ViTS: Video Tagging System From Massive Web Multimedia Collections

Dèlia Fernández, David Varas, Joan Espadaler, Issey Masuda, Jordi Ferreira, Alejandro Woodward, David Rodríguez, Xavier Giró-i-Nieto, Juan Carlos Riveiro and Elisenda Bou
Motivation
Motivation

- Exponential and constant growth of visual content in the Internet.
- Unfeasible to manually labeling these data.
  - Automatic methods for large-scale annotation are needed.
Motivation

- New concepts are constantly appearing.
- Relations between concepts are change over time.

→ System that learns and updates concept relations on real time.
ViTS: Video Tagging System

- ViTS is...
  - Vocabulary of over 2.5M KG concepts
  - Indexing more than 150k videos per month
  - 80,87% of accuracy
ViTS Pipeline

- Video Action Summarization
- Contextual Tagging Algorithm
- Video Dataset
- Knowledge Graph (Relations Matrix)

Production
Real Time Learning
ViTS Pipeline

- Video Action Summarization
- Contextual Tagging Algorithm
- Knowledge Graph (Relations Matrix)

- Video
- Web
- Social Networks

Production
Real Time Learning

Video summaries
Contextual Tags
Video Dataset
ViTS Pipeline

Video

Video Action Summarization

Video summaries

Video Dataset

Web

Contextual Tagging Algorithm

Contextual Tags

Knowledge Graph (Relations Matrix)

Social Networks

Production

Real Time Learning

The Jeremy Lin Show Vs. Dallas Mavericks (2/19/12)

889.3MB visualizations

Jeremy Lin leads the Knicks to a victory against the Mavericks in this memorable NBA game.

Categories: Sports

License: Creative Commons Attribution-ShareAlike 3.0 Unported License
ViTS Pipeline

- Video Action Summarization
- Contextual Tagging Algorithm
- Knowledge Graph (Relations Matrix)
- Video summaries
- Contextual Tags
- Video Dataset
- Production
- Real Time Learning
Knowledge Graph

- **Concepts** are universal semantic representation of worlds
- 2.5M concepts vocabulary
- Multilingual alias
- Freebase / Google Knowledge Graph → Wikidata

<table>
<thead>
<tr>
<th>Wikidata ID</th>
<th>Freebase ID</th>
<th>Description</th>
<th>Types</th>
<th>Alias</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q60</td>
<td>/m/02_286</td>
<td>City in New York</td>
<td>Place, City, Administrative Area</td>
<td>New York City, The Big Apple, New York, NYC, City of New York, New Amsterdam, Nueva York, Ciudad de Nueva York, Nova Iorque, New York</td>
<td>en, es, pt, tr</td>
</tr>
</tbody>
</table>
Knowledge Graph

- $R \rightarrow$ Concept relations matrix
The Knowledge Graph is useful for:

- **Synonymous keywords association**
- **Multilingual tagging**
- **Concept disambiguation**

<table>
<thead>
<tr>
<th>Wikidata ID</th>
<th>Freebase ID</th>
<th>Description</th>
<th>Types</th>
<th>Alias</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q60</td>
<td>/m/02_286</td>
<td>City in New York</td>
<td>Place, City, Administrative Area</td>
<td>New York City, The Big Apple, New York, NYC, City of New York, New Amsterdam, Nueva York, Ciudad de Nueva York, Nova Iorque, New York</td>
<td>en, en, en, en, en, es, es, pt, tr</td>
</tr>
</tbody>
</table>
The Knowledge Graph is useful for:
- Synonymous keywords association
- Multilingual tagging
- Concept disambiguation

Tag: Queen
Description: Music band
Freebase ID: /m/0bk1p

Tag: Elizabeth II
Description: Queen of England
Freebase ID: /m/0d1_f

Tag: Evil Queen
Description: Fictional Character
Freebase ID: /m/02r4l2
The Knowledge Graph is useful for...
- Synonymous keywords association
- Multilingual tagging
- Concept disambiguation → CONTEXT

