Exploiting Multimodal Data for Image Understanding

Matthieu Guillaumin

Supervised by Cordelia Schmid and Jakob Verbeek

27/09/2010

Multimodal data

- Webpages with images, videos, ...
- Videos with sound, scripts and subtitles, ...

Images with user tags

- Leverage user tags available on flickr or other sources:

Tags

- wow
- **San Fransisco**
- **Golden Gate Bridge**
- SBP2005
- top-f50
- **fog**
- SF Chronicle 96 hours

News images with captions

- Exploit YAHOO! NEWS to identify persons, retrieve images, ...

An Iranian reads the last issue of the Farsi-language Nowruz in Tehran, Iran Wednesday, July 24, 2002. An appeals court on Wednesday confirmed the sentence banning Iran's leading reformist daily Nowruz from publishing for six months and its publisher, Mohsen Mirdamadi, who is President Mohammad Khatami's ally, from reporting for four years. Mirdamadi is head of the National Security and Foreign Policy Committee of the Iranian parliament. (AP Photo/Hasan Sarbakhshian)

Chanda Rubin of the United States returns a shot during her match against Elena Dementieva of Russia at the Hong Kong Ladies Challenge January 1, 2003. Rubin beat Dementieva 6-4 6-1. (REUTERS/Bobby Yip)

Use of multimodal data

- As additional features for classification,
- As labels for training (weak supervision),
- Or to build large collections of images automatically.

Outline

1. Introduction
2. Face verification
   - Logistic discriminant metric learning
   - Experiments
3. News images with captions
   - Graph-based approach for face naming
   - Multiple-instance metric learning
4. Images with user tags
   - Nearest neighbor image auto-annotation
   - Experiments
5. Multimodal classification
6. Conclusion

Outline

Visual verification

- Decide whether two faces images depict the same individual.

Visual verification

- Decide whether two faces images depict the same individual.

Related work

On face recognition:

- Eigenfaces [Turk and Pentland, 1991]
- Fisherfaces [Belhummeur *et al.*, 1997]

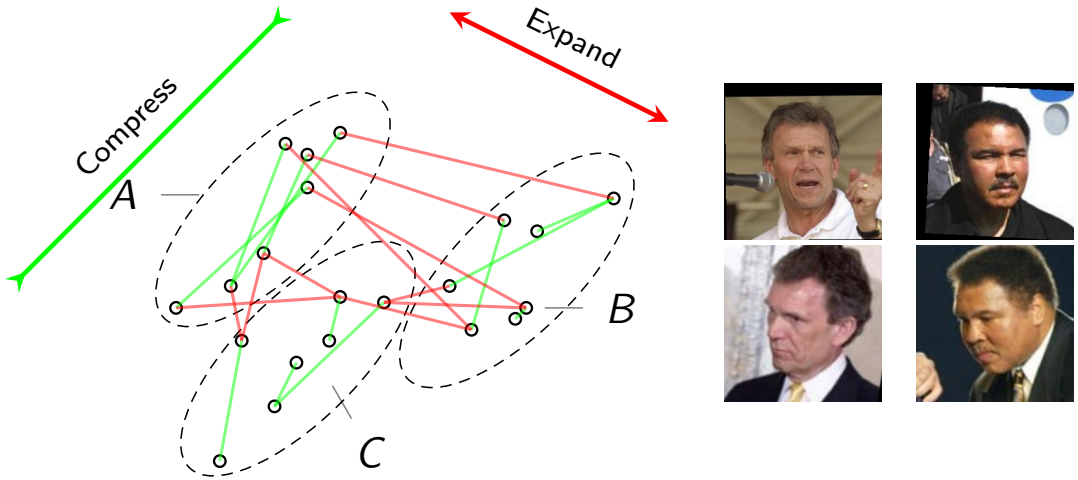
On visual verification:

- Patch sampling + Forest + SVM [Nowak and Jurie, 2007]
- One-shot similarities [Wolf *et al.*, 2008]
- Many low-level kernels + MKL [Pinto *et al.*, 2009]

"Is that you? Metric learning approaches for face identification" [Guillaumin, Verbeek and Schmid, ICCV 2009]

Mahalanobis metric learning

- Make positive pairs closer than negative pairs



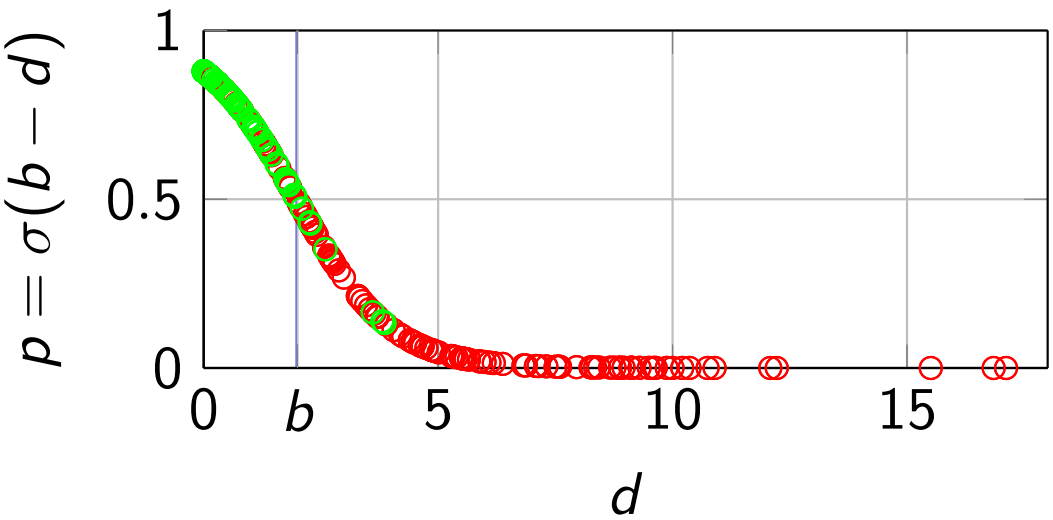
- Mahalanobis metrics $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)$, where $\mathbf{M}$ is positive semidefinite (PSD).
- LMNN [Weinberger *et al.*, 2005], ITML [Davis *et al.*, 2007], MCML [Globerson and Roweis, 2005], ...

Logistic discriminant metric learning (LDML)

- Model the probability of $(\mathbf{x}_i, \mathbf{x}_j)$ to have the same label as:

$$p_{ij} = p(y_i = y_j | \mathbf{x}_i, \mathbf{x}_j; \mathbf{M}, b) = \sigma(b - d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j))$$

where $\sigma(z) = 1/(1 + \exp(-z))$.



Logistic discriminant metric learning (LDML)

- Find $\mathbf{M}$ and b to maximize the likelihood on training data:

$$\mathcal{L}(\mathbf{M}, b) = \prod_{(i,j)} p_{ij}^{[y_i = y_j]} (1 - p_{ij})^{[y_i \neq y_j]}$$

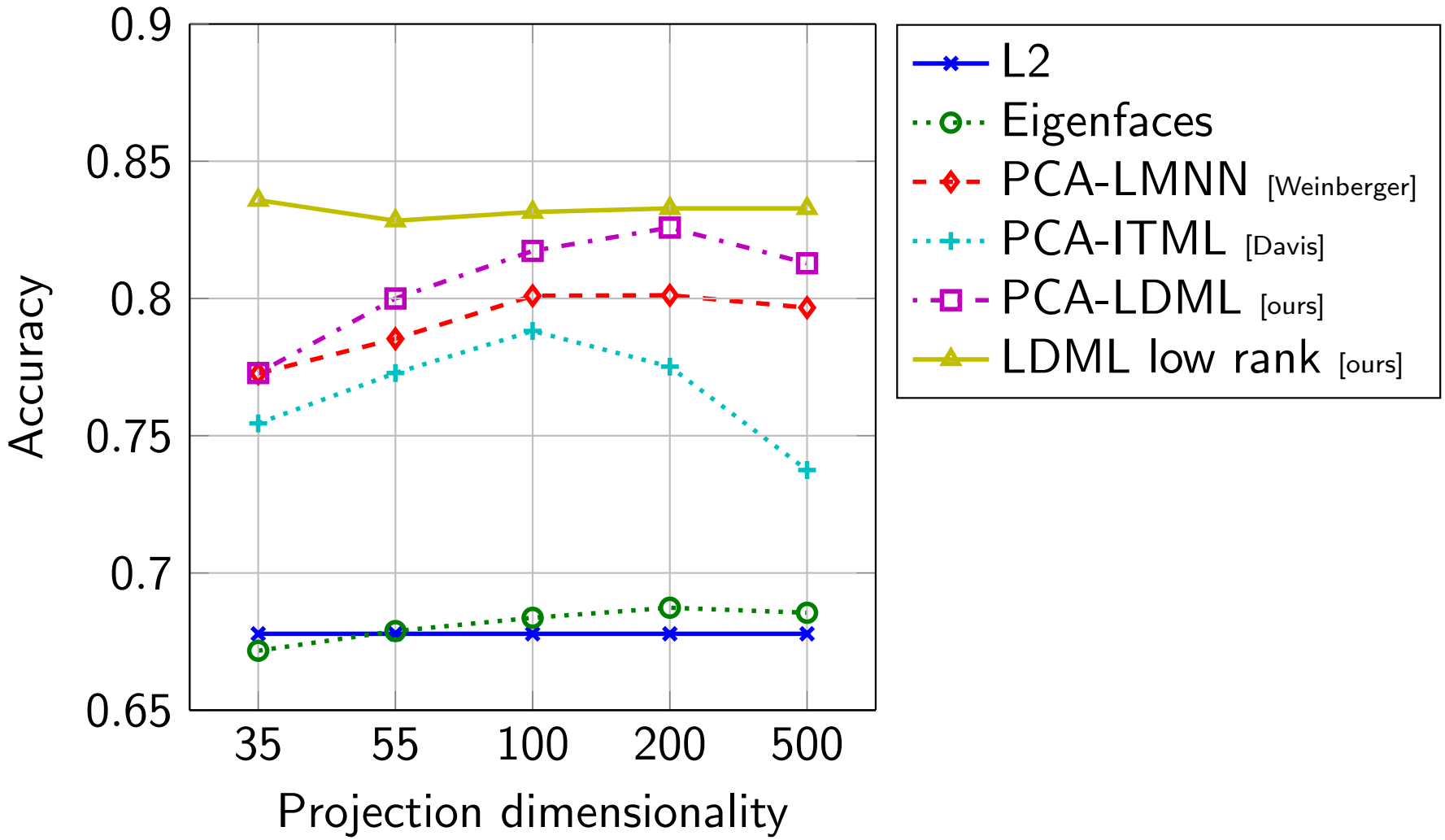
- Convex and smooth objective and convex PSD constraint:
  - Very effective optimization methods.
- Kernelizable:
  - Can handle very high dimensional data.
- Low-rank regularization:
  - Reduces the number of parameters (linear),
  - Defines a PSD matrix,
  - Supervised dimensionality reduction,
  - But: objective becomes non-convex.
- Desktop machine: $\sim 10^4$ instances of 3500d in an hour.

Outline

Data set of uncontrolled face images

- *Labeled Faces in the Wild* data set,
- 13233 images, 5749 individuals, standard evaluation protocol.
- Features: 9 locations $\times$ 3 scales $\times$ 128d SIFT $\rightarrow$ 3456d. [Everingham *et al.*, 2006]

Comparison to other metric learning



Comparison to the state of the art

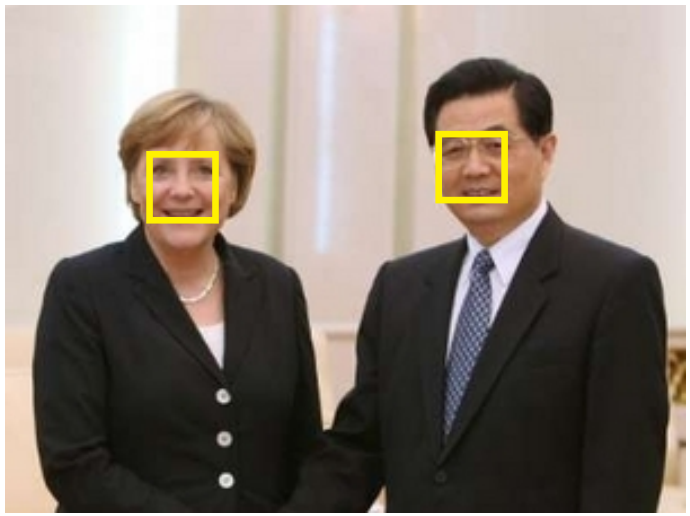
Method	Setting	Accuracy
Eigenfaces	restricted	0.600 ± 0.8
[Nowak, 2007]	restricted	0.739 ± 0.5
[Wolf, 2008]	restricted	0.785 ± 0.5
[Pinto, 2009]	restricted	$\mathbf{0.794 \pm 0.6}$
LDML [ours]	restricted	$\mathbf{0.793 \pm 0.6}$
[Kumar, 2009]	restricted*	0.853 ± 1.2
[Wolf, 2008]	unrestricted	0.793 ± 0.3
LDML [ours]	unrestricted	0.838 ± 0.6
LDML+MkNN [ours]	unrestricted	$\mathbf{0.875 \pm 0.4}$
Combined multishot [Wolf, 2009]	aligned	$\mathbf{0.895 \pm 0.5}$

* relies on additional training data.

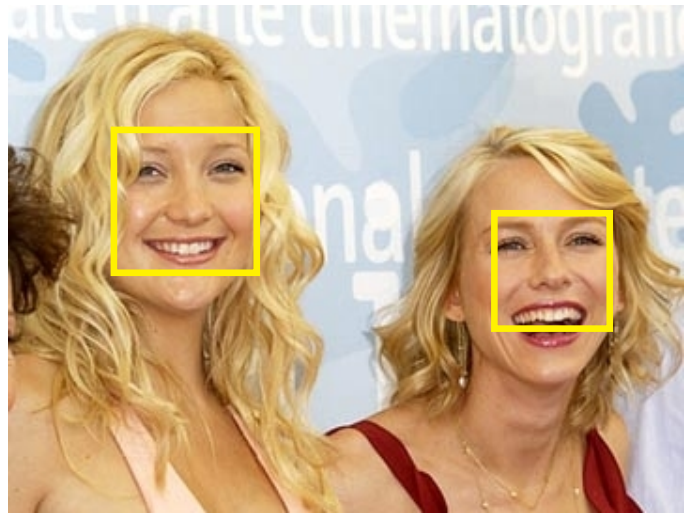
Outline

Face naming from news images

- The goal is to recover the names of the faces:



German Chancellor **Angela Merkel** shakes hands with Chinese President **Hu Jintao** (. . .)

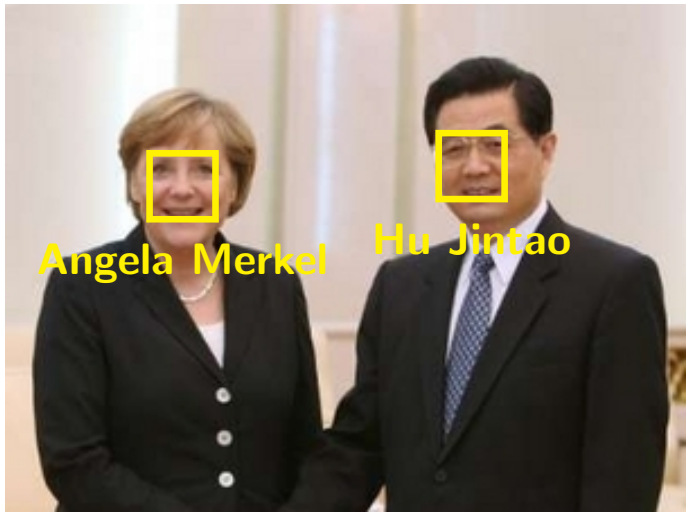


Kate Hudson and **Naomi Watts**, Le Divorce, Venice Film Festival - 8/31/2003.

