The science of labeling

- To understand anything in science, things have to have a name that is recognized and is universal

- naming chemical elements
- naming human genome
- naming 'categories'
- naming textual information
- naming rocks and minerals
- What about naming video information?
- naming living organisms
**Problem statement**

Different low-level features

Each feature yields a vector representation of the visual data

- Regularity
- Coarseness
- Directionality

**Concept-based Video Retrieval**

Cees G.M. Snoek and Marcel Worring

SSIP 2008

www.mediamill.nl
**Basic example: color histogram**

380 pixels

640 pixels

Total 243200 pixels

Histogram is a summary of the data summarizing in this case color characteristics

---

**Advanced example: codebook model**


---
The goal: semantic video indexing

- Is the process of automatically detecting the presence of a semantic concept in a video stream

Semantic indexing

- The computer vision approach
  - Building detectors one-at-the-time

One (or more) PhD for every new concept

Airplane
So how about these?

And the > 1000 others ……..

Generic concept detection in a nutshell

It is an aircraft probability 0.7
It is outdoor probability 0.95
K nearest neighbor

Linear classification
Support vector machine

SVM usually is a good choice

Support Vector Machine
- Learns from provided examples
- Maximizes margin between two classes

Supervised Learner
- Support Vector Machine
  - Learns from provided examples
  - Maximizes margin between two classes
- Depends on many parameters
  - Select best of multiple parameter combinations
  - Using cross validation

$\psi(w) = \frac{1}{1+\exp(-w(\overline{d}) + b)}$
How to improve concept detection?

- Feature Extraction
- Supervised Learner

Feature fusion: multimodal

- Vector concatenation & normalization
- + Only one learning phase
- + Truly a multimedia representation
- - Multimodal combination often ad hoc
- - One modality may dominate
- - Feature vectors become too large easily

References:
- Snoek, ACM Multimedia 2005
- Magelhaes, CIVR 2007
Feature fusion: unimodal

+ Codebook model reduces dimensionality
- Combination still ad hoc
- One feature may dominate

Classifier fusion: multimodal

+ Focus on modality strength
+ Fusion in semantic space
- Expensive in terms of learning effort
- Possible loss of feature space correlation
Classifier fusion: unimodal

- Aggregation functions reduce learning effort
- Offers opportunity to use all available examples efficiently
- Linear function likely to be sub-optimal

Modeling relations

- Exploitation of conceptual co-occurrence
  - Concepts do not occur in vacuum
  - In contrast, they are related

- What is sports?
  - Answer: a combination of various individual sports
Modeling relations

- Learning co-occurrence
  - Explicitly model relations: using graphical models
    - Computationally complex
    - Limited scalability
  - Implicitly learn relations: using SVM, or data mining tools
    - Assumes classifier learns relations
    - Suffers from error propagation

IBM’s pipeline

References: IBM 2003
Qi, ACM Multimedia 2007
Liu, IEEE TMM 2008
Naphade and Huang, TMM 3(1) 2001
IBM’s pipeline

Feature Fusion Classifier Relations

Annotation and Data Preparation

Feature Extraction

Classification Learner

Content Features Extraction

Semantic Features Combination

Malicious Features Extraction

Supervised Learner

Supervised Learner

Supervised Learner

Post-processing

Semantic Pathfinder

Content Analysis Step

Style Analysis Step

Context Analysis Step

Select Best of 3 Paths Validation

Hu Jintao

Weather news

Fire

Flag

Vehicle

Sports

Animal

News

Monologue

Entertainment

Weather

Weather

News

News

News
Semantic Pathfinder

Feature Fusion

Classifier Fusion

Modeling Relations

Context Analysis Step

Select Best of 3 Paths after Validation

Tsinghua University

Feature Extraction

SVN module

Concept Level Fusion

Concept Concept Level Fusion

China
Tsinghua University

Feature Fusion

Classifier

Modeling Relations

Fragmented research efforts...

Video analysis researchers
✓ Until 2001 everybody defined her or his own concepts
✓ Using specific and small data sets
✓ Hard to compare methodologies
Since 2001 worldwide evaluation by NIST
NIST TRECVID benchmark

Benchmark objectives
- Promote progress in video retrieval research
- Provide common dataset (shots, recognized speech, key frames)
- Use open, metrics-based evaluation

Large international field of participants
- Carnegie Mellon
- IBM
- TU Delft
- FX PAL
- MIT
- Microsoft
- and the 70 others...

