Indoor Location Estimation using a Wearable Camera

with Application to the Monitoring of Persons at Home

Vladislavs DOVGALECS

Composition of jury

Atilla BASKURT
François BREMOND
Matthieu CORD
Jenny BENOIS-PINEAU
Yannick BERTHOUMIEU
Rémi MÉGRET

Context: *Wearable Sensors*

"Lifelogging is the process of tracking personal data generated by our own behavioral activities."

Sports performance evaluation

SenseWear
Health monitoring

Zeo sleep logger

Wearable sensors

Fall detection

... and many more!

Wearable Sensors: *Visual Lifelogging*

- Memory aid for recalling daily activities
- Passive creation of visual diaries
- Personal security applications

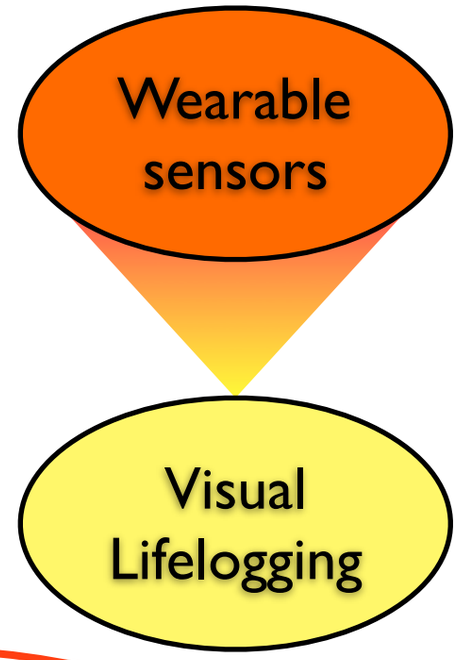


Image lifelogging

ViconRevue

Sousveillance
ExisTech

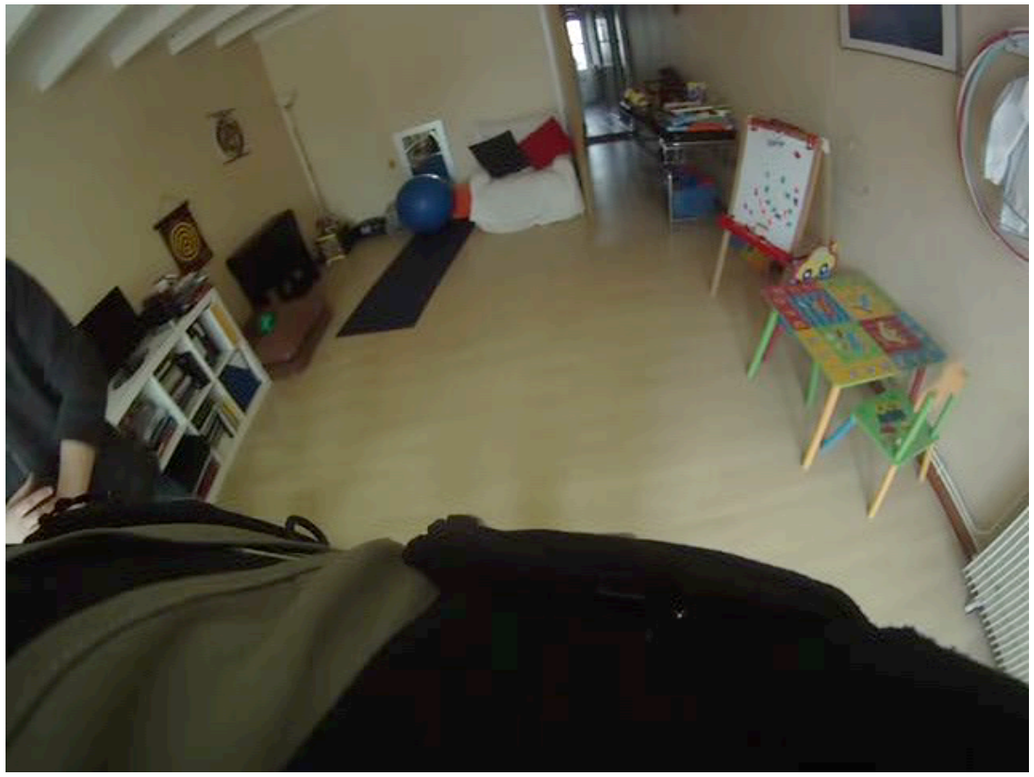
Video lifelogging

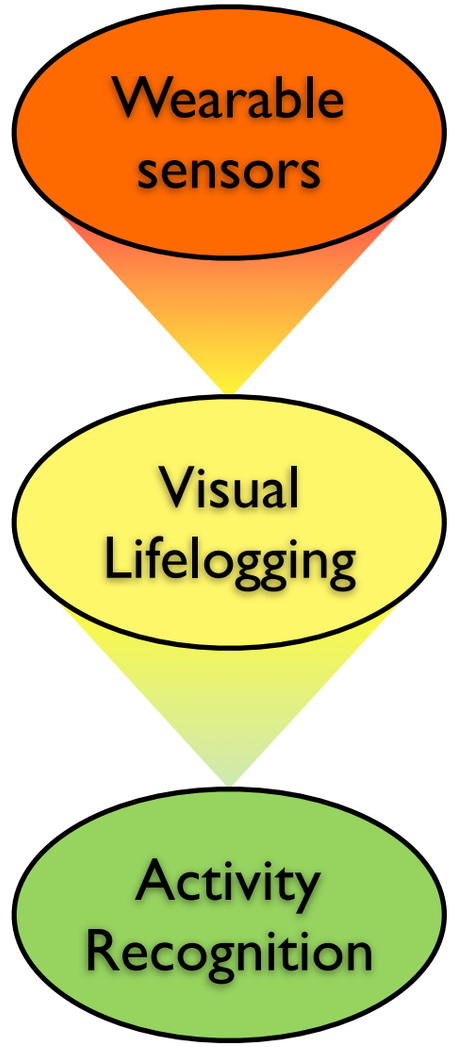
ZionEyez

Looxcie

Wearable Video Monitoring

Light and autonomous
video recorder





Recording activities of a patient indoors

Device developed in the IMMED project

Application to the IMMED project

1

Recording session

Video & Sound

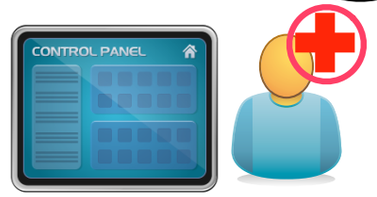
2

Annotation

3

Automatic indexing algorithms

Indexed content



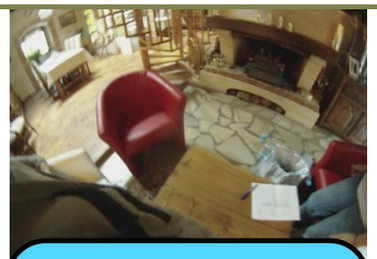
4

Indexed content browsing

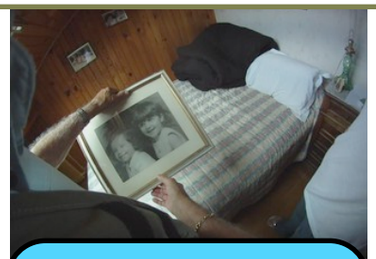
Analysis of results

Video Indexing problem

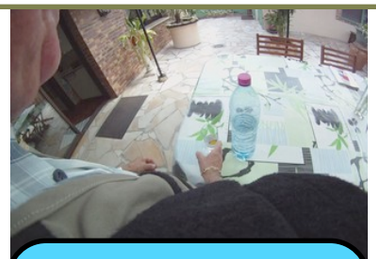
Manually annotated **bootstrap video** with topological location tags



Lobby



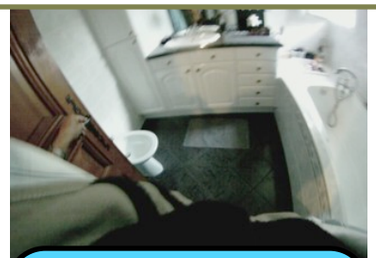
Bedroom



Saloon

Unlabeled video with unknown topological location tags

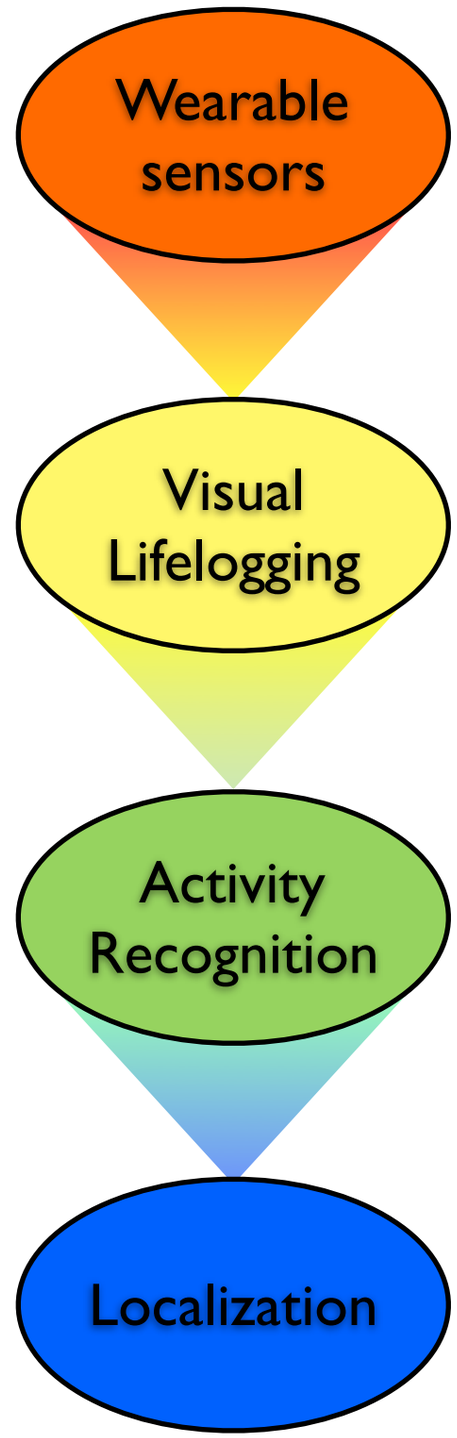
???



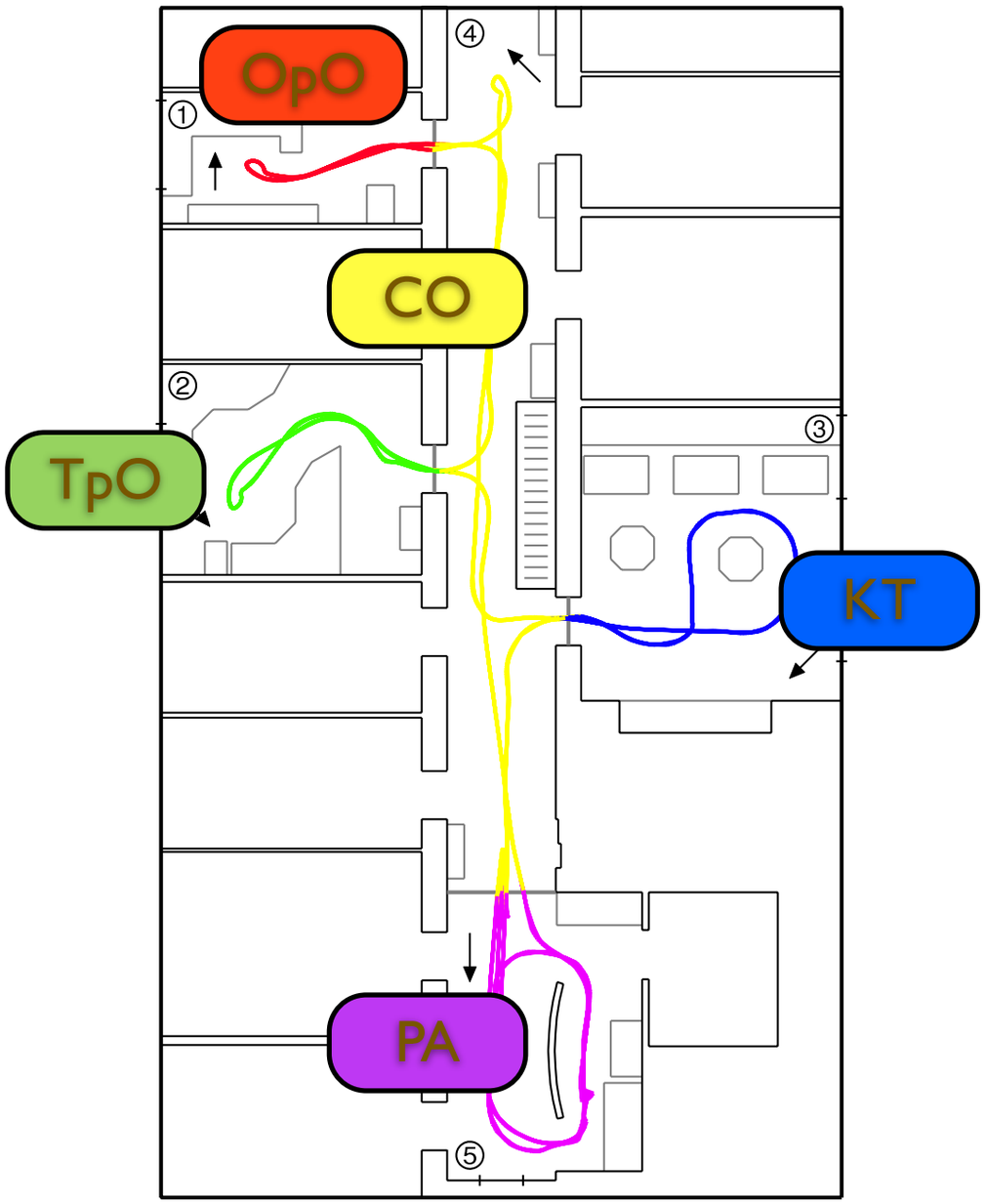
???

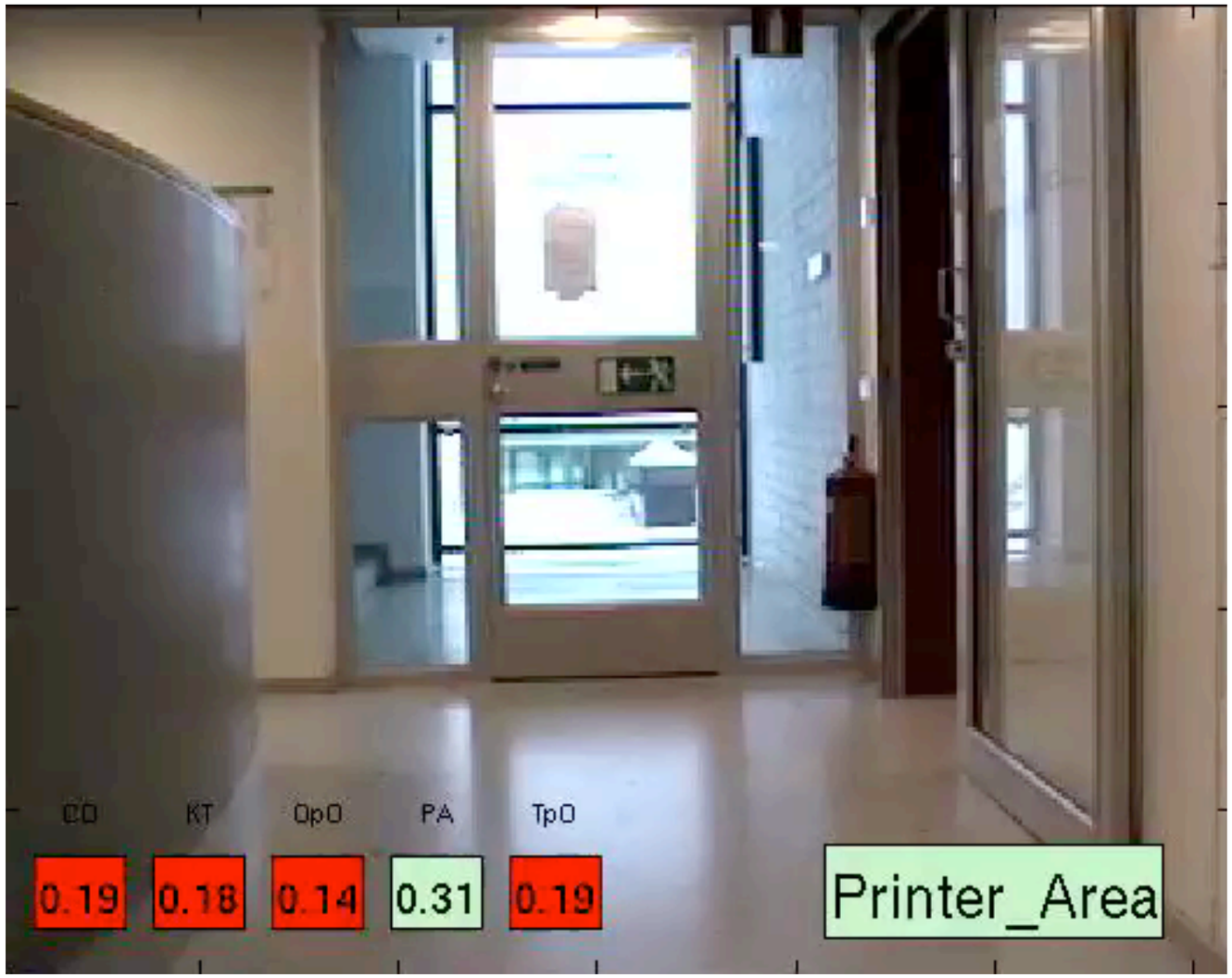
???

QUESTION: How to estimate topological locations?



Indoors Location Estimation





Automatically annotated sample video

J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "**The KTH-IDOL2 Database**". Technical Report CVAP304, 2006.

Challenges

Deal with great variability and complexity of visual content

Work with small amounts of manually annotated videos

Leveraging additional information about the problem

Contributions

Answering the Challenges

Deal with great variability and complexity of visual content

1. Relevant visual information extraction
2. Multiple information sources utilization
   Comparing Early and Late information fusion strategies for topological localization

Work with small amounts of manually annotated videos

1. Utility of unlabeled image data
   Study of Semi-Supervised methods for topological localization
2. Exploiting temporal continuity and unified framework
   Proposition of temporal accumulation schemes
3. Exploiting invariance to spatial transformations

Overview

Baseline Location Estimation
Visual Feature Extraction and Localization

Improving Discrimination Power
Multiple Information Sources

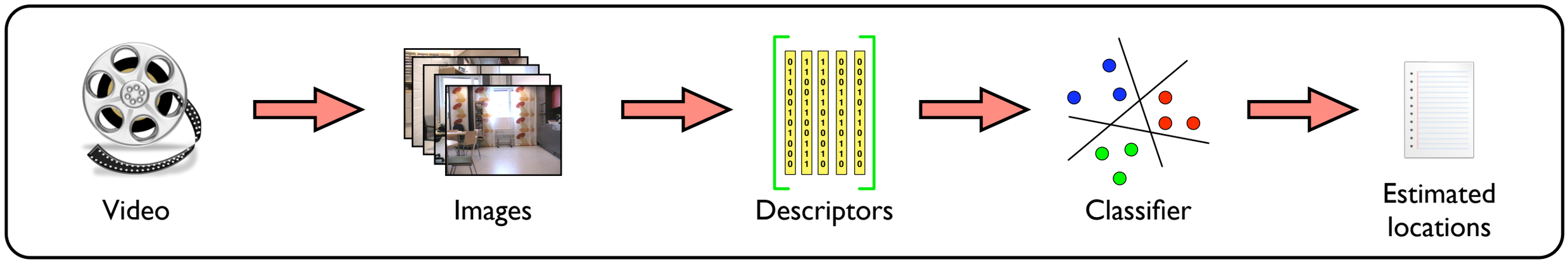
Main Proposition
Time-Aware Co-Training Framework

Other Prior Information
Invariance

Experimental Results

Baseline Location Estimation

Visual Feature Extraction and Localization



From video to descriptors

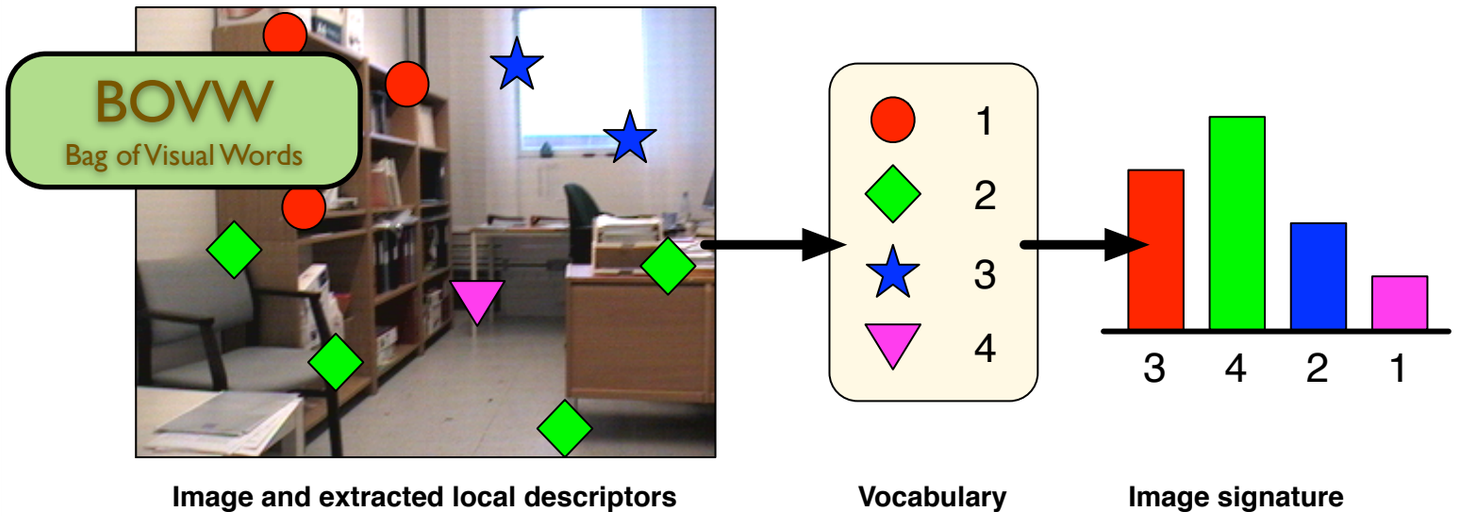
Video Extracted Frames Global Descriptors

Global descriptors characterize or describe image contents

Ideally, descriptors for the same topological location should be the same

Descriptors

Three Visual Features

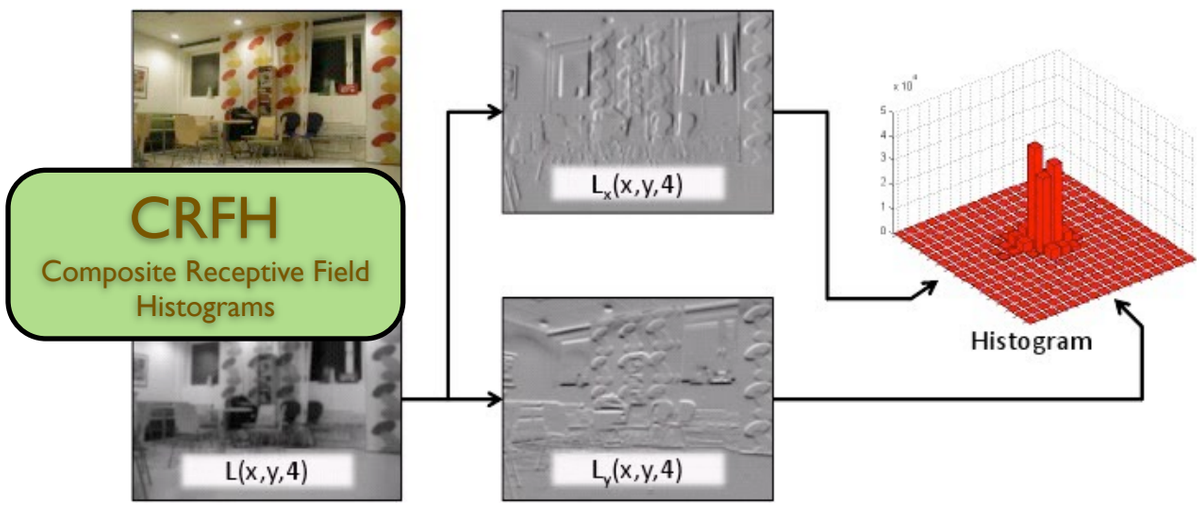


Extracting local features (e.g. SIFT or SURF)

Quantizing local features in discrete classes

$$\mathbf{x}_{\mathrm{BOVW}} \in \mathbb{R}^{1111}$$

[Csurka2004]

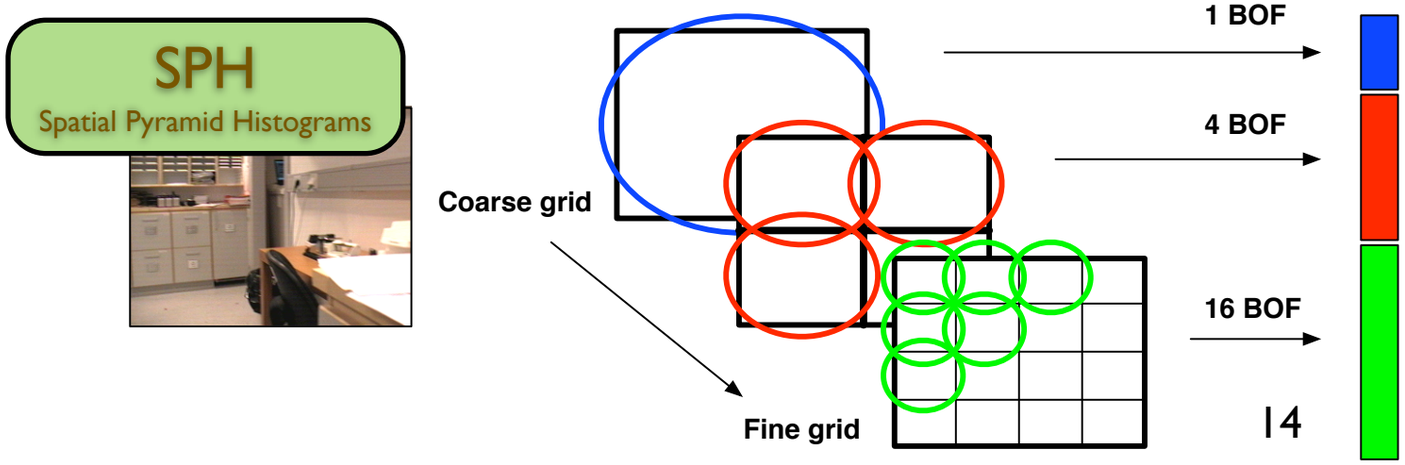


Multi-dimensional representation of an image counting occurrences of filter responses

Histogram counts the number of pixels sharing the same response

$$\mathbf{x}_{\mathrm{CRFH}} \in \mathbb{R}^{300'000'000}$$

[Linde2004]



Concatenation of BOVW histograms for different grids

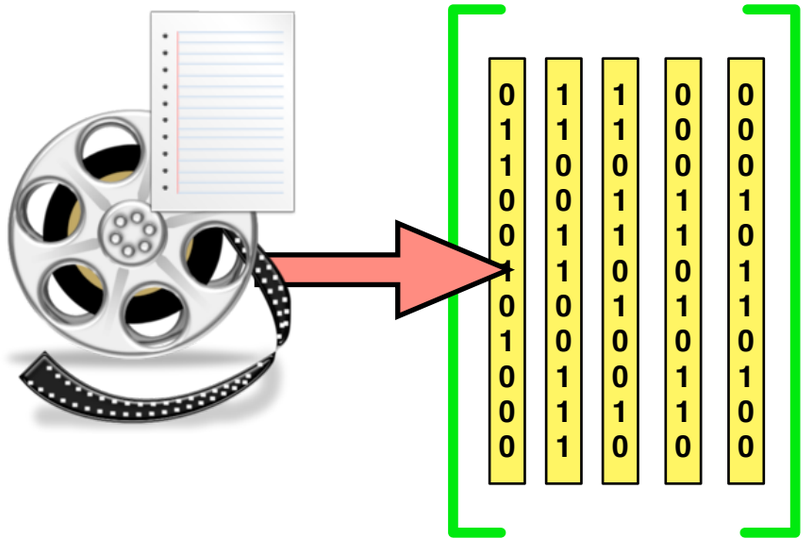
Captures spatial information which is missing in BOVW

$$\mathbf{x}_{\mathrm{SPH}} \in \mathbb{R}^{4'200}$$

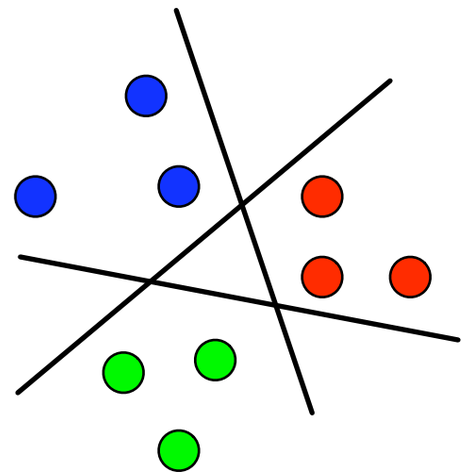
[Lazebnik2006]

Generic location estimation

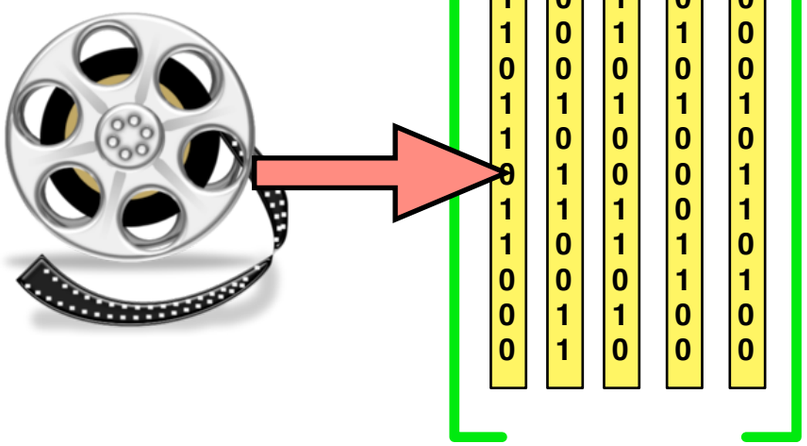
Bootstrap
video
+
annotation



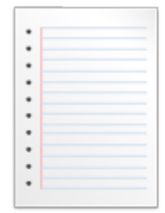
Train a
model



Unlabeled
video



Query the
model



Estimated
locations

Similarity Measure

Notion of a Kernel

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$

$\mathbf{x}$ - a descriptor or pattern

Notion of similarity

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

$$(\mathbf{x}_i, \mathbf{x}_j) \mapsto k(\mathbf{x}_i, \mathbf{x}_j)$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

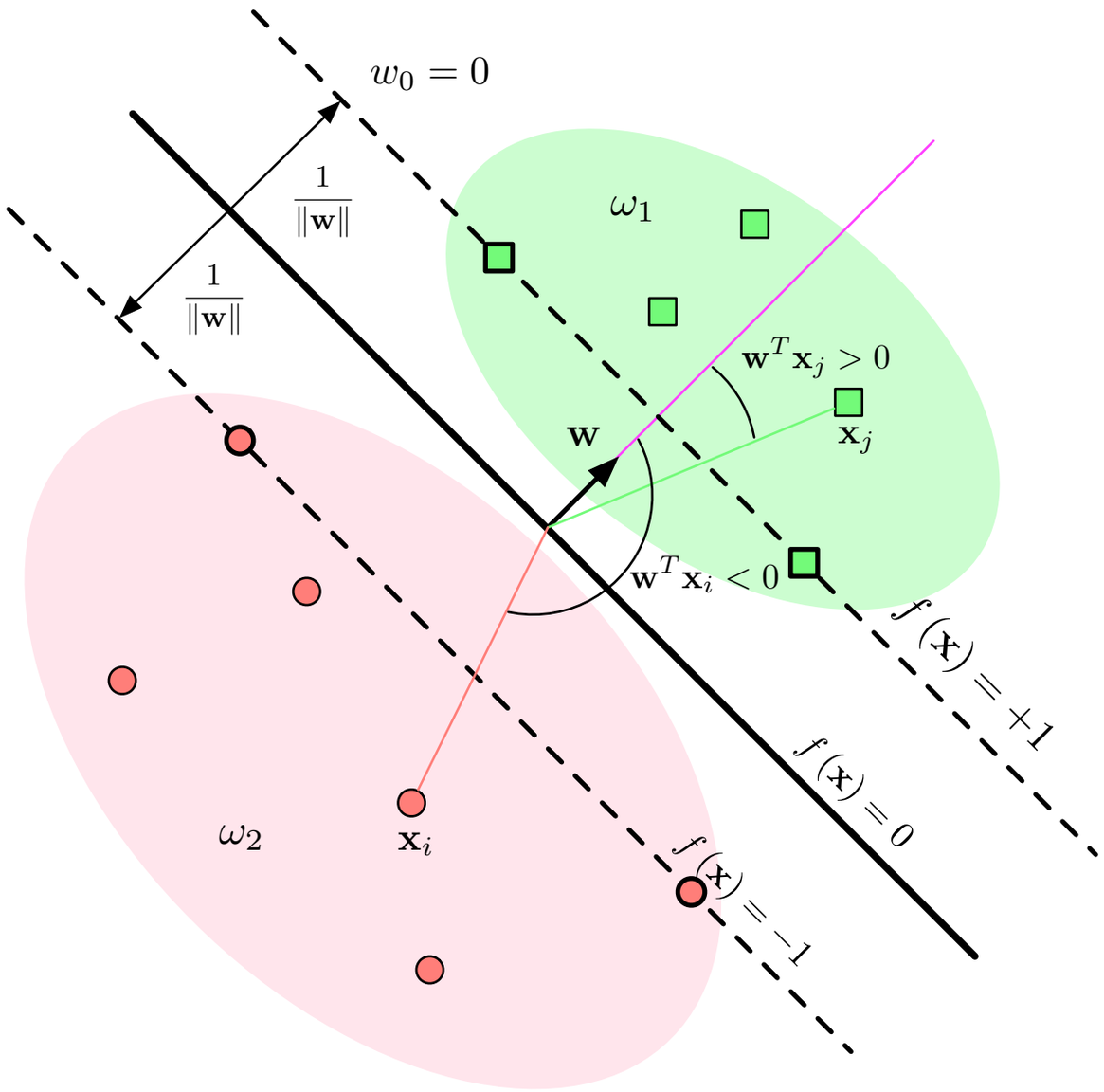
Dot product

$$k(\mathbf{x}_i, \mathbf{x}_j) = 1 - 2 \sum_{k=1}^{n} \frac{(\mathbf{x}_i[k] - \mathbf{x}_j[k])^2}{(\mathbf{x}_i[k] + \mathbf{x}_j[k])}$$

Chi-Square - adapted measure

Classification - Linear Support Vector Machines

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R} \times \{\pm 1\}$$



For linearly separable data

Idea
Maximum margin classifier generalizes the best for new data

Find a function

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

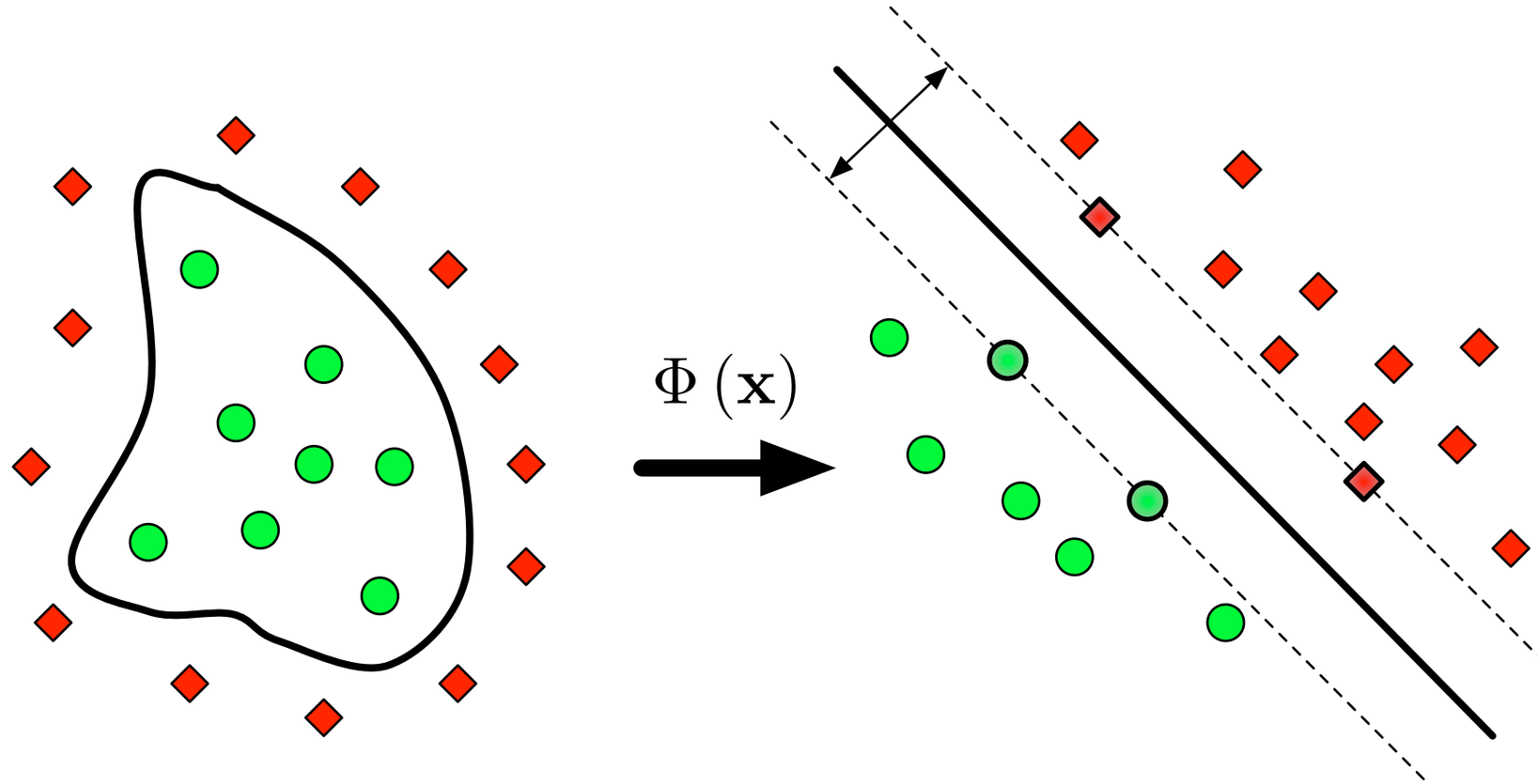
Estimations

$$\hat{y} = \begin{cases} +1, & \text{if } f(\mathbf{x}) > 0 \\ -1, & \text{if } f(\mathbf{x}) < 0 \end{cases}$$

$$\max_{\mathbf{w}} \|\mathbf{w}\|^2$$

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 0, \forall i$$

Classification - Non-Linear SVM



descriptor space

feature space

Decision function
in descriptor space

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$$

Decision function
in feature space

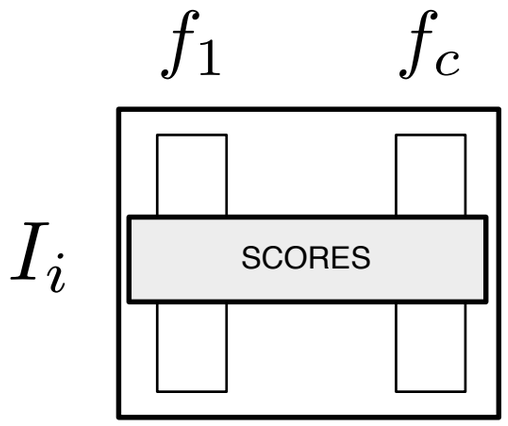
$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle + b$$

Multi-class Classification

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$$

SVM is a binary classifier

One-vs-All approach for multiple class classification



1. Train "c" binary classifiers

$$f_1, \ldots, f_c$$

2. Compute the scores for a test pattern

$$\mathbf{s}_i = [f_1(\mathbf{x}_i), \ldots, f_c(\mathbf{x}_i)]$$

3. Assign a class of largest score value

$$\hat{y}_i = \arg \max_{j=1,\ldots,c} f_j(\mathbf{x}_i)$$

I_i - *i*th frame of the video

Test Database (IDOL2)

Printer Area

Corridor

Two Person Office

One Person Office

Kitchen

Recorded using a mobile robot platform

Five topological locations

Three lighting conditions

Half of the videos recorded across a span of 6 weeks

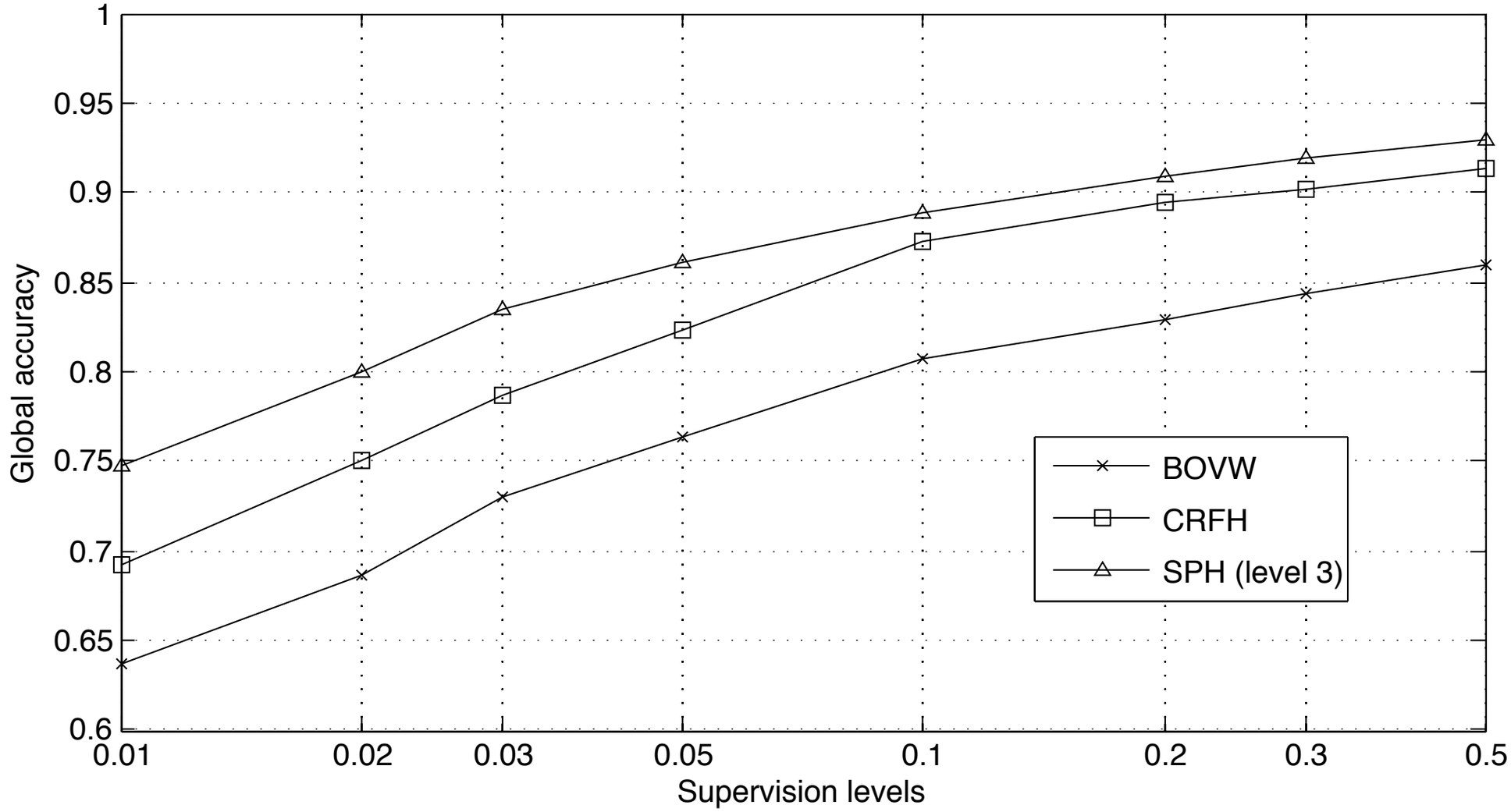
Cloudy
4 videos

Sunny
4 videos

Night
4 videos

Baseline approach consider the whole database as an orderless set of images

Baseline evaluation results



➡Varying performance of different visual features

➡Need for annotated data at low supervision levels

Outline

Principal Contributions

Multiple information sources

Leveraging unlabeled data

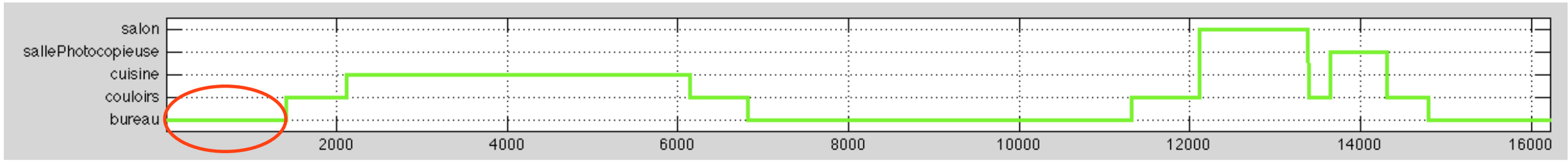
Integrating temporal information

Improving a visual descriptor

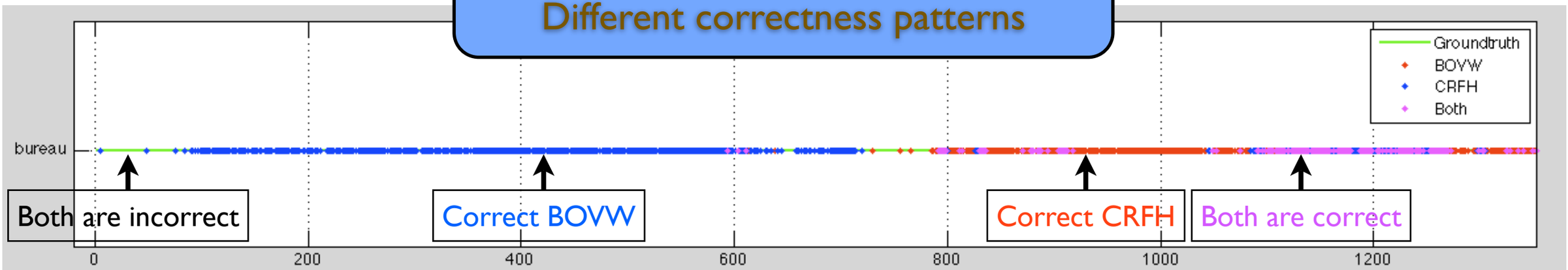
Improving Discrimination Power

Multiple Feature Fusion

Complementarity of visual descriptors



Different correctness patterns



BOVW and CRFH features are complementary

Can we leverage the discrimination power of multiple features?