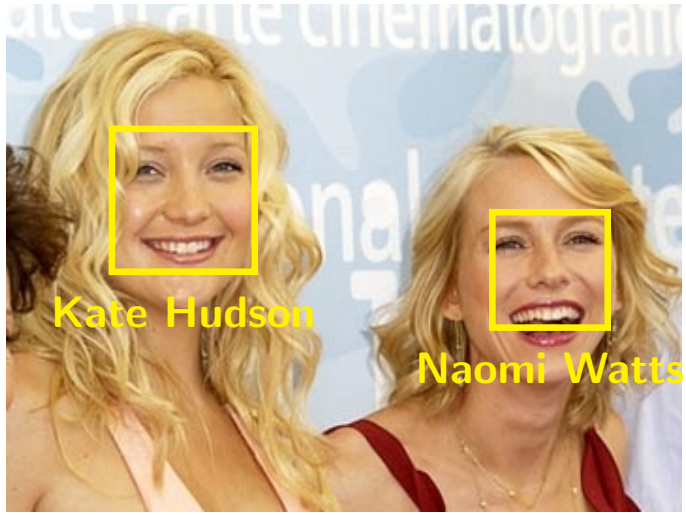
- Images as sets of faces (using face detector [Viola and Jones, 2004]),
- Captions as sets of labels (using NLP [Deschacht and Moens, 2006]).

Face naming from news images

- The goal is to recover the names of the faces:



German Chancellor **Angela Merkel** shakes hands with Chinese President **Hu Jintao** (. . .)



Kate Hudson and **Naomi Watts**, Le Divorce, Venice Film Festival - 8/31/2003.

- Images as sets of faces (using face detector [Viola and Jones, 2004]),
- Captions as sets of labels (using NLP [Deschacht and Moens, 2006]).

Related work

On associating names and faces (videos):

- Name-It system [Satoh *et al.*, 1999]
- Video Google Faces and automatic naming in videos [Everingham, Sivic and Zissermann, 2006–2009]

For still images:

- Gaussian mixture model (GMM) [Berg *et al.*, 2004–2007]
- Multimodal clustering [Pham *et al.*, 2008–2010]
- Identities and actions [Luo *et al.*, 2009]
- Graph-based method for retrieval [Ozkan and Duygulu, 2006–2010]

"Automatic face naming using caption-based supervision"
[Guillaumin, Mensink, Verbeek and Schmid, CVPR 2008]

Outline

Graph-based approach

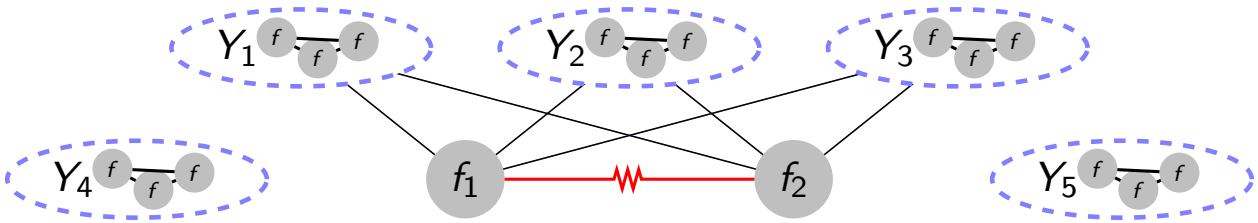
- Build a similarity graph:
  - One vertex f_i per face image,
  - Edges are weighted with a similarity w_{ij},
  - One sub-graph Y_n for each name n.
- Find the sub-graphs Y_n that maximize the sum of inner similarities:

$$\max_{\{Y_n\}} \sum_n \sum_{f_i \in Y_n} \sum_{f_j \in Y_n} w_{ij}$$

Optimization

As such, the global problem is intractable:

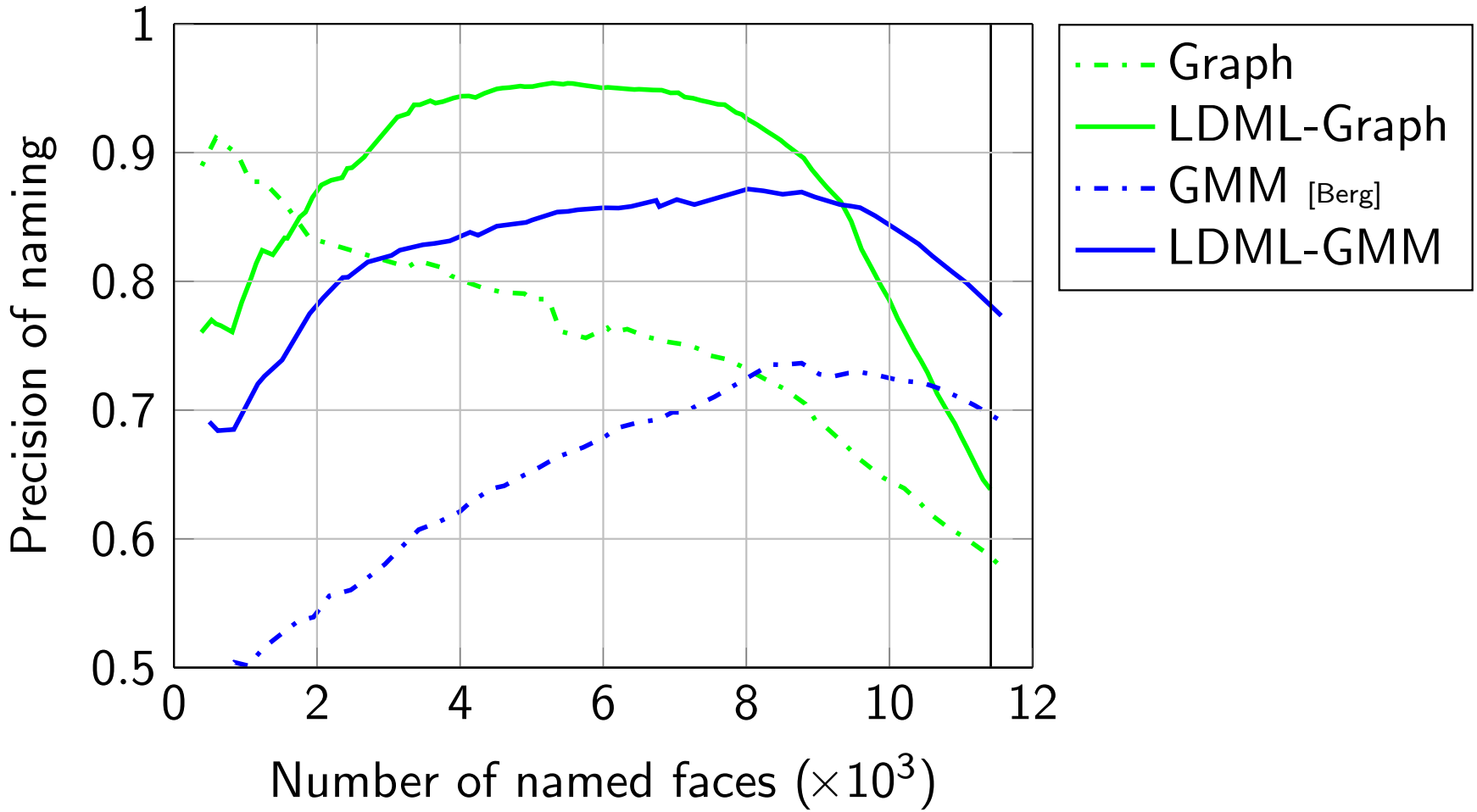
- Generally, the following holds:
  1. Faces can only be assigned to at most one name.
  2. Faces can only be assigned to a name detected in the caption.
  3. Names can only be assigned to at most one face.
- Approximate solution:
  - At document level, match detected faces with detected names,
  - Can be solved exactly and efficiently,
  - Iteration over documents until convergence.



