Representation Learning for Image/Video Understanding

Nicolas Thome

Université **P**ierre et **M**arie **C**urie
Laboratoire d'**I**nformatique de **P**aris **6**

12 Septembre 2014

Web Science Workshop
GDRI

MultiMedia group at LIP6/DAPA/MALIRE

People

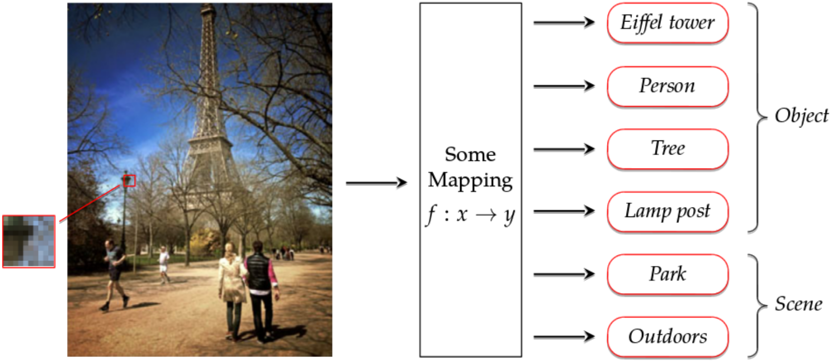
1. LIP6 lab in Paris
   - $\sim$ 150 permanent researchers, $\sim$ 250 Phd students
2. DAPA department: Databases and Machine learning
   - $\sim$ 35 permanent researchers, $\sim$ 50 Phd students
3. MLIA team: MAchine Learning and Information Acess (P. Gallinari)
   - $\sim$ 10 permanent researchers, $\sim$ 20 Phd students
4. MultiMedia group: Matthieu Cord
   - 2 permanent researchers (M. Cord, N.Thome), $\sim$ 10 Phd/Post-docs

Outline

Context

Semantic annotation of visual data

- Holy Grail of computer vision
- Filling the semantic gap: extremely challenging



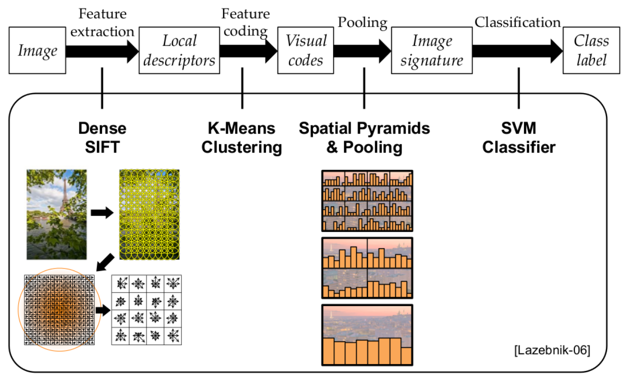
Set of image pixels: x

Image labels: y

Semantic annotation

Handcrafted features

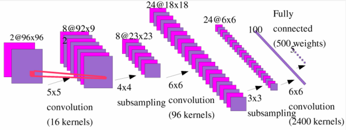
- Last decade : supremacy of robust local features: SIFT, STIP, *etc*
- Edge-based features
- Embedded into a coding/pooling framework: BoW model



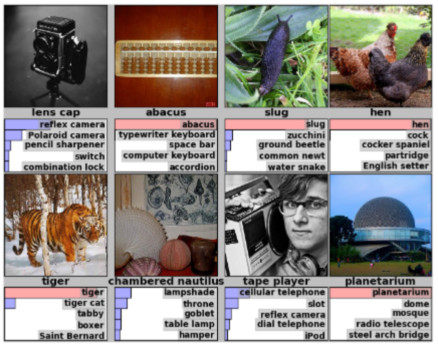
Semantic annotation

Deep Learning: Learning Representations from data

- Image/Video : Convolutionnal Neural Networks (CNN)



