HUMAN SHAPE AND MOTION FROM VIDEO

Pascal Fua
CVLab EPFL
Switzerland
cvlab.epfl.ch

MODELING PEOPLE

Media technologies
- Electronic publishing in 2 and 3—D
- Education and training
- Scientific visualization
- Database retrieval

Entertainment
- Special effects
- Video Games

Medicine and sports
- Motion Analysis
- Outcome evaluation
- Plastic surgery

Smart interfaces
- Gesture recognition
- Facial motion understanding

Surveillance
- Incident detection
- Automated recognition
- Behavioral analysis

Currently, and in the foreseeable future, a hot R&D area.

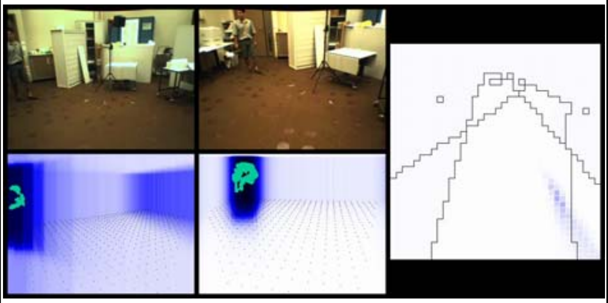
USING IMAGES

Good news:
- Images are readily available and can be acquired using ever cheaper sensors.

Bad news:
- Images provide noisy and incomplete information.

→ **Use models to overcome poor data quality.**

SURVEILLANCE

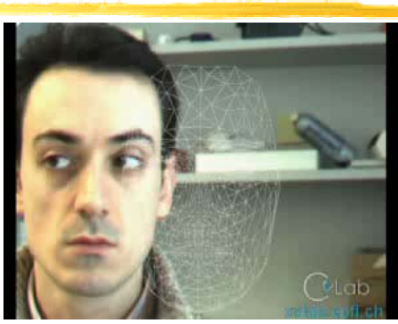


TALK OUTLINE

3—D models for
- Tracking and Detection
- Head Modeling
- Body Modeling

→ Allow the use of powerful constraints.

HEAD DETECTION AND TRACKING



Real-time tracking at 25Hz

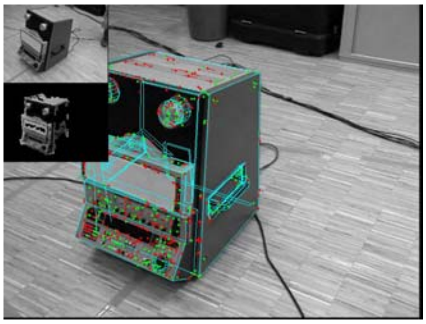
3—D TRACKING

Feature based tracking that combines:

- Short-baseline matching with previous frames
- Wide-baseline matching with keyframes

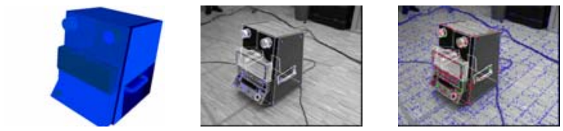
→ Tracking at 25Hz without drift or jitter.

PROJECTOR



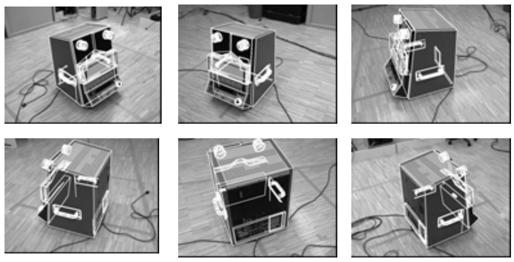
FEATURE-BASED TRACKING

- Interest points detection and matching;
- Robust viewpoint estimation from 2D-3D correspondences;
- Accounting for appearing/disappearing points.



