Performance evaluation of registration (and segmentation?) algorithms in the absence of Gold Standard

X. Pennec

EPIDAURE Project
2004, route des Lucioles B.P. 93
06902 Sophia Antipolis Cedex (France)

Overview

➨ **Registration performances**
- Types of errors
- Quantification of errors...
- …with no gold standard

❍ **Performances estimation**

❍ **Error prediction**

❍ **Segmentation**

❍ **Conclusion**

Classification of registration algorithms

Registration = finding correspondences between homologous points (duality matches / transformation)

Feature space

- 0D: points, landmarks, frames
- 1D: curves
- 2D: surfaces
- 3D: volumes (i.e. intensity-based methods)

Transformation space

- Rigid, affine, locally affine, deformable
- Dimensionality reduction (e.g. 3D/2D)

Similarity metric (criterion), optimization scheme

Variability of a registration algorithm

Fixed internal parameters

- Multiscale resolution
- (Typical variance…)

External parameters

- Data (image) 1
- Data (image) 2
- Acquisition noise
- Patient effects

Varying internal parameters

- Initial transformation
- (…)

Registration algorithm

Final transformation

Robustness: ability to find the right transformation (success/failure)

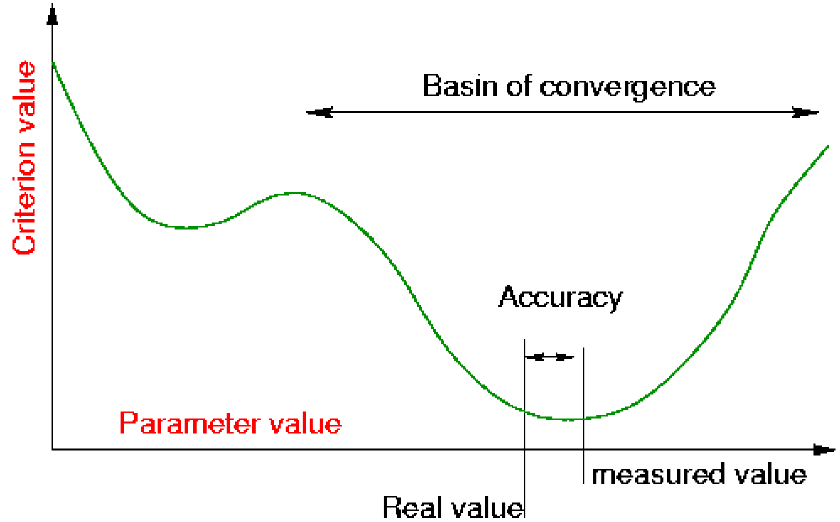
Repeatability: Variability w.r.t. varying internal parameters

Accuracy: Variability w.r.t. the ground truth for typical data

Types of errors for an energy minimization

Robustness

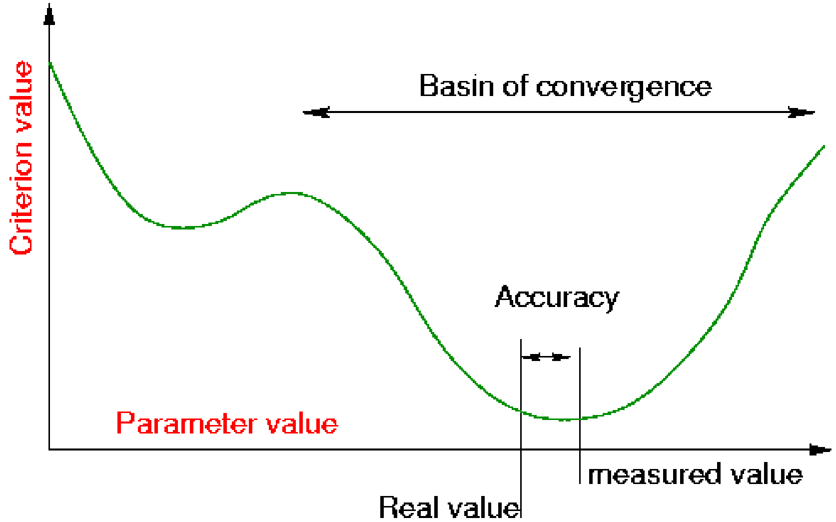
- Local minima at a global scale



Uncertainty = deviation from the real transformation

- Bias (features, method, adequacy of the criterion)
- Accuracy
  - Extrinsinc (sensitivity to the noise on the features)
  - Intrinsic or precision (optimization, interpolation, local minima)

Quantifying the registration errors



Robustness:

- size of the basin of attraction
- Probability of convergence

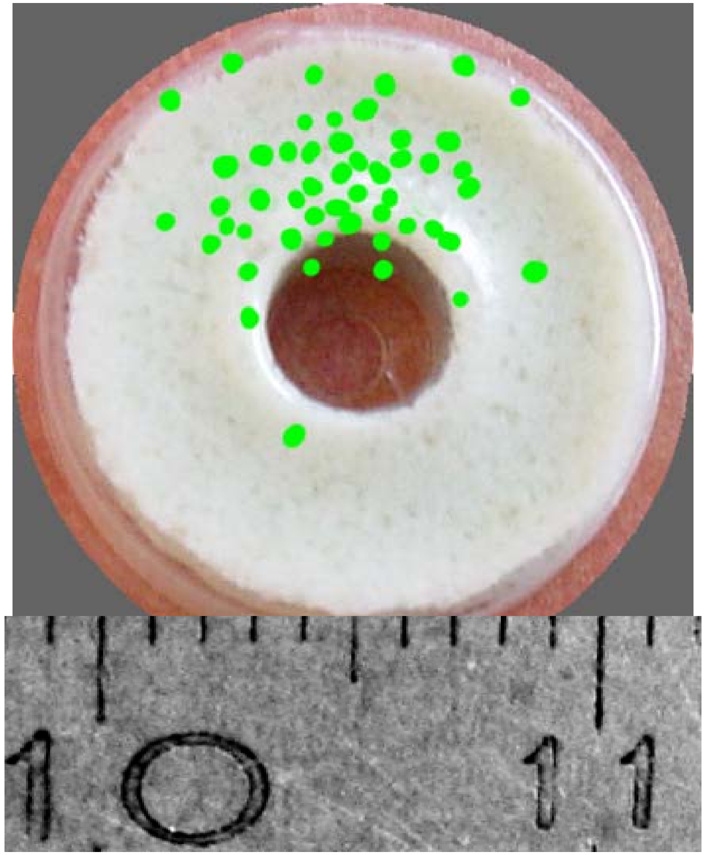
Uncertainty = deviation from the real transformation

- Maximum error: bound
- Mean Error: covariance matrix, std dev.
  - On the transformation (rotation σ_r [rad], translation σ_t [mm])
  - On test points (TRE σ_x)

Targeting using Augmented reality

User 1 (50 trials):

- Repeatability:
  σ = 2.2 mm
- Bias: 3.0 mm
- Accuracy:
  σ = 3.7 mm

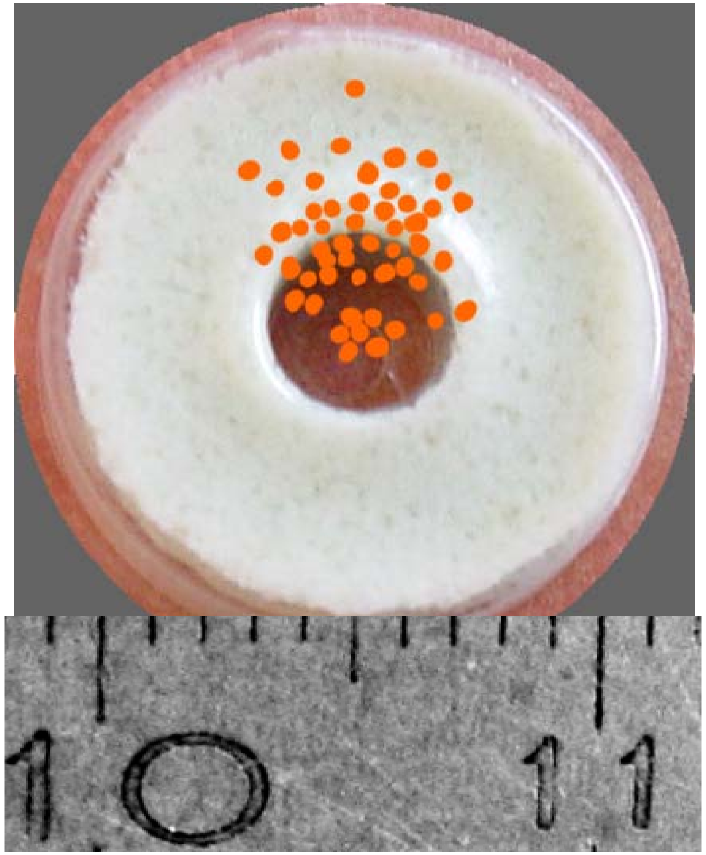


[S. Nicolau, A. Garcia et al., Aug. & Virtual Reality Workshop, Geneva, 2003]

Targeting using Augmented reality

User 2 (50 trials):

- Repeatability:
  $\sigma = 1.9$ mm
- Bias: 1.3 mm
- Accuracy:
  $\sigma = 2.3$ mm

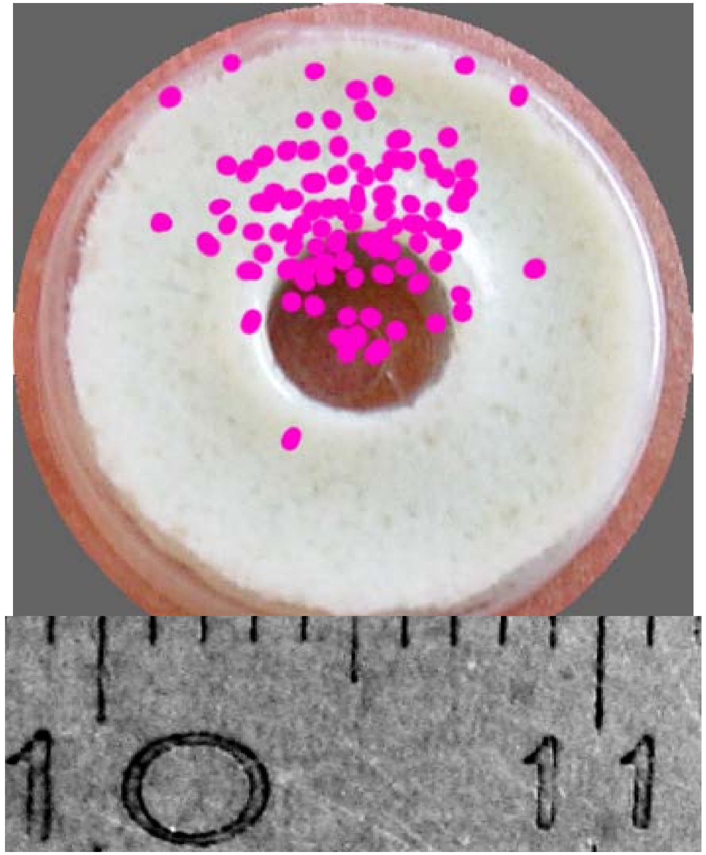


[S. Nicolau, A. Garcia et al., Aug. & Virtual Reality Workshop, Geneva, 2003]

Targeting using Augmented reality

Both users (100 trials):

- Repeatability:
  $\sigma = 2.2$ mm
- Bias: 1.7 mm
- Accuracy:
  $\sigma = 2.8$ mm



[S. Nicolau, A. Garcia et al., Aug. & Virtual Reality Workshop, Geneva, 2003]

Performance evaluation and validation

Synthetic data (simulation):

- Available ground truth
- Difficult to identify and model all sources of variability

Real data in a controlled environment (Phantom):

- Possible gold standard
- Performances evaluation in specific conditions
  - Difficult to test all clinical conditions
  - May hide a bias

Image database representative of the clinical application

- Usually no ground truth
- Should span all sources of variability

Performance evaluation without Gold Standard

Registration or consistency loops

- Pennec et al. IJCV 25(3) 1997 & MICCAI 1998.
- Holden et al. TMI 19(2), 2000
- Roche et al MICCAI 2000 & TMI 20(10), 2001.

Cross-comparison of criterions

- Hellier et al MICCAI 2001 & TMI 22(9), 2003.

Ground truth as a hidden variable (EM like algorithms)

- Granger, MICCAI 2001 & ECCV 2002,
- Warfield, MICCAI 2002, [Staple, segmentation]
- Nicolau, IS4TM 2003

Error prediction

- Pennec et al. ICCV 1995, IJCV 25(3) 1997 & MICCAI 1998.
- Fitzpatrick et al, MedIm 1998, TMI 17(5), 1999.
- Nicolau et al, INRIA Research Report 4993, 2003

Overview

- ✔ **Registration performances**
- ➪ **Performances estimation**
  - ➪ MR/US intensity-based registration
  - ❍ Surface-based registration
- ❍ **Error prediction**
- ❍ **Segmentation**
- ❍ **Conclusion**

