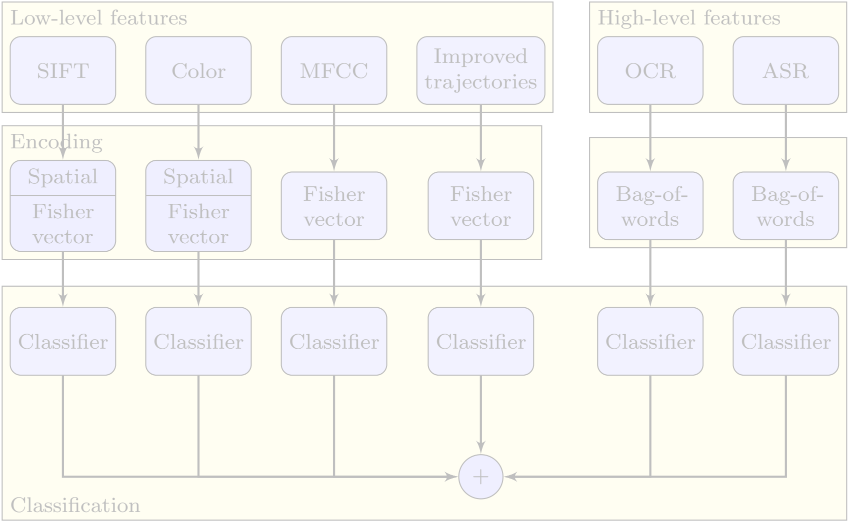
AXES @ TRECVid MED 2013

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

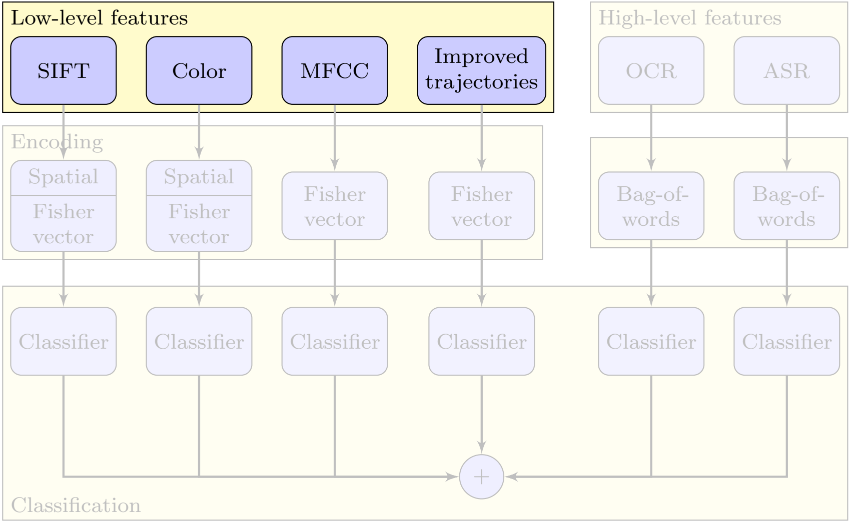
[1]INRIA–LEAR, Grenoble, France
[2]Fraunhofer Sankt Augustin, Germany

