Quadruplet-wise constraints for visual metric learning

Journée scientifique LIMA2

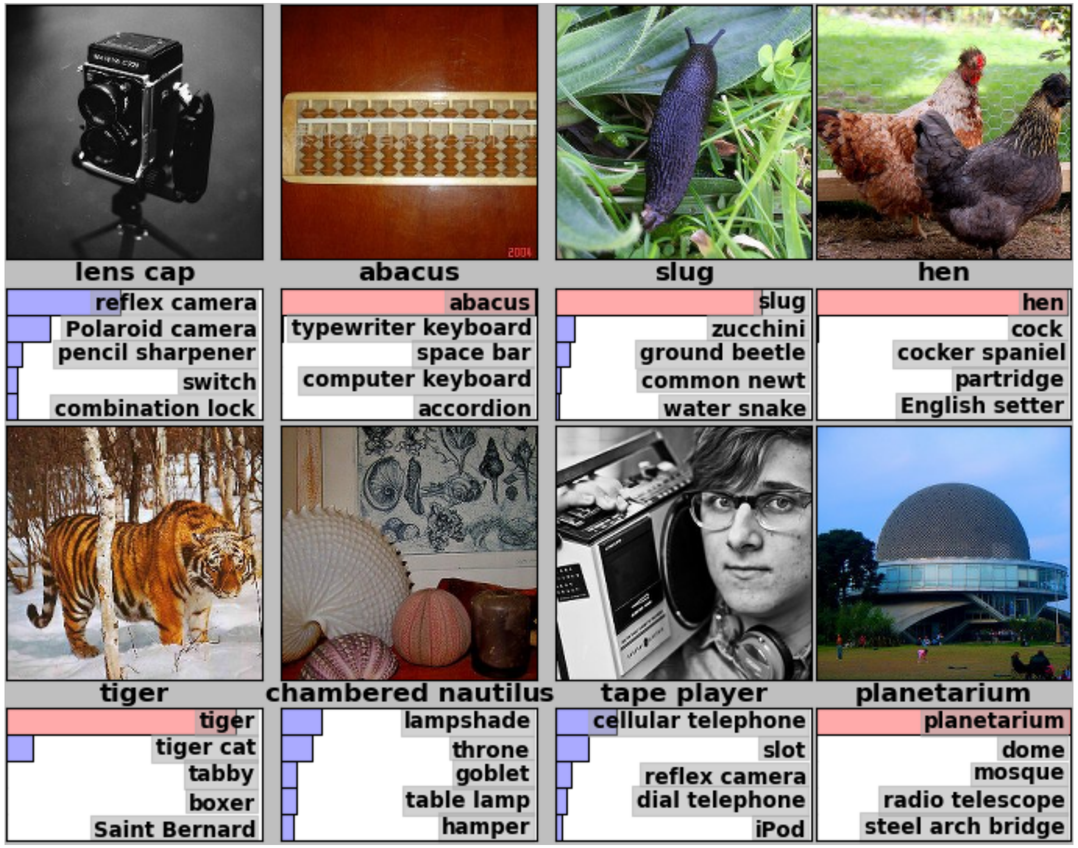
Région Rhône-Alpes, pôle **Imageinove**

Matthieu Cord

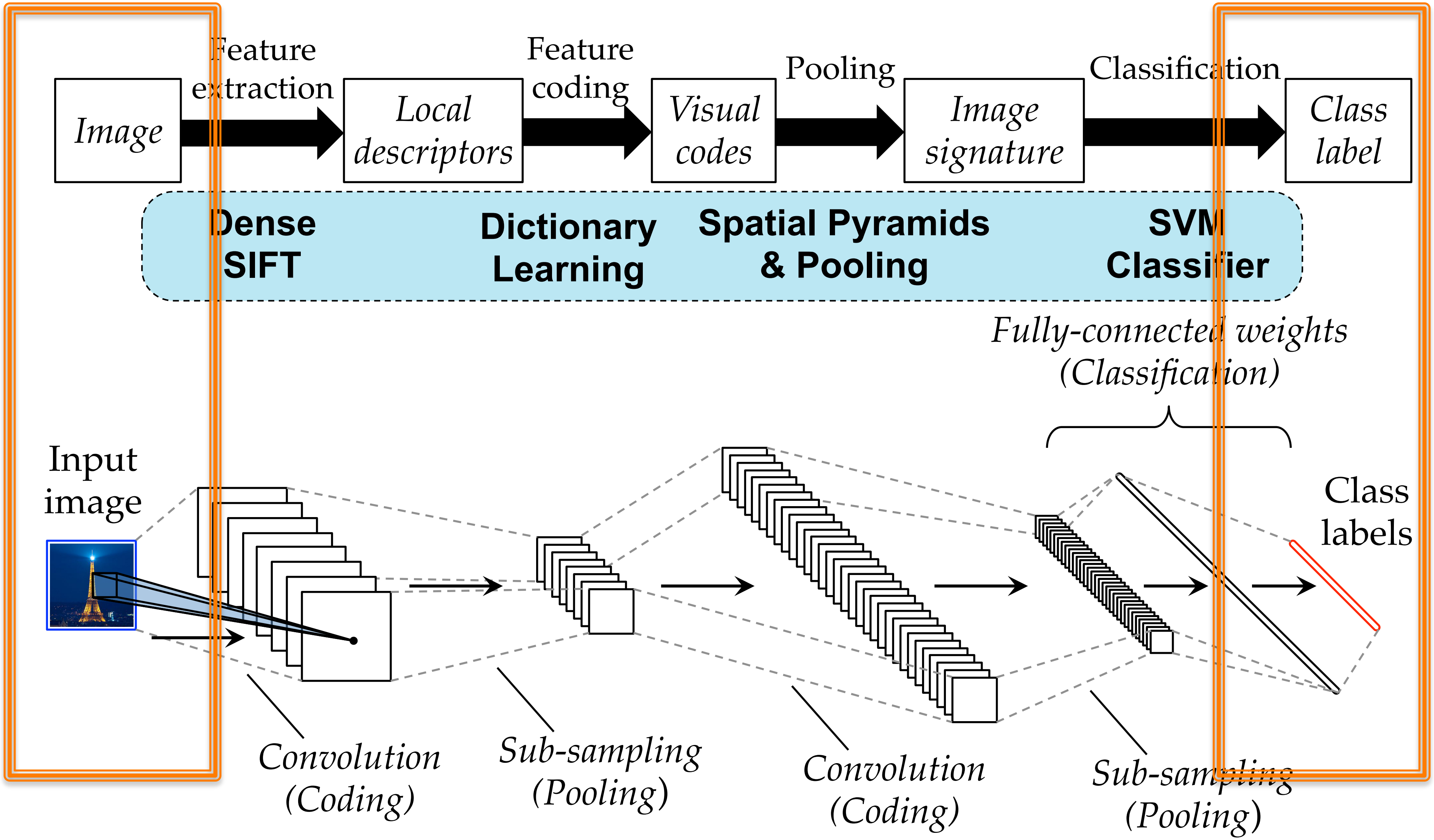
April 15, 2014

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

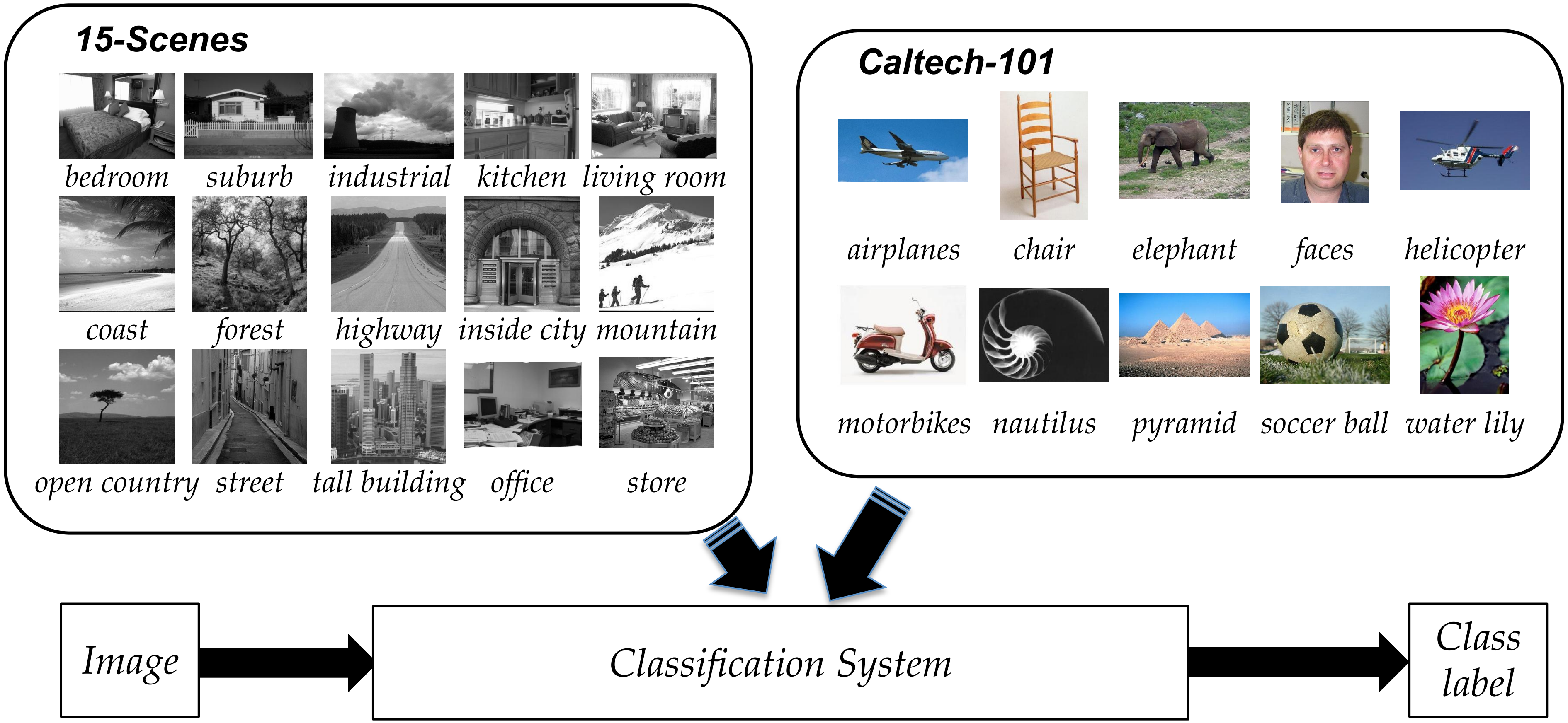


Introduction: Visual learning



Introduction: Visual learning

- Data for training



Introduction: Scarlett and others

- Joint work at ICCV 2013:

  Quadruplet-wise Image Similarity Learning,

  M.T. Law, N. Thome, M. Cord

- Inspired by: 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 furrier 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-

What are attributes?