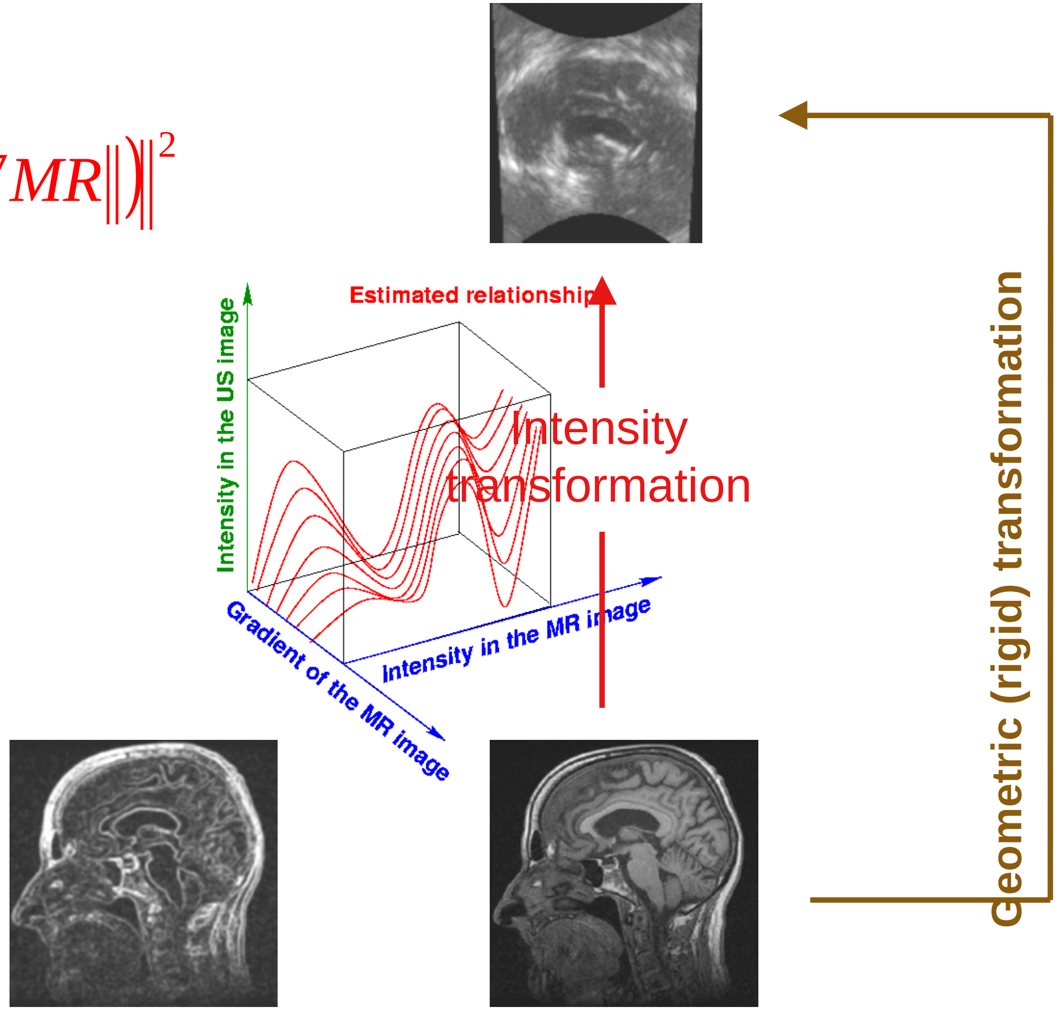
MR / US registration

➨ *Bivariate correlation ratio*

$$C(T,f) = \left\| T(US) - f\left(MR, \|\nabla MR\|\right)\right\|^2$$

Alternated search

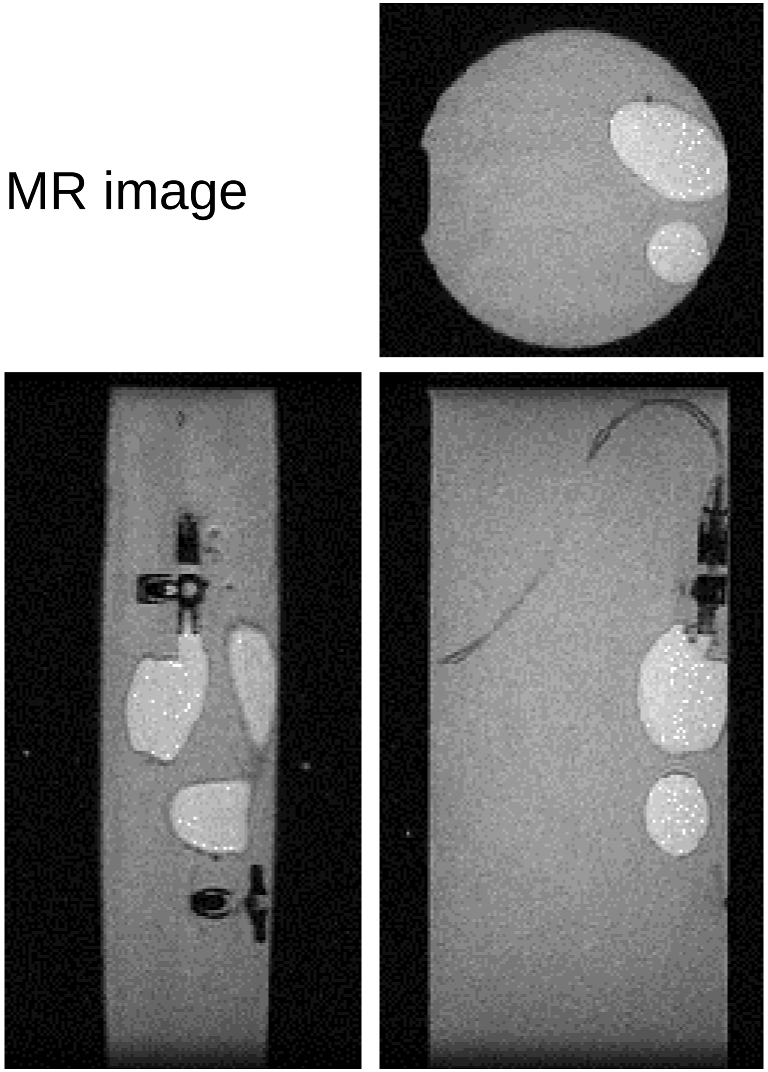
- function f
- transformation T



[A. Roche, X. Pennec et al., MICCAI 2000, TMI 20 (10) 2001.]

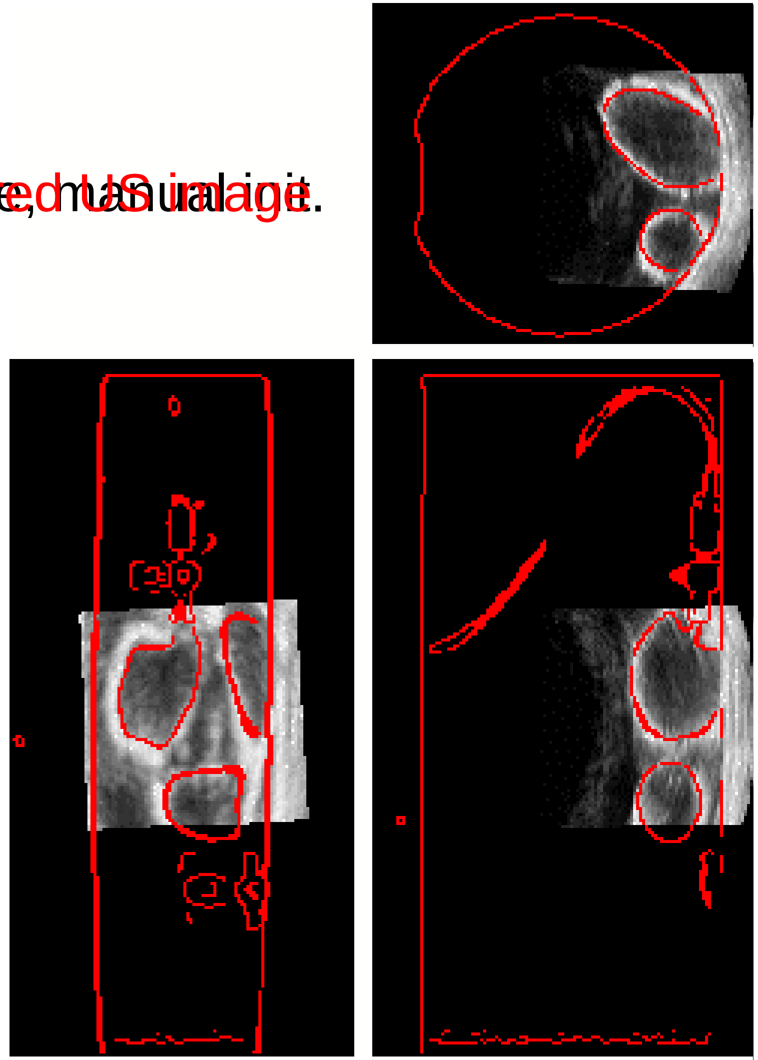
Results on a phantom

MR image



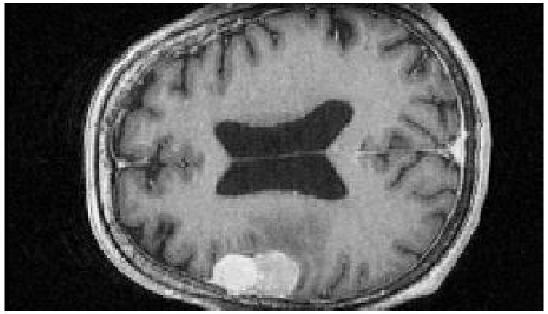
US image, manual init.

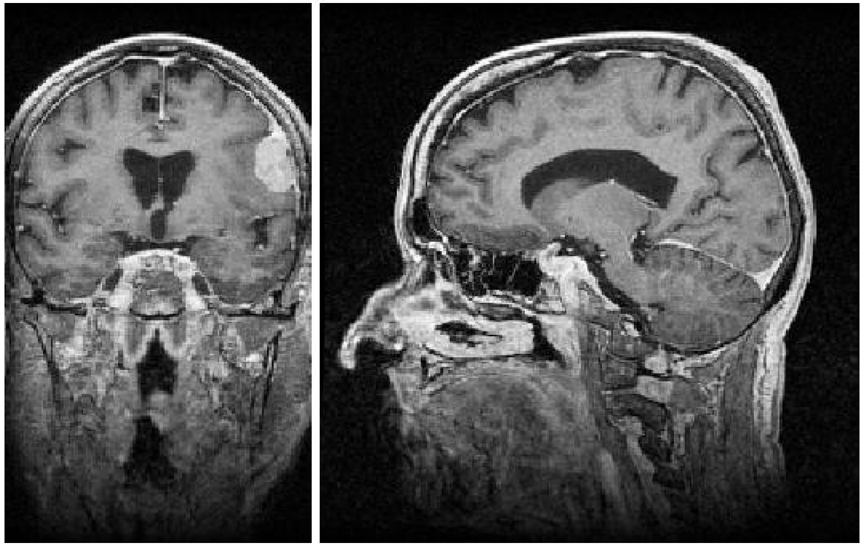
Registered US image



Results on per-operative patient images

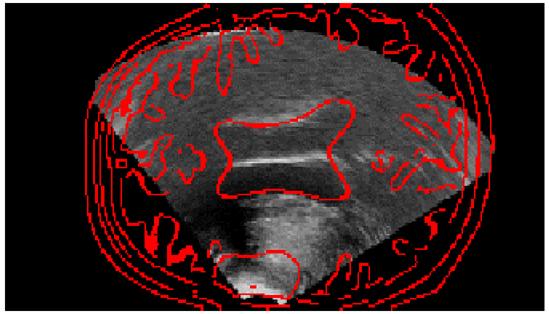
MR Image

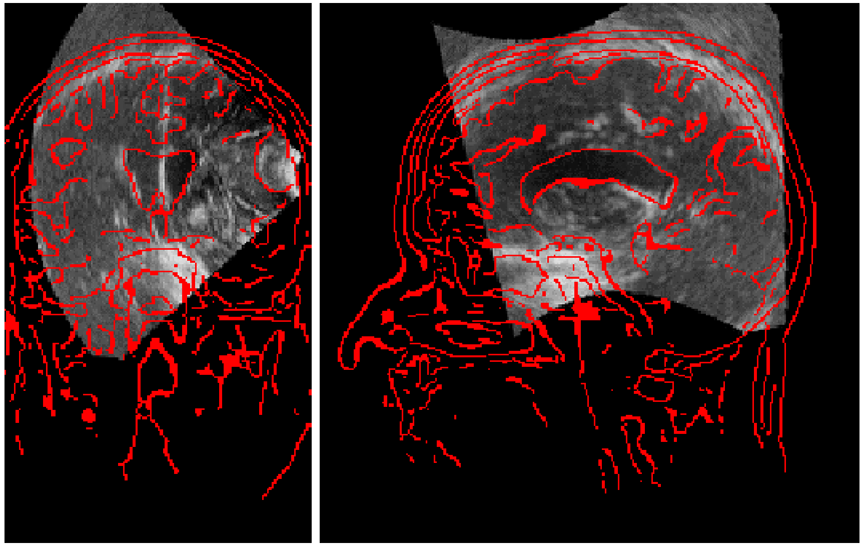




Registered US

Initial US





Evaluation of MR / US registration

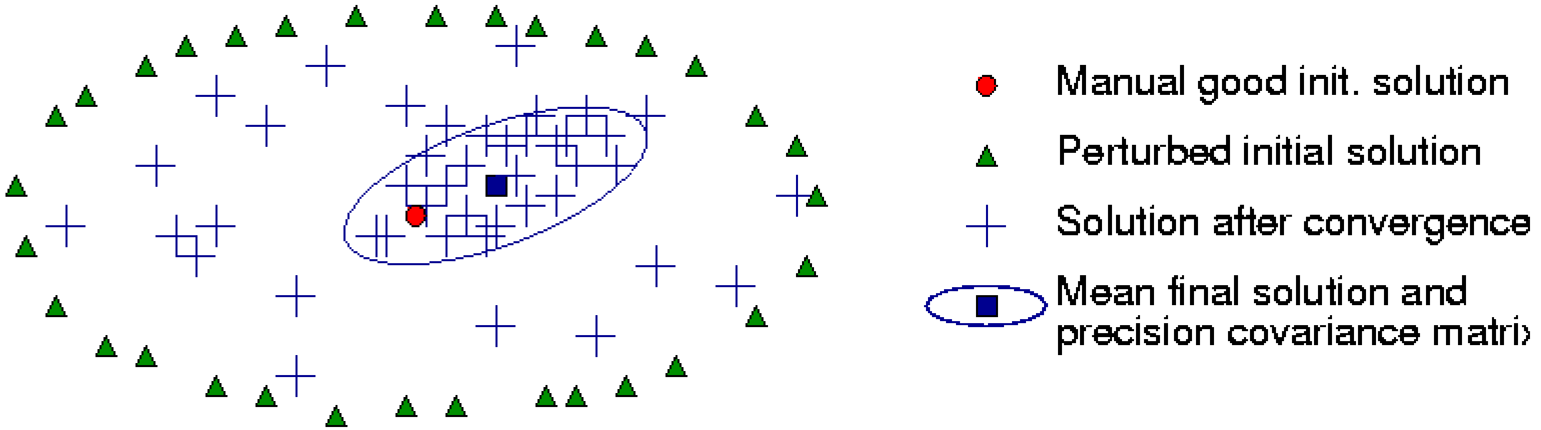
Robustness w.r.t. initial transformation

- Manual initial transformation is easy up to 20mm and 15 deg.
- What is the probability to converge to the right transformation?
  - Determine the right transformation
  - What is inlier (CV to the right transfo) and outlier ? Determine repeatability

Consistency (accuracy?) of the registration

- Registration loops versus comparison of transformations
- Toward a “bronze standard” registration