Outline

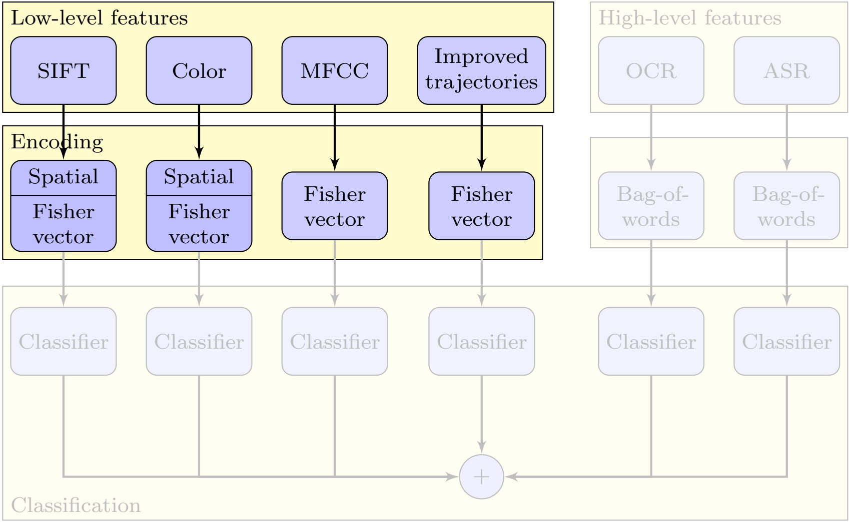


Low-level features
SIFT
Color
MFCC
Improved trajectories
High-level features
OCR
ASR
Encoding
Spatial
Fisher vector
Spatial
Fisher vector
Fisher vector
Fisher vector
Bag-of-words
Bag-of-words
Classifier
Classifier
Classifier
Classifier
Classifier
Classifier
+
Classification

Outline

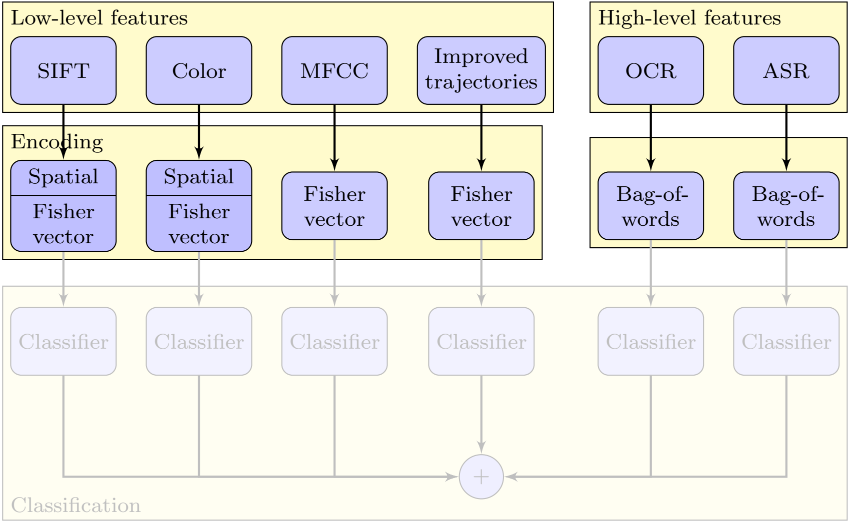

Low-level features
SIFT
Color
MFCC
Improved trajectories
High-level features
OCR
ASR
Encoding
Spatial
Fisher vector
Spatial
Fisher vector
Fisher vector
Fisher vector
Bag-of-words
Bag-of-words
Classifier
Classifier
Classifier
Classifier
Classifier
Classifier
+
Classification

Outline



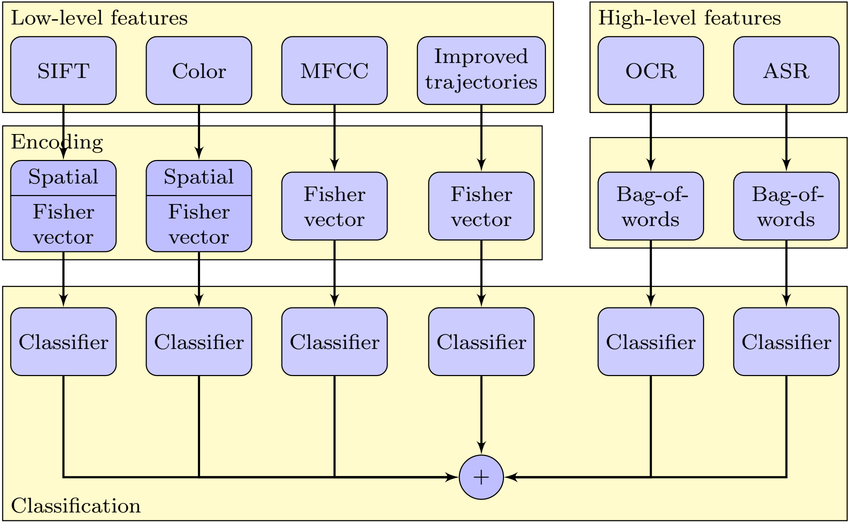
Low-level features
SIFT
Color
MFCC
Improved trajectories
High-level features
OCR
ASR
Encoding
Spatial
Fisher vector
Spatial
Fisher vector
Fisher vector
Fisher vector
Bag-of-words
Bag-of-words
Classifier
Classifier
Classifier
Classifier
Classifier
Classifier
+
Classification