- Mid-level concepts
  - Higher than low-level features
  - Lower than high-level categories
- Shared across categories
- Human-understandable (semantic)
- Machine-detectable (visual)

otter

black:	yes
white:	no
brown:	yes
stripes:	no
water:	yes
eats fish:	yes

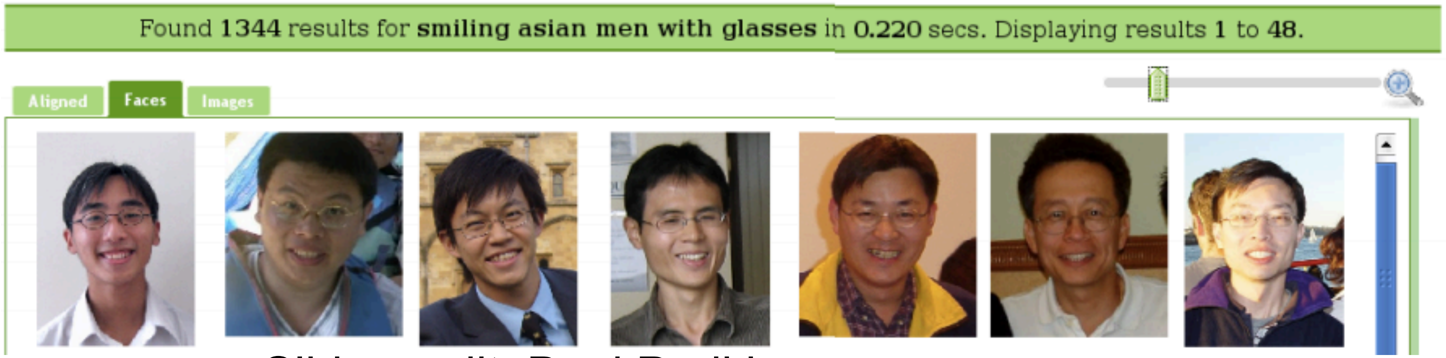
polar bear

black:	no
white:	yes
brown:	no
stripes:	no
water:	yes
eats fish:	yes

zebra

black:	yes
white:	yes
brown:	no
stripes:	yes
water:	no
eats fish:	no

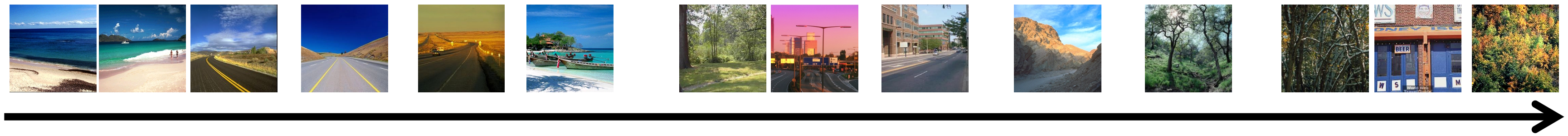
Face Tracer Image Search (Kumar 08) “Smiling Asian Men With Glasses”



Slide credit: Devi Parikh

Introduction: Attribute Models

$\boldsymbol{x}_i \rightarrow$ Real value



Density,
Smiling,
….

