Telefónica Research @ Trecvid 2011

Xavier Anguera, Daru Xu\textsuperscript{1} and Tomasz Adamek

(With the collaboration of Juan Manuel Barrios, Prisma Group)

\textsuperscript{1}Daru Xu is a graduate student at the Ming-Hsieh Department of Electrical Engineering, University of Southern California, USA
Outline of the talk

• Telefonica 2011 Video-copy detection system
  – Overall system
  – Video-copy detection
  – Audio-copy detection
  – Fusion algorithm
  – Results

• Multi-systems fusion experiment
Multimodal Video-copy detection

Audio ref. → Local audio system (MASK) → MASK results

Video ref. → Local video system (DART) → Multimodal results (2 systems)

Video ref. → Global video system (PRISMA) → Joint results (3 systems)

Features extraction and indexing

Video query
Video-based System

DART* local features extraction

Key-frame extraction

DART* extraction

Inserted static text & banners filtering

Subtitle filtering

Temporal stability & scale filtering

Matched video segments

Key-frame matching

Temporal consistency post-processing

Ref. Video indexing info.

Video query

Differences from last year:
• Software refactoring
• Elimination of temporary files

Audio-based System

- Time-to-frequency transformation
- Salient spectral points search
- Local mask application
- MASK fingerprint encoding and storage

0110101000110110
0001101110110111
0001001101101010
0001011111010101
1111011000100110
0010010100110101
MASK fingerprint extraction (I)

1) Audio track extraction using FFMPEG
Acoustic fingerprint extraction (I)

1) Audio track extraction using FFMPEG
   - 10ms, 100ms window

2) FFT, bandwidth limited to 300-3KHz
   - 32 MEL-spectrum bands
Acoustic fingerprint extraction (I)

1) Audio track extraction using FFMPEG

2) FFT, bandwidth limited to 300-3KHz

3) Find spectrogram peaks.

32 MEL-spectrum bands
Acoustic fingerprint extraction (II)

4) Apply a mask in each maxima location
Acoustic fingerprint extraction (II)

5) Construct the fingerprint
Multimodal Fusion Algorithm

• Fusion of different modalities at decision level
  – Agnostic of internal system’s behaviors

• No limit on the number of systems to be combined
  – provided each system is better than random

• To work optimally it needs N-best matches from each system. It returns the best fused matches (N=20)
  – Makes use of the individual scores and the rank within each modality.

Paper on ACM MM 2011: “Multimodal Fusion for Video Copy Detection”, Xavier Anguera, Juan Manuel Barrios, Tomasz Adamek and Nuria Oliver
Data preprocessing

Audio scores histogram

Local video scores histogram

Global video scores histogram
N-best flooring and L1 Normalization (I)

L1 normalization

$$\frac{\text{MScore}_i}{\sum_{j=1}^{N_{\text{best}}} \text{MScore}_j}$$
N-best flooring and L1 Normalization (II)

\[
MScore_i = \frac{MScore_i}{\sum_{j=1}^{N\text{best}} MScore_j}
\]
The median score is a good trade-off between normalizations:

- Normalizing the scores distributions

Although we could do much more complicated things by the median score of the scores for all queries in a given modality.

In order to avoid problems in the subsequent steps, the distribution of scores for every modality will usually be given query-ordered by their matching score.

Fusion:

Next we describe in detail each of the steps involved in the multimodal fusion algorithm.

The input of the algorithm can be seen as the NoFa case but we believe it is more usable to have some modalities have been merged or their initial scores.

When all modalities are normalized to the range [-1, 1] to make it easier to later get normalized with the accompanying reference.

The result of the fusion algorithm is a ranked list of the N best matches from the available video matches for many different applications.

We are more prone to make those modalities that do not provide any result, i.e., their scores will be close to zero.

On the contrary, much higher values than the rest in a particular modality, they will stand out across the others.

To void this problem we apply a preprocessing step:

1. We just show two for convenience and space limitations:

   - Overlapping
   - Segments
   - Merge

   Example:

   Segment Q

   Segment R

   Merged segment

   min\{\hat{E}_k^Q(r), \hat{E}_k^R(r)\} - max\{B_k^Q(r), B_k^R(r)\}

   max\{\hat{E}_k^Q(r), \hat{E}_k^R(r)\} - min\{B_k^Q(r), B_k^R(r)\}

   > 0.5

   Examples:

   Mutimodal overlap

   Missing modality

   Non-overlapping modalities
Output score computation

\[
S(c_l) = \frac{\sum_{c_k(r) \in c_l} W_k \cdot \frac{N_k-r+1}{N_k} \cdot \hat{S}_k(r)}{\sum_{k=1}^{K} (W_k \cdot \hat{S}_k(1))}
\]

- Resulting score for fused match
- Number of matches
- A-priori weight for each modality
- Rank \([1 \text{ to } N_k]\)
- Normalized matching score at rank \(r\)
- Best normalized matching score for each modality
# Official evaluation results

Optimum scores, balanced profile:

<table>
<thead>
<tr>
<th>Profile</th>
<th>Min NDCR</th>
<th>FA count</th>
<th>Miss count</th>
<th>True positives</th>
<th>Opt F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio system</td>
<td>BALANCED</td>
<td>0.662</td>
<td>0.66</td>
<td>54.75</td>
<td>54.78</td>
</tr>
<tr>
<td>Multimodal</td>
<td>BALANCED</td>
<td>0.610</td>
<td>0.80</td>
<td>11.73</td>
<td>63.69</td>
</tr>
<tr>
<td>Joint</td>
<td>BALANCED</td>
<td>0.268</td>
<td>0.23</td>
<td>4.71</td>
<td>101.4</td>
</tr>
</tbody>
</table>

Choosing only 1\textsuperscript{st}-best results:

<table>
<thead>
<tr>
<th>Profile</th>
<th>Min NDCR</th>
<th>FA count</th>
<th>Miss count</th>
<th>True positives</th>
<th>Opt F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio system</td>
<td>BALANCED</td>
<td>0.477</td>
<td>0.14</td>
<td>55.89</td>
<td>72.05</td>
</tr>
</tbody>
</table>
Multi-systems fusion experiment

• We tested the fusion algorithm with many system outputs

• We asked participants in TRECVID 2011 for their submitted runs
  – 10 teams contributed their results: PKU-IDM, CRIM, INRIA-TEXMEX/LEAR, FT, prisma, ATTLabs, kddi, iupr-dfki, brno, Telefonica Research
  – I used the “Balanced” runs: 17 runs
Status of the runs

• The fusion algorithm works optimally when Nbest results are available for each fused output.
  – Results for the used systems had (many times) only 1best results, resulting suboptimal for the fusion.
Individual results (Min NDCR)

- Labeled from 1 to 17, to anonymize them.
Individual results (optimum F1)
We incrementally added systems and computed the fusion
Systems 5 and 15 are the only ones making the fusion worse
Final Min_NDCR=0.0333
Fusion of all minus 1

We obtain an order from worse to best in the fusion (worse in here is system 15)
Incremental elimination

- With only 5 systems we achieve pretty decent results
- The best result is 0.0195, although this is “cheating”
Conclusions

• The fusion algorithm can extract knowledge and make results better
  – Even if fusing systems which have weaker NDCR results, the fusion results in good scores.

• FUTURE WORK: automatically identify which modalities bring novelty.