Learning TRECVID'08 High-level Features from YouTube

Adrian Ulges*, Markus Koch,
Christian Schulze, Thomas M. Breuel

Image Understanding and Pattern Recognition
DFKI & TU Kaiserslautern / Germany

2008/07/07

Outline

Motivation

Online Video Concept Detection

TRECVID'08 Experiments

More Experiments

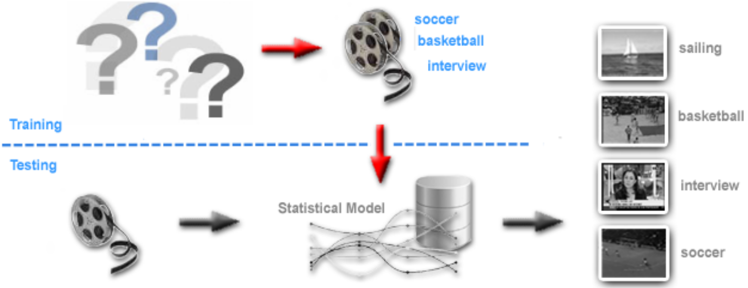
Discussion

Concept Detection

Detection of generic semantic concepts in video

- objects ("US flag"), locations ("desert"), events ("interview")
- main application: video search

Concept Detection



Key issue - training data acquisition

- training sets must be large-scale and annotated

Training Data: State-of-the-art

- high-quality manual annotations
- TRECVID [Smeaton06], Mediamill [Snoek06], LSCOM [naphade06], ...
- detectors exist for 100s of concepts

Training Data: State-of-the-art

- high-quality manual annotations
- TRECVID [Smeaton06], Mediamill [Snoek06], LSCOM [naphade06], ...
- detectors exist for 100s of concepts

Limitations

- need to scale up further
  ($1,000$s of concepts [Hauptmann07])
- annotations are bound to a dataset
- annotations are static

Outline

Motivation

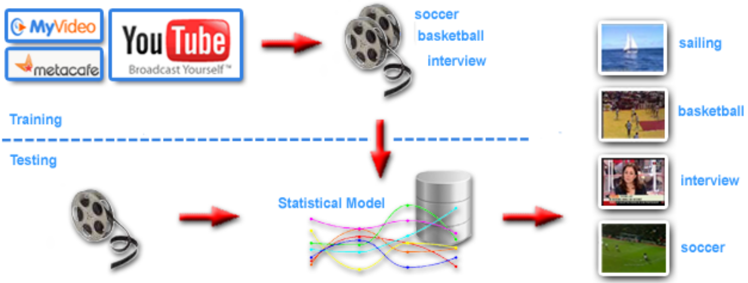
Online Video Concept Detection

TRECVID'08 Experiments

More Experiments

Discussion

Online Video Concept Detection



Idea: use **online video** as training data

- tags provided by users are used as annotations
- video taggers can learn autonomously

Online Video Concept Detection

Benefits

- scalability: can scale up to $1,000$s of concepts
- flexibility: web community keeps content up-to-date

Online Video Concept Detection

Benefits

- scalability: can scale up to $1,000$s of concepts
- flexibility: web community keeps content up-to-date

Problems

- web video is a mixture of domains with varying **production style** (TV news, home video, music clips, ...)
- annotations are **coarse** and **weak**
- (*for benchmarking*) potential **mismatch** between TRECVID and YouTube concepts.

YouTube

YouTube (filtered)

TRECVID

How Well Do Concept Detectors Trained on YouTube Work?

Key Idea

- use a standard concept detection approach (visual words + SVM)
- train it on YouTube and on a standard dataset (TRECVID-devel)
- benchmark both detectors

Experiments

1. participation in TRECVID'08
2. further experiments: TV05, TV07, YouTube

Outline

Motivation

Online Video Concept Detection

TRECVID'08 Experiments

More Experiments

Discussion

Approach

- Keyframe Extraction
  - adaptive clustering [Borth08]
- Features: Bag-of-visual-words
  - dense sampling over several scales (ca. $3,600$ features / frame)
  - SIFT descriptors
  - $2,000$-means clustering to codebook
- Classifier: SVMs
  - χ^2 kernel
  - cross-validation for γ and C maximizing avg. prec.
  - roughly balanced training sets (downsample negative class)
- Fusion over keyframes
  - simple averaging

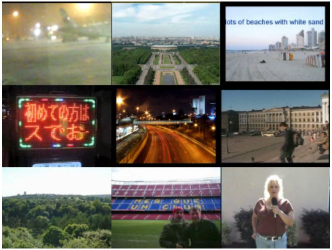
Datasets

- Test
  - standard TV'08 test data
- Training 1: TV'08
  - standard TV'08 training data
- Training 2: YouTube
  - downloaded using the YouTube API
  - 100 videos per concept of up to 3 min. length
  - two refinements:
    1. **by category**: mountain $\rightarrow$
       mountain[travel&places]
    2. **manually**: mountain[travel&places] $\rightarrow$
       mountain+panorama[travel&places]