Buckingham Palace
London

Queen

England
Parliament

Tag: Queen
Description: Music band
Freebase ID: /m/0bk1p

Tag: Elizabeth II
Description: Queen of England
Freebase ID: /m/0d1_f

Tag: Evil Queen
Description: Fictional Character
Freebase ID: /m/02r4l2

Theresa May
1. **Keyword Extractor**
   - Download all related text sources
   - Extract keywords with NLP techniques

2. **Keyword to KG Entity Mapping**
   - Get keyword concept candidates
   - For each concept compute a Concept Score
     - Concept Score = Intra-score + Inter-score
Experiments

- Quality of ViTS tags?
  - Subset from the **YouTube-8M Dataset** [1]
    - 14k videos randomly selected from YouTube-8M
    - Videos are summarized and tagged using ViTS

Experiments

- Quality of ViTS tags?
  - Human raters: Amazon Mechanical Turk (AMT) [2]
  - Random subset of 1.400 English videos

Are these tags correct?

<table>
<thead>
<tr>
<th>#</th>
<th>TAG</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Do not know</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Video Game</td>
<td>✔️</td>
<td></td>
<td>✗</td>
</tr>
<tr>
<td>2</td>
<td>Pet Rescue</td>
<td>✔️</td>
<td></td>
<td>✗</td>
</tr>
<tr>
<td>3</td>
<td>Saga</td>
<td>✔️</td>
<td></td>
<td>✗</td>
</tr>
<tr>
<td>4</td>
<td>Level</td>
<td>✔️</td>
<td></td>
<td>✗</td>
</tr>
<tr>
<td>5</td>
<td>Video Game Culture</td>
<td>✔️</td>
<td></td>
<td>✗</td>
</tr>
</tbody>
</table>

Title: "Pet Rescue Saga Level 627"

### Results

- **Comparison between ViTS and YouTube-8M**
  - Average number of tags in YouTube-8M subset is **3.64 tags/video** VS **10.04 tags/video** extracted by ViTS

<table>
<thead>
<tr>
<th>ViTS</th>
<th>YouTube-8M</th>
<th>ViTS</th>
<th>YouTube-8M</th>
<th>ViTS</th>
<th>YouTube-8M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseball</td>
<td>Game</td>
<td>Thomas Robinson</td>
<td>Basketball</td>
<td>Minecraft</td>
<td>Game</td>
</tr>
<tr>
<td>Alex Rodriguez</td>
<td>Arena</td>
<td>Sacramento Kings</td>
<td>Video game</td>
<td>Video game</td>
<td>Minecraft</td>
</tr>
<tr>
<td>New York Yankees</td>
<td>Athlete</td>
<td>New Jersey</td>
<td>Server</td>
<td>Server</td>
<td></td>
</tr>
<tr>
<td>New York City</td>
<td>Baseball park</td>
<td>Sport</td>
<td>Browser extension</td>
<td>Tutorial</td>
<td></td>
</tr>
<tr>
<td>Yankee Stadium</td>
<td>Stadium</td>
<td>2012 NBA Draft</td>
<td>Download</td>
<td>Download</td>
<td></td>
</tr>
<tr>
<td>SportHit</td>
<td>Home run</td>
<td></td>
<td>Video game culture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home run</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

- Tags Quality Evaluation

Human Rating Results

<table>
<thead>
<tr>
<th>#Videos</th>
<th>#Tags Total</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,400</td>
<td>14,024</td>
<td>80.87%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>% Correct</th>
<th>% Incorrect</th>
<th>% Discarded</th>
</tr>
</thead>
<tbody>
<tr>
<td>77.81%</td>
<td>18.27%</td>
<td>3.90%</td>
</tr>
</tbody>
</table>
Conclusions

- Base of Industrial Product:
  - Video search and indexation
  - Video content recommendation
  - Trending video detection
THANKS FOR YOUR ATTENTION

Dèlia Fernandez
delia@vilyn.com

Extracted tags and video summaries are available at: