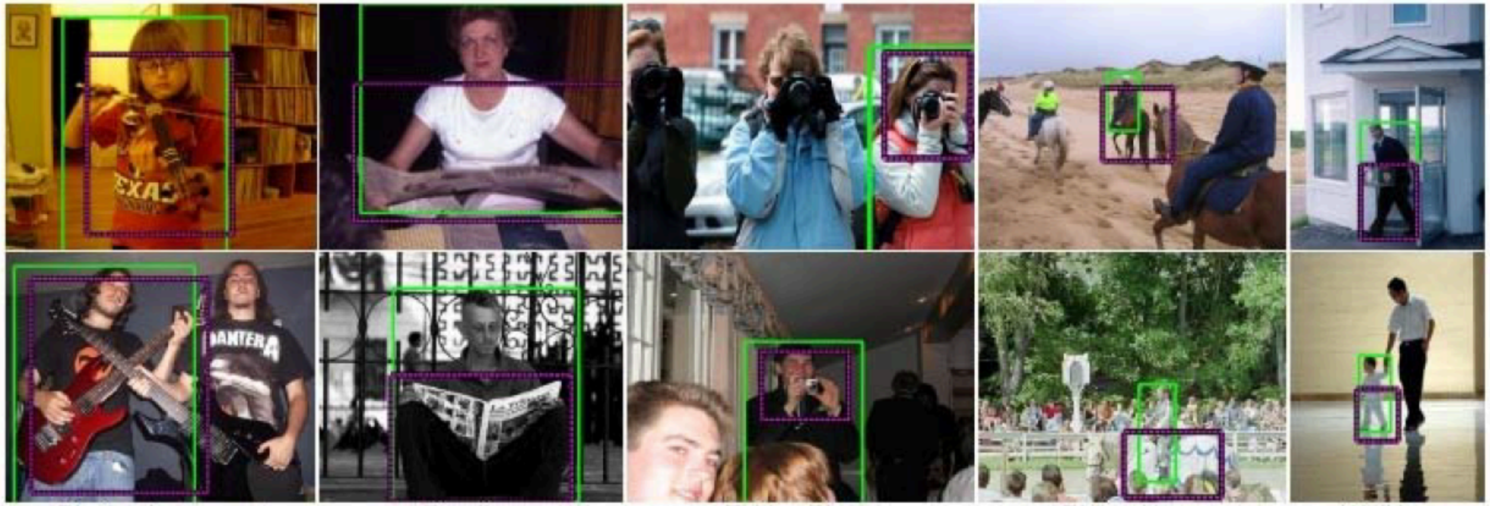
- Large variations in appearance
  - Viewpoint changes
  - Intra-class variation
  - Camera motion

- Manual collection of training data is difficult
  - Many action classes, rare occurrence
  - Pose and object annotation often a plus

- Action vocabulary is not well defined
  - What is the action granularity?
  - How to represent composite actions?

