TubeTagger
—— YouTube-based Concept Detection ——

Adrian Ulges, Markus Koch
Damian Borth, Thomas M. Breuel

German Research Center for Artificial Intelligence (DFKI) &
University of Kaiserslautern

December 6 2009
Outline

Motivation

TubeTagger System

TubeTagger Web Demo

Summary
Data, data, data

"...TV, video on demand, Internet video, and P2P video will account for over 91 percent of global consumer traffic by 2013..."

"Cisco Visual Networking Index", 2009

"...more than 20 hours of video uploaded to YouTube every minute, 1 billion views per day..."

, 2009

Increasing importance of video live streams: Obama's inauguration, Michael Jackson's memorial service, music video debuts...

D.Borth: : TubeTagger 3 December 6 2009
"...TV, video on demand, Internet video, and P2P video will account for over 91 percent of global consumer traffic by 2013..."

"Cisco Visual Networking Index", 2009
"...TV, video on demand, Internet video, and P2P video will account for over 91 percent of global consumer traffic by 2013..."

"Cisco Visual Networking Index", 2009

"...more than 20 hours of video uploaded to YouTube every minute, 1 billion views per day..."
Data, data, data

"...TV, video on demand, Internet video, and P2P video will account for over 91 percent of global consumer traffic by 2013..."

"Cisco Visual Networking Index", 2009

"...more than 20 hours of video uploaded to YouTube every minute, 1 billion views per day..."

, 2009

Increasing importance of video live streams: Obama’s inauguration, Michael Jackson’s memorial service, music video debuts...

facebook USTREAM, 2009
Content-based Video Retrieval

*How to search in large video databases?*
Content-based Video Retrieval

How to search in large video databases?
Content-based Video Retrieval

How to search in large video databases?

▶ Video Concept Detection [MediaMill, Columbia, IBM,...]
→ as key building block of CBVR
Content-based Video Retrieval

How to search in large video databases?

- Video Concept Detection [MediaMill, Columbia, IBM,...] → as key building block of CBVR
- Supervised Machine Learning → need labeled training data
Training Data

State-of-the-art Datasets

- TRECVID
- LSCOM

Acquisition

- collaborative [Ayache07]
- precise & high quality
- time consuming effort
Training Data

State-of-the-art Datasets

▶ TRECVID
▶ LSCOM

Acquisition

▶ collaborative [Ayache07]
▶ precise & high quality
▶ time consuming effort

▶ vocabulary size does not scale
→ from hundreds to thousands of concepts [Hauptmann07]
▶ missing new emerging concepts of interest
→ e.g. "Michael Jackson", "Obama", "iPod" ¹

¹top ranked searches 2009-Q3 by "Google Insights for Search" for web search, news search and product search respectively
Web-video as training data

Pros
- available at large scale
- high variability of data
- enriched with tags
- comments
- ratings
- could allow automatic concept learning

Cons
- weakly labeled
- incomplete / subjective
- coarse
- focus of interest
- domain change
- video portal as black box
Web-video as training data

Pros

▶ available at large scale
▶ high variability of data
▶ enriched with
  ▶ tags
  ▶ comments
  ▶ ratings
▶ could allow automatic concept learning
Web-video as training data

Pros
- available at large scale
- high variability of data
- enriched with
  - tags
  - comments
  - ratings
- could allow automatic concept learning

Cons
- weakly labeled
  - incomplete / subjective
  - coarse
- focus of interest
- domain change
- video portal as black box
TubeTagger System Overview

- **Notation**
  - $t \in T := \text{tags}$
  - $x \in X_{db} := \text{keyframe of a video}$
  - $P(t| x) := \text{probability of tag } t \text{ for keyframe } x$
TubeTagger System Overview

Notation

- $t \in T$ := tags ($\approx$ concepts)
- $x \in \mathcal{X}_{db}$ := keyframe of a videos
- $P(t|x)$ := probability of tag $t$ for keyframe $x$
TubeTagger System Overview

- Use YouTube as primary training data source.
- Use tags as weak labels.
  - Pos. = tagged.
  - Neg. = all non-tagged.

```
soccer
junior match funny tricks ronaldo 2008 soccer cup city
stadium beckham barcelona
```

Visual Model

Semantic Model
TubeTagger System Overview

Training Data Acquisition

➤ use YouTube as primary training data source
➤ use tags as weak labels
  ➤ pos. = tagged
  ➤ neg. = all non-tagged
TubeTagger System Overview

Training Data Acquisition
- use YouTube as primary training data source
- use tags as weak labels
  - pos. = tagged
  - neg. = all non-tagged

Learning Pipelines
- Visual Concept Learning
- Semantic Concept Learning
Visual Concept Learning

Visual Concept Learning involves the following steps:

1. **Keyframe Extraction**
   - $x \in X_{db}$

2. **Bag-of-Visual-Words Descriptors**
   - SIFT [Lowe99]
   - SURF [Bay06]

3. **Statistical Models**
   - SVM [Schölkopf01]
   - PAMIR [Grangier08, Paredes09]
   - Max. Entropy [Deselaers05]

4. **Output**
   - $P(t|x)$
Visual Concept Learning

Features

- keyframe extraction
  \[ x \in X_{db} \]
- bag-of-visual-words descriptors [Sivic06]
  - SIFT [Lowe99]
  - SURF [Bay06]
Visual Concept Learning