Two approaches

Multiple Kernel Learning
Learning a New Kernel

Late Classifier Fusion
Fusion at classifier level

Learned similarity measure

S. Sonnenburg et al., "**A general and efficient multiple kernel learning algorithm**," *Advances in Neural Information Processing Systems*, vol. 18, p. 1273, 2006.

Combining multiple classifiers

M. Antenreiter et al., "**Combining Classifiers for Improved Multilabel Image Classification**," *European Conference on Machine Learning*, 2009.

Early Fusion : *Multiple Kernel Learning*

Images

$$k_1 = k\left(\mathbf{x}_i^{\mathrm{BOVW}}, \mathbf{x}_j^{\mathrm{BOVW}}\right)$$

$$k_2 = k\left(\mathbf{x}_i^{\mathrm{CRFH}}, \mathbf{x}_j^{\mathrm{CRFH}}\right)$$

$$k_3 = k\left(\mathbf{x}_i^{\mathrm{SPH}}, \mathbf{x}_j^{\mathrm{SPH}}\right)$$

Learning

k_{MKL}

New Kernel

New kernel : Sum of *m* individual kernels

$$k_{\mathrm{MKL}}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum_{\ell=1}^{m} \beta_\ell k_\ell\left(\mathbf{x}_i, \mathbf{x}_j\right) \qquad \beta_k \text{ - } k\text{th Kernel weight}$$

Idea

Weight an individual kernel w.r.t. its importance

Dual SVM Objective

$$\max_{\alpha,\beta} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \underbrace{\sum_{\ell=1}^{m} \beta_\ell k_\ell\left(\mathbf{x}_i, \mathbf{x}_j\right)}_{k_{\mathrm{MKL}}}$$

Subject To

$$\sum_{i=1}^{n} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1 \ldots, n$$

$$\sum_{\ell=1}^{m} \beta_\ell = 1, \beta_\ell \geq 0, \ell = 1, \ldots, m$$

A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "**SimpleMKL**," Journal of Machine Learning Research, vol. 9, pp. 2491–2521, 2008.

Late Fusion : *Classifier fusion*

Images

Train *m* independent multi-class SVM classifiers

$$\mathbf{s}_i^1 = \left[f_1^1\left(\mathbf{x}_i\right), \ldots, f_c^1\left(\mathbf{x}_i\right)\right] \quad \cdots \quad \mathbf{s}_i^m = \left[f_1^m\left(\mathbf{x}_i\right), \ldots, f_c^m\left(\mathbf{x}_i\right)\right]$$

$$f_j^{\mathrm{DAS}}\left(\mathbf{x}_i\right) = \sum_{k=1}^{m} \beta_k f_j^k\left(\mathbf{x}_i\right)$$

$$\mathbf{s}_i^{\mathrm{DAS}} = \left[f_1^{\mathrm{DAS}}\left(\mathbf{x}_i\right), \ldots, f_c^{\mathrm{DAS}}\left(\mathbf{x}_i\right)\right]$$

c - # classes

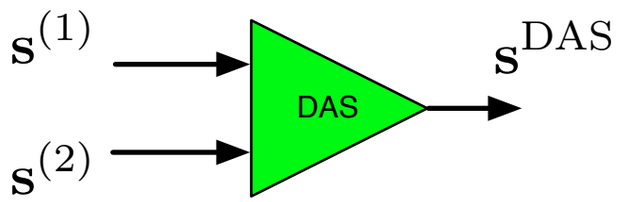
m - # descriptors

k_{BOVW} k_{CRFH} k_{SPH}

Classifier Classifier Classifier

Fusion of results

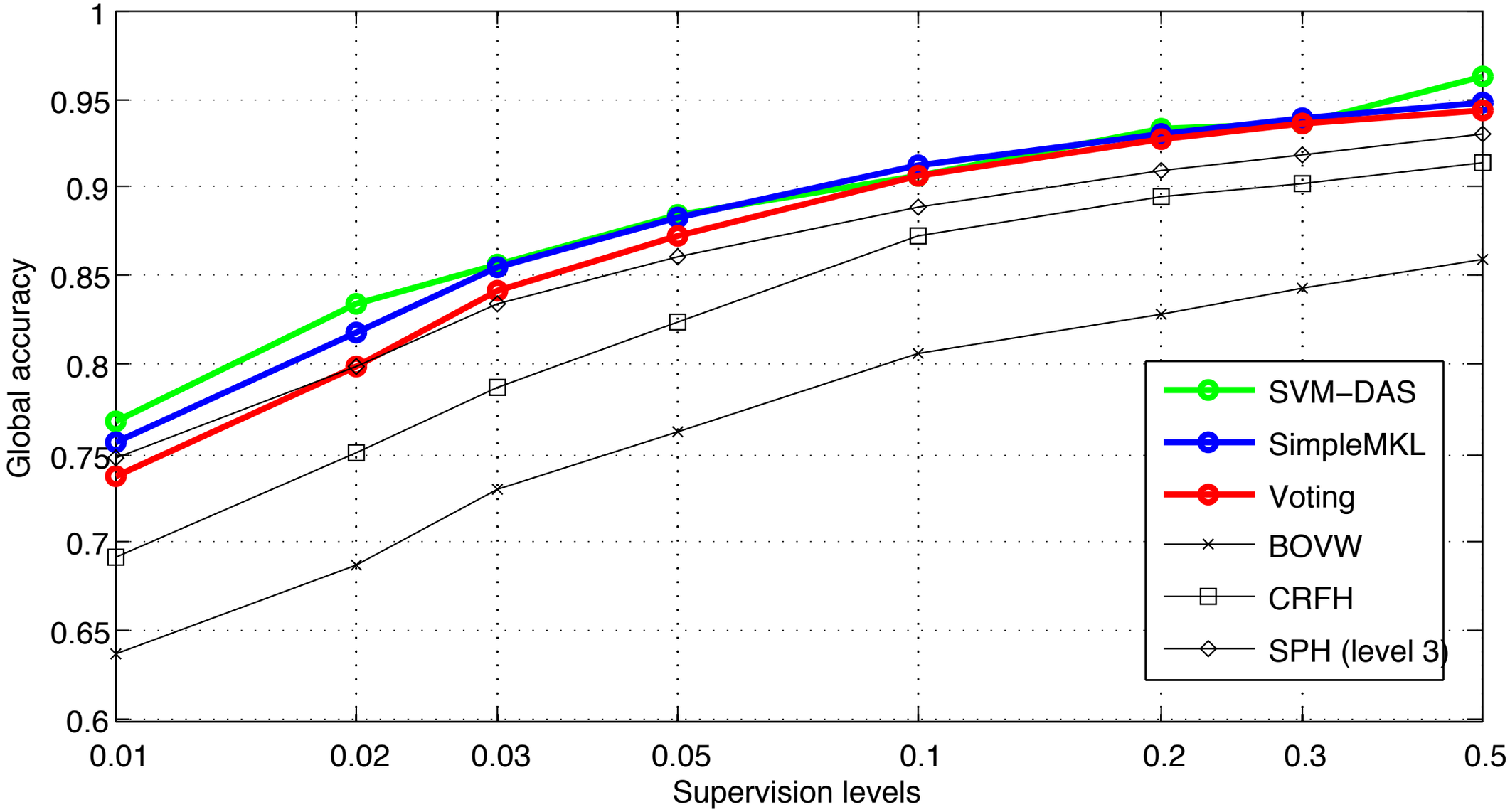
Linear kernel
Combine linearly the scores



Learning of the weights
Using Cross-validation

A. Pronobis, M. Mozos, and B. Caputo, "**SVM-based discriminative accumulation scheme for place recognition**," Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on, pp. 522–529, 2008.

Experimental Evaluation



Both strategies allow to improve the baselines

MKL is more computationally expensive than high-level fusion

Main Proposition

Time-Aware Co-Training Framework

✓ Leverage unlabeled data

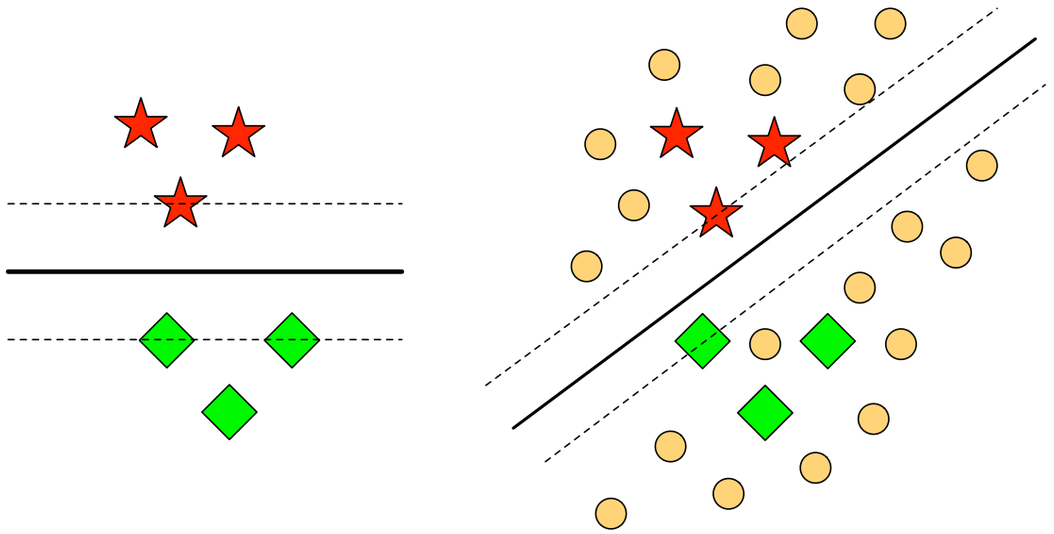
✓ Utilize multiple visual features

✓ Integrate time information

Leveraging unlabeled data

Idea - Visual Appearance model can be learned using also unlabeled images

Semi-Supervised SVM



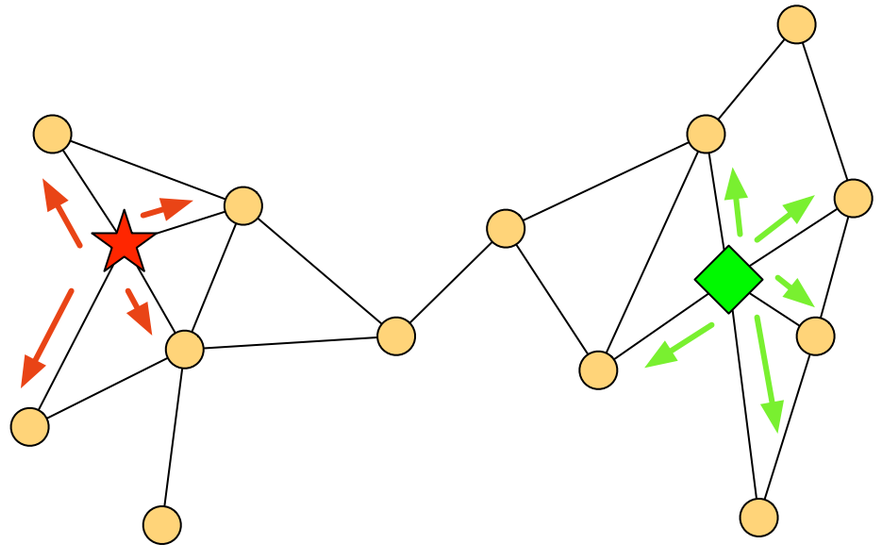
(a) Supervised SVM (b) Semi-Supervised SVM

Low Density Assumption

- Irrelevant unlabeled data can hurt
- Can be computationally expensive

S. Melacci, and M. Belkin, "**Laplacian Support Vector Machines Trained in the Primal**," *JMLR*, 2011.

Label Propagation

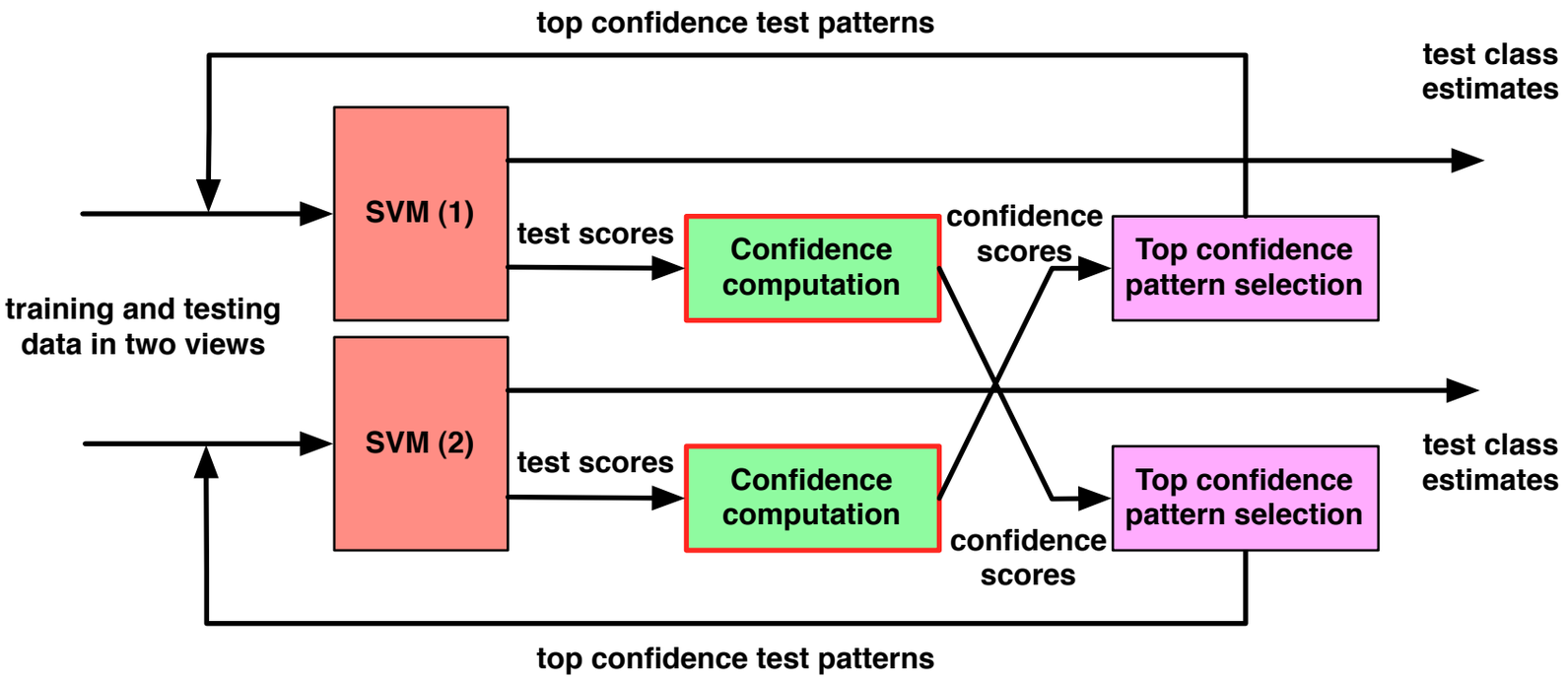


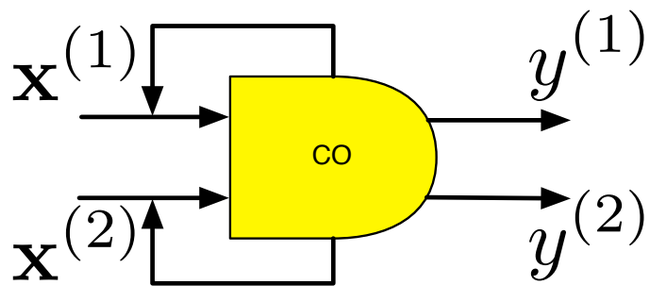
Cluster Assumption

- Prior knowledge about graph structure
- Adapted for sparse graphs
- Sensitive to class imbalance

X. Zhu and Z. Ghahramani, "**Learning from Labeled and Unlabeled Data with Label Propagation**," *Technical Report CMU-CALD-02-107*, 2003.

Standard Co-Training





Idea (1)
High confidence patterns can be used to improve the trained model

Idea (2)
Use two independent views on the data to diversify the outputs

Idea (3)
Hard to classify images are left for later Co-Training iterations

[1] A. Blum and T. Mitchell, "**Combining Labeled and Unlabeled Data with Co-Training**," *Conference on Computational Learning Theory*, Oct. 1998.

Integrating temporal continuity

Can we leverage the temporal continuity of the video?

$$\mathbf{s}_t = \left(f^1\left(\mathbf{x}_t\right), f^2\left(\mathbf{x}_t\right), \ldots, f^c\left(\mathbf{x}_t\right)\right)$$

$$\mathbf{s}_t = \sum_{t=-\tau}^{\tau} h\left(k\right) \mathbf{s}_{t+k}$$

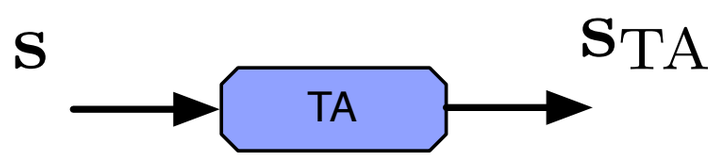
Averaging filter

$$h\left(k\right) = \frac{1}{2\tau+1}, k = -\tau, \ldots, \tau$$

Temporal Accumulation Scheme

Idea
Scores for temporally close images should be similar

Idea
Occasional misclassifications can be removed using this filtering

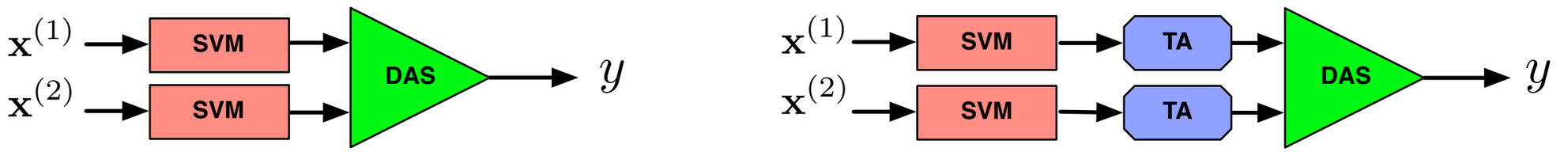


How to take into account time information into the learning framework?