# Action recognition – approaches

- Action recognition from still images
  - Human pose + interaction with objects

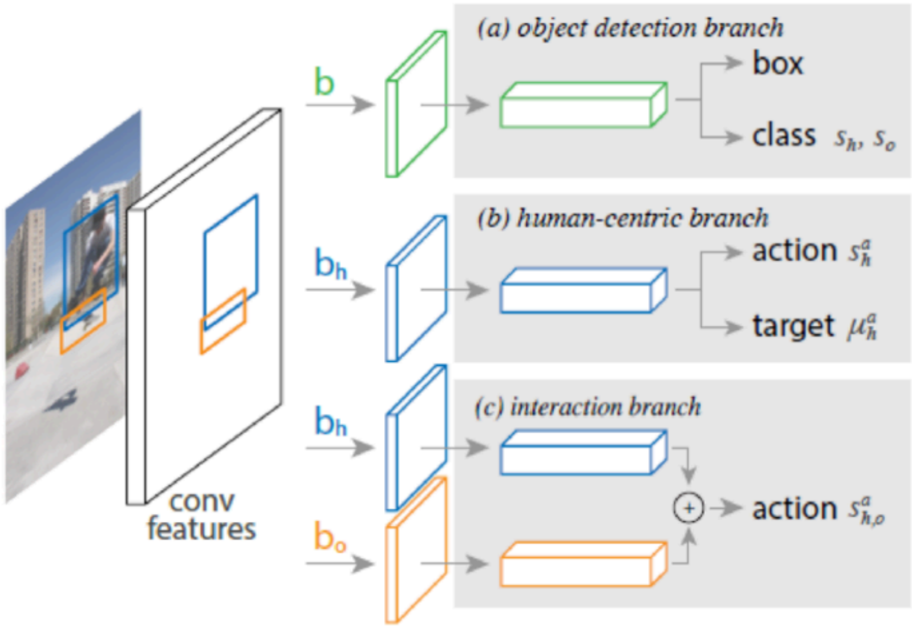


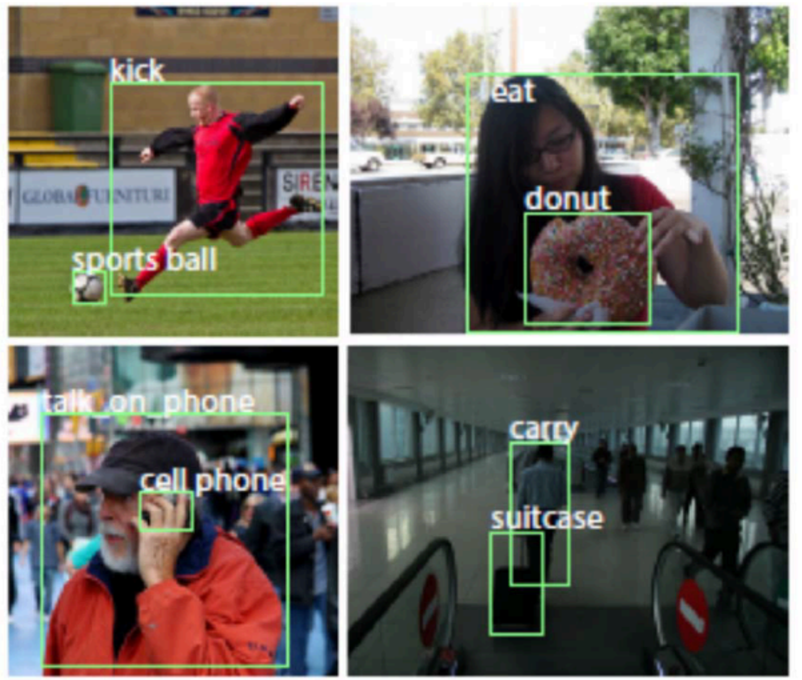
Playing Instrument | Reading | Taking Photo | Riding Horse | Walking

Results on PASCAL VOC 2010 Human action classification dataset [Prest et al., PAMI 2012]

# Action recognition – approaches

- Action recognition from still images
  - Human pose + interaction with objects




V-COCO

[Detecting and Recognizing Human-Object Interactions.
G. Gkioxari, R. Girshick, P. Dollar and K. He. CVPR 2018]

# Action recognition – approaches

- Motion information necessary to disambiguate actions

Open or close door?

- Motion often sufficient by itself

# Motion perception

- Gunnar Johansson [1973] pioneered studies on sequence based human motion analysis

- Moving light displays enable identification of motion, familiar people and gender

# Action classification in videos

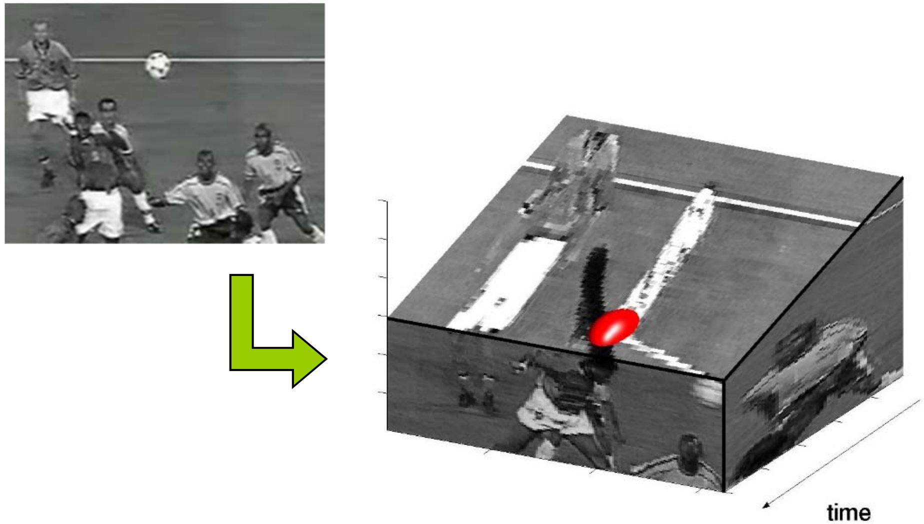
- Space-time interest points [Laptev, IJCV'05]
- Dense trajectories [Wang and Schmid, ICCV'13]
- Video-level CNN features

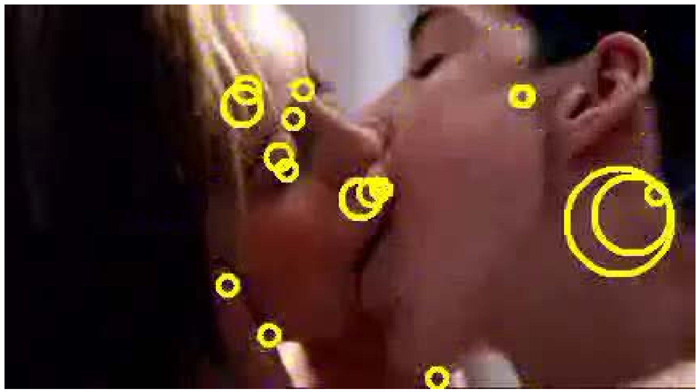
# Space-time interest points (STIP)

- Space-time corner detector
  [Laptev, IJCV 2005]

$$H \;=\; \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

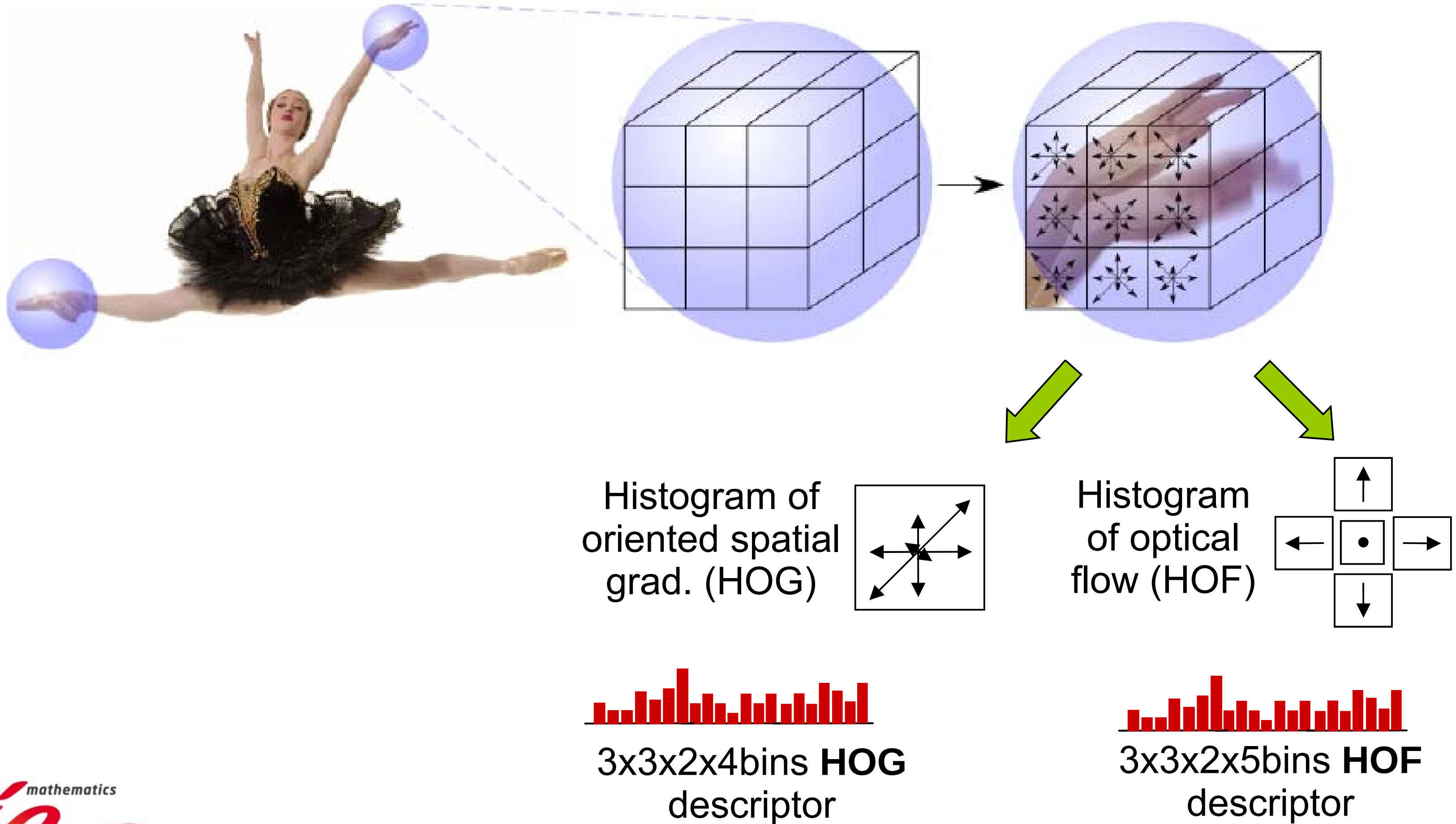
$$\mu = \begin{pmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_x I_y & I_y I_y & I_y I_t \\ I_x I_t & I_y I_t & I_t I_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$