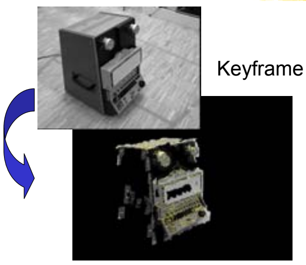
→ Robust but tends to drift.

KEYFRAMES



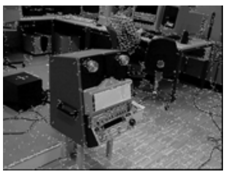
Using reference frames to eliminate drift.

WIDE BASE-LINE MATCHING



Keyframe

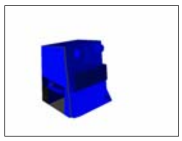
Keyframe rendered from the previous viewpoint estimate



Current frame

KEYFRAME CHOICE

Appearance-based criterion:

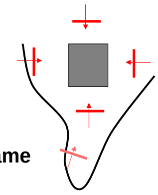


$$\min_{Keyframe} \sum_{f \in \text{Model}} \left(\text{Area}(f, A_P[R_P \mid T_P]) - \text{Area}(f, A_K[R_K \mid T_K])\right)^2$$

→ Can be estimated quickly using OpenGl

ADDING KEYFRAMES ONLINE

If the number of points between the current and closest key frame falls below a threshold, the previous frame becomes a keyframe.



Online Keyframe

KEYFRAMES ONLY

Jitter is clearly visible

COMPLETE METHOD

Keyframes eliminate drift but, if used alone, introduce jitter → 2-step process:

1. Initial estimate using closest keyframe;
2. Refinement using previous frame.

→ Tracking without drift or jitter.

KEYFRAME + RECURSIVE TRACKING

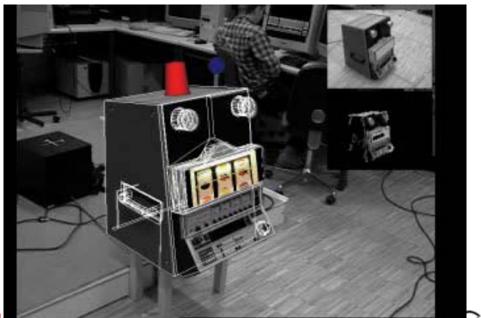
→ No jitter and robust to aspect changes.

PERFORMANCE

Images 384x288:
- 25 fps on Pentium 4 2.6 GHz

Possible improvements:
- Faster processor (3.04 GHz already available)
- Dual processor
- Code optimization

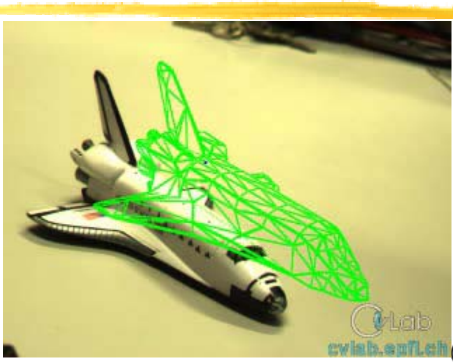
SLOT MACHINE



VIDEO AUGMENTATION

VIDEO AUGMENTATION

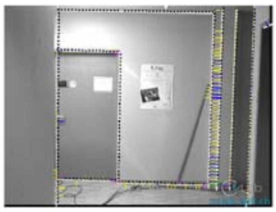
SHUTTLE



COMBINING EDGE AND TEXTURE INFORMATION

- Improved accuracy for untextured objects.
- Reduced number of keyframes.

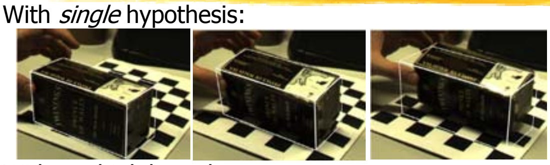
Keypoints only



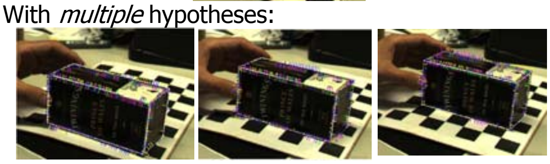
With edge information

MULTIPLE HYPOTHESES WHEN TRACKING EDGES

With *single* hypothesis:

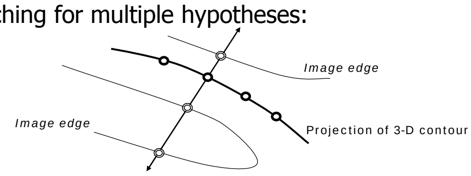


With *multiple* hypotheses:



ROBUST ESTIMATOR FOR MULTIPLE HYPOTHESES

Searching for multiple hypotheses:



Edge contribution: $v_i = \frac{1}{N_e}\sum_i\sum_j \left(\Delta_i(E_i, e'_{i,j,1}), \ldots, \Delta_i(E_i, e'_{i,j,K_{i,j}})\right)$

where ρ^* is our robust estimator for multiple hypotheses:

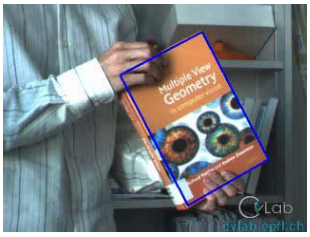
$$\rho^*(x_1, \ldots, x_n) = \min_i \rho(x_i) \quad \text{where} \quad \rho(x) = \begin{cases} \frac{c^2}{6}\left[1-\left(1-\left(\frac{x}{c}\right)^2\right)^3\right] & \text{if } |x| \le c \\ \frac{c^2}{6} & \text{if } |x| > c \end{cases}$$

CORRIDOR

CORRIDOR

DETECTION AT FRAME RATE

Target Object



25 frames/sec, 640×480 images,
on a Pentium 4 2.6 GHz

AFFINE INVARIANT MATCHING

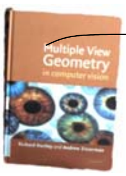
- [Schmid and Mohr 97]
- [Tuytelaars and VanGool 00]
- [Mikolajczyck and Schmid 02]
- [Lowe 04]
- ...

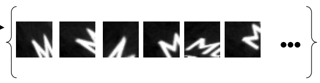
Affine deformation estimation ⇨ *Feature point characterization* ⇨ *Search for the corresponding point in the database*

→ By contrast, we propose to introduce a training stage to speed up online detection.

Direct Classification of the patch around the feature point

VIEW SET





- **View Set:** Set of all possible appearances of a keypoint under different viewing conditions.
- **Approach:** Train a classifier to recognize the viewsets build by synthesizing new views of the object keypoints.

→ Fast keypoint recognition in the input image.

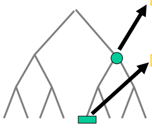
LOCAL PLANARITY CONSTRAINTS

- Warping the patch under an affine transformation:
  $(\mathbf{n} - \mathbf{n}_0) = \mathbf{A}(\mathbf{m} - \mathbf{m}_0) + \mathbf{t}$
  with $\mathbf{A} = \mathbf{R}_\theta (\mathbf{R}_\phi)^{-1} \mathbf{S} \mathbf{R}_\phi$, $\mathbf{t} = (t_u, t_v)$
- Robustness to localization error:
  t *is allowed to vary in the range of a few pixels.*
- Invariance to illumination changes:
  Normalization of synthesized patch intensities.

DECISION TREES

They naturally handle multi-class problems:

- Fast classification
- High recognition rate
- Provide a probability for the matches



- Nodes: very simple test: $I(x_1, y_1) - I(x_2, y_2) > s$ with x_1, y_1, x_2, y_2, s chosen to minimize the entropy $\sum_i P_i \log P_i$
- Leaves give the match probability.

Remove the 2D rotation in the Lowe manner: good compromise speed / classification.

