UNITN
Deliverables Year 1
(D1.4.1, D2.2.1, D3.1.1, D3.2.1)
Trento is in the hearth of Europe
Venice: 2.5 hours by train

Rome: 5 hours by train
Trento
Living in Italy!!!!
Research Group

- M-HUG: Multimedia-assisted Human Understanding Group
  - 12 PhD students
  - 5 Researchers
Joining the protest!!!!

[Image of a protest with signs and participants]

- Support Vector Machines
- Never have demo-
- Country or we
Research Subjects

• Multimedia analysis
  – Visual-based Object Recognition (Pascal VOC, ImageNet) (with UvA)
  – Visual-based Distributional Semantics (with Marco Baroni – Cimec)
  – Automatic image annotation
  – (Ad-hoc) Multimedia Event Detection (with Alex Hauptmann - CMU)
  – Implicit video summaries using viewer’s affective information

• Behavior understanding (facial expressions, gaze estimation, affective state analysis)
The Big Picture

Detection → Tracking

Facial Expression → Head/Body Pose → Eye Location/Tracking → Visual Gaze

Applications
Object-based Image Classification

Given an image collection
Object-based Image Classification

Given an image collection: Find the objects containing a specific object
Object-based Image Classification

Problems:
- Viewpoint changes
- Location
- Illumination conditions
Object-based Image Classification

Problems:
- Viewpoint changes
- Location
- Illumination conditions
- Different manifestations
Object-based Image Classification: Bag-of-Words

- No segmentation
- No location

Sivic et al. 2003, Csurka et al. 2004
How Bag-of-Words classifies images

- Bag-of-Words works really on local details, although details are slightly larger than patches.

- Bag-of-Words uses details from both an object and its surroundings.

- Individual details are not very object or surrounding specific.

Uijlings, Smeulders, Scha, IJCV 2012
Segmentation as Selective Search

Need to bring back notion of object location

Sande, Uijlings, Gevers, Smeulders, ICCV 2011
Localisation: Exhaustive Search

~100,000-1,000,000 locations (huge computational constraints on subsequent methods)
Localisation: Segmentation

10-100 locations. But captures few objects.
Segmentation as Selective Search

Rethink segmentation:

- High Recall
- Coarse locations are sufficient (boxes)
- Fast to compute
An image is intrinsically hierarchical. A segmentation at a single scale cannot find all objects.
Segmentation as Selective Search

~Use all locations from a hierarchical grouping

Oversegmentation (Felzenszwalb 2004) → Hierarchical grouping of segments → Object hypotheses from all hierarchy levels
Segmentation as Selective Search

~ Hierarchical Grouping
~ Use of a variety of colour spaces with complementary invariance properties
~ Different grouping criteria: Colour, Texture, Size, Insideness
~ 2 methods:
  ~ Fast uses 8 different hierarchical groupings
  ~ Quality uses 80 different hierarchical groupings

Sande, Uijlings, Gevers, Smeulders, ICCV 2011
Evaluation of Locations

Pascal Overlap Criterion

Correctly localised if best overlap > 50%

Recall is the % of objects for which there is a location with > 50% overlap
Evaluation of Locations

What does a .88 Best Overlap score look like?

Overlap 88.4%  
Overlap 87.9%  
Overlap 87.4%
Selective Search in Object Localisation

Goal: To identify and find the location of the objects. An object is found if the Pascal Overlap score > 50%.
Selective Search in Object Localisation

Pascal VOC 2011
- Best results for 9 out of 20 object classes
- Works especially well on non-rigid object classes
- All competing methods are based on exhaustive search with HOG-features
Conclusions Selective Search

- Results in a small yet high quality set of potential object locations.
- Works by rethinking segmentation:
  - Focus more on Recall than Precision
  - Hierarchical grouping to deal with objects at multiple scales
  - Multiple complementary strategies to deal with high variety in image conditions
- Enables use of more expensive features
Automatic Image Annotation

- Exponential image increase requires effective management
- Images are represented by various features
- Feature selection seeks the discriminative subset by eliminating noise and redundancy, thus it can improve annotation for both classification accuracy and computational efficiency
Methodology

• Efficient feature analysis
  o Sparse feature selection
  o Semi-supervised via graph Laplacian

• Advantages:
  o Batch-mode: evaluating features jointly across all data points
  o Semi-supervised: not so expensive as supervised learning
    leverages both labeled and unlabeled data
    boosted performance when properly designed
Algorithm 1: The SFSS algorithm.