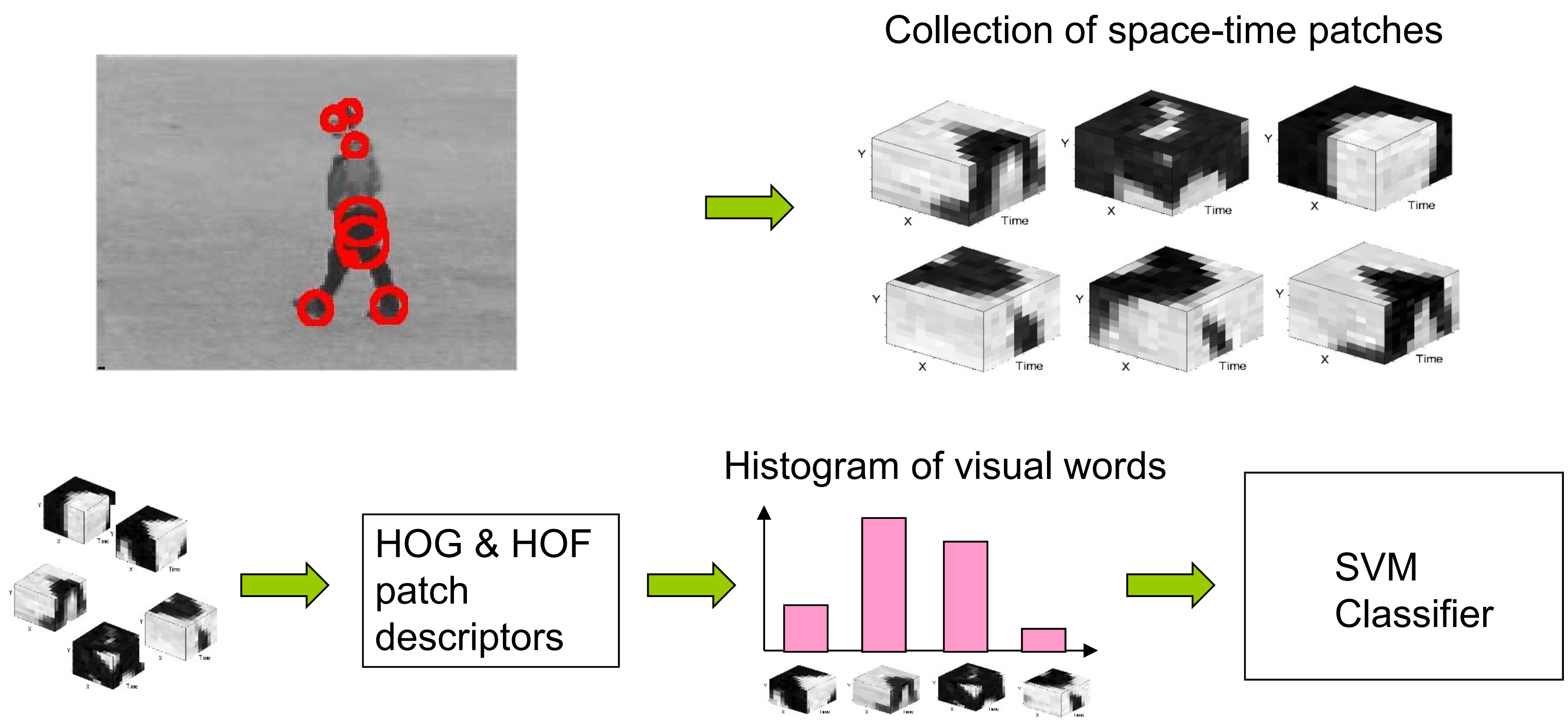
# STIP descriptors

Space-time interest points



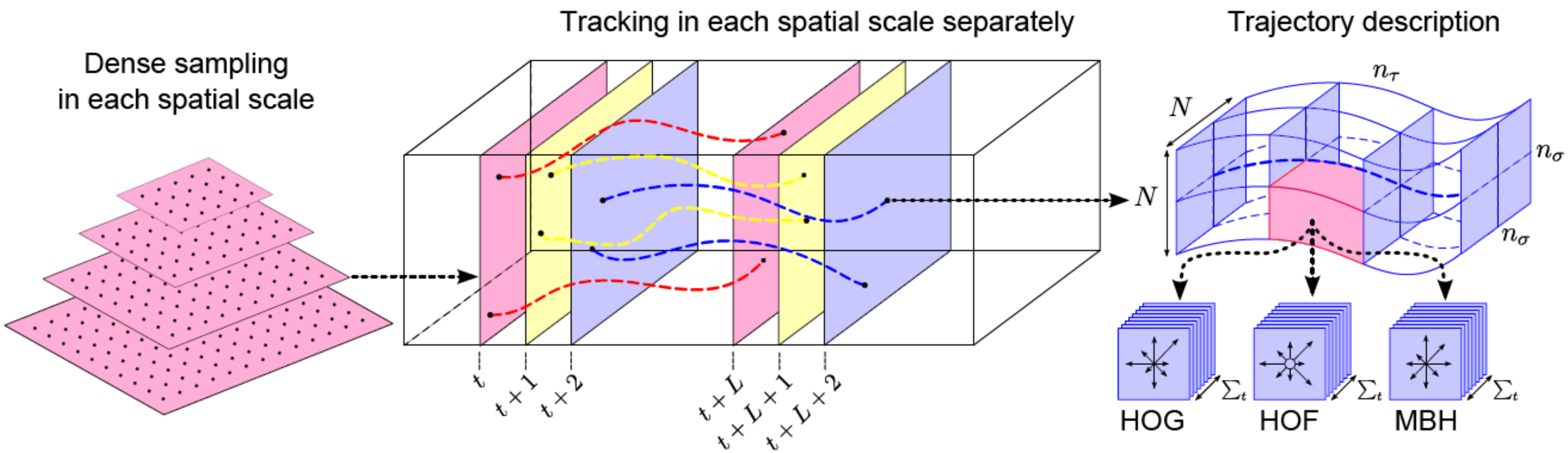
# Action classification

- Bag of space-time features + SVM [Schuldt'04, Niebles'06, Zhang'07]



# State of the art for video description

- Dense trajectories [Wang et al., IJCV'13] and Fisher vector encoding [Perronnin et al. ECCV'10]