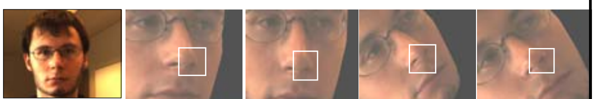
DETECTING A SAIL

Training image

GENERIC 3D OBJECTS

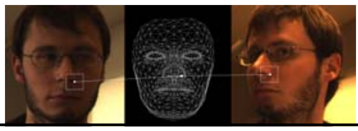
Use 3D model it to generate the viewsets:

- Capture complex appearance changes



Training image

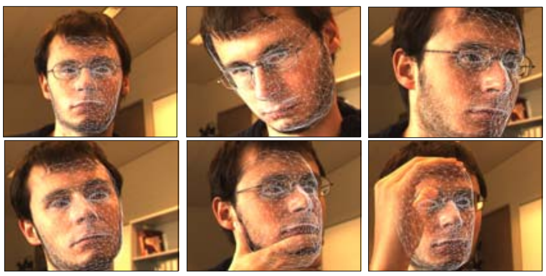
- Merge information from several training images:



FACE POSE ESTIMATION TRAINING

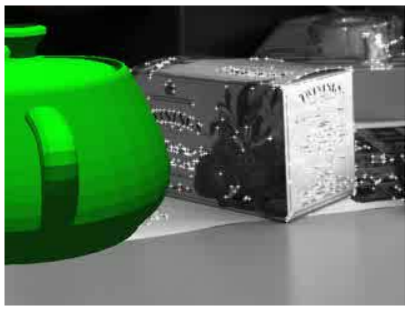
3 training images

FACE POSE ESTIMATION RESULTS



Without glasses *Partial occlusions*

AUTOMATED INITIALIZATION



NATURAL INTERACTION WITH MOBILE DEVICES

- Tourist photographs building using camera attached to PDA/Phone.
- System superposes 3—D model of the target object onto the image.
- Tourist can now point at any part of the image and obtain information about it.

CONTRIBUTIONS

Real-time algorithms for

- 3D tracking robustly and without drift.
- 3D detection and pose estimation.

→ Numerous potential applications in the fields of AR, man-machine interfaces, visual servoing ….

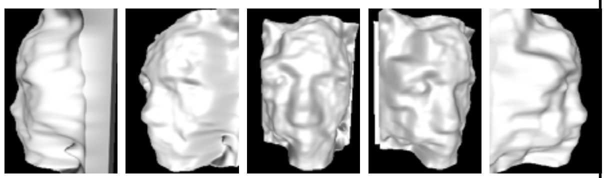
FACES FROM MONOCULAR SEQUENCES

•No calibration data

•Relatively little texture

•Difficult lighting

MAXFLOW RESULTS



PCA FACE MODEL

$$S = \bar{S} + \sum_{i=1}^{99} \alpha_i S_i$$

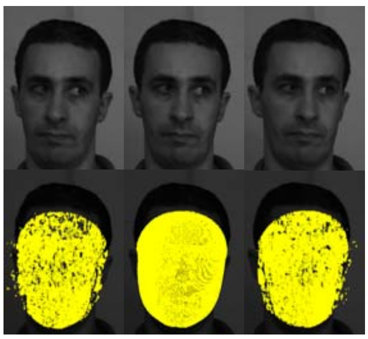
$\bar{S}$: Average shape

S_i: Shape vector

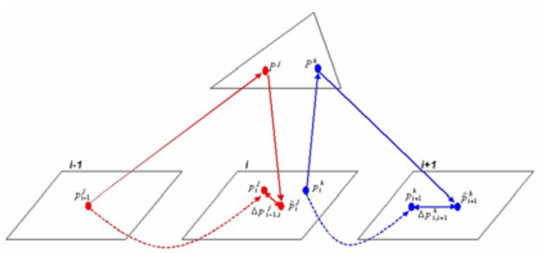
α_i: Shape coefficients

V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3-D Faces" in Computer Graphics, SIGGRAPH Proceedings, Los Angeles, CA, August 1999.

