Conclusions and questions from last year

- What are the individual contributions of audio and video?
  - Audio weaker than video, apparently
    - But complementary to image
    - Further improvement possible?
  - Fusion step is critical
    - Is early fusion an option?

- Scoring strategies to optimize NDCR looks critical
  - Keep maximum 1 result per query?
Our Runs at Trecvid

<table>
<thead>
<tr>
<th>Run</th>
<th>Profile</th>
<th>Visual</th>
<th>Audio</th>
<th>Fusion</th>
<th>Cut</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEAF</td>
<td>balanced</td>
<td>yes</td>
<td>no</td>
<td>N/A</td>
<td>yes</td>
</tr>
<tr>
<td>AUDIOONLY</td>
<td>balanced</td>
<td>no</td>
<td>yes</td>
<td>N/A</td>
<td>yes</td>
</tr>
<tr>
<td>THEMIS</td>
<td>balanced</td>
<td>yes</td>
<td>yes</td>
<td>late</td>
<td>yes</td>
</tr>
<tr>
<td>ZOZO</td>
<td>balanced</td>
<td>yes</td>
<td>yes</td>
<td>late</td>
<td>no</td>
</tr>
<tr>
<td>DODO\textsubscript{bal}</td>
<td>balanced</td>
<td>yes</td>
<td>yes</td>
<td>early</td>
<td>yes</td>
</tr>
<tr>
<td>TYCHE</td>
<td>nofa</td>
<td>yes</td>
<td>yes</td>
<td>late</td>
<td>yes</td>
</tr>
<tr>
<td>DODO\textsubscript{nofa}</td>
<td>nofa</td>
<td>yes</td>
<td>yes</td>
<td>early</td>
<td>yes</td>
</tr>
</tbody>
</table>

- 5 runs to measure the individual contributions of our system
- 2 runs designed for “best” search quality: the DODO runs
Video visual system: ingredients (same as in 2010)

- Local descriptors: CS-LBP
- Weak geometric consistency
- Hamming Embedding
  - Improve bag-of-features
- Burstiness strategy
  + Multi-probe
Audio system: basic ingredients (same as in 2010)

- Base descriptor: Filter banks
- Overlapping temporal analysis
- Compounding
- Matching: product quantization

Overall bandwidth
500 Hz 3000 Hz

Time (ms)
0 10 20 30

500 Hz
3000 Hz

\[ f_1, f_2, f_3, \ldots, f_N \]

\[ d_1, d_2, d_3, d_4, \ldots, d_m \]

\[ \text{Time (ms)} \]

\[ \text{Overall bandwidth} \]

Inria
INVENTEURS DU MONDE NUMÉRIQUE
New ingredient 1: temporal shift

DB descriptors

10ms

Query misaligned: Not lucky!

5ms

Query descriptors: query all shifts (5 * slower!)

2ms shift

4ms shift

6ms shift

8ms shift
New ingredient 2: reciprocal nearest neighbors

- Audio matches: k-nearest neighbors
- Pb: if X neighbor of Y, Y not necessarily neighbor of X
- Weighted **Reciprocal nearest neighbors**
Audio: improvement over last year

- 6 times slower in total, for a limited improvement
New ingredient 3: early fusion audio/video

- Early fusion:
  - Input: image & audio raw Hough hypotheses
  - Robust time warping to align query frames with DB frames

Example of time warping matrix:
Early fusion

- Input: image & audio raw Hough hypotheses

1. Robust time warping - align query frames with db frames

2. Description of matching segments
   - segment length, number of audio/image frame matches, …
   - surface of the image recognized on the database side
   - KL-divergence between db keypoints distribution / matches distribution
   - relative support of image & audio for the hypothesis
   - etc.

3. Classifier produces a score
Early fusion: training the classifier

- Boosting scheme:
  - Each iteration, addition of a new feature
    - Criterion: maximize AP on validation set
  - Classifier: Logistic regression (better than SVM here)
    - 40,000 positive samples
    - 150,000 negative examples

- Result: selected features (sorted)
  - Detected area
  - Nb of audio & image frame matches
  - KL divergence between keypoints distribution
  - Length of matching segment in seconds
  - Etc…
2010 vs 2011 approach

![Precision vs Recall Graph](image)

- **Late fusion (2010 baseline)**
- **Early fusion + classifier (2011 approach)**
2010 vs 2011 approach
### Analysis: balanced profile (opt-NDCR)

<table>
<thead>
<tr>
<th>RUN</th>
<th>AUDIO</th>
<th>VIDEO</th>
<th>FUSION</th>
<th>NDCR  (avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEAF</td>
<td>no</td>
<td>yes</td>
<td>n/a</td>
<td>0.258</td>
</tr>
<tr>
<td>AUDIOONLY</td>
<td>yes</td>
<td>no</td>
<td>n/a</td>
<td>0.406</td>
</tr>
<tr>
<td>THEMIS</td>
<td>yes</td>
<td>yes</td>
<td>late</td>
<td>0.211</td>
</tr>
<tr>
<td>ZOZO</td>
<td>yes</td>
<td>yes</td>
<td>late</td>
<td>0.194</td>
</tr>
<tr>
<td>DODO</td>
<td>yes</td>
<td>yes</td>
<td>early</td>
<td>0.144</td>
</tr>
</tbody>
</table>

- One surprise: ZOZO > THEMIS
  - Keeping more than 1 result is better if scores are ties
Overview of results (opt-NDCR)

<table>
<thead>
<tr>
<th>RANK</th>
<th>INRIA</th>
<th>PKU</th>
<th>CRIM</th>
<th>NTT-CSL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>31</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>23</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>2</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>0</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RANK</th>
<th>INRIA</th>
<th>PKU</th>
<th>CRIM</th>
<th>NTT-CSL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>14</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>31</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>10</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>1</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

- PKU and CRIM are much better with Actual-NDCR
  - We don’t know how to set the threshold
  - This problem may be inherent to our system
Introduction of Babaz audio matching system


- Well… PQ-codes replaced by k-means LSH (licensing issue)
  - Requires more memory (40GB instead of 5GB) and slower
  - But PQ-codes Matlab implementation available

- All Trecvid queries: query times (16 cores), memory, mAP
  - Pqcodes – heavy: 20H 5GB mAP = 80.7 %
  - Pqcodes – light: 3H 5GB mAP = 78.9 %
  - K-means LSH: 25H 40GB mAP = 78.8 %

Questions?