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

"This person is smiling 60%"
(Attribute Strength)

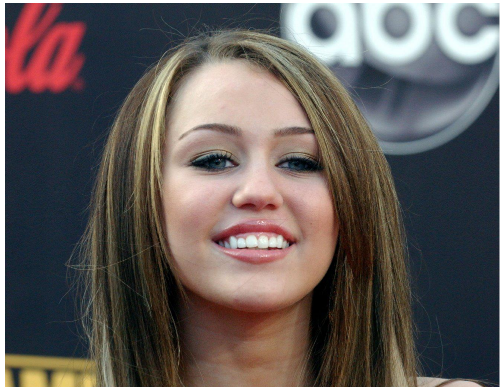
Slide credit: Devi Parikh

Introduction: 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	T I S H C O M F	
natural	0 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 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 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 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 0	$F \prec O \sim M \prec C \prec I \sim S \prec H \prec T$
close-depth	1 1 1 1 0 0 0 1	$C \prec M \prec O \prec T \sim I \sim S \sim H \sim F$
PubFig	A C H J M S V Z	
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

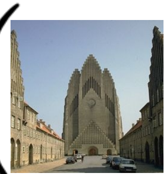
Introduction: Attribute Models

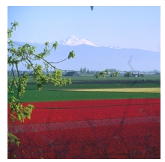
- Ranking functions for relative attributes

For each attribute a_m, open

Supervision = all pairs as:

	Binary	Relative
OSR	TI S HC OMF	
natural	0 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 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 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 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 0	$F \prec O \sim M \prec C \prec I \sim S \prec H \prec T$
close-depth	1 1 1 1 0 0 0 1	$C \prec M \prec O \prec T \sim I \sim S \sim H \sim F$
PubFig	ACHJ MS V Z	
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	$J \prec V \prec H \prec A \prec C \prec S \prec Z \prec M$
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	$H \prec V \prec J \prec C \prec Z \prec A \prec S \prec M$

$O_m : \{($  $\succ$  $), \ldots\},$

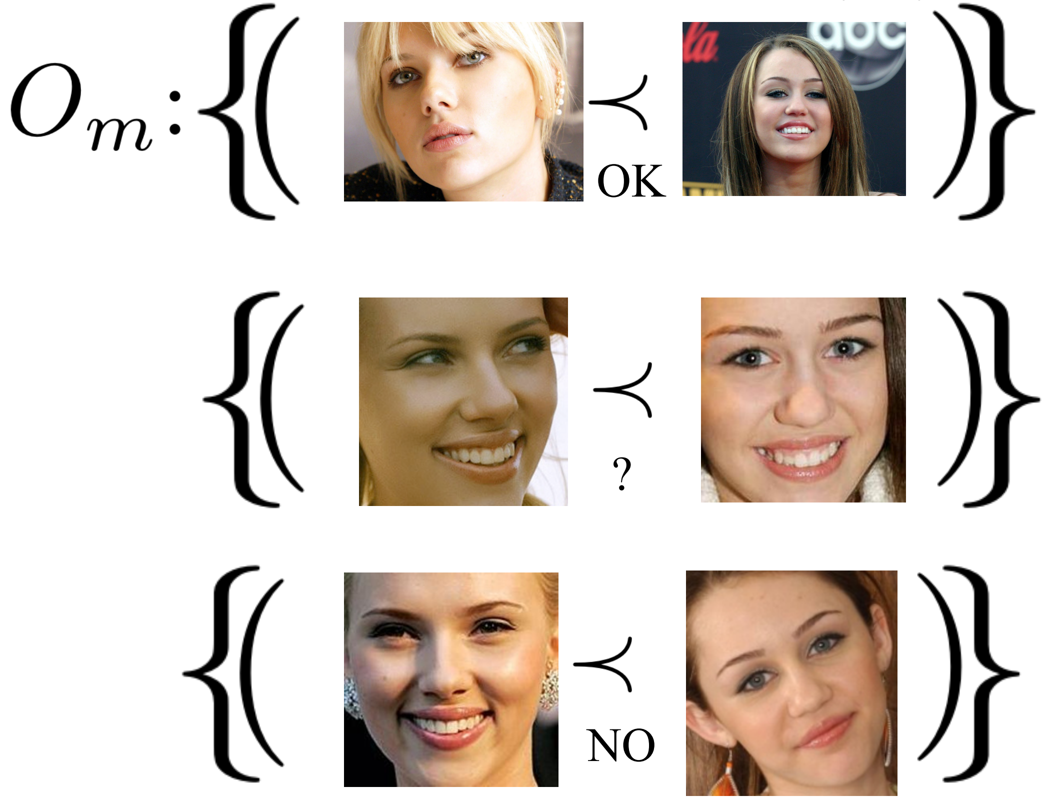
$S_m : \{\{$  $\sim$  $\}, \ldots\}$

Introduction: pairwise ranking

- Coarse labeling at category level => noisy pair sampling

	Binary	Relative
OSR	TI S HC OMF	
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

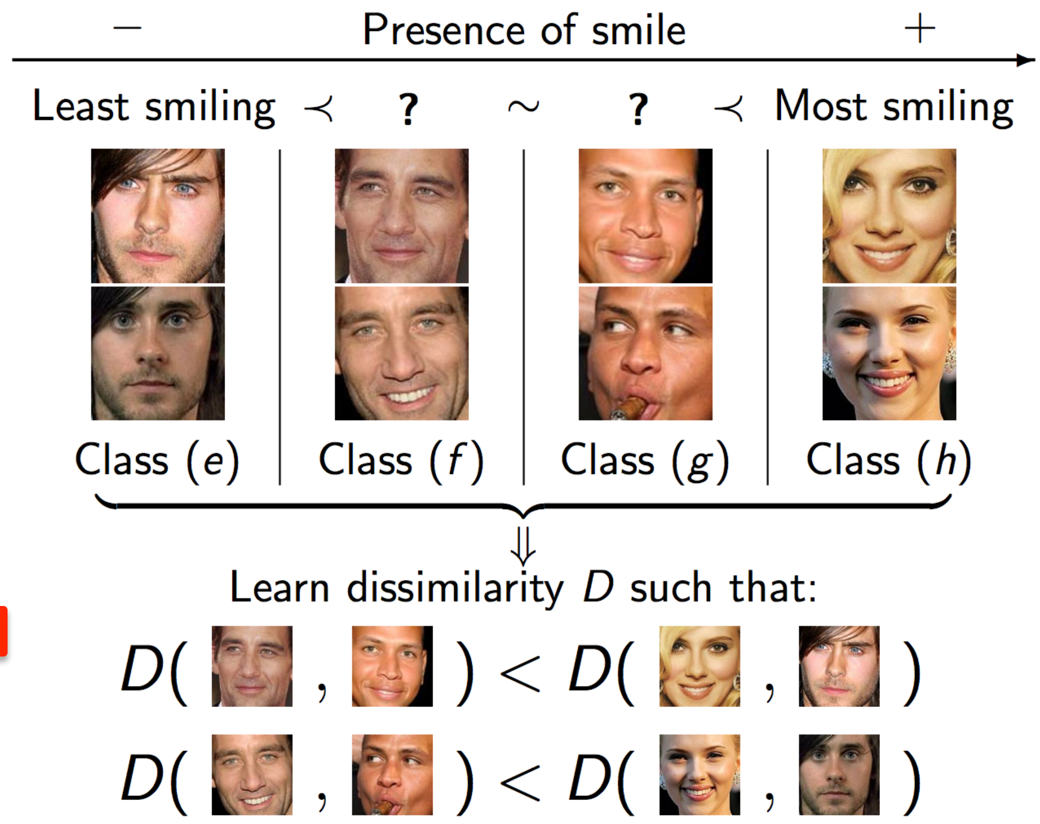
Scarlett Johansson *vs* Miley Cyrus



- Proposition: see the problem as a specific Metric Learning problem with exotic supervision data