Used since the 80's

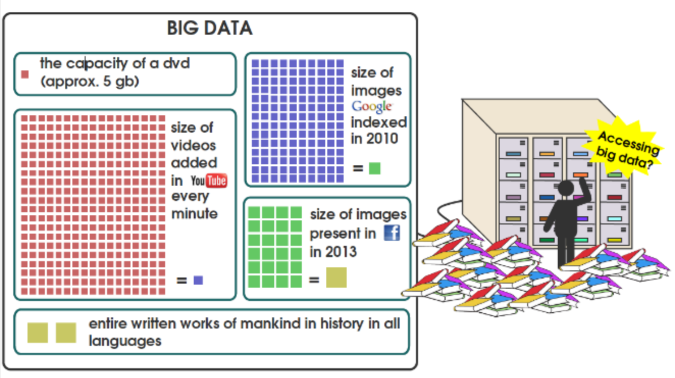


- ⊕ deep models
- ⊖ difficult to train
  - Many parameters, requires lots of data
  - Overfitting

- 2012:Big data (10^6 images, 10^3 classes)
- Computational resources (GPU)

Representation Learning

- Importance of learning representation from data (transfer learning)
- Supervised *vs* unsupervised learning
- big data: huge number of unlabeled data, many (but fewer) labeled data

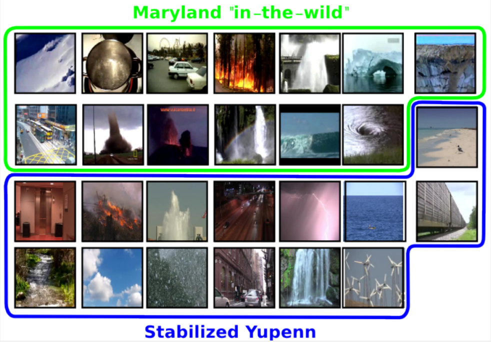


Outline

Dynamic Scene Classification

Context

- Recognition of complex dynamic natural scenes

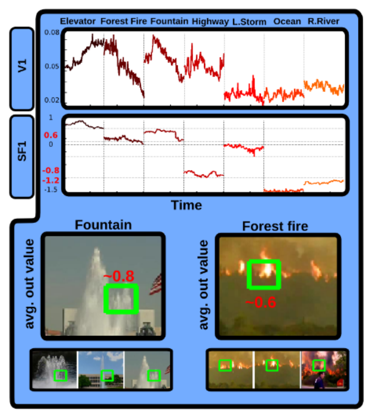


- Computer vision descriptors such as HOF [MLS09], LDS [DCW+03] not adapated to such context [DLD+12]
  - HOF: Constant illumination constraints
  - LDS: 1st order markovian assumption
- Our idea: unsupervised learning of motion descriptors

Dynamic Scene Classification

Unsupervised learning of motion descriptors

- Manifold Untangling



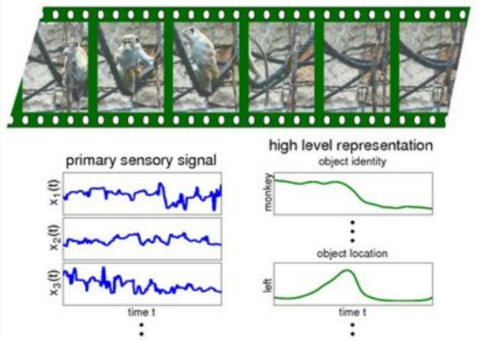
Contributions:

- Using Slow Feature Analysis (SFA) for learning stable motion descriptors
  - Compact description (low dimensional space)
- Embedded into a coding/pooling architecture
- Outperforming state-of-the-art performances in 2 challenging dynamic scenes databases

Slow Feature Analysis

Intuition

- Measurements are noisy/chaotic, perceptions are stable [WS02, BW05]



- Idea: learning data representations that "slow down" the signal
- Goal: slow component capture relevant motion features

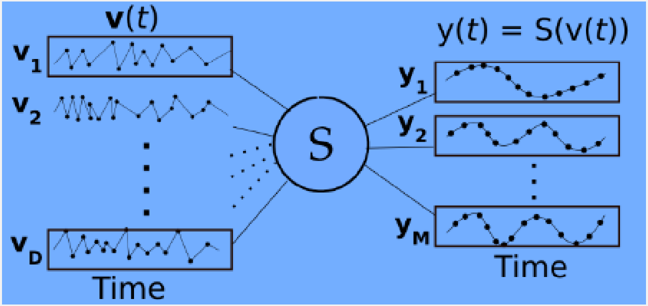
Source : http://www.scholarpedia.org/article/Slow_feature_analysis

[WS02] L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. Neural Computl, 2002.
[BW05] P.Berkes and L. Wiskott . Slow feature analysis yields a rich repertoire of complex cell properties J.Vision, 2005.

