LEAR @ TrecVid MED 2012

Dan Oneaţă[1], Matthijs Douze[1], Jérôme Revaud[1],
Jochen Schwenninger[2], Heng Wang[1], Danila Potapov[1],
Zaïd Harchaoui[1], Jakob Verbeek[1], Cordelia Schmid[1]

[1]LEAR team, INRIA Grenoble, France
[2]Fraunhofer Sankt Augustin, Germany

Outline

1. Low-level features: appearance, motion, audio
2. Feature encoding: Fisher vectors
3. High-level features: text
4. Fusion strategies
5. Experiments and results

Outline

1. Low-level features: appearance, motion, audio
2. Feature encoding: Fisher vectors
3. High-level features: text
4. Fusion strategies
5. Experiments and results

Appearance and audio features

Scale-invariant feature transform (SIFT, Lowe 2004):

- 21×21 patches at 4 pixel steps on 5 scales
- Every 60-th frame.

Appearance and audio features

Scale-invariant feature transform (SIFT, Lowe 2004):

- 21×21 patches at 4 pixel steps on 5 scales
- Every 60-th frame.

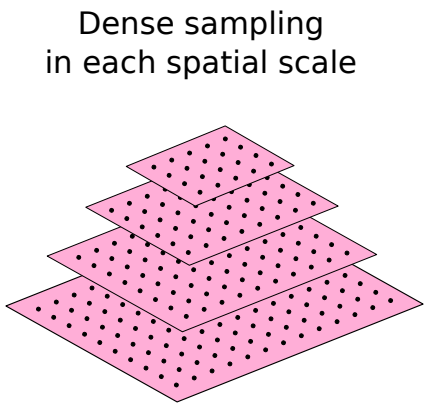
Mel-frequency cepstral coefficients (MFCC, Rabiner and Schafer 2007).

- Window of 25 ms and a step-size of 10 ms
- 39 coefficients: 12 MFCC and energy of the signal, first and second derivative
- Optionally: Speech/non-speech separation.

Motion features

Dense trajectories (Wang et al., 2011)

- Strong performance on many action recognition datasets: Hollywood2, Youtube, UCF Sports.
- Idea: MBH descriptors computed across short densely sampled trajectories.

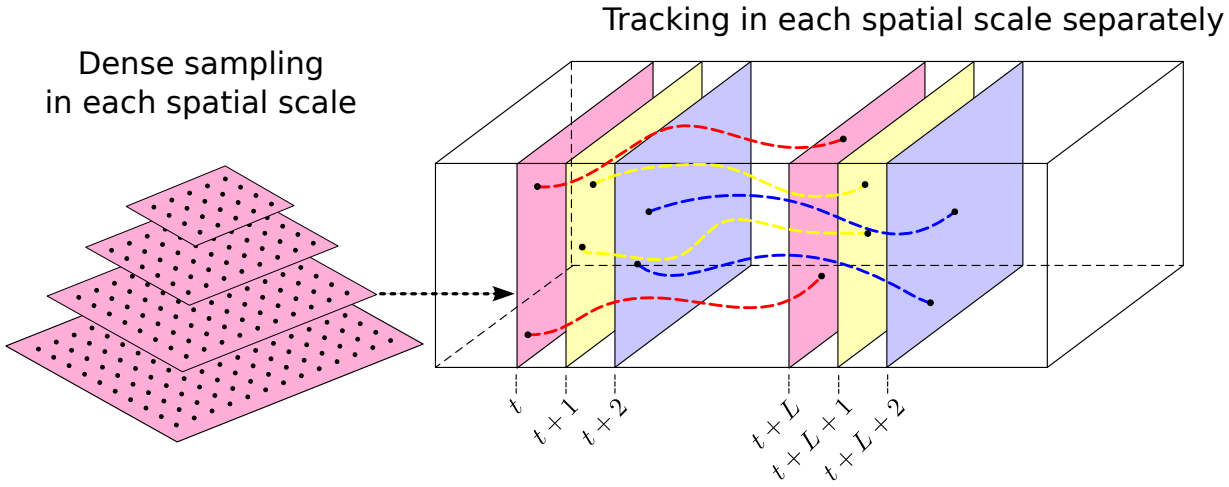


Wang, H., Kläser, A., Schmid, C., and Cheng-Lin, L. (2011). Action recognition by dense trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States

Motion features

Dense trajectories (Wang et al., 2011)

- Strong performance on many action recognition datasets: Hollywood2, Youtube, UCF Sports.
- Idea: MBH descriptors computed across short densely sampled trajectories.

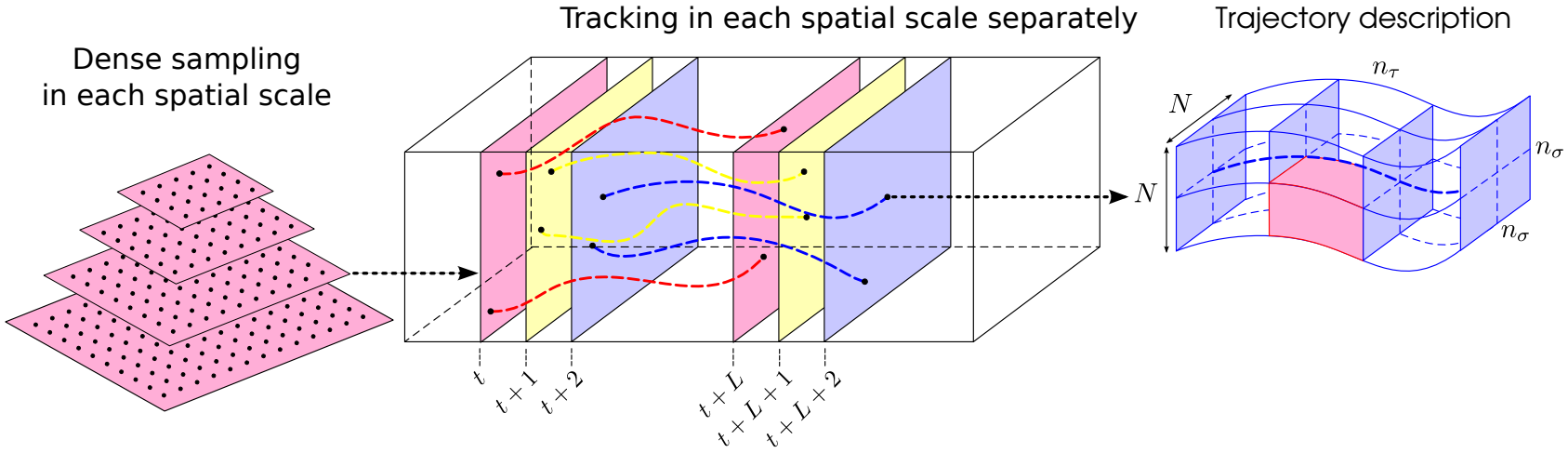


Wang, H., Kläser, A., Schmid, C., and Cheng-Lin, L. (2011). Action recognition by dense trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States

Motion features

Dense trajectories (Wang et al., 2011)

- Strong performance on many action recognition datasets: Hollywood2, Youtube, UCF Sports.
- Idea: MBH descriptors computed across short densely sampled trajectories.