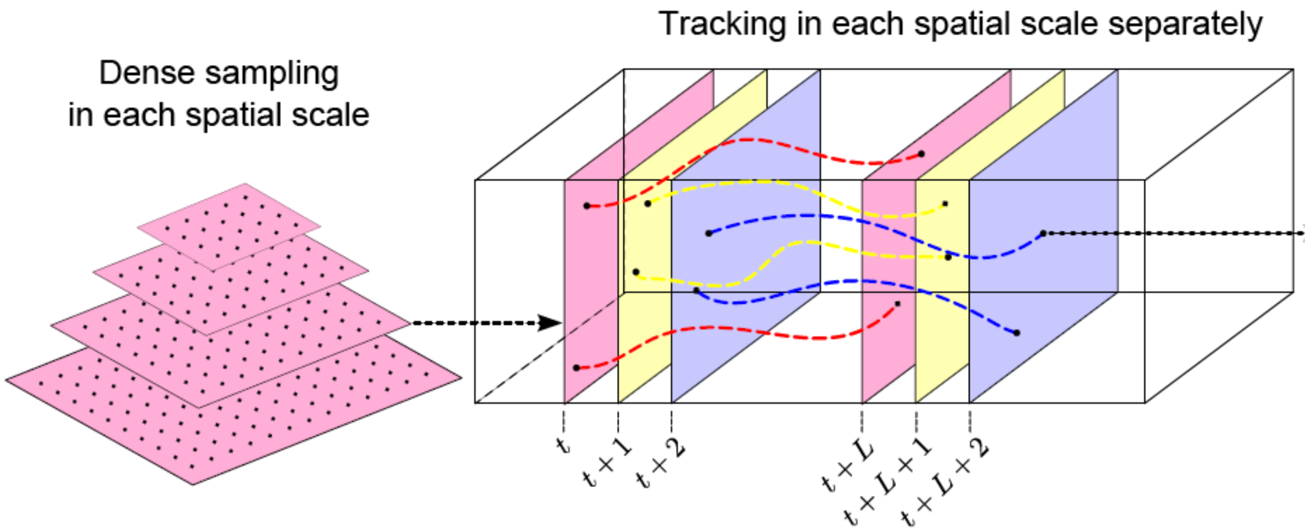



- Orderless representation

informatics mathematics Inria

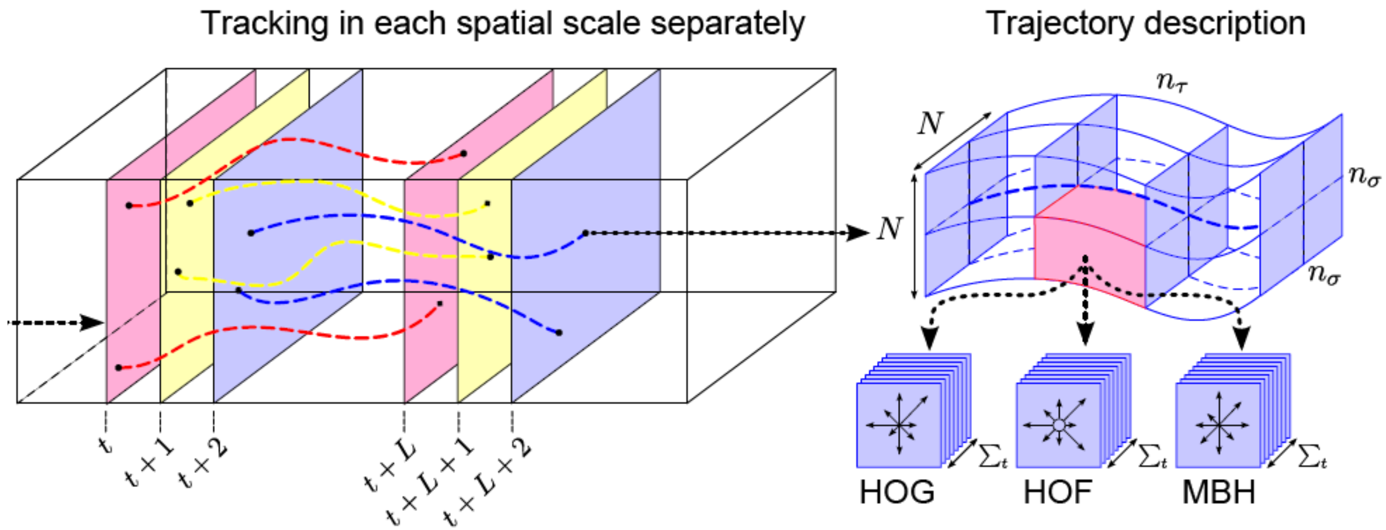
# Dense trajectories [Wang et al., IJCV'13]

- Dense sampling at several scales
- Feature tracking based on optical flow for several scales
- Length 15 frames, to avoid drift