Robustness and repeatability



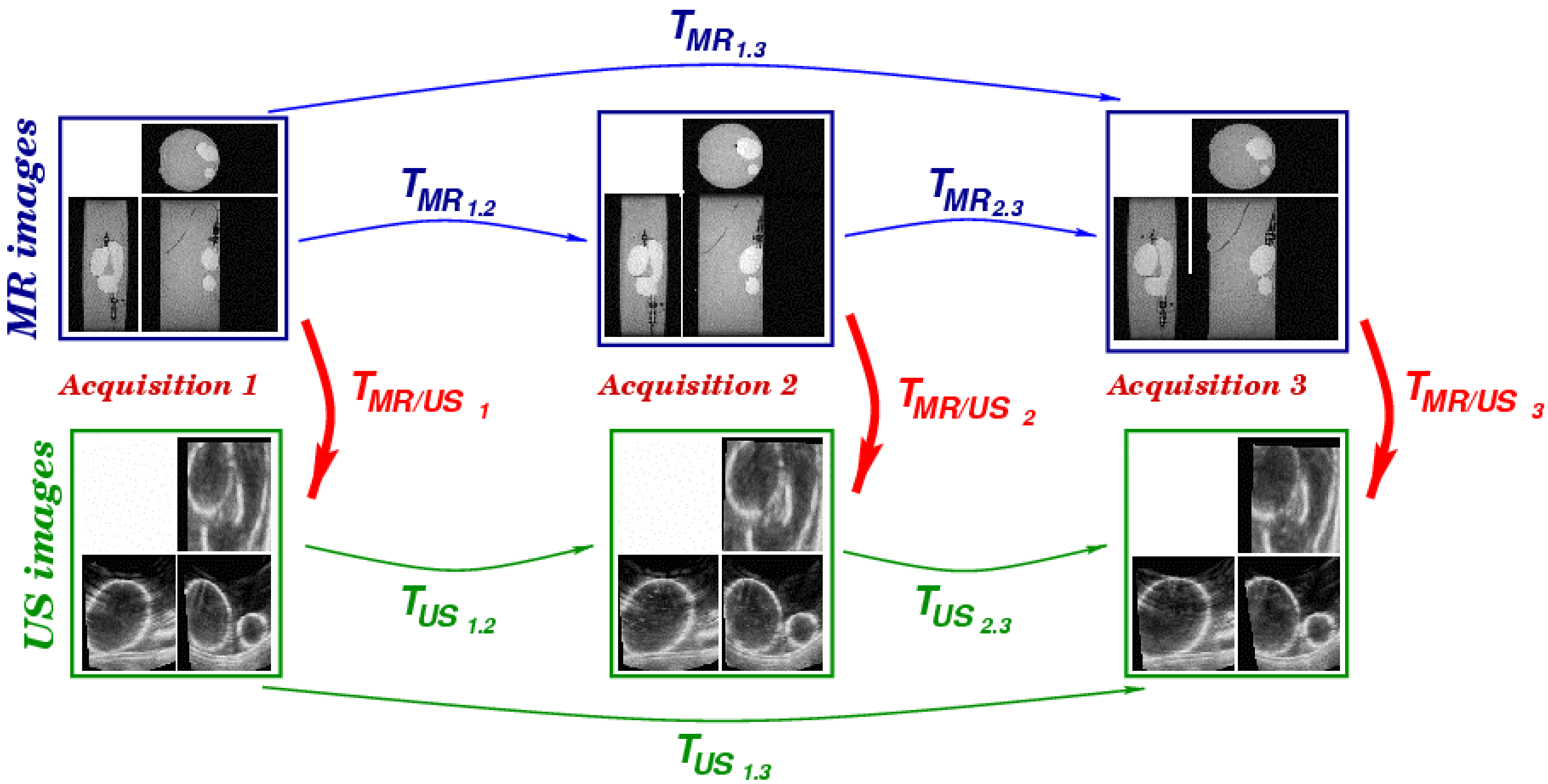
- Find the mean of good results and its variability

$$\boxed{\text{Success rate,} \quad \sigma_{rot}, \sigma_{trans}}$$

- Robust Fréchet mean $\bar{T} = \arg\min_{T}\left(\sum_i \min\left(\mu^2(T_i, T), \chi^2\right)\right)$

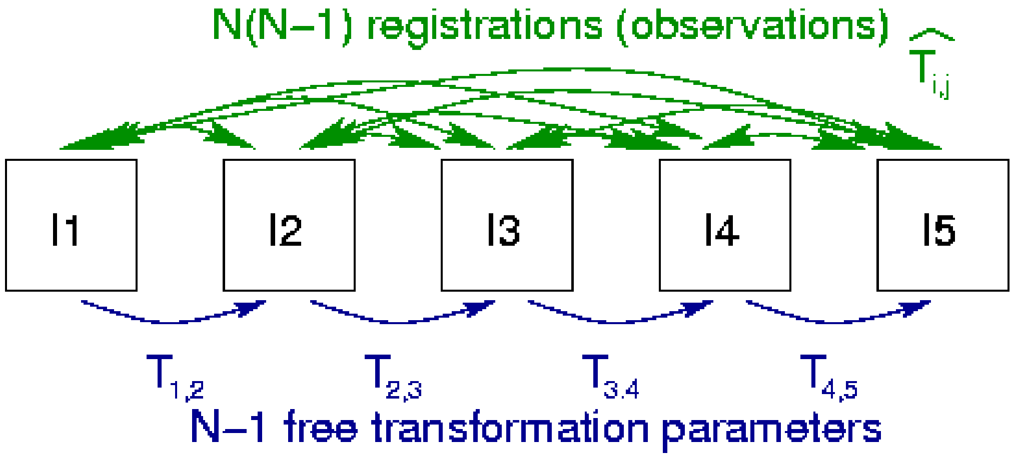
- ML estimation of a mixture Gaussienne + uniforme (EM)

Accuracy Evaluation (Consistency)



$$\sigma_{loop}^2 = 2\sigma_{MR/US}^2 + \sigma_{MR}^2 + \sigma_{US}^2$$

Multiple a posteriori registration



Best explanation of the observations (ML) :

- Robust Fréchet mean $$d^2(T_1, T_2) = \min\left(\mu^2(T_i, T), \chi^2\right)$$
- Robust initialisation and Newton gradient descent

Result $$\boxed{T_{i,j}, \sigma_{rot}, \sigma_{trans}}$$

Results on the phantom dataset

Data (varying balloons volumes)

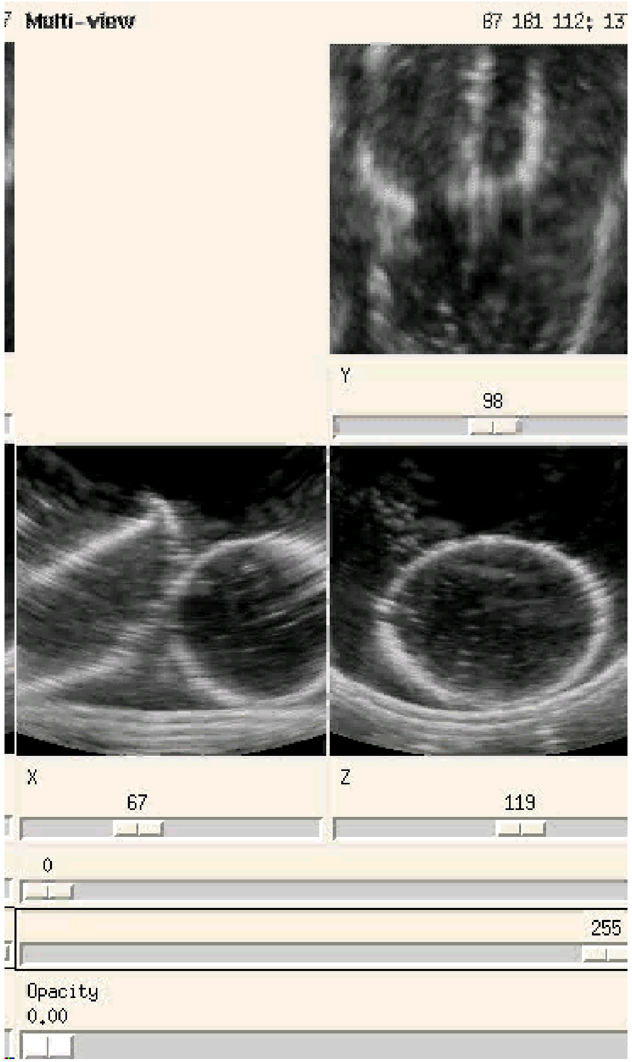
- 8 MR (0.9 x 0.9 x 1.0 mm)
- 8 US (0.4 x 0.4 x 0.4 mm)
- 54 loops

Robustness and repeatability

	Success	var rot (deg)	var trans (mm)
MI	39%	0.40	0.27
CR	52%	0.43	0.25
BCR	**76%**	**0.14**	**0.09**

Consistency of BCR

	var rot (deg)	var trans (mm)	var test (mm)
Multiple MR	0.06	0.1	0.13
Multiple US	0.60	0.4	0.71
Loop	1.62	1.43	2.07
MR/US	**1.06**	**0.97**	**1.37**



Results on per-operative patient images

Data (per-operative US)

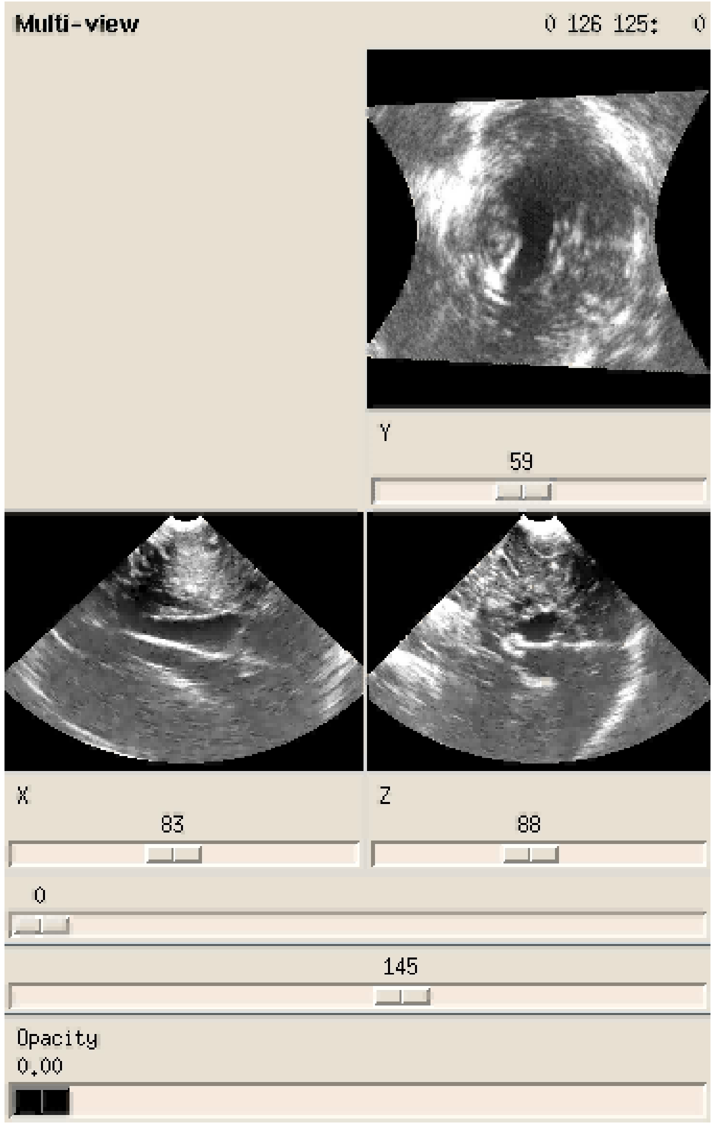
- 2 pre-op MR (0.9 x 0.9 x 1.1 mm)
- 3 per-op US (0.63 and 0.95 mm)
- 3 loops

Robustness and precision

	Success	var rot (deg)	var trans (mm)
MI	29%	0.53	0.25
CR	90%	0.45	0.17
BCR	**85%**	**0.39**	**0.11**

Consistency of BCR

	var rot (deg)	var trans (mm)	var test (mm)
Multiple MR	0.06	0.06	0.10
Loop	2.22	0.82	2.33
MR/US	**1.57**	**0.58**	**1.65**



Overview

- ✔ **Registration performances**

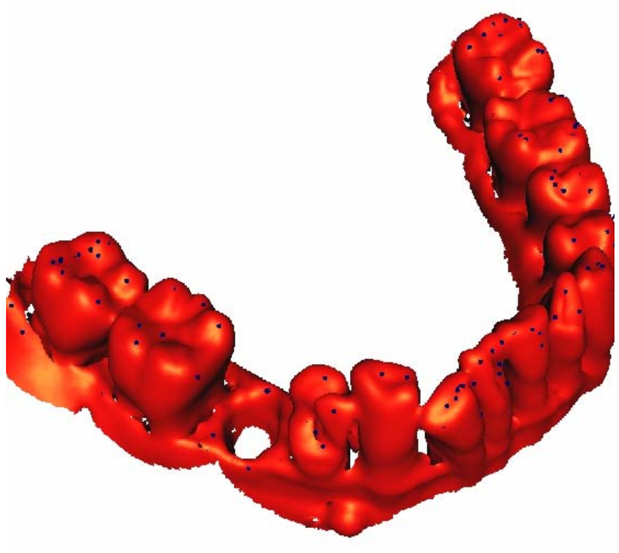
- ➪ **Performances estimation**
  - ✔ MR/US intensity-based registration
  - ➪ Surface based registration

- ❍ **Error prediction**

- ❍ **Segmentation**

- ❍ **Conclusion**

Points and surface registration for computer guided oral implantology

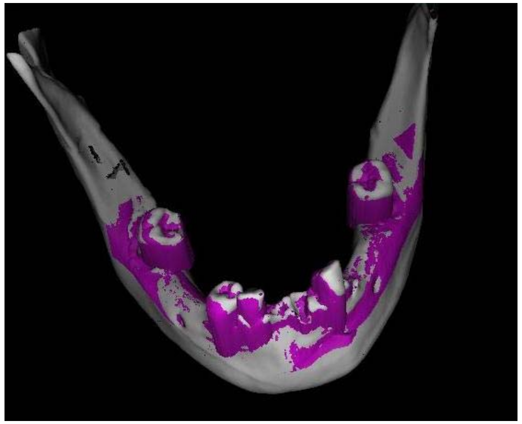


- Three point/surface registration pb.
- Robustness wrt the initial transfo
- Detection of failures
- Accuracy

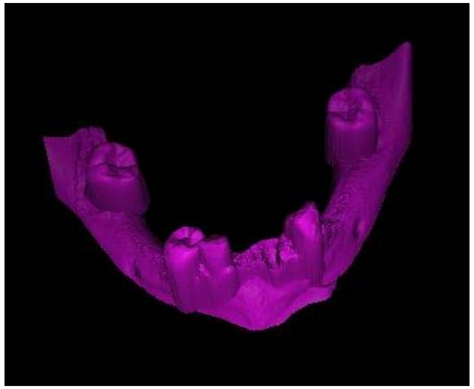
[AREALL, PhD Thesis S. Granger, MICCAI’01, ECCV’02]

Three points / surface registrations problems

Pre-operative CT Scan
Bone and teeth surfaces



Optical scan of dental cast
Teeth and gum surfaces

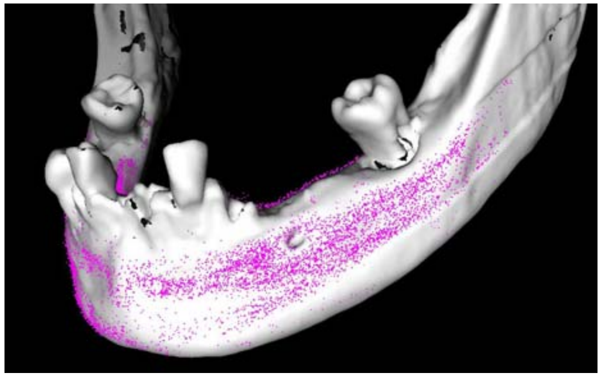


Surface / surface

Per-operative
consistency loop

Large clouds of points
on bone surface

Isolated points
on teeth surface



Per-operative points

Points / surface registration algorithm

Data modeling

- Surface : Gaussian mixtures
- Independent point measures

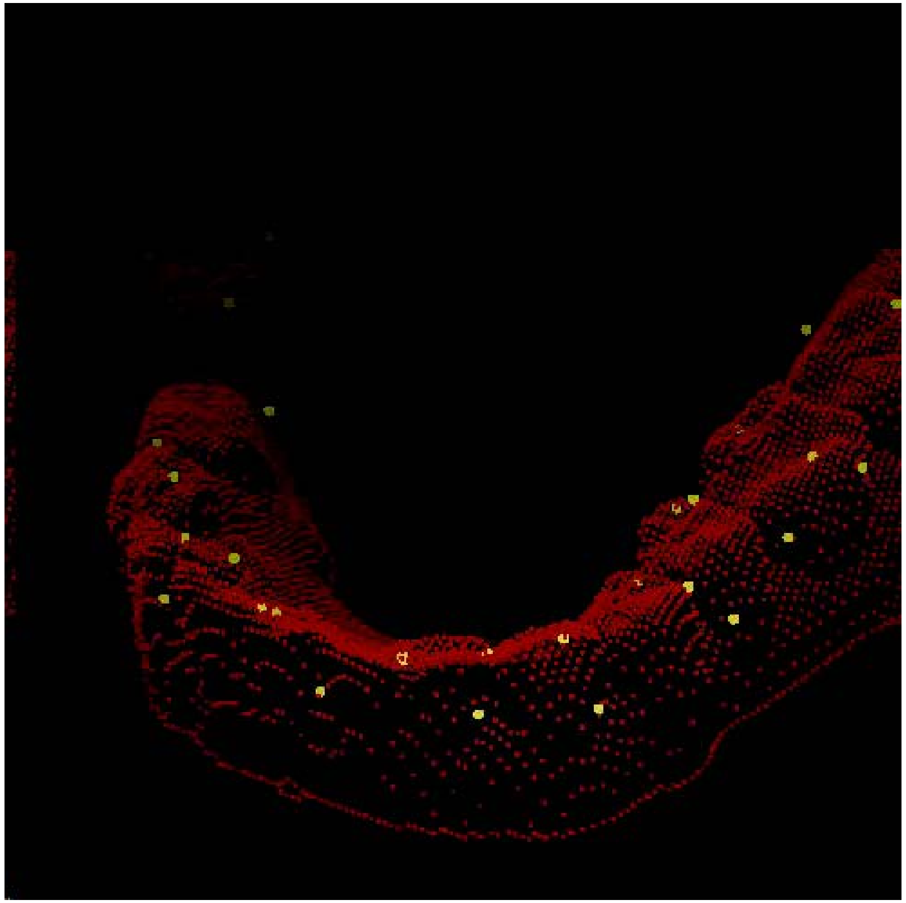
Algorithm

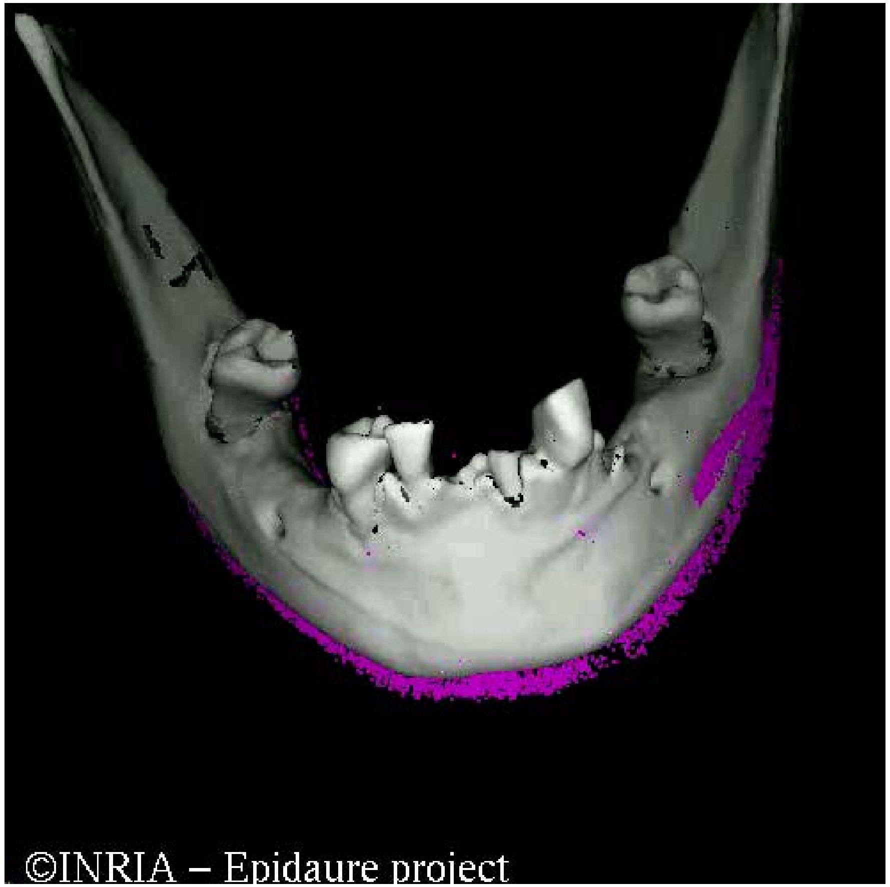
- ML estimation of the transformation: EM variant of ICP
- Multi-scale scheme on variance: principal axis -> ICP
- Multi-resolution scheme on points: computation time

	Sucess	Repeat. (mm)	Time (s)
ICP	15%	0	320
Basic EM	27%	0	530
Multiscale ICP	79%	0,1	25
Multiscale EM	88%	0	58

[S. Granger, X. Pennec et al., MICCAI 2001, ECCV 2002]

Points / surface registration





Points / surface registration performances

- Isolated points to surface registration

	CV basin (translation)	Repeatability	Fault detection (criterion value)
ICP	2 mm	0.2 mm	difficult
Multiscale EM	10 mm	0.007 mm	easy (100%)

- Surface to surface registration

	CV basin (translation)	Repeatability	Fault detection (criterion value)
ICP	6 mm	0.02 mm	easy
Multiscale EM	> 8 mm	0.005 mm	easy (100%)

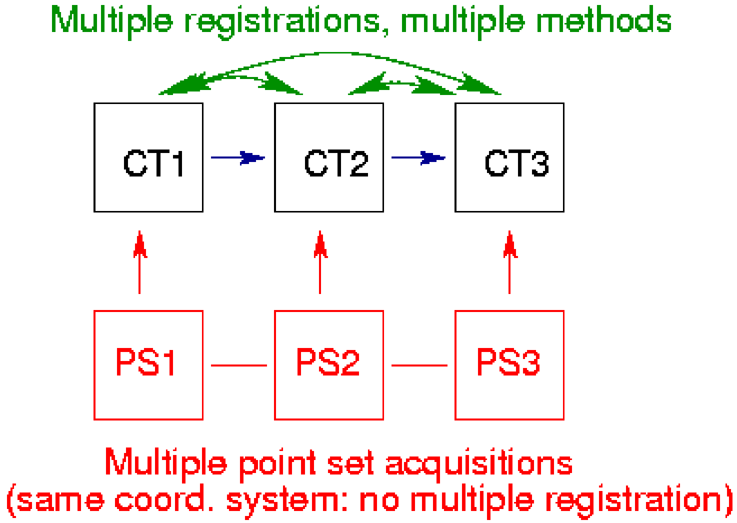
- Large clouds of points to surface registration

	CV basin (translation)	Repeatability	Fault detection (criterion value)
ICP	2 mm	0.04 mm	possible
Multiscale EM	5 mm	0.001 mm	easy

Points / surface registration accuracy

Measuring the uncertainty

- Multiple data acquisitions: independent transfo.
- Repeat on many subjects
- Bias will have to be tested on the real system.



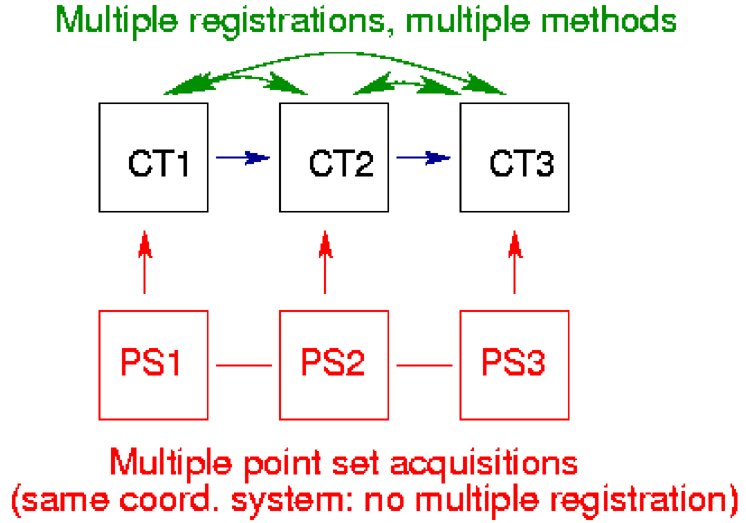
Establish a “bronze standard” based on several CT images

• Crest lines (extraction at two scales):	0.12 deg / 0.20 mm
• Yasmina (SSD and correlation ratio):	0.14 deg / 0.18 mm
• Aladin (correlation by block matching):	0.13 deg / 0.20 mm
• Multiscale EM-ICP:	0.08 deg / 0.10 mm
• Bronze standard:	0.03 deg / 0.04 mm

Points / surface registration accuracy

Measuring the uncertainty

- Multiple data acquisitions: independent transfo.
- Repeat on many subjects
- Bias will have to be tested on the real system.



Isolated points to surface registration accuracy

• ICP:	0.35 mm (50 pts)	0.19 mm (200 pts)
• Multiscale EM-ICP:	0.25 mm (50 pts)	0.16 mm (200 pts)

Point clouds to surface registration accuracy

• ICP:	0.17 mm (1500 pts)
• Multiscale EM-ICP:	0.17 mm (1500 pts)

Overview

✔ **Registration performances**

✔ **Performances estimation**

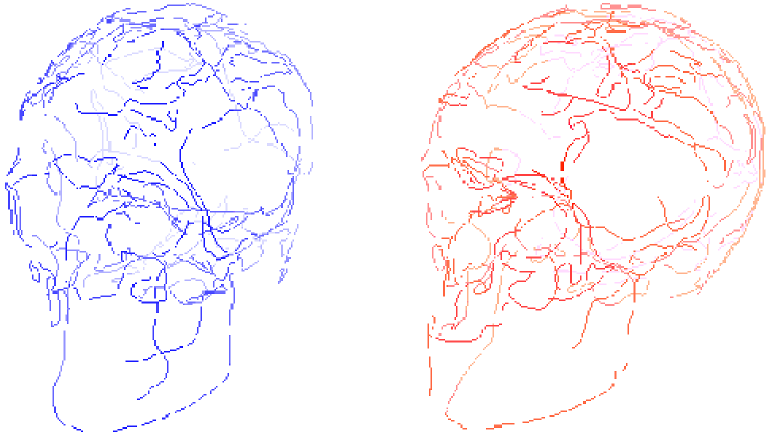
➪ **Error prediction**

- ➪ 3D/3D feature-based registration
- ❍ 3D/2D registration for Augmented Reality

