From local descriptors to holistic models: a general framework for video processing and object representation

Antoine Manzanera
ENSTA-ParisTech

IPTA 2014, Paris
Oct, 16, 2014

Context and Motivations

Purpose

Reduce the computational gap between "low level" and "high level" processing tasks using unified visual models and processing framework.

Context

Robotics and Computer Vision Lab at ENSTA-ParisTech/U2IS
$\rightarrow$ Embedded / Mobile / Hybrid / Autonomous Systems.

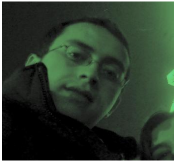
Context and Motivations

The global objective of our research is to design efficient *and* versatile vision systems, by addressing the problem *globally*, from the image models to the parallel implementation.

- High Performance Computer Vision: Performance *and* Versatility
- Embedded Vision Systems: Opportunism and Redundancy Tracking
- Smart Sensors: Hybrid and Active

This presentation focuses on the fundamental parts: image and processing models

Acknowledgements



Fabio Martínez Carrillo

PhD student

Motion analysis

Matthieu Garrigues

PhD student

Video processing

Thanh Phuong Nguyen

Post-doc researcher

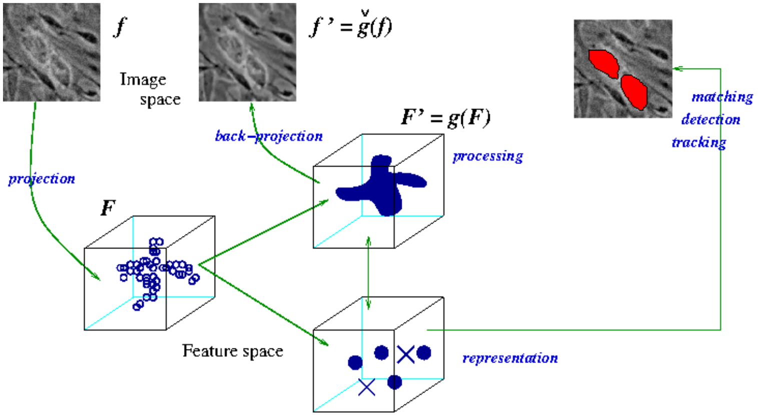
Visual recognition

Antoine Tran

PhD student

Image modelling

General view of the Approach



Presentation Outline

Introduction

Local Jet Feature Space

Image Processing
- NL-Means Filtering
- Optical Flow
- Lines and Circles Detection

Object Representations
- Basic representations
- Background Modelling
- Kernel Based Tracking
- Dense Implicit Shape Models

Conclusions

Related works (1)

Manifold Image Processing

The projected data form a manifold in the feature space [Peyré 09]. Image processing operates on the manifold, then back-project the data onto the image space.

Scale Space Derivatives

The multiscale derivative representation is biologically [Koenderink 87] and mathematically [Lindeberg 98] founded.

Filter Banks and Codebook

Many visual representation frameworks, for texture (e.g. textons) or objects (e.g. visual bag of features) are based on filter banks and clustering [Freeman 91], [Rubner 99].

Related works (2)

Nearest Neighbour Search

Efficient data coding may be necessary to represent sparse feature spaces with few memory, and efficient Nearest Neighbour Search [Arya and Mount 07].

Hough Transforms

Accumulation techniques in a parameter space are relevant for modelling and detection using feature space representation [O'Gorman 76], [Valenti 08].

Kernel Based Methods

The distribution of significant feature components can be used to represent objects [Comaniciu 03].

Presentation Outline

- Introduction
- Local Jet Feature Space
- Image Processing
  - NL-Means Filtering
  - Optical Flow
  - Lines and Circles Detection
- Object Representations
  - Basic representations
  - Background Modelling
  - Kernel Based Tracking
  - Dense Implicit Shape Models
- Conclusions

Multiscale Gaussian Derivatives

Our basic feature space is the Multiscale Gaussian Local Jet:

Projection in the Local Jet space

$$\mathbf{p} = (x_{\mathbf{p}}, y_{\mathbf{p}}) \mapsto \hat{\mathbf{p}} = (a_{ij}^{\sigma} f_{ij}^{\sigma}(\mathbf{p}))_{i+j \leq r; \sigma \in S}$$

Gaussian multiscale local jet

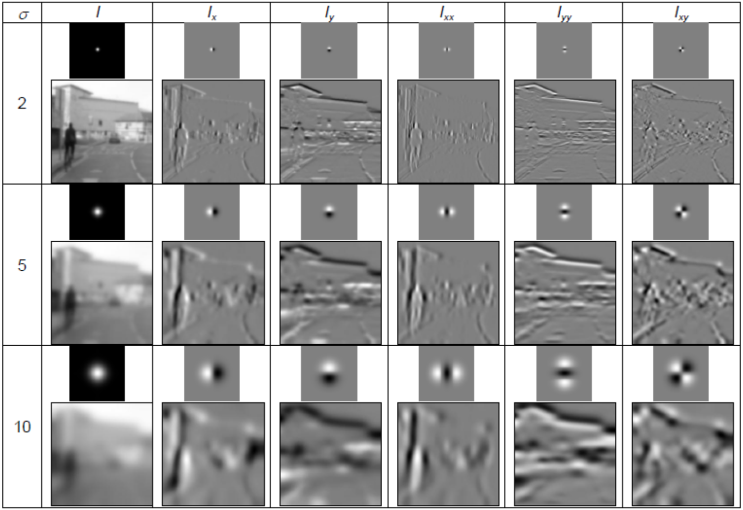
$$f_{ij}^{\sigma} = f \star \frac{\partial^{i+j} G_{\sigma}}{\partial x^i \partial y^j}$$

Normalised local jet

$$a_{ij}^{\sigma} = \frac{\sigma^{i+j}}{i+j+1}$$

- G_{σ} is the 2d Gaussian function with variance σ^2.
- σ^{i+j} is the scale space normalisation [Lindeberg 98].
- $i + j + 1$ is the number of $(i + j)^{\text{th}}$ order derivatives.