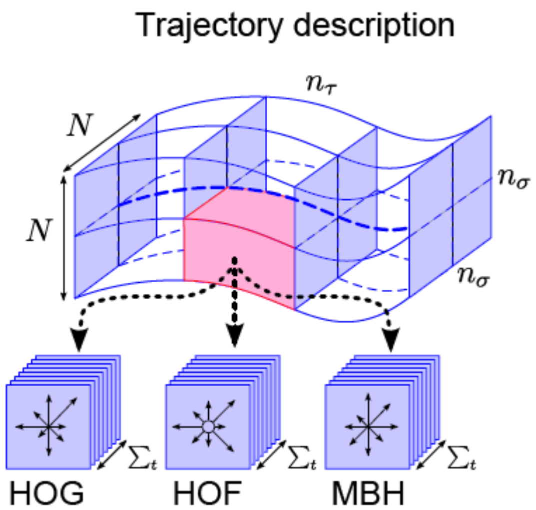
# Descriptors for dense trajectory

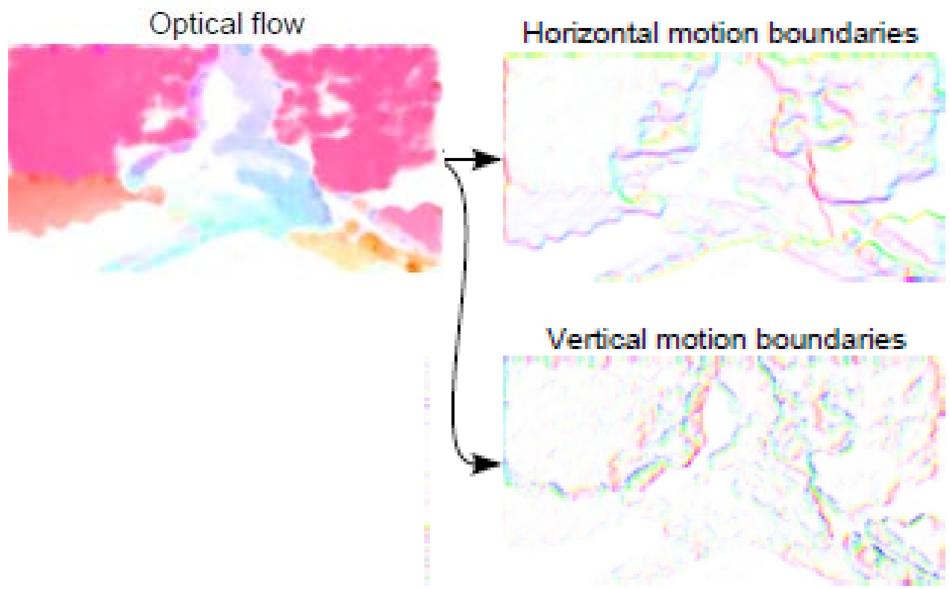
- Histogram of gradients (HOG: 2x2x3x8)
- Histogram of optical flow (HOF: 2x2x3x9)



# Descriptors for dense trajectory

- Motion-boundary histogram (MBHx + MBHy: 2x2x3x8)
  - spatial derivatives are calculated separately for optical flow in x and y, quantized into a histogram
  - captures relative dynamics of different regions
  - suppresses constant motions



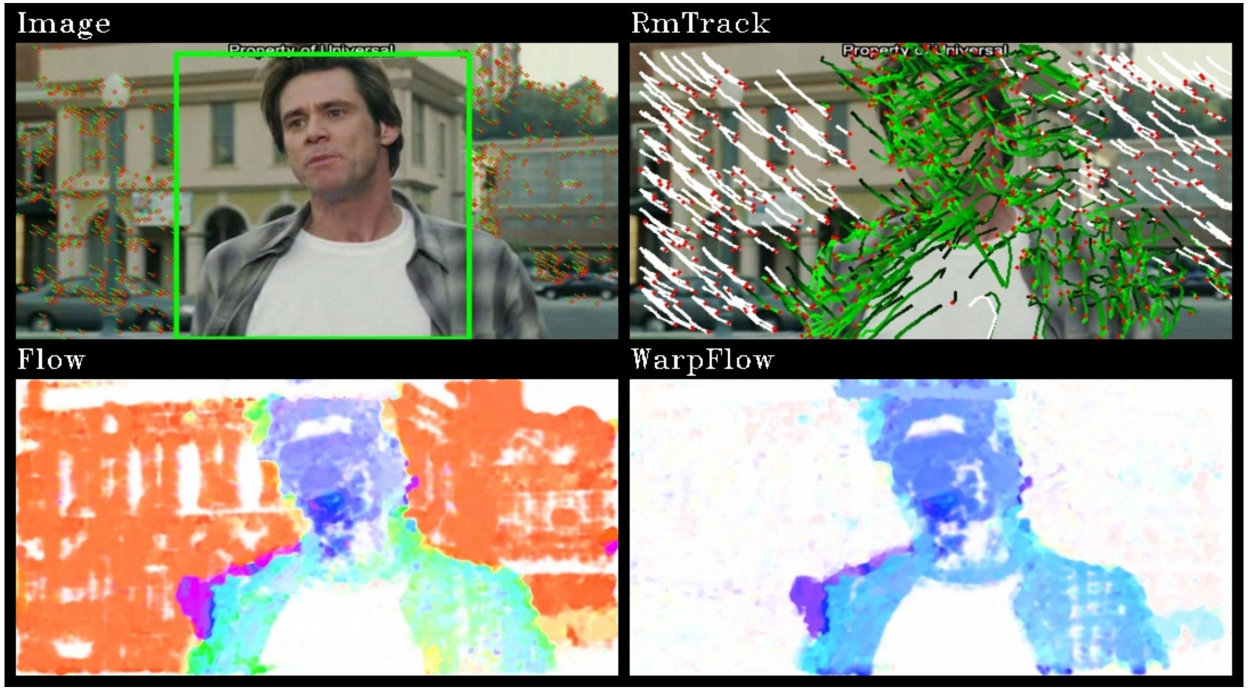
# Dense trajectories

- Advantages:
  - Captures the intrinsic dynamic structures in videos
  - MBH is robust to certain camera motion