❍ **Segmentation**

❍ **Conclusion**

Uncertainty of feature-based registration



Matches estimation

- Alignment
- Geometric hashing
- ICP

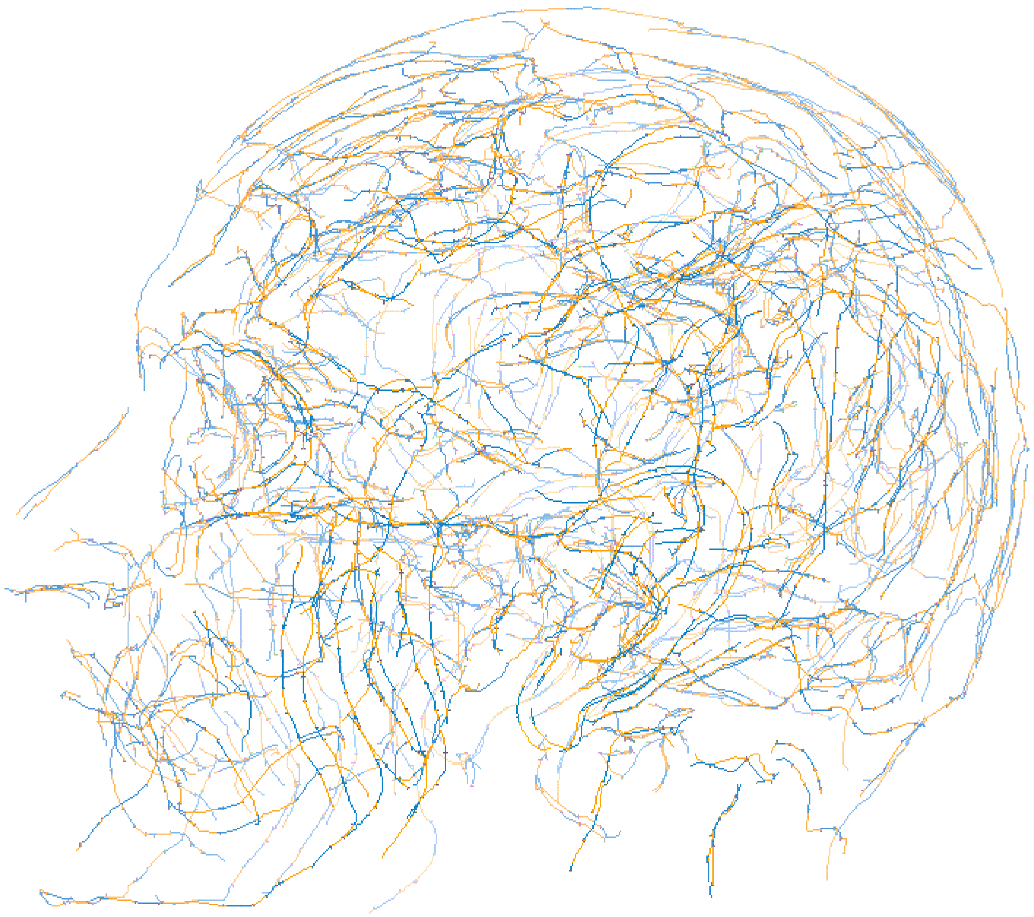
Least square registration

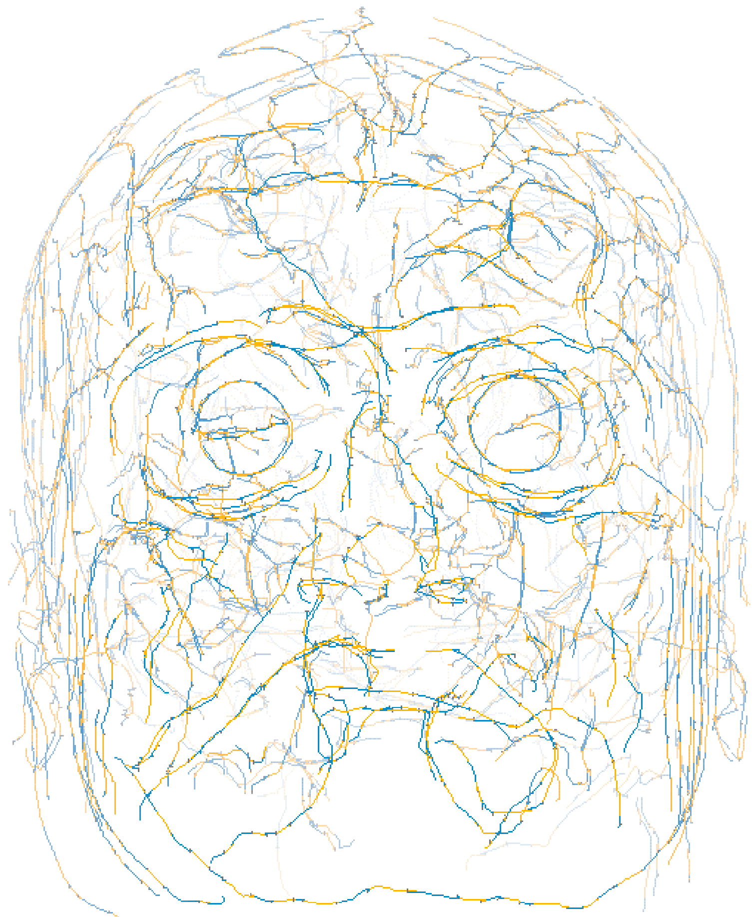
$$C(T,\chi) = \sum_i \left\| y_i - T * x_i \right\|^2$$

- Propagation of the errors from the data to the optimal transformation at the first order (implicit function theorem):

$$\Sigma_{\chi\chi} = \sigma^2.Id \quad \Rightarrow \quad \boxed{\Sigma_{TT} = \sigma^2.H^{-1}} \quad \text{with} \quad H = \frac{\partial^2 C(T,\chi)}{\partial T^2}$$

Registration of MR T1 images of the head





Typical object accuracy: 0.06 mm

860 matched frames among 3600

Typical corner accuracy: 0.125 mm

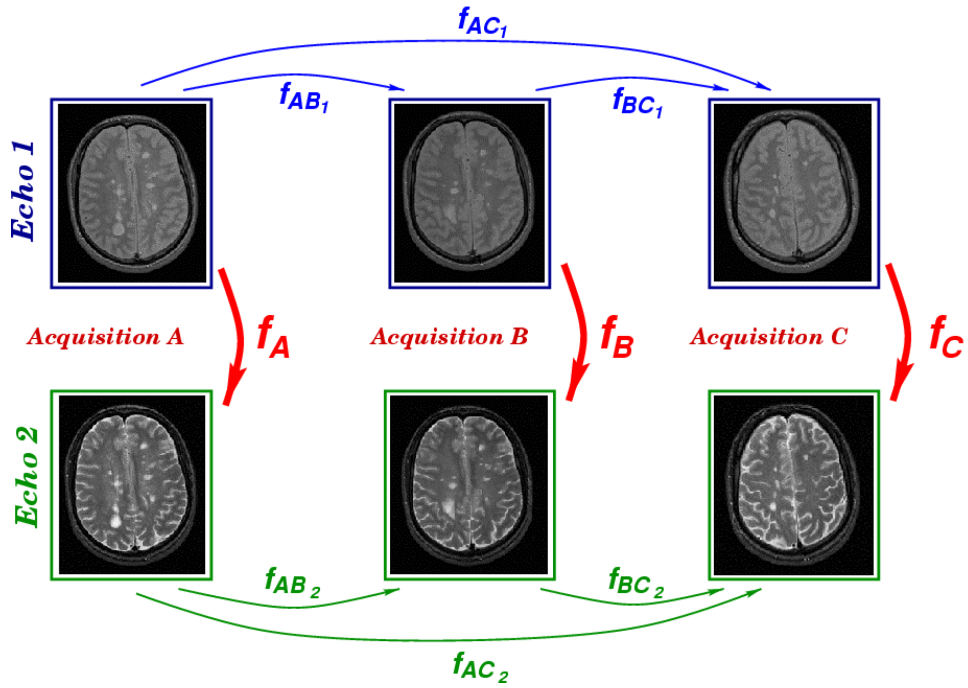
Validation of the error prediction

Comparing two transformations and their Covariance matrix :

$$\mu^2(T_1, T_2) \approx \chi_6^2$$

Mean: 6, Var: 12

KS test



Brigham and Women's Multiple sclerosis database

- 24 3D acquisitions over one year per patient
- T2 weighted MR, 2 different echo times, voxels 1x1x3 mm
- Predicted object accuracy: 0.06 mm.

[X. Pennec et al., Int. J. Comp. Vis. 25(3) 1997, MICCAI 1998]

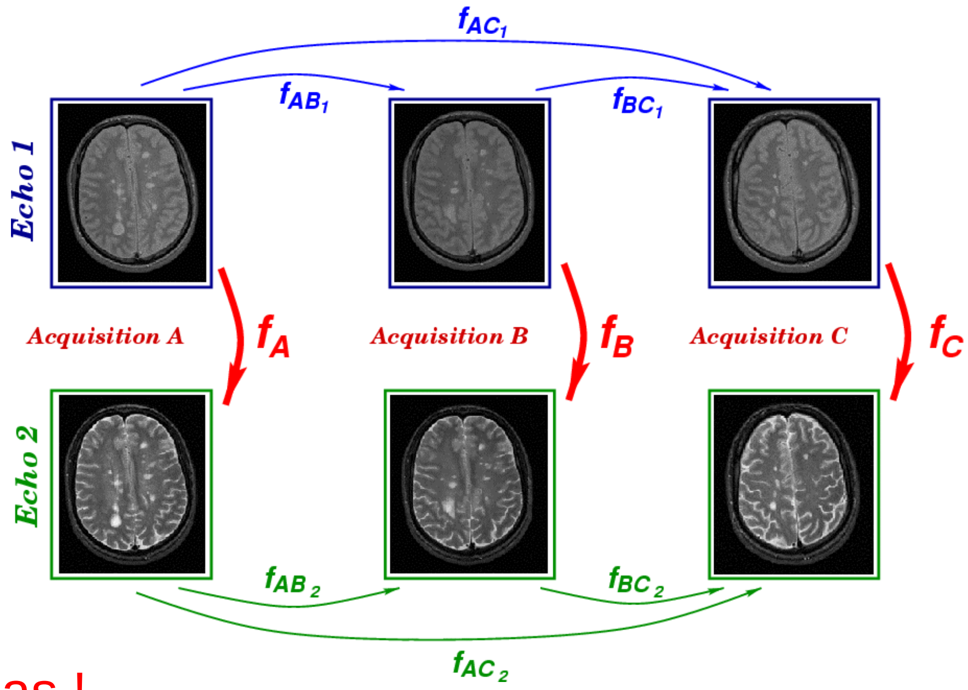
Validation of the error prediction

Comparing two transformations and their Covariance matrix :

$$\mu^2(T_1, T_2) \approx \chi_6^2$$

Mean: 6, Var: 12

KS test



Intra-echo: $\mu^2 \approx 6$, KS test OK

Inter-echo: $\mu^2 > 50$, KS test failed, Bias !

Bias esimation: (chemical shift, susceptibility effects)

- $\sigma_{rot} = 0.06$ deg (not significantly different from the identity)
- $\sigma_{trans} = 0.2$ mm (significantly different from the identity)

Inter-echo with bias corrected: $\mu^2 \approx 6$, KS test OK

[X. Pennec et al., Int. J. Comp. Vis. 25(3) 1997, MICCAI 1998]

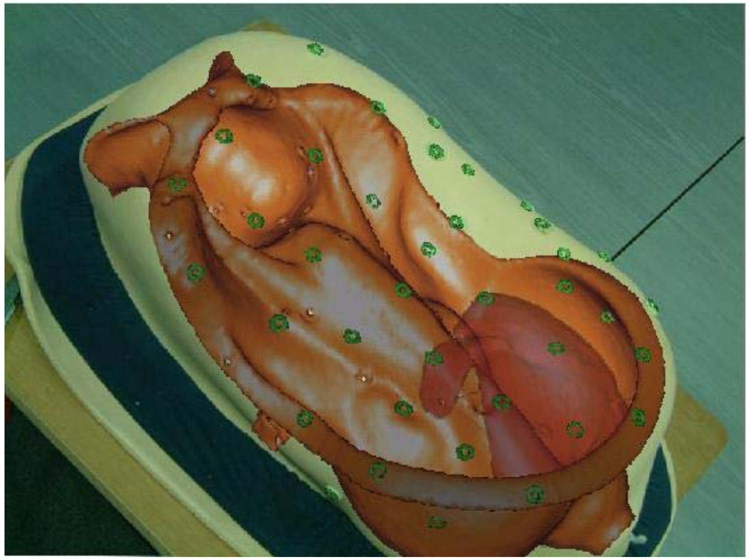
Overview

- ✔ **Registration performances**
- ✔ **Performances estimation**
- ➪ **Error prediction**
  - ✔ 3D/3D feature-based registration
  - ➪ 3D/2D registration for Augmented Reality
- ❍ **Segmentation**
- ❍ **Conclusion**

Augmented reality guided radio-frequency tumor ablation

Current operative setup at IRCAD (Strasbourg, France)

- Per-operative CT “guidance”
- Respiratory gating
  - ➨ Marker based 3D/2D rigid registration



S. Nicolau, X.Pennec, A. Garcia,L. Soler, N. Ayache

Increase 3D/2D registration accuracy:
A new Extended Projective Point Criterion

Standard criterion:

$$\sum_{l=1}^{M} \sum_{i=1}^{N} \left\| P^{l}(T*M_{i}) - m_{i}^{l} \right\|^{2}$$

- image space minimization (ISPPC)
- noise only on 2D data

Complete statistical assumptions + ML estimation

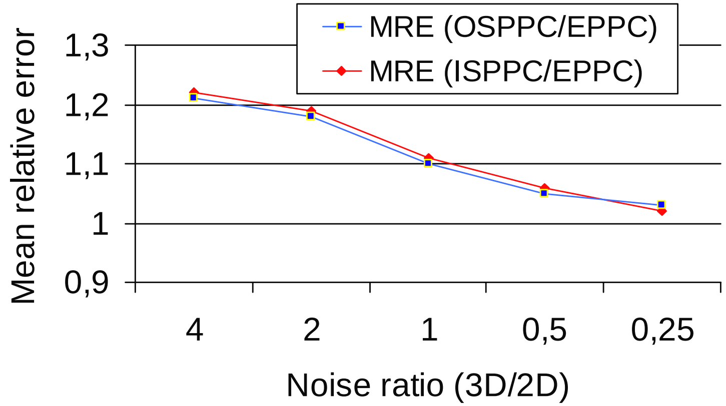
- Gaussian noise on 2D *and* 3D data
- Hidden variables M_i (exact 3D positions)

$$\sum_{l=1}^{M} \sum_{i=1}^{N} \frac{\left\| P^{l}(T*M_{i}) - \tilde{m}_{i}^{l} \right\|^{2}}{2\sigma_{2D}^{2}} + \sum_{i=1}^{N} \frac{\left\| M_{i} - \tilde{M}_{i} \right\|^{2}}{2\sigma_{3D}^{2}}$$

