AXES © TRECVID MED 2013

Matthijs Douze\textsuperscript{1}, Zaid Harchaoui\textsuperscript{1}, Dan Onea\textsuperscript{3}a\textsuperscript{1}, Danila Potapov\textsuperscript{1}, Jérôme Revaud\textsuperscript{1}, Cordelia Schmid\textsuperscript{1}, Jochen Schwenninger\textsuperscript{2}, Jakob Verbeek\textsuperscript{1}, Heng Wang\textsuperscript{1}

\textsuperscript{1}INRIA–LEAR, Grenoble, France
\textsuperscript{2}Fraunhofer Sankt Augustin, Germany
Outline

Low-level features
- SIFT
- Color
- MFCC
- Improved trajectories

Encoding
- Spatial
- Fisher vector
- Spatial
- Fisher vector
- Fisher vector
- Fisher vector

High-level features
- OCR
- ASR
- Bag-of-words
- Bag-of-words

Classifier
- Classifier
- Classifier
- Classifier
- Classifier
- Classifier
- Classifier

Classification

+
Outline

Low-level features
- SIFT
- Color
- MFCC
- Improved trajectories

Encoding
- Spatial Fisher vector
- Spatial Fisher vector
- Fisher vector
- Fisher vector

High-level features
- OCR
- ASR
- Bag-of-words
- Bag-of-words

Classifier

Classification
Outline

Low-level features
- SIFT
- Color
- MFCC
- Improved trajectories

Encoding
- Spatial Fisher vector
- Spatial Fisher vector
- Fisher vector
- Fisher vector

High-level features
- OCR
- ASR

Classifier

Classification
Outline

Low-level features
- SIFT
- Color
- MFCC
- Improved trajectories

Encoding
- Spatial Fisher vector
- Spatial Fisher vector
- Fisher vector
- Fisher vector

High-level features
- OCR
- ASR
- Bag-of-words
- Bag-of-words

Classifier
- Classifier
- Classifier
- Classifier
- Classifier

Classification
Table of Contents

1. Low-level features: static, motion, audio

2. Feature encoding: Fisher vector

3. High-level features

4. Experiments and results
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).

\[
\begin{array}{ccc}
\mu, \sigma & \mu, \sigma & \mu, \sigma \\
\end{array}
\]

Mean and variance... 2
... of RGB values... 3
... in 4 \times 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 successive frames
Improved motion features (Wang and Schmid, ICCV, 2013)

- Idea: stabilize camera motion before computing optical flow.

Two successive frames

Optical flow
**Improved motion features** (Wang and Schmid, ICCV, 2013)

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

Two successive 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 successive frames

Warped optical flow

Optical flow

Removed trajectories
Removed trajectories under various camera motions
Table of Contents

1. Low-level features: static, motion, audio

2. Feature encoding: Fisher vector

3. High-level features

4. Experiments and results
Fisher vector for appearance

- Generalization of the bag-of-words.
- Strong performance across multiple tasks:
  - action recognition, action detection, event recognition
    (Oneață 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 (Oneață 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: $\mu$ and $\sigma$ 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: $\mu$ and $\sigma$ 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: $\mu$ and $\sigma$ for $(x, y, t)$ coordinates.

- **Compared to spatial pyramids:**
  (Oneață et al., ICCV, 2013)
  - similar performance gain

Schematic illustration of the spatial Fisher vector for three types of visual words ($\bigcirc$, $\times$, $\square$) 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: $\mu$ and $\sigma$ 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:
    $\mu$ and $\sigma$ for $(x, y, t)$ coordinates.

- Compared to spatial pyramids:
  
  (Oneață 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

1. **Low-level features: static, motion, audio**

2. Feature encoding: Fisher vector

3. **High-level features**

4. Experiments and results
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
  - $\ell_2$ normalization.
Low-level features

SIFT  Color  MFCC  Improved trajectories

Encoding

Spatial  Spatial  Fisher  Fisher  Fisher
vector  vector  vector

High-level features

OCR  ASR

Bag-of-words  Bag-of-words

Classifier  Classifier  Classifier  Classifier  Classifier  Classifier

Classification

+
Table of Contents

1 Low-level features: static, motion, audio

2 Feature encoding: Fisher vector

3 High-level features

4 Experiments and results
Initial experiments on TREC Vid '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.
Initial experiments on TREC Vid ’11 subset

- Spatial Fisher vectors improve for color and SIFT.
Initial experiments on TREC Vid ’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.

\[ \times \text{Real} \]

<table>
<thead>
<tr>
<th>Modality</th>
<th>Descriptor</th>
<th>Encoding</th>
<th>Dimensions</th>
<th>Processing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>SIFT</td>
<td>FV+SFV</td>
<td>34k</td>
<td>2</td>
</tr>
<tr>
<td>Image</td>
<td>Color</td>
<td>FV+SFV</td>
<td>73k</td>
<td>10</td>
</tr>
<tr>
<td>Audio</td>
<td>MFCC</td>
<td>FV</td>
<td>20k</td>
<td>0.05</td>
</tr>
<tr>
<td>Image</td>
<td>OCR</td>
<td>BoW (sparse)</td>
<td>110k</td>
<td>1.5</td>
</tr>
<tr>
<td>Audio</td>
<td>ASR</td>
<td>BoW (sparse)</td>
<td>110k</td>
<td>3</td>
</tr>
</tbody>
</table>
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

Total processing time was 27 times slower than real-time on a single core.