Multiscale Gaussian Derivatives



Local Jet Feature Space: Interest and Justification (1)

Representation Continuum:

The multiscale derivatives form a *continuum* between the local (geometric) and the global (statistical) levels.

Similarity Space:

Neighbouring points in the LJ space represent visually similar points in the image space.

Local Jet Space Metrics

Single scale distance

$$d_f^\sigma(\mathbf{p},\mathbf{q}) = \sum_{i+j\leq r} (a_{ij}^\sigma f_{ij}^\sigma(\mathbf{p}) - a_{ij}^\sigma f_{ij}^\sigma(\mathbf{q}))^2$$

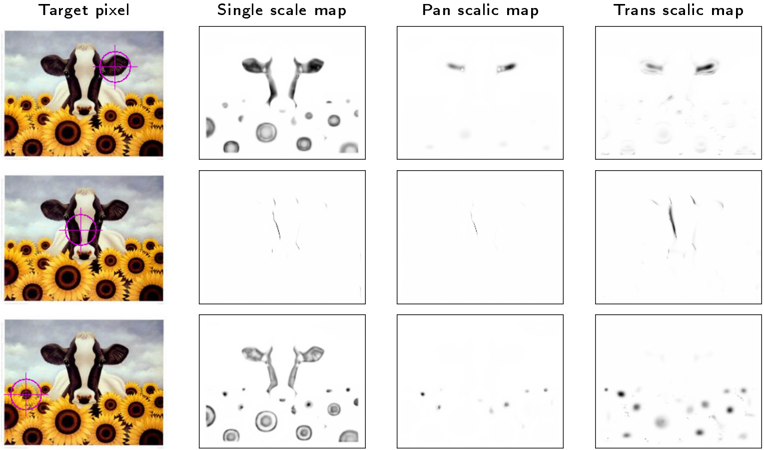
Pan-scalic distance

$$D_f^S(\mathbf{p},\mathbf{q}) = \sum_{i+j\leq r,\sigma\in S} (a_{ij}^\sigma f_{ij}^\sigma(\mathbf{p}) - a_{ij}^\sigma f_{ij}^\sigma(\mathbf{q}))^2$$

Trans-scalic pseudo-distance

$$d_f^S(\mathbf{p},\mathbf{q}) = \min_{(\sigma_1,\sigma_2)\in S^2} \sum_{i+j\leq r} (a_{ij}^{\sigma_1} f_{ij}^{\sigma_1}(\mathbf{p}) - a_{ij}^{\sigma_2} f_{ij}^{\sigma_2}(\mathbf{q}))^2$$

An example of Similarity Maps



Local Jet Feature Space: Interest and Justification (2)

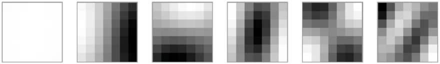
Local Geometry:

The local behaviour of the image may be predicted by its derivatives:

$$f(x_0+\varepsilon, y_0+\eta) = \sum_{k=0}^{r}\sum_{i=0}^{k}\binom{k}{i}\varepsilon^{k-i}\eta^{i}\frac{\partial^k f}{\partial x^{k-i}\partial y^{i}}(x_0,y_0) + o\left((\varepsilon^2+\eta^2)^{r/2}\right)$$

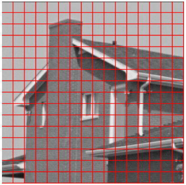
Natural image Statistics:

The first eigen vectors in patch based PCAs ressemble the derivative convolution kernels [Orchard 08]:



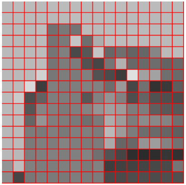
Local Jet Based Representation

The Local jet descriptor corresponds to the local structure outline made by the Taylor expansion:



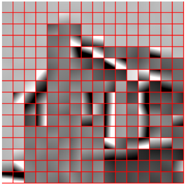
Original:

16 × 16 values per block



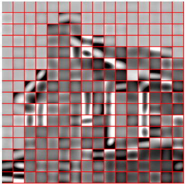
Order 0:

1 value per block



Order 1:

3 values per block

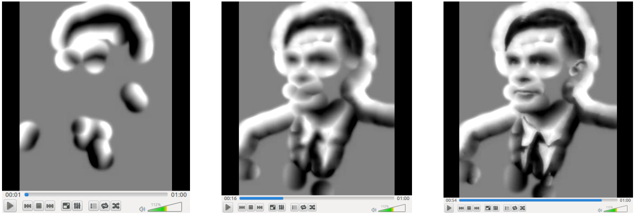


Order 2:

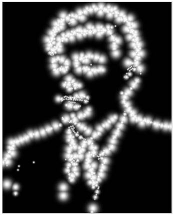
6 values per block

Local Jet Sparse Sampling

The representation may also be limited to a sparse set of points, selected according to their local structure (salience):



Salience (LJ norm) based multiscale Taylor reconstruction.

