

LIRIS – 02/2015

Does Scarlett smile more than her French lover ? Constraints and regularization in visual metric learning

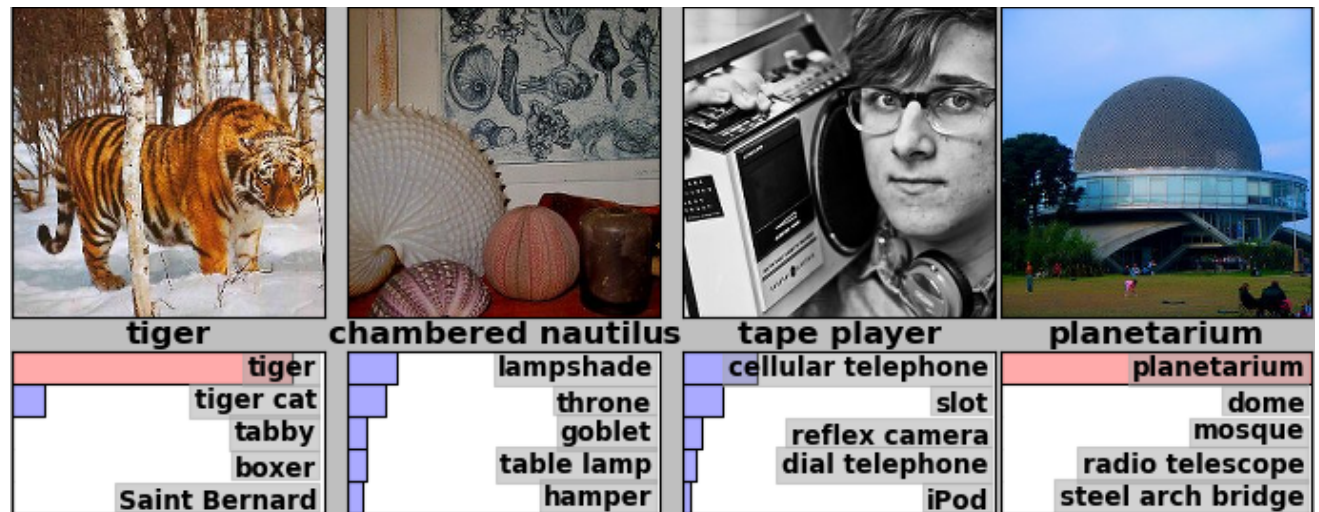
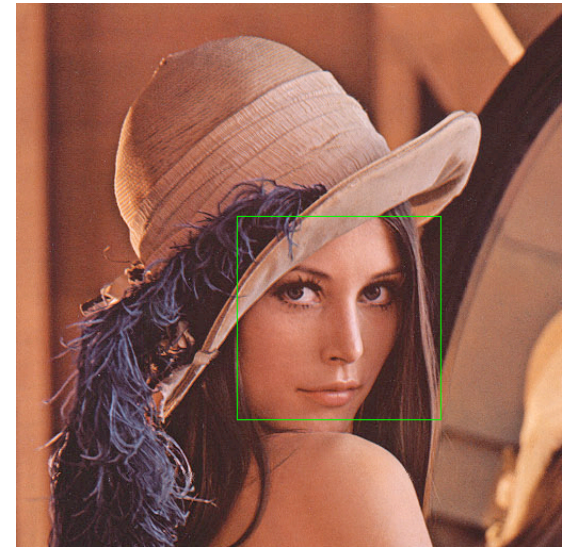
Matthieu Cord

MLIA/DAPA/LIP6
Université Pierre et Marie Curie
Comue Sorbonne Univ.

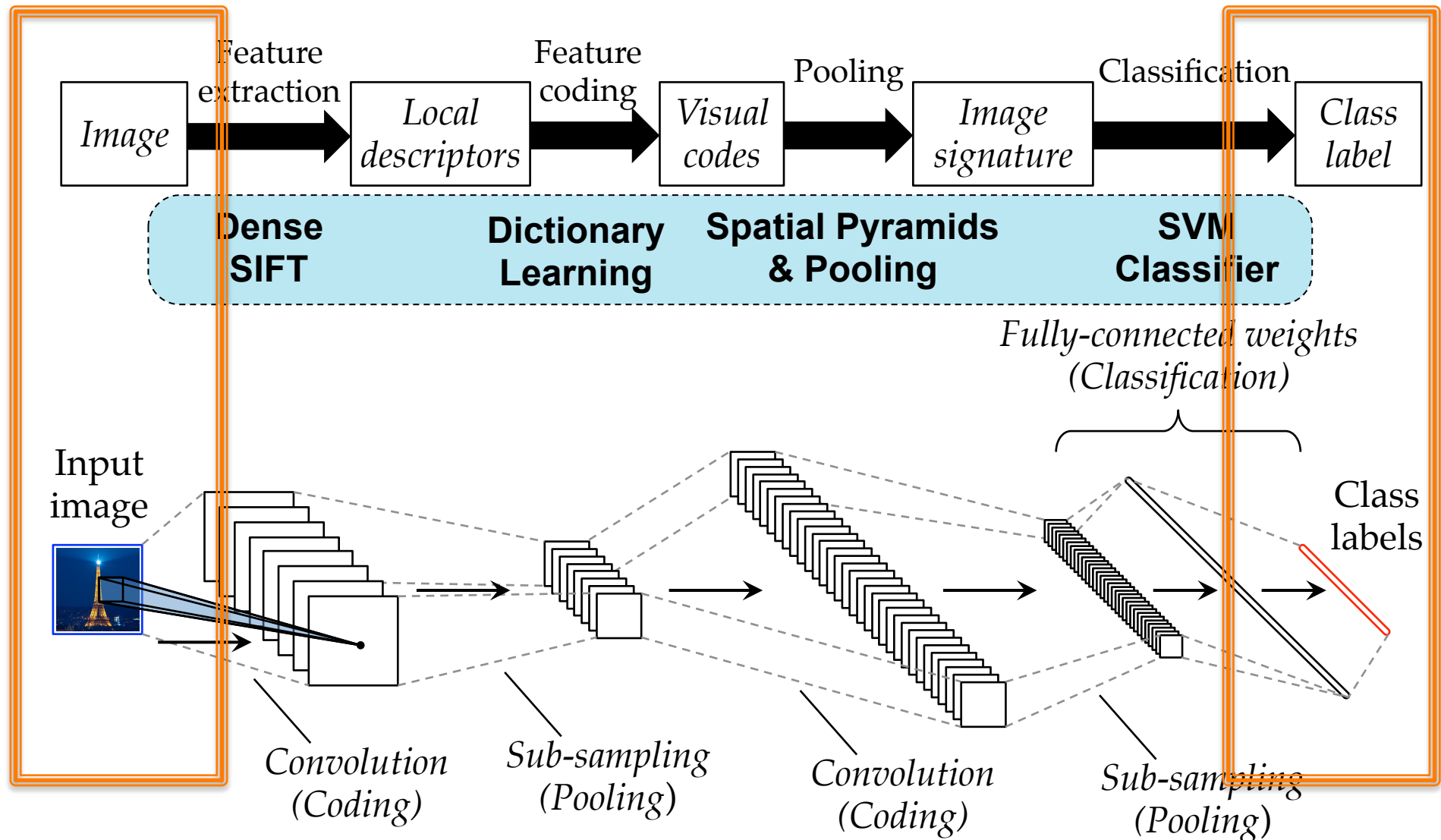


Introduction: Visual learning

- A lot of recent successful applications of Machine Learning to Visual Understanding
- Supervised classification on large dataset ImageNet
 - 1M images
 - 1000 classes



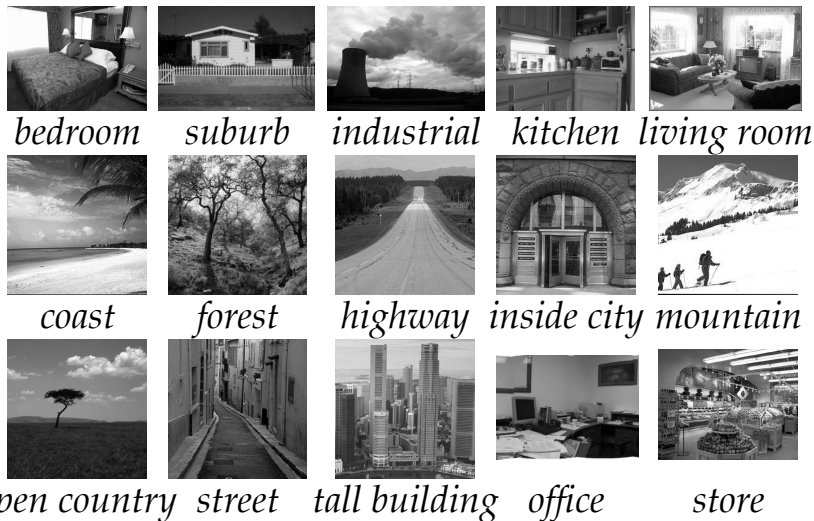
Introduction: Visual learning



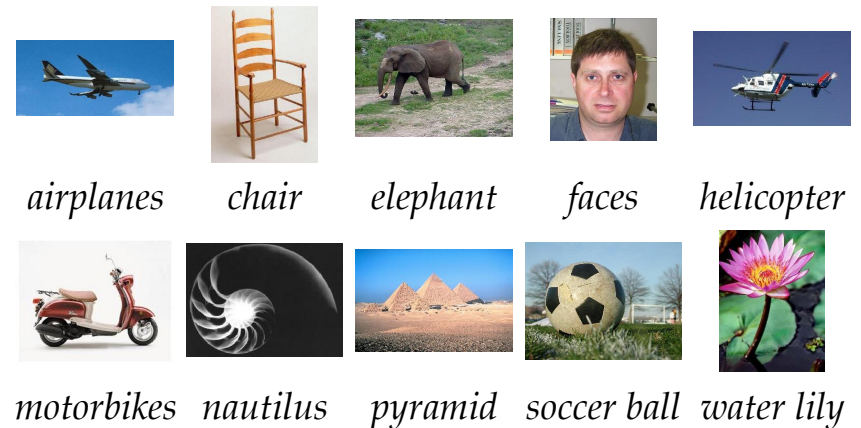
Introduction: Visual learning

- Data for training

15-Scenes

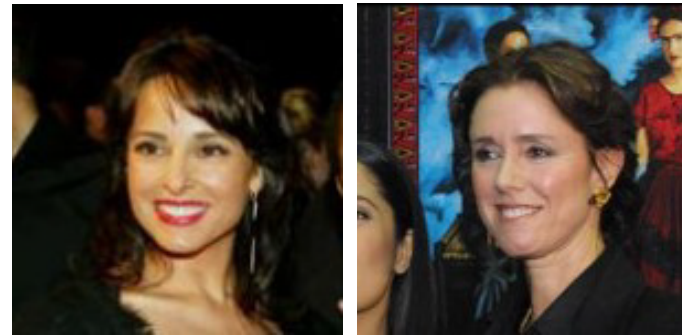


Caltech-101



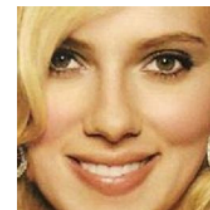
Introduction: Visual learning

- Beyond classification image+label
- Data for training : image pairs, triplets, ...
 - Pairs+label YES/NO (LFW)

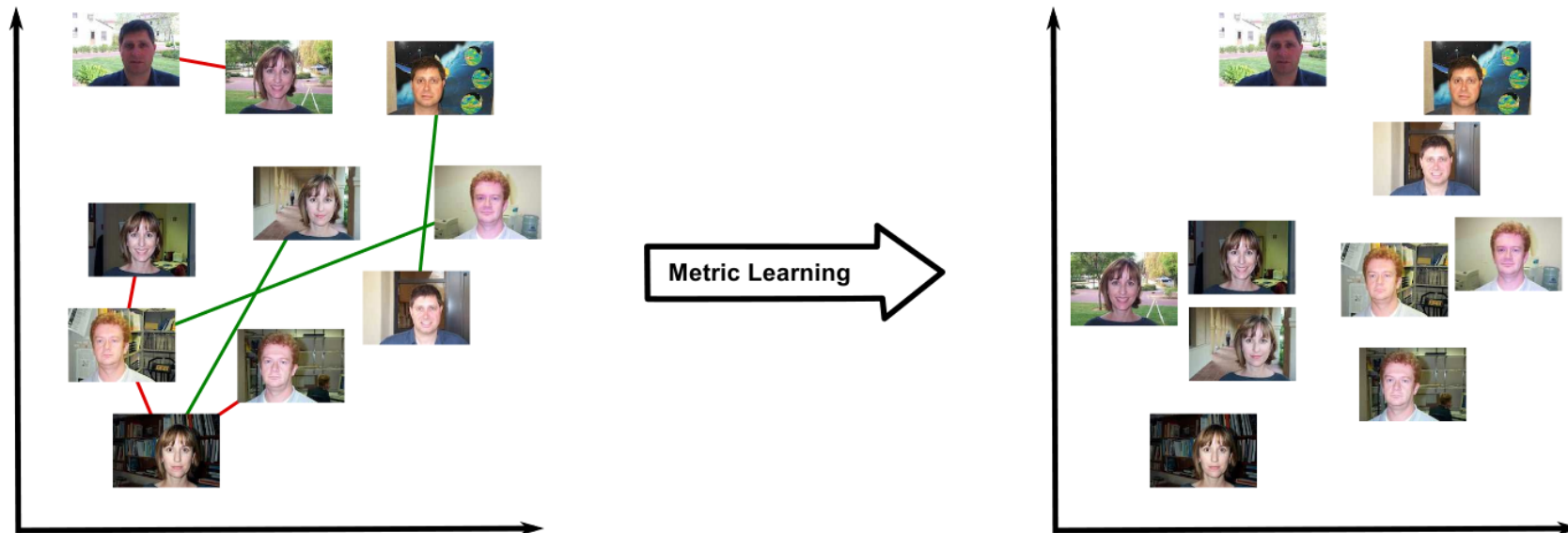


- Class information

Least smiling \prec ? \sim ? \prec Most smiling



Introduction: Metric learning



Metrics in Machine Learning and Computer Vision

- Clustering
- Information/Image retrieval
- kNN classification, Kernel methods

Commonly used metrics: Euclidean distance, chi2 for histograms, ...

[Bellet et al., A Survey on Metric Learning for Feature Vectors and Structured Data, Tech. report, 2013]

Outline

1. Introduction

2. Metric Learning

- Data and Metric models: Mahalanobis, ...
- Learning schemes:
 - ▶ Constraints :Pairs, triplets ...
 - ▶ Objective function: regularization, optimization ...
 - ▶ Examples

3. Computer Vision Applications

- Relative attribute learning
- Web page comparison

Metric Learning

- Key ingredients of metric/similarity learning:

- Data representation including both:

- ▶ Feature space

- » Gist

- » Bag of visual word representation BoW

- » Deep features

IMAGE REPRESENTATION == VECTOR

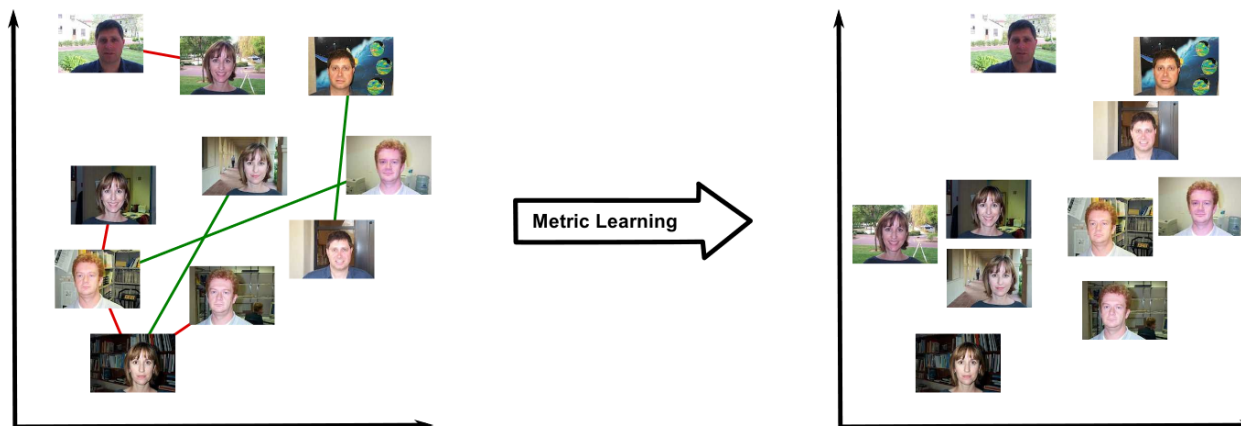
- ▶ Similarity function / Metric

- Learning framework

- ▶ training data, type of labels and relations,

- ▶ Optimization formulation

- ▶ Solvers



Metric Learning

Notations:

Vector representations $\mathbf{x} \in \mathbb{R}^d$ (visual BoWs)

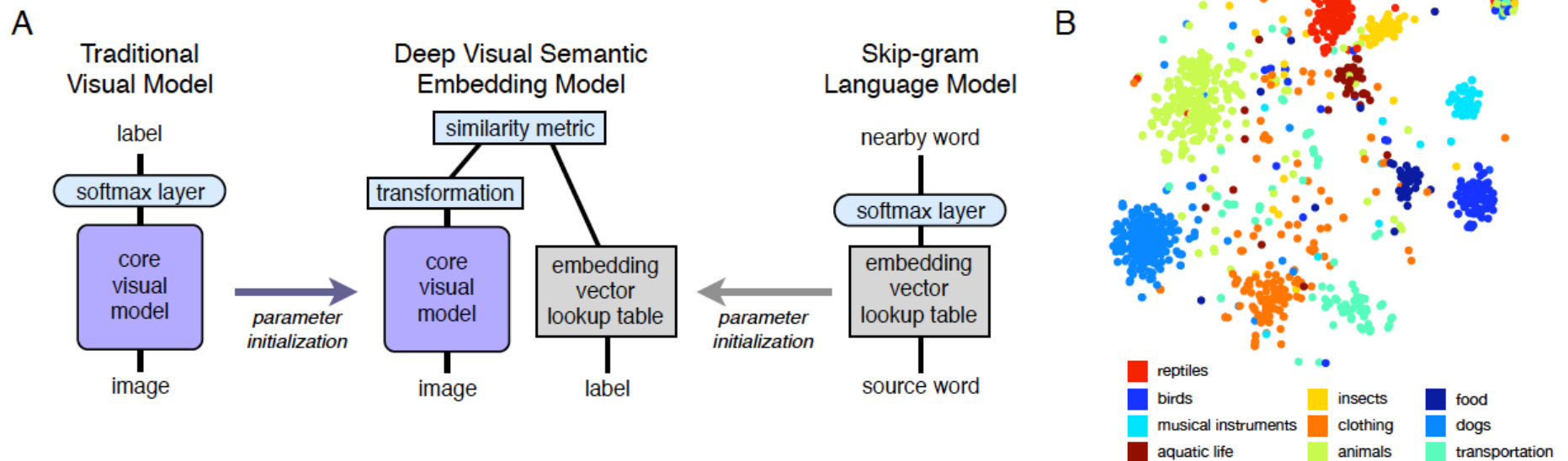
Widely used approach: **Mahalanobis-like Distance Metric Learning**

$$\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d, \mathbf{M} \in \mathbb{S}_+^d, D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) \quad (1)$$

Since for all $\mathbf{M} \in \mathbb{S}_+^d$ with $\text{rank}(\mathbf{M}) = e \leq d$, there exists $\mathbf{L} \in \mathbb{R}^{e \times d}$ such that $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$:

$$\begin{aligned} \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d, \mathbf{M} \in \mathbb{S}_+^d, D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{L}^\top \mathbf{L} (\mathbf{x}_i - \mathbf{x}_j) \\ &= \|\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j\|_2^2 \end{aligned} \quad (2)$$

Metric Learning



DeVISE system (google NIPS 2013)

- Non-linear extension
- Comparison of heterogeneous objects

Metric Learning

- PairWise Constraints for learning

Similar pairs




Dissimilar pairs




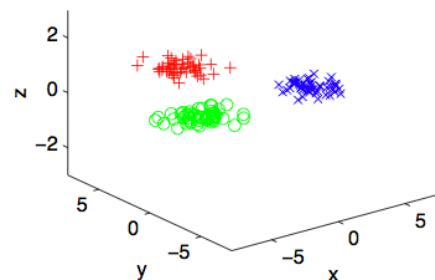
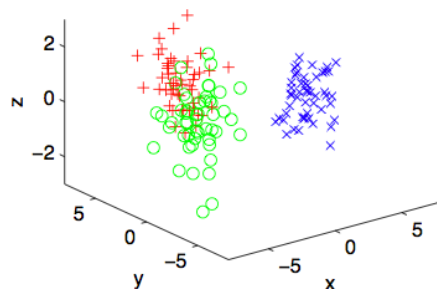
Metric Learning

- Learning scheme for pairwise constraints
- Xing et al: *Distance metric learning, with application to clustering with side-information, NIPS 2002*

$$\min_{\mathbf{M} \in \mathbb{S}_+^d} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) \quad s.t. \quad \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \sqrt{D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j)} \geq 1$$

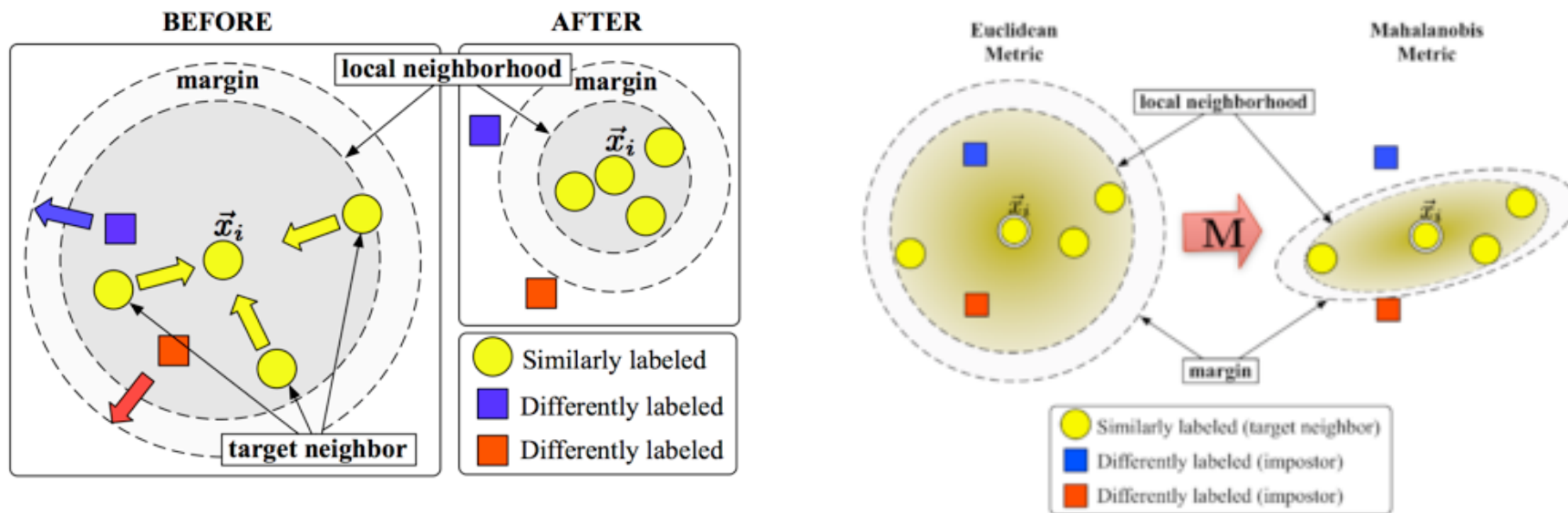






Metric Learning

- TripletWise [Weinberger LMNN NIPS06]

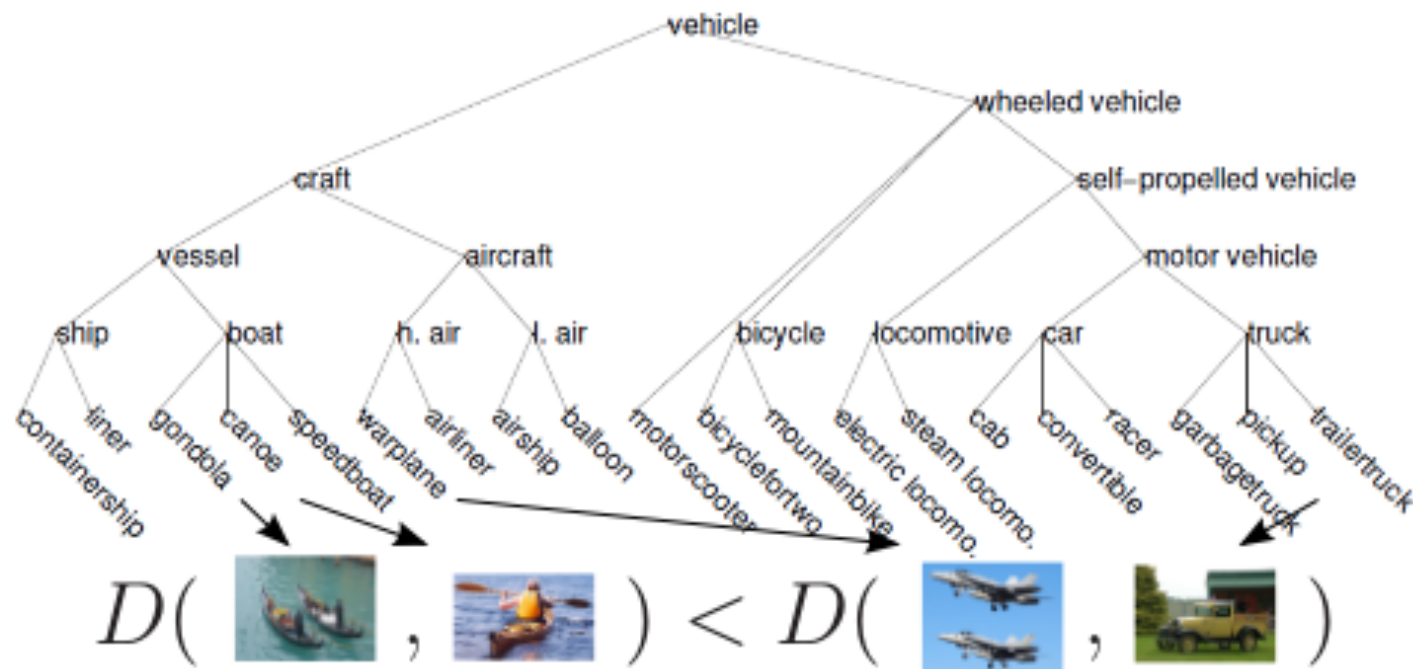


$$\min_{\mathbf{M} \in \mathbb{S}_+^d} \sum_{(\mathbf{x}_i, \mathbf{x}_i^+) \in \mathcal{S}} D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_i^+)$$

$$\text{s.t. } \forall (\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-) \in \mathcal{T}, D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_i^-) \geq \delta + D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_i^+)$$

Metric Learning

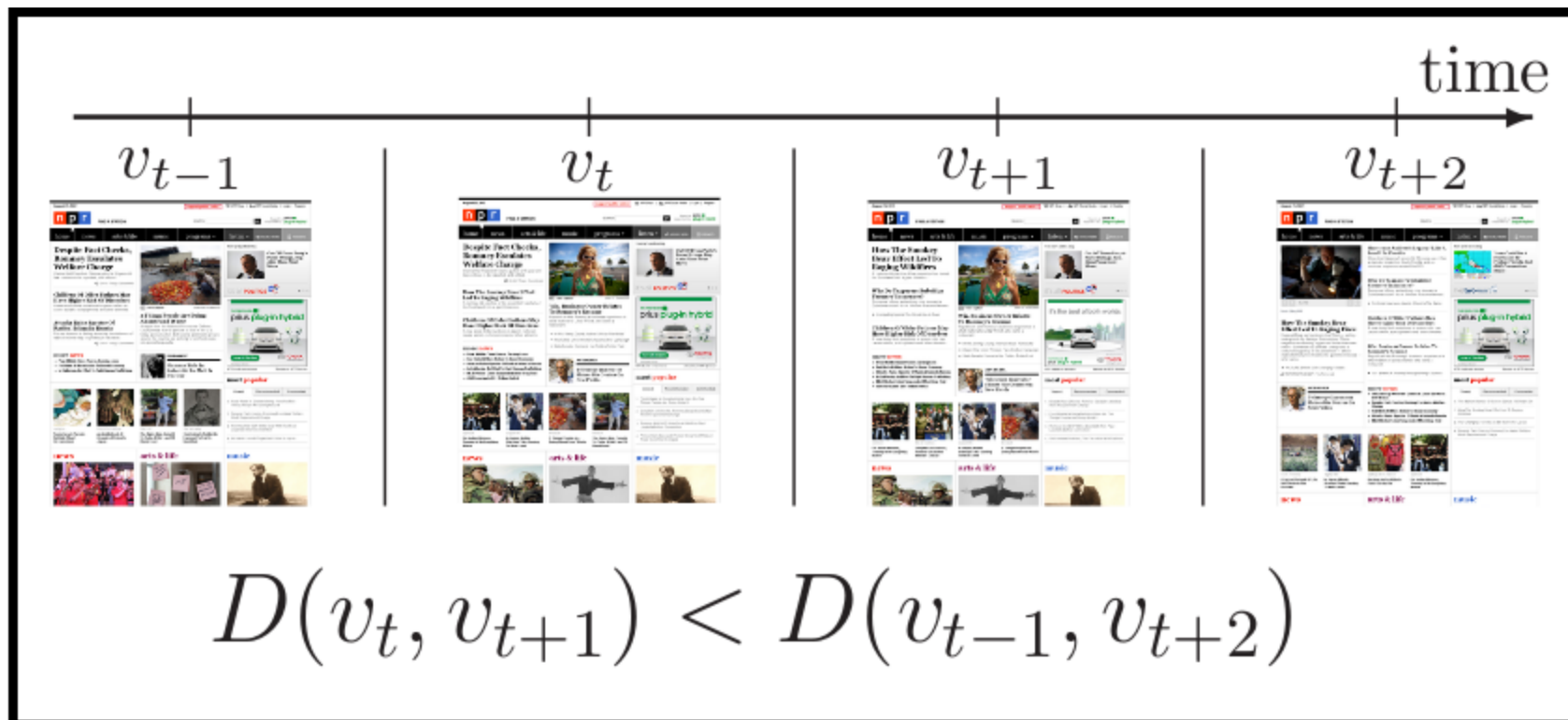
- QuadrupletWise [Law ICCV 2013] (from taxonomy):



$$\forall q = (\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) \in \mathcal{N}, \quad D(\mathbf{x}_i, \mathbf{x}_j) + \delta_q \leq D(\mathbf{x}_k, \mathbf{x}_l)$$

Web page ML

- Qwise Constraints:
 - Fully unsupervised ML, but temporal information available
 - Constraints by comparing screenshots of successive webpage versions



Metric Learning

$\ell(\mathbf{M}, \mathcal{N})$ loss over set of constraints \mathcal{N}

- **Pairs:**

$$\mathcal{N} = \mathcal{S} \cup \mathcal{D} \implies \begin{cases} \forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} & D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) < 1 \\ \forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} & D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) > 1 \end{cases}$$

- **Triples:**

$$\mathcal{N} = \{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)\}_{i=1}^N \implies \forall (\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-) \in \mathcal{N}, D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_i^+) + \delta \leq D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_i^-)$$

- **Quadruplets:**

$$\mathcal{N} = \{q = (\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)\} \implies \forall q \in \mathcal{N}, D_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) + \delta_q \leq D_{\mathbf{M}}(\mathbf{x}_k, \mathbf{x}_l)$$

Classic Mahalanobis-like distance metric problem formulation:

$$\min_{\mathbf{M} \in \mathbb{S}_+^d} \mu R(\mathbf{M}) + \ell(\mathbf{M}, \mathcal{N})$$

With $R(\mathbf{M})$: regularizer

Metric Learning

Large margin optimization

- Qwise optimization framework

$$\min_{\mathbf{M} \in \mathbb{S}_+^d} \mu R(\mathbf{M}) + C_Q \sum_{q \in \mathcal{N}} \xi_q$$

$$\text{s.t. } \forall q = (\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) \in \mathcal{N}, D_{\mathbf{M}}^2(\mathbf{x}_k, \mathbf{x}_l) \geq D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) + \delta_q - \xi_q$$
$$\forall q \in \mathcal{N}, \xi_q \geq 0$$

- $R(\mathbf{M})$: regularization term
- $C_Q > 0$: trade-off between fitting and regularization.
- Triplet optim:

$$\min_{\mathbf{M} \in \mathbb{S}_+^d} \sum_{(\mathbf{x}_i, \mathbf{x}_i^+) \in \mathcal{S}} D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_i^+) + C_t \sum_{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-) \in \mathcal{T}} \xi_i$$

$$\text{s.t. } \forall (\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-) \in \mathcal{T}, D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_i^-) \geq 1 + D_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_i^+) - \xi_i$$

Metric Learning

- Objective function

$$\min_{\mathbf{M} \in \mathbb{S}_+^d} \mu R(\mathbf{M}) + \ell(\mathbf{M}, \mathcal{N})$$

- Regularization term to express *prior*, to control complexity ...
- Low rank solution?
 - Controlling overfitting
 - Sparsity of the singular values
 - Exploiting correlation between features
 - Fast/efficient solution