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

<table>
<thead>
<tr>
<th>Modality</th>
<th>Descriptor Encoding</th>
<th>Dimensions</th>
<th>Processing Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion</td>
<td>HOG+HOF+MBH FV+H3</td>
<td>51k</td>
<td>10</td>
</tr>
<tr>
<td>Image</td>
<td>SIFT FV+SFV</td>
<td>34k</td>
<td>2</td>
</tr>
<tr>
<td>Image</td>
<td>Color FV+SFV</td>
<td>73k</td>
<td>10</td>
</tr>
<tr>
<td>Audio</td>
<td>MFCC FV</td>
<td>20k</td>
<td>0.05</td>
</tr>
<tr>
<td>Image</td>
<td>OCR BoW (sparse)</td>
<td>110k</td>
<td>1.5</td>
</tr>
<tr>
<td>Audio</td>
<td>ASR BoW (sparse)</td>
<td>110k</td>
<td>3</td>
</tr>
</tbody>
</table>
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

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

<table>
<thead>
<tr>
<th>Modality Descriptor Encoding</th>
<th>D_time</th>
<th>Motion</th>
<th>HOG+HOF+MBH FV+H3</th>
<th>51k</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image SIFT FV+SFV</td>
<td></td>
<td></td>
<td></td>
<td>34k</td>
<td>2</td>
</tr>
<tr>
<td>Image Color FV+SFV</td>
<td></td>
<td></td>
<td></td>
<td>73k</td>
<td>10</td>
</tr>
<tr>
<td>Audio MFCC FV</td>
<td></td>
<td></td>
<td></td>
<td>20k</td>
<td>0.05</td>
</tr>
<tr>
<td>Image OCR BoW (sparse)</td>
<td></td>
<td></td>
<td></td>
<td>110k</td>
<td>1.5</td>
</tr>
<tr>
<td>Audio ASR BoW (sparse)</td>
<td></td>
<td></td>
<td></td>
<td>110k</td>
<td>3</td>
</tr>
</tbody>
</table>

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.
Initial experiments on TREC Vid ’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 TRECVis ’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.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Descriptor</th>
<th>Encoding</th>
<th>(D)</th>
<th>(\times) Real time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion</td>
<td>HOG+HOF+MBH</td>
<td>FV+H3</td>
<td>51k</td>
<td>10</td>
</tr>
<tr>
<td>Image</td>
<td>SIFT</td>
<td>FV+SFV</td>
<td>34k</td>
<td>2</td>
</tr>
<tr>
<td>Image</td>
<td>Color</td>
<td>FV+SFV</td>
<td>73k</td>
<td>10</td>
</tr>
<tr>
<td>Audio</td>
<td>MFCC</td>
<td>FV</td>
<td>20k</td>
<td>0.05</td>
</tr>
<tr>
<td>Image</td>
<td>OCR</td>
<td>BoW (sparse)</td>
<td>110k</td>
<td>1.5</td>
</tr>
<tr>
<td>Audio</td>
<td>ASR</td>
<td>BoW (sparse)</td>
<td>110k</td>
<td>3</td>
</tr>
</tbody>
</table>
Comparison to our earlier systems.

<table>
<thead>
<tr>
<th></th>
<th>DCR</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best TV ’11</td>
<td>0.437</td>
<td></td>
</tr>
<tr>
<td>AXES 2011</td>
<td>0.642</td>
<td></td>
</tr>
<tr>
<td>AXES 2012</td>
<td>0.411</td>
<td>44.5</td>
</tr>
<tr>
<td>AXES 2013</td>
<td>0.379</td>
<td>52.6</td>
</tr>
</tbody>
</table>
Results on TRECVid ’11 data

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

<table>
<thead>
<tr>
<th></th>
<th>DCR</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best TV’11</td>
<td>0.437</td>
<td></td>
</tr>
<tr>
<td>AXES 2011</td>
<td>0.642</td>
<td></td>
</tr>
<tr>
<td>AXES 2012</td>
<td>0.411</td>
<td>44.5</td>
</tr>
<tr>
<td>AXES 2013</td>
<td>0.379</td>
<td>52.6</td>
</tr>
<tr>
<td>Motion + SIFT</td>
<td>46.4</td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td>27.7</td>
<td></td>
</tr>
<tr>
<td>Audio</td>
<td>18.2</td>
<td></td>
</tr>
<tr>
<td>ASR</td>
<td>8.2</td>
<td></td>
</tr>
<tr>
<td>OCR</td>
<td>10.8</td>
<td></td>
</tr>
</tbody>
</table>
### Results on TRECVID ’13 data

<table>
<thead>
<tr>
<th>MED pre-specified</th>
<th>mAP</th>
<th>MED ad-hoc</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
<td></td>
<td>Team</td>
<td></td>
</tr>
<tr>
<td>AXES (1/15)</td>
<td>34.6</td>
<td>AXES (1/14)</td>
<td>36.6</td>
</tr>
<tr>
<td>BBNVISER (2/15)</td>
<td>33.0</td>
<td>CMU (2/14)</td>
<td>36.3</td>
</tr>
<tr>
<td>median</td>
<td>24.7</td>
<td>median</td>
<td>23.3</td>
</tr>
</tbody>
</table>

MED results, for the PROGAll, 100Ex challenge.
Results on TRECVID ’13 data

<table>
<thead>
<tr>
<th>Team</th>
<th>MED pre-specified</th>
<th></th>
<th></th>
<th></th>
<th>MED ad-hoc</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP</td>
<td></td>
<td></td>
<td></td>
<td>mAP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AXES (1/15)</td>
<td>34.6</td>
<td></td>
<td></td>
<td></td>
<td>AXES (1/14)</td>
<td>36.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBNVISER (2/15)</td>
<td>33.0</td>
<td></td>
<td></td>
<td></td>
<td>CMU (2/14)</td>
<td>36.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>median</td>
<td>24.7</td>
<td></td>
<td></td>
<td></td>
<td>median</td>
<td>23.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MED results, for the PROGAll, 100Ex challenge.

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

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.