Currently the de facto standard for evaluation

http://trecvid.nist.gov/

TRECVID Evolution: data, tasks, participants...

Source: Paul Over, NIST
Concept-based Video Retrieval
Cees G.M. Snoek and Marcel Worring

SSIP 2008

www.mediamill.nl

Concept detection task

- Given:
  - a video dataset segmented into set of \( S \) unique shots
  - set of \( N \) semantic concept definitions:

- Task:
  - How well can you detect the concepts?
  - Rank \( S \) based on presence of concept from \( N \)

Measuring uncertainty

Set of relevant items
Set of retrieved items

Results
1.
2.
3.
4.
5.

- Precision
- Recall

inverse relationship
TRECVID evaluation measures

- **Classification procedure**
  - Training: many hours of (partly) annotated video
  - Testing: many hours of *unseen* video

- **Evaluation measure: Average Precision**
  - Combines precision and recall
  - Averages precision after every relevant shot
  - Top of the ranked list most important

\[
AP = \frac{1/1 + 2/3 + 3/4 + \ldots}{\text{Total Number of correct shots}}
\]

Semantic Pathfinder @ TRECVID

With the MediaMill team

The Good
- With the MediaMill team

The Bad
- ill-defined / few examples

The Ugly
- exploit TV repetition
491 detectors, a closer look

The number of labeled image examples used at training time seems decisive in concept detector accuracy.

Demo time!
Concept detector: requires examples

- TRECVID’s collaborative research agenda has been pushing manual concept annotation efforts

![TRECVID-Driven Annotation Efforts]

Publicly Available

Concept definition

- MM078-Police/Security Personnel
  - Shots depicting law enforcement or private security agency personnel.
Collaborative annotation tool

- Manual annotation by 100+ TRECVID participants
  - Incomplete, but reliable

Manual annotations: LSCOM-lite

- LSCOM:
  - Large Scale Annotation for Multimedia
  - Aims for ontology of 1,000 annotated concepts
- LSCOM-Lite: annotations for 39 semantic concepts
  - Used in TRECVID 2005 and 2006
TRECVID Criticism

- Focus is on the final result
  - TRECVID judges relative merit of indexing methods
  - Ignores repeatability of intermediate analysis steps
- Systems are becoming more complex
  - Typically combining several features and learning methods
- Component-based optimization and comparison impossible

What is the contribution of these components?

MediaMill Challenge

- The Challenge provides
  - Manually annotated lexicon of 101 semantic concepts
  - Pre-computed low-level multimedia features
  - Trained classifier models
  - Five experiments
  - Baseline implementation together with baseline results

- The Challenge allows to
  - Gain insight in intermediate video analysis steps
  - Foster repeatability of experiments
  - Optimize video analysis systems on a component level
  - Compare and improve upon baseline

• The Challenge lowers threshold for novice multimedia researchers

Online available: http://www.mediamill.nl/challenge/
MediaMill Challenge

Advantages

- For research
  - People can focus on the experiment for which they have the expertise without having to do all the processing
    - Pure computer vision
    - Pure natural language processing
    - Pure machine learning
    - ……………………..
  
- For education
  - Students can do
    - large scale experiments
    - compare themselves to each other
    - …… and to the state-of-the-art

Columbia374

Baseline for 374 concept detectors

- Focus is on visual analysis experiments

Online available: http://www.ee.columbia.edu/ln/dvmm/columbia374/
Case study

Fabchannel.com

- Fabchannel narrowcasts concerts from Amsterdam Paradiso and Melkweg venues
  - Currently +/- 700 concerts online

- Fabchannel request
  - What can you do with 45 hours of live concerts?

