Combining Features at Search Time: PRISMA at TRECVID 2011

Juan Manuel Barrios\textsuperscript{1}, Benjamin Bustos\textsuperscript{1}, and Xavier Anguera\textsuperscript{2}

\textsuperscript{1} PRISMA Research Group, Department of Computer Science, University of Chile.

\textsuperscript{2} Telefónica Research, Barcelona, Spain.

Content-Based Video Copy Detection Task, TRECVID.
December 7, 2011
P-VCD Overview

- P-VCD System developed for TRECVID 2010. [1]

- **2010**: Visual-only detection.
  - Global descriptors.
  - Approximate k-NN search using pivots.

- **2011**: Audio+Visual detection.
  - Fusion of audio and global descriptors at the similarity search: “distance fusion”.
  - Approximate search as a filtering step.
  - Sequential (exact) A+V search.

Fusion at Decision Level

Reference Videos

Query Videos

Independent Systems

System A (audio-only)
Descriptor A

System B (visual-only)
Descriptor B

System C (visual-only)
Descriptor C

Copy Candidates

Candidates A
ref start/end score
... ... ...

Candidates B
ref start/end score
... ... ...

Candidates C
ref start/end score
... ... ...

Fusion

Intersection, Union, Score aggregation, etc....

Detection List
ref start/end score
ref start/end score
... ... ...

CCD TASK
PRISMA (University of Chile)
Fusion at Similarity Search Level

**Similarity Search**
for each query segment locates the $k$ closest reference segments

**Distance Fusion**
$\delta(\text{segment}_i, \text{segment}_j)$
Linear fusion $= \sum w_i \cdot d_i(,)$

**Copy Localization**
for each query video locates chains of segments with temporal consistency

**k-NN list**
```
segment1 1NN dist1 2NN dist2
segment2 1NN dist1 2NN dist2
segment3 1NN dist1 2NN dist2
segment4 1NN dist1 2NN dist2
...... ...... ..... 
```

**Detection List**
```
ref start/end score ref start/end score ref start/end score 
...... ...... ..... 
```
P-VCD 2011 Overview

Reference Videos → Query Videos

Preprocessing → Video Segmentation → Feature Extraction → Descriptors

Exact k-NN Search → Filtering Step

Distance Fusion → k-NN list

... segment1 1NN dist1 2NN dist2
... segment2 1NN dist1 2NN dist2
... segment3 1NN dist1 2NN dist2
... segment4 1NN dist1 2NN dist2
...

Copy Localization

Detection List

ref start/end score
ref start/end score
...

PRISMA (University of Chile)

CCD TASK
1. Preprocessing

- Removes black borders and noisy frames from each query and reference video.
- For each query video, it creates a flipped version and detects and reverts PIP and camcording.

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>Visual</th>
<th>Audio+Visual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Queries</td>
<td>1,407</td>
<td>1,608</td>
<td>11,256</td>
</tr>
<tr>
<td>New Queries</td>
<td>-</td>
<td>3,539</td>
<td>-</td>
</tr>
<tr>
<td>Total Queries</td>
<td>1,407</td>
<td>5,147</td>
<td>36,029</td>
</tr>
</tbody>
</table>
2. Video Segmentation

- Partitions every query and reference video into segments of 0.333 ms length (visual and audio track).

<table>
<thead>
<tr>
<th></th>
<th>Audio segments</th>
<th>Visual segments</th>
<th>Audio+Visual segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query collection</td>
<td>306,304</td>
<td>1,120,455</td>
<td>7,840,587</td>
</tr>
<tr>
<td>Reference collection</td>
<td>4,441,717</td>
<td>4,522,262</td>
<td>4,387,633</td>
</tr>
</tbody>
</table>
3. Feature Extraction

- Three Visual-Global descriptors per segment:
  - Edge Histogram (Ehd): $4 \times 4 \times 10 = 160d$.  
  - Gray Histogram (Gry): $4 \times 4 \times 12 = 192d$.  
  - Color Histogram (Rgb): $4 \times 4 \times 12 = 192d$.  

- The descriptor for a visual segment is the average descriptor for every frame.

- One Audio Descriptor (Aud), 160d.
4. Distance Fusion

- Distance between two descriptors: Manhattan distance (city-block)
  \[ L_1(\vec{x}, \vec{y}) = \sum_{i=0}^{dim} |x_i - y_i| \]

- Distance between any two Audio+Visual segments:
  \[ d_{av}(q, r) = \frac{w_1}{\tau_1} \cdot L_1(\text{Ehd}(q), \text{Ehd}(r)) + \frac{w_2}{\tau_2} \cdot L_1(\text{Rgb}(q), \text{Rgb}(r)) + \frac{w_3}{\tau_3} \cdot L_1(\text{Aud}(q), \text{Aud}(r)) \]

- Normalization factors \( \tau_i \) and weighting factors \( w_i \) are calculated by the “\( \alpha \)-Normalization” and “weighting by max-\( \tau \)” algorithms. [1]
4. Distance Fusion (cont.)