YouTube Dataset: Quality

TRECVID | YouTube

mountain



cityscape

YouTube Dataset: Quality cont'd

TRECVID

YouTube

singing

telephone

Results 1

Top detections of YouTube-based detector

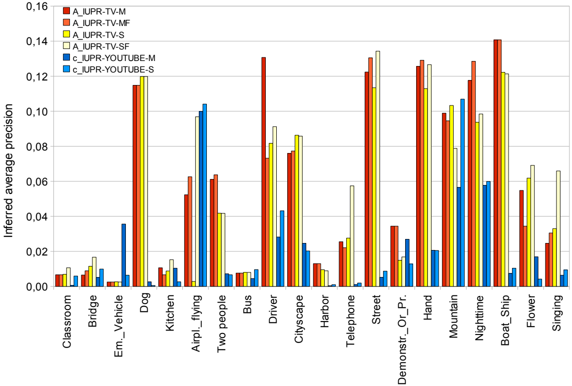
mountain

cityscape

singing

telephone

Results 2

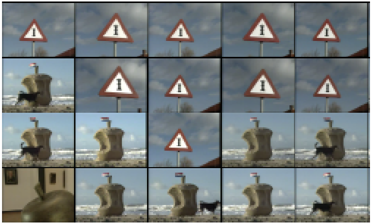


- infMAP for TRECVID runs: 5.3-6.3 %
- infMAP for YouTube runs: 2.1-2.2 %
- *performance strongly depends on the concept*

Results 3

Concept "Dog":

TRECVID training "dogs"



detected TRECVID test "dogs"

- specialized detectors make use of **duplicates** in the dataset
- the YouTube-based tagger cannot do this

if annotations on the target domain are given, specialized detectors outperform YouTube-based ones in terms of MAP. Influence of Duplicates?

Outline

Idea

Goal: Compare YouTube-based detectors with standard ones on a third target domain where no annotations are given!

- Approach / Concepts: *see last experiments*
- Datasets:
  1. **TV05**: TRECVID'05 video data with LSCOM annotations
  2. **TV07**: TRECVID'07 video data with TRECVID'08 annotations
  3. **YouTube**: *see last experiment*

Setup

- split each dataset for training and testing
- train on all datasets $\rightarrow$ 3 detectors
- **test each detector on all** 3 **datasets**

Results 1

MAP[%]			
training / testing	TV05	TV07	YOUTUBE
TV05	**18.40**	3.82	14.68
TV07	3.32	**9.65**	16.49
YOUTUBE	2.83	3.51	**31.33**

- specialized detectors always perform best! *(also for YouTube)*
- all detectors generalize poorly!
- in-depth analysis: duplicates in all datasets

Results 2

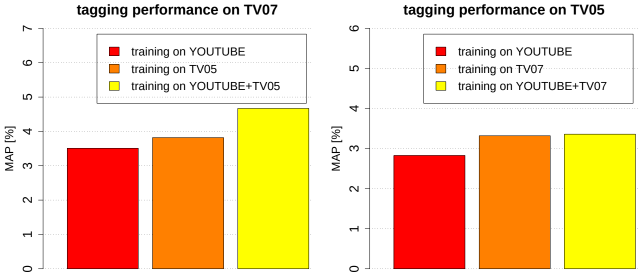
MAP[%]			
training / testing	TV05	TV07	YOUTUBE
TV05	18.40	**3.82**	14.68
TV07	**3.32**	9.65	16.49
YOUTUBE	**2.83**	**3.51**	31.33

- **the relative performance loss for the YouTube-based detector is moderate (11.4%)**

Results 3

Enhancing standard training sets with YouTube material

- join two datasets, test on third one



- **Combining training sets with YouTube material slightly increases generalization performance (11.7%)**

Outline

Motivation

Online Video Concept Detection

TRECVID'08 Experiments

More Experiments

Discussion

Conclusions

YouTube helps on domains with **no training annotations** when...

- ... **replacing** standard datasets (11.4% performance loss, but autonomous training)
- ... **complementing** standard datasets (11.7% increase in generalization capabilities)
- more: [TRECVID Notebook Paper], [adrian.ulges@dfki.de]

Conclusions

YouTube helps on domains with **no training annotations** when...

- ... **replacing** standard datasets (11.4% performance loss, but autonomous training)
- ... **complementing** standard datasets (11.7% increase in generalization capabilities)
- more: [TRECVID Notebook Paper], [adrian.ulges@dfki.de]

Issues

- Scaling to 1000 tags?
- Adapting YouTube-based detectors to other target domains?

Fine

Thanks for Your Attention!

(thanks also to Marcel Worring and Alexander Hauptmann for helpful discussions!)

References

- [Smeaton06]: A. Smeaton, P. Over, W. Kraaij. *Evaluation Campaigns and TRECVID*. MIR 2006.
- [Snoek06]: C. Snoek, M. Worring, J. van Gemert, J. Geusebroek, A. Smeulders. *The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia*. Multimedia 2006.
- [Naphade06]: M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis. *Large-Scale Concept Ontology for Multimedia*. IEEE Multimedia, 2006.
- [Hauptmann07]: A. Hauptmann, R. Yan, W. Lin. *How many High-Level Concepts will Fill the Semantic Gap in News Video Retrieval?*. CIVR, 2007.
- [Ulges08]: A. Ulges, C. Schulze, D. Keysers, T. Breuel. *A System that Learns to Tag Videos by Watching Youtube*. ICVS, Santorini, 2008.
- images taken from: [youtube,TRECVID datasets]