Data set and features

- *Labeled Yahoo! News*, with around 28.000 documents.
- Manually annotated.
- Same features as previous section.
- Study influence of LDML on both GMM and Graph-based approach.

Results


Precision of naming
1
0.9
0.8
0.7
0.6
0.5
0
2
4
6
8
10
12
Number of named faces ($\times 10^3$)
Graph
LDML-Graph
GMM [Berg]
LDML-GMM

Outline

MildML

"Multiple-instance ML from automatically labeled bags of faces" [Guillaumin, Verbeek and Schmid, ECCV 2010]

- Extends LDML to handle sets of faces with sets of labels:
  - Define distance between pairs of images D and E:

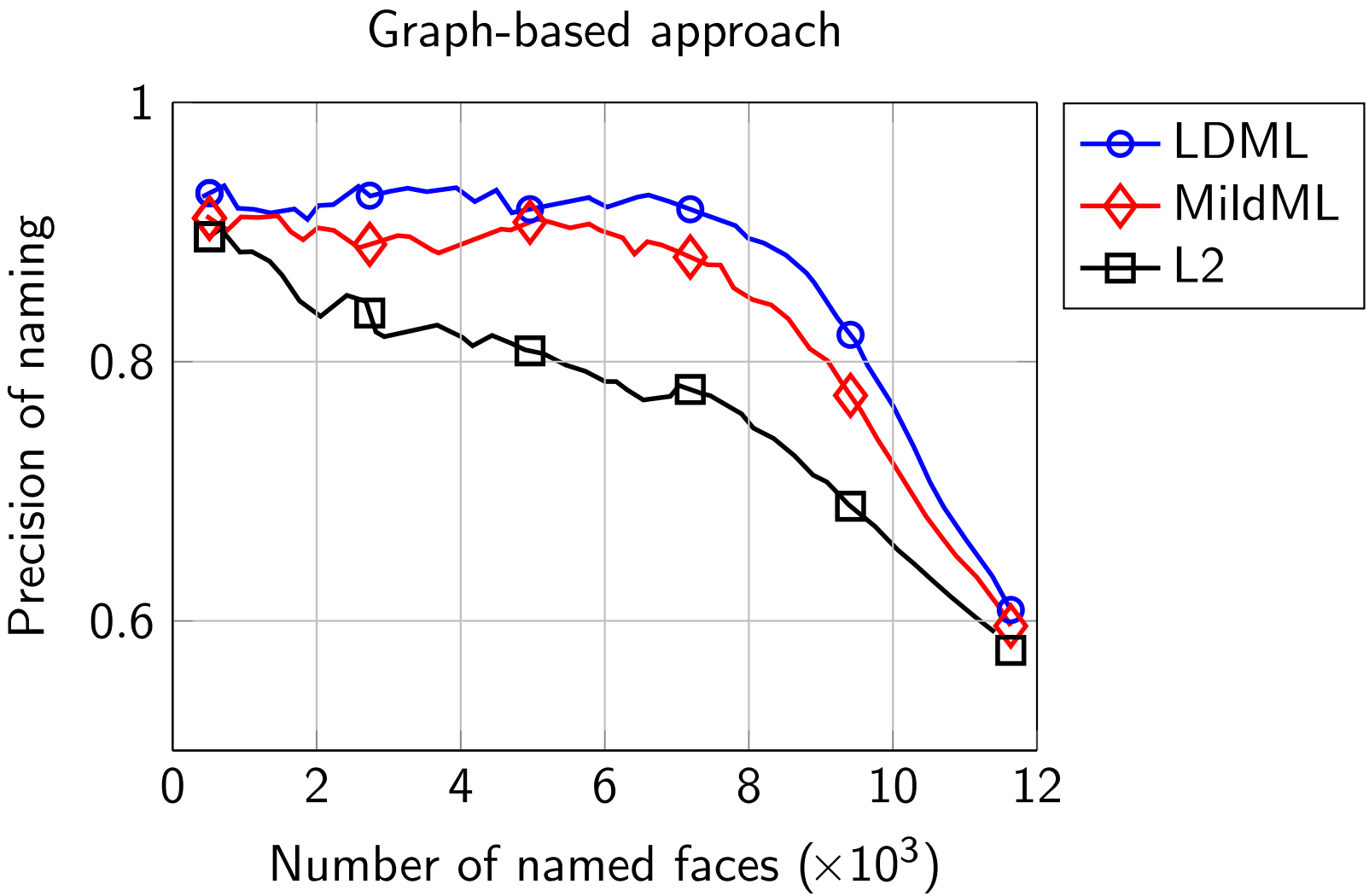
$$d_{\mathbf{M}}(D, E) = \min_{(i,j) \in D \times E} d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)$$

  - Define positivity and negativity of pairs of images by intersecting their label sets.

Johnny Depp, **Orlando Bloom**.

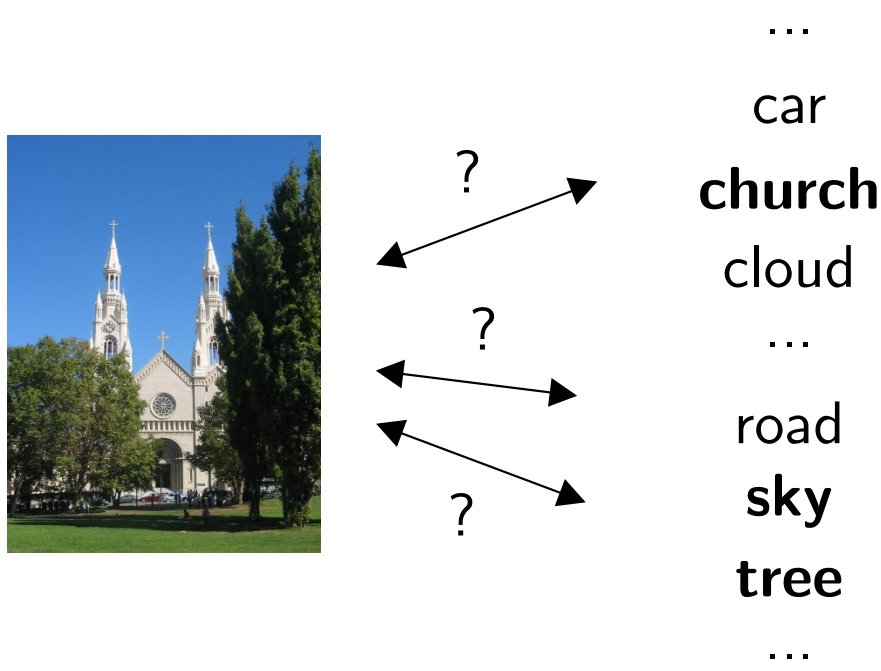
Gore Verbinski, Jerry Bruckheimer, **Johnny Depp**, Keira Knightley, **Orlando Bloom**.

Results


Graph-based approach
Precision of naming
1
0.8
0.6
0
2
4
6
8
10
12
Number of named faces ($\times 10^3$)
LDML
MildML
L2

Outline

Predicting relevance of keywords for images



- Application 1: Image annotation and retrieval
  - Propose a list of relevant keywords to assist human annotator.
  - Given one or more keywords, propose a list of relevant images.
- Application 2: Semantic embedding
  - An image is represented as a vector of word relevances.