Qwise: Quadruplet-wise ML

	Binary	Relative
OSR	TI S HC OMF	
natural	0 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 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 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 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 0	$F \prec O \sim M \prec C \prec I \sim S \prec H \prec T$
close-depth	1 1 1 1 0 0 0 1	$C \prec M \prec O \prec T \sim I \sim S \sim H \sim F$
PubFig	ACHJ MS V Z	
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	$H \prec V \prec J \prec C \prec Z \prec A \prec S \prec M$



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

Outline

1. Introduction
   - ICCV paper on relative attributes
   - Other approach: from pairwise to Qwise
2. **Quadruplet-wise Metric Learning Model**
   - Training data
   - Distance and objective function
   - Optimization scheme
3. Application to relative attribute learning
4. Qwise for hierarchical classification
5. Qwise for Web page comparison

Qwise Metric Learning

- Key ingredients of (our) similarity learning:
  - Data representation including both the feature space and the similarity function
  - Learning framework
    - training data, type of labels and relations,
    - Optimization formulation
    - Solvers

Qwise Metric Learning

- Data representation:
  - Image (p) and features (x): GIST, Bag of Words, BossaNova, Bio-inspired, Deep …
  - Similarity function:
    - Most popular: Mahalanobis-like distance metric
    - M symmetric matrix

$$D^2_{\mathbf{M}}(p_i, p_j) = \Phi(p_i, p_j)^\top \mathbf{M} \Phi(p_i, p_j), \mathbf{M} \succeq 0$$

Qwise Metric Learning

- Constraints (strict) on quadruplets $q = (p_i, p_j, p_k, p_l)$ using margin τ:

$$D(p_k, p_l) \geq \tau + D(p_i, p_j)$$

- 2 different learning frameworks
  - Decomposing $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$ and optimizing over the rows $\mathbf{w}_m$ of $\mathbf{L}$.
  - Diagonal PSD matrix $\mathbf{M} = \mathrm{Diag}(\mathbf{w})$, $\mathbf{w} \geq 0$

- In both cases, metric learning expressed as a linear combination with $\mathbf{w}$ (Ψ equal to Φ or Φ^2):

$$D_{\mathbf{w}}(p_i, p_j) = \mathbf{w}^\top \, \Psi(p_i, p_j)$$

- Constraints (again):

$$D(p_k, p_l) - D(p_i, p_j) = \mathbf{w}^\top [\Psi(p_k, p_l) - \Psi(p_i, p_j)] = \mathbf{w}^\top \mathbf{z}_q \geq \tau$$

Qwise Metric Learning

- L_1^h loss function differentiable approximation of the hinge loss inspired by the Huber Loss function (as described in [Chapelle NeurComp. 07]) with $t = \mathbf{w}^\top \mathbf{z}_q$:

$$L_1^h(t) = \begin{cases} 0 & \text{if} \quad t > 1+h \\ \frac{(1+h-t)^2}{4h} & \text{if} \quad |1-t| \leq h \\ 1-t & \text{if} \quad t < 1-h \end{cases}$$

- Usually $h \in [0.01, 0.5]$, here h set to 0.05

- Optimization scheme:

$$\min_{\mathbf{w}} \sum_{q \in \mathcal{A}} L_1^h(\mathbf{w}^\top \mathbf{z}_q) + \lambda \|\mathbf{w}\|_2^2 \tag{1}$$

with a regularization term over $\mathbf{w}$

- with additional light constraints:

$$\min_{\mathbf{w}} \sum_{q \in \mathcal{A}} L_1^h(\mathbf{w}^\top \mathbf{z}_q) + \sum_{q \in \mathcal{B}} L_0^h(\mathbf{w}^\top \mathbf{z}_q) + \lambda \|\mathbf{w}\|_2^2$$

Qwise Metric Learning

- Solver:
  - **Convex optimization** problem
  - With such a regularization, scheme similar to ranking SVM, except loss functions on quadruplets and constraints on w
  - Differentiable => Solving a primal problem using Newton's method [Chapelle10]
  - Complexity linear in the nb constraints => efficiently solved even with a large number of constraints
  - "Small" number of parameters (grows linearly with the input space) => limiting overfitting

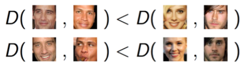
- T. Joachims, "Optimizing Search Engines using Clickthrough Data", ACM Conference on Knowledge Discovery and Data Mining, 2002
- O. Chapelle, S. Keerthi. Efficient algorithms for ranking with svms. Inf. Retrieval, 2010

Outline

1. Introduction
2. Quadruplet-wise Metric Learning Model
3. **Application to relative attribute learning**
4. Qwise for hierarchical classification
5. Qwise for Web page comparison

Relative attribute learning

$$\min_{\mathbf{w}} \|\mathbf{w}\|_2^2 + C \sum_{(p_i, p_j, p_k, p_l)} L_1^h \left(\mathbf{w}^\top \left[\Psi(p_k, p_l) - \Psi(p_i, p_j)\right]\right)$$



- $\mathbf{x}_i \in \mathbb{R}^d$: GIST (+ color) descriptor
- $\Psi(p_i, p_j) = \mathbf{x}_i - \mathbf{x}_j$
- Relative attributes a_m for $m \in \{1, \ldots, M\}$: 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 $\mathbf{L} \in \mathbb{R}^{M \times d}$:

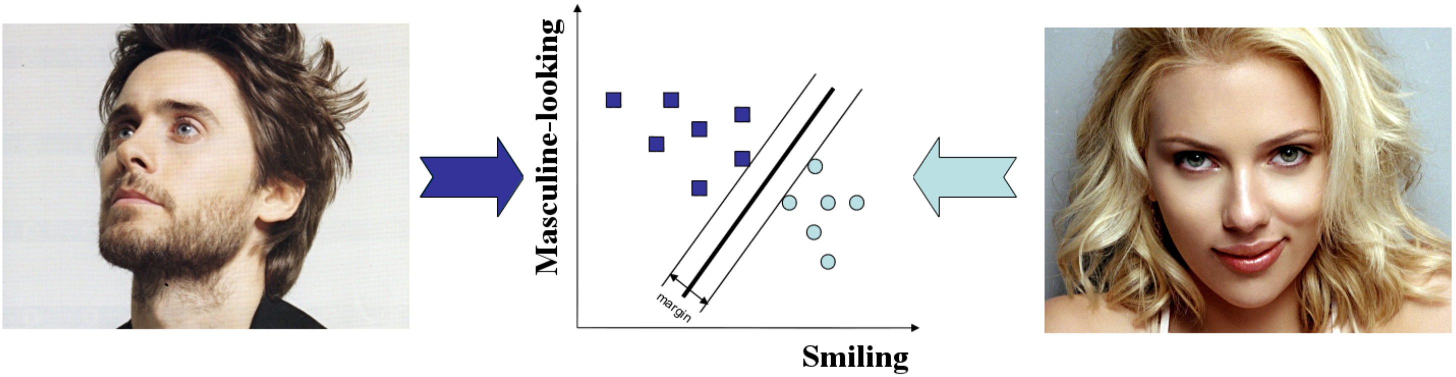
$$\mathbf{L} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,d} \\ \vdots & \vdots & \vdots \\ w_{M,1} & \cdots & w_{M,d} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1^\top \\ \vdots \\ \mathbf{w}_M^\top \end{bmatrix}, \quad \mathbf{w}_m^\top : m\text{-th row}$$