Slow Feature Analysis

Formulation

- Input : D-dimensional temporal signal $\mathbf{v}(t) = [v_1(t)v_2(t)...v_D(t)]^T$
- Output : M-dimensional temporal signal $\mathbf{y}(t) = [y_1(t)y_2(t)..y_M(t)]^T$



- Linear model $y_j(t) = S_j v(t)$, $\forall t$ et $\mathbf{S} \in \mathbb{R}^{D \times M}$

Slow Feature Analysis

Formulation

- $y_j(t) = S_j v(t)$, $\forall t$ et $\mathbf{S} \in \mathbb{R}^{D \times M}$. Let us define:
  - $\langle y \rangle_t$ temporal average of y
  - $\dot{y}$ temporal derivative of y
- SFA objective function:

$$\min_{S_j} \langle \dot{y_j}^2 \rangle_t \tag{1}$$

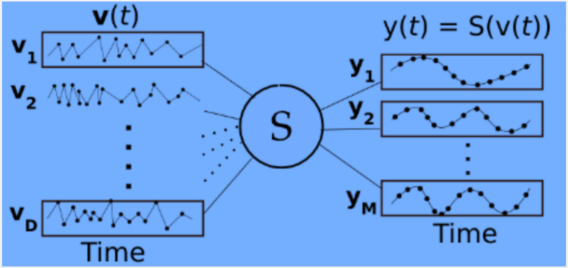
  Under the constraints:

  1. $\langle y_j \rangle_t = 0$ (zero mean)
  2. $\langle y_j^2 \rangle_t = 1$ (unit variance)
  3. $\forall j < j' :\ \langle y_j, y_{j'} \rangle_t = 0$ (decorrelation)
- Can be rewritten as:

$$\langle \dot{\mathbf{v}} \dot{\mathbf{v}}^T \rangle_t S_j = \lambda_j S_j \tag{2}$$

Slow Feature Analysis

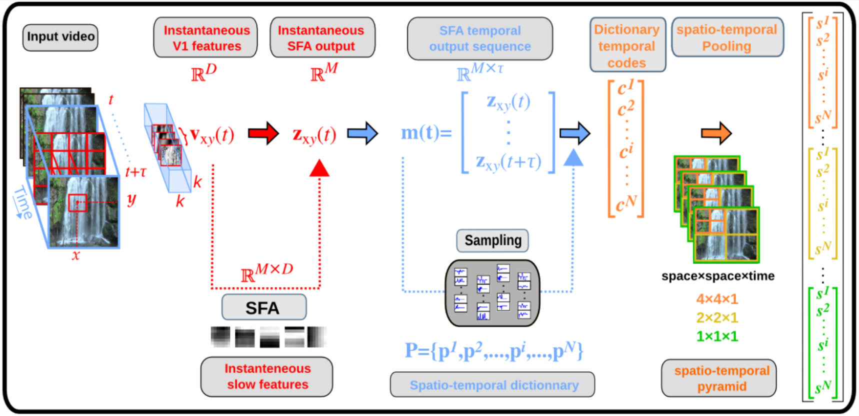
Formulation



- Can be rewritten as: $\langle \dot{\mathbf{v}}\dot{\mathbf{v}}^T \rangle_t S_j = \lambda_j S_j$
- $\dot{\mathbf{v}}\dot{\mathbf{v}}^T$ diagonalization
- Keeping M eigenvectors associated with the **smallest eigenvalues**

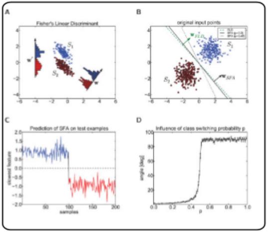
Global Video Representation

SFA embedded into a coding/pooling scheme

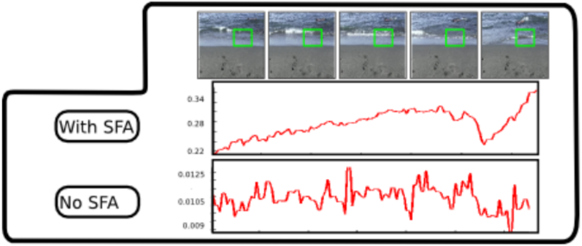


Slow Feature Analysis

Connection SFA $\leftrightarrow$ LDA



Credit [KM09]



- Small variations ignored
- Dominant/stable components of the motion encoded

[KM09] Klampfl S, Maass W. Replacing supervised classification learning by Slow Feature Analysis in spiking neural networks, Advances in Neural Information Processing Systems 22, 988-996, 2010. MIT Pres.