CORRESPONDENCES



TRANSFER FUNCTION



$$F_3(A, C_{i-1}, C_i, C_{i+1}) = \sum_{j \in Q_{i-1}} \left\| \Delta p^j_{i-1,i} \right\|^2 + \sum_{k \in Q_i} \left\| \Delta p^k_{i,i+1} \right\|^2$$

MODEL BASED BUNDLE ADJUSTMENT

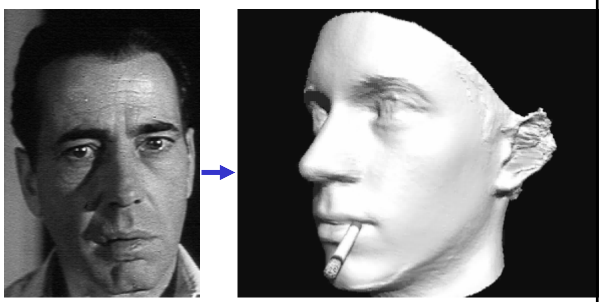
→ Median accuracy greater than 0.5mm

ROBUSTNESS TO LIGHTING CHANGES

2 3 4

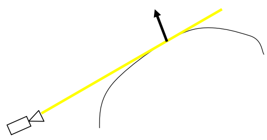
Reprojection of frame 4 into frame 3

MODEL FROM OLD MOVIE



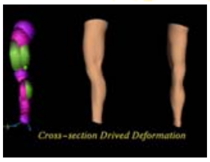
NECK AND SHOULDERS

MODELING DEFORMATIONS



Must use silhouette information
→ Express constraints on surface normals

IMPLICIT SURFACES IN COMPUTER GRAPHICS



J.F. Blinn. A Generalization of Algebraic Surface Drawing. *ACM Transactions on Graphics*, 1982.

M.P. Gascuel and A. Verroust and C. Puech. A Modeling System for Complex Deformable Bodies Suited to Animation and Collision Processing. *Journal of Visualization and Computer*, 1991.

D. Thalmann, J. Shen, and E. Chauvineau. Fast Realistic Human Body Deformations for Animation and VR Applications. In *Computer Graphics International*, June 1996.

The volumetric primitives melt like mercury drops

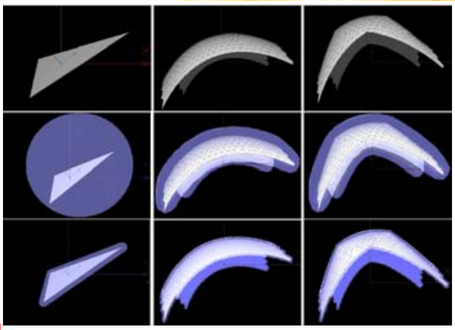
IMPLICIT VERSUS EXPLICIT SURFACES

Explicit surfaces
- Easy to deform & render

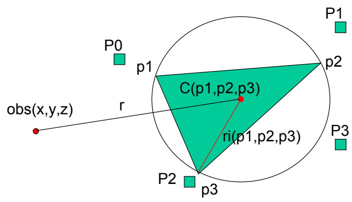
Implicit surfaces
- Direct distance evaluation
- Differentiable distance function
- Surface normals and curvatures are well defined

-> Get best of both worlds by using the explicit mesh to build an implicit surface.

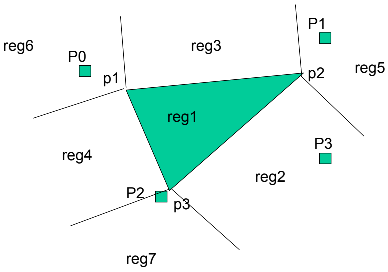
IMPLICIT MESHES