Input:
The training data $X \in \mathbb{R}^{d \times n}$;
The training data labels $Y \in \mathbb{R}^{n \times c}$;
Parameters $\mu$ and $\gamma$.

Output:
Converged $W \in \mathbb{R}^{d \times c}$.

1: Compute the graph Laplacian matrix $L \in \mathbb{R}^{n \times n}$;
2: Compute the decision rule matrix $U \in \mathbb{R}^{n \times n}$;
3: $H = I - \frac{1}{n}1_n1_n^T$;
4: $P = (L + U + \mu H)^{-1}$;
5: $A = XH(\mu I - \mu^2 P)HX^T$;
6: $B = \mu XHPUY$;
7: Set $t = 0$ and initialize $W_0 \in \mathbb{R}^{d \times c}$ randomly;
8: repeat
   Compute the diagonal matrix $D_t$ as:
   $$D_t = \begin{bmatrix}
   2 \|w_t^1\|_2 \\
   \vdots \\
   2 \|w_t^d\|_2
   \end{bmatrix};$$
   Update $W_{t+1}$ as: $W_{t+1} = (D_t A + \gamma I)^{-1} D_t B$;
   $t = t + 1$.
until Convergence;
9: Return $W$. 
## Experiments

### Datasets Description

<table>
<thead>
<tr>
<th></th>
<th>Corel-5K</th>
<th>MSRA-MM 2.0</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class Number</strong></td>
<td>50</td>
<td>100</td>
<td>81</td>
</tr>
<tr>
<td><strong>Training Set Size</strong></td>
<td>2500</td>
<td>10,000</td>
<td>10,000</td>
</tr>
<tr>
<td><strong>Testing Set Size</strong></td>
<td>2500</td>
<td>32,266</td>
<td>199,347</td>
</tr>
</tbody>
</table>

- **Corel-5K**: real-world images from COREL image CDs, single labeled
- **MSRA-MM 2.0**: collected from the web, diverse, multi-labeled, real-world
- **NUS-WIDE**: all images are from Flickr, diverse, multi-labeled, real-world, large scale
Experiments

Results comparison – all cases

(a) Corel-5K  (b) MSRA-MM  (c) NUS-WIDE
Experiments

• Labeled training samples improve performance
• Our method has consistently high scores on all three datasets.
• When > 25% training data are labeled, we are competitive with the best supervised algorithms
• When < 25% of the data are labeled, we consistently outperform other supervised methods on all three datasets
• Our advantage is especially visible when there are only few labeled training data, which is desirable for real world scenarios
KNOWLEDGE ADAPTATION FOR AD HOC MULTIMEDIA EVENT DETECTION WITH FEW EXEMPLARS

Zhigang Ma, Yi Yang, Yang Cai, Nicu Sebe and Alexander G. Hauptmann
An important topic

• Multimedia Event Detection:
  To detect the occurrence of an event within a video clip based on an Event Kit, which contains some text description and some example videos.

  National Institute of Standards and Technology
the abstract or general idea inferred from specific instances; usually describable by a single shot
Annotation vs Detection

To detect the existence of concept/event through pre-trained detectors

To associate multimedia content with semantic labels (tags)
The Reality

How to get respectable performance for un-predefined events with few positive examples

Ad Hoc MED
Inspiration – Knowledge Adaptation

- Why?
  The information from few positive examples is limited.

- So?
  We propose to borrow strength from other multimedia resources.