Experiments

Classification results

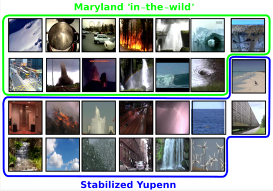
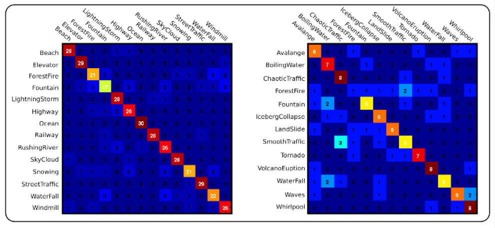


Table: Recognition Rate (%) on dynamic scene datasets

	HOF	GIST	Chaos	SOE	Ours
Maryland	17	38	36	41	**60**
Yupenn	59	56	20	74	**85.5**

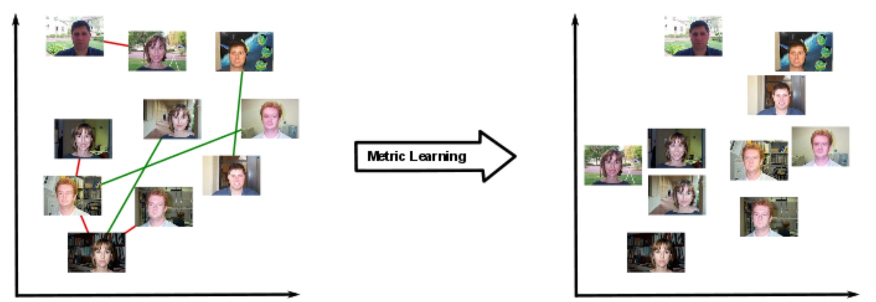


- Based on V1 features
- Both SFA learning and coding/pooling scheme improve performances
- Very competitive wrt state-of-the-art methods (mono-feature results)

Outline

Metric Learning

Context



- Learning a metric: important for many applications
- Difference wrt standard classification contexts:
  - Notion of similar/dissimilar $\neq$ class labels
  - Large scale:
    - Adding new classes does not require to retrain the whole model
    - Zero-shot learning

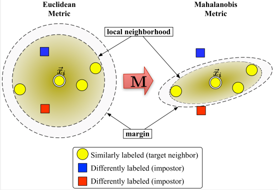
Metric Learning

Context

- Mahalanobis-like Metric Parametrization (matrix $\mathbf{M}$ SDP): $D^2_{\mathbf{M}}(\mathcal{I}_i, \mathcal{I}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}\ (\mathbf{x}_i - \mathbf{x}_j) = \langle \mathbf{M}, \mathbf{x}_{ij}\mathbf{x}_{ij}^\top \rangle = \langle \mathbf{M}, \mathbf{C}_{ij} \rangle$
- Supervised metric learning: training set $\mathcal{A}$ with elements e

$$\min_{\mathbf{M}} \mu R(\mathbf{M}) + \sum_{e \in \mathcal{A}} \ell(\mathbf{M}, e) \tag{3}$$

- R regularization term, $\ell(\mathbf{M}, e)$ data-dependent, *e.g.* based on:
  - Pairs: $e = (\mathcal{I}_i, \mathcal{I}_j)$. e similar $\Rightarrow D^2_{\mathbf{M}}(\mathcal{I}_i, \mathcal{I}_j) < u$, e dissimilar $\Rightarrow D^2_{\mathbf{M}}(\mathcal{I}_i, \mathcal{I}_j) > l$
  - Triplets: $e = (\mathcal{I}_i, \mathcal{I}_i^+, \mathcal{I}_i^-)$, *e.g.* LMNN [WS09]: $D_{\mathbf{M}}(\mathcal{I}_i, \mathcal{I}_i^+) < D_{\mathbf{M}}(\mathcal{I}_i, \mathcal{I}_i^-) + 1$



[WS09] Weinberger, K. Q.; Saul L. K. Distance Metric Learning for Large Margin Classification. Journal of Machine Learning Research 10: 207244, 2009.