Metric Learning

Formulation of $R(\mathbf{M})$

- Frobenius norm $R(\mathbf{M}) = \|\mathbf{M}\|_F^2 = \sum M_{ij}^2$
 - matrix analog of the standard ℓ_2 regularizer in SVM
 - does not promote low-rank solutions
 - useful when \mathbf{M} is a diagonal matrix
- Schultz, Learning a Distance Metric from Relative Comparisons, NIPS 2003
- Nuclear norm regularization $R(\mathbf{M}) = \|\mathbf{M}\|_* = \text{tr}(\mathbf{M})$:
 - rank NP-hard to optimize
 - convex envelope of $\text{rank}(\mathbf{M})$ on the set $\{\mathbf{M} \in \mathbb{R}^{d \times d} : \|\mathbf{M}\| \leq 1\}$
 - ℓ_1 norm of vector of singular values $\sigma(\mathbf{M})$

- McFee, Metric Learning to Rank, ICML 2010

Metric Learning

- M Law, Fantope regularization in ML, CVPR 2014:
 - Explicit control of the rank of \mathbf{M}
By noting, $\forall \mathbf{M} \in \mathbb{S}_+^d$, $R(\mathbf{M})$: sum of the k smallest eigenvalues of \mathbf{M}

$$R(\mathbf{M}) = 0 \iff \text{rank}(\mathbf{M}) \leq d - k$$

- Reformulation

$$\min_{\mathbf{M} \in \mathbb{S}_+^d} \mu R(\mathbf{M}) + \ell(\mathbf{M}, \mathcal{N}) \implies \min_{\mathbf{M} \in \mathbb{S}_+^d} \mu \langle \mathbf{W}, \mathbf{M} \rangle + \ell(\mathbf{M}, \mathcal{N})$$

with \mathbf{W} rank- k projector on the eigenvectors of \mathbf{M} with k smallest eigenvalues

Metric Learning

Construction of \mathbf{W}

- $\mathbf{M} = \mathbf{V}_{\mathbf{M}} \text{Diag}(\lambda(\mathbf{M})) \mathbf{V}_{\mathbf{M}}^{\top}$ eigendecomposition of $\mathbf{M} \in \mathbb{S}_{+}^d$, $\mathbf{V}_{\mathbf{M}}$ orthogonal matrix
- We construct $\mathbf{w} = (w_1, \dots, w_d)^{\top} \in \mathbb{R}^d$:

$$w_i = \begin{cases} 0 & \text{if } 1 \leq i \leq d - k \text{ (the first } d - k \text{ elements)} \\ 1 & \text{if } d - k + 1 \leq i \leq d \text{ (the last } k \text{ elements)} \end{cases}$$

$$\mathbf{W} = \mathbf{V}_{\mathbf{M}} \text{Diag}(\mathbf{w}) \mathbf{V}_{\mathbf{M}}^{\top} \quad (1)$$

$$\min_{\mathbf{M} \in \mathbb{S}_{+}^d} \mu R(\mathbf{M}) + \ell(\mathbf{M}, \mathcal{N}) \implies \min_{\mathbf{M} \in \mathbb{S}_{+}^d} \mu \langle \mathbf{W}, \mathbf{M} \rangle + \ell(\mathbf{M}, \mathcal{N}) \text{ s.t. } \mathbf{W} = \mathbf{V}_{\mathbf{M}} \text{Diag}(\mathbf{w}) \mathbf{V}_{\mathbf{M}}^{\top}$$

Metric Learning

- Algorithm: alternating optimization procedure

Input: Training constraints \mathcal{N} , hyper-parameter μ and step size $\eta > 0$

Output: $\mathbf{M} \in \mathbb{S}_+^d$

Initialize $\mathbf{M}^1 \in \mathbb{S}_+^d$, iteration $n = 1$

Repeat until convergence

1. $\mathbf{W}^n \leftarrow \mathbf{V}_{\mathbf{M}^n} \text{Diag}(\mathbf{w}) \mathbf{V}_{\mathbf{M}^n}^\top$
2. Fix \mathbf{W}^n in Eq. (1)
3. $\mathbf{W}^n \in \partial(\langle \mathbf{W}^n, \mathbf{M}^n \rangle)$
4. $\mathbf{G}^n \in \partial\ell(\mathbf{M}^n, \mathcal{N})$
5. $\mathbf{M}^{n+1} \leftarrow \Pi_{\mathbb{S}_+^d} (\mathbf{M}^n - \eta(\mu \mathbf{W}^n + \mathbf{G}^n))$
6. $n \leftarrow n + 1$

Results on face verification pb

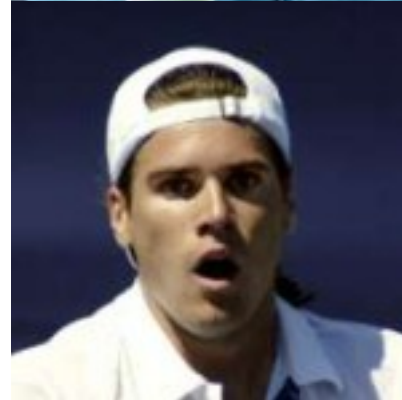
2 images => same face ?

Labeled Faces in the Wild (LFW)-- 27 SIFT descriptors concatenated
10-fold Cross Validation (600 pairs per fold)

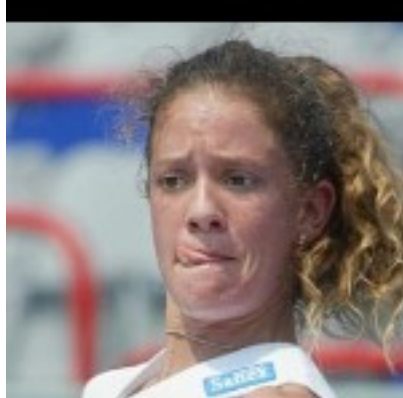


Method	Accuracy (in %)
ITML	76.2 ± 0.5
LDML	77.5 ± 0.5
PCCA	82.2 ± 0.4
Proposed method	83.5 ± 0.5

Bad results: Should be similar

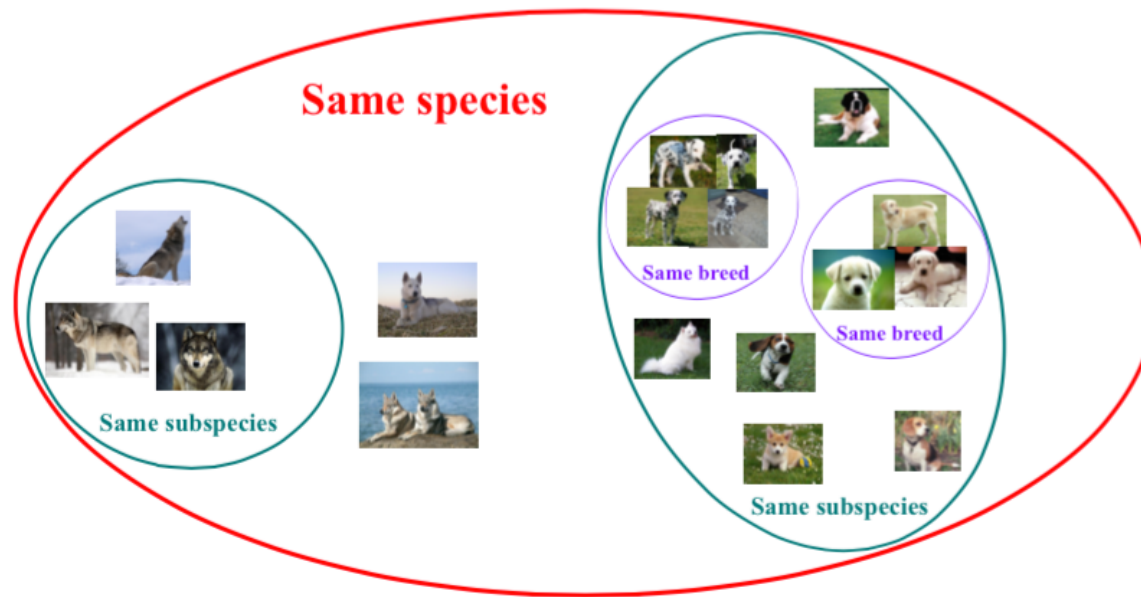


Bad results: Should be dissimilar



Hierarchical Image Classification

Rich relationships in taxonomies can be described with relative distances
Information richer than “is similar” or “is dissimilar”
Different levels of similarity



Learn dissimilarity D such that:

$$D(\text{img}_1, \text{img}_2) < D(\text{img}_3, \text{img}_4)$$

$$D(\text{img}_5, \text{img}_6) < D(\text{img}_7, \text{img}_8)$$

Taxonomy ML

- Qwise constraints sampling:
 1. Images in the same class more similar than images in sibling classes
 2. Images in sibling classes more similar than images in cousin classes
- $\mathbf{x}_i \in \mathbb{R}^d$: 1,000 dimensional SIFT BoW descriptor
- Diagonal PSD matrix framework: $\mathbf{w} \geq 0$
- **Convex Optimization Problem:**

$$\min_{\mathbf{w}} \mu \|\mathbf{w}\|_2^2 + \sum_{(p_i, p_j, p_k, p_l)} \ell(\mathbf{w}^\top [\Psi(p_k, p_l) - \Psi(p_i, p_j)])$$

with $\Psi(p_i, p_j) = (\mathbf{x}_i - \mathbf{x}_j) \circ (\mathbf{x}_i - \mathbf{x}_j)$ Hadamard product

Taxonomy ML

Subtree Dataset	[Verma 2012]	Qwise
Amphibian	41%	43.5%
Fish	39%	41%
Fruit	23.5%	21.1%
Furniture	46%	48.8%
Geological Formation	52.5%	56.1%
Musical Instrument	32.5%	32.9%
Reptile	22%	23.0%
Tool	29.5%	26.4%
Vehicle	27%	34.7%
Global Accuracy	34.8%	36.4%

Table 1: Standard classification accuracy for the various datasets.

- **9 datasets** from ImageNet, for each dataset: from 8 to 40 different classes, from 8,000 to 54,000 images for training

Outline

1. Introduction

2. Metric Learning

- Data and Metric models: Mahalanobis, ...
- Learning schemes:
 - ▶ Constraints :Pairs, triplets ...
 - ▶ Objective function: regularization, optimization ...
 - ▶ Examples

3. **Computer Vision Applications**

- Relative attribute learning
- Web page comparison

CV app: Scarlett and others

- Best Paper (Marr Prize) at ICCV 2011:

Relative attributes,

D. Parikh (TTI Chicago) and
K. Grauman (Texas Univ)

To appear, Proceedings of the International Conference on Computer Vision (ICCV), 2011.

Relative Attributes

Devi Parikh
Toyota Technological Institute Chicago (TTIC)
dparikh@ttic.edu

Kristen Grauman
University of Texas at Austin
grauman@cs.utexas.edu

Abstract

Human-nameable visual “attributes” can benefit various recognition tasks. However, existing techniques restrict these properties to categorical labels (for example, a person is ‘smiling’ or not, a scene is ‘dry’ or not), and thus fail to capture more general semantic relationships. We propose to model relative attributes. Given training data stating how object/scene categories relate according to different attributes, we learn a ranking function per attribute. The learned ranking functions predict the relative strength of each property in novel images. We then build a generative model over the joint space of attribute ranking outputs, and propose a novel form of zero-shot learning in which the supervisor relates the unseen object category to previously seen objects via attributes (for example, ‘bears are further than giraffes’). We further show how the proposed relative attributes enable richer textual descriptions for new images, which in practice are more precise for human interpretation. We demonstrate the approach on datasets of faces and natural scenes, and show its clear advantages over traditional binary attribute prediction for these new tasks.