Outline



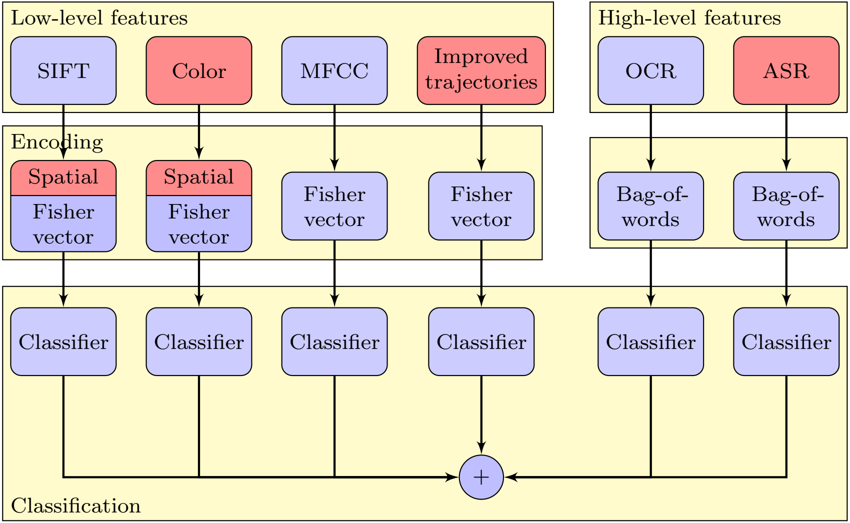
Low-level features
SIFT
Color
MFCC
Improved trajectories
High-level features
OCR
ASR
Encoding
Spatial
Fisher vector
Spatial
Fisher vector
Fisher vector
Fisher vector
Bag-of-words
Bag-of-words
Classifier
Classifier
Classifier
Classifier
Classifier
Classifier
+
Classification

Outline



Low-level features
SIFT
Color
MFCC
Improved trajectories
High-level features
OCR
ASR
Encoding
Spatial
Fisher vector
Spatial
Fisher vector
Fisher vector
Fisher vector
Bag-of-words
Bag-of-words
Classifier
Classifier
Classifier
Classifier
Classifier
Classifier
+
Classification

Outline



Low-level features
SIFT
Color
MFCC
Improved trajectories
High-level features
OCR
ASR
Encoding
Spatial
Fisher vector
Spatial
Fisher vector
Fisher vector
Fisher vector
Bag-of-words
Bag-of-words
Classifier
Classifier
Classifier
Classifier
Classifier
Classifier
+
Classification

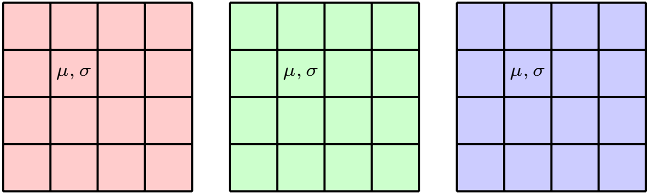
Table of Contents

Static and audio features

- Scale-invariant feature transform (SIFT, Lowe 2004)
- Mel-frequency cepstral coefficients (MFCC, Rabiner and Schafer 2007)

Static and audio features

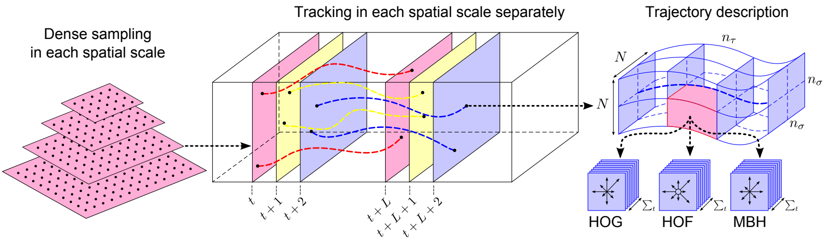
- Scale-invariant feature transform (SIFT, Lowe 2004)
- Mel-frequency cepstral coefficients (MFCC, Rabiner and Schafer 2007)
- Color descriptors (Clinchant et al., 2007).