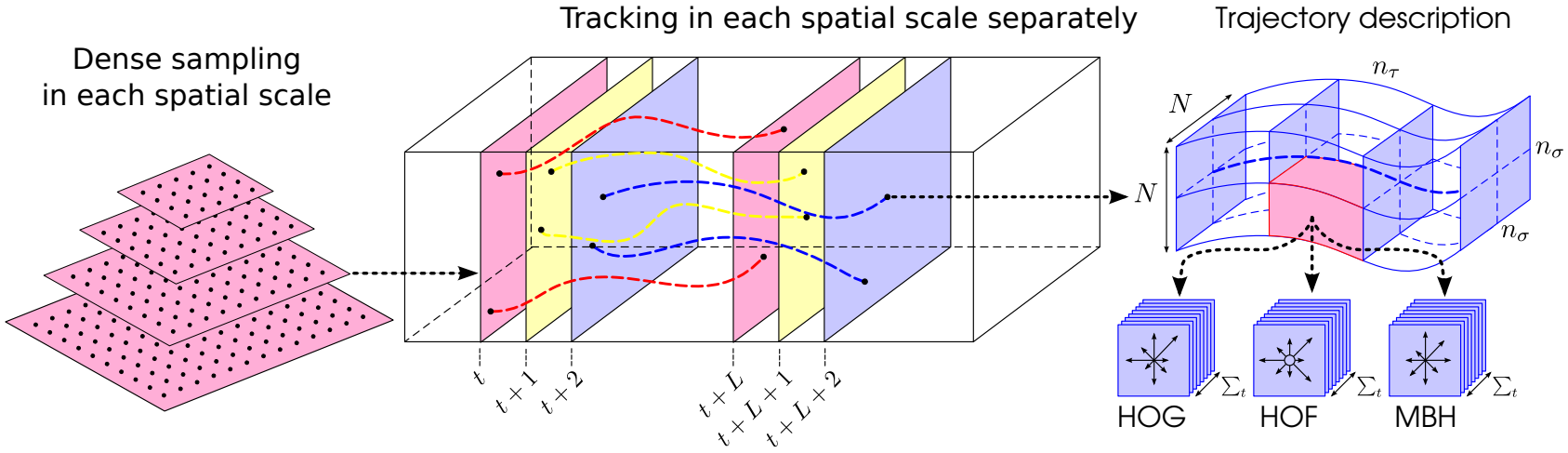


Wang, H., Kläser, A., Schmid, C., and Cheng-Lin, L. (2011). Action recognition by dense trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States

Motion features

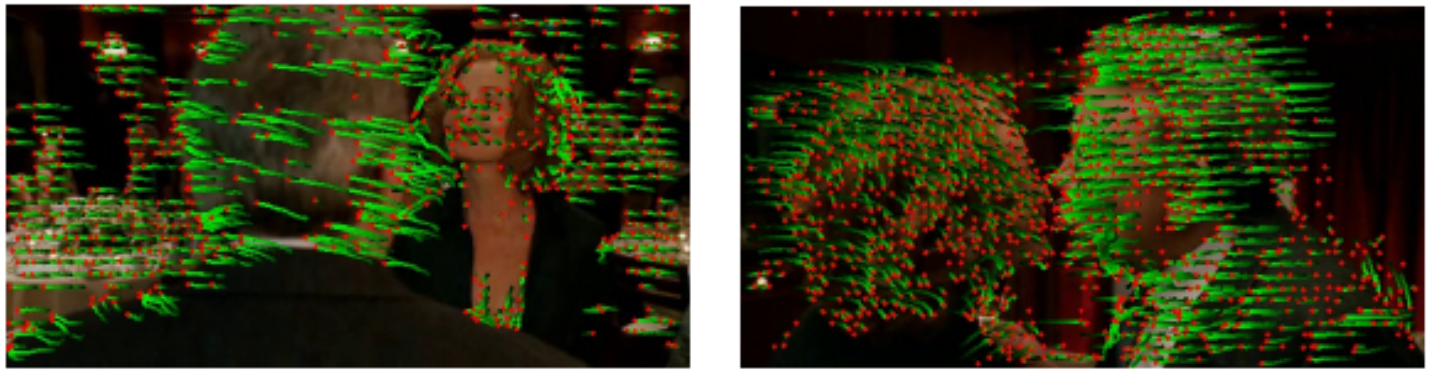
Dense trajectories (Wang et al., 2011)

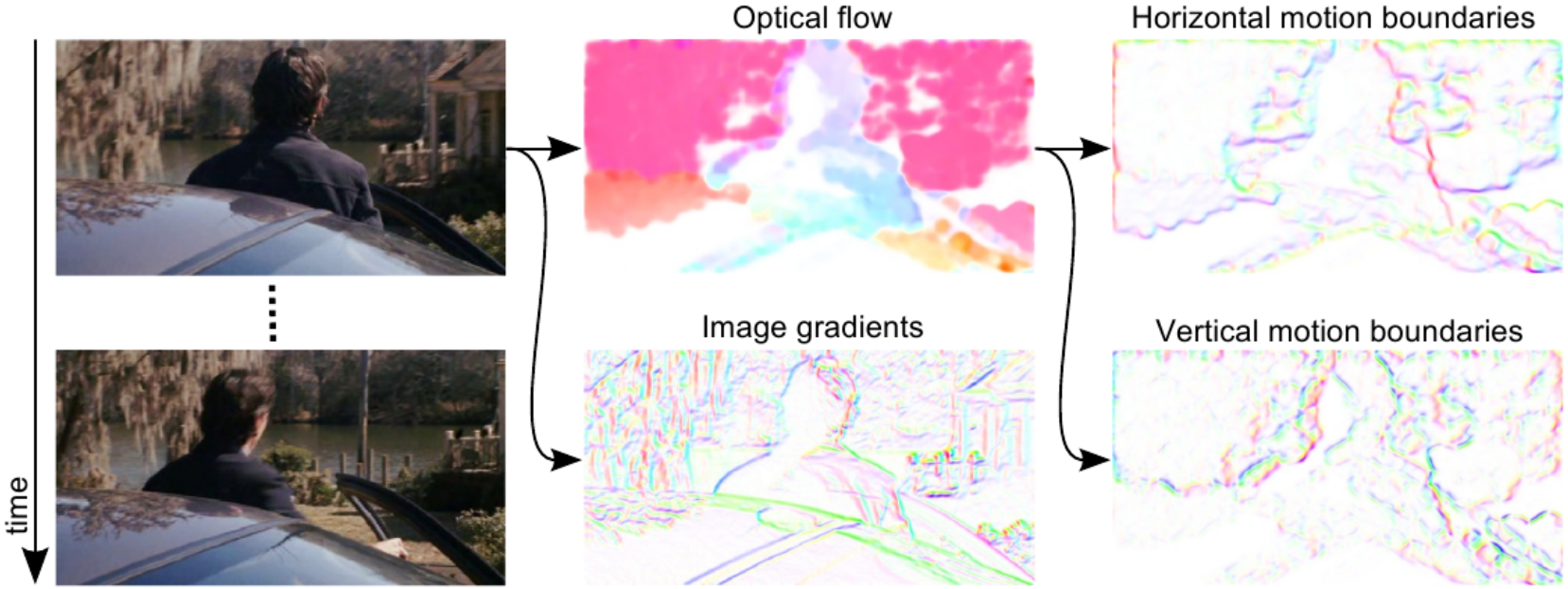
- Strong performance on many action recognition datasets: Hollywood2, Youtube, UCF Sports.
- Idea: MBH descriptors computed across short densely sampled trajectories.