Relative attribute learning

- Learning a feature space

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

- 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:



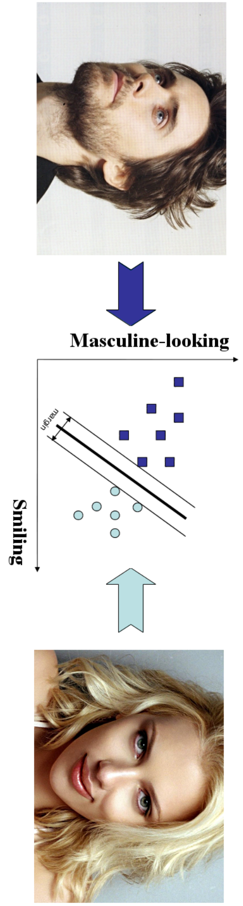
Relative attribute experiments

- 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, shubby …

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



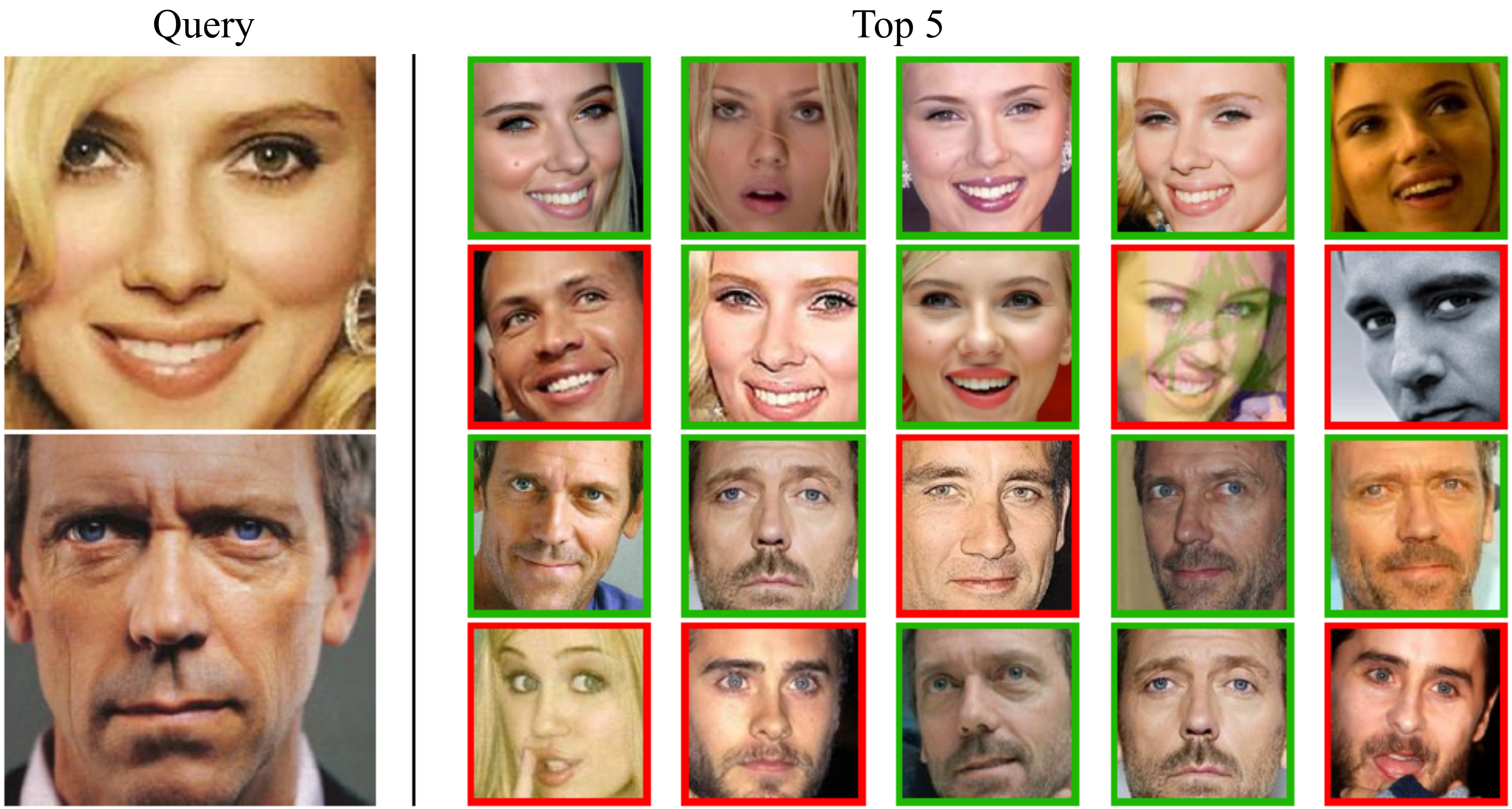
[Wein.09] K.Q. Weinberger, and L.K. Saul, Distance metric learning for large margin nearest neighbor classication, In JMLR 2009

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	$\mathbf{74.6} \pm \mathbf{1.7}\%$	$76.5 \pm 1.2\%$
Qwise + LMNN	$74.3 \pm 1.9\%$	$\mathbf{77.6} \pm \mathbf{2.0}\%$

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

Relative attribute experiments



Relative attribute experiments

Outline

1. Introduction
2. Quadruplet-wise Metric Learning Model
3. Application to relative attribute learning
4. **Qwise for hierarchical classification**
5. Qwise for Web page comparison