Mean and variance...	2
...of RGB values...	3
...in 4×4 cells	16
Descriptor dimensionality	96

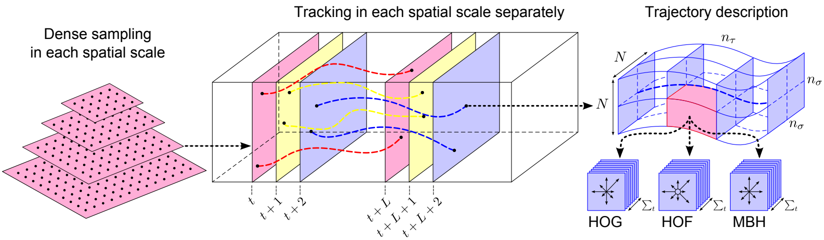
Improved motion features (Wang and Schmid, ICCV, 2013)

- Builds upon dense trajectory features (?, CVPR, ?)



Improved motion features (Wang and Schmid, ICCV, 2013)

- Builds upon dense trajectory features (?, CVPR, ?)
- Dense trajectories can be affected by camera motion.

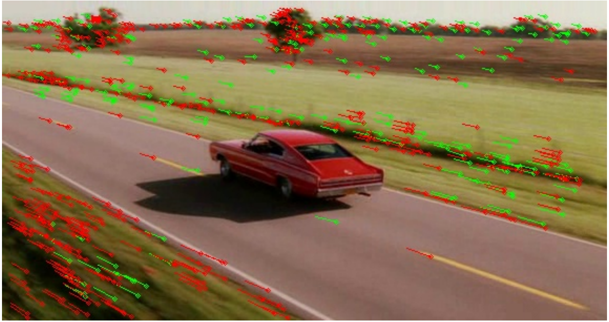


Improved motion features (Wang and Schmid, ICCV, 2013)

- Idea: stabilize camera motion before computing optical flow.

Improved motion features (Wang and Schmid, ICCV, 2013)

- Idea: stabilize camera motion before computing optical flow.
- Method:
  1. extract feature points (SURF descriptors and dense optical flow)
  2. match feature points and estimate homography with RANSAC
  3. warp the optical flow.



Improved motion features (Wang and Schmid, ICCV, 2013)

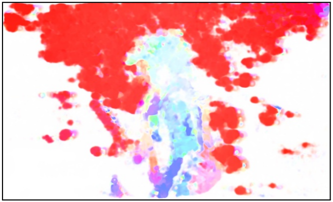
- Idea: stabilize camera motion before computing optical flow.

Two succesive frames

Improved motion features (Wang and Schmid, ICCV, 2013)

- Idea: stabilize camera motion before computing optical flow.

Two succesive frames

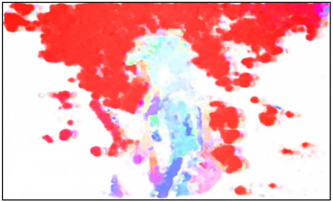


Optical flow

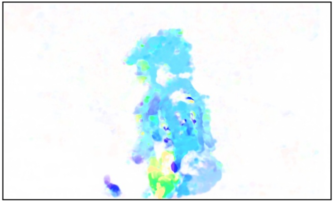
Improved motion features (Wang and Schmid, ICCV, 2013)

- Idea: stabilize camera motion before computing optical flow.
  - improves flow estimation

Two succesive frames



Optical flow

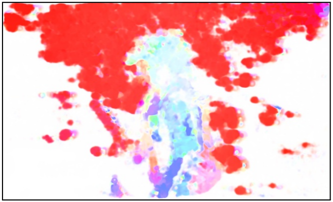


Warped optical flow

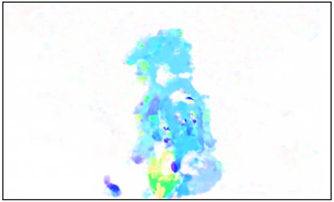
Improved motion features (Wang and Schmid, ICCV, 2013)

- Idea: stabilize camera motion before computing optical flow.
  - improves flow estimation
  - removes background tracks.