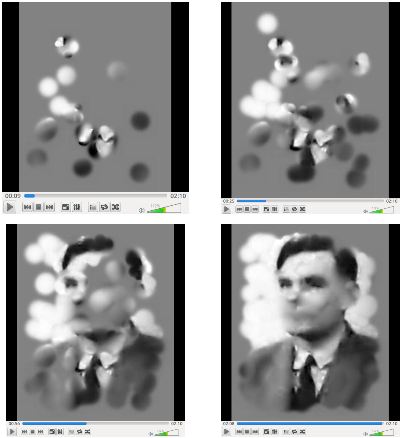


Final weight map

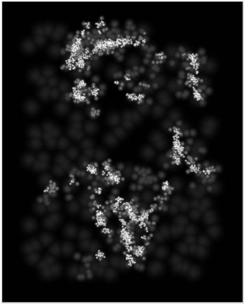
[PhD Antoine Tran]

Active Sensing Representation

The set of representing points may also be acquired iteratively during an active sensing process:



Active scale-recursive Taylor reconstruction.

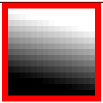
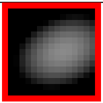
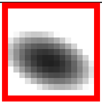
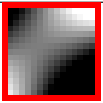
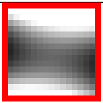


Final weight map

[PhD Antoine Tran]

Local Characterisation

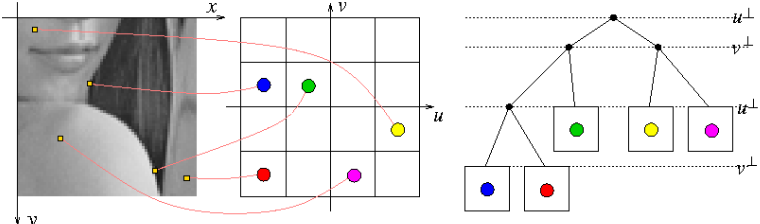
A reduced descriptor may be obtained by categorising the local jet according to the dominant order and polarity. See also [Crosier 10].

0	1	2			
$\|\|\nabla_f\|\| \simeq 0$	$\|\|\nabla_f\|\| \gg 0$	$\|\|H_f\|\|_F \gg 0$			
$\|\|H_f\|\|_F \simeq 0$	$\|\|H_f\|\|_F \simeq 0$	$\Lambda_f \lambda_f > 0$		$\Lambda_f \lambda_f < 0$	
Plateau	Contour	Elliptic curvature		Tubular curvature	
					
		$\Lambda_f < 0$	$\Lambda_f > 0$	$\Lambda_f < 0$	$\Lambda_f > 0$
		$\lambda_f < 0$	$\lambda_f > 0$	$\lambda_f > 0$	$\lambda_f < 0$

2nd order pixel characterisation. $||\nabla_f||$ is the norm of the gradient, $||H_f||_F$, Λ_f and λ_f are the Frobenius norm and the eigen values of the Hessian matrix.

Data representation (basics)

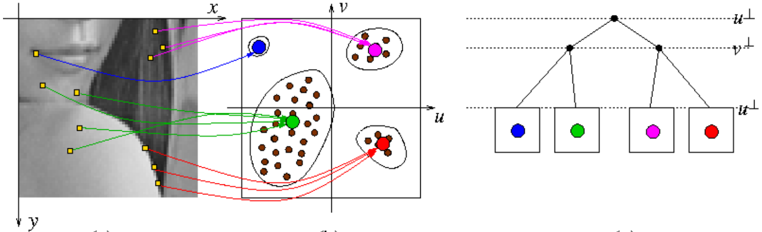
- The pixel data are projected in the feature space.
- The feature vectors may be coded in a binary search tree.
- Every feature vector may be attached to a pixel index.



(a) (b) (c)

Data representation (with clustering)

- The feature space can be quantised through clustering.
- Every code word may keep record of a set of pixel indices.



(a) (b) (c)

Presentation Outline

Introduction

Local Jet Feature Space

Image Processing
- NL-Means Filtering
- Optical Flow
- Lines and Circles Detection

Object Representations
- Basic representations
- Background Modelling
- Kernel Based Tracking
- Dense Implicit Shape Models

Conclusions

NL-means in the feature space

- The NL-means denoising filter [Buades 05] calculates a weighted average of every pixel, with the weights defined as a function of local similarity.
- Here we simply use the distance in the feature space:
$$\omega(u,v) = e^{-\frac{||u-v||^2}{h^2}}, \text{ with:}$$
$||.||$ a norm in the feature space, and h a decay parameter.
- Then we perform the average on a neighbourhood of $\mathbf{p}$ in the image space (Limited Range)...
- ...or on a neighbourhood of $\hat{\mathbf{p}}$ in the feature space (Unlimited range).