Related work

Parametric topic models:

- Extension of PLSA or LDA [Barnard *et al.*, 2003]

Non-parametric topic models:

- Multiple Bernoulli Relevance Model [Feng *et al.*, 2004]

Discriminative methods:

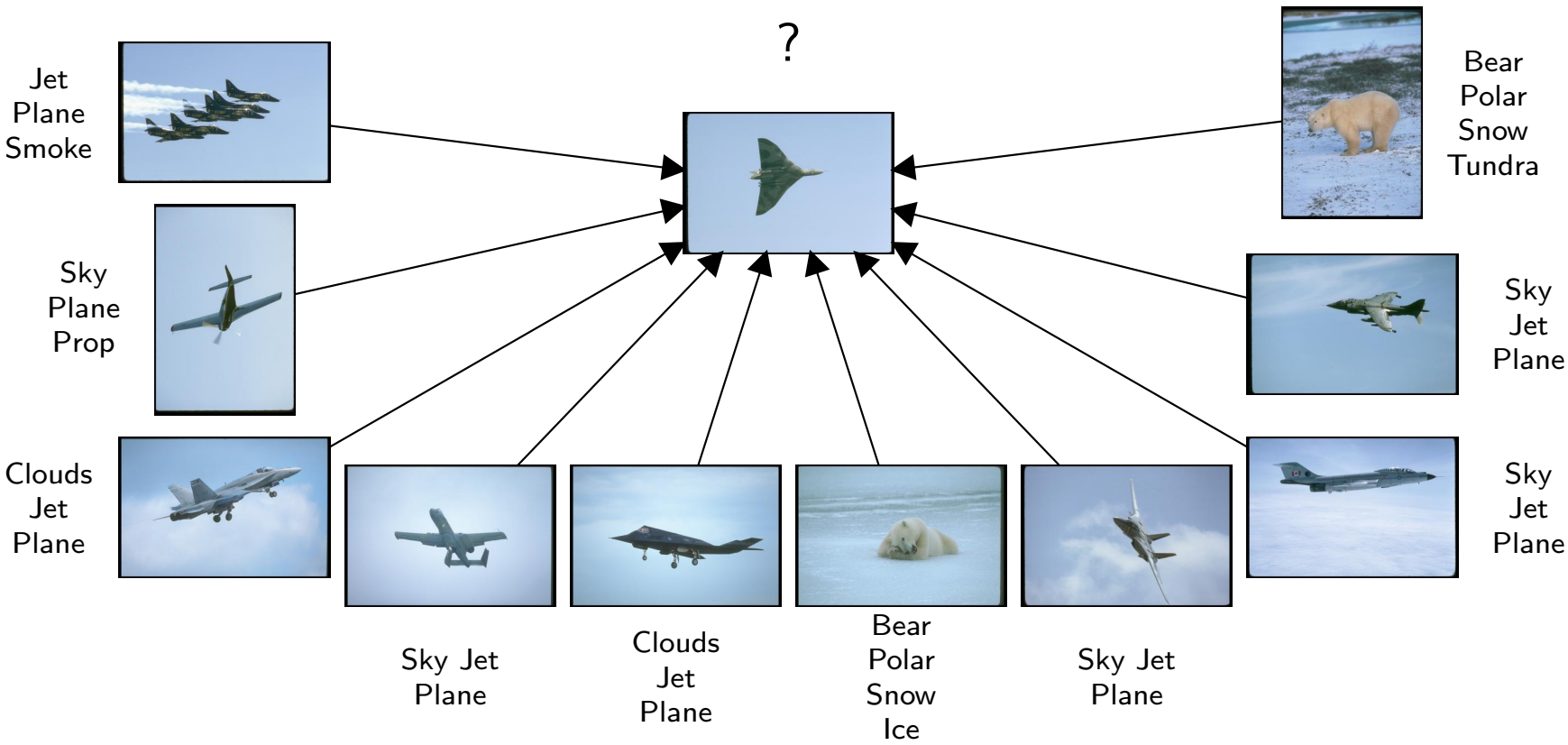
- Multiclass labeling [Carneiro *et al.*, 2007]
- PAMIR [Grangier and Bengio, 2008]

Local approaches:

- Diffusion of labels on similarity graph [Liu *et al.*, 2009]
- Nearest neighbor tag transfer [Makadia *et al.*, 2008]

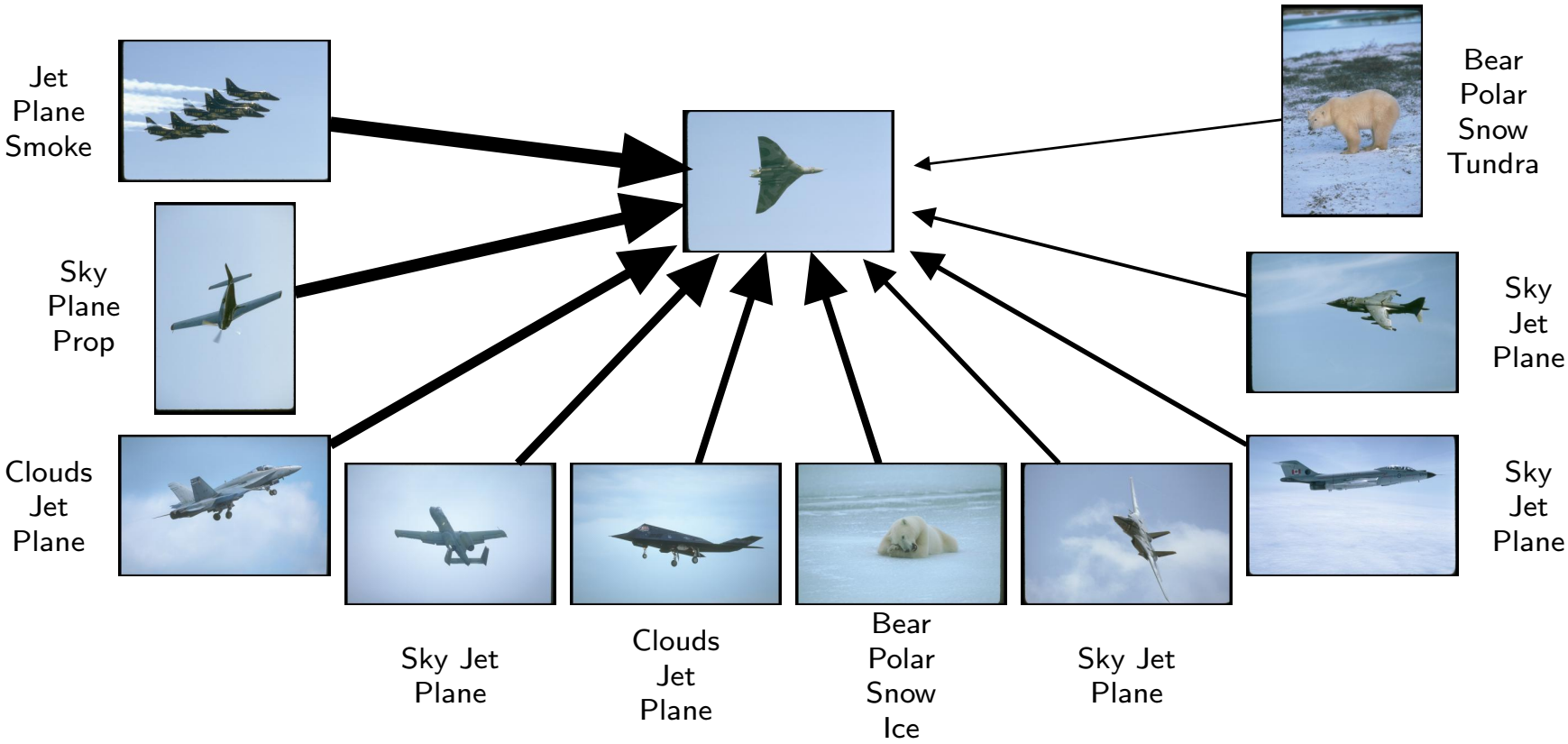
"TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation"
[Guillaumin, Mensink, Verbeek and Schmid, ICCV 2009]

TagProp: Nearest neighbor image annotation



- Learns the optimal visual distance to use to define neighbors,
- Effectively sets the number of neighbors to consider.