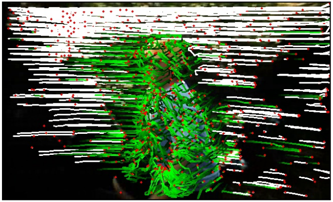
Two succesive frames



Optical flow



Warped optical flow



Removed trajectories

Removed trajectories under various camera motions

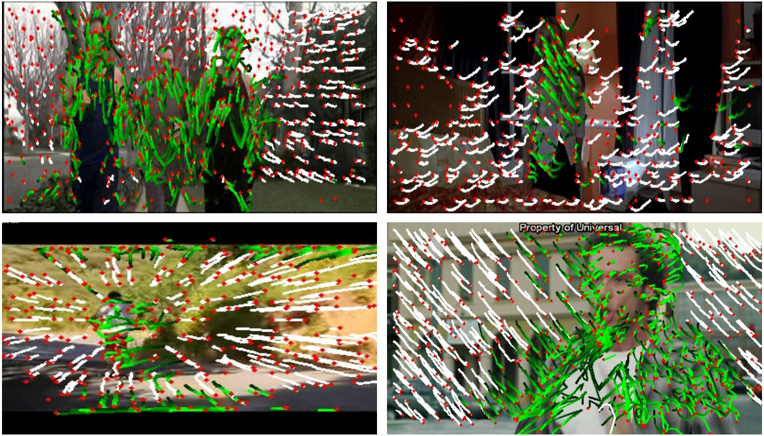


Table of Contents

Fisher vector for appearance

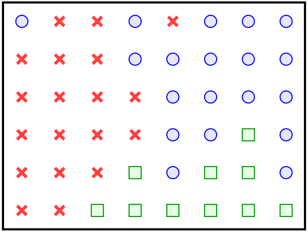
- Generalization of the bag-of-words.
- Strong performance across multiple tasks:
  - action recognition, action detection, event recognition
    (Oneaţă et al., ICCV, 2013)

Fisher vector for appearance

- Generalization of the bag-of-words.
- Strong performance across multiple tasks:
  - action recognition, action detection, event recognition (Oneaţă et al., ICCV, 2013)
  - image classification (Chatfield et al., BMVC, 2011)
  - image retrieval (Jégou et al., PAMI, 2012)
  - fine-grained image classification (Gavves et al., ICCV, 2013)
  - face verification (Simonyan et al., BMVC, 2013)
  - word spotting (Almazán et al., ICCV, 2013).

Fisher vector for location

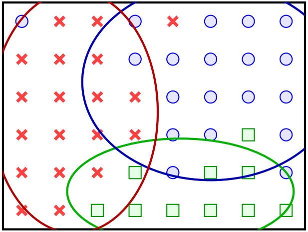
- Spatial Fisher vector (SFV)
  (Krapac et al., ICCV, 2011)
  - encodes first and second moments of visual word locations
  - adds 6 entries for each visual word: μ and σ for (x, y, t) coordinates.



Schematic illustration of the spatial Fisher vector for three types of visual words (○, ×, □) in an image.

Fisher vector for location

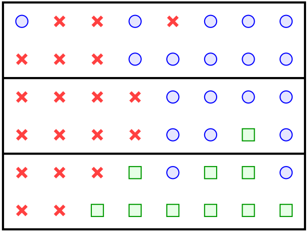
- Spatial Fisher vector (SFV)
  (Krapac et al., ICCV, 2011)
  - encodes first and second moments of visual word locations
  - adds 6 entries for each visual word: μ and σ for (x, y, t) coordinates.



Schematic illustration of the spatial Fisher vector for three types of visual words (○, ×, □) in an image.

Fisher vector for location

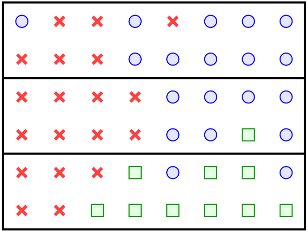
- Spatial Fisher vector (SFV)
  (Krapac et al., ICCV, 2011)
  - encodes first and second moments of visual word locations
  - adds 6 entries for each visual word: μ and σ for (x, y, t) coordinates.
- Compared to spatial pyramids:
  (Oneaţă et al., ICCV, 2013)
  - similar performance gain



Schematic illustration of the spatial Fisher vector for three types of visual words (○, ×, □) in an image.