Limited Range LJ-NL-Means

The weights (in the LJ space) are calculated in a limited neighbourhood of $\mathbf{p}$ *in the image space*:

Limited range NL-means

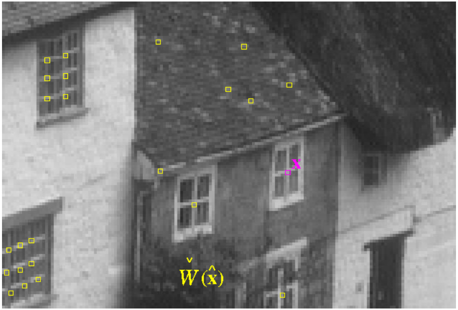
$$f_{LR}^{NL}(\mathbf{p}) = \frac{1}{\zeta(\mathbf{p})} \sum_{\mathbf{q} \in \mathcal{N}(\mathbf{p})} f(\mathbf{q})\omega(\hat{\mathbf{p}}, \hat{\mathbf{q}})$$

Unlimited Range LJ-NL-Means

The weights (in the LJ space) are calculated in a limited neighbourhood of $\hat{\mathbf{p}}$ *in the LJ space*:

Unlimited range NL-means

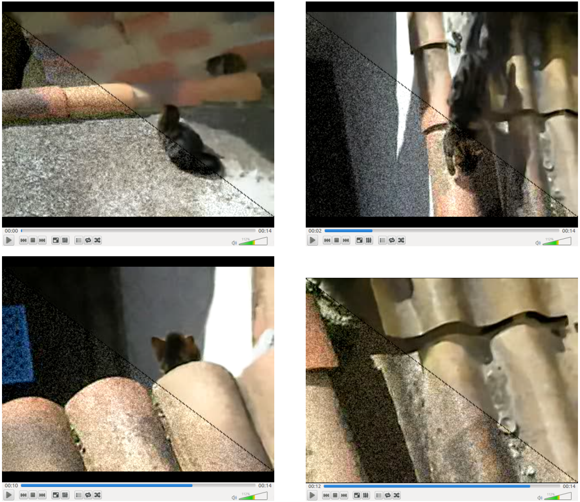
$$f_{UR}^{NL}(\mathbf{p}) = \frac{1}{\xi(\mathbf{p})} \sum_{u \in \mathcal{W}(\hat{\mathbf{p}})} f(\check{\mathbf{u}})\omega(\hat{\mathbf{p}}, \mathbf{u})$$



Local jet based NL-Means Video Denoising

Example:
Space-time Limited Range Local Jet based NL-Means filtering (colour, one single scale).

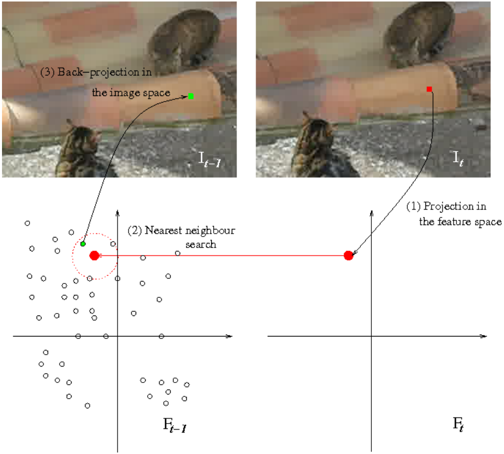
Pros and Cons:
+ Continuum anisotropic filtering $\rightarrow$ NL-means.
- Unlimited range remains costly.



Apparent motion in the feature space

$$u(f_{t-1}, f_t, \mathbf{p}) = \arg \min_{\mathbf{v} \in \mathcal{F}_{f_{t-1}}} d^F(\hat{\mathbf{p}}_{f_t}, \mathbf{v})$$

In our framework the optical flow estimation is simply expressed by a nearest neighbour search in the feature space:

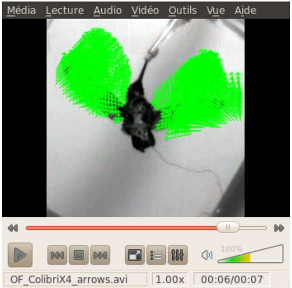


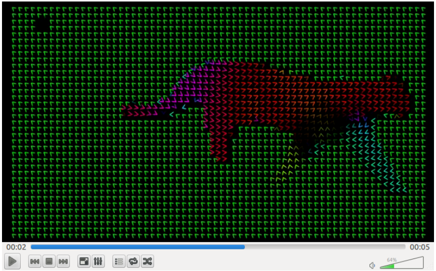
Infinite range optical flow

Pros and Cons:

+ Conceptual simplicity
+ Implicit spatial regularisation
+ Infinite Range motion

- Computational complexity
- Spatial accuracy

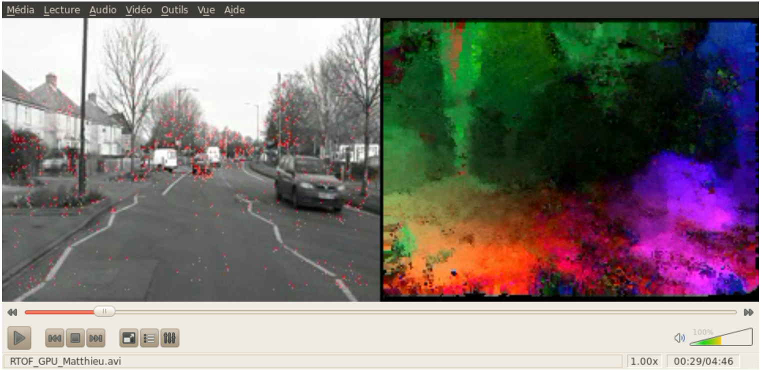




Biological Motion Quantification based on dense optical flow [PhD Fabio Martínez].