1. Introduction

While traditional visual recognition approaches map low-level image features directly to object category labels, recent work proposes models using *visual attributes* [1–8]. Attributes are properties observable in images that have human-designated names (e.g., ‘striped’, ‘four-legged’), and they are valuable as a new semantic cue in various problems. For example, researchers have shown their impact for strengthening facial verification [5], object recognition [6, 8, 16], generating descriptions of unfamiliar objects [1], and to facilitate “zero-shot” transfer learning [2], where one trains a classifier for an unseen object simply by specifying which attributes it has.

Problem: Most existing work focuses wholly on attributes as binary predicates indicating the presence (or absence) of a certain property in an image [1–8, 16]. This may suffice for part-based attributes (e.g., ‘has a head’) and some

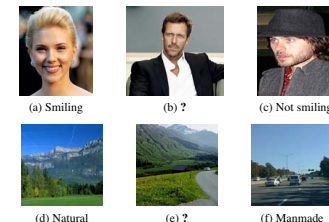


Figure 1. Binary attributes are an artificially restrictive way to describe images. While it is clear that (a) is smiling, and (c) is not, the more informative and intuitive description for (b) is via *relative* attributes: he is smiling more than (a) but less than (c). Similarly, scene (e) is less natural than (d), but more so than (f). Our main idea is to model relative attributes via learned ranking functions, and then demonstrate their impact on novel forms of zero-shot learning and generating image descriptions.

binary properties (e.g., ‘spotted’). However, for a large variety of attributes, not only is this binary setting restrictive, but it is also unnatural. For instance, it is not clear if in Figure 1(b) Hugh Laurie is smiling or not; different people are likely to respond inconsistently in providing the presence or absence of the ‘smiling’ attribute for this image, or of the ‘natural’ attribute for Figure 1(e).

Indeed, we observe that *relative* visual properties are a semantically rich way by which humans describe and compare objects in the world. They are necessary, for instance, to refine an identifying description (“the ‘rounder’ pillow”, “the same except ‘bluer’”), or to situate with respect to reference objects (“‘brighter’ than a candle; ‘dimmer’ than a flashlight”). Furthermore, they have potential to enhance active and interactive learning—for instance, offering a better guide for a visual search (“find me similar shoes, but ‘shinier.’” or “refine the retrieved images of downtown Chicago to those taken on ‘sunnier’ days”).

Proposal: In this work, we propose to model *relative attributes*. As opposed to predicting the presence of an attribute, a relative attribute indicates the strength of an attribute in an image with respect to other images. For exam-

CV app: Attribute Models

$x_i \rightarrow$ Real value



Density,
Smiling,

....

“I am 60% sure this person is smiling”
(Binary Classifier Confidence)

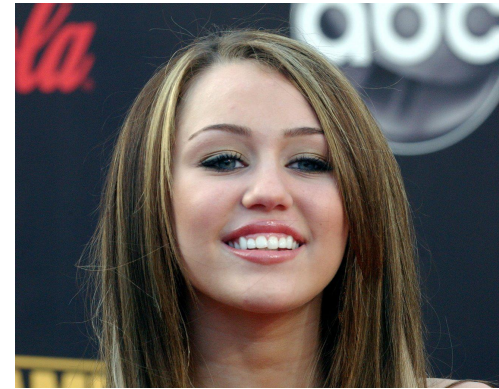
“This person is smiling 60%”
(Attribute Strength)

CV app: Relative Attributes

“Person A is smiling more than Person B”
(Relative Attribute, Parikh and Grauman ICCV 2011)



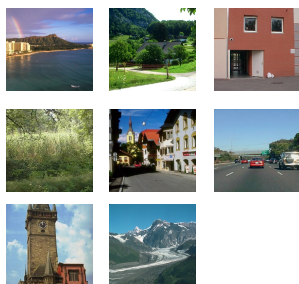
<
smiling



>
natural



- Training sets:
Attributes labeled
at category level



	Binary	Relative
OSR	TI SHC OMF	
natural	0 0 0 1 1 1 1	$T \prec I \sim S \prec H \prec C \sim O \sim M \sim F$
open	0 0 0 1 1 1 0	$T \sim F \prec I \sim S \prec M \prec H \sim C \sim O$
perspective	1 1 1 1 0 0 0	$O \prec C \prec M \sim F \prec H \prec I \prec S \prec T$
large-objects	1 1 1 0 0 0 0	$F \prec O \sim M \prec I \sim S \prec H \sim C \prec T$
diagonal-plane	1 1 1 1 0 0 0	$F \prec O \sim M \prec C \prec I \sim S \prec H \prec T$
close-depth	1 1 1 1 0 0 1	$C \prec M \prec O \prec T \sim I \sim S \sim H \sim F$
PubFig	ACHJ MSVZ	
Masculine-looking	1 1 1 1 0 0 1 1	$S \prec M \prec Z \prec V \prec J \prec A \prec H \prec C$
White	0 1 1 1 1 1 1 1	$A \prec C \prec H \prec Z \prec J \prec S \prec M \prec V$
Young	0 0 0 0 1 1 0 1	$V \prec H \prec C \prec J \prec A \prec S \prec Z \prec M$
Smiling	1 1 1 0 1 1 0 1	$J \prec V \prec H \prec A \sim C \prec S \sim Z \prec M$
Chubby	1 0 0 0 0 0 0 0	$V \prec J \prec H \prec C \prec Z \prec M \prec S \prec A$
Visible-forehead	1 1 1 0 1 1 1 0	$J \prec Z \prec M \prec S \prec A \sim C \sim H \sim V$
Bushy-eyebrows	0 1 0 1 0 0 0 0	$M \prec S \prec Z \prec V \prec H \prec A \prec C \prec J$
Narrow-eyes	0 1 1 0 0 0 1 1	$M \prec J \prec S \prec A \prec H \prec C \prec V \prec Z$
Pointy-nose	0 0 1 0 0 0 0 1	$A \prec C \prec J \sim M \sim V \prec S \prec Z \prec H$
Big-lips	1 0 0 0 1 1 0 0	$H \prec J \prec V \prec Z \prec C \prec M \prec A \prec S$
Round-face	1 0 0 0 1 1 0 0	$H \prec V \prec J \prec C \prec Z \prec A \prec S \prec M$

Table 1. Binary and relative attribute assignments used in our experiments. Note that none of the relative orderings violate the binary memberships. The OSR dataset includes images from the following categories: coast (C), forest (F), highway (H), inside-city (I), mountain (M), open-country (O), street (S) and tall-building (T). The 8 attributes shown above are listed in [11] as the properties subjects used to organize the images. The PubFig dataset includes images of: Alex Rodriguez (A), Clive Owen (C), Hugh Laurie (H), Jared Leto (J), Miley Cyrus (M), Scarlett Johansson (S), Viggo Mortensen (V) and Zac Efron (Z). The 11 attributes shown above are a

CV app: Attribute Models

- Ranking functions for relative attributes
For each attribute a_m , **open**

Supervision = all pairs as:

	Binary	Relative
OSR	TI SHC OMF	
natural	0 0 0 1 1 1 1	T < I ~ S < H < C ~ O ~ M ~ F
open	0 0 0 1 1 1 0	T ~ F < I ~ S < M < H ~ C ~ O
perspective	1 1 1 1 0 0 0	O < C < M ~ F < H < I < S < T
large-objects	1 1 1 0 0 0 0	F < O ~ M < I ~ S < H ~ C < T
diagonal-plane	1 1 1 1 0 0 0	F < O ~ M < C < I ~ S < H < T
close-depth	1 1 1 1 0 0 1	C < M < O < T ~ I ~ S ~ H ~ F
PubFig	ACHJ MS VZ	
Masculine-looking	1 1 1 1 0 0 1 1	S < M < Z < V < J < A < H < C
White	0 1 1 1 1 1 1 1	A < C < H < Z < J < S < M < V
Young	0 0 0 1 1 0 1	V < H < C < J < A < S < Z < M
Smiling	1 1 1 0 1 1 0 1	J < V < H < A < C < S ~ Z < M
Chubby	1 0 0 0 0 0 0 0	V < J < H < C < Z < M < S < A
Visible-forehead	1 1 1 0 1 1 1 0	J < Z < M < S < A < C ~ H ~ V
Bushy-eyebrows	0 1 0 1 0 0 0 0	M < S < Z < V < H < A < C < J
Narrow-eyes	0 1 1 0 0 0 1 1	M < J < S < A < H < C < V < Z
Pointy-nose	0 0 1 0 0 0 0 1	A < C < J ~ M ~ V < S < Z < H
Big-lips	1 0 0 0 1 1 0 0	H < J < V < Z < C < M < A < S
Round-face	1 0 0 0 1 1 0 0	H < V < J < C < Z < A < S < M

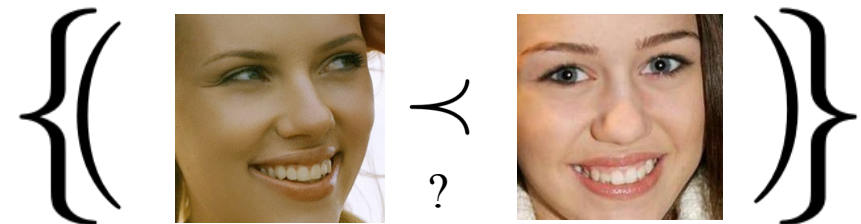
$$O_m: \left\{ \left(\text{img}_1 \succ \text{img}_2 \right), \dots \right\},$$

$$S_m: \left\{ \left\{ \text{img}_1 \sim \text{img}_2 \right\}, \dots \right\}$$

CV app: pairwise ranking

Scarlett Johansson vs Miley Cyrus

- Coarse labeling at category level => noisy pair sampling

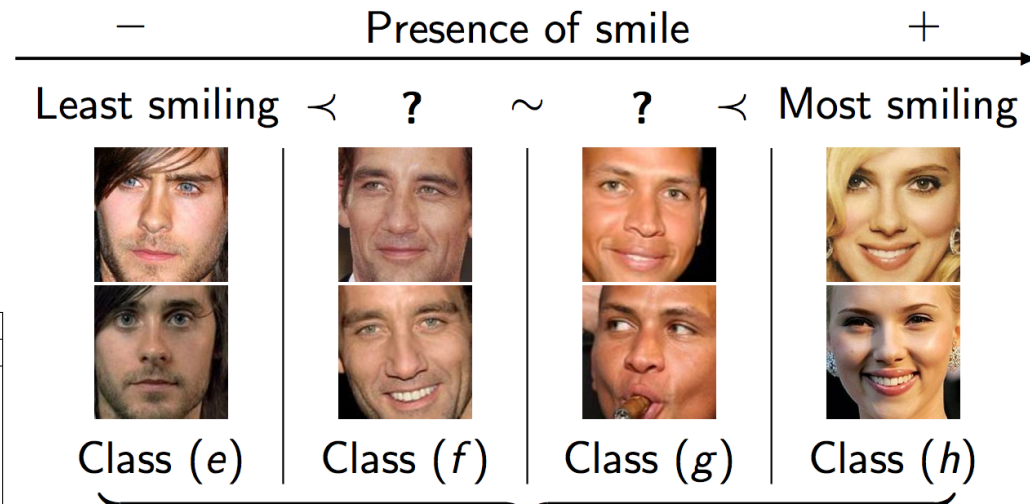


	Binary	Relative
OSR	TI SHC OMF	
natural	00001111	T<I~S<H<C~O~M~F
open	00011110	T~F<I~S<M<H~C~O
perspective	11110000	O<C<M~F<H<I~S<T
large-objects	11100000	F<O~M<I~S<H~C<T
diagonal-plane	11110000	F<O~M<C<I~S<H<T
close-depth	11110001	C<M<O<T~I~S~H~F
PubFig	ACHJ MSVZ	
Masculine-looking	11110011	S<M<Z<V<J<A<H<C
White	01111111	A<C<H<Z<J<S<M<V
Young	00001101	V~H~C~J~A~S~Z~M
Smiling	11101101	J<V<H<A~C<S~Z~M
Chubby	10000000	V~J~H~C<Z~M~S~A
Visible-forehead	11101110	J<Z<M<S<A~C~H~V
Bushy-eyebrows	01010000	M<S<Z<V<H<A<C<J
Narrow-eyes	01100011	M<J<S<A<H~C~V~Z
Pointy-nose	00100001	A<C<J~M~V~S<Z~H
Big-lips	10001100	H<J~V~Z<C~M~A~S
Round-face	10001100	H~V~J~C~Z~A~S~M