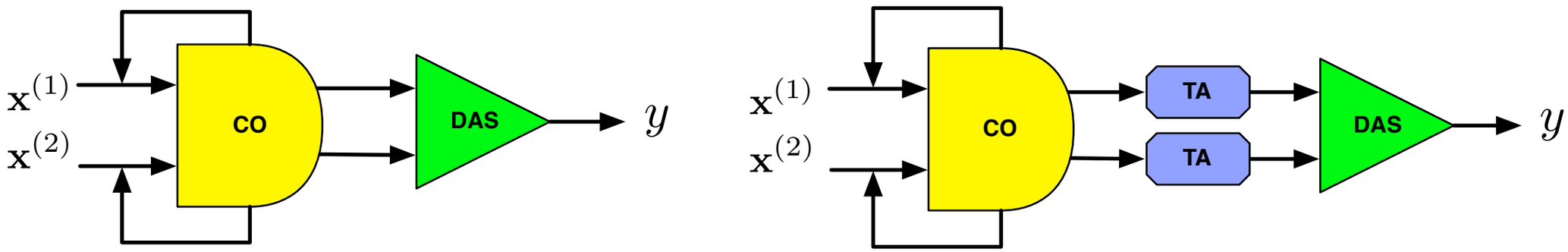
Outline of the Methods

Single descriptor

Multiple descriptor

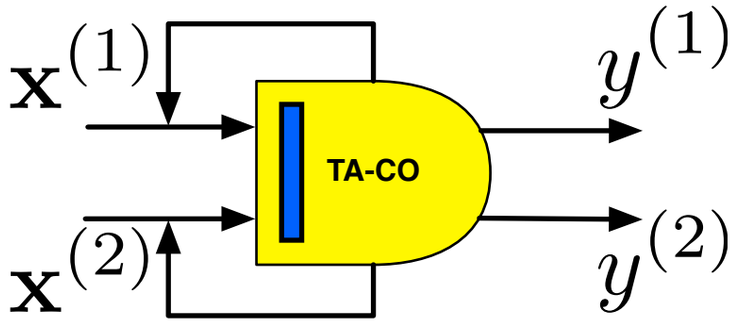


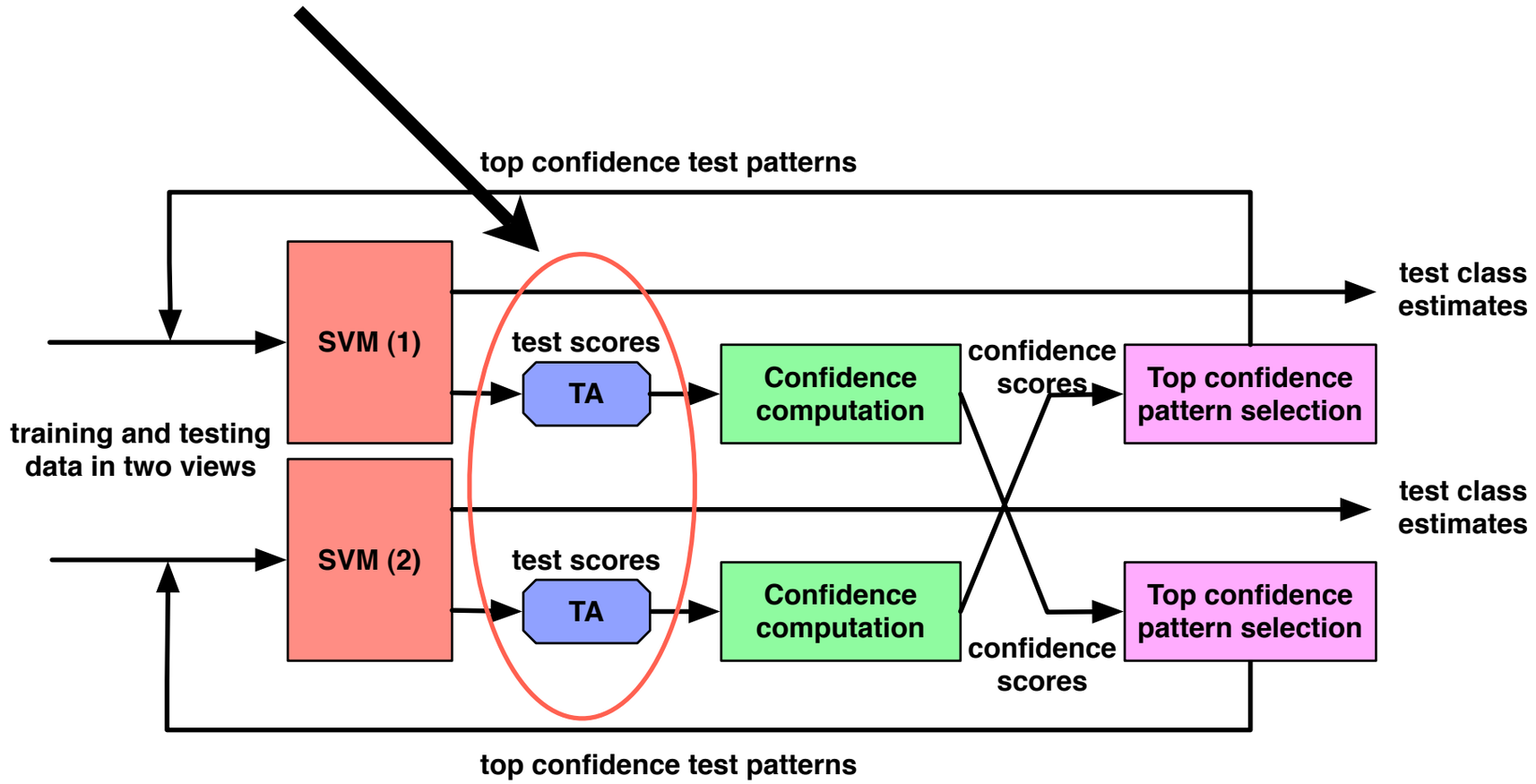
Co-Training: Semi-Supervised with Multiple descriptors



Time-Aware Co-Training

Idea
Internal TA introduces novel high-confidence patterns



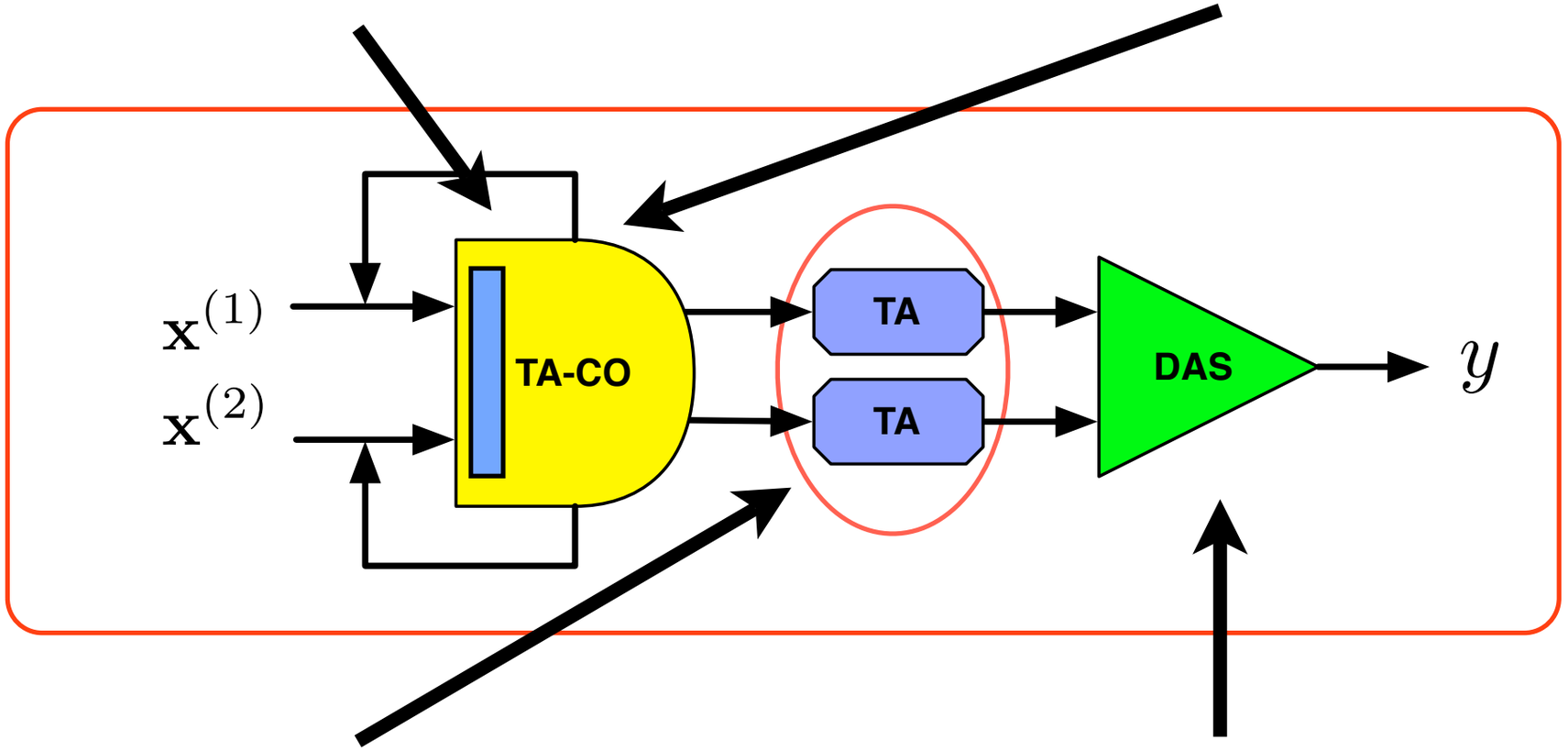


V. Dovgalecs, R. Mégret, and B. Yannick, "**Time-aware Co-Training for Indoors Localization in Visual Lifelogs**," *ACM International Conference on Multimedia*, pp. 1–4, Jul. 2011.

Proposed learning framework

Idea (1)
Co-Training leverages multiple feature data

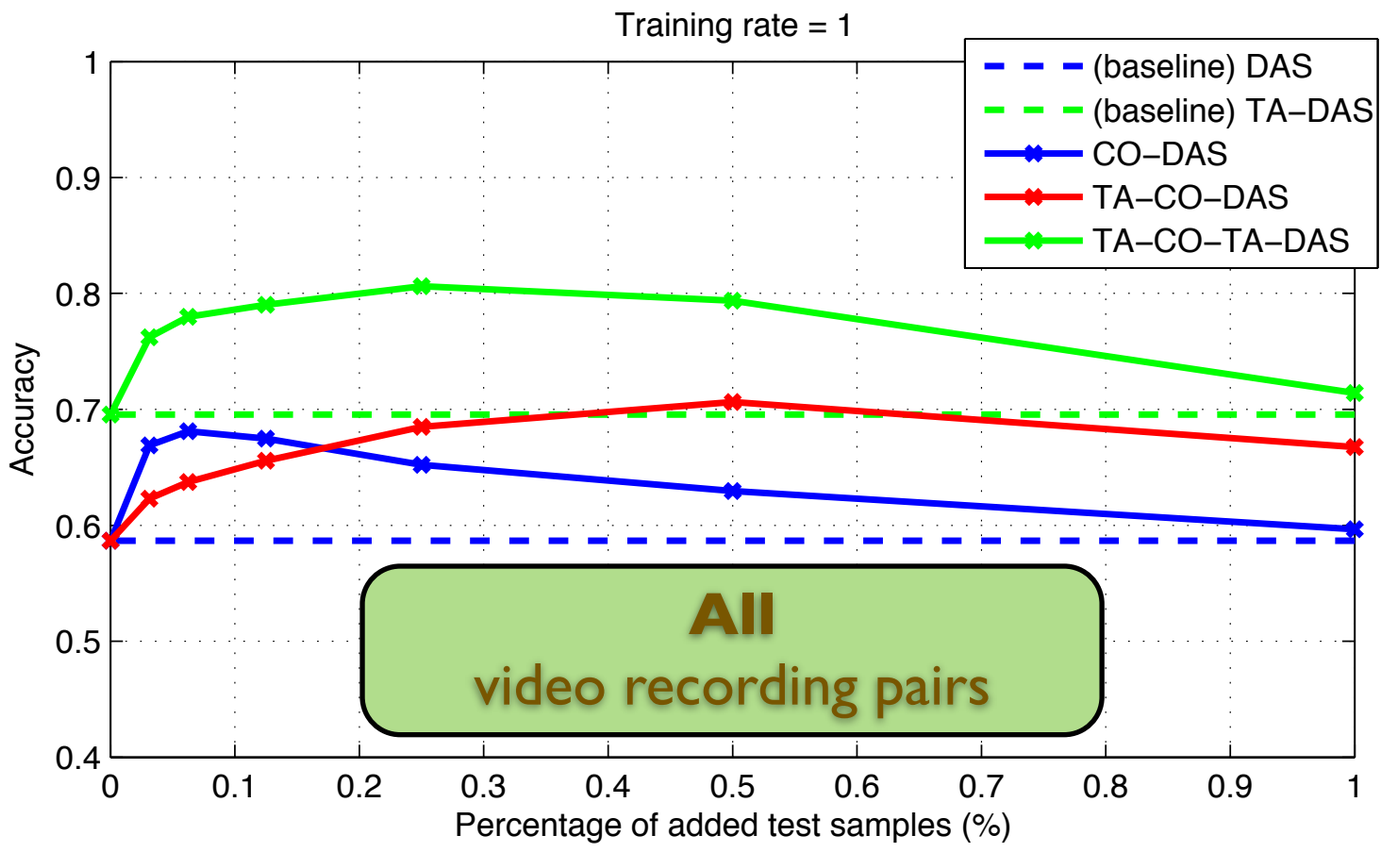
Idea (2)
Internal TA introduces novel high-confidence patterns

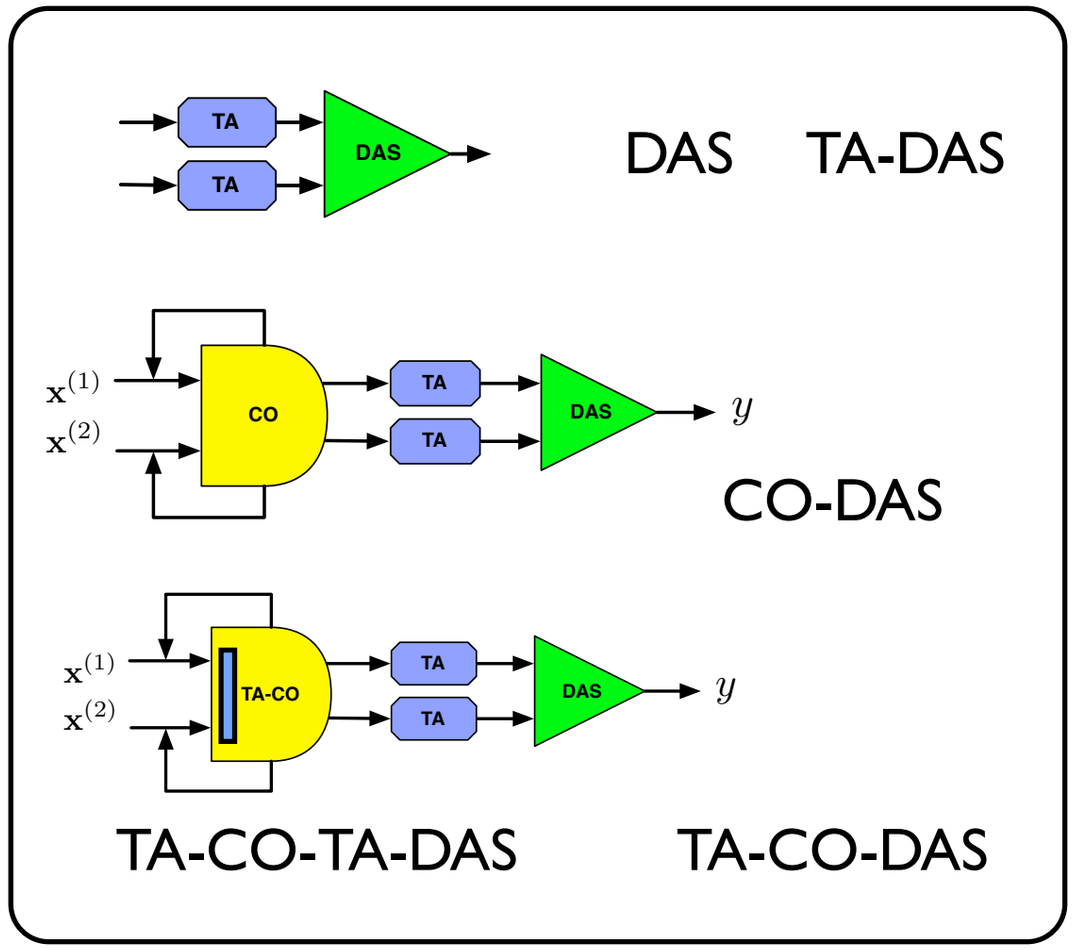


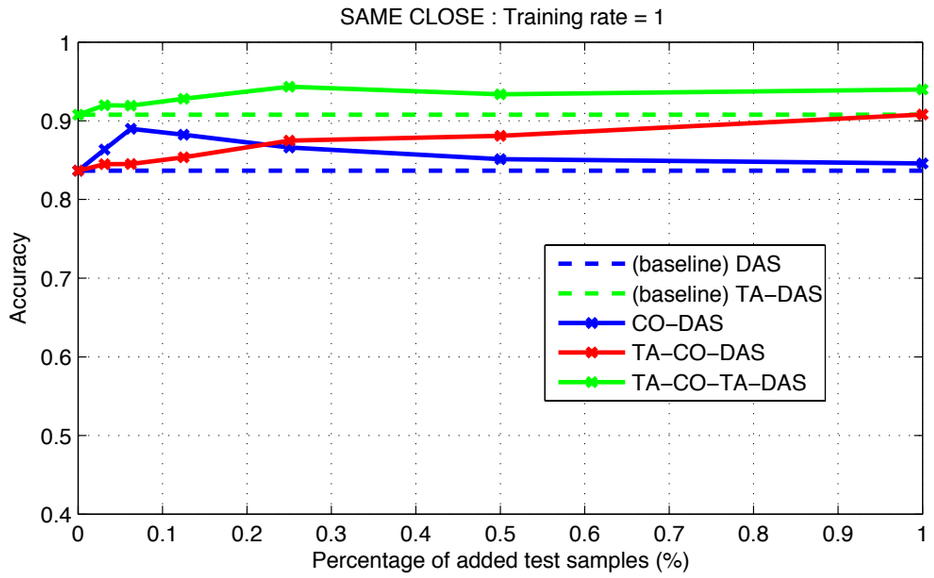
Idea (3)
External TA removes occasional misclassifications

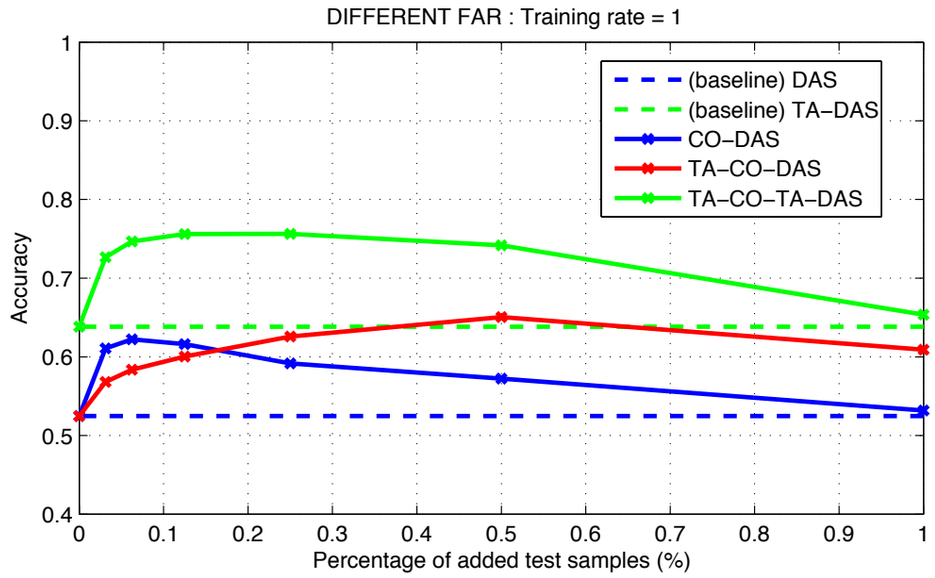
Idea (4)
DAS fusion provides a single high-level decision

Experimental Evaluation









Highlights of Co-Training methods

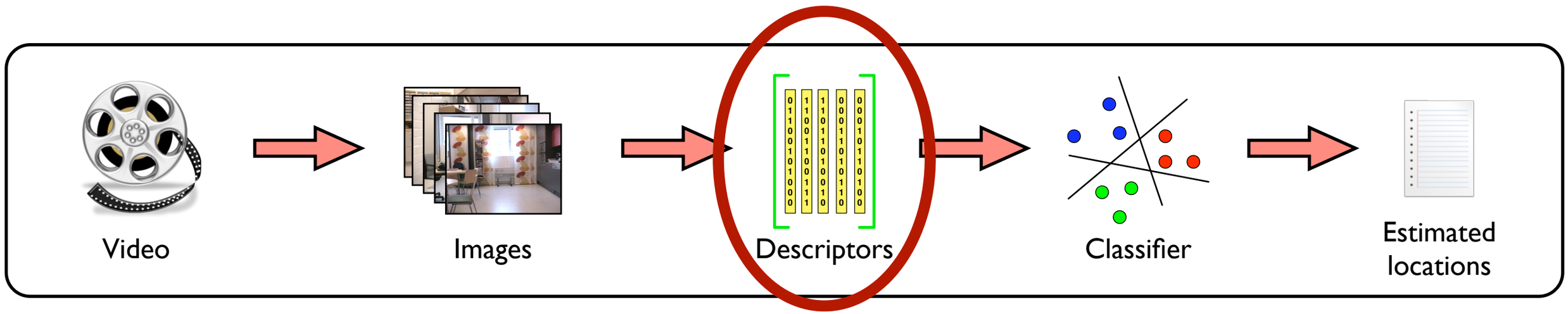
1. **Unlabeled data** appears to help in the localization task
   ➡ Co-Training is adapted for the task

2. Use of **multiple visual features** should be encouraged
   ➡ Late fusion scheme is adapted for the task

3. Leveraging **temporal information** has a significant impact
   ➡ Use in post-processing and in the feedback loop

Improving Visual Descriptor

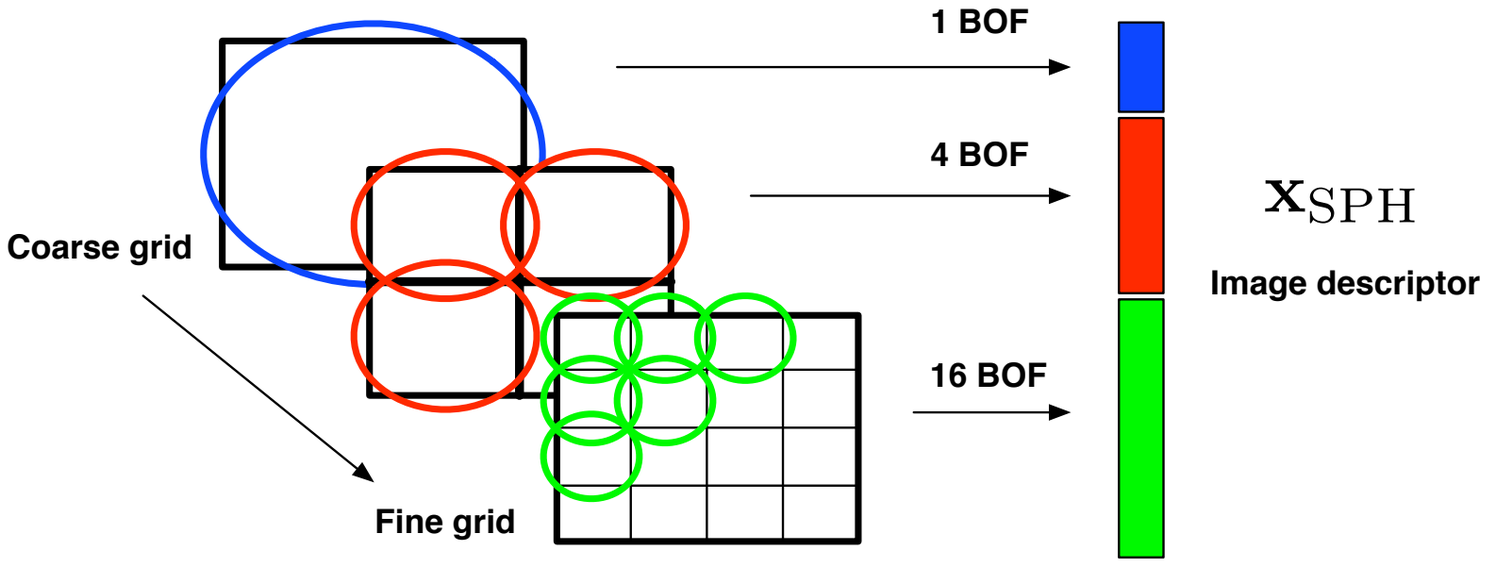
Translation Invariant SPH

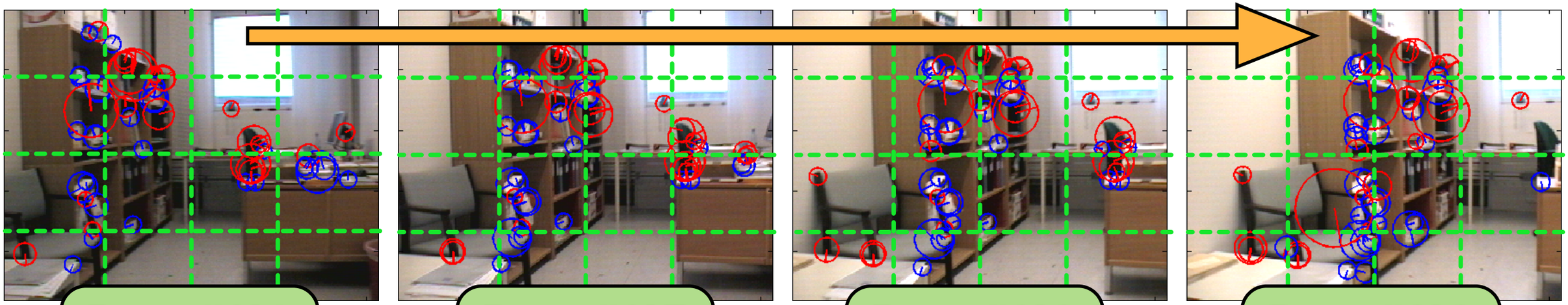


Spatial Pyramid Histograms

50 strongest response SURF descriptors per image.