Performance evaluation of EPPC vs ISPPC

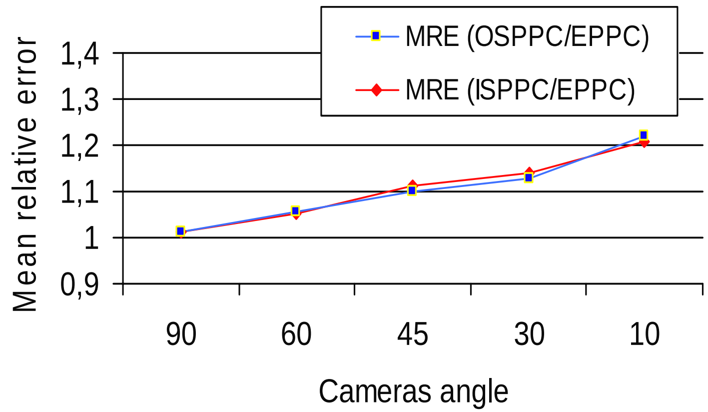
Simulation

- ❑ Better TRE (up to 20%)
  - large 3D/2D noise ratio
  - small cameras angle



- ❑ Better robustness

	Proba of CV	
ISPPC	57%	91%
OSPPC	85%	96%
EPPC	94%	99%



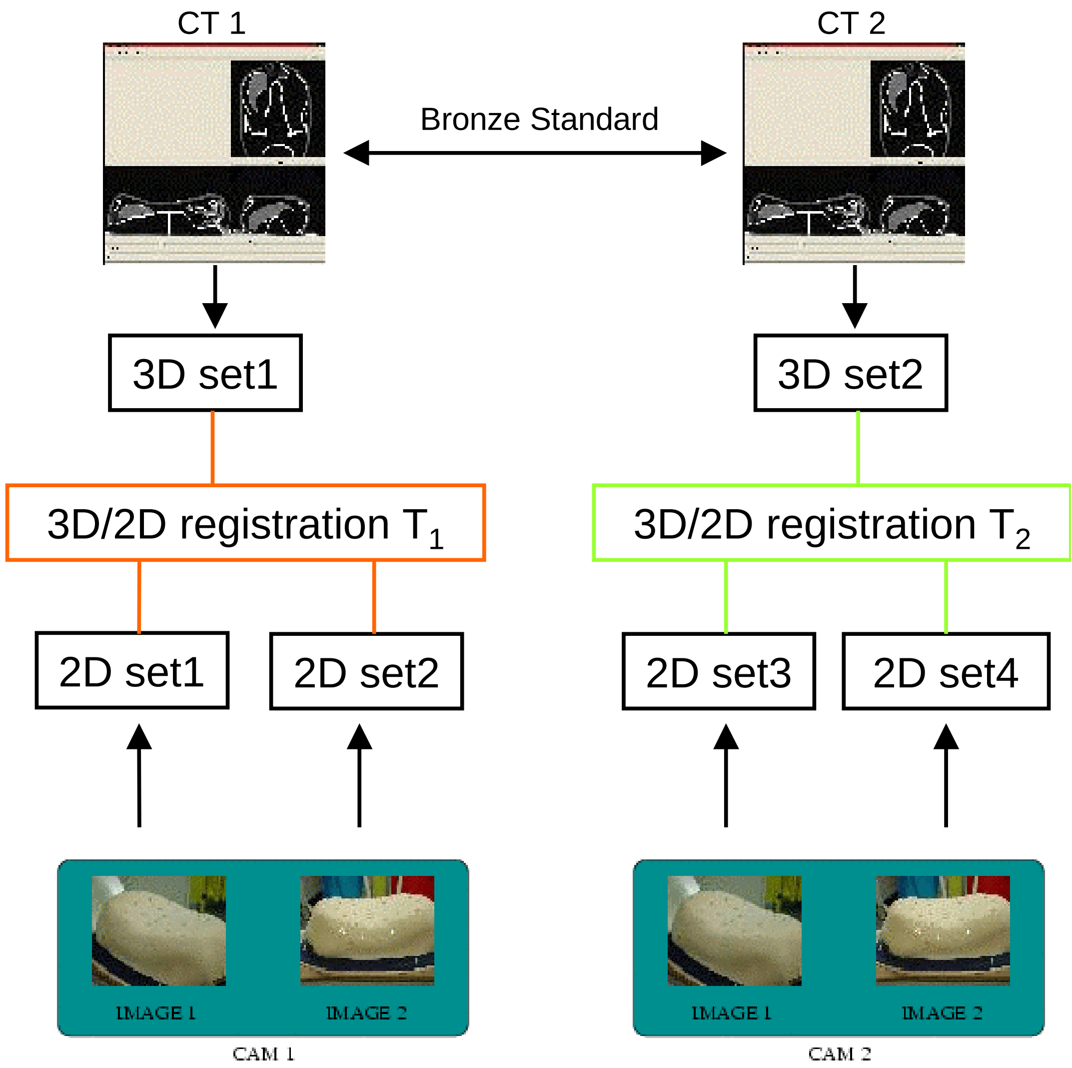
- ❑ Com. Time 10 to 20 larger

[S. Nicolau et al., IS4TM 2003]

Phantom acquisition protocol

Registration protocol:

- Real 2D and 3D images of a phantom
- Multiple CT acquisitions: independent transfo.
- 2 pairs of cameras calibrated together
- Soft skin (unwanted motion ~1.5 mm).

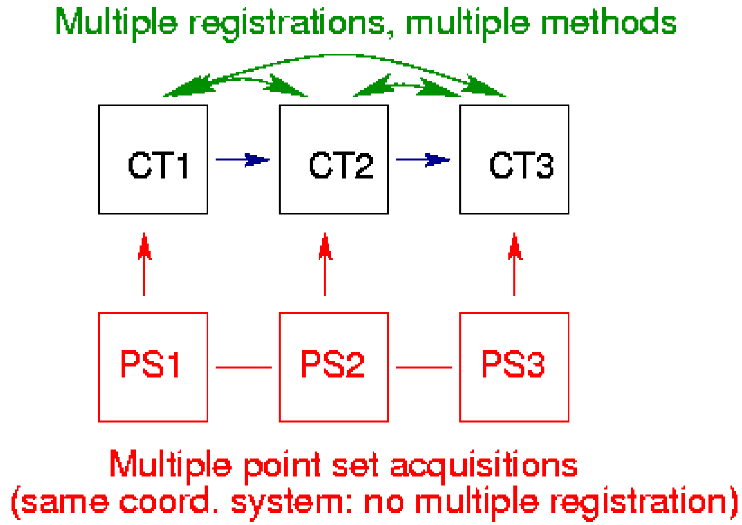


4 cameras calibrated together

Phantom bronze standard

Measuring the uncertainty

- Multiple CT acquisitions: independent transfo.
- 2x2 cameras calibrated together
- Soft skin (motion ~1.5 mm).



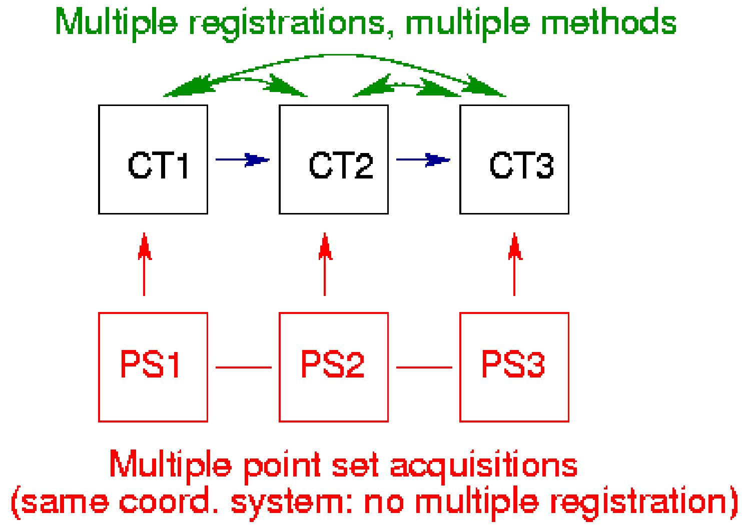
Establish a "bronze standard" based on several CT images

• Crest lines (extraction at two scales):	0.04 deg / 0.27 mm
• Yasmina (SSD and correlation ratio):	0.04 deg / 0.41 mm
• Aladin (correlation by block matching):	0.09 deg / 0.56 mm
• Multiscale EM-ICP:	0.08 deg / 0.68 mm
• Bronze standard:	0.01 deg / 0.07 mm
• Fiducials:	0.15 deg / 0.85 mm

Phantom bronze standard

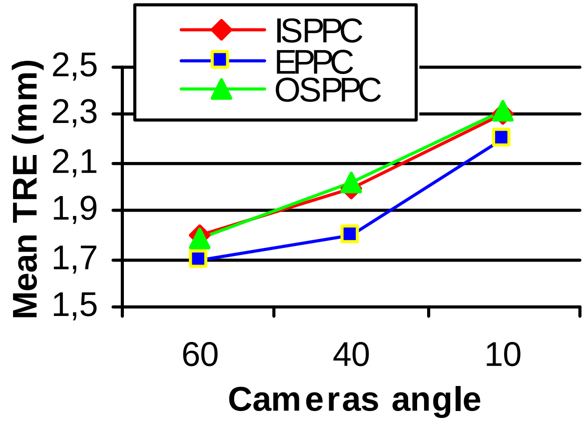
Measuring the uncertainty

- Multiple CT acquisitions: independent transfo.
- 2x2 cameras calibrated together
- Soft skin (motion ~1.5 mm).



Performances evaluation

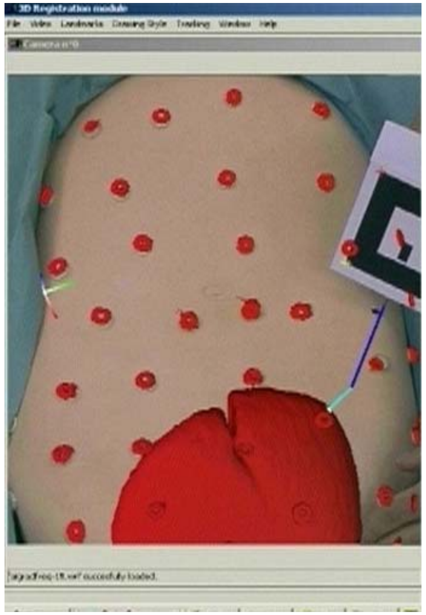
- EPPC TRE up to 9 % better:
  24 -> 20 markers
- Mean TRE of 2 mm
- Comp. Time: 0.2 s



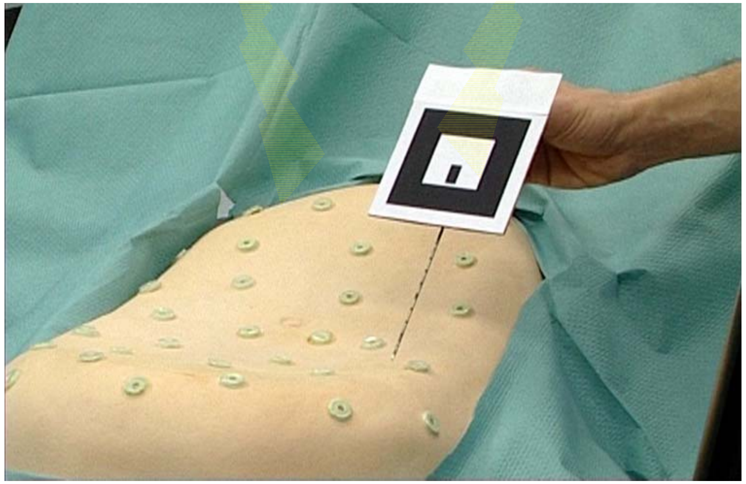
[S. Nicolau et al., IS4TM 2003]