- Quadruplet to minimize this artefact

CV app: Quadruplet-wise ML

	Binary	Relative
OSR	TI SHC OM F	
natural	0 0 0 0 1 1 1 1	T<I~S<H<C~O~M~F
open	0 0 0 1 1 1 1 0	T~F<I~S<M<H~C~O
perspective	1 1 1 1 0 0 0 0	O<C<M~F<H<I<S<T
large-objects	1 1 1 0 0 0 0 0	F<O~M<I~S<H~C<T
diagonal-plane	1 1 1 1 0 0 0 0	F<O~M<C<I~S<H<T
close-depth	1 1 1 1 0 0 0 1	C<M<O<T~I~S~H~F
PubFig	ACHJ MS V Z	
Masculine-looking	1 1 1 1 0 0 1 1	S<M<Z<V<J<A<H<C
White	0 1 1 1 1 1 1 1	A<C<H<Z<J<S<M<V
Young	0 0 0 0 1 1 0 1	V<H<C<J<A<S<Z<M
Smiling	1 1 1 0 1 1 0 1	J<V<H<A~C<S~Z<M
Chubby	1 0 0 0 0 0 0 0	V<J<H<C<Z<M<S<A
Visible-forehead	1 1 1 0 1 1 1 0	J<Z<M<S<A~C~H~V
Bushy-eyebrows	0 1 0 1 0 0 0 0	M<S<Z<V<H<A<C<J
Narrow-eyes	0 1 1 0 0 0 1 1	M<J<S<A<H<C<V<Z
Pointy-nose	0 0 1 0 0 0 0 1	A<C<J~M~V<S<Z<H
Big-lips	1 0 0 0 1 1 0 0	H<J<V<Z<C<M<A<S
Round-face	1 0 0 0 1 1 0 0	H<V<J<C<Z<A<S<M



Learn dissimilarity D such that:

$$D(\text{Class (f) image 1}, \text{Class (g) image 1}) < D(\text{Class (h) image 1}, \text{Class (e) image 1})$$

$$D(\text{Class (f) image 2}, \text{Class (g) image 2}) < D(\text{Class (h) image 2}, \text{Class (e) image 2})$$

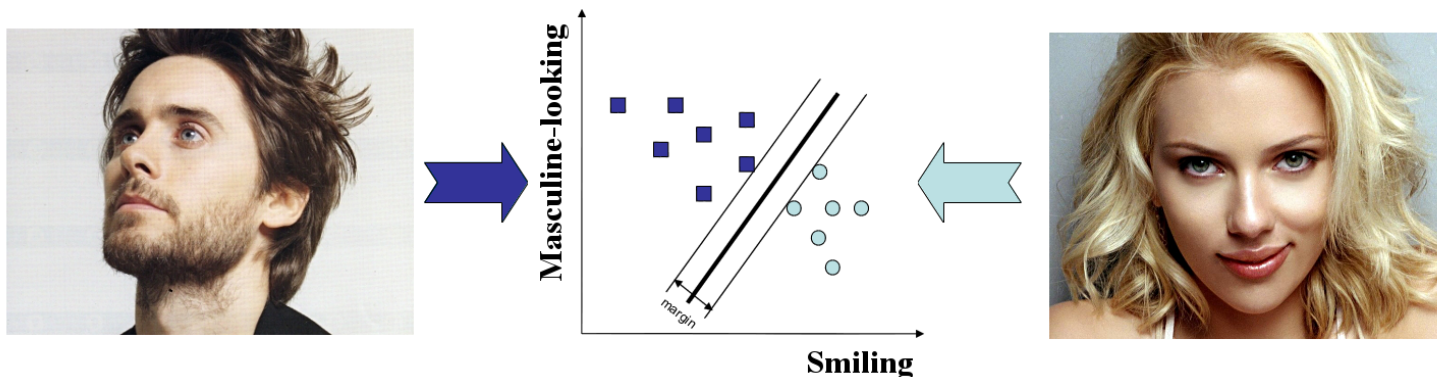
- Relative attributes => (Dis)similarity Learning under Qwise constraints

Relative attribute learning

- Learning a feature space

$$\begin{aligned} D_{\mathbf{M}}^2(p_i, p_j) &= \Phi(p_i, p_j)^\top \mathbf{M} \Phi(p_i, p_j) \\ &= (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{L}^\top \mathbf{L} (\mathbf{x}_i - \mathbf{x}_j) \end{aligned}$$

- Corresponds to learn a linear transformation parameterized by $\mathbf{L} \in \mathbb{R}^{M \times d}$ such that $\mathbf{h}_i = \mathbf{L} \mathbf{x}_i$ where the m -th row of \mathbf{L} is \mathbf{w}_m^\top
- Application to Actor retrieval and classification:



Relative attribute learning

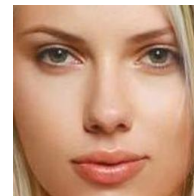
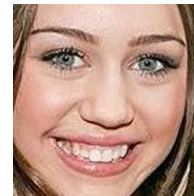
$$\min_{\mathbf{w}} \mu \|\mathbf{w}\|_2^2 + \sum_{\substack{(p_i, p_j, p_k, p_l) \\ D(\text{img}_i, \text{img}_j) < D(\text{img}_k, \text{img}_l) \\ D(\text{img}_i, \text{img}_j) < D(\text{img}_k, \text{img}_l)}} \ell(\mathbf{w}^\top [\Psi(p_k, p_l) - \Psi(p_i, p_j)])$$

- $\mathbf{x}_i \in \mathbb{R}^d$: GIST (+ color) descriptor
- $\Psi(p_i, p_j) = \mathbf{x}_i - \mathbf{x}_j$
- Relative attributes a_m for $m \in \{1, \dots, M\}$: smiling, masculine-looking young...
- Learning a \mathbf{w}_m for each attribute a_m using Qwise optimization
- Resulting in learning a linear transformation parameterized by $\mathbf{L} \in \mathbb{R}^{M \times d}$

$$\mathbf{L} = \begin{bmatrix} w_{1,1} & \dots & w_{1,d} \\ \vdots & \vdots & \vdots \\ w_{M,1} & \dots & w_{M,d} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1^\top \\ \vdots \\ \mathbf{w}_M^\top \end{bmatrix}, \quad \mathbf{w}_m^\top : m\text{-th row}$$

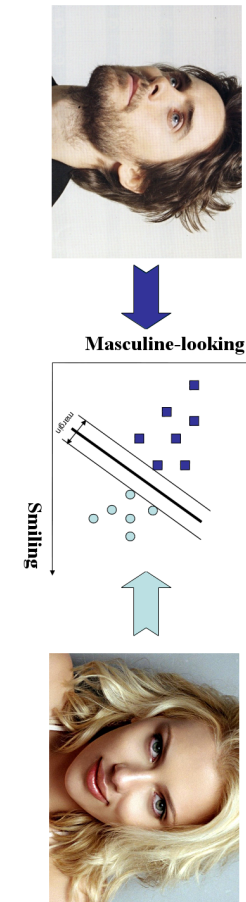
Relative attribute experiments

- Outdoor Scene Recognition
OSR [Oliva 01]
- 8 classes, ~2700 images, GIST
- 6 attributes: open, natural ...
- Public Figures Faces PubFig
[Kumar 09]
- 8 classes, ~800 images, GIST
+color
- 11 attributes: smiling, shabby ...



Relative attribute experiments

- Baselines
 - RA Relative attribute method (Parikh and Grauman)
 - annotations on class relationships with pairwise constraints
 - LMNN Linear transformation learned [Wein.09]
 - class membership information used only unlike RA
 - RA + LMNN: Combination of the first two baselines
 1. Relative attribute annotations to learn attribute space
 2. Metric in attribute space with LMNN
- Qwise Method:
 - Qwise constraints generated as pairwise
 - Qwise output alone or combined Qwise + LMNN



[Wein.09] K.Q. Weinberger, and L.K. Saul, Distance metric learning for large margin nearest neighbor classification, In JMLR 2009

Relative attribute experiments

	OSR	Pubfig
Parikh's code	$71.3 \pm 1.9\%$	$71.3 \pm 2.0\%$
LMNN-G	$70.7 \pm 1.9\%$	$69.9 \pm 2.0\%$
LMNN	$71.2 \pm 2.0\%$	$71.5 \pm 1.6\%$
RA + LMNN	$71.8 \pm 1.7\%$	$74.2 \pm 1.9\%$
Qwise	$74.1 \pm 2.1\%$	$74.5 \pm 1.3\%$
Qwise + LMNN-G	$74.6 \pm 1.7\%$	$76.5 \pm 1.2\%$
Qwise + LMNN	$74.3 \pm 1.9\%$	$77.6 \pm 2.0\%$