- For efficiency, we define two more distances:
  - Between two audio segments:
    \[ d_a(q, r) = L_1(\text{Aud}(q), \text{Aud}(r)) \]
  - Between two visual segments:
    \[ d_v(q, r) = \frac{w_1}{\tau_1} * L_1(\text{Ehd}(q), \text{Ehd}(r)) + \frac{w_2}{\tau_2} * L_1(\text{Rgb}(q), \text{Rgb}(r)) \]
    \[ d_v(q, r) = \frac{w_1}{\tau_1} * L_1(\text{Ehd}(q), \text{Ehd}(r)) + \frac{w_2}{\tau_2} * L_1(\text{Gry}(q), \text{Gry}(r)) \]
5. Search Domain Filtering

- It performs approximate k-NN searches [1] using visual-only distance and audio-only distance.
  - Requirement: \( d \) complies the triangle inequality.

- Distance approximation: \( d(a, b) \approx |d(a, p) - d(p, b)| \)

- For many pivots: \( d(a, b) \approx \max_{p \in P} |d(a, p) - d(p, b)| \)

- It evaluates the actual distance only for the pairs with lowest approximated distance.
5. Search Domain Filtering

- Perform approximate k-NN searches for each query segment using visual-only distance and audio-only distance (k=30).

<table>
<thead>
<tr>
<th>Query Video</th>
<th>1st NN</th>
<th>2nd NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query1-Segm1</td>
<td>Vid07-Segm54 dist</td>
<td>Vid08-Segm73 dist</td>
</tr>
<tr>
<td>Query1-Segm2</td>
<td>Vid09-Segm13 dist</td>
<td>Vid02-Segm34 dist</td>
</tr>
<tr>
<td>Query1-Segm3</td>
<td>Vid07-Segm34 dist</td>
<td>Vid03-Segm54 dist</td>
</tr>
<tr>
<td>Query1-Segm4</td>
<td>Vid09-Segm15 dist</td>
<td>Vid02-Segm13 dist</td>
</tr>
<tr>
<td>Query1-Segm5</td>
<td>Vid01-Segm88 dist</td>
<td>Vid01-Segm12 dist</td>
</tr>
<tr>
<td>Query1-Segm6</td>
<td>Vid09-Segm54 dist</td>
<td>Vid09-Segm17 dist</td>
</tr>
<tr>
<td>Query1-Segm7</td>
<td>Vid01-Segm45 dist</td>
<td>Vid03-Segm43 dist</td>
</tr>
<tr>
<td>Query1-Segm8</td>
<td>Vid09-Segm19 dist</td>
<td>Vid01-Segm12 dist</td>
</tr>
</tbody>
</table>

For each query video, it selects the $D$ reference videos that have more segments in the k-NN lists ($D=40$).

Query1 ➔ {Vid01, Vid02, Vid03, Vid07, Vid08, Vid09}
Query2 ➔ {Vid02, Vid04, Vid06, Vid07}
6. Exact k-NN Search

- For each query segment performs an exact k-NN search using the audio+visual distance (k=10).
- The search space domain depends on each query video.

Query1 → {Vid01, Vid02, Vid03, Vid07, Vid08, Vid09}

<table>
<thead>
<tr>
<th>Query Video</th>
<th>1st NN</th>
<th>2nd NN</th>
<th>3rd NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query1-Segm1</td>
<td>Vid07-Segm54</td>
<td>Vid08-Segm73</td>
<td>Vid01-Segm68</td>
</tr>
<tr>
<td></td>
<td>dist</td>
<td>dist</td>
<td>dist</td>
</tr>
<tr>
<td>Query1-Segm2</td>
<td>Vid09-Segm13</td>
<td>Vid02-Segm34</td>
<td>Vid02-Segm33</td>
</tr>
<tr>
<td></td>
<td>dist</td>
<td>dist</td>
<td>dist</td>
</tr>
<tr>
<td>Query1-Segm3</td>
<td>Vid07-Segm34</td>
<td>Vid03-Segm54</td>
<td>Vid09-Segm14</td>
</tr>
<tr>
<td></td>
<td>dist</td>
<td>dist</td>
<td>dist</td>
</tr>
<tr>
<td>Query1-Segm4</td>
<td>Vid09-Segm15</td>
<td>Vid02-Segm13</td>
<td>Vid03-Segm65</td>
</tr>
<tr>
<td></td>
<td>dist</td>
<td>dist</td>
<td>dist</td>
</tr>
<tr>
<td>Query1-Segm5</td>
<td>Vid01-Segm88</td>
<td>Vid01-Segm12</td>
<td>Vid07-Segm58</td>
</tr>
<tr>
<td></td>
<td>dist</td>
<td>dist</td>
<td>dist</td>
</tr>
<tr>
<td>Query1-Segm6</td>
<td>Vid09-Segm54</td>
<td>Vid09-Segm17</td>
<td>Vid07-Segm59</td>
</tr>
<tr>
<td></td>
<td>dist</td>
<td>dist</td>
<td>dist</td>
</tr>
<tr>
<td>Query1-Segm7</td>
<td>Vid01-Segm45</td>
<td>Vid03-Segm43</td>
<td>Vid03-Segm20</td>
</tr>
<tr>
<td></td>
<td>dist</td>
<td>dist</td>
<td>dist</td>
</tr>
<tr>
<td>Query1-Segm8</td>
<td>Vid09-Segm19</td>
<td>Vid01-Segm12</td>
<td>Vid07-Segm61</td>
</tr>
<tr>
<td></td>
<td>dist</td>
<td>dist</td>
<td>dist</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
7. Copy Localization