- Answer:
  - Let’s try the semantic pathfinder to detect concert concepts

Results for singer
Results for drummer

Conclusions

- An international community is building a bridge to narrow the semantic gap
  - Currently detects more than 500 concepts in broadcast video
  - Generalizes outside news domain

- Important lessons
  - No superior method for all concepts exists,
  - Best to learn optimal approach per concept
  - Best methods cover variation in features, classifiers, and concepts
**Concept detection challenges**

- Show generality of approach over several domains
  - Show benefit of web-based image/video and annotations
- Show that concept classes work with less analysis
  - People, objects, setting
- Show benefit of using dynamic nature of video
  - Events
- Show that an ontology can help
  - How to connect logical relations to uncertain detectors?
- Show that ‘iconological’ concepts can be detected
  - E.g. funny, sarcastic, cozy, …

**Using concept detectors**

- “We are now seeing researchers starting to use the confidence values from concept detectors, within the shot retrieval process and this appears to be the roadmap for future work in this area.”
Measure concept detector influence

- Hypothesis 1:
  ✓ Increasing the number of concept detectors in a lexicon improves video retrieval accuracy.

- Hypothesis 2:
  ✓ Combining concept detectors from a lexicon improves video retrieval accuracy.

TRECVID automatic search task

- Automatically solve search topic
- Return 1,000 ranked shot-based results
- Evaluate using Average Precision

- TRECVID 2005
  ✓ 85 hrs test set – Chinese, Arabic, English TV News
  ✓ 24 search topics
**Topic examples**

- Find shots of one or more helicopters in flight.
- Find shots of a hockey rink with at least one of the nets fully visible from some point of view.
- Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people.
- Find shots of a group including at least four people dressed in suits, seated, and with at least one flag.

**Influence of lexicon size**

- **Lexicon** = 363 machine learned concept detectors

- **Procedure**
  1. Set bag size $B = 10$;
  2. Select random bag of $B$ detectors from lexicon
  3. Determine maximum performance for each search topic
  4. $B += 10$;
  5. Go back to step 2.

- **Repeat the process 100 times**
  - Reduces random positive and negative effects
Influence of lexicon size

- Size matters
  - Lexicon of 150 detectors comes close to maximum performance
- Some detectors perform well for specific topics
  - Tennis game detector for “find two visible tennis players”
- Substantial number of detectors not accurate enough yet
  - Only small increase when more than 70 detectors are used

Influence of detector combination

- How to combine multiple detectors?
  - Experiment: pair-wise oracle fusion

  \[
  \lambda \rightarrow \text{Office} \quad \text{Computer} \rightarrow 1 - \lambda
  \]

  - Improvement for 20 out of 24 topics
  - Increase per topic as high as 89%
  - Overall increase 10%
Typical results

Find shots of a graphic map of Iraq, location of Baghdad marked - not a weather map.

Best
Maps

2nd Best
Overlayed Text

How to select relevant detectors automatically?

Find shots of George Bush entering or leaving a vehicle (e.g., car, van, airplane, helicopter, etc) (he and vehicle both visible at the same time)

Best
rocket propelled grenades

2nd Best
Iyad Allawi

Problem statement

Find shots of an office setting

Topics

Query

Search Engine

Result

How to translate query topic to concept detectors?
Detector selection strategies

- Video Query
  - Find shots of an office setting

- Semantic Visual Querying

- Ontology Querying

- Fusion

Data flow conventions:
- Multimedia raw data
- Textual description
- Links to WordNet
- Visual model
- Concept detection result

Multimedia Thesaurus

Influence of detector selection combi

- Individual selection strategies seem comparable
  - But, oracle combination of selection strategies pays off!

Find shots of a tall building (with more than 5 floors above the ground)

Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people

Find shots of one or more palm trees.
TRECVID interactive search task

- So many choices for retrieval…
  - Why not let user decide interactively?

http://trecvid.nist.gov/

‘Classic’ Informedia system

- First multimodal video search engine
MediaMagic

- Focus on the story level

IBM MARVel

- A web based system

http://mp7.watson.ibm.com/marvel/
Cluster-temporal browsing

- Using that result are typically similar/close in time

Físchlár

- Optimized for use by "real" users
Concept-based Video Retrieval
Cees G.M. Snoek and Marcel Worring

VisionGo

- Extremely fast and efficient

References:
NUS & ICT-CAS

Extreme video retrieval

- Observation
  - Correct results are retrieved, but not optimally ranked
  - If user has time to scan results exhaustively, retrieval is a matter of watching, selecting, and sorting quickly

- Push the user to the max = very demanding!
  - Rapid-serial visual presentation
  - Adjust browser to depth of results