Fisher vector for location

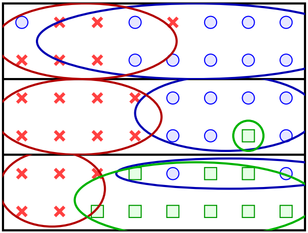
- Spatial Fisher vector (SFV)
  (Krapac et al., ICCV, 2011)
  - encodes first and second moments of visual word locations
  - adds 6 entries for each visual word: μ and σ for (x, y, t) coordinates.
- Compared to spatial pyramids:
  (Oneaţă et al., ICCV, 2013)
  - similar performance gain
  - SFV are more compact



Schematic illustration of the spatial Fisher vector for three types of visual words (○, ×, □) in an image.

Fisher vector for location

- Spatial Fisher vector (SFV)
  (Krapac et al., ICCV, 2011)
  - encodes first and second moments of visual word locations
  - adds 6 entries for each visual word: μ and σ for (x, y, t) coordinates.
- Compared to spatial pyramids:
  (Oneaţă et al., ICCV, 2013)
  - similar performance gain
  - SFV are more compact
  - complementary.

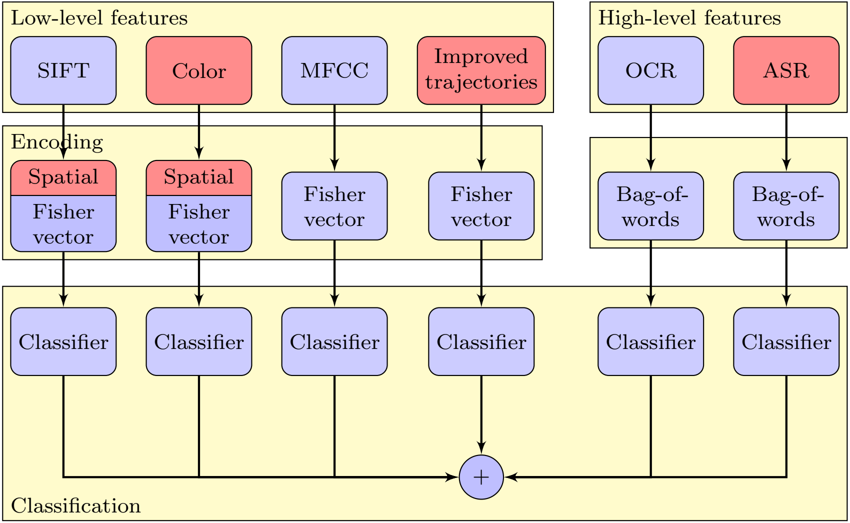


Schematic illustration of the spatial Fisher vector for three types of visual words (○, ×, □) in an image.

Table of Contents

High-level features: OCR and ASR

- Optical character recognition (OCR)
- Automatic speech recognition (ASR) (from Fraunhofer IAIS)
  - trained on 100 hours of English broadcasts
  - language model trained on news articles and patents
- For both systems:
  - bag-of-words encoding with $110,000$ words.
  - tf-idf weighting
  - ℓ_2 normalization.



Low-level features
SIFT
Color
MFCC
Improved trajectories
High-level features
OCR
ASR
Encoding
Spatial
Fisher vector
Spatial
Fisher vector
Fisher vector
Fisher vector
Bag-of-words
Bag-of-words
Classifier
Classifier
Classifier
Classifier
Classifier
Classifier
+
Classification

Table of Contents

Initial experiments on TRECVid '11 subset

Initial experiments on TRECVid '11 subset

- Spatial Fisher vectors improve for color and SIFT.