TagProp: Nearest neighbor image annotation

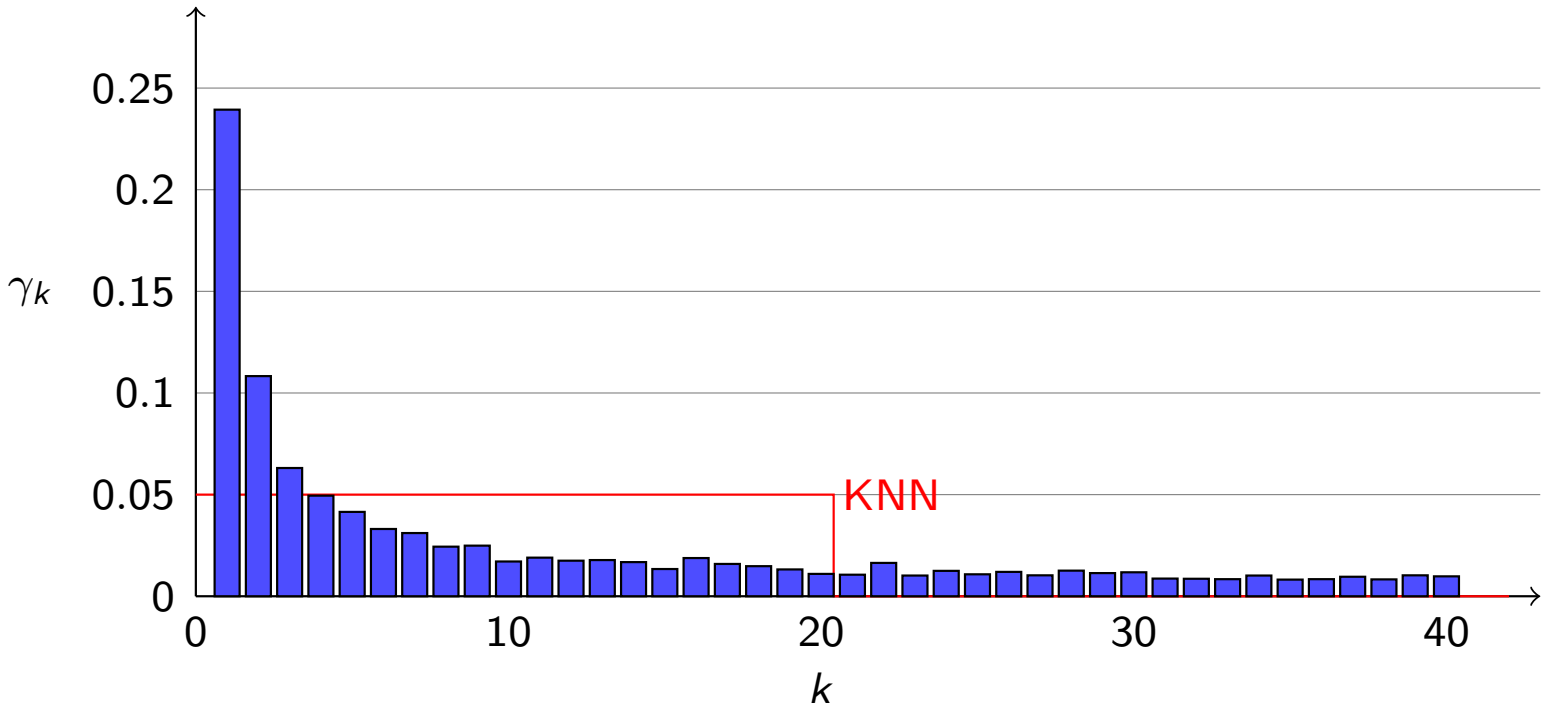


- TagProp predictions: weighted sum over neighbor images,

$$p(y_{iw} = +1) = \sum_j \pi_{ij} y_{jw}.$$

Rank-based weights

- For every image, the k-th neighbor gets fixed weight γ_k,
- There are K parameters, where K is the neighborhood size,
- Effective neighborhood size set automatically.



Distance-based weights

- Weights π_{ij} depend smoothly on d_{ij} with exponential decrease:

$$\pi_{ij} = \frac{\exp(-\lambda d_{ij})}{\sum_m \exp(-\lambda d_{im})}$$

  λ: decay rate, effective neighborhood size.
- TagProp can optimize d_{ij}, *e.g.* combining n "base" distances:

$$d_{ij} = \lambda^{(1)} d_{ij}^{(1)} + \lambda^{(2)} d_{ij}^{(2)} + \cdots + \lambda^{(n)} d_{ij}^{(n)}$$

  One parameter $\lambda^{(k)}$ for each base distance $d^{(k)}$.
- Small number of parameters, shared by all keywords.

Optimization

- Maximize the log-likelihood of predictions of training data:

$$\mathcal{L} = \sum_i \sum_w \log p(y_{iw})$$

- Rank-based: convex objective with constraints:

$$\forall k,\ \gamma_k \geq 0 \qquad \sum_{k=1}^{K} \gamma_k = 1.$$

- Distance-based: non-convex objective with constraints:

$$\forall k,\ \lambda^{(k)} \geq 0.$$

- Optimized using projected gradient descent.
- With pre-computed distances and nearest neighbors, training takes only minutes with 20000 images and 300 keywords.

Outline

Features and data sets

Set of 15 standard features with base distances:

- Color histograms / Bag-of-SIFT,
- Dense sampling / Interest points,
- Spatial pyramids: global, landscape layout,
- GIST.

Three benchmark data sets:

- Corel 5k: 5000 images, 260 words,
- ESP Game: 20000 images, 268 words,
- IAPR TC-12: 20000 images, 291 words.