Quadruplet-wise Metric Learning

Quadruplets

- Constraints involving up to 4 images: $e = (\mathcal{I}_i, \mathcal{I}_j, \mathcal{I}_k, \mathcal{I}_l)$
- $D^2_{\mathbf{M}}(\mathcal{I}_k, \mathcal{I}_l) \geq D^2_{\mathbf{M}}(\mathcal{I}_i, \mathcal{I}_j) + \delta$
- Any pair or triplet constraint can be expressed with quadruplets
- However, converse not true $\Rightarrow$ only relative distances with quadruplets
  - More general/flexible constraints, useful in various applicative contexts

Optimization Scheme

Objective function:

$$\min_{\mathbf{M} \in \mathbb{S}^d_+} \boldsymbol{R}(\mathbf{M}) + C_q \sum_{q \in \mathcal{A}} \xi_q$$

$$\text{s.t.} \forall \boldsymbol{q} \in \mathcal{A} : D^2_{\mathbf{M}}(\mathcal{I}_k, \mathcal{I}_l) \geq D^2_{\mathbf{M}}(\mathcal{I}_i, \mathcal{I}_j) + \delta_q - \xi_q$$

$$\xi_q \geq 0 \tag{4}$$

- Eq. 4 with full matrix $\mathbf{M}$: solved using projected (PSD cone) gradient descent
- Simplification for diagonal matrices ($\sim$ ranking SVM)

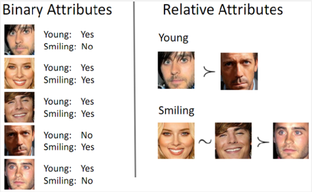
Which contexts can benefit from QWise constraints ?

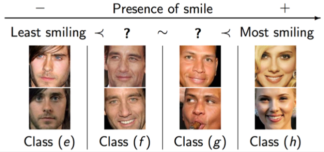
Application: Relative Attributes

- Attributes: Mid-level concepts (higher than low-level features, lower than high-level categories)

- RA datasets: annotation provided at the class level

- Relative Attributes (RA) [PG11]: Ranking two images wrt attributes easier than binary labeling

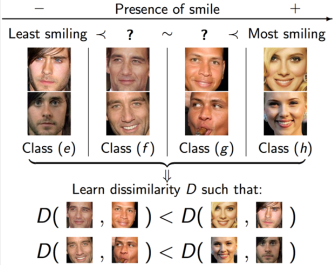




[PG11] Devi Parikh, Kristen Grauman. Relative attributes, ICCV, pp.503-510, 2011.

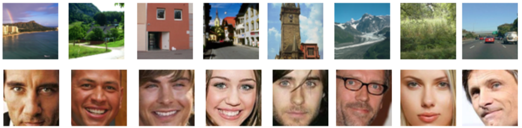
Which contexts can benefit from QWise constraints ?

QWise constraints for learning Relative Attributes



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

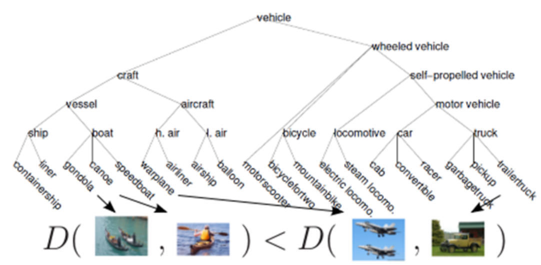
- QWise constraints more robust to noise in the labeling: second row, ranking should rather be (g) $\prec$ (f) $\sim$ (h)
- Learning $\mathbf{M} = \mathbf{L}^T\mathbf{L}$: each row of $\mathbf{L}$ is a parameter vector for learning RA's
- Experiments on OSR and PubFig datasets
  - QWise outperforms baseline [PG11] based on pairs
  - Complementary to class labels used in LMNN



Which contexts can benefit from QWise constraints ?

Hierarchical classification

- Qwise to learn taxonomy:
  - 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 classe



- Learning a full matrix **M**
- Improved classification performances

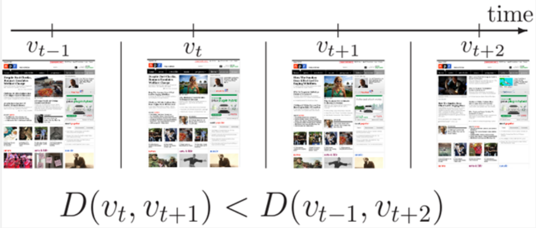
[Verma12] N. Verma, D. Mahajan, S. Sellamanickam, and V. Nair. Learning hierarchical similarity metrics. In CVPR, 2012.

Which contexts can benefit from QWise constraints ?

Web archiving: change detection

- Web crawling: useful to understand the change behavior of websites over time
  - Significant changes between successive versions of a same webpage ⇒ revisit the page
- Focus on news websites
  - Advertisements or menus not significant
  - News content significant

- Qwise Constraints:
  - Fully unsupervised, but temporal information available
  - Comparing screenshots of successive versions

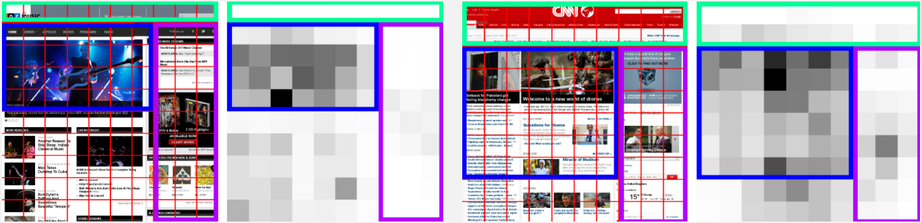


Which contexts can benefit from QWise constraints ?

Web archiving: change detection

- Evaluation: 50 days on CNN, NPR, BBC, NYT
- GT annotation for change detection (news updates) on 5 days
- Features: GIST on a 10x10 grid
- Metric: MAP on succ. Web pages

Site	CNN			NPR		
Eval.	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
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
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**
	New York Times			BBC		
	AP_S	AP_D	MAP	AP_S	AP_D	MAP
	69.8 ±0.9	79.5 ±0.4	74.6 ±0.5	91.1 ±0.3	76.7 ±0.6	83.9 ±0.4
	83.2 ±1.4	89.1 ±2.7	86.1 ±2.0	92.5 ±0.4	**80.1 ±1.0**	**86.3 ±0.6**
	85.5 ±5.4	**92.3 ±4.1**	**88.9 ±4.6**	**92.8 ±0.4**	79.3 ±1.3	86.1 ±0.8



Conclusion

Representation Learning

- Two Methods for learning representations:
  - An unsupervised method for learning motion descriptors (SFA)
  - A supervised metric learning scheme that can encompass exotic (beyond binary labels) annotations and tackles various applications
- Extension of our metric learning work on the regularization side $\Rightarrow$ explicit control over the rank of the learned matrix
- Joint work with **C. Thériault**, **M.T. Law**, **M. Cord** and **P. Pérez**.

Publications

- Slow Feature Analysis

C. Thériault, N. Thome and M. Cord, P. Pérez. Perceptual principles for video classification with Slow Feature Analysis, IEEE Journal of Selected Topics in Signal Processing, p. 1-10, vol 99, April 2014

C. Thériault, N. Thome and M. Cord. Dynamic Scene Classification: Learning Motion Descriptors with Slow Features Analysis, CVPR 2013

- 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, DocEng, 2012

Conclusion

Projects

- ANR
  - Finished: ASAP (deep learning), ITOWNS, GeoPeuple
  - VISIIR started on oct. 2013 on interative learning with eye-tracker
- European SCAPE Project
- Bilateral franco-brazilian CAPES-COFECUB. Collaborations::
  - UNICAMP: E. Valle, R. Torres, J. Stolfi
    - R. Minetto Phd Thesis
  - UFMG: A. de Albuquerque, S. Jamil,
    - S. Avila Phd Thesis

Questions ?