Wang, H., Kläser, A., Schmid, C., and Cheng-Lin, L. (2011). Action recognition by dense trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States

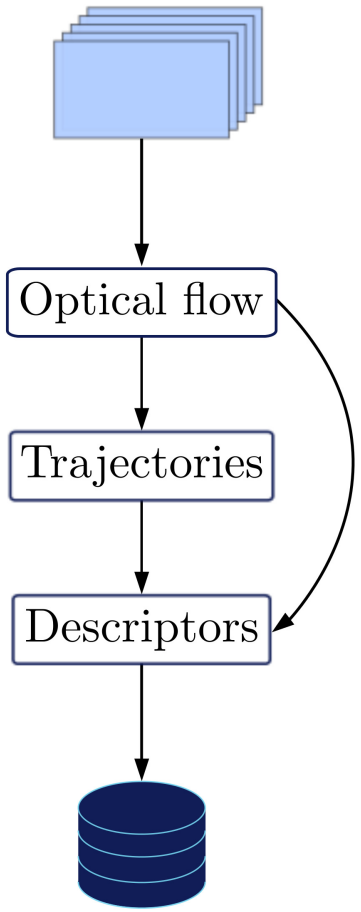
Motion features





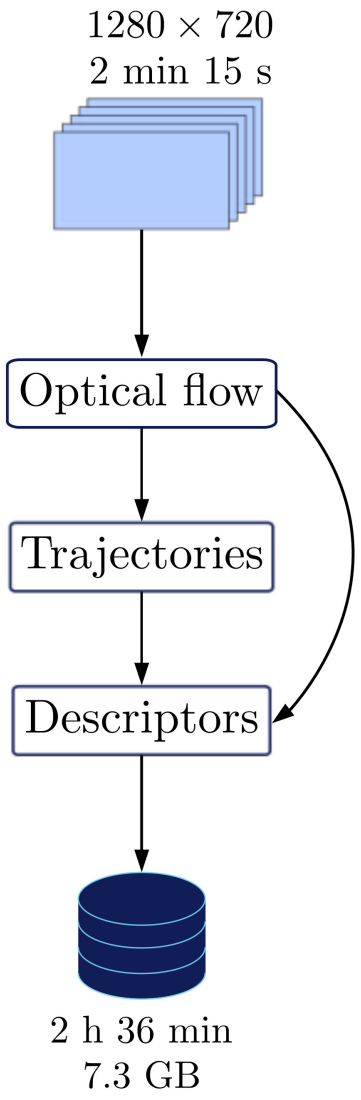
Video rescaling for dense trajectories

- Computationally expensive: cost scales linearly with the size of the video (time $\times$ resolution)



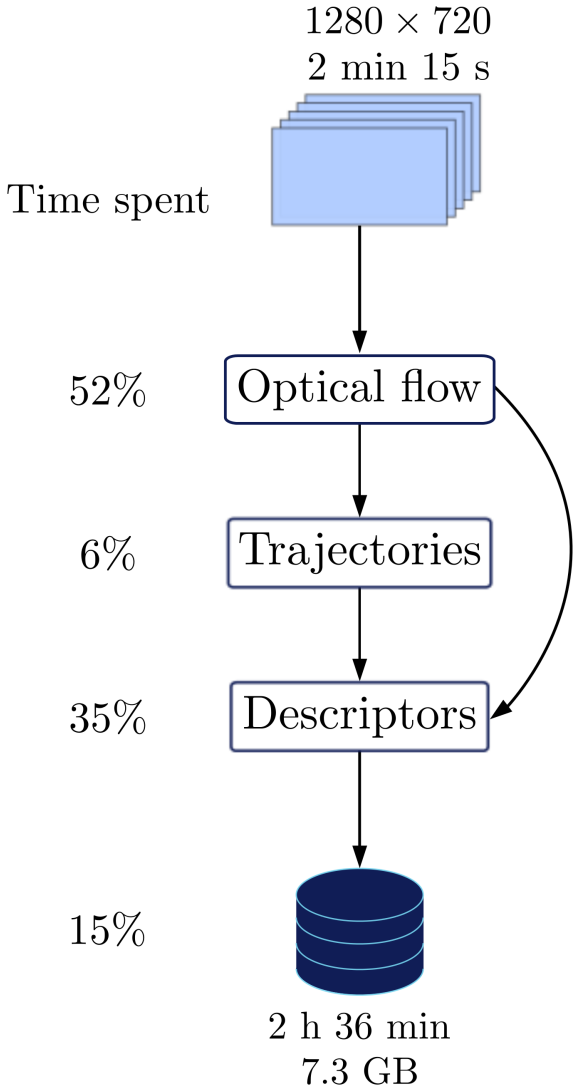
Video rescaling for dense trajectories

- Computationally expensive: cost scales linearly with the size of the video (time × resolution)



