Multi-modal query expansion for video object instances retrieval

**Objectives**
- Retrieve object instances from a large video repository starting from minimum, user-provided textual information
- Leverage on users’ affinity for textual queries and crawl images from the Internet
- Remove outliers from retrieved data and identify representative instances for the topic given by the user
- Build visual descriptors from filtered representative instances and use them for querying the video repository

**Approach overview**
1. Issue textual query
2. Extract local features
3. Match images
4. Build query graph
5. Determine representative images
6. Build query descriptors
7. Aggregate query descriptors
8. Aggregate query results

**Evaluation**
- Trecvid 2012 Instance Search Task Flickr dataset
- 74,958 videos mined from Flickr
- 22 query topic s with up to 9 example images with precise object annotation and basic textual description: 102 query images
- Hessian Affine regions + RootSIFT descriptors from 683,433 keyframes
- Bag-of-Words with vocabulary of 1M visual words

**Examples**
- Eiffel Tower
- Baldachin in Saint Peter’s Basilica
- U.S. Capitol exterior
- Stonehenge
- Empire State Building
- Mercedes star

**Results**
- Retrieval performance (mean Average Precision)
  - Centered representative query: Consists of the union of all points from the representative image that have been matched/shared with at least one neighboring image from the query graph.
  - Distributed representative query: Consists of the union of all points from every neighboring image that have been matched with points from the representative image

<table>
<thead>
<tr>
<th>Expansion method</th>
<th>Number of mined images</th>
<th>Aggregation strategy</th>
<th>mean Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centered representative query</td>
<td>25</td>
<td>A posteriori</td>
<td>0.0455</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>A posteriori</td>
<td>0.0585</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>A posteriori</td>
<td>0.0689</td>
</tr>
<tr>
<td>Distributed representative query</td>
<td>25</td>
<td>A posteriori</td>
<td>0.0540</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>A posteriori</td>
<td>0.0756</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>A posteriori</td>
<td>0.0871</td>
</tr>
</tbody>
</table>

**Conclusion and perspectives**
- Novel multi-modal query definition and expansion method:
  - text ⇨ image ⇨ video
- Good object retrieval performance even when using only textual data
- Distributed query descriptors with a priori aggregation provide better results while reducing the number of query operations
- Extend method for multiple Internet sources
- Use an ad-hoc SVM classifier on representative images
- Integrate other image metadata for validating positive instances (geotags, image popularity, uploader reputation)