Taken from IDOL2 cloudy1 sequence





Variability due to spatial movement
Notice no class change

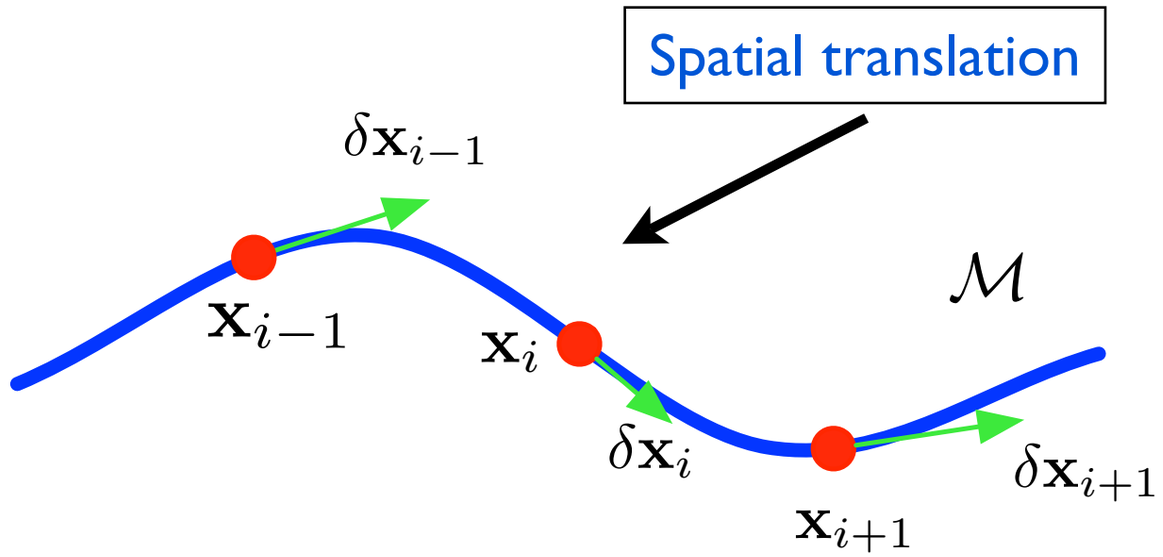
How to render the descriptor invariant to horizontal spatial translation?

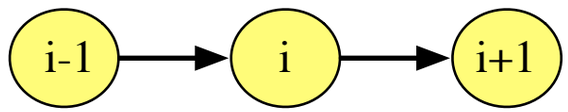
Notion of tangent vectors

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$$

Idea

There may exist **directions** in feature space along which the SVM decision value **should not change**





$$\delta\mathbf{x}_i = \lim_{t \to 0} \frac{1}{t} (\mathcal{L}_t \mathbf{x}_i - \mathbf{x}_i) = \left. \frac{\partial}{\partial t} \right|_{t=0} \mathcal{L}_t \mathbf{x}_i$$

$\mathbf{x}_i$ - original pattern

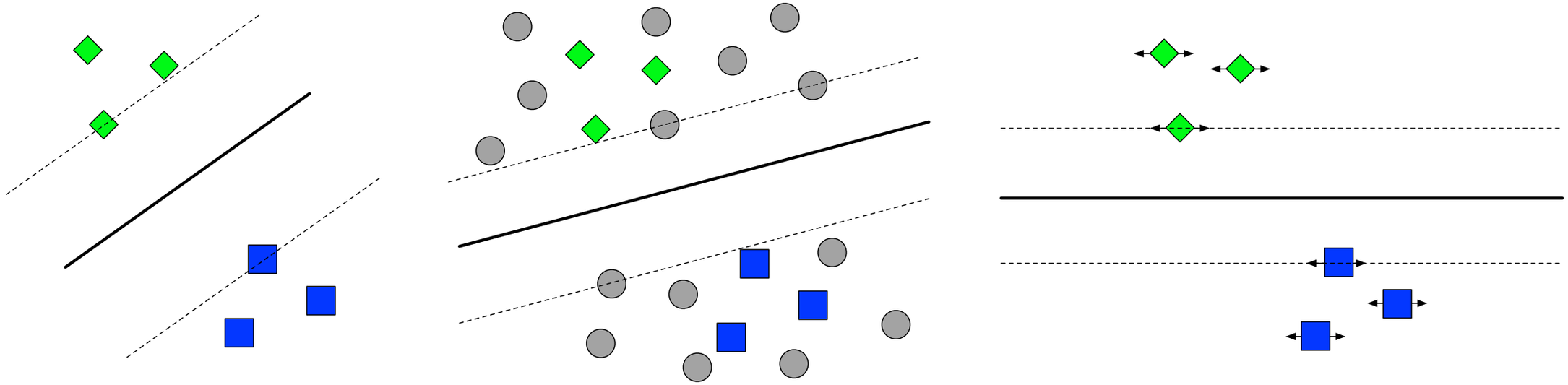
$\mathcal{L}_t \mathbf{x}_i$ - transformed pattern

Invariance - Intuition

Standard Supervised SVM

Semi-Supervised SVM

Invariant Supervised SVM



-Lack of labeled data

Risk of overfitting!
(if low supervision)

+Unlabeled data may help

Computationally expensive!

Invariant directions in feature space provide hints

Standard and Invariant SVM

Training and Testing data

$$L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l} \qquad U = \{\mathbf{x}_i\}_{i=l+1}^{l+u}$$

IDEA

$$\min_{\mathbf{w}} (1-\gamma) \|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{l} \langle \mathbf{w}, \delta\mathbf{x}_i \rangle^2$$

Penalize tangent vector margin crossing

Information about invariant directions

$$S_\gamma = (1-\gamma) I + \gamma \sum_{i=1}^{l} \delta\mathbf{x}_i \delta\mathbf{x}_i^T$$

$0 \le \gamma \le 1$ γ - amount of invariance

$$\mathbf{x}_i^\gamma = S_\gamma^{-\frac{1}{2}} \mathbf{x}_i \qquad \mathbf{w}^\gamma = S_\gamma^{\frac{1}{2}} \mathbf{w}$$

Standard SVM

$$\min_{\mathbf{w}} \|\mathbf{w}\|^2$$

$$\text{s.t. } y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \ge 1$$

$$f^{\text{SVM}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$$

Invariant SVM

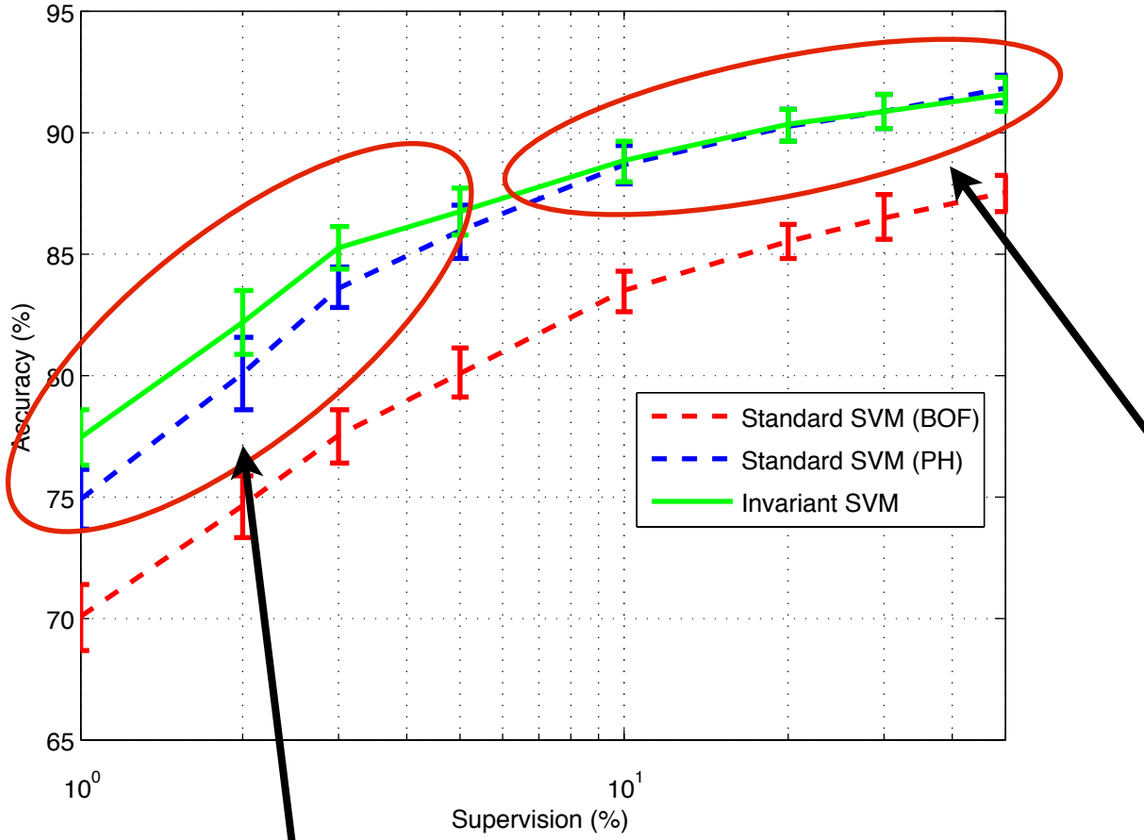
$$\min_{\mathbf{w}^\gamma} \|\mathbf{w}^\gamma\|^2$$

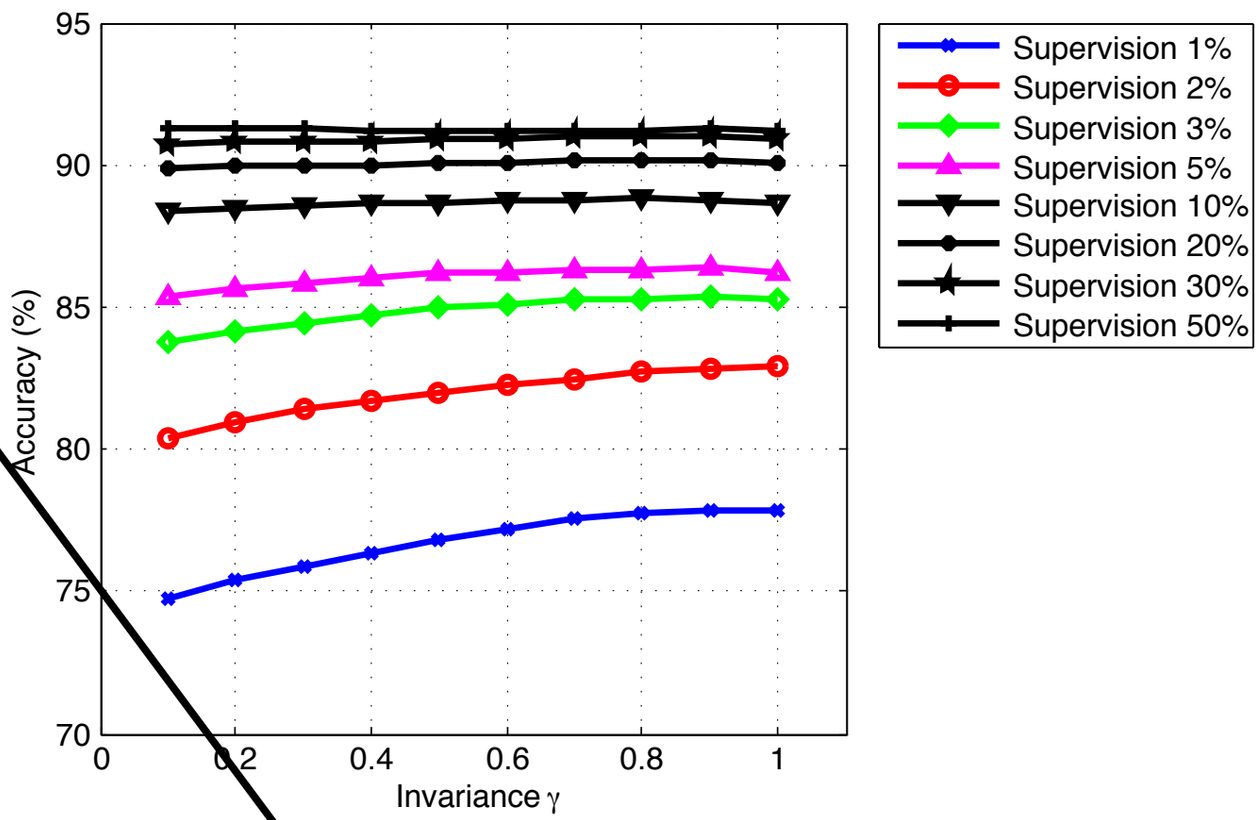
$$\text{s.t. } y_i \left(\langle \mathbf{w}^\gamma, \mathbf{x}_i^\gamma \rangle + b \right) \ge 1$$

B. Scholkopf and A. J. Smola, "***Learning with Kernels***". MIT Press, Cambridge, 2002.

Experimental Evaluation

Test setup : Orderless image collection





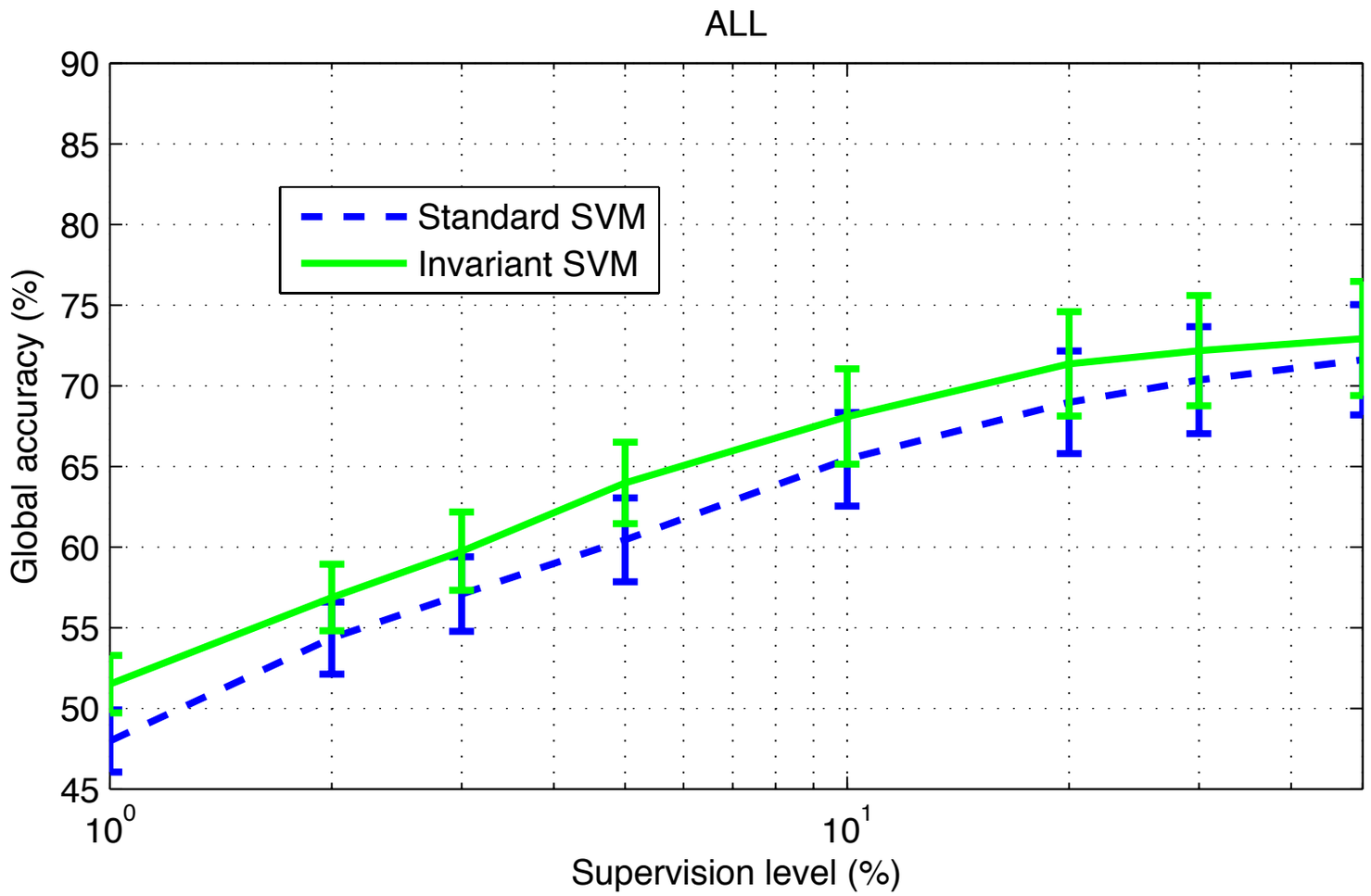
Improvement at low supervision levels

Invariance already included in training

V. Dovgalecs, S. Ilcus, R. Megret and Y. Berthoumieu, "**Pyramides spatiales d'histogrammes invariantes aux transformations pour la reconnaisance des lieux**", *18éme congres nationale en Reconnaisance des Formes et Intelligence Artificielle (RFIA)*, LIRIS, Lyon, 2012.(accepted)

Experimental Evaluation

Test setup : Video vs Video



Lack of training data

Improved discrimination power

10 folds of training sets

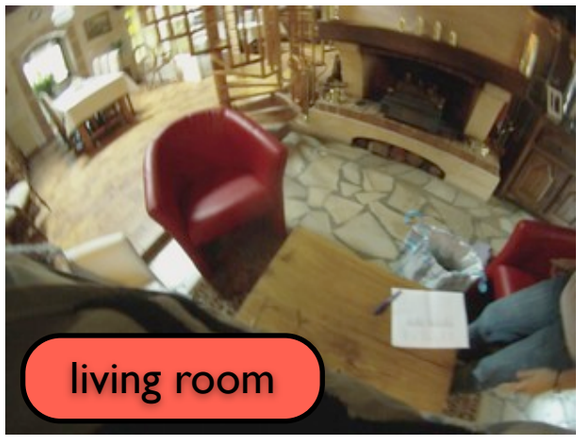
Constant performance increase in all conditions

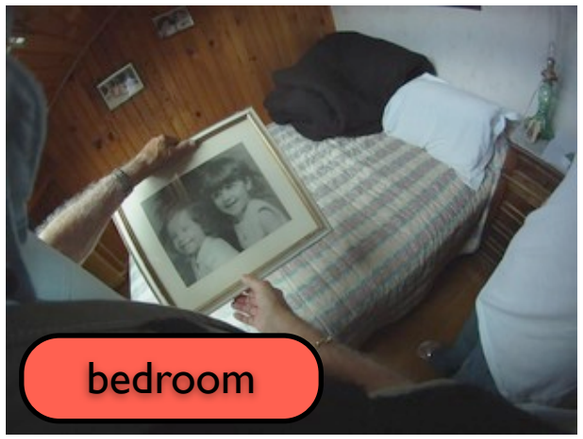
Better knowledge transfer to new videos!