Initial experiments on TRECVid '11 subset

- Spatial Fisher vectors improve for color and SIFT.
- Comparison of the motion features (HOG, HOF, MBH):

Initial experiments on TRECVid '11 subset

- Spatial Fisher vectors improve for color and SIFT.
- Comparison of the motion features (HOG, HOF, MBH):
  - MBH > HOG > HOF

Initial experiments on TRECVid '11 subset

- Spatial Fisher vectors improve for color and SIFT.
- Comparison of the motion features (HOG, HOF, MBH):
  - MBH > HOG > HOF
  - HOG+MBH > HOF+MBH

Initial experiments on TRECVid '11 subset

- Spatial Fisher vectors improve for color and SIFT.
- Comparison of the motion features (HOG, HOF, MBH):
    - MBH > HOG > HOF
    - HOG+MBH > HOF+MBH
    - The combination of all the three channels is the best.

Initial experiments on TRECVid '11 subset

- Spatial Fisher vectors improve for color and SIFT.
- Comparison of the motion features (HOG, HOF, MBH):
  - MBH $>$ HOG $>$ HOF
  - HOG+MBH $>$ HOF+MBH
  - The combination of all the three channels is the best.
- SIFT descriptors are complementary to the motion features.

Initial experiments on TRECVid '11 subset

- Spatial Fisher vectors improve for color and SIFT.
- Comparison of the motion features (HOG, HOF, MBH):
  - MBH > HOG > HOF
  - HOG+MBH > HOF+MBH
  - The combination of all the three channels is the best.
- SIFT descriptors are complementary to the motion features.
- Total processing time was 27 times slower than real-time on a single core.

Overview of our system: descriptors' dimensions and processing time.

Modality	Descriptor	Encoding	D	$\times$Real time
Motion	HOG+HOF+MBH	FV+H3	51k	10
Image	SIFT	FV+SFV	34k	2
Image	Color	FV+SFV	73k	10
Audio	MFCC	FV	20k	0.05
Image	OCR	BoW (sparse)	110k	1.5
Audio	ASR	BoW (sparse)	110k	3

Results on TRECVid '11 data

- Comparison to our earlier systems.

	DCR	mAP
Best TV'11	0.437	
AXES 2011	0.642	
AXES 2012	0.411	44.5
AXES 2013	**0.379**	**52.6**

Results on TRECVid '11 data

- Comparison to our earlier systems.
- Performance for individual channels

	DCR	mAP
Best TV'11	0.437	
AXES 2011	0.642	
AXES 2012	0.411	44.5
AXES 2013	**0.379**	**52.6**
Motion + SIFT		46.4
Color		27.7
Audio		18.2
ASR		8.2
OCR		10.8

Results on TRECVid '13 data

MED pre-specified		MED ad-hoc	
Team	mAP	Team	mAP
AXES (1/15)	34.6	AXES (1/14)	36.6
BBNVISER (2/15)	33.0	CMU (2/14)	36.3
median	24.7	median	23.3

MED results, for the PROGAll, 100Ex challenge.

Results on TRECVid '13 data

MED pre-specified		MED ad-hoc	
Team	mAP	Team	mAP
AXES (1/15)	34.6	AXES (1/14)	36.6
BBNVISER (2/15)	33.0	CMU (2/14)	36.3
median	24.7	median	23.3

MED results, for the PROGAll, 100Ex challenge.

Team	Full system	ASR	Audio	OCR	Visual
AXES	**36.6**	1.0	12.4	1.1	**29.4**
BBNVISER	32.2	**8.0**	15.1	**5.3**	23.4
CMU	36.3	5.7	**16.1**	3.7	28.4
Genie	20.2	4.3	10.1	—	16.9
IBM-Columbia	2.8	—	0.2	—	2.8
MediaMill	25.3	—	5.6	—	23.8
NII	24.9	—	8.8	—	19.9
ORAND	3.8	—	—	—	3.8
PicSOM	0.6	—	0.1	—	0.6
SRIAURORA	24.2	3.9	9.6	4.3	20.4
Sesame	25.7	3.9	5.6	0.2	23.2
VisQMUL	0.2	—	0.2	—	0.2

Per-channel results on the MED ad-hoc 100Ex, challenge.

Conclusions

- Key components of our system:
  - Improved motion features
  - Spatial Fisher vector.
- Code available on our web-site
  `http://lear.inrialpes.fr/software`
- Check out our posters:
  - Action recognition with improved trajectories.
  - Action and event recognition with Fisher vectors on a compact feature set.