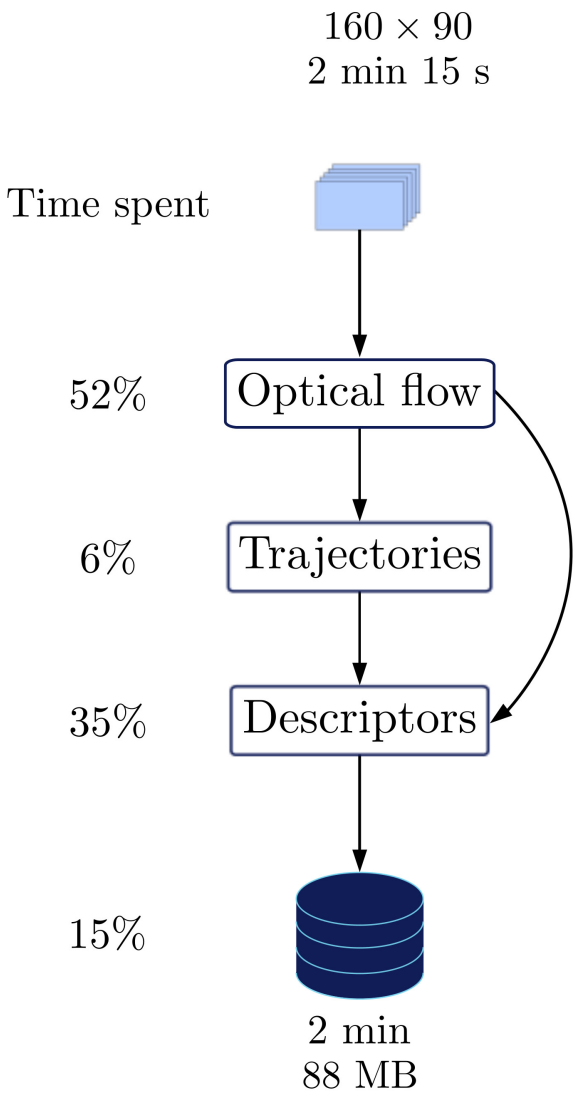
Video rescaling for dense trajectories

- Computationally expensive: cost scales linearly with the size of the video (time × resolution)



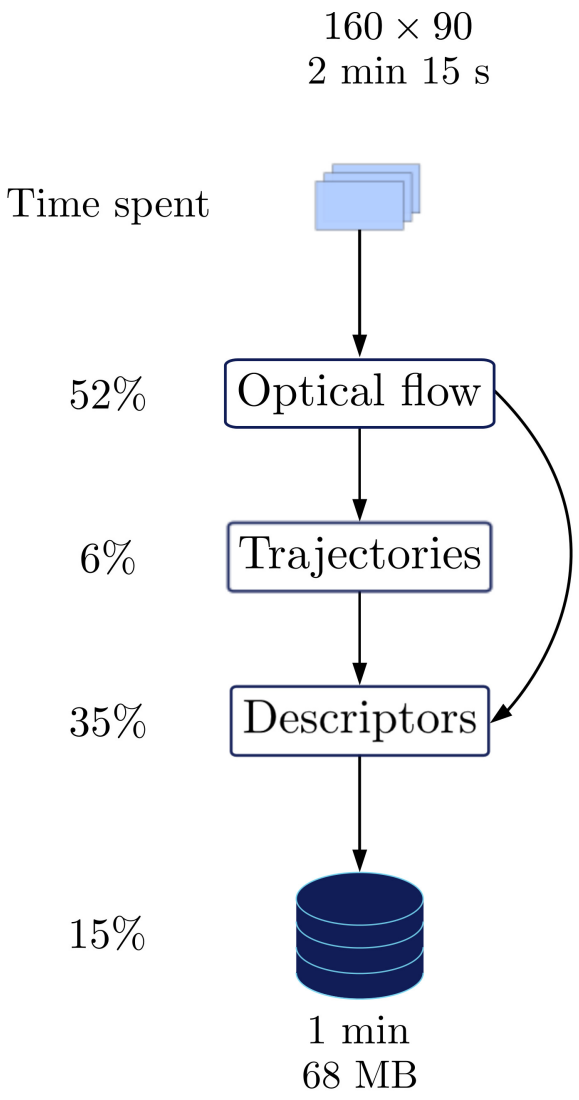
Video rescaling for dense trajectories

- Computationally expensive: cost scales linearly with the size of the video (time × resolution)
- Speed-ups:
  - Rescale videos: width at most 200 px.



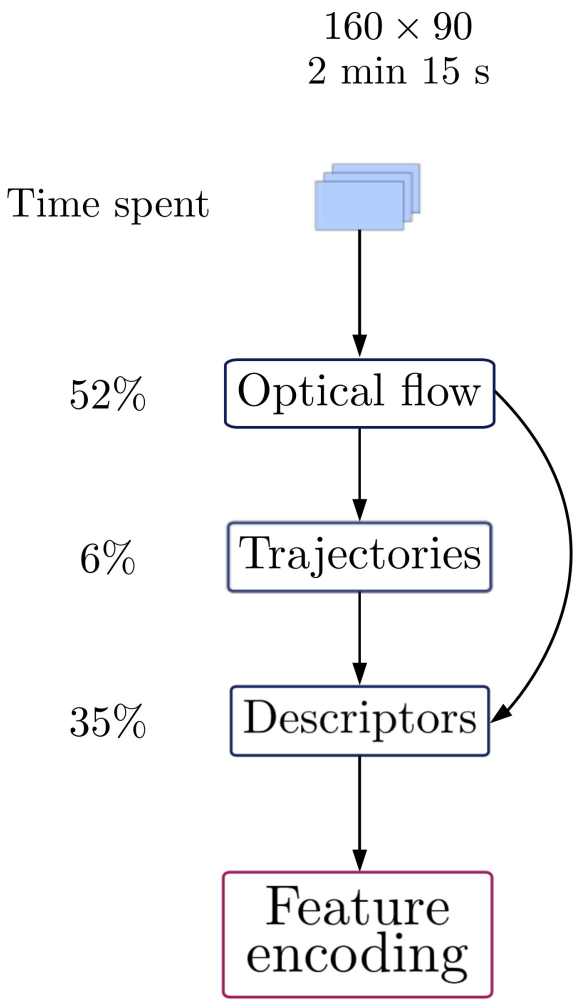
Video rescaling for dense trajectories

- Computationally expensive: cost scales linearly with the size of the video (time × resolution)
- Speed-ups:
  - Rescale videos: width at most 200 px.
  - Skip every second frame



Video rescaling for dense trajectories

- Computationally expensive: cost scales linearly with the size of the video (time × resolution)
- Speed-ups:
  - Rescale videos: width at most 200 px.
  - Skip every second frame
  - Process descriptors on-the-fly.