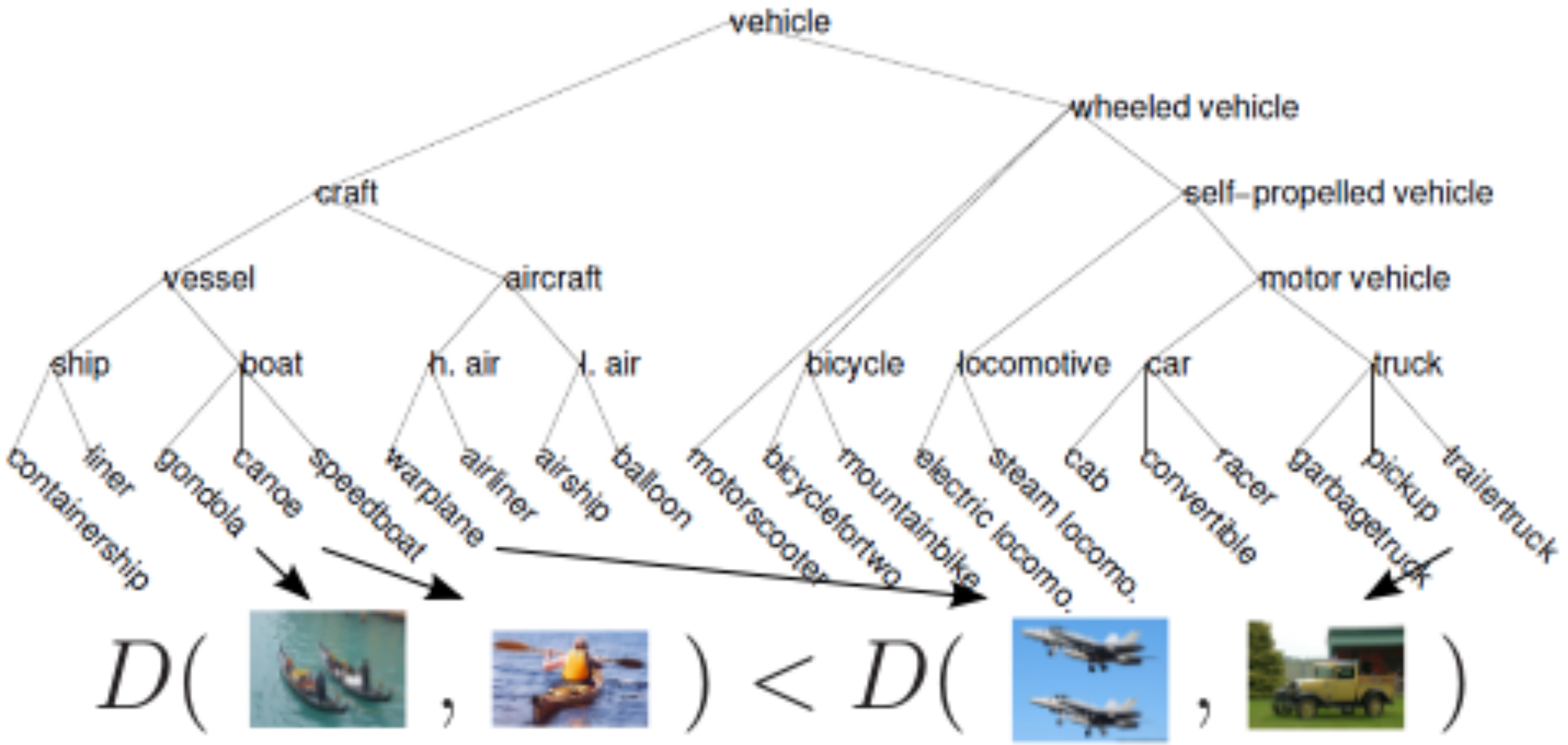
Taxonomy ML

- Hierarchical image classification:
  - Qwise to learn taxonomy
- Context:
  - Rich annotations using a semantic taxonomy structure
  - How to exploit complex relations from a class hierarchy as proposed in [Verma12]:
    - Learn a metric such that images from close (sibling) classes with respect to the class semantic hierarchy are more similar than images from more distant classes

[Verma12] N. Verma, D. Mahajan, S. Sellamanickam, and V. Nair. Learning hierarchical similarity metrics. In *CVPR,* 2012.

Taxonomy ML

- Qwise constraint generation:



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 (provided by ImageNet)
- Diagonal PSD matrix framework: $\mathbf{w} \geq 0$
- **Convex Optimization Problem**:

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

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.

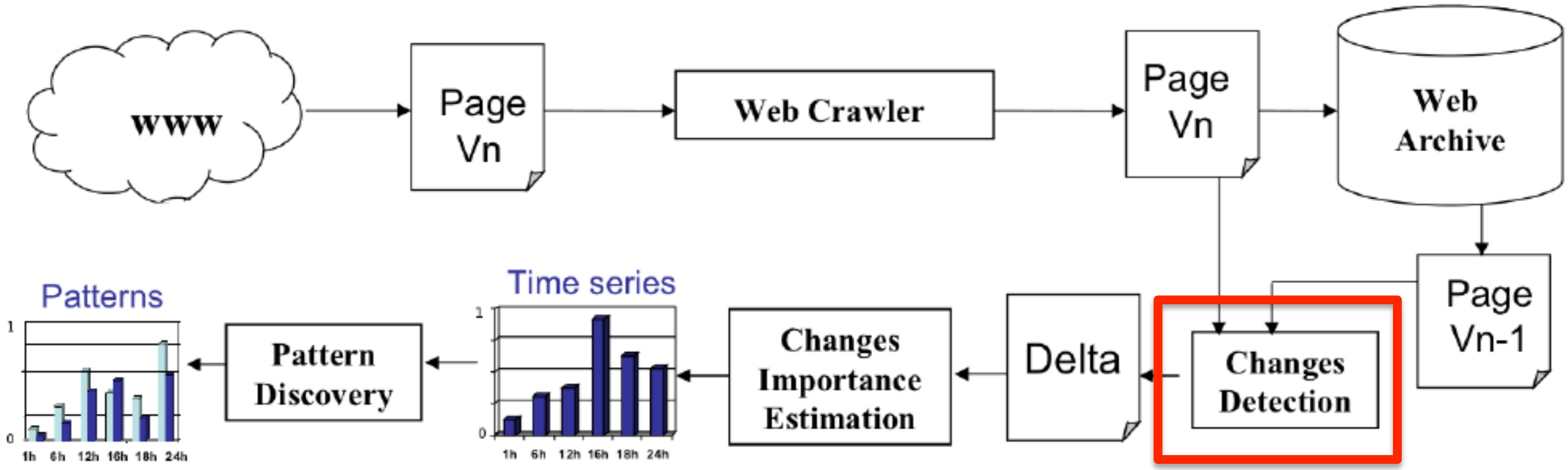
- **9 datasets** from ImageNet, for each dataset: from 8 to 40 different classes, from 8,000 to 54,000 images for training

Outline

1. Introduction
2. Quadruplet-wise Metric Learning Model
3. Application to relative attribute learning
4. Qwise for hierarchical classification
5. **Qwise for 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