Features and data sets

arctic
den
fox
grass

iguana
lizard
marine
rocks

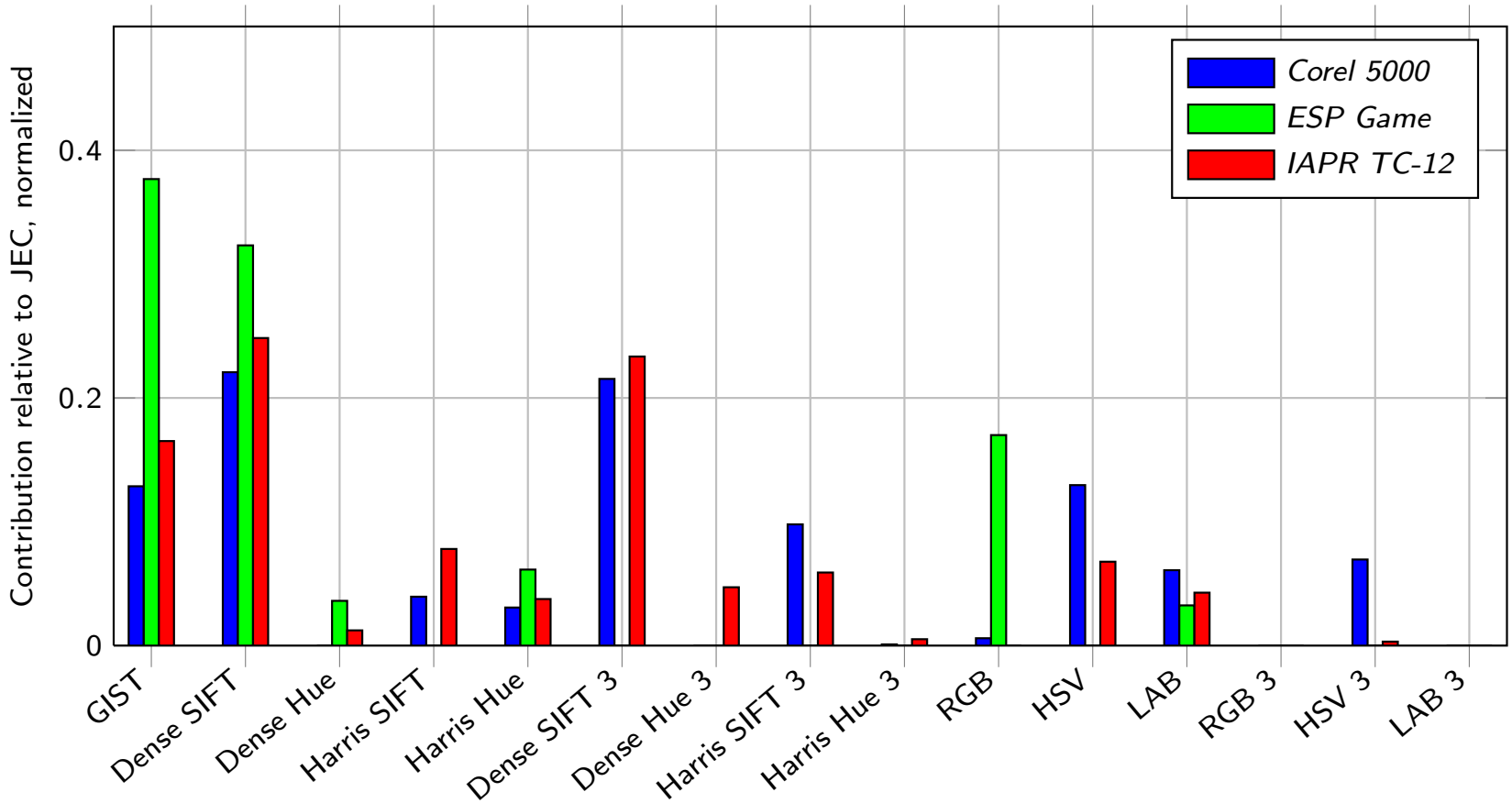
box
brown
square
white

blue
cartoon
man
woman

glacier
mountain
people
tourist

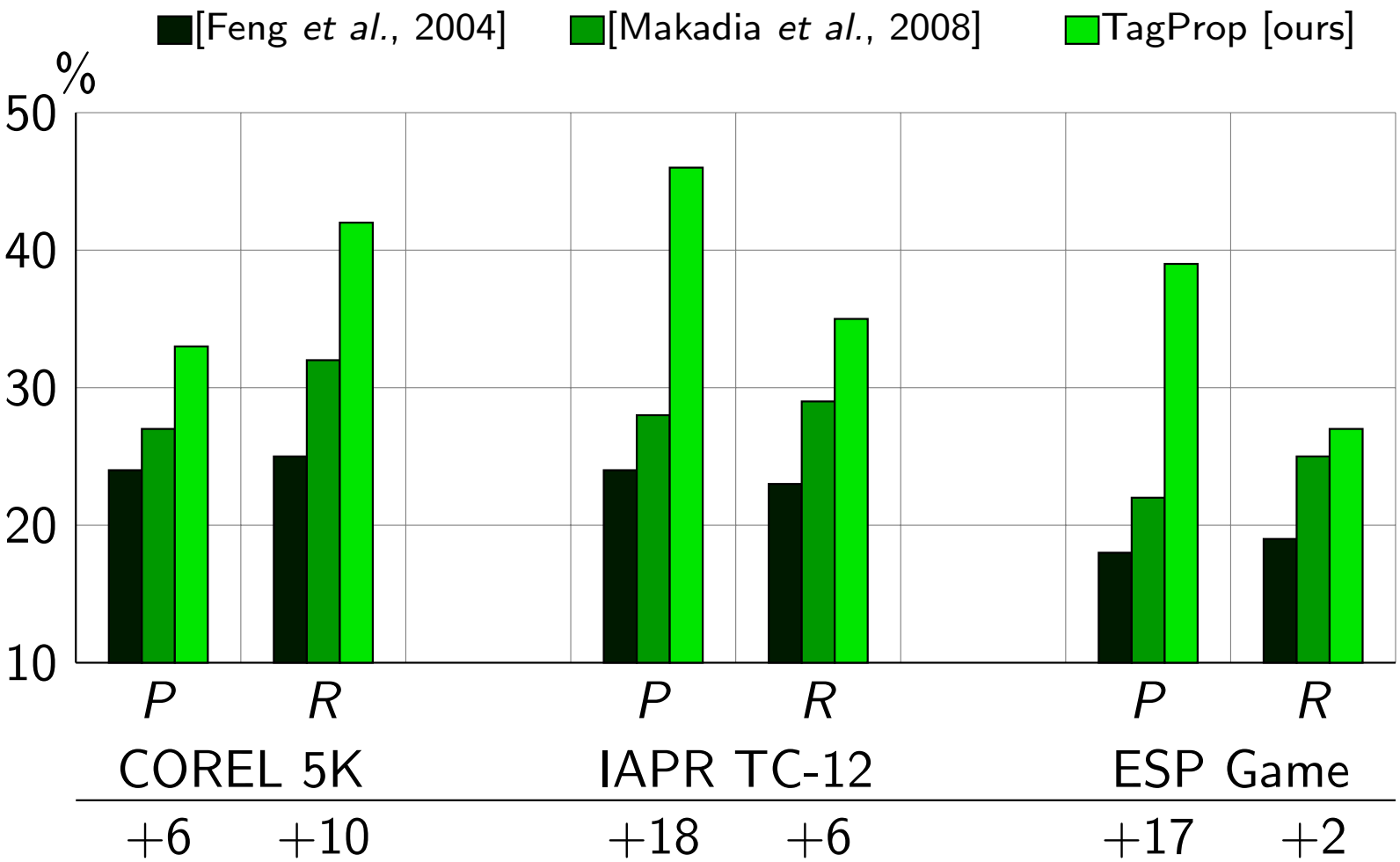
landscape
lot
meadow
water

TagProp: Learned combinations of distance



- Distance combination tends to be sparse,
- Weights differ between data sets.

Comparison to state-of-the-art



Multi-word queries

- PAMIR [Grangier and Bengio, 2008],
- 2241 queries using one or several keywords (COREL),
- Easy (≥ 3 images) vs. difficult (≤ 2).

mAP	All	Single	Multiple	Easy	Difficult
PAMIR	26%	34%	26%	43%	22%
TagProp	**36%**	**46%**	**35%**	**55%**	**32%**

- Mean average precision: +10% globally.

Outline

Multimodal classification

- Use as additionnal features for classification.
- Combine visual and textual kernels in SVM.

DOG (+1)

greyhound
running
athlete
sport

dog
rottweiler
pets

canine pet

not DOG (−1)

horse
vermont

computer
dual
monitor

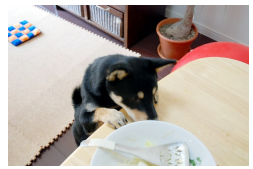
locomotive

→

DOG?