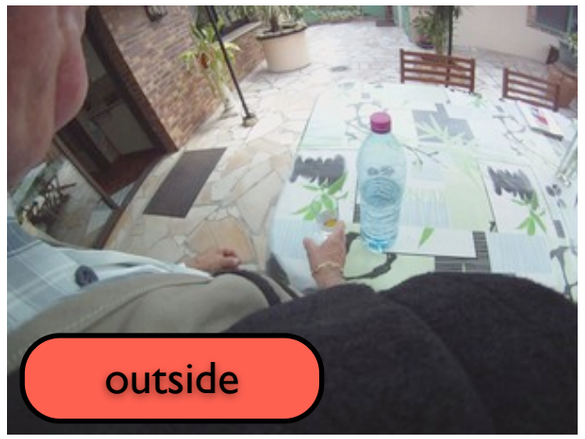
Experimental Results

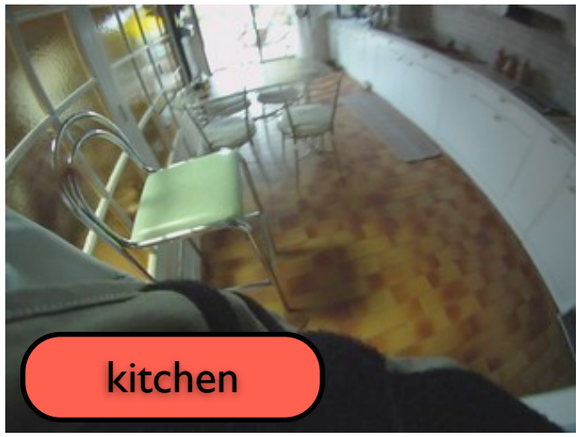
Large-Scale evaluation (IMMED)

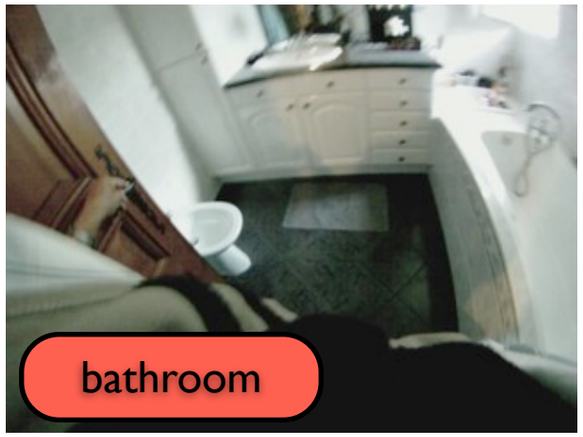
The Database

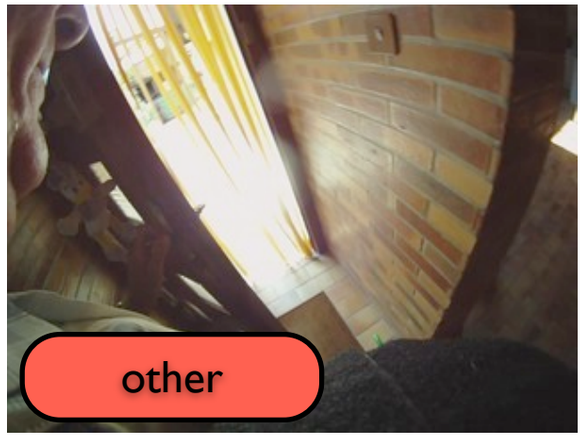












BOVW
Bag of Visual Words

SPH
Spatial Pyramid Histograms

All parameters were set using Cross-Validation

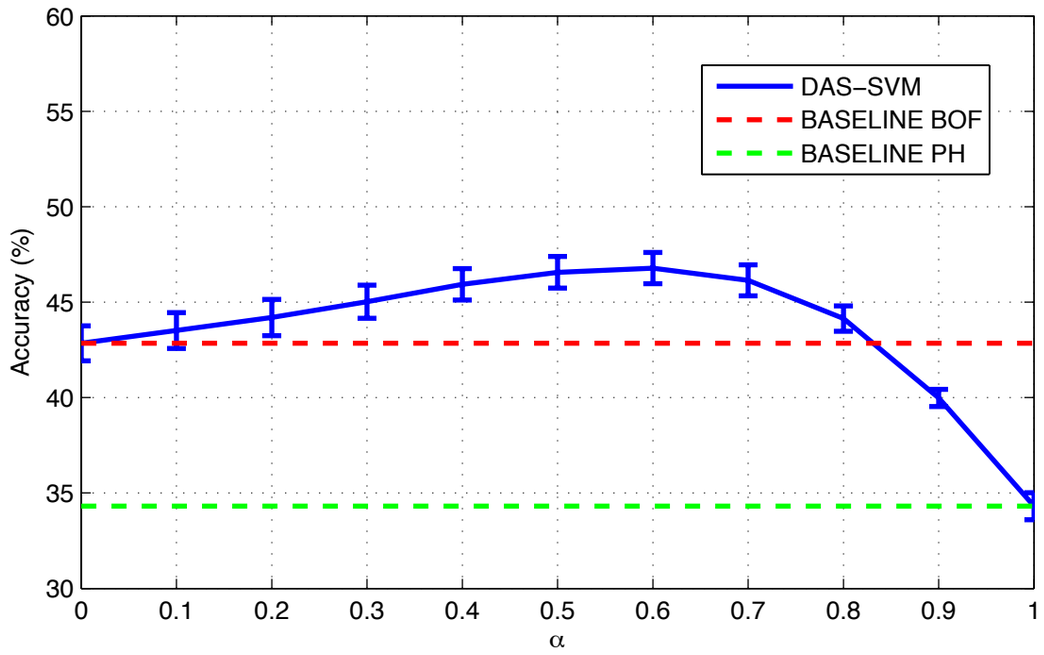
Bootstrap video - **6 minutes**

Unlabeled video - **33 minutes**

bathroom	bedroom	kitchen	living room	outside	other
hygiene, teeth brushing	sleep, rest	food preparation, dish washing, complex machines	eating, watching TV, reading	outdoors	other indoor locations

Baselines

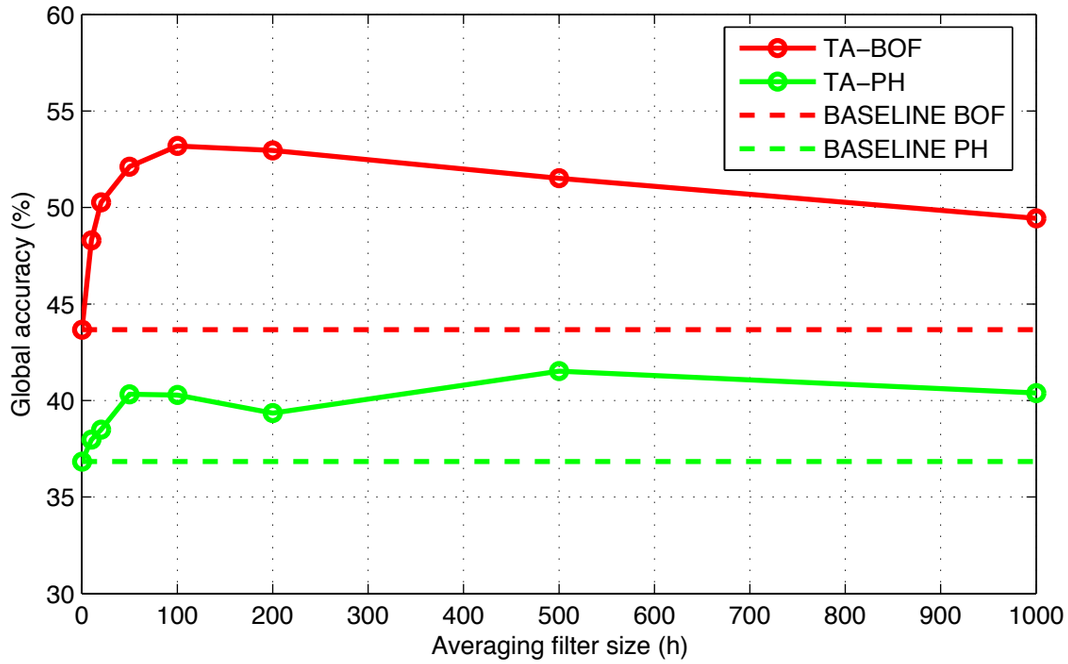
What are the baseline classification performances?



Relatively **low** starting performance

Visual features **are** complementary

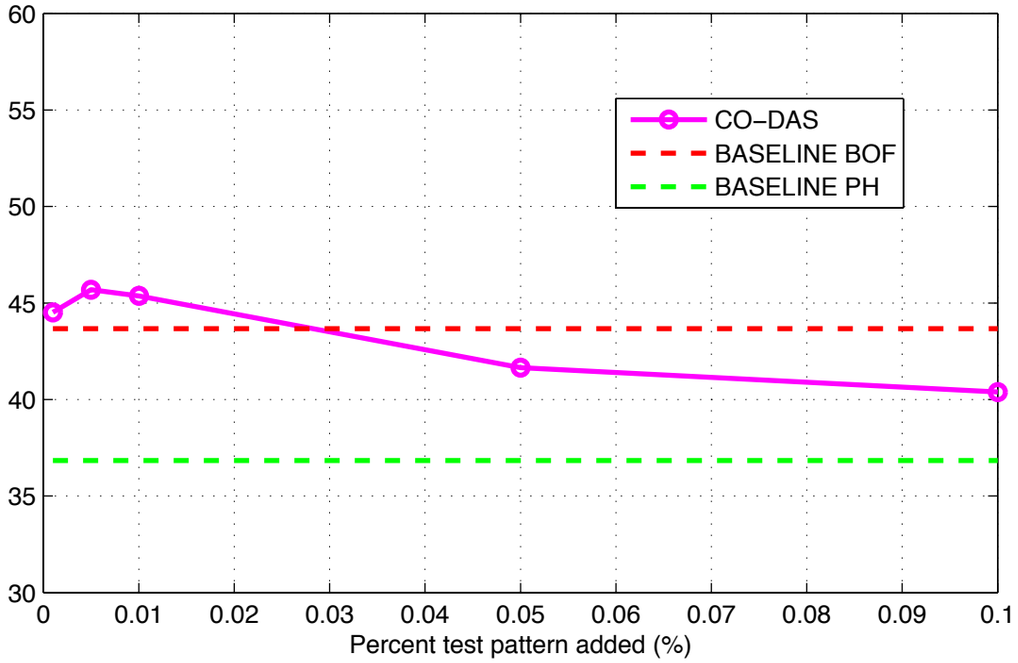
What is the impact of TA scheme?



TA boosts the performance

Time-Aware classification

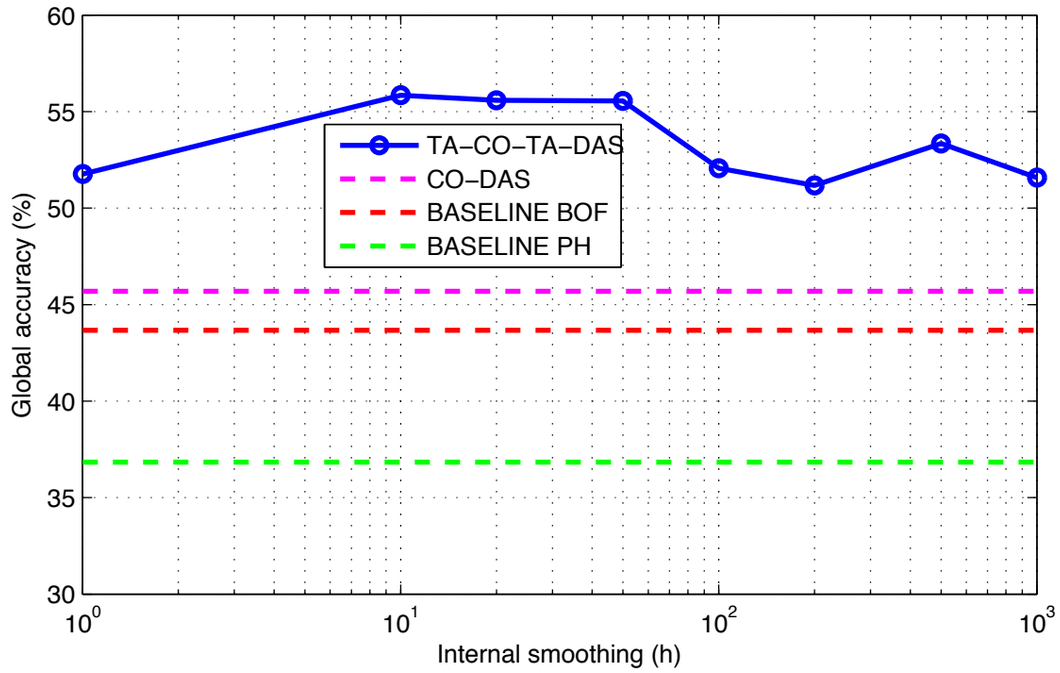
Can Co-Training improve the performance?



Small amounts of top confidence estimates are reliable even for **complicated** data

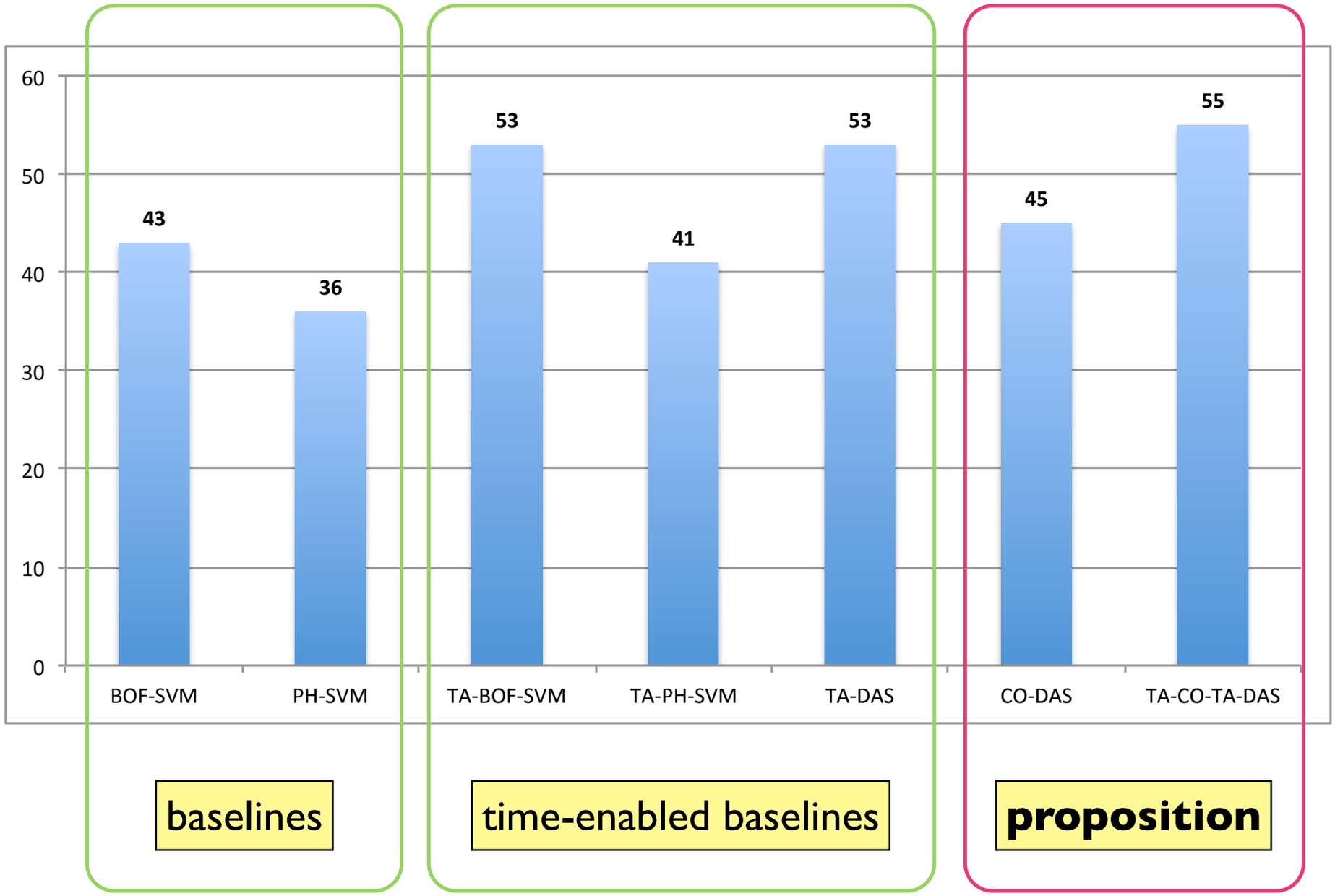
We used **one** Co-Training iteration

Time-Aware Co-Training



Internal TA **diversifies** input for the next round of Co-Training

Summary



Low baseline performance

Utility of temporal info

The role of Co-Training

Conclusions

Multiple information sources
Early and Late fusion

Leveraging unlabeled data
Semi-Supervised Learning

Integrating temporal information
Temporal Accumulation

Improving visual descriptor
Translation Invariant SPH

Visual content variability and complexity

✓ Study of relevant image content **descriptors** and similarity measures

➡BOVW, CRFH,SPH
➡Kernels

✓ Increasing discrimination power using **multiple** features

➡Early or late fusion

✓ Building spatial translation **invariant** PH descriptor

➡Improved generalization capacity

Small amounts of training data

✓ Study of applicability of **Semi-Supervised** learning for image-based localization

➡Unlabeled data helps to improve place recognition rates
➡Proposition of a confidence measure

✓ Proposed an **unified** learning framework

➡Multiple features
➡Efficient semi-supervised learning
➡Temporal constraint integration

Perspectives

Perspectives

- Semi-Automatic video indexing approach

  Active learning for reliable bootstrapping

  Expected Improve robustness and the speed of the learning process

- Integration with other information sources

  Coupling vision + inertial and environment sensor information

  Expected Provide additional cues when pure vision is insufficient

Thank You! *

* Questions?

Annex

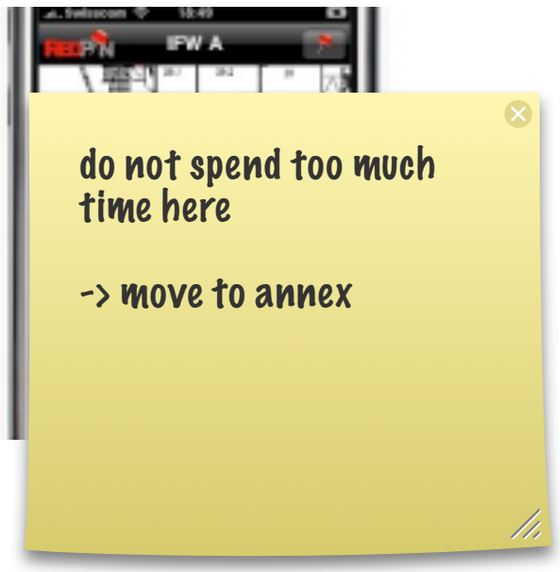
Indoor localization

GPS

WiFi / RFID

✓Indoor localization only
✓Requires relatively high precision
✓No additional equipment to install

➡Wearable, image-based approach



Video Indexing Problem

Text indexing

Invention of press in 1400AD. Need for text indexing solutions.

Challenge: discovering semantic meaning of the text.

Sphinx

up to **6TB** of structured text (~5 billion documents)

Im…g

Prolif… …s in 20…

Large… …ions.

Chall… …re interpretations than text.

Flickr hosts ~6 billion photos (August 2011)

Facebook hosts ~3 billion new photos monthly (2010)

hundreds of TB

Vide… …ing

Video recording devices in 20th century.
Ever growing video archives.

Challenge: videos are non-static and richer in content.

Youtube - 48/hrs of video uploaded every minute (2011)

Petabyte storage

Why we use visual localization?

It should stick well with visual lifelogging

Wearable sensors in intro -> position wrt fixed sensors to explain complementarity (specificity of each)

In what the work is original? other solutions are similar but the problem is not the same

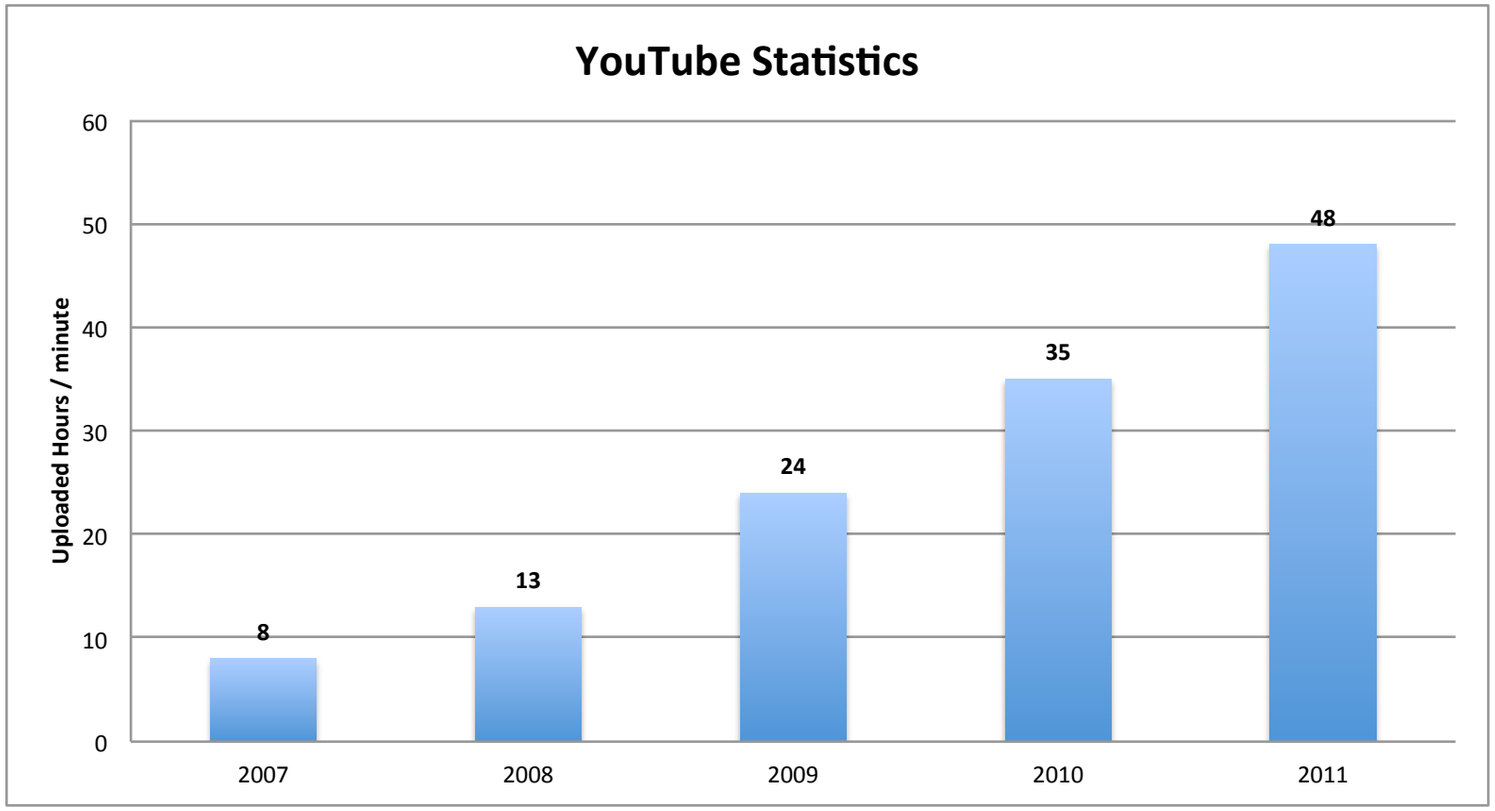
MIssing link between indexing and localization