$$D(v_t, v_{t+1}) < D(v_{t-1}, v_{t+2})$$

Web page ML

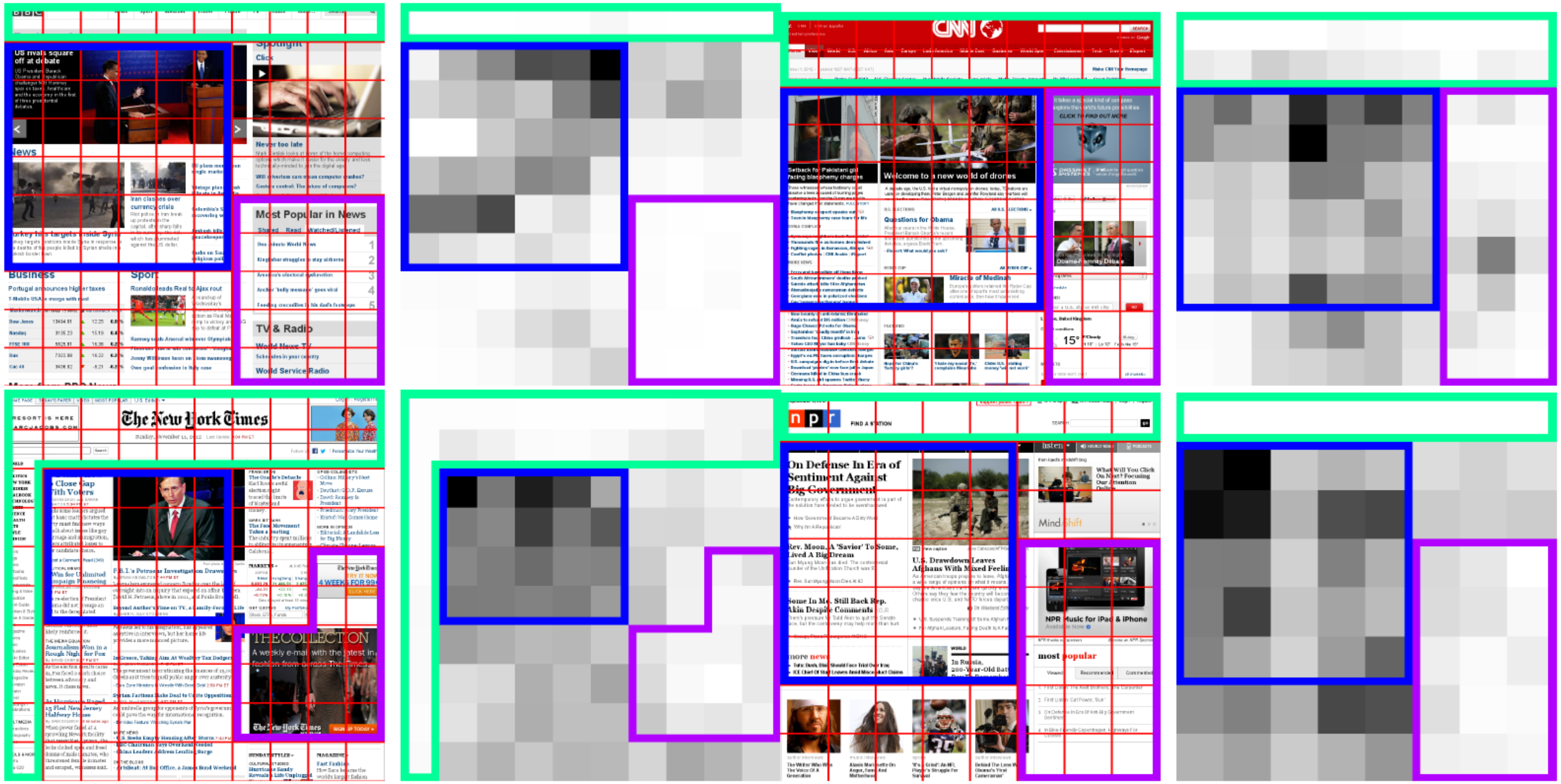
- Descriptors: GIST on m-by-m grid over screenshots
- **Ψ** is a m-by-m vector of Euclidean distance between blocks
- Diagonal PSD matrix: **w** represents block weights
- Optimization over **w**
  - Learning of spatial weights of webpage regions using temporal relationships
  - Automatically
    - Discovering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

Web page ML

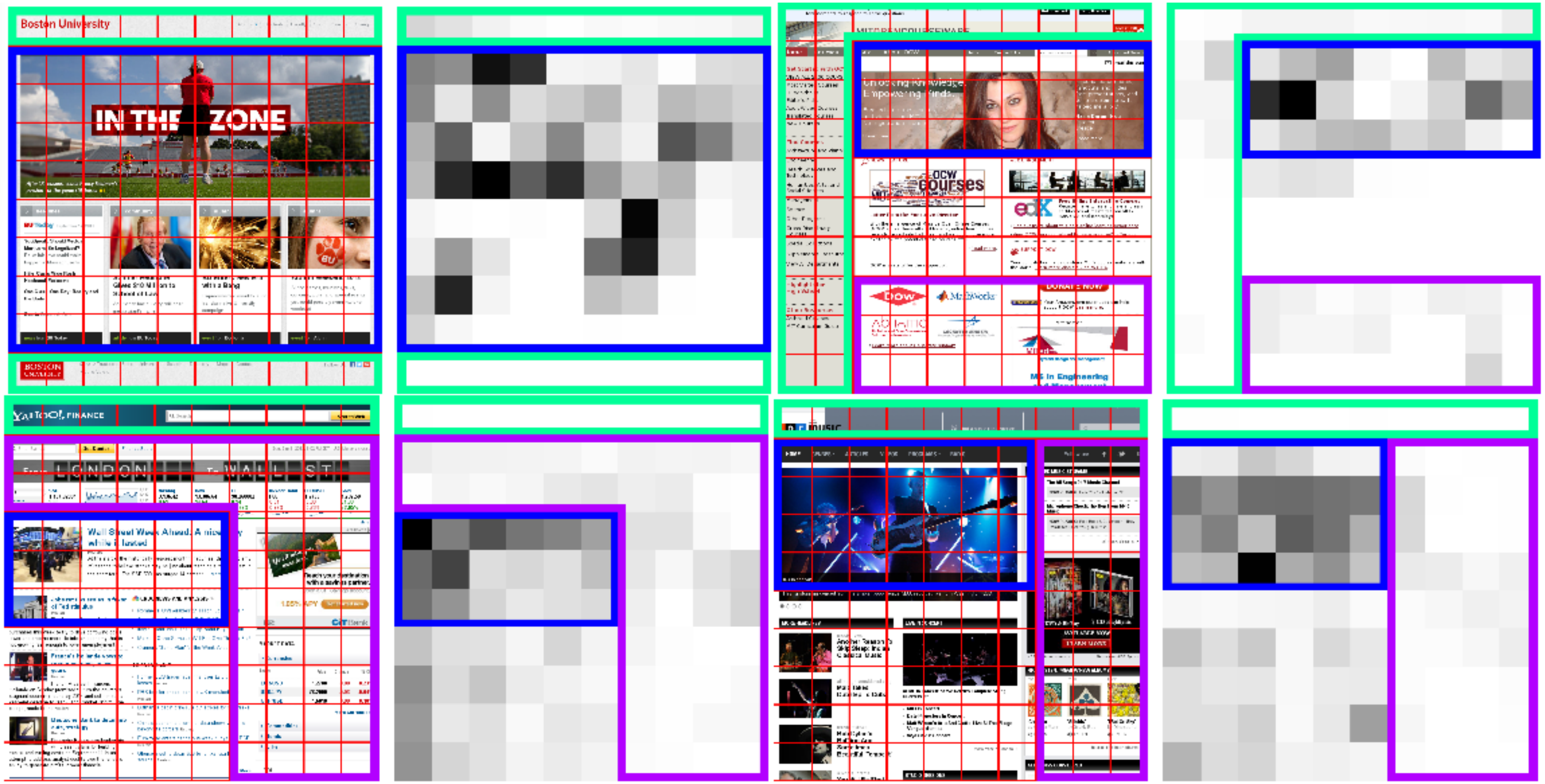
- Evaluation and Comparison
  - Crawling 50 days Several sites CNN, NPR, BBC, …
  - Manual change detection (news updates) for GT on 5 days
  - Baselines: Euclidean Dist, LMNN
  - GIST on 10x10
  - Mean Average Precision on succ. Web page Metric scores