Outline

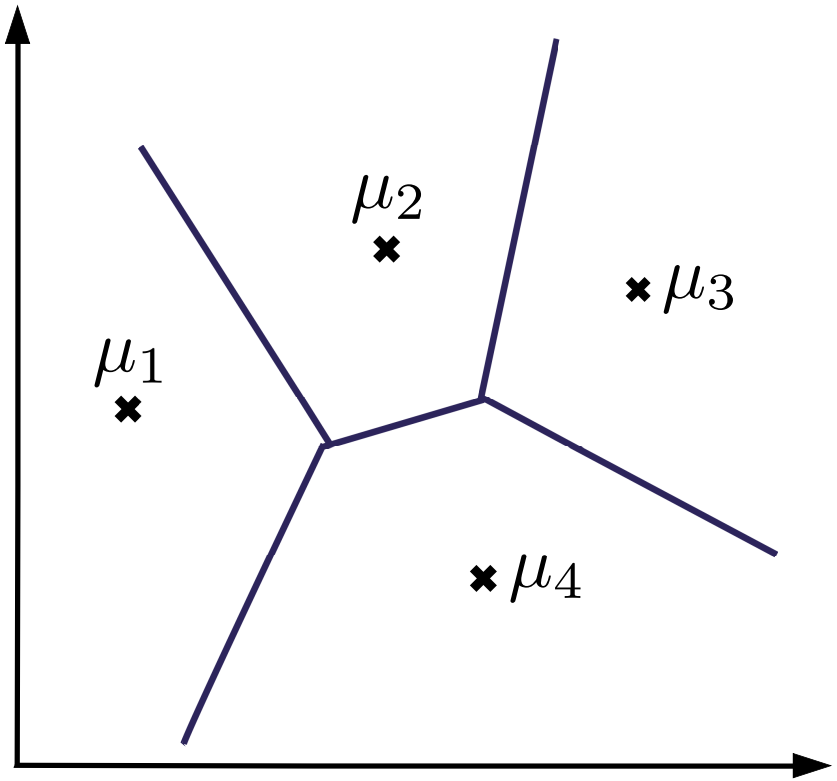
1. Low-level features: appearance, motion, audio
2. Feature encoding: Fisher vectors
3. High-level features: text
4. Fusion strategies
5. Experiments and results

Feature encoding: Fisher vectors (Perronnin et al., 2010)

- Top feature encoding technique for:
  - object recognition (Chatfield et al., 2011)
  - action recognition (Wang et al., 2012).

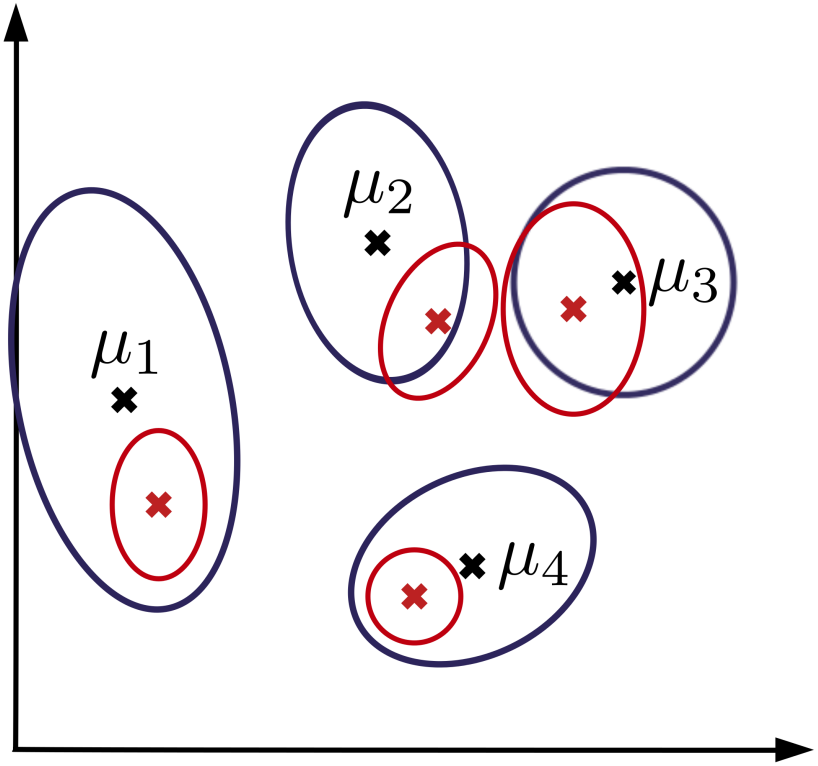
Feature encoding: Fisher vectors (Perronnin et al., 2010)

- Top feature encoding technique for:
  - object recognition (Chatfield et al., 2011)
  - action recognition (Wang et al., 2012).
- Fisher vectors (FV) for GMM:



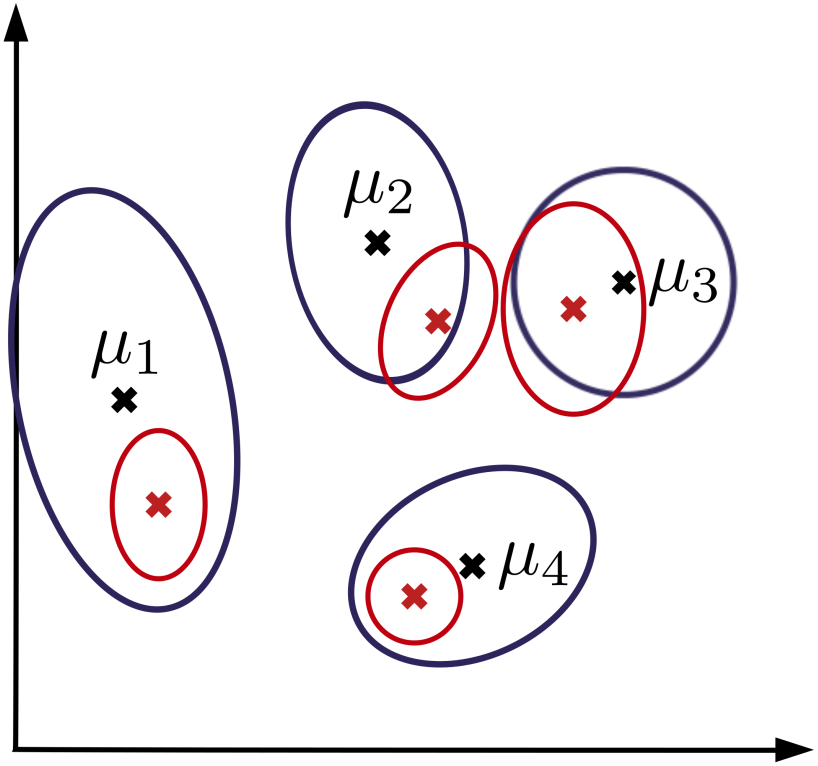
Feature encoding: Fisher vectors (Perronnin et al., 2010)

- Top feature encoding technique for:
  - object recognition (Chatfield et al., 2011)
  - action recognition (Wang et al., 2012).
- Fisher vectors (FV) for GMM:
  - soft bag-of-words: $\sum_x p(k|\mathbf{x})$
  - first moment: $\sum_x p(k|\mathbf{x})(\mathbf{x} - \mu_k)$
  - second moment: $\sum_x p(k|\mathbf{x})(\mathbf{x} - \mu_k)^2$.