Title -> 2nd slide transition issue

REMOVE IT

Video Indexing Problem

New video content in hrs / min (YouTube)



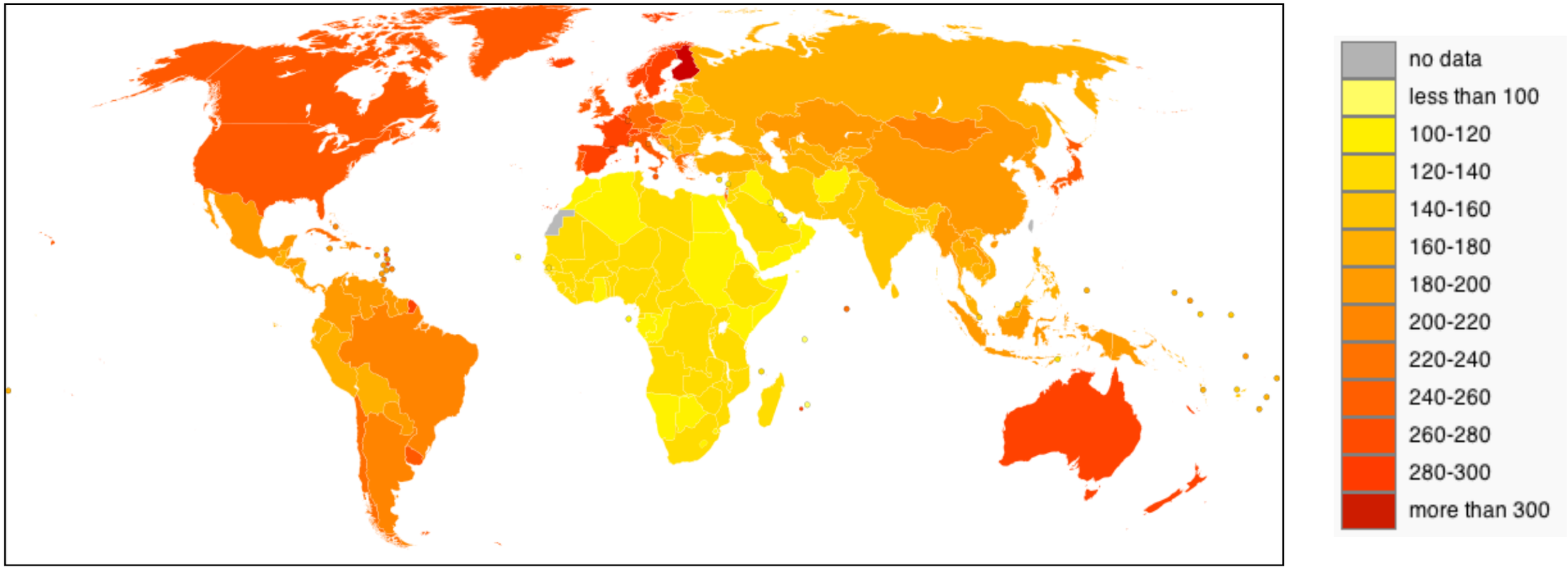
Applications

- film archives
- surveillance
- news videos
- sports video analysis
- medical applications

How can we search videos by content efficiently?

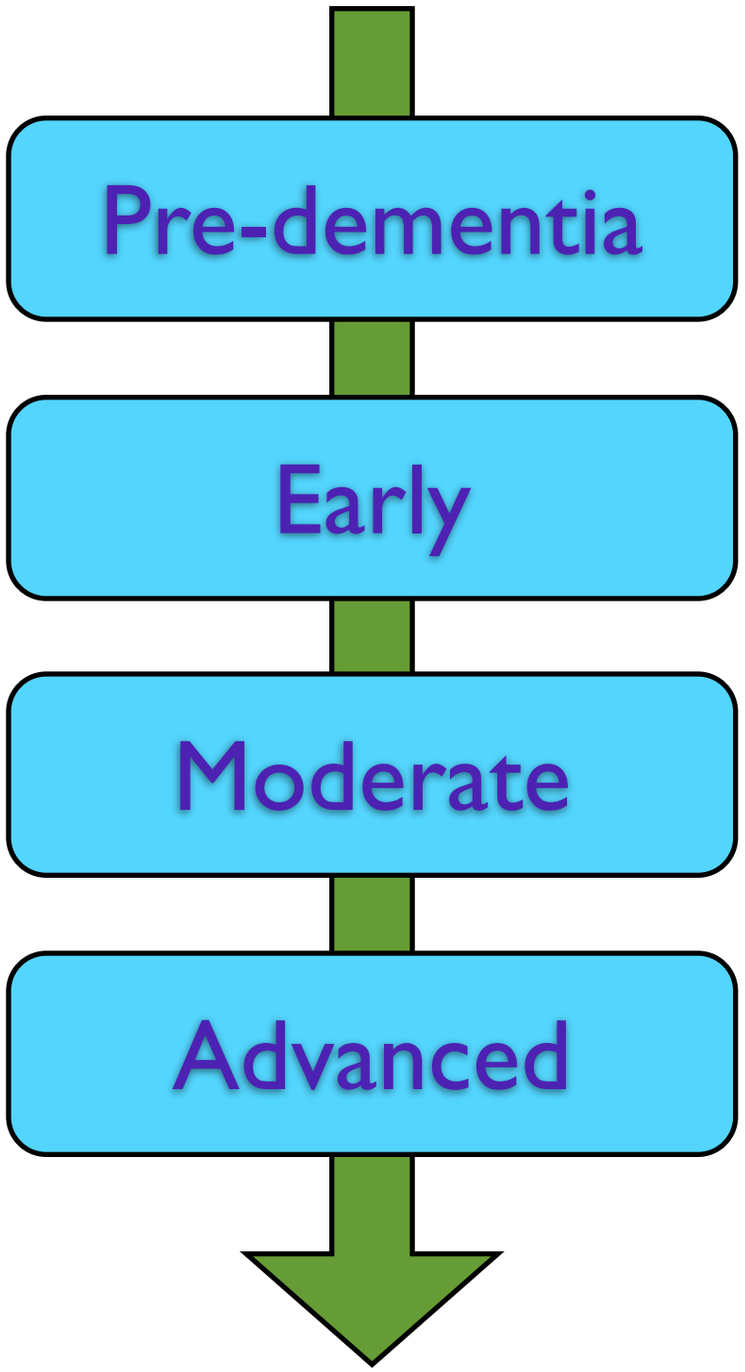
Alzheimer's Disease

AD is dementia type disease that develops cognitive and functional disorders mostly for elderly people (>65 years) in developed countries



Disability-adjusted life year for Alzheimer and other dementia's per 100'000 inhabitants (2004)

Development of AD



Mild cognitive impairments. Difficulty remembering recently learned facts. Up to 8 years before clinical diagnosis.

Memory reduces to older memories and learned facts. Decreased oral fluency and shrinking word vocabulary.

Loss of independence as many daily activities. Loosing communication skills. May fail to recognize close relatives.

Completely dependent on caregivers. Very simple daily tasks cannot be accomplished without help.

The disease cannot be cured nor its cause is known.

Treatment

Pharmaceutical

Delaying and relieving certain AD side-effects

Psychosocial

Cognition-oriented treatment aimed to reduced cognitive deficits

Caregiving

Help in daily activities by close relatives or professional caregivers

Importance of early diagnosis

Early changes and adaptation of the lifestyle

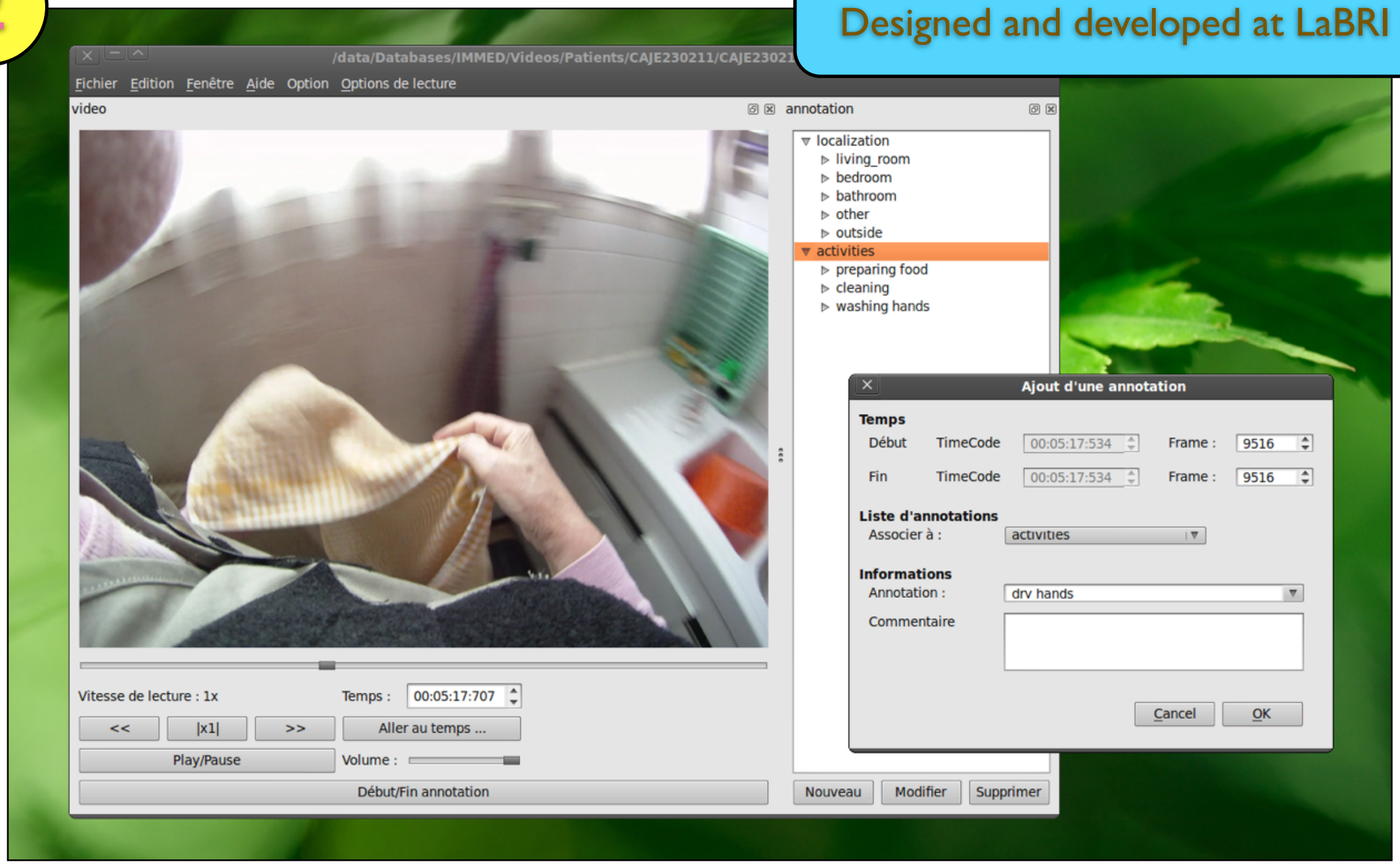
Improved acceptance by family and close relative

Significantly reduces direct and indirect costs

Annotation task

2

Designed and developed at LaBRI

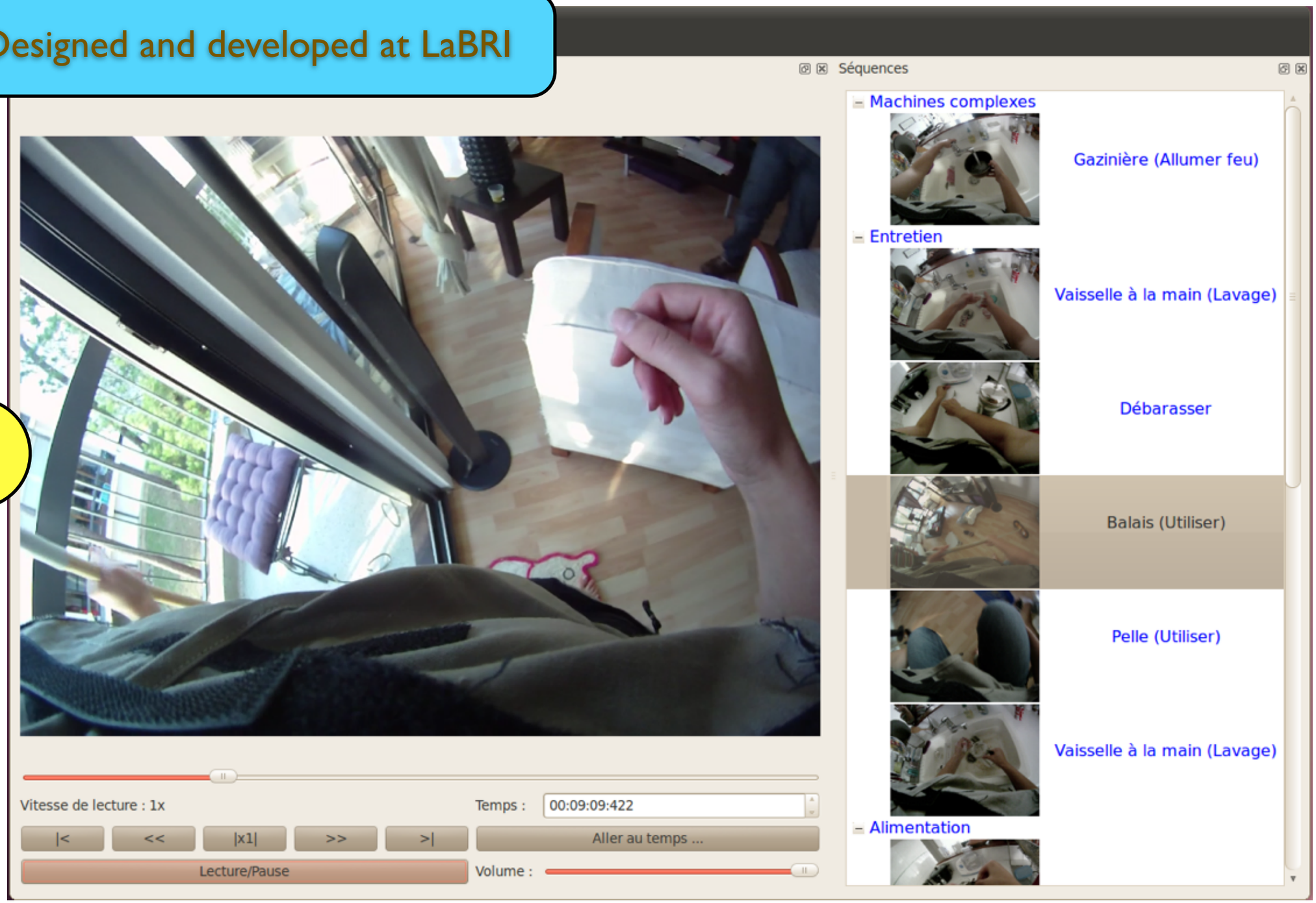


Manual annotation is done only for a short bootstrap video!

LaBRI

Automatically indexed video browsing

Designed and developed at LaBRI

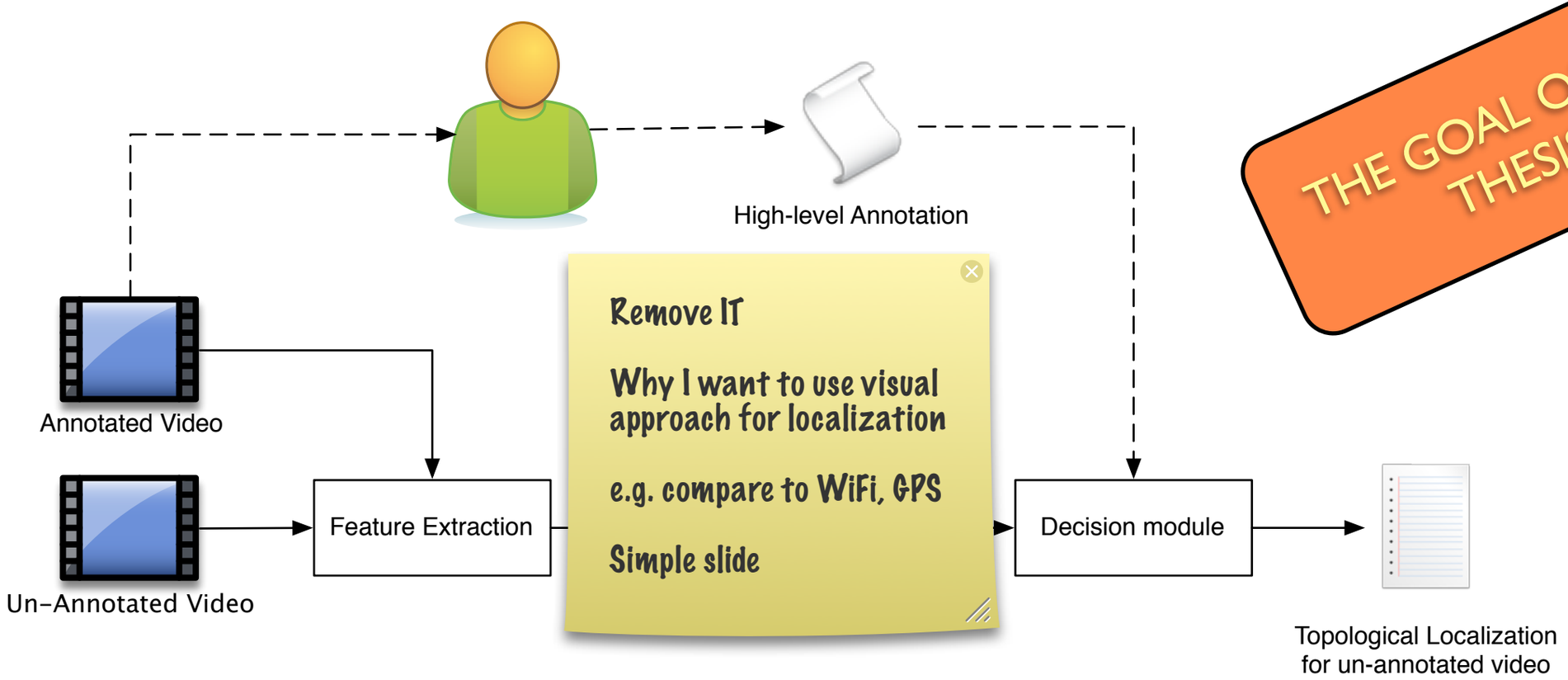


4

Automatically indexed visual content allows fast video browsing

LaBRI

Automatic Video Indexing



THE GOAL OF THE THESIS

➡ Classical Pattern Recognition problem - extract features and provide high-level decision

➡ Adapt Machine Learning techniques for Decision in our context

Relevant Information Extraction

Extracted visual features are of high dimensionality and redundant

Can we reduce dimensionality of the data and adapt it for the task?

Unsupervised dimensionality reduction

Data-Adapted kernel function!

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left\langle \Phi\left(\mathbf{x}_i\right), \Phi\left(\mathbf{x}_j\right) \right\rangle$$

$$\Phi \rightarrow \mathcal{H}$$

$$\mathbf{x} \mapsto \Phi\left(\mathbf{x}\right)$$

Variance (Kernel PCA)

Highlight that compact and adapted features are now found!