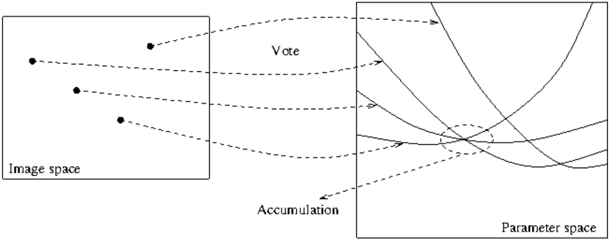
Real-Time Local Jet based Optical Flow on GPU

Real-Time dense optical flow without explicit spatial regularisation is obtained by implementing the limited range version on GPU [PhD Matthieu Garrigues]:



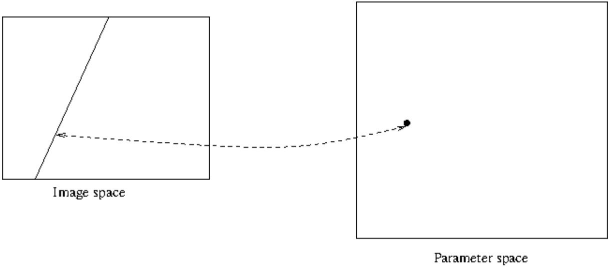
Hough Transform: global view

- One of the oldest applications of Computer Vision (End 50's, bubble chamber images)
- Adapted to both analytical (curves) or non analytical (objects) shapes
- Based on accumulation (vote) mechanism from image space (pixels) to multidimensional parameter space



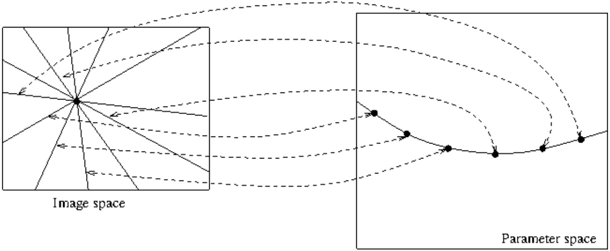
Hough Transform: details (1)

Every point of the parameter space corresponds to one shape in the image space.
Example : One (θ, ρ) polar coordinates point correspond to one line.



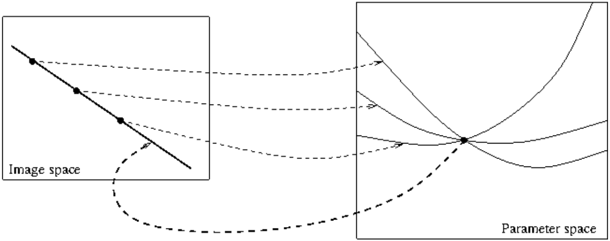
Hough Transform: details (2)

Every single curve of the parameter space corresponds to one point, or equivalently to one beam of shapes in the image space.
Example: One sine curve corresponds to one beam of line, i.e. one point.



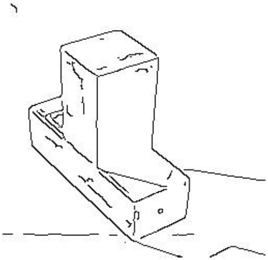
Hough Transform: details (3)

Reciprocally, different points from the same shape in the image space form a beam of curves in the parameter space, converging to one point defining the right shape.



Hough Transform: practice

So classically, the Hough transform (i.e. the result of the projection of all image points in the parameter space) is calculated from a limited set of points: the contours.
The best candidat shapes are then detected by computing the local maxima of the Hough transform.



Contour image

Classical Hough transform: different accumulation points are visible

Partial derivatives and 1-to-1 Hough transforms

Classical approaches

- sparse: Only a few points (contours, key points) are voting.
- 1-to-many: Every point from the image space is voting uniformly on a n dimensional surface in the parameter space.
- many-to-1: (AKA Probabilistic) Every n-tuple of points from the image space is voting for one unique point in the parameter space.

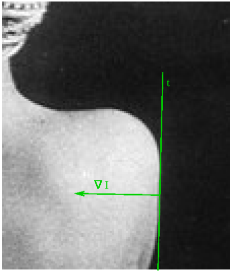
Partial derivatives and 1-to-1 Hough transforms

Hough transfoms based on partial derivatives

- dense: All the points are voting...
- inegalitarian: ...but their votes don't have the same weight!
- 1-to-1: Every point from the image space is voting for one unique point in the parameter space.

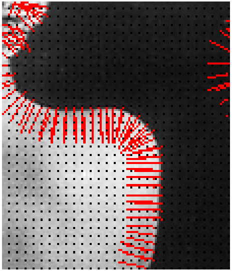
1-to-1 transform: order 1

At order 1, the gradient defines the isophote direction, and then the direction of the candidate line. The weight of the vote is the norm of the gradient.



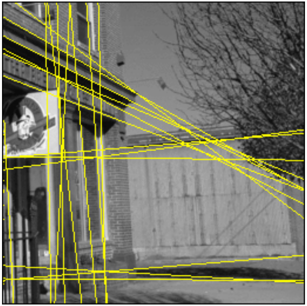
Gradient and line

Weight of the vote

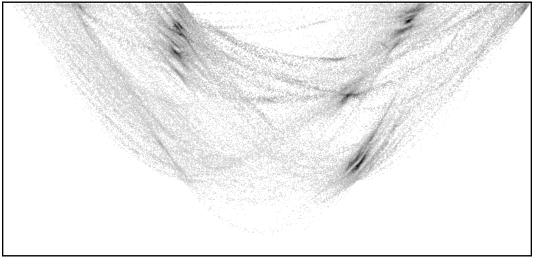


Main votes

1-to-1 transform: order 1



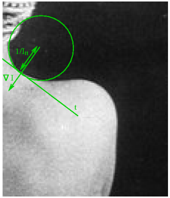
20 best lines



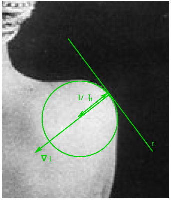
(ρ, θ) 1-to-1 transform

1-to-1 transform: order 2

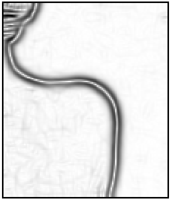
At order 2, the gradient direction and the isophote curvature define the radius and the centre of the osculating circle to the isophote curve, and then the equation of the candidate circle. The weight of the vote is the Frobenius norm of the Hessian matrix.



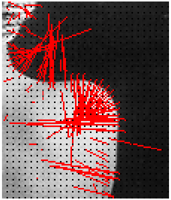
Positive curvature



Negative curvature



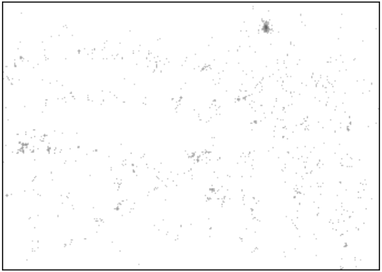
Weight of the vote



Main votes

1-to-1 transform: order 2

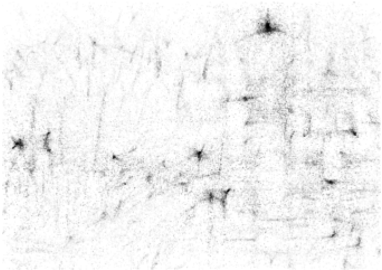
10 best circles



(ρ, x, y) 1-to-1 transform (level $\rho = 19$)

1-to-1 transform: order 2 in two passes

10 best circles



First pass (x, y) 1-to-1 transform (centre votes)

Presentation Outline

Introduction

Local Jet Feature Space

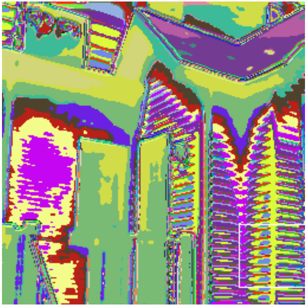
Image Processing
- NL-Means Filtering
- Optical Flow
- Lines and Circles Detection

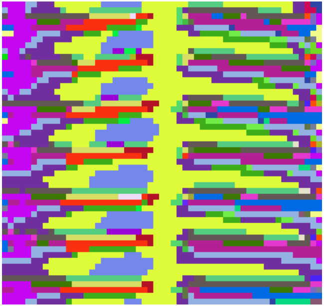
Object Representations
- Basic representations
- Background Modelling
- Kernel Based Tracking
- Dense Implicit Shape Models

Conclusions

Codebook distribution

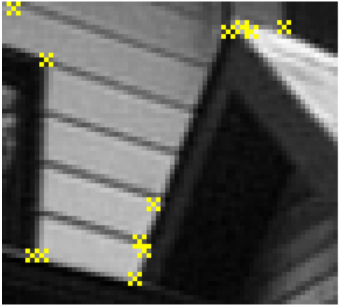
The distribution of words in the codebook provides a possible representation for an image or any visual category. See for example [Rubner 99].





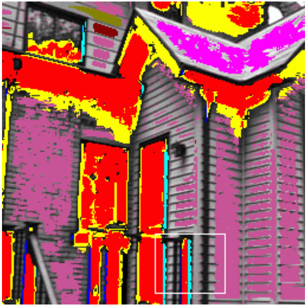
Singularities of the feature space

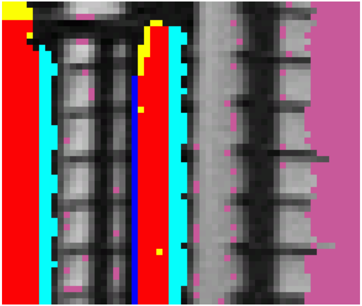
Finding the isolated points in the feature space is a way to fuse the detection of salient point and the calculation of attached descriptors. See also [Kervrann 08].



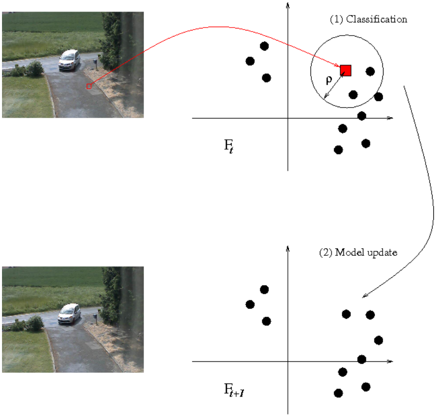
Modes of the feature space

The modes (clusters) in the feature space (see [Burman 09]), back-projected in the image space, correspond to large homogeneous coulours, long straight edges or regular texture elements.





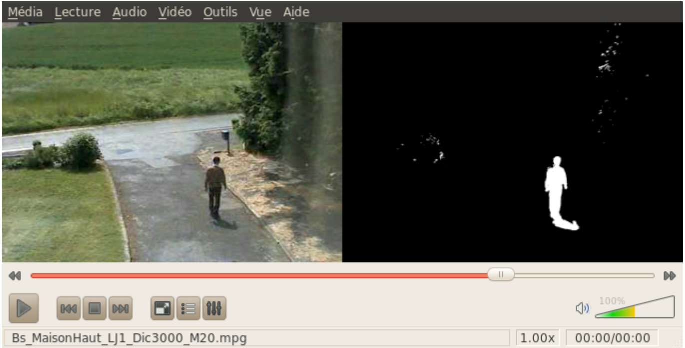
Sampling and Consensus in the feature space



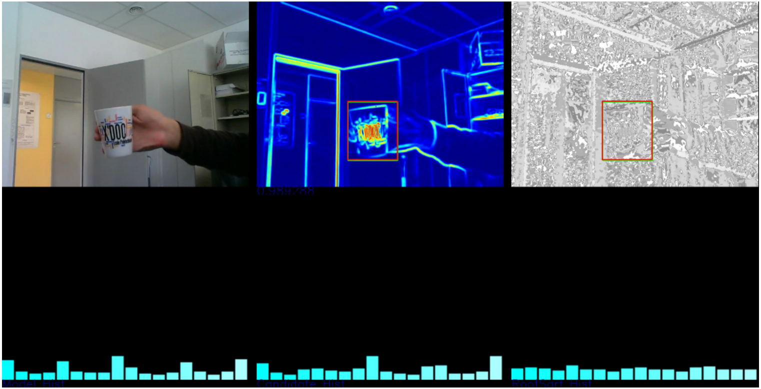
- We model the static Background by sampling the values of every pixel in the local jet space.
- The non static Foreground is classified according to a consensus vote in the local jet space.
- The Background model is updated accordingly.

Foreground classification results

- Feature space: Local jet of order 2, single scale, 3 colours.
- Vector quantization: 3,000 words code book.
- Number of samples: $M = 20.0$.



Local Jet based Mean-shift Object Tracking

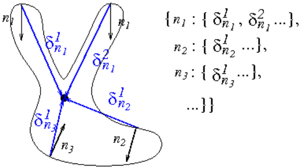


Ongoing work on LJ based object tracking [PhD Antoine Tran].

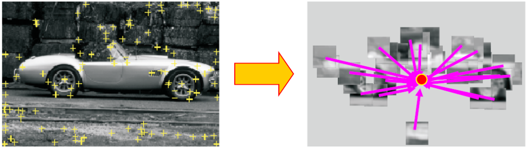
Object representation by R-Tables

The classical generalised Hough transforms are *sparse*: they are calculated from a reduced set of feature points: contour [Ballard 81], or salient points [Leibe 04].

$$\text{R-Table}: \{i, \{\vec{\delta_i^j}\}_j\}_i$$



Contour

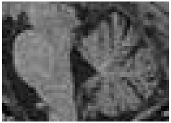


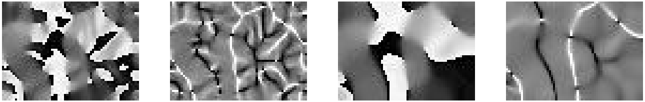
Salient points

Dense R-Tables indexed by derivatives

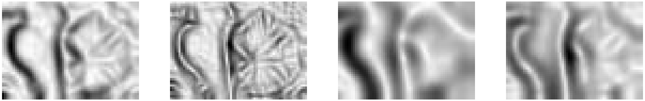
In the *dense* approach, the indices i of the R-table are the quantised mutiscale derivatives, available everywhere.

The R-Table is weighted as: $\{i, \{\vec{\delta}_i^j, \omega_i^j\}_j\}_i$





$(o = 1; \sigma = 2,0)$ $(o = 2; \sigma = 2,0)$ $(o = 1; \sigma = 4,0)$ $(o = 2; \sigma = 4,0)$



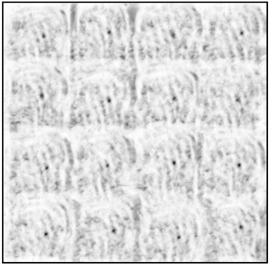
Generalised Hough Transform: Object Detection

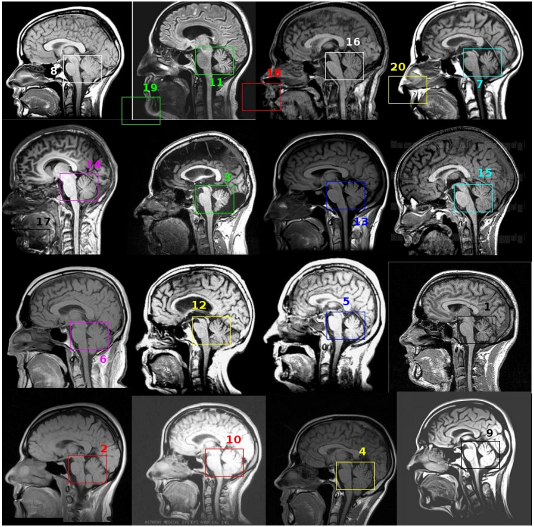
Initial: $H(\mathbf{x}) = 0$ everywhere.
Forall $\mathbf{x}$ in image,
let $\lambda(\mathbf{x})$ the quantised derivative.
    Forall occurrence j
    of the R-Table associated to
    $\lambda(\mathbf{x})$, do:
        $H(\mathbf{x} + \delta^j_{\lambda(\mathbf{x})}) \mathrel{+}= \omega^j_{\lambda(\mathbf{x})}$

The best candidate objects are then localised on the maxima of H.

Generalised Hough Transform: Object Detection

Hough Transform (left, reduced 50%), and MRI image mosaic with the 20 best cerebellum candidates (right).





Presentation Outline

- Introduction
- Local Jet Feature Space
- Image Processing
  - NL-Means Filtering
  - Optical Flow
  - Lines and Circles Detection
- Object Representations
  - Basic representations
  - Background Modelling
  - Kernel Based Tracking
  - Dense Implicit Shape Models
- Conclusions

Conclusions: Contribution Outline

We proposed a unified framework based on local jet feature for image representation and processing.

- LJ-NL-Means: Fast, Missing link between True NL-Means and Bilateral Filtering, May be combined with other optimisations.
- LJ-NN-Optical flow: Simple, Dense and Smooth without explicit Regularisation.
- LJ singularities and cluster: Unify Detection and Description.
- Dense 1-to-1 Hough transforms: Fast, Generic, Adaptive level of Density/Sparsity.

Conclusions: Prospective Works

- Optimise and Evaluate LJ based Background and Object Modelling.
- Optimise and Evaluate Dense LJ based General Hough Transforms.
- Improve and Unify LJ Quantisation and Clustering.
- Combine Sparse and Active LJ representations with Object Modelling.

Bibliography

[Manzanera 10] A. MANZANERA
Local Jet Based Similarity for NL-Means Filtering
Int. Conf. on Pattern Recognition 2668-2671 (2010)

[Manzanera 11] A. MANZANERA
Local Jet Feature Space Framework for Image Processing and Representation
Int. Conf. on Signal Image Technology and Internet Based Systems 261-268 (2011)

[Manzanera 12] A. MANZANERA
Dense Hough transforms on gray level images using multi-scale derivatives (invited paper)
Int. Work. on Medical and Healthcare Applications (AMINA'12) 55-62 (2012)

References (Concept and models)

- **[Peyré 09]** G. PEYRE
  Manifold Models for Signals and Images
  Computer Vision and Image Understanding 113(2), 249-260. (2009)
- **[Lindeberg 98]** T. LINDEBERG
  Feature detection with automatic scale selection
  International Journal of Computer Vision 30(2), 77-116. (1998)
- **[Koenderink 87]** J.J. KOENDERINK and A.J. VAN DOORN
  Representation of Local Geometry in the Visual System
  Biological Cybernetics 55, 367-375. (1987)

References (Filter Banks and Codebooks)

[Freeman 91] W.T. FREEMAN and E.H. ADELSON
The design and use of Steerable Filters
IEEE Trans. on Pattern Analysis and Machine Intelligence 13(9), 891-906. (1991)

[Crosier 10] M. CROSIER and L.D. GRIFFIN
Using Basic Image Features for Texture Classification
International Journal of Computer Vision 88(3), 447-460. (2010)

[Rubner 99] Y. RUBNER and C. TOMASI
Texture-Based Image Retrieval Without Segmentation
IEEE International Conference on Computer Vision, Kerkyra, Greece 1018-1024. (1999)

References (Tools)

- **[Mount 97]** D.M. MOUNT and S. ARYA
  ANN: A Library for Approximate Nearest Neighbor Searching
  CGC Workshop on Computational Geometry (1997)
  http://www.cs.umd.edu/~mount/ANN/

- **[Burman 09]** P. BURMAN and W. POLONIK
  Multivariate mode hunting: Data analytic tools with measures of significance
  Journal of Multivariate Analysis 100(6), 1198-1218. (2009)

- **[Van Vliet 98]** L.J. VAN VLIET, I.T. YOUNG and P.W. VERBEEK
  Recursive Gaussian derivative filters
  Proc. Int. Conf. on Pattern Recognition vol. 1, 509-514. (1998)

References (NL-Means)

- **[Buades 05]** A. BUADES, B. COLL and J.M. MOREL
  A non-local algorithm for image denoising
  Proc. IEEE Conf. on Computer Vision and Pattern Recognition vol. 2, 60-65. (2005)
- **[Orchard 08]** J. ORCHARD, M. EBRAHIMI and A. WONG
  Efficient Non-Local Means Denoising using the SVD
  Proc. Int. Conf. on Image Processing 1732-1735. (2008)
- **[Kervrann 08]** C. KERVRANN and J. BOULANGER
  Local adaptivity to variable smoothness for exemplar-based image denoising and representation
  International Journal of Computer Vision 79(1), 45-69. (2008)

References (Background and object modelling)

[Barnich 09] O. BARNICH and M. VAN DROGENBROECK
ViBe: a powerful random technique to estimate the background in video sequences
International Conference on Acoustics, Speech, and Signal Processing 945-948. (2009)

[Wang 07] H. WANG and D. SUTER
A consensus-based method for tracking: Modelling background scenario and foreground appearance
Pattern Recognition 40(3), 1091-1105. (2007)

[Comaniciu 03] D. COMANICIU, V. RAMESH and P. MEER
Kernel-Based Object Tracking
IEEE Trans. on Pattern Analysis and Machine Intelligence 25(5), 564-575 (2005)

References (Hough Transforms)

[O'Gorman 76] F. O'GORMAN AND B. CLOWES
Finding picture edges through collinearity of feature points
IEEE Trans. on Computers C-25 449-456 (1976)

[Leibe 04] B. LEIBE, A. LEONARDIS and B. SCHIELE
Combined object categorization and segmentation with an implicit shape model
ECCV Workshop on Statistical Learning in Computer Vision (2004)

[Valenti 08] R. VALENTI and T. GEVERS
Accurate eye center location and tracking using isophote curvature
Int. Conf. on Computer Vision and Pattern Recognition (2008)