- Locates chains of NN with temporal consistency. [1]
  
  **k-NN list** $d_{av}$
  
<table>
<thead>
<tr>
<th>Query Video</th>
<th>1st NN</th>
<th>2nd NN</th>
<th>3rd NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query1-Segm1</td>
<td>Vid07-Segm54 dist</td>
<td>Vid08-Segm73 dist</td>
<td>Vid01-Segm68 dist</td>
</tr>
<tr>
<td>Query1-Segm2</td>
<td>Vid07-Segm34 dist</td>
<td>Vid02-Segm34 dist</td>
<td>Vid02-Segm33 dist</td>
</tr>
<tr>
<td>Query1-Segm3</td>
<td><strong>Vid09-Segm13 dist</strong></td>
<td>Vid03-Segm54 dist</td>
<td>Vid03-Segm65 dist</td>
</tr>
<tr>
<td>Query1-Segm4</td>
<td>Vid07-Segm34 dist</td>
<td>Vid02-Segm13 dist</td>
<td>Vid03-Segm65 dist</td>
</tr>
<tr>
<td>Query1-Segm5</td>
<td><strong>Vid09-Segm15 dist</strong></td>
<td>Vid01-Segm12 dist</td>
<td>Vid07-Segm58 dist</td>
</tr>
<tr>
<td>Query1-Segm6</td>
<td>Vid01-Segm88 dist</td>
<td>Vid01-Segm12 dist</td>
<td>Vid07-Segm59 dist</td>
</tr>
<tr>
<td>Query1-Segm7</td>
<td>Vid09-Segm54 dist</td>
<td>Vid03-Segm43 dist</td>
<td>Vid03-Segm20 dist</td>
</tr>
<tr>
<td>Query1-Segm8</td>
<td><strong>Vid09-Segm19 dist</strong></td>
<td>Vid01-Segm12 dist</td>
<td>Vid07-Segm61 dist</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- No False Alarms profile:
  - It reports the candidate with the highest score.

- Balanced profile:
  - It reports the two candidates with highest scores.
TRECVID 2011 Results
No False Alarms profile

- Analysis focused on optimal threshold and average result for all transformations.

- No False Alarms profile:
  - One candidate per query.
  - EhdGry: Combination of two global descriptors
    - Average Optimal NDCR=0.374
    - Average Optimal F1=0.938
    - Average Processing Time=50 s
  - EhdRgbAud: Combination of two global descriptors and audio
    - Average Optimal NDCR=0.286
    - Average Optimal F1=0.946
    - Average Processing Time=64 s

- TRECVID 2010
  - Avg.Opt.NDCR=0.611
  - Avg.Opt.F1=0.828
  - Avg.Proc.Time=128 s
No False Alarms profile

- Multimodal detection outperforms visual-only detection.

- The exact search step increases the accuracy for copy localization.

- Good tradeoff between effectiveness and efficiency.

- Global descriptors can achieve good performance in NoFA profile.
Balanced profile

- Balanced profile:
  - Two candidates per query.
  - **EhdGry**: Combination of two global descriptors
    - Average Optimal NDCR= **0.412**
    - Average Optimal F1= **0.938**
    - Average Processing Time= **50 s**
  - **EhdRgbAud**: Combination of two global descriptors and audio
    - Average NDCR= **0.300**
    - Average F1= **0.955**
    - Average Processing Time= **64 s**
  - **Joint** submission with Telefonica team.
    - **EhdRgb** with twenty candidates per query.
    - Late fusion with Telefonica’s audio and local descriptors.
Balanced profile

- Good localization accuracy.
- Good tradeoff between effectiveness and efficiency.
- Global descriptors achieve better performance in NoFA profile than in Balanced profile.

All these tests were run on a desktop computer:
- Intel Core i7-2600k
- 8 GB RAM
Conclusions

- We have presented the “distance fusion” approach for combining global and audio descriptors.
  - It automatically fixes a good set of weights.
- The approximate search can avoid most of the distance evaluations while achieving a good detection performance.
  - The analysis of the approximate search is in [1].
- The exact search step increases the accuracy for the copy localization.

Future work:
- Fuse audio, global and local descriptors following this approach.
- Test non-metric distances at the exact search step.
- Test a segmentation with overlaps.
Thank you!