- How?
  Concepts-based videos are used as auxiliary resource since many labeled video corpora are available, and events are usually related to certain concepts.
Formulation

landing a fish

sparse model
Auxiliary Concept-based Videos

Target Training Videos

Target Testing Videos

feature extraction

feature extraction

feature extraction

Hilbert Space

New representation of auxiliary videos

Labels of auxiliary videos

New representation of target training videos

Labels of target training videos

New representation of target testing videos

Concept classifier

Event detector

Shared components mining

Knowledge Adaptation

Optimizing by minimizing the negative impact of irrelevance and noise jointly

Optimized event detector

Decision values
Objective function

\[
\min_{w_a, w_t, b_a, b_t} \left( \left\| \tilde{X}_a^T W_a + 1_a b_a - Y_a \right\|_{2,1} + \left\| \tilde{X}_t^T W_t + 1_t b_t - y_t \right\|_{2,1} + \alpha \left\| W_a, W_t \right\|_{2,\rho} + \beta (\left\| W_a \right\|_F^2 + \left\| W_t \right\|_F^2) \right)
\]

Algorithm 1: Optimizing the event detector.

Input:
- The auxiliary data \( \tilde{X}_a \in \mathbb{R}^{d_a \times n_a}, Y_a \in \mathbb{R}^{n_a \times c_a} \);
- The target training data \( \tilde{X}_t \in \mathbb{R}^{d_t \times n_t}, y_t \in \mathbb{R}^{n_t \times 1} \);
- Parameters \( \alpha \) and \( \beta \).

Output:
- Optimized \( W_t \in \mathbb{R}^{d_t \times 1} \) and \( b_t \in \mathbb{R}^1 \).

1: Set \( t = 0 \), initialize \( W_a \in \mathbb{R}^{d_a \times c_a} \) and \( W_t \in \mathbb{R}^{d_t \times 1} \) randomly;
2: repeat
   1. Compute \( \tilde{X}_a^T W_a - Y_a = [w_1^T, \ldots, w_{n_a}^T]^T \), \( \tilde{X}_t^T W_t - y_t = [v_1^T, \ldots, v_{n_t}^T]^T \), and \( W = [w_1^T, \ldots, w_{d_t}^T]^T \);
   2. Compute the diagonal matrix \( D_a^w, D_t^v \) and \( D_t^y \) according to \( D_a^w = \frac{1}{\|w_i\|_2^2}, D_t^v = \frac{1}{\|v_i\|_2^2} \), and \( D_t^y = \frac{1}{\|y_i\|_2^2} \) respectively;
   3. Update \( W_t^{t+1} \) as:
      \( W_t^{t+1} = (\tilde{X}_a H_a D_a \tilde{X}_a^T + \alpha D + \beta I_d)^{-1} \tilde{X}_a H_a D_a H_a Y_a^T \);
   4. Update \( b_t^{t+1} \) as:
      \( b_t^{t+1} = \frac{1}{n_t} 1_T Y_t - \frac{1}{n_t} 1_T \tilde{X}_t^T W_t^{t+1} \);
   5. Update \( W_t^{t+1} \) as:
      \( W_t^{t+1} = (\tilde{X}_t H_t D_t \tilde{X}_t^T + \alpha D + \beta I_d)^{-1} \tilde{X}_t H_t D_t H_t y_t^T \);
   6. Update \( b_t^{t+1} \) as:
      \( b_t^{t+1} = \frac{1}{n_t} 1_T y_t - \frac{1}{n_t} 1_T \tilde{X}_t^T W_t^{t+1} \);
   7. \( t = t + 1 \),
   until Convergence;
3: Return \( W_t \) and \( b_t \).
Experiments

- Datasets:
  - TRECVID MED 2010
  - TRECVID MED 2011 development set
  - TRECVID 2011 semantic indexing task development set.
    - auxiliary videos
    - concepts with few positive examples are removed
    - 65 concepts related to human, environment and objects
    - 2529 video frames.
  - 9822 video clips
Mission Events

Making a cake
Batting a run
Assembling a shelter
Mission Events

- Attempting a board trick
- Feeding an animal
- Working on a woodworking project
- Landing a fish
- Wedding ceremony
Mission Events

- Changing a vehicle tire
- Birthday party
- Getting a vehicle unstuck
- Flash mob gathering
- Grooming an animal
Mission Events