- Disadvantages:
  - Generates irrelevant trajectories in background due to camera motion
  - Motion descriptors are modified by camera motion, e.g., HOF, MBH

# Improved dense trajectories

- Improve dense trajectories by explicit camera motion estimation
- Detect humans to remove outlier matches for homography estimation
- Stabilize optical flow to eliminate camera motion




[Wang and Schmid. Action recognition with improved trajectories. ICCV'13]

# Camera motion estimation

- Find the correspondences between two consecutive frames:
  - Extract and match SURF features (robust to motion blur)
  - Use optical flow, remove uninformative points
- Combine SURF (green) and optical flow (red) results in a more balanced distribution
- Use RANSAC to estimate a homography from all feature matches


Inlier matches of the homography

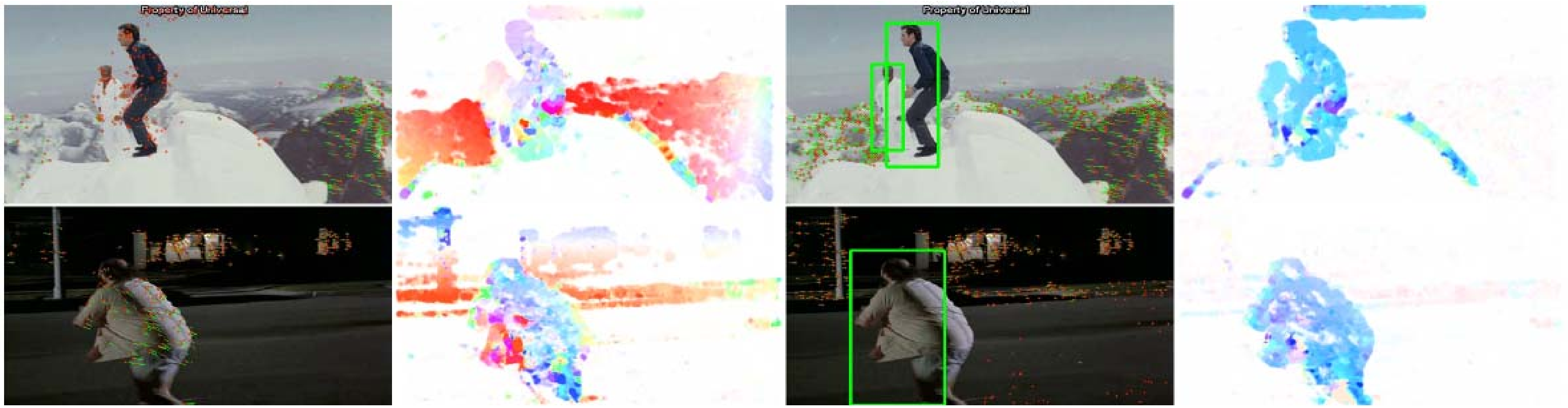
informatics mathematics
Inria

# Remove inconsistent matches due to humans

- Human motion is not constrained by camera motion, thus generates outlier matches

- Apply a human detector in each frame, and track the human bounding box forward and backward to join detections

- Remove feature matches inside the human bounding box during homography estimation