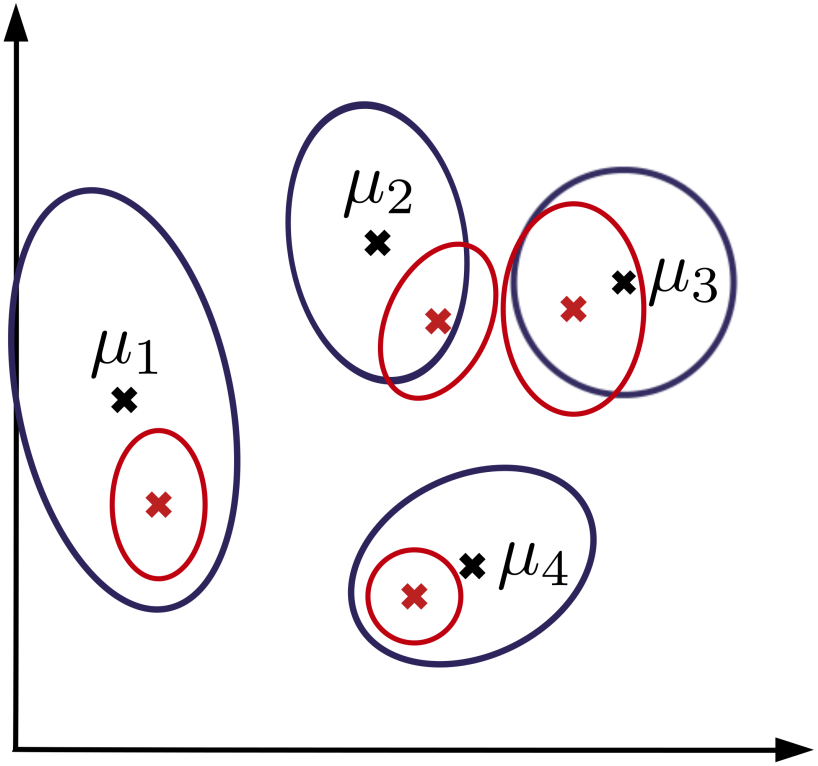
Feature encoding: Fisher vectors (Perronnin et al., 2010)

- Top feature encoding technique for:
  - object recognition (Chatfield et al., 2011)
  - action recognition (Wang et al., 2012).
- Fisher vectors (FV) for GMM:
  - soft bag-of-words: $\sum_x p(k|\mathbf{x})$
  - first moment: $\sum_x p(k|\mathbf{x})(\mathbf{x} - \mu_k)$
  - second moment: $\sum_x p(k|\mathbf{x})(\mathbf{x} - \mu_k)^2$.
- FV size: $K + 2KD$
  - K: number of Gaussians
  - D: descriptor dimension.



Feature encoding: Fisher vectors (Perronnin et al., 2010)

- Top feature encoding technique for:
  - object recognition (Chatfield et al., 2011)
  - action recognition (Wang et al., 2012).
- Fisher vectors (FV) for GMM:
  - soft bag-of-words: $\sum_x p(k|\mathbf{x})$
  - first moment: $\sum_x p(k|\mathbf{x})(\mathbf{x} - \mu_k)$
  - second moment: $\sum_x p(k|\mathbf{x})(\mathbf{x} - \mu_k)^2$.
- FV size: $K + 2KD$
  - K: number of Gaussians
  - D: descriptor dimension.
- Normalization:
  - zero mean, unit variance
  - signed square-rooting
  - ℓ_2 normalization.



Outline

1. Low-level features: appearance, motion, audio
2. Feature encoding: Fisher vectors
3. High-level features: text
4. Fusion strategies
5. Experiments and results

High-level features. Optical character recognition

Feature extraction:

- Maximally stable extremal regions (MSER; Matas et al. 2004)

High-level features. Optical character recognition

Feature extraction:

- Maximally stable extremal regions (MSER; Matas et al. 2004)

High-level features. Optical character recognition

Feature extraction:

- Maximally stable extremal regions (MSER; Matas et al. 2004)
- Filtering based on boundary gradients and aspect ratio.

High-level features. Optical character recognition

Feature extraction:

- Maximally stable extremal regions (MSER; Matas et al. 2004)
- Filtering based on boundary gradients and aspect ratio.
- HOG descriptor (Dalal and Triggs, 2005).

High-level features. Optical character recognition

Feature extraction:

- Maximally stable extremal regions (MSER; Matas et al. 2004)
- Filtering based on boundary gradients and aspect ratio.
- HOG descriptor (Dalal and Triggs, 2005).

Recognition:

- RBF-kernel SVM trained on Windows fonts.

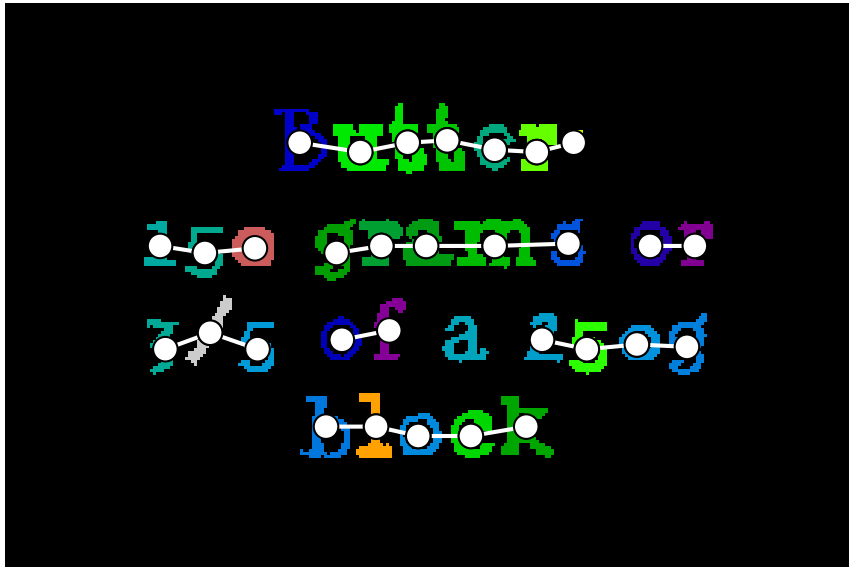
High-level features. Optical character recognition

Feature extraction:

- Maximally stable extremal regions (MSER; Matas et al. 2004)
- Filtering based on boundary gradients and aspect ratio.
- HOG descriptor (Dalal and Triggs, 2005).