Parade

Making a sandwich

Working on a sewing project

Parkour

Repairing an appliance
Experiments

• Setup:
  - Data representation: SIFT + CSIFT
  - Training: 10 positive example + 300 negative examples
  - Evaluation metrics:
    (1) Minimum NDC (MinNDC)
    (2) Probability of Miss-Detection based on the Detection Threshold 12.5 (Pmd@TER=12.5)
    (3) Average Precision (AP)
Experiments

(a) Notations

- $\chi^2$-SAR
- G-SAR
- $\chi^2$-SVM
- G-SVM

MinNDC: Minimum NDC
The LOWER, the BETTER.

Pmd: $\text{Pmd@TER}_{12.5}$
The LOWER, the BETTER.

AP: Average Precision
The HIGHER, the BETTER.

(b) Attempting a board trick

(c) Feeding an animal

(d) Landing a fish

(e) Wedding ceremony
Experiments

(a) Notations

MinNDC: Minimum NDC
The LOWER, the BETTER.

Pmd: Pmd@ TER= 12.5
The LOWER, the BETTER.

AP: Average Precision
The HIGHER, the BETTER.

(b) Batting a run

(c) Average

(s) Assembling a shelter
Conclusion

• First attempt on Ad Hoc MED
• More generic, complicated and meaningful events
• Only few positive examples
• Knowledge adaptation from concepts-based videos
• Assumption: shared structures
• Future direction: how to judge the structural commonness
Automatic Assessment of Video Footage
Automatic Assessment of Video Footage

Status: Face Found!

Instructions:
Press "Emotion" to analyse the face.
Footage Assessment
Affective Cinema

- Adaptive System
- Understands emotion of the viewer
- Chooses the suitable way of narration
Different possible paths for the movie to be played chosen implicitly based on the user affective feeling
xLiMe?

Hi I'm Abbey

Play video
Role in xLiMe

• WP3 leader: Cross-lingual Multimedia Semantic Annotation
• WP8 leader: Dissemination, Exploitation, and Community Building
• Task leader:
  – T1.4 Functional Requirements Analysis
  – T2.2 Extraction of Text from Video
  – T3.1 Audio Annotation (objects and emotions)
  – T3.2 Video Annotation (objects and emotions)
Deliverables (Year 1)

- D1.4.1 Requirements for Early Prototype (M3)

  “This deliverable will provide functional specifications for the early prototype, using the pre-existing technologies identified in T1.2 and the technology [to be] developed in the first year of the project”
  - gather requirements based on the case studies
  - integrate a solution with the business of the companies
  - define a set of possible services built on top of the xLiMe technology to be used within case studies
  - define relevant languages and media types for the case studies
Deliverables (Year 1)

• D2.2.1 Early Text from Video Prototype (M12)
  “This deliverable will provide a first version of the video OCR system that will be robust to the size, colour, and orientation of the text”
  – gather requirements based on the case studies
  – take an unsupervised approach based on character energy computed using the similarity of stroke edges and link energy
  – will the text be used for matching? Augmenting the existing metadata?
Deliverables (Year 1)

• D3.1.1 Early Prototype for Audio Annotation (M12)
  “This deliverable will provide the first implementation of the audio annotation prototype including the annotation of affect. The usage of simple lists or, respectively, rules for mapping features to emotions will be evaluated”
  – gather requirements based on the case studies
  – define emotional features in terms of arousal and valence levels
  – extract low-level audio features from the sources available and map these to arousal and valence levels of the associated audio streams
  – Support personalization (like vs. dislike)
Deliverables (Year 1)

• D3.2.1 Early Prototype for Video Annotation (M12)
  “This deliverable will provide the first implementation of the video annotation prototype dealing with the features that are specific to affective visual content classification.”
  – gather requirements based on the case studies
  – investigate the features that are specific to affective visual content classification
  – build a large-scale ontology of semantic concepts reflecting strong positive and negative sentiments that complement a textual sentiment dictionary [built on top of the one developed in the RENDER project?]
  – predict sentiment directly from the visual content