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

Web page ML



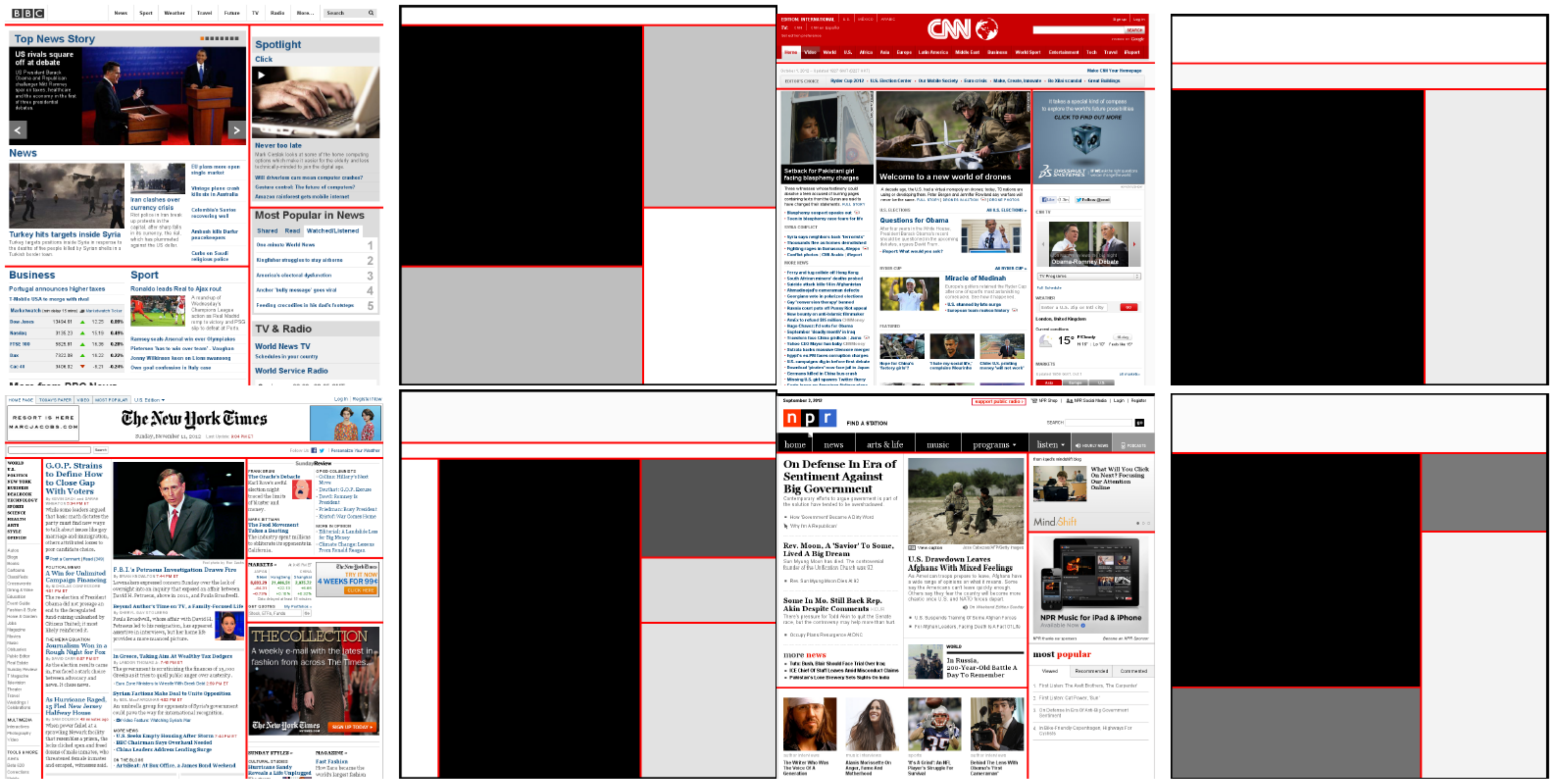
Web page ML



- Not connected to the structural layout of the Web page

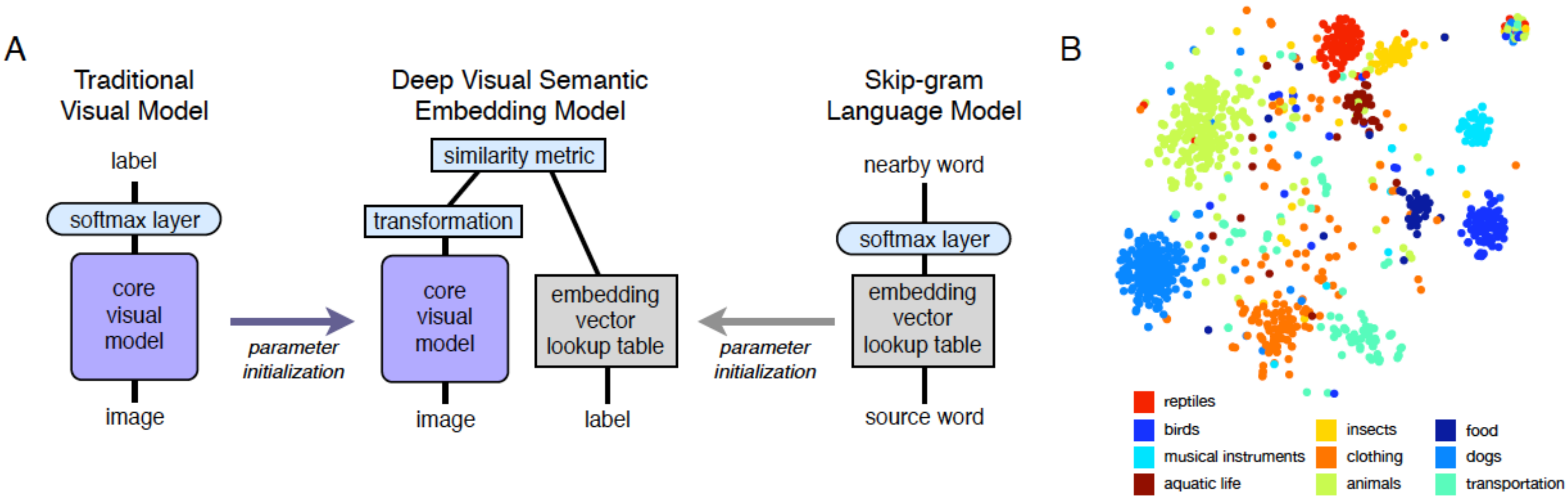
Web page ML

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



Qwise Metric Learning

- Similarity function
  - Full M Regularization => trace(M), early stopping [Law CVPR14]
- Scalability
- Temporal/spatial relationships, class relationships => rich context to learn metrics or semantic embedding



DeVISE system (google NIPS 2013)

Ref.

Matthieu Cord
Joint work with Marc T. Law and Nicolas Thome
LIP6, Computer Science Department
UPMC Paris 6 - Sorbonne University
Paris FRANCE
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

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

Web page Segmentation:

- A. Sanoja, S. Gancarski, Yet another hybrid segmentation tool, In International Conference on Preservation of Digital Objects. 2012