cars
racing

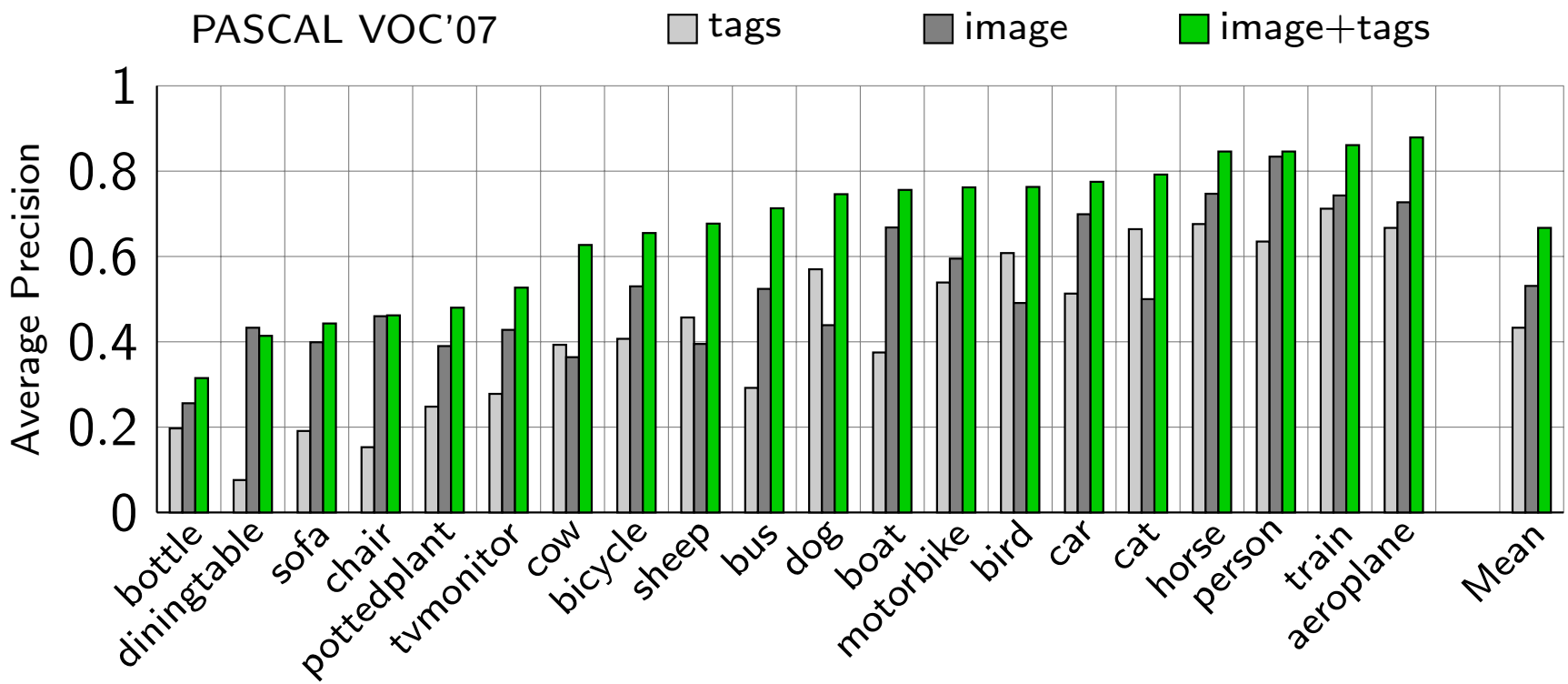
yacht



black
puppy

cute
dog

Results on PASCAL VOC 2007



- Tags (0.43) $<$ Image (0.53) $<$ Image+tags (0.67)
- Winner of PASCAL VOC'07 [Marszałek *et al.*]: 0.59.

Multimodal semi-supervised learning

- Large pool of additional unlabeled images with tags.
- Tags **NOT** available at test time: visual categorization.

DOG (+1)

greyhound
running
athlete
sport

dog
rottweiler
pets

not DOG (−1)

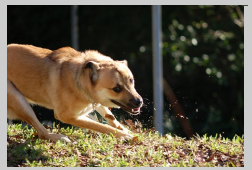
computer
dual
monitor

locomotive

Unlabeled

vermont
horse

canine
pet



puppy
dog

railroads
train

car
auto

→

DOG?

Multimodal semi-supervised learning

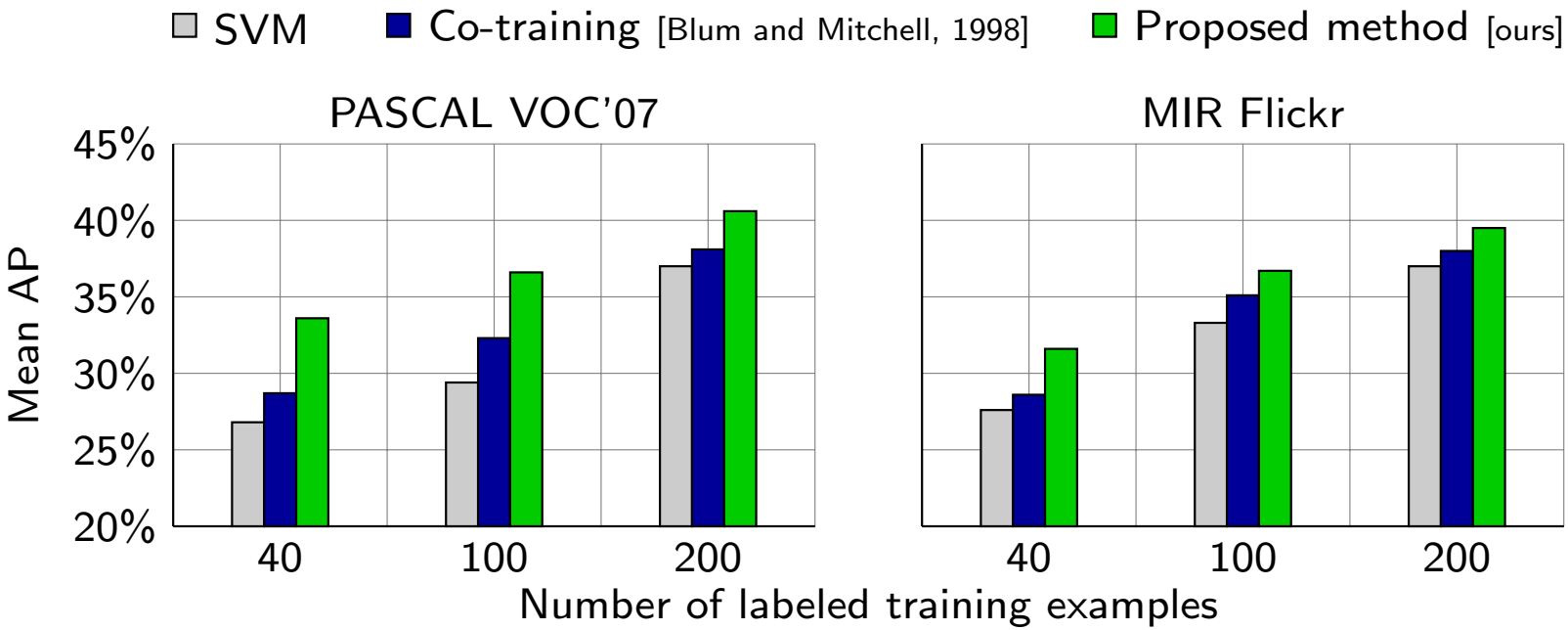
Handful of methods that can exploit this setting explicitly:

- Co-training [Blum and Mitchell, 1998]

"Multimodal semi-supervised learning for image classification" [Guillaumin, Verbeek and Schmid, CVPR 2010]

- Our proposed method, in a nutshell:
  - Learn a combined image+tags SVM on labeled data,
  - Score the unlabeled multimodal data,
  - Regress these scores with a visual scoring function on the entire training data.
- Compare with baseline visual SVM on labeled data.

Results of semi-supervised learning



- Proposed method improve over SVM and co-training.
- Especially when few training examples are used.

Outline

Contributions

- New approach for face naming using graph-based method:
  "Automatic face naming with caption-based supervision" (CVPR 2008)
- New methods for face verification using metric learning (LDML) and nearest neighbor approaches (MkNN):
  "Is that you? Metric learning approaches for face identification" (ICCV 2009)
- LDML applied to the naming problem:
  "Face recognition from caption-based supervision" (Technical Report, 2010)
- ML extended to the multiple instance learning framework:
  "Multiple instance metric learning from automatically labeled bags of faces" (ECCV 2010)

Contributions

- New model for image auto-annotation (TagProp):

  "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation" (ICCV 2009)

  "Apprentissage de distance pour l'annotation d'images par plus proches voisins" (RFIA 2010)
- Obtained excellent results at the ImageCLEF 2009 and 2010 competitions, and on the MIR Flickr data set:

  "INRIA-LEARs participation to ImageCLEF 2009" (CLEF workshop 2009)

  "Image Annotation with TagProp on the MIRFLICKR set" (ACM MIR 2010)
- Proposed method to improve visual classification using multimodal training data:

  "Multimodal semi-supervised learning for image classification" (CVPR 2010)

Conclusion

- Learning adapted similarities significantly improves recognition, clustering and classification,
- To some extent, these similarities can be learned from weak text-based supervision,
- More generally, multimodal data indeed improves visual recognition and image understanding.

Future challenges

- Improve textual analysis to "denoise" labels,
- Explore other multimodal data (*e.g.*, videos),
- Design methods for web-scale data sets: at constant annotation cost, can weakly supervised learning outperform supervised learning?

Exploiting Multimodal Data for Image Understanding

Matthieu Guillaumin

Supervised by Cordelia Schmid and Jakob Verbeek

27/09/2010