Graph (Laplacian Eigenmaps)

$$S_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} \Phi\left(\mathbf{x}_i\right) \Phi\left(\mathbf{x}_i\right)^T$$

$$S_{\mathcal{H}} \mathbf{e}_i = \lambda_i \mathbf{e}_i$$

$$\mathbf{e}_i = \left[\Phi\left(\mathbf{x}_1\right), \ldots, \Phi\left(\mathbf{x}_n\right)\right] \mathbf{a}_i$$

$$K\mathbf{a}_i = \lambda_i \mathbf{a}_i \qquad \mathbf{z} = A_k^T K$$

z - the new representation

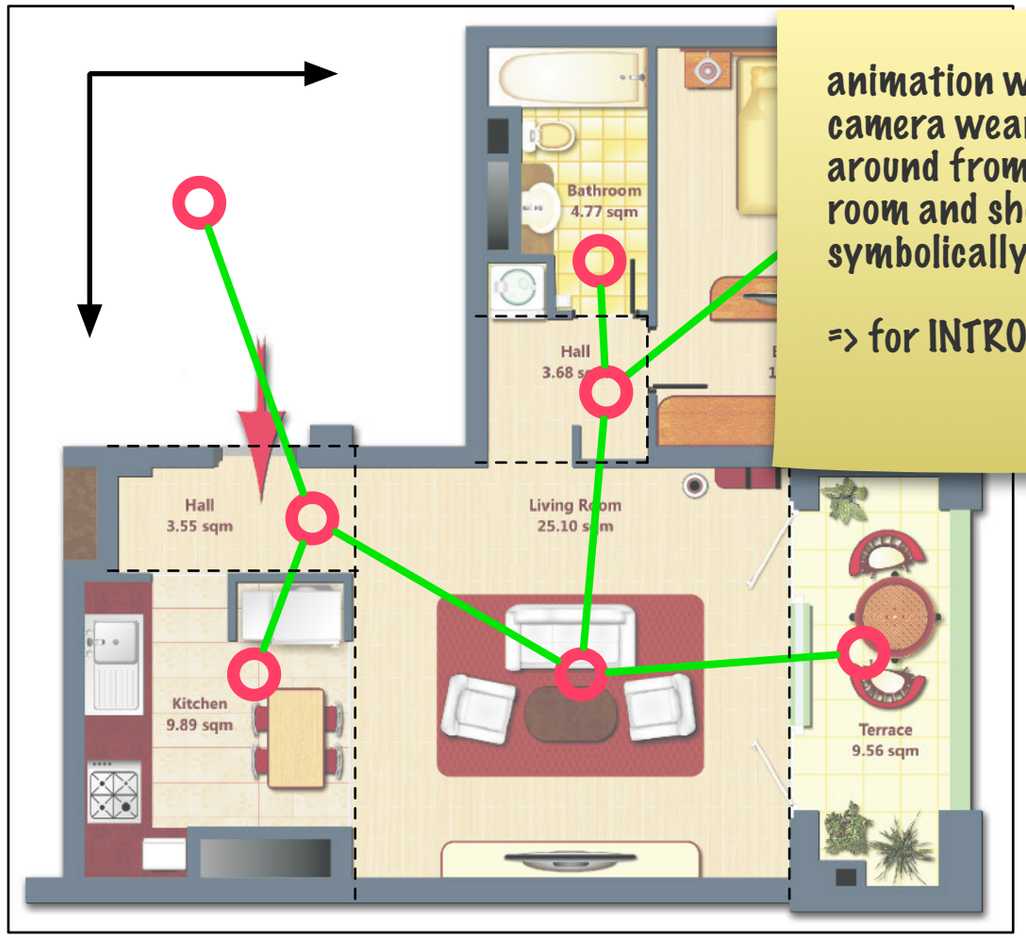
$$E\left(\mathbf{z}\right) = \frac{1}{2} \sum_{i,j} w_{ij} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|^2$$

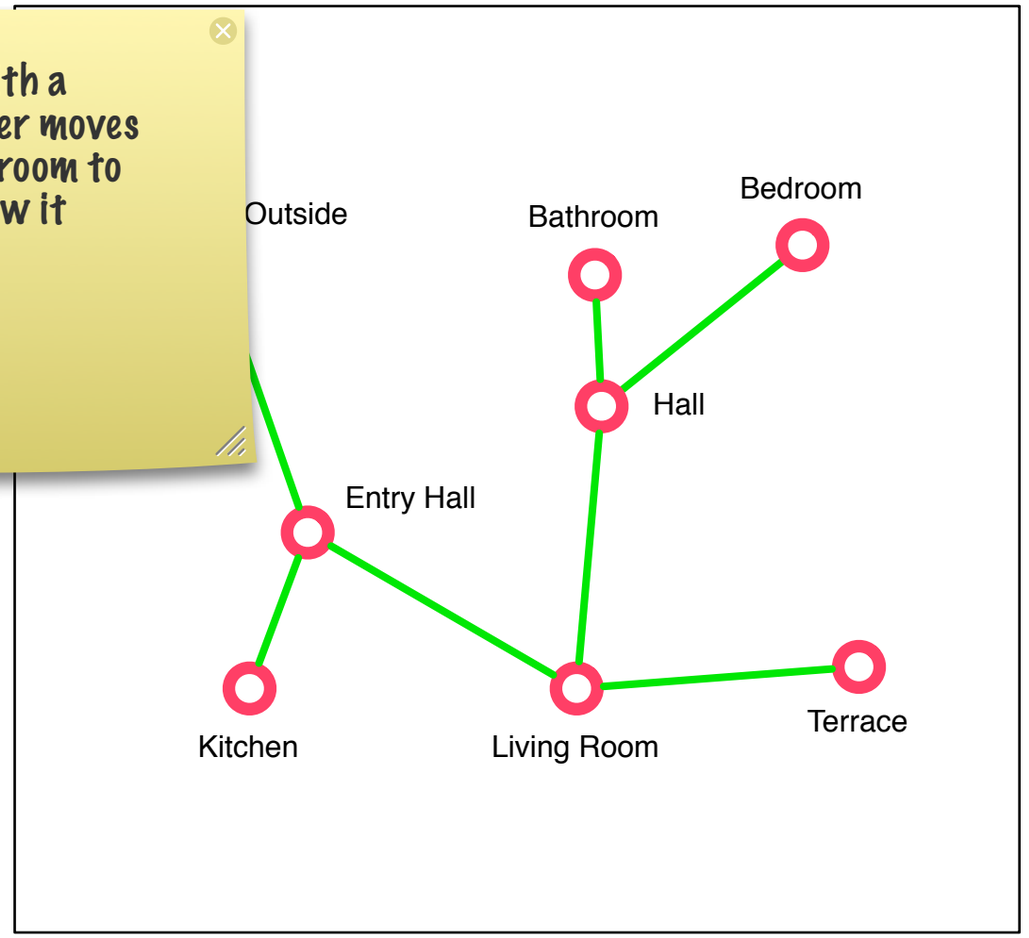
$$L\mathbf{e} = \lambda \mathbf{e}$$ L - Graph Laplacian

$$\mathbf{z}_i = \left[\mathbf{e}_1\left(i\right), \ldots, \mathbf{e}_k\left(i\right)\right]^T$$

Thesis Objective

Develop automatic algorithms for video indexing with the goal of topological localization indoors





Metric Localization

Topological Localization

Notion of similarity

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

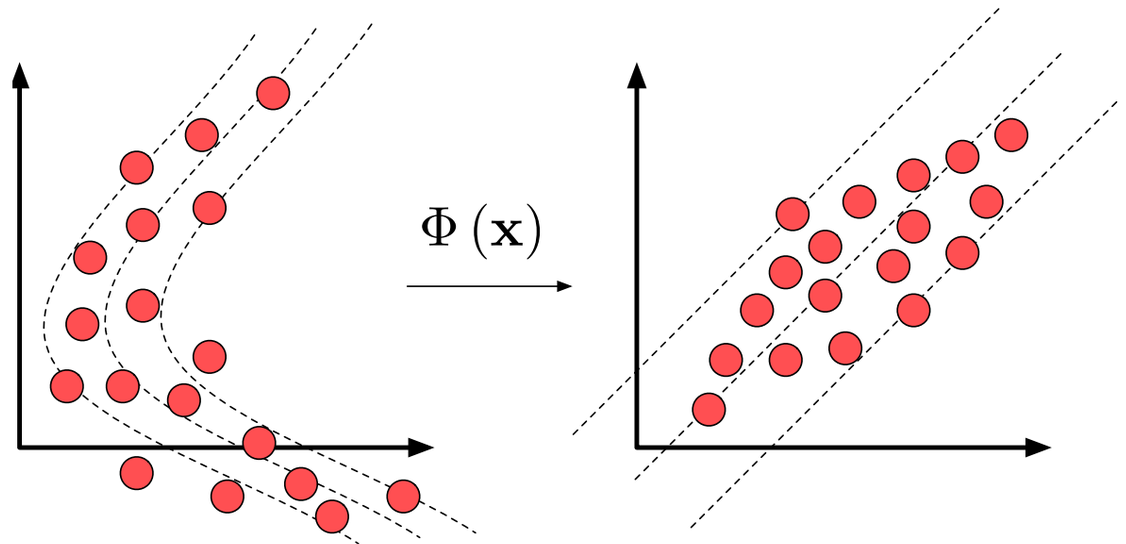
$$(\mathbf{x}_i, \mathbf{x}_j) \mapsto k(\mathbf{x}_i, \mathbf{x}_j)$$

Passage to Feature Space

$$\Phi : \mathcal{X} \rightarrow \mathcal{H}$$

$$\mathbf{x} \mapsto \Phi(\mathbf{x})$$

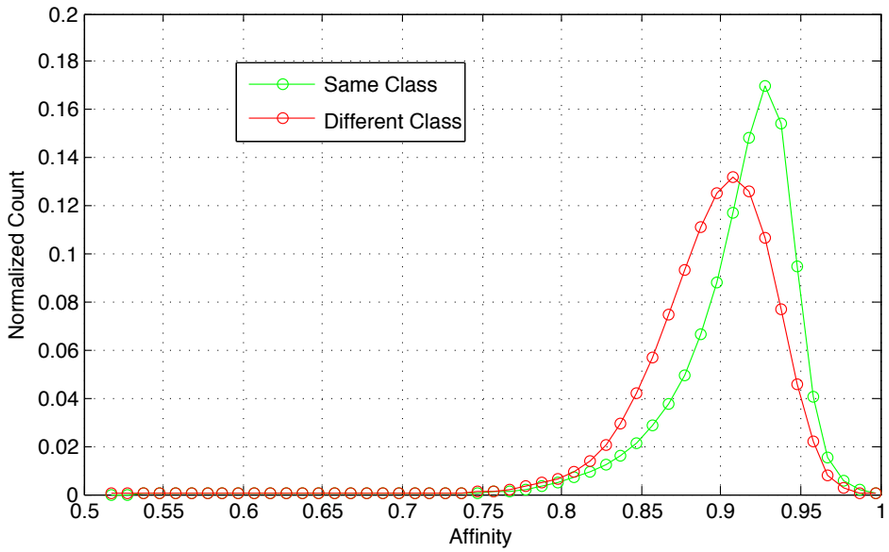
$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$$



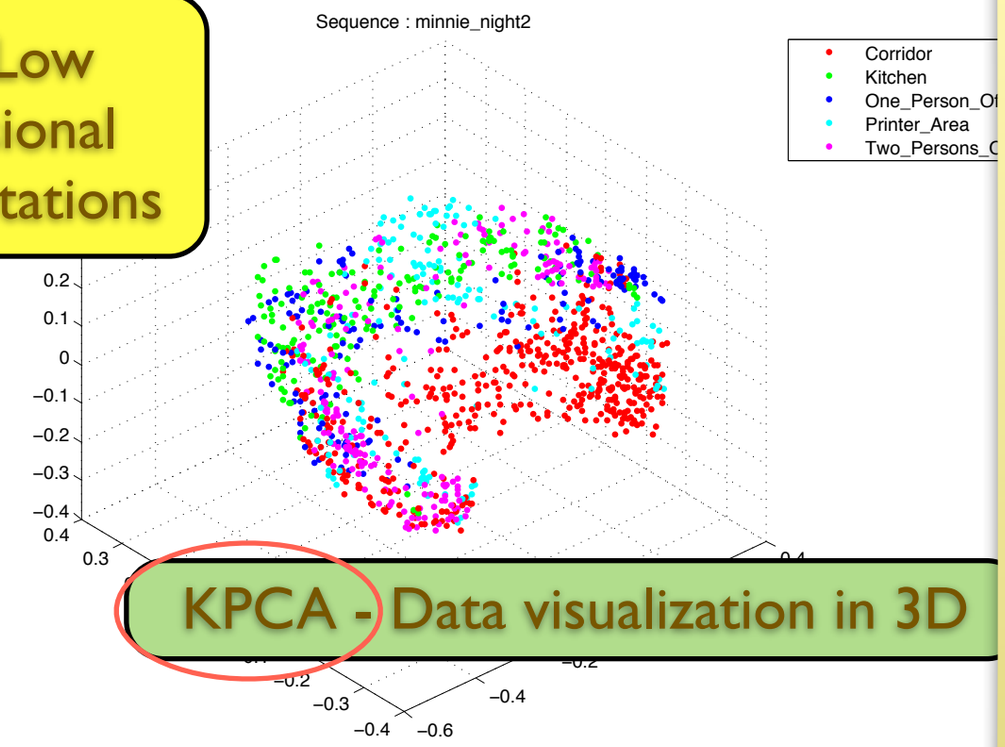
Original space Feature space

Experimental Evaluation

Kernel Gram Matrix values



Using Low Dimensional Representations



KPCA - Data visualization in 3D

(1) Atten
differen
matrix a

(2) How
the spac
figure

(3) Speal
kernel f
features
=> chi-sq
=> KPCA

(4) Point
intrinsic
dimensio
=> 500-1
represen
used

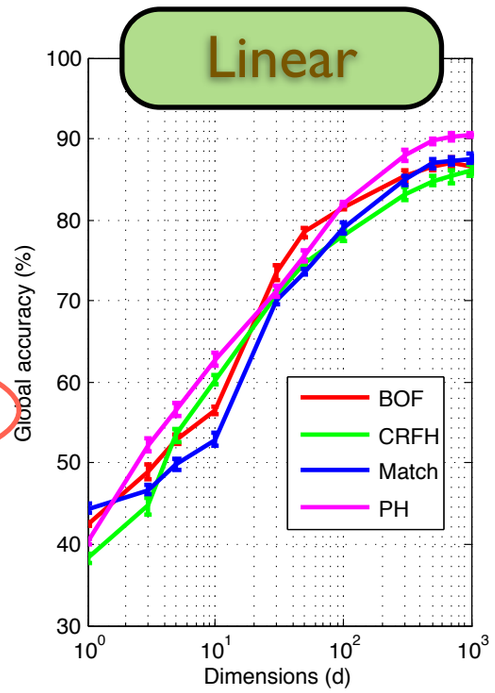
Search for adapted kernel

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle$$

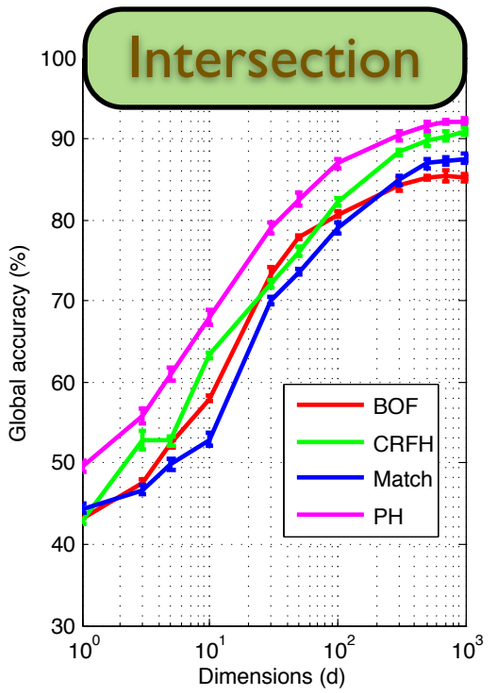
$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = 1 - 2\sum_{k=1}^{n} \frac{\left(\mathbf{x}_i\left[k\right] - \mathbf{x}_j\left[k\right]\right)^2}{\left(\mathbf{x}_i\left[k\right] + \mathbf{x}_j\left[k\right]\right)}$$

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum_{k=1}^{n} \min\left(\mathbf{x}_i\left[k\right], \mathbf{x}_j\left[k\right]\right)$$

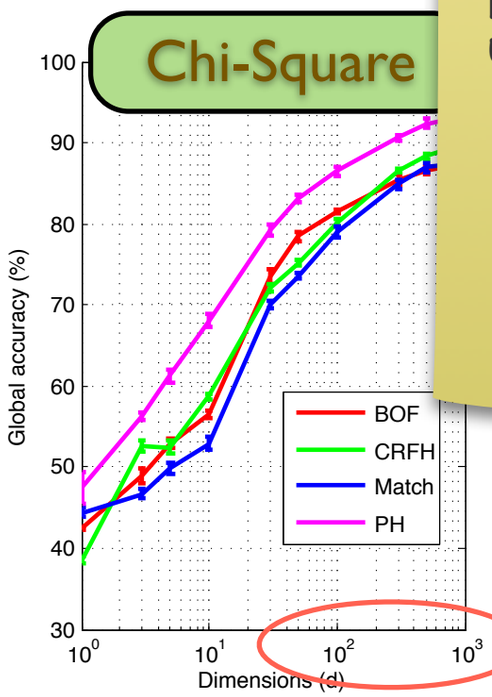
Linear



Intersection



Chi-Square



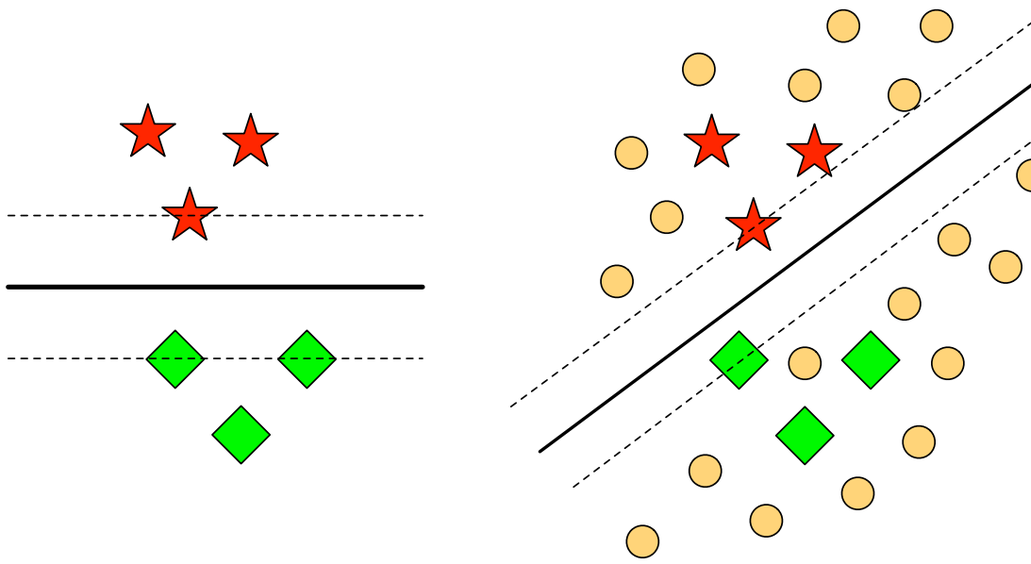
Leveraging unlabeled data

Idea - Visual Appearance model can be learned using also unlabeled images

Semi-Supervised SVM

l labeled images

u unlabeled images



(a) Supervised SVM

(b) Semi-Supervised SVM

Empirical loss

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} c\left(\mathbf{x}_i, y_i, f\right) +$$

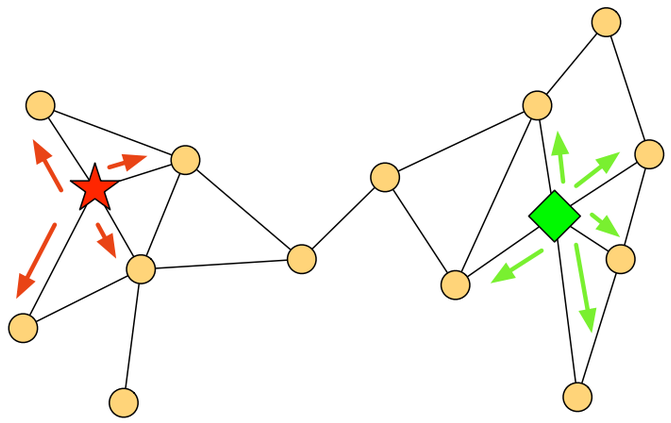
$$\gamma_A \left\| f \right\|_K^2 +$$

Ambient regularization, complexity control

$$\frac{\gamma_I}{\left(l+u\right)^2} \sum_{i,j=1}^{l+u} W_{ij} \left(f\left(\mathbf{x}_i\right) - f\left(\mathbf{x}_j\right)\right)^2$$

Locality term

Label Propagation



$$W_{\text{symm}} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

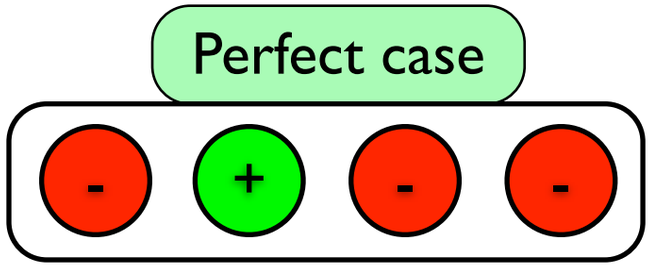
$$\mathbf{f}^{t+1} = \alpha W_{\text{symm}} \mathbf{f}^t + (\alpha - 1) Y^0$$

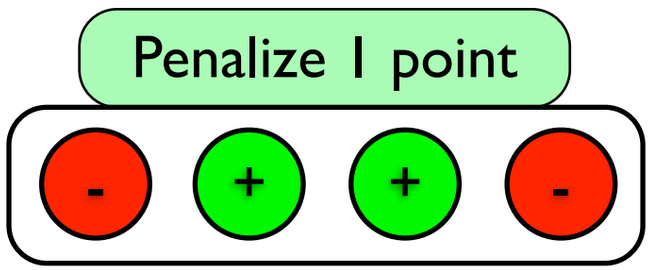
$$D_{ii} = \sum_j W_{ij}$$

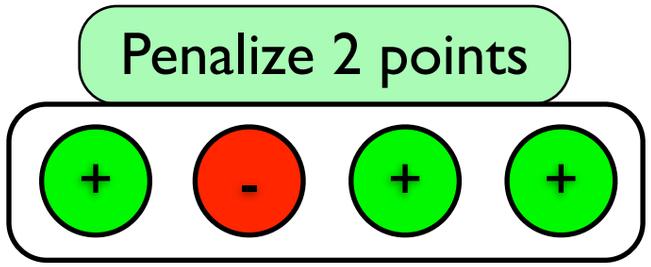
$$\hat{y}_{t+1} = \arg\max_{j \le c} \mathbf{f}^{t+1}$$

$$\mathbf{f}_i = (f_{i1}, \ldots, f_{ic})$$

Proposed Confidence Measure







SVM scores

$$f^k(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b$$

$$k^* = \arg \max_{k=1,\ldots,c} f^k(\mathbf{x}) \quad k = 1, \ldots, c$$

Confidence value

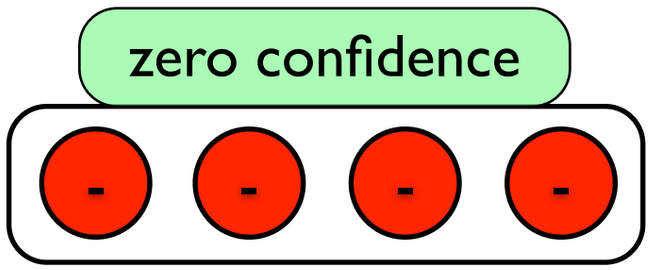
$$z_i^0 = f^{j^*}(\mathbf{x}_i) - \max_{k=1,\ldots,c,k\neq j^*} f^k(\mathbf{x}_i)$$

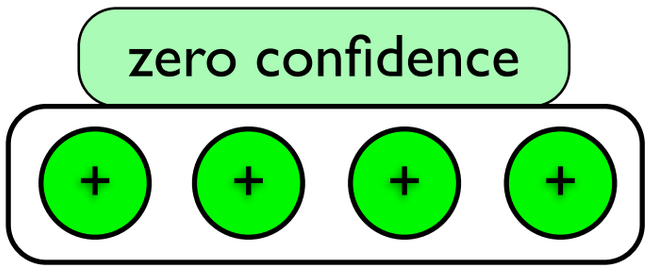
$$z_i = z_i^0 \max\left(0, 1 - \frac{p_i - 1}{c}\right)$$

$$p_i = \text{card}\left(\left\{k = 1, \ldots, c | f^k(\mathbf{x}_i) > 0\right\}\right)$$

Deducing confidence

Penalizing class overlap





Class overlap penalization allows to exclude confusing estimations

V. Dovgalecs, R. Mégret, and B. Yannick, "**Time-aware Co-Training for Indoors Localization in Visual Lifelogs**," *ACM International Conference on Multimedia*, pp. 1–4, Jul. 2011.