Recognition:

- RBF-kernel SVM trained on Windows fonts.
- n-gram model over characters.



High-level features. Optical character recognition

Feature extraction:

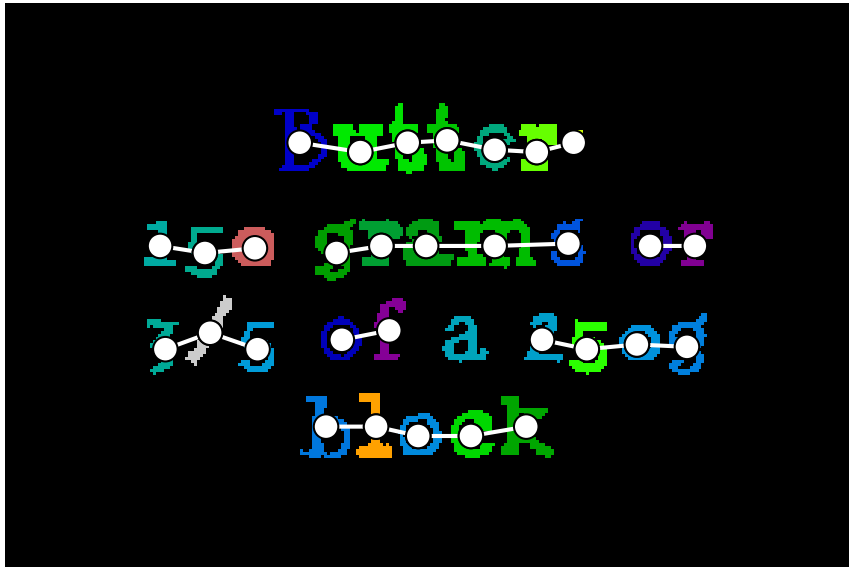
- Maximally stable extremal regions (MSER; Matas et al. 2004)
- Filtering based on boundary gradients and aspect ratio.
- HOG descriptor (Dalal and Triggs, 2005).

Recognition:

- RBF-kernel SVM trained on Windows fonts.
- n-gram model over characters.

Video representation:

- Bag-of-words.



Outline

1. Low-level features: appearance, motion, audio
2. Feature encoding: Fisher vectors
3. High-level features: text
4. Fusion strategies
5. Experiments and results

Fusion strategies

- Early fusion: concatenate feature vectors.
  - Sum of kernels.

Fusion strategies

- Early fusion: concatenate feature vectors.
  - Sum of kernels.
- Late fusion: linear combination of scores.
  - Learn classifiers for each channel
  - Learn weights using grid-search or logistic regression.

Outline

1. Low-level features: appearance, motion, audio
2. Feature encoding: Fisher vectors
3. High-level features: text
4. Fusion strategies
5. Experiments and results

MinNDC error on the TrecVid '11 data

- Best result in 2011 by BBN-Viser team (Natarajan et al., 2012).

Channel	birthday party	changing vehicle tire	flash mob gathering	unstuck vehicle	grooming an animal	making a sandwich	parade	parkour	repairing an appliance	sewing project	Average
Best 2011	0.45	0.47	0.28	0.38	0.62	0.57	0.45	0.31	0.38	0.57	0.45

MinNDC error on the TrecVid '11 data

- Best result in 2011 by BBN-Viser team (Natarajan et al., 2012).

Channel	birthday party	changing vehicle tire	flash mob gathering	unstuck vehicle	grooming an animal	making a sandwich	parade	parkour	repairing an appliance	sewing project	Average
Best 2011	0.45	0.47	0.28	0.38	0.62	0.57	0.45	0.31	0.38	0.57	0.45
MBH	0.77	0.79	0.34	0.59	0.75	0.77	0.52	0.25	0.53	0.65	0.60

MinNDC error on the TrecVid '11 data

- Best result in 2011 by BBN-Viser team (Natarajan et al., 2012).

Channel	birthday party	changing vehicle tire	flash mob gathering	unstuck vehicle	grooming an animal	making a sandwich	parade	parkour	repairing an appliance	sewing project	Average
Best 2011	0.45	0.47	0.28	0.38	0.62	0.57	0.45	0.31	0.38	0.57	0.45
MBH	0.77	0.79	0.34	0.59	0.75	0.77	0.52	0.25	0.53	0.65	0.60
SIFT	0.71	0.63	0.40	0.45	0.75	0.69	0.71	0.57	0.61	0.77	0.63

MinNDC error on the TrecVid '11 data

- Best result in 2011 by BBN-Viser team (Natarajan et al., 2012).

Channel	birthday party	changing vehicle tire	flash mob gathering	unstuck vehicle	grooming an animal	making a sandwich	parade	parkour	repairing an appliance	sewing project	Average
Best 2011	0.45	0.47	0.28	0.38	0.62	0.57	0.45	0.31	0.38	0.57	0.45
MBH	0.77	0.79	0.34	0.59	0.75	0.77	0.52	0.25	0.53	0.65	0.60
SIFT	0.71	0.63	0.40	0.45	0.75	0.69	0.71	0.57	0.61	0.77	0.63
MFCC	0.65	0.93	0.70	0.77	0.96	0.94	0.80	0.94	0.55	0.82	0.81

MinNDC error on the TrecVid '11 data

- Best result in 2011 by BBN-Viser team (Natarajan et al., 2012).

Channel	birthday party	changing vehicle tire	flash mob gathering	unstuck vehicle	grooming an animal	making a sandwich	parade	parkour	repairing an appliance	sewing project	Average
Best 2011	0.45	0.47	0.28	0.38	0.62	0.57	0.45	0.31	0.38	0.57	0.45
MBH	0.77	0.79	0.34	0.59	0.75	0.77	0.52	0.25	0.53	0.65	0.60
SIFT	0.71	0.63	0.40	0.45	0.75	0.69	0.71	0.57	0.61	0.77	0.63
MFCC	0.65	0.93	0.70	0.77	0.96	0.94	0.80	0.94	0.55	0.82	0.81
OCR	0.95	0.94	0.91	0.99	0.93	0.85	0.95	1.00	0.68	0.88	0.91

MinNDC error on the TrecVid '11 data

- Best result in 2011 by BBN-Viser team (Natarajan et al., 2012).