References:
Carnegie Mellon University
Futuristic video retrieval

Jonathan Wang, Carnegie Mellon University

CrossBrowsing through results

With the MediaMill team
ForkBrowser

Demo time!
NIST TRECVID benchmark

- Benchmark objectives
  - Promote progress in video retrieval research
  - Provide common dataset (shots, recognized speech, key frames)
  - Use open, metrics-based evaluation

- Large international field of participants

- Currently the de facto standard for evaluation

http://trecvid.nist.gov/

Experimental Setup

- Experiment 1
  - TRECVID 2004 (64 hrs test set – English TV News)
  - Lexicon with 32 learned concepts (where others use max. 10)
  - All other components “standard”

- Experiment 2
  - TRECVID 2005 (85 hrs test set – Chinese, Arabic, English TV News)
  - Lexicon with 101 learned concepts (where others use max. 39)
  - Added advanced display (CrossBrowser)
Concept-based Video Retrieval
Cees G.M. Snoek and Marcel Worrin

SSIP 2008

Lexicon indexing

Semantic Pathfinder

Select best of 3 paths after validation

Generic Video Indexing!
but performance varies...

Content Analysis Step

Monologue
Weather

Animals
Sports
Vehicle
Flag
Fire

Context Features Extraction

Supervised Learner

Multimodal Features Combination

Visual Features Extraction

Supervised Learning

Semantic Features Combination

Auditory Features Extraction

Supervised Learning

Visual Features Extraction

Supervised Learning

Textual Features Extraction

Content Analysis Step

Context Analysis Step

Style Analysis Step

Entertainment
Monologue

Learned lexicon of 32 concepts

TRECVID 2004

Animal
Football
Road
Beach
Stock Quotes
Golf
Financial Anchor
Cartoon

Building
Airplane Take Off
Boat
Graphic
People
Car
Vegetation
Overlaid Text

Basket Scored
Bill Clinton
Sporting Event
Studio Setting
Physical Violence
Train
Baseball
News Subject Monologue

Anchor
Outdoor
Ice Hockey
People Walking
Madeleine Albright
Soccer
Bicycle
Weather News

www.mediamill.nl

www.mediamill.nl 39
Query selection

... yields a ranking of the data

TRECVID 2004

Experiment 1

Lexicon = 32 concepts

The Good
The Bad
The Ugly

Exploit TV repetition
TRECVID 2005

Learned lexicon of 101 concepts

Lexicon = 101 concepts

The Good
Almost all topics solvable by using concept lexicon only!

The Bad
concept specification fails
concept combination fails

The Beautiful
Exploit common sense!
**Lexicon = 101 concepts**

**The Good**
Almost all topics solvable by using concept lexicon only!

**The Bad**

**The Beautiful**
Exploit common sense!

---

**TRECVID criticism**

- Need to realize that benchmark performance cannot be the sole criterion
  - Quality of concept detectors counts
  - Experience of searcher counts
  - Visualization of the interface counts
  - Ease of use counts
  - ...
**VideOlympics**

- Promote multiple facets of interactive video search
  - Real-time interactive video search ‘competition’
  - Simultaneous exposure of multiple video retrieval systems
  - Audience obtains complete overview of possibilities and limitations of state-of-the-art

**Participants**

**Showcase Event**

- Columbia University
- Carnegie Mellon
- National Institute of Informatics
- Intelligent Multimedia Group
- IBM
- MediaMill
- VIS Multimedia
Setup

Showcase Event

One display

TRECVID like queries
A result is submitted as soon as it is found

Video trailer

http://www.VideOlympics.org
Conclusions

- We should use
  - Generic semantic indexing
  - Large lexicons

- For browsing
  - The large lexicons are the best entrance
  - Ranking and time are most important dimensions

- TRECVID = priceless
  - Fosters a common research agenda and international collaboration
  - Many valuable resources available online
  - Offers you an opportunity to do it yourself easily

Concept retrieval challenges

- How to leverage concept detectors for search?
  - How to select the best detectors automatically?
  - How to combine concept selection methods?
  - How to balance semantic coverage and anticipated performance of detectors for a specific query?

- How to help the user in browsing
  - How to present detectors and their uncertainty to users?
  - How to select the best detector on the fly?
  - How to present thread space in an intuitive manner?

- Generalization to unstructured domains
  - Consider YouTube domain for example
  - How to use contextual metadata?
Further information

www.MediaMill.nl