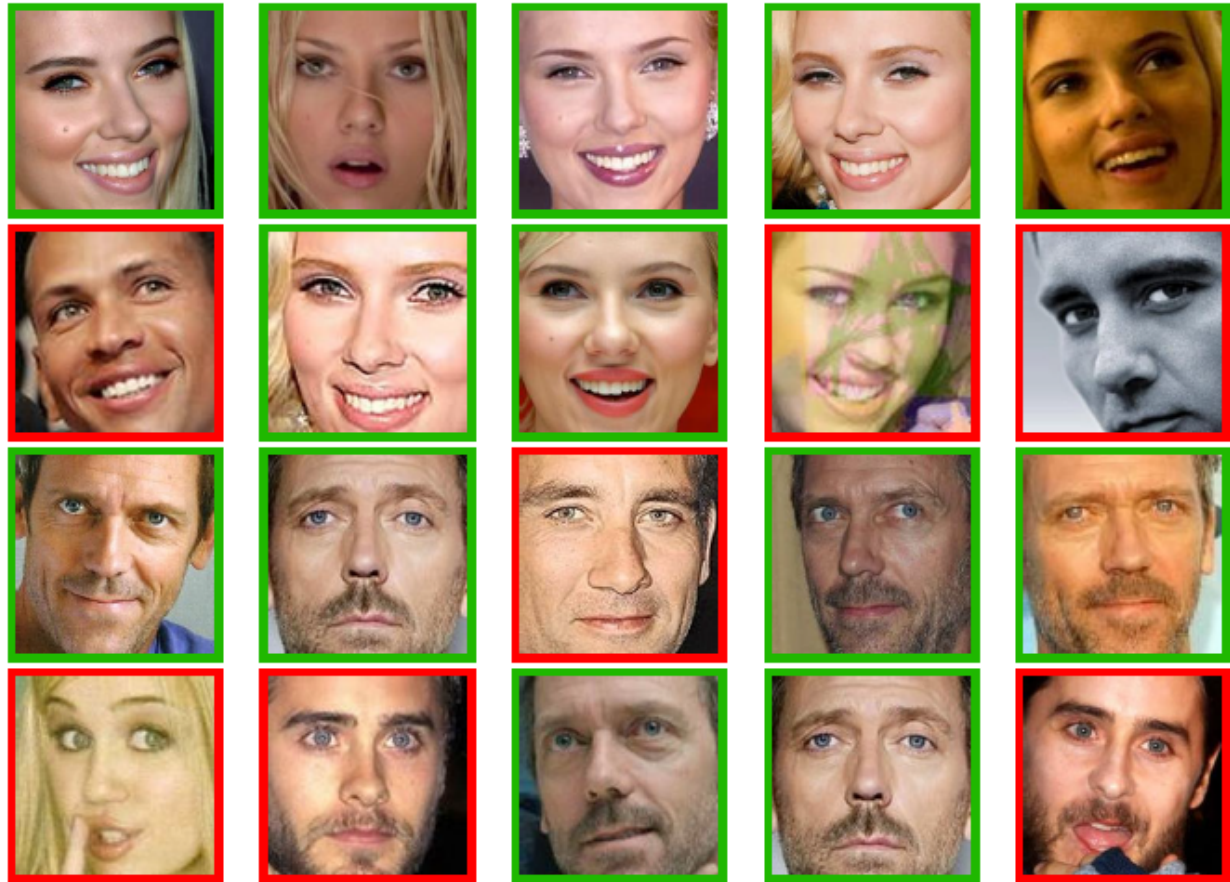
Table 1: Test classification accuracies on the OSR and Pubfig datasets for different methods.

Relative attribute experiments

Query

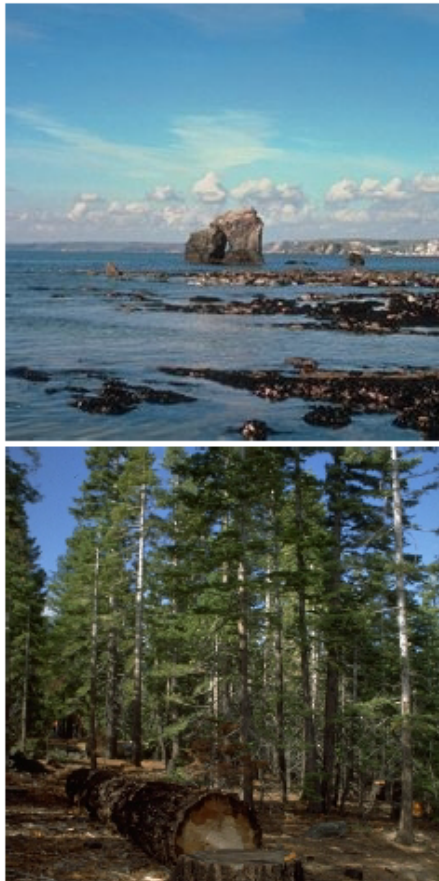


Top 5



Relative attribute experiments

Query



Top 5



Outline

1. Introduction

2. Metric Learning

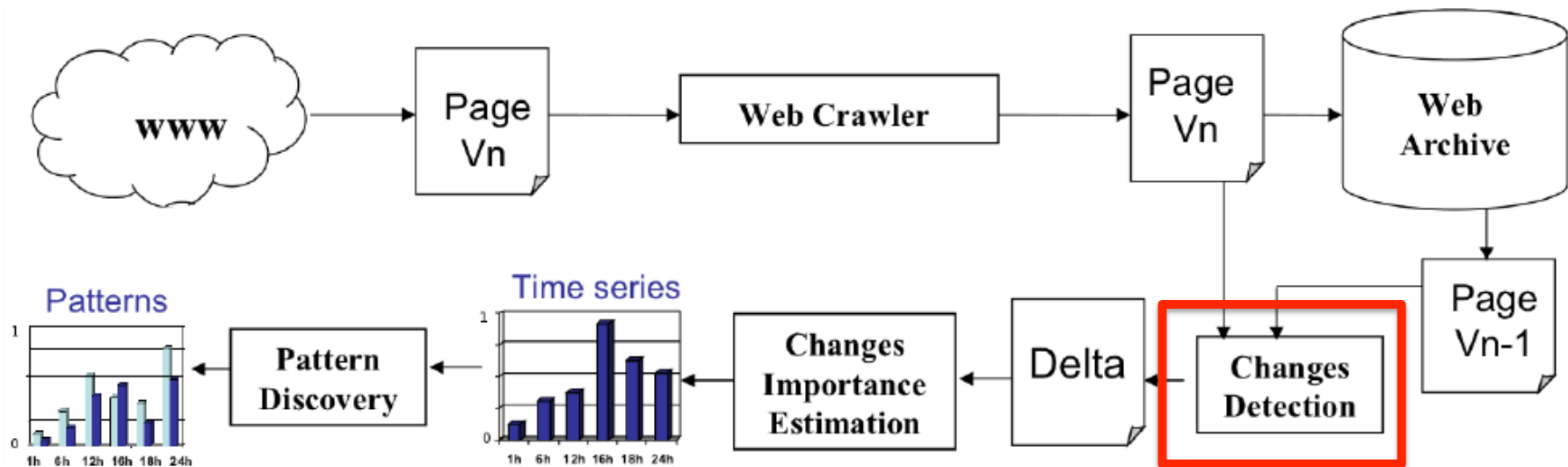
- Data and Metric models: Mahalanobis, ...
- Learning schemes:
 - ▶ Constraints :Pairs, triplets ...
 - ▶ Objective function: regularization, optimization ...
 - ▶ Examples

3. Computer Vision Applications

- Relative attribute learning
- **Web page comparison**

Web page ML

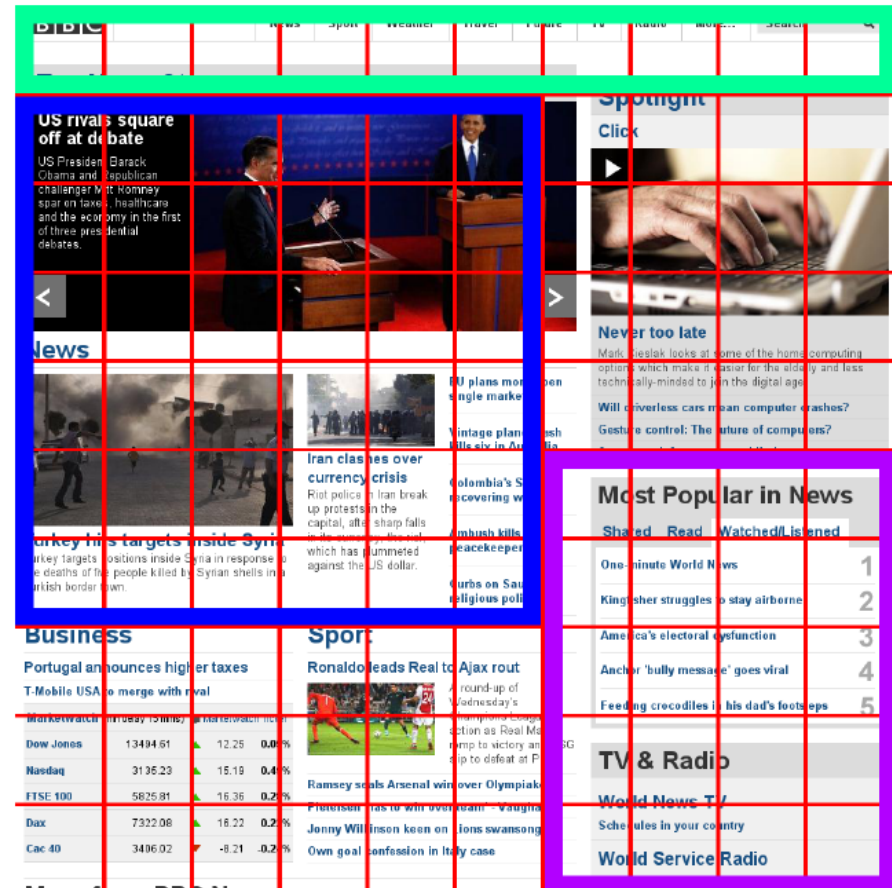
- Context:
 - For Web crawling purpose, useful to understand the change behavior of websites over time [AWUPCP11]



- Significant changes between successive versions of a same webpage => revisit the page
- Web page comparison
 - Qwise to learn Web page metric and significant webpage regions

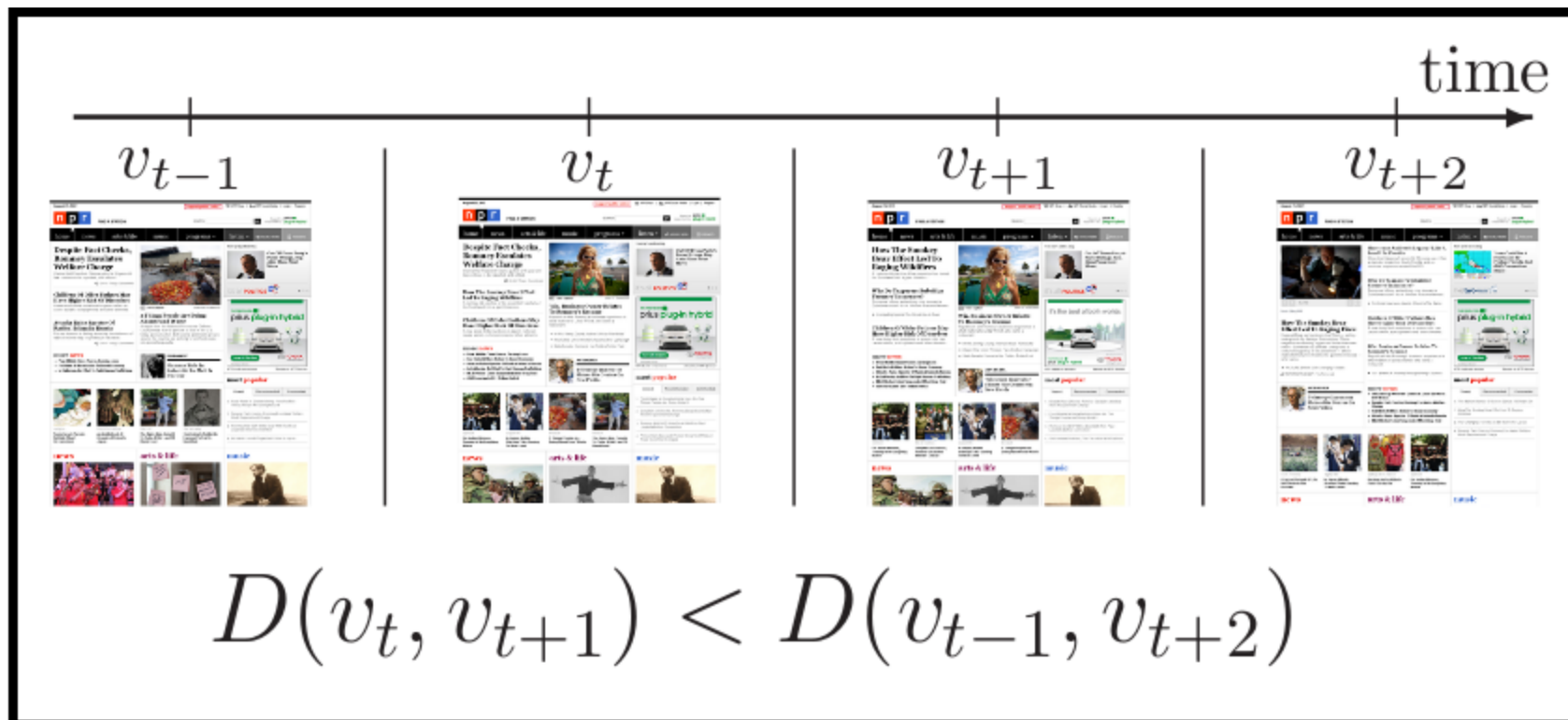
Web page ML

- Focus on news websites
 - Advertisements or menus not significant
 - News content significant
- Find a metric able to properly identify **significant** changes between webpage versions
- Localize changes inside pages [Song04]:
 - semantic spatial structure
 - significant to capture