Whole loop accuracy evaluation

Left monitor with AR & needle tracking

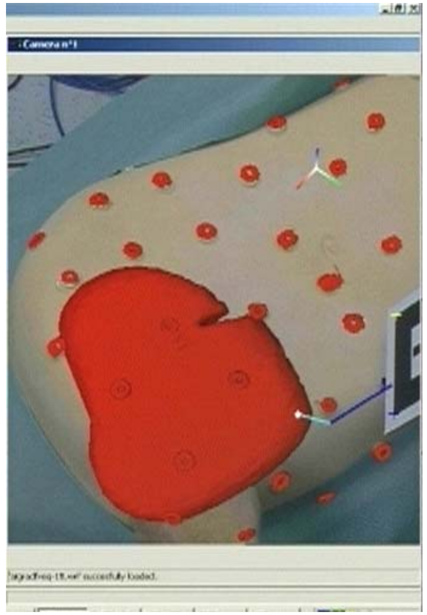


Stereoscopic HD video acquisition



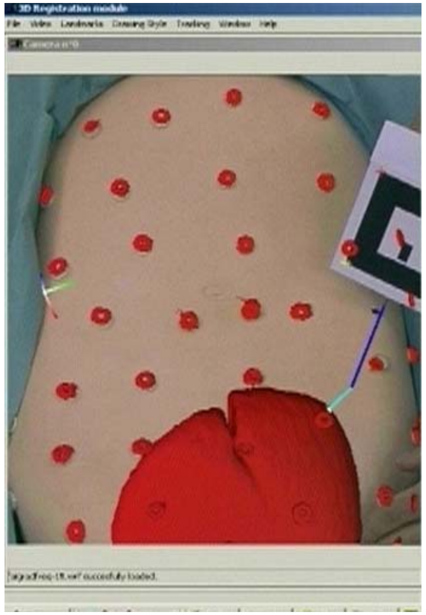
Real scene: phantom + needle

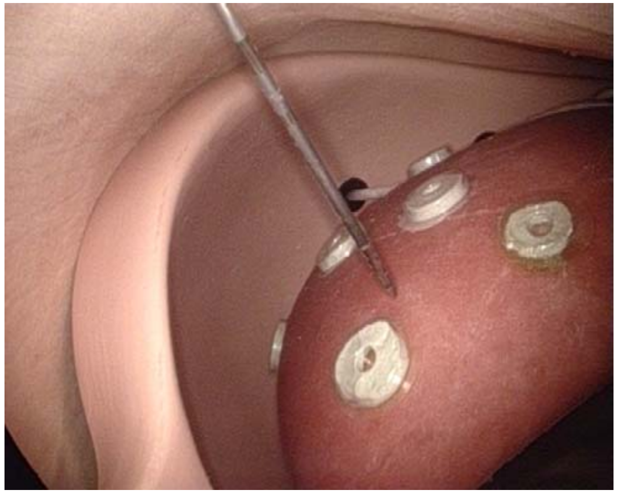
Right monitor with AR & needle tracking



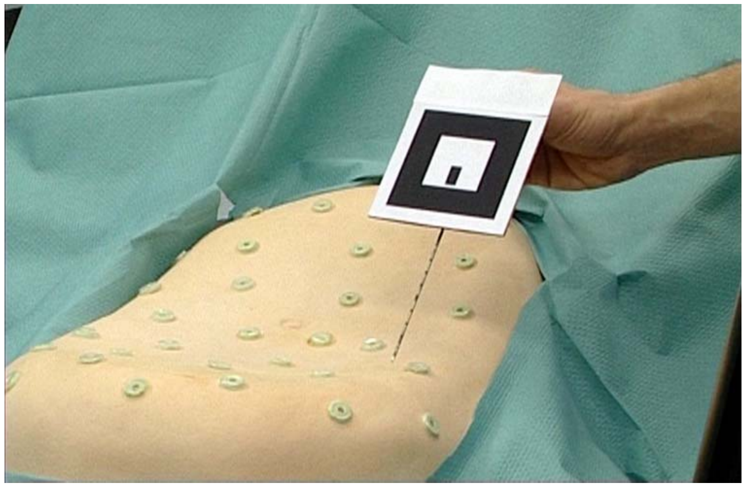
Whole loop accuracy evaluation

Left monitor with AR & needle tracking



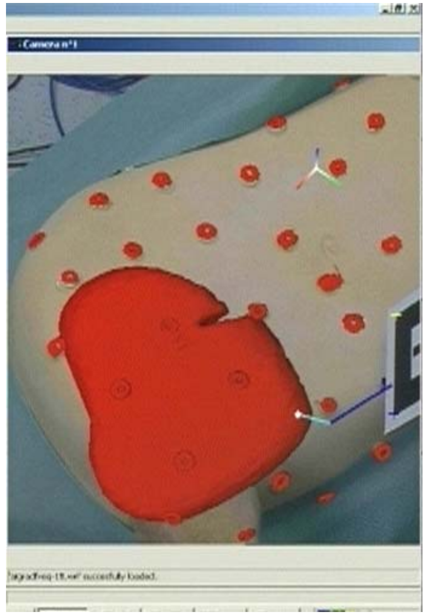


Endoscopic control



Real scene: phantom + needle

Right monitor with AR & needle tracking

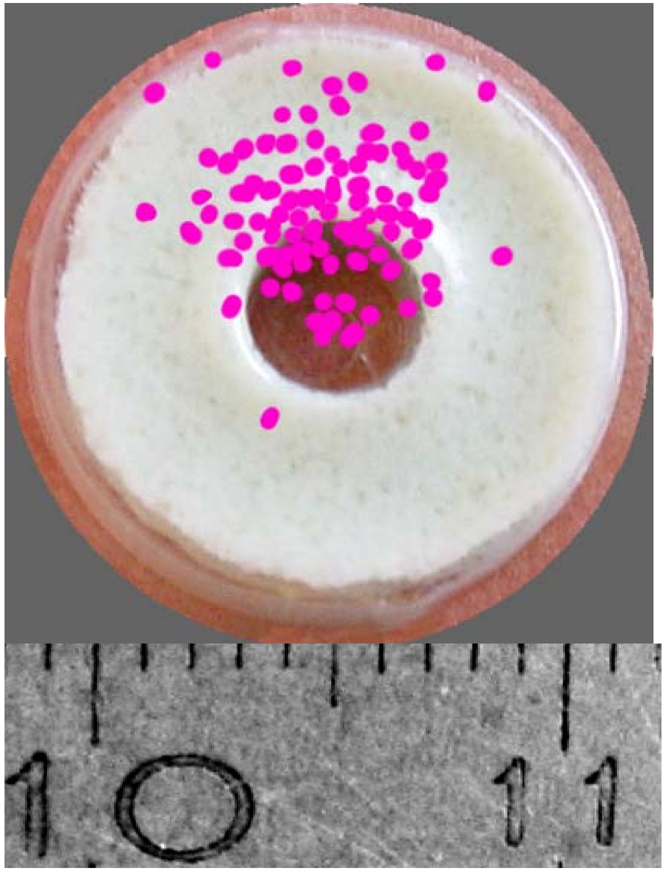


DivX Video File

Whole loop accuracy evaluation

Experimental setup

- Two participants (comp. sci. + surgeon)
- 100 needle targetings



Measures

- Distribution of hits
  (endoscopic view, video recording)

- Average deviation from target
  2.8 mm $\pm$ 1.4

- Average targeting time:
  46.6 sec. $\pm$ 24.64

[S. Nicolau, A. Garcia et al., Aug. & Virtual Reality Workshop, Geneva, 2003]

Improving the safety

Performances evaluation

- Hundreds of simulations
- Multiple phantom experiments

➪ Difficult to span the whole range of parameters

Error propagation

- Predict target registration error for a given configuration
- 1st order propagation of the 2D and 3D noise on points

➪Need to validate for the absence of biases

[S. Nicolau et al., INRIA RR 4993, 2003]

Validation of the error prediction

Incremental validation protocol

- Check non linearities of the criterion (1st order propag.)
  - Simulation

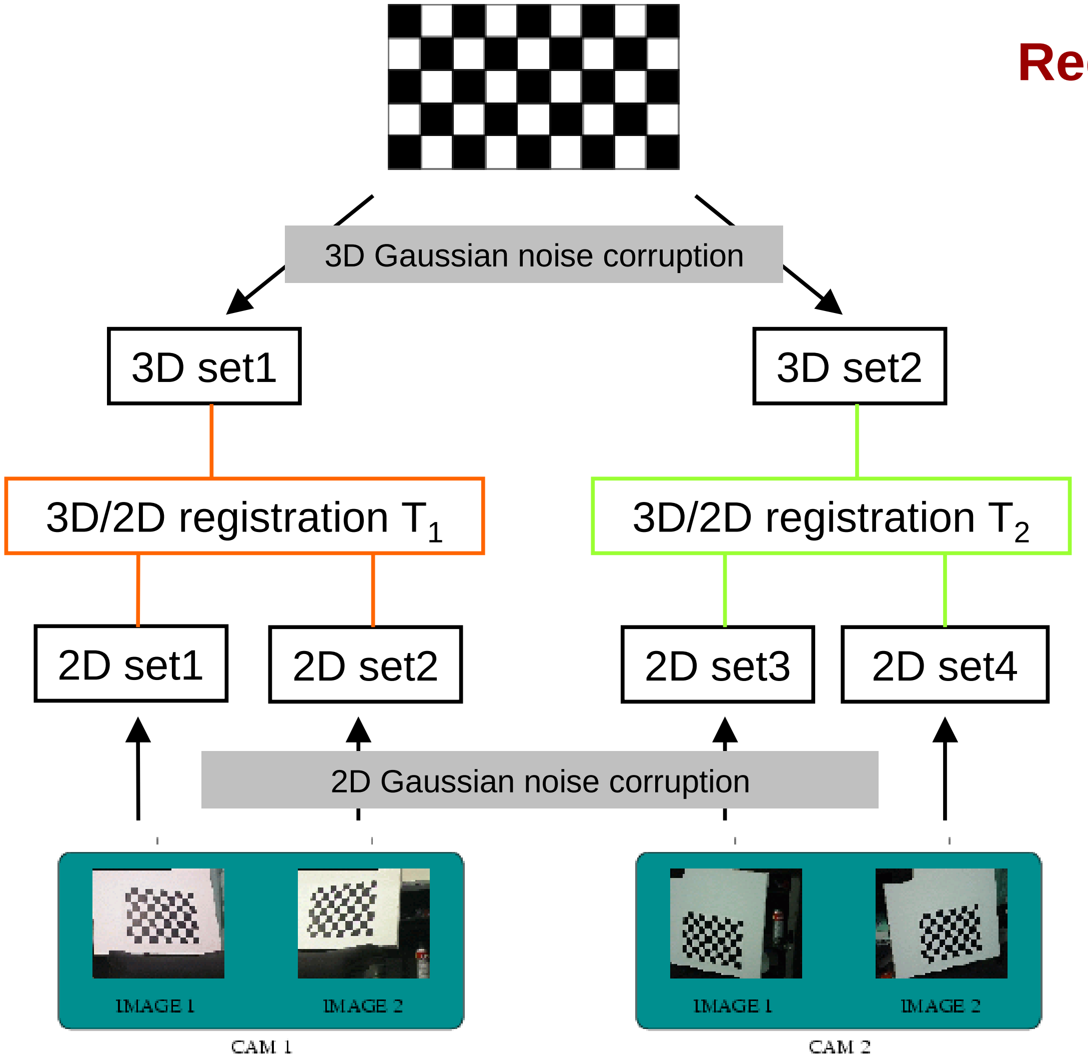
	μ^2 (3)	σ (2.45)	KS-conf. (0.01…1.0)
SPPC	**3.020**	**2.506**	**0.353**
EPPC	**3.016**	**2.486**	**0.647**

- Check unbiased and accurate calibration
  - Real 2D images with synthetic (Gaussian) noise
- Check Gaussian 3D and 2D noise on points
  - Real 2D and 3D images of a phantom

Validation methodology

- Mahalanobis distances (1D χ^2 measures)
- KS tests

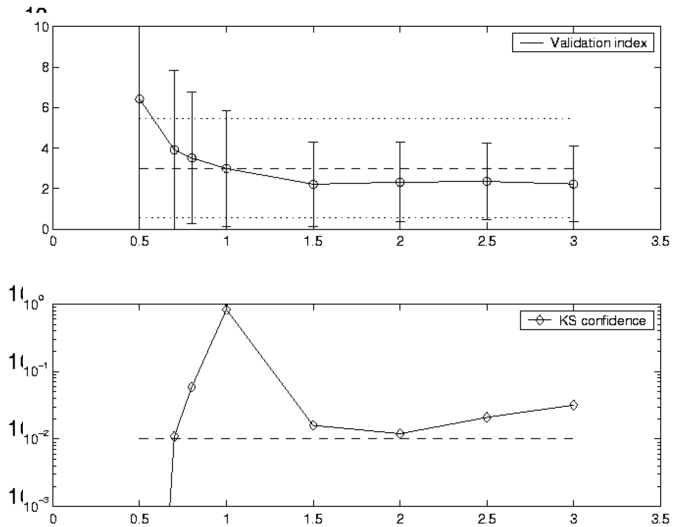
Check unbiased and accurate calibration



4 cameras calibrated together

Registration loops:

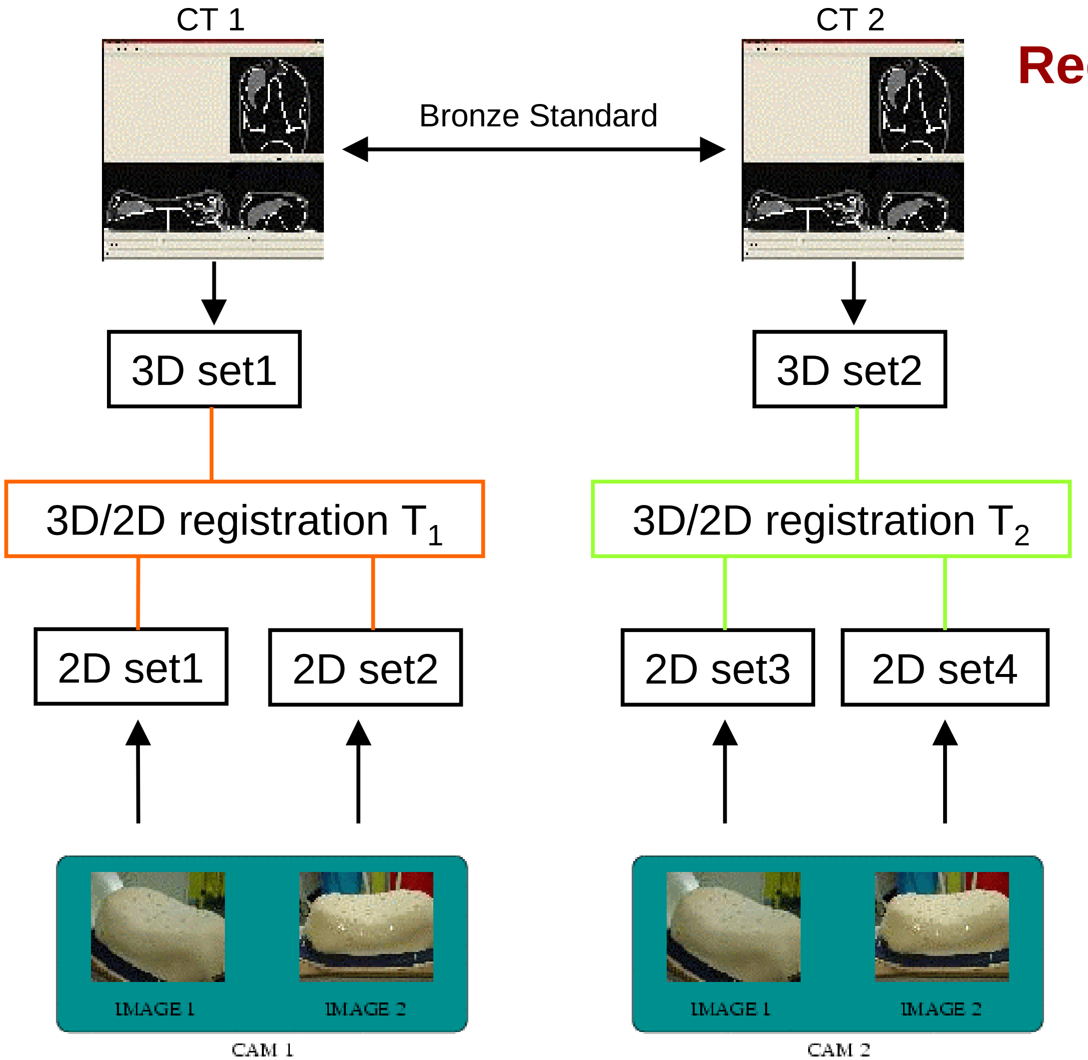
- Real 2D images with synthetic (Gaussian) noise



- Number of points < 36
- 2D and 3D noise > 0.7

Successful validation if TRE > 1.5 mm

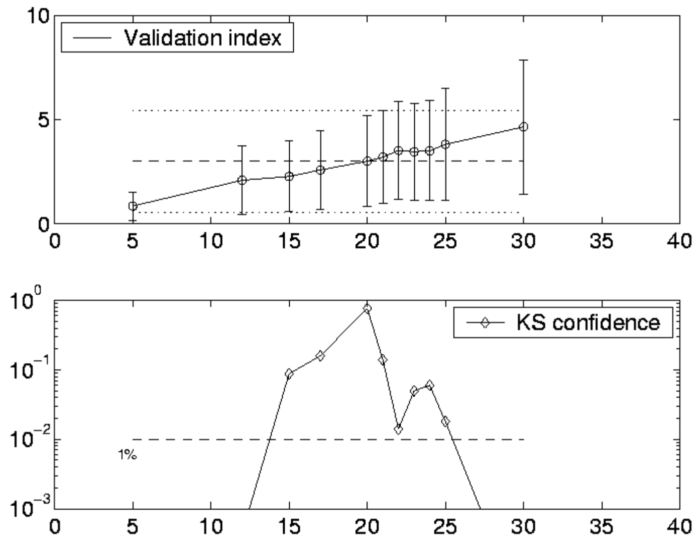
Check Gaussian 3D and 2D noise on points



4 cameras calibrated together

Registration loops:

- Real 2D and 3D images of a phantom



Consistent motion of the skin (about 1 mm)

Good error prediction in our application range

Validation of the error prediction

Incremental validation protocol

- Check non linearities of the criterion
  - OK
- Check unbiased and accurate calibration
  - OK if registration TRE > calibration accuracy (~ 1.5 mm)
- Check Gaussian 3D and 2D noise on points
  - Consistent motion of the skin of ~ 1.5 mm (bias)
  - OK within our application range if overestimation of the 3D noise

[S. Nicolau et al., INRIA RR 4993, 2003]

Overview

- ✔ **Registration performances**
- ✔ **Performances estimation**
- ✔ **Error prediction**
- ➪ **Segmentation**
- ❍ **Conclusion**

Validating Automatic Segmentation

Segmentation result: what measure to use?