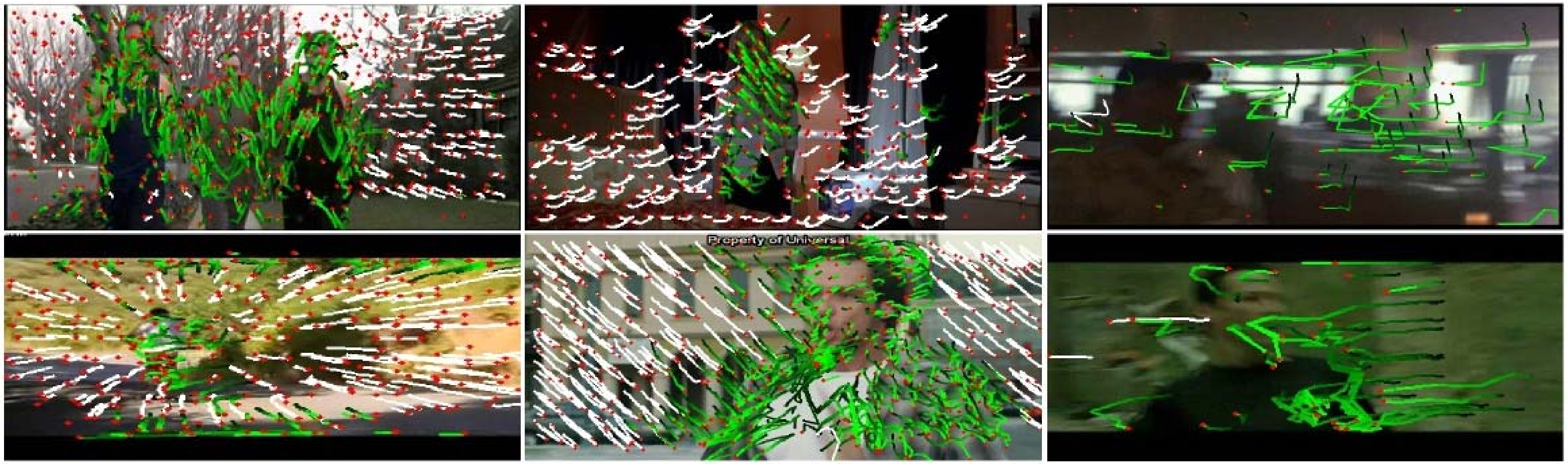

Inlier matches and warped flow, without or with HD

# Remove background trajectories

- Remove trajectories by thresholding the maximal magnitude of stabilized motion vectors

- Our method works well under various camera motions, such as pan, zoom, tilt

Successful examples	Failure cases




Removed trajectories (white) and foreground ones (green)

- Failure due to severe motion blur; the homography is not correctly estimated due to unreliable feature matches

## Experimental setting

- Motion stabilized trajectories and features (HOG, HOF, MBH)
- Normalization for each descriptor, then PCA to reduce its dimension by a factor of two
- Use Fisher vector to encode each descriptor separately, set the number of Gaussians to K=256
- Use Power+L2 normalization for FV, and linear SVM with one-against-rest for multi-class classification

## Datasets

- Hollywood2: 12 classes from 69 movies, report mAP
- HMDB51: 51 classes, report accuracy on three splits
- UCF101: 101 classes, report accuracy on three splits

# Evaluation of the intermediate steps

| | HOG | HOF | MBH | HOF+MBH | Combined |
|---|---|---|---|---|---|
| DTF | 38.4% | 39.5% | 49.1% | 49.8% | 52.2% |
| ITF | 40.2% | 48.9% | 52.1% | 54.7% | 57.2% |

Results on HMDB51 using Fisher vector

- Baseline: DTF = "dense trajectory feature"
- ITF = "improved trajectory feature”

- HOF improves significantly and MBH somewhat
- Almost no impact on HOG

- HOF and MBH are complementary, as they represent zero and first order motion information

informatics mathematics

# Impact of feature encoding on improved trajectories

| Datasets | Fisher vector | | |
| --- | --- | --- | --- |
| | DTF | ITF wo human | ITF w human |
| Hollywood2 | 63.6% | 66.1% | 66.8% |
| HMDB51 | 55.9% | 59.3% | 60.1% |
| UCF101 | 83.5% | 85.7% | 86.0% |

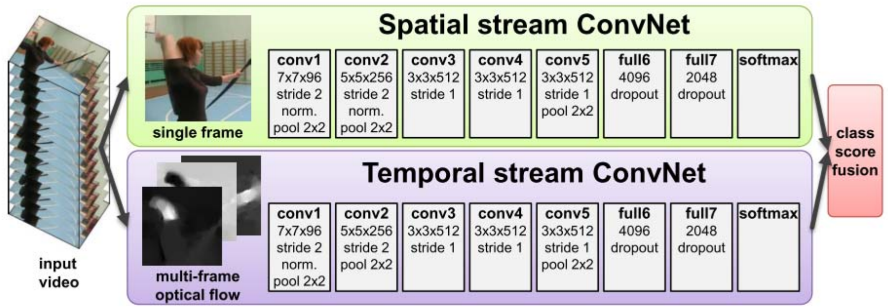
Compare DTF and ITF with and without human detection using HOG+HOF+MBH and Fisher encoding

- IDT significantly improvement over DT
- Human detection always helps. For Hollywood2 and HMDB51, the difference is more significant, as there are more humans present.
- Source code: http://lear.inrialpes.fr/~wang/improved_trajectories

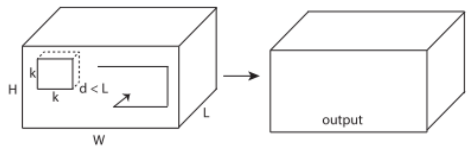
informatics mathematics

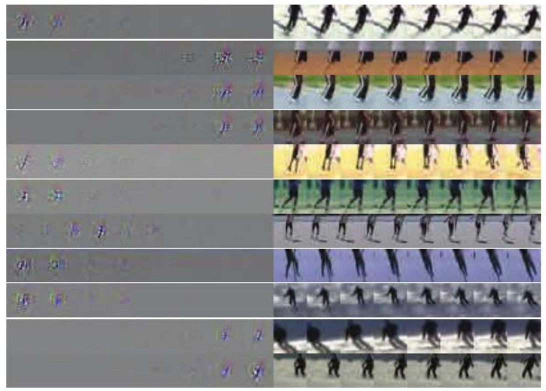
# Recent CNN methods

Two-Stream Convolutional Networks
for Action Recognition in Videos
[Simonyan and Zisserman NIPS14]




Learning Spatiotemporal Features with
3D Convolutional Networks
[Tran et al. ICCV15]



Quo vadis action recognition? A new
model and the Kinetics dataset
[Carreira et al. CVPR17]