Channel	birthday party	changing vehicle tire	flash mob gathering	unstuck vehicle	grooming an animal	making a sandwich	parade	parkour	repairing an appliance	sewing project	Average
Best 2011	0.45	0.47	0.28	0.38	0.62	0.57	0.45	0.31	0.38	0.57	0.45
MBH	0.77	0.79	0.34	0.59	0.75	0.77	0.52	0.25	0.53	0.65	0.60
SIFT	0.71	0.63	0.40	0.45	0.75	0.69	0.71	0.57	0.61	0.77	0.63
MFCC	0.65	0.93	0.70	0.77	0.96	0.94	0.80	0.94	0.55	0.82	0.81
OCR	0.95	0.94	0.91	0.99	0.93	0.85	0.95	1.00	0.68	0.88	0.91
MBH+SIFT	0.62	0.54	0.26	0.37	0.67	0.62	0.44	0.22	0.46	0.60	0.48

MinNDC error on the TrecVid '11 data

- Best result in 2011 by BBN-Viser team (Natarajan et al., 2012).

Channel	birthday party	changing vehicle tire	flash mob gathering	unstuck vehicle	grooming an animal	making a sandwich	parade	parkour	repairing an appliance	sewing project	Average
Best 2011	0.45	0.47	0.28	0.38	0.62	0.57	0.45	0.31	0.38	0.57	0.45
MBH	0.77	0.79	0.34	0.59	0.75	0.77	0.52	0.25	0.53	0.65	0.60
SIFT	0.71	0.63	0.40	0.45	0.75	0.69	0.71	0.57	0.61	0.77	0.63
MFCC	0.65	0.93	0.70	0.77	0.96	0.94	0.80	0.94	0.55	0.82	0.81
OCR	0.95	0.94	0.91	0.99	0.93	0.85	0.95	1.00	0.68	0.88	0.91
MBH+SIFT	0.62	0.54	0.26	0.37	0.67	0.62	0.44	0.22	0.46	0.60	0.48
$\cdots$+MFCC	0.49	0.48	0.26	0.38	0.59	0.65	0.41	0.21	0.35	0.52	0.43

MinNDC error on the TrecVid '11 data

- Best result in 2011 by BBN-Viser team (Natarajan et al., 2012).

Channel	birthday party	changing vehicle tire	flash mob gathering	unstuck vehicle	grooming an animal	making a sandwich	parade	parkour	repairing an appliance	sewing project	Average
Best 2011	0.45	0.47	0.28	0.38	0.62	0.57	0.45	0.31	0.38	0.57	0.45
MBH	0.77	0.79	0.34	0.59	0.75	0.77	0.52	0.25	0.53	0.65	0.60
SIFT	0.71	0.63	0.40	0.45	0.75	0.69	0.71	0.57	0.61	0.77	0.63
MFCC	0.65	0.93	0.70	0.77	0.96	0.94	0.80	0.94	0.55	0.82	0.81
OCR	0.95	0.94	0.91	0.99	0.93	0.85	0.95	1.00	0.68	0.88	0.91
MBH+SIFT	0.62	0.54	0.26	0.37	0.67	0.62	0.44	0.22	0.46	0.60	0.48
$\cdots$+MFCC	0.49	0.48	0.26	0.38	0.59	0.65	0.41	0.21	0.35	0.52	0.43
$\cdots$+OCR	0.46	0.45	0.26	0.38	0.54	0.55	0.39	0.23	0.34	0.51	0.41

Our submissions

	Actual NDC			
Run	Features	Late fusion	PreSpec	AdHoc
`c-LFdnsmall`	small	grid search	0.544	0.711
`c-LFjrlrsmall`	small	logistic regression	0.536	0.749
`p-LFdnbig`	big	grid search	0.516	0.559
`c-LFjrlrbig`	big	logistic regression	**0.515**	**0.536**

Our submissions

	Actual NDC			
Run	Features	Late fusion	PreSpec	AdHoc
`c-LFdnsmall`	small	grid search	0.544	0.711
`c-LFjrlrsmall`	small	logistic regression	0.536	0.749
`p-LFdnbig`	big	grid search	0.516	0.559
`c-LFjrlrbig`	big	logistic regression	**0.515**	**0.536**

Modality	Descriptor	small		big	
		dim	× RT	dim	× RT
Motion	MBH	33k	2.4	131k	3.0
Image	SIFT	16k	2.5	66k	6.6
Audio	MFCC	40k	0.2	81k	0.2
Text	OCR	200k	1.4	200k	1.4
Total		289k	6.5	478k	11.2

RT: single CPU computation × real time.

Our submissions

Actual NDC

Run	Features	Late fusion	PreSpec	AdHoc
`c-LFdnsmall`	small	grid search	0.544	0.711
`c-LFjrlrsmall`	small	logistic regression	0.536	0.749
`p-LFdnbig`	big	grid search	0.516	0.559
`c-LFjrlrbig`	big	logistic regression	**0.515**	**0.536**

Modality	Descriptor	small		big	
		dim	× RT	dim	× RT
Motion	MBH	33k	2.4	131k	3.0
Image	SIFT	16k	2.5	66k	6.6
Audio	MFCC	40k	0.2	81k	0.2
Text	OCR	200k	1.4	200k	1.4
Total		289k	6.5	478k	11.2

RT: single CPU computation × real time.

Computation on MED test set: $4,000$ h video $\times 11.2/400$ CPUs ≈ 4 days.

Conclusions

Excellent results while being compact:

- Small set of low-level features: MBH, SIFT, MFCC
- High-dimensional FV encoding
- One type of high level features: OCR
- Linear classifiers + late fusion.

Code for MBH, SIFT and Fisher vectors available at
`http://lear.inrialpes.fr/software/`

Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767.

Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., and Prasad, R. (2012). Multimodal feature fusion for robust event detection in web videos. In *CVPR*.

Perronnin, F., Sánchez, J., and Mensink, T. (2010). Improving the Fisher Kernel for Large-Scale Image Classification. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *European Conference on Computer Vision (ECCV '10)*, volume 6314 of *Lecture Notes in*

Computer Science (LNCS), pages 143–156, Heraklion, Greece. Springer-Verlag.

Rabiner, L. R. and Schafer, R. W. (2007). Introduction to digital speech processing. *Found. Trends Signal Process.*

Wang, H., Kläser, A., Schmid, C., and Cheng-Lin, L. (2011). Action recognition by dense trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States.

Wang, X., Wang, L., and Qiao, Y. (2012). A comparative study of encoding, pooling and normalization methods for action recognition. ACCV '12.