Features

- keyframe extraction
  \[ x \in \mathcal{X}_{db} \]
- bag-of-visual-words descriptors [Sivic06]
  - SIFT [Lowe99]
  - SURF [Bay06]

Statistical Models

- SVM [Schölkopf01]
- PAMIR [Grangier08, Paredes09]
- Max. Entropy [Deselaers05]
Visual Concept Learning

Features

- keyframe extraction
  \[ \rightarrow x \in \mathcal{X}_{db} \]
- bag-of-visual-words descriptors [Sivic06]
  - SIFT [Lowe99]
  - SURF [Bay06]

Statistical Models

- SVM [Sölkopf01]
- PAMIR [Grangier08, Paredes09]
- Max. Entropy [Deselaers05]

\[ \rightarrow \text{output: } P(t|x) \]
Semantic Concept Learning

- Query Formulation
  - concepts $\neq$ textual queries
  - mapping of text queries to concept vocabulary
  - "computer" $\rightarrow$ Monitor, Windows-Desktop, iPhone
  - "funny" $\rightarrow$ Muppets, Cats, Commercials

- Learning of tag co-occurrences

D.Borth: TubeTagger 14 December 6 2009
Semantic Concept Learning

Query Formulation

- concepts ≠ textual queries
- mapping of text queries to concept vocabulary
  - "computer" → Monitor, Windows-Desktop, iPhone
  - "funny" → Muppets, Cats, Commercials
- learning of tag co-occurrences
Semantic Concept Learning

Feature

- bag-of-words features
  - $t \in T \rightarrow h_t$
  - $q := h_q$
- mapping to concepts as weights
  - $w(q, t) := \langle h_t, h_q \rangle$
Semantic Concept Learning

Feature
- bag-of-words features
  - \( t \in T \rightarrow h_t \)
  - \( q := h_q \)
- mapping to concepts as weights
  - \( w(q, t) := \langle h_t, h_q \rangle \)

Approach
- \( T_5 = w(q, t)_{top5} \)
- final fusion
  - \( P(q|x) = \sum_{t \in T_5} \frac{w(q, t)}{\sum_{t \in T_5} w(q, t)} P(t|x) \)
Experiments
Experiments

Dataset

- 1200 hrs. (≈ 750k keyframes)
- 233 concepts
- per concept:
  - 150 videos for training
  - 50 videos for testing
Experiments

Dataset

- 1200 hrs. (= 750k keyframes)
- 233 concepts
- per concept:
  - 150 videos for training
  - 50 videos for testing

Clips/Concept Evaluation

- 10 concepts
- trained on $N$ clips
- tested on 200k keyframes
- saturation at 100 – 150 clips/concepts for SURF+PAMIR
Experiments

Feature & Classifier Evaluation

- 81 concepts
- Video level testing (avg. fusion)
- Results:
  - MAP: 22.4%
  - MAP: 15.4% (6 times faster)

<table>
<thead>
<tr>
<th>model</th>
<th>feature</th>
<th>SURF</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVMs</td>
<td></td>
<td>20.4</td>
<td>22.4</td>
</tr>
<tr>
<td>PAMIR</td>
<td></td>
<td>15.4</td>
<td>18.4</td>
</tr>
<tr>
<td>MAXENT</td>
<td></td>
<td>14.1</td>
<td>15.5</td>
</tr>
</tbody>
</table>
Experiments

Feature & Classifier Evaluation

- 81 concepts
- video level testing (avg. fusion)
- results:
  - MAP: **22.4%**
  - MAP: **15.4%** (6 times faster)

<table>
<thead>
<tr>
<th>model</th>
<th>feature</th>
<th>SURF</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVMs</td>
<td></td>
<td>20.4</td>
<td><strong>22.4</strong></td>
</tr>
<tr>
<td>PAMIR</td>
<td></td>
<td>15.4</td>
<td>18.4</td>
</tr>
<tr>
<td>MAXENT</td>
<td></td>
<td>14.1</td>
<td>15.5</td>
</tr>
</tbody>
</table>

Performance Distribution

![Bar chart showing performance distribution of concepts]

Showing 78 representative concepts out of 233
Insights in Web-based Concept Detection

Good Concepts
- "boat/ship", "pyramids"
- broad community of YouTube users
- often "interesting", "spectacular"

Redundant Concepts
- "drummer", "fencing"
- series of clips, not sufficiently diverse
- problem: generalization

Bad Concepts
- "fence", "operation-room"
- not regularly used as a tag
User Interface

1. selected a concept
2. or enter a query
3. switch between video or keyframe level
Deep Tagging

- frame-accurate concept detection beyond coarse tags
- find concept (e.g. "Christmas Tree") within a video clip
Text-based Search

▶ matches queries to concepts
▶ e.g. "diving" → "underwater", "shipwreck", "fish", ...
TubeTagger - YouTube-based Concept Detection

- utilize YouTube videos as training source for visual concept detector learning
- utilize YouTube tags for semantic model generation
- web demo providing:
  - deep tagging
  - tag recommendation
  - text-based search
questions?

project site: www.dfki.de/moonvid

web demo at: http://madm.dfki.de/demo/tubetagger/