Web page ML

- Qwise Constraints:
 - Fully unsupervised ML, but temporal information available
 - Constraints by comparing screenshots of successive webpage versions



Web page ML

- Descriptors: GIST on m-by-m grid over screenshots
- Ψ is a m-by-m vector of Euclidean distance between blocks
- Diagonal PSD matrix: w represents block weights
- Optimization over w
 - ▶ Learning of spatial weights of webpage regions using temporal relationships
 - ▶ Automatically
 - » Discovering important change regions
 - » Ignoring menus and advertisements



Web page ML

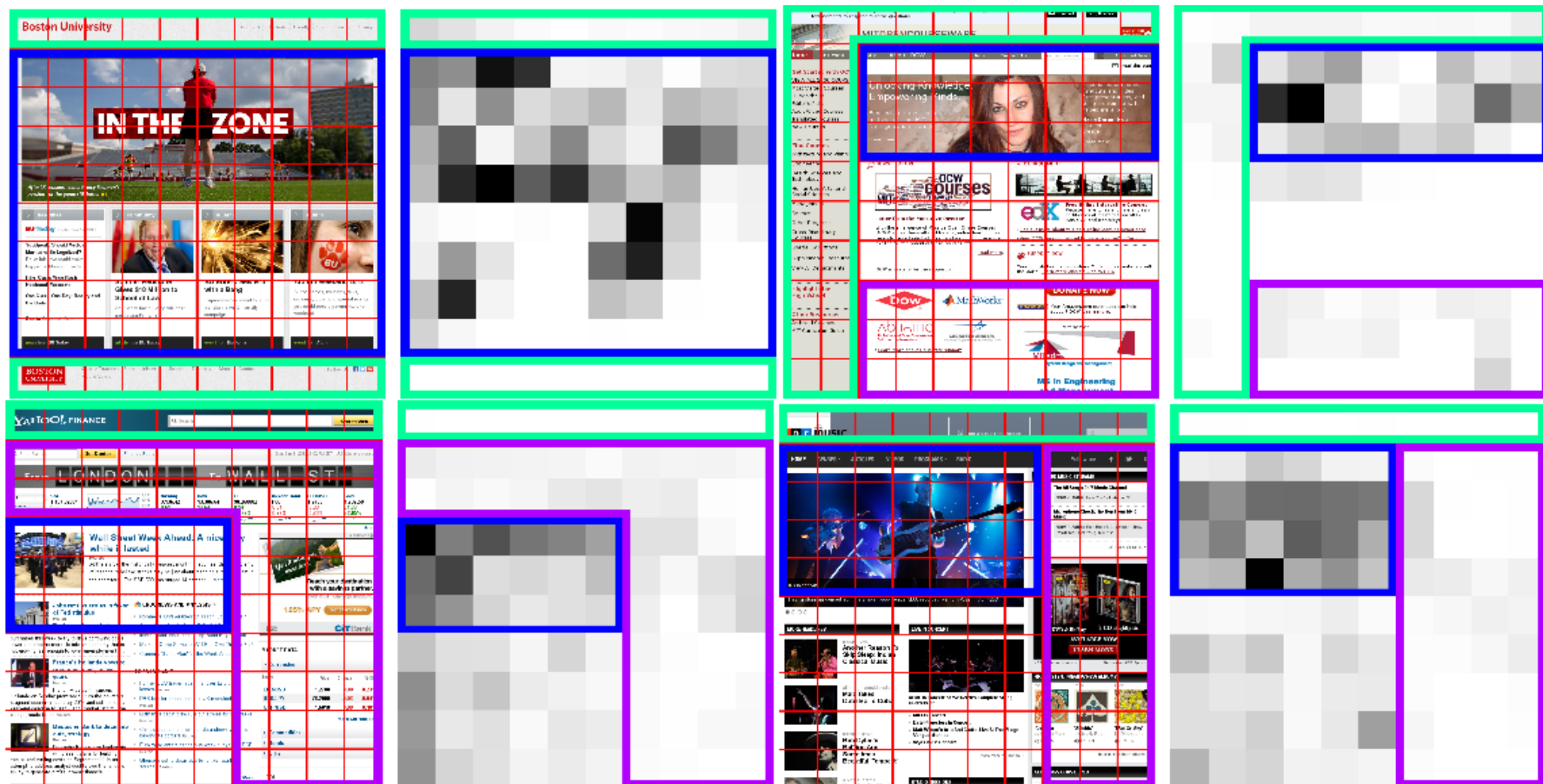
- Evaluation and Comparison
 - Crawling 50 days Several sites CNN, NPR, BBC, ...
 - Manual change detection (news updates) for GT on 5 days
 - Baselines: Euclidean Dist, LMNN
 - GIST on 10x10
 - Mean Average Precision on succ. Web page Metric scores

Site	CNN			NPR			New York Times			BBC		
Eval.	AP_S	AP_D	MAP	AP_S	AP_D	MAP	AP_S	AP_D	MAP	AP_S	AP_D	MAP
Eucl.	68.1	85.9	77.0	96.3	89.5	92.9	69.8	79.5	74.6	91.1	76.7	83.9
Dist.	± 0.6	± 0.6	± 0.5	± 0.2	± 0.5	± 0.3	± 0.9	± 0.4	± 0.5	± 0.3	± 0.6	± 0.4
LMNN	78.8	91.7	85.2	98.0	92.5	95.2	83.2	89.1	86.1	92.5	80.1	86.3
	± 1.9	± 1.7	± 1.8	± 0.6	± 1.1	± 0.9	± 1.4	± 2.7	± 2.0	± 0.4	± 1.0	± 0.6
Qwise	82.7	94.6	88.6	98.6	94.3	96.5	85.5	92.3	88.9	92.8	79.3	86.1
	± 4.1	± 1.8	± 2.9	± 0.2	± 0.6	± 0.4	± 5.4	± 4.1	± 4.6	± 0.4	± 1.3	± 0.8

Web page ML



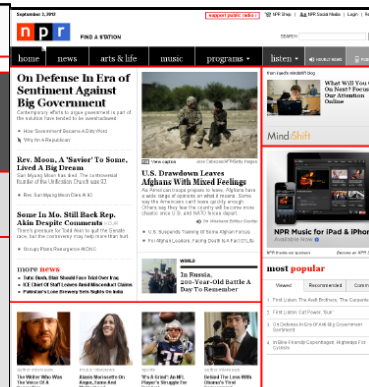
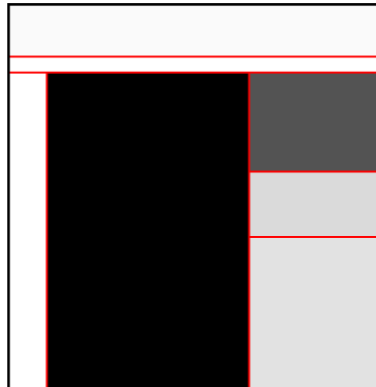
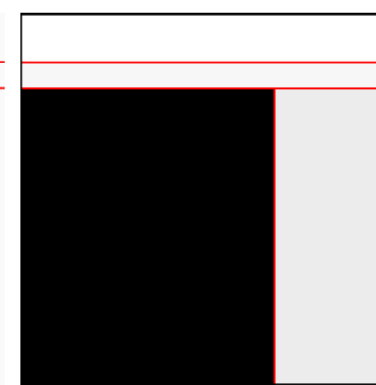
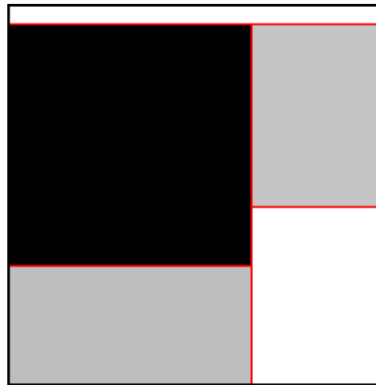
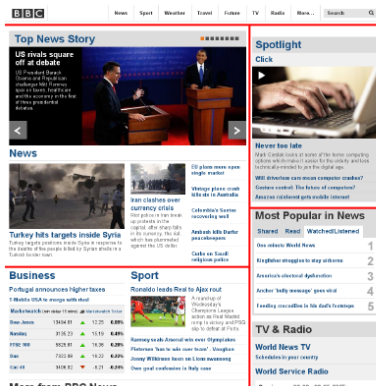
Web page ML



- Not connected to the structural layout of the Web page

Web page ML

- Detect significant changes using the source code of pages (Segmentation) + Qwise



Conclusion

- Key issues in Metric Learning:
 - Modeling: Data representation, form of the metric (linear, non lin., local)
 - Learning Paradigm: unsupervised, semi-supervised, transfer, type of constraints
 - ▶ Temporal/spatial relationships, class relationships => rich context to learn metrics or semantic embedding
 - Optimization issues: Global/local solution, Convexity, Scalability, dim. Reduction

Thx

Matthieu Cord

Joint work with Marc T. Law and Nicolas Thome

LIP6, Computer Science Department UPMC Paris 6 - Sorbonne University

<http://webia.lip6.fr/~cord>

Metric learning:

- M.T. Law, N. Thome and M. Cord. Fantope Regularization in Metric Learning, CVPR 2014
- M.T. Law, N. Thome and M. Cord. Quadruplet-wise Image Similarity Learning, ICCV 2013
- M.T. Law, N. Thome, S. Gancarski and M. Cord. Structural and Visual Comparisons for Web Page Archiving, ACM DocEng, 2012

M.T. Law PhD doc available: Distance Metric Learning for Image and Webpage Comparison, (defense Jan 2015 with F. Bach, P. Gallinari, P. Perez, J. Ponce, F. Precioso, A. Rakotomamonjy) Code available on demand

Image representation:

- S. Avila, N. Thome, M. Cord, E. Valle, A. araujo, Pooling in Image Representation: the Visual Codeword Point of View, CVIU 2012
- H. Goh, , N. Thome, M. Cord, JH. Lim, Top-Down Regularization of Deep Belief Networks, NIPS 2013