- Volume (set of voxels) / surface / probability map
- important qualitative parameters for the application
  - Absolute or relative volume of the organ
  - Localization

Quantification of errors on end-points parameters

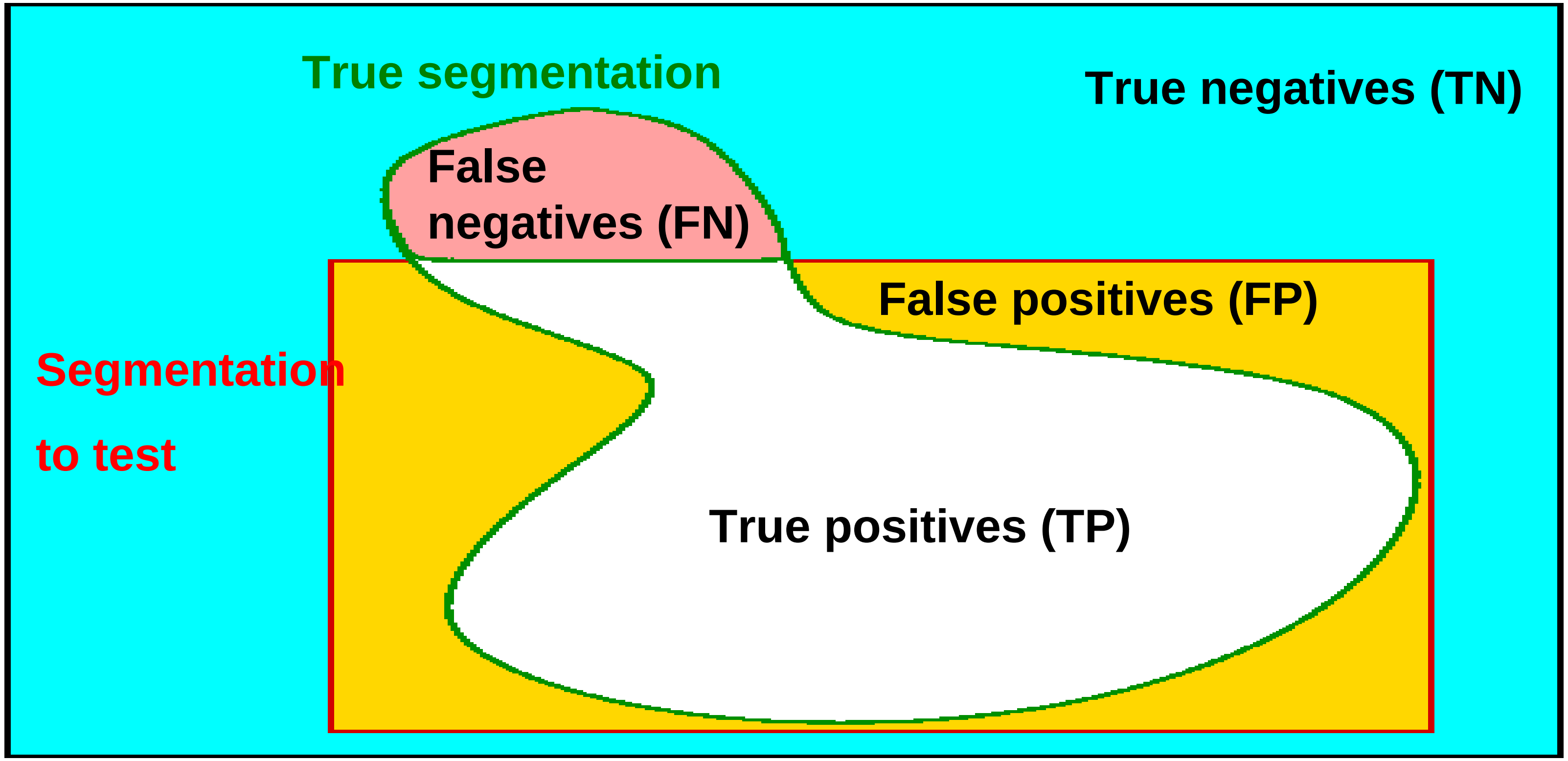
  - Gross errors: outliers
  - Small errors: accuracy

Ground truth segmentation by clinical experts

- intra and inter operator variability
  - limited accuracy (of the order of what we want to evaluate)

Sensitivity / specificity on volume

Sensitivity: $p = \frac{TP}{TP + FN}$ **Specificity:** $q = \frac{TN}{TN + FP}$



- Depends on the normalization volume !

Staple algorithm

Principle

- True segmentation volume = hidden variable
- Model of a segmentation method or operator:
  - Unknown Specificity
  - Unknown Sensitivity

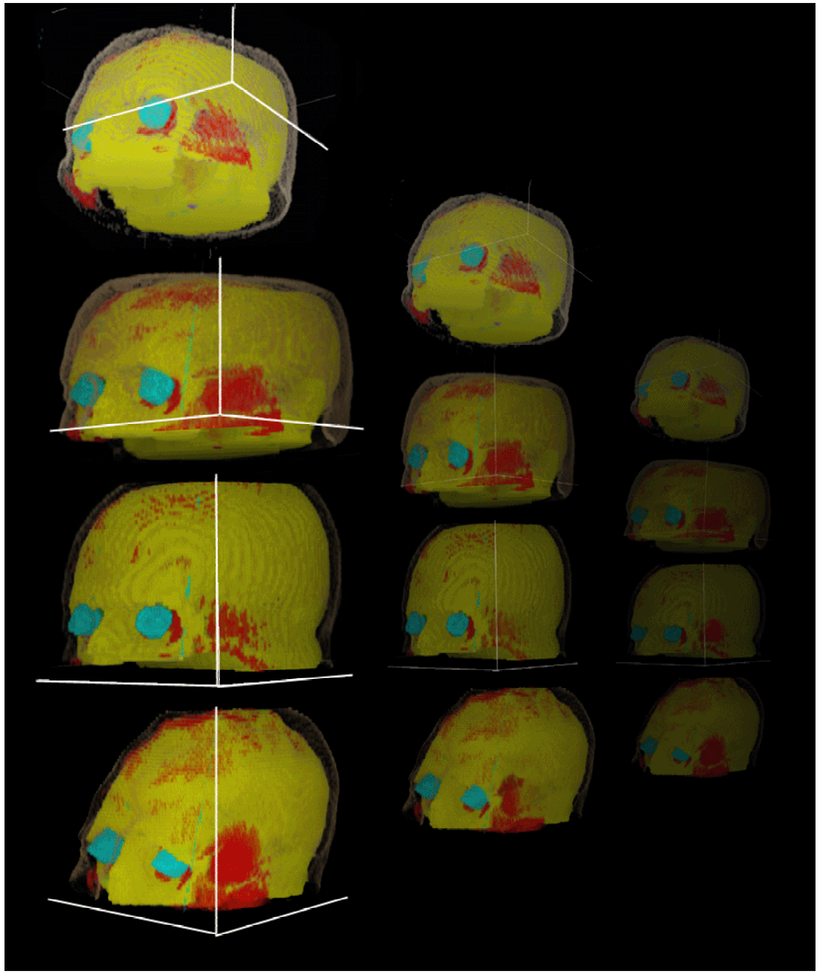
EM algorithm

- E-step: determine the “true” segmentation
- M-step: determine quality parameters for each expert

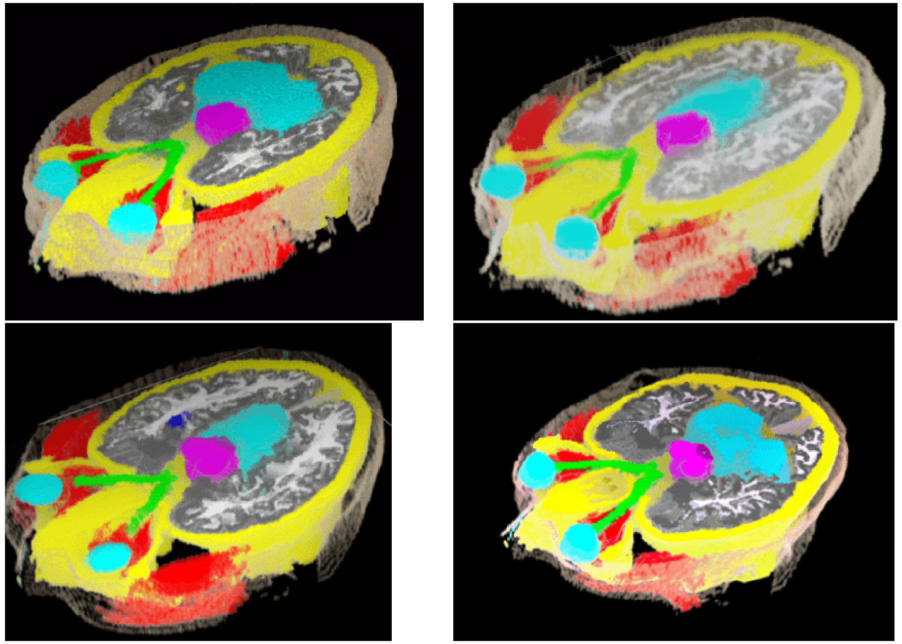
[S. Warfield, MICCAI 2002]

Atlas-based segmentation for radio-therapy

CAL Radiotherapy Atlas (P-Y Bondiau, MD, CAL, Nice)



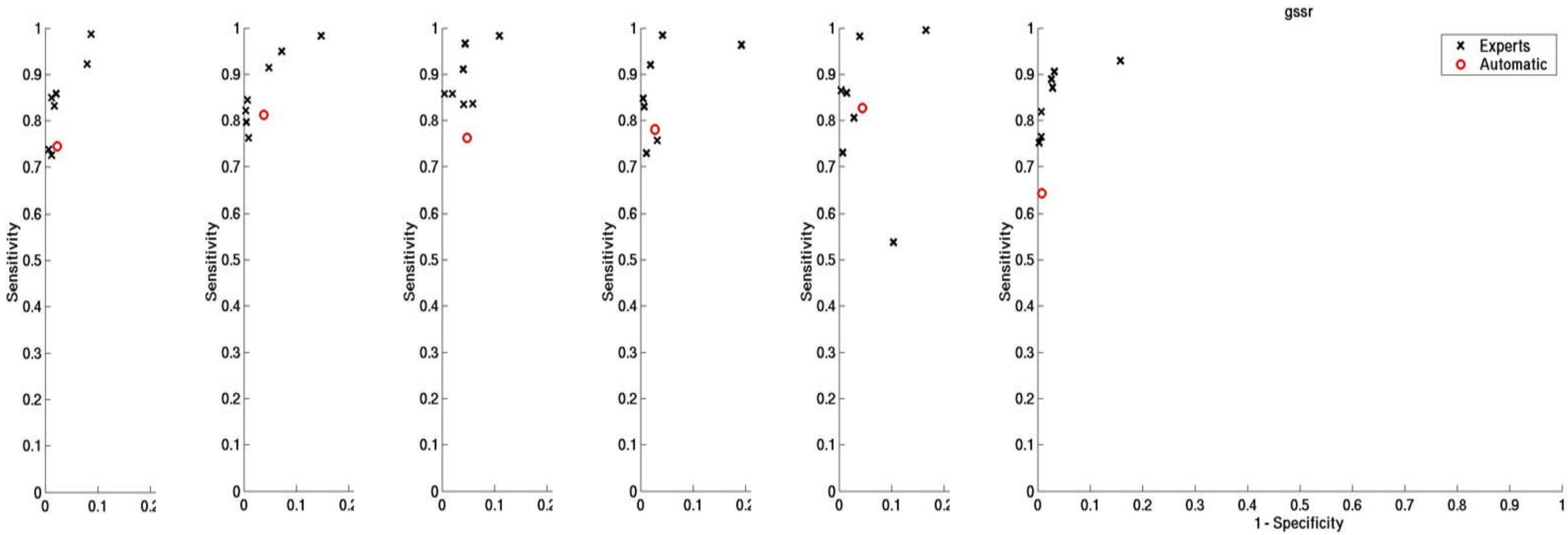
- Artificial MR image (MNI simulator)
- Segmentation of structures of interest
- Non-rigid registration with Pasha [Cachier, CVIU 2003]



Validation of the Brain-stem segmentation

"Ground truth"

- Brainstem volume for 6 patients, 7 experts



Ongoing work

- Choice of the optimal parameters (learning/test data)
- Incorporate a tumor model

Conclusion

Tools for performance evaluation without gold standard

- Registration or consistency loops
- Ground truth as a hidden variable
- Error prediction
- Cross comparison of criteria
- Leave-one-out methods

Methodological guidelines

- Identify (and model if possible) all sources of variability
- Incremental validation/evaluation setup to test for successive noise assumptions
- A database of test images should be representative of all clinical conditions

Acknowledgments

- A. Roche
- P. Cathier
- S. Nicolau
- S. Granger
- J.P. Thirion
- O. Commowitz
- P.Y. Bondiau
- and the remaining of the EPIDAURE Team

- L. Soler, A. Garcia (IRCAD, Strasbourg)
- R. Derycke (AREALL, Paris)
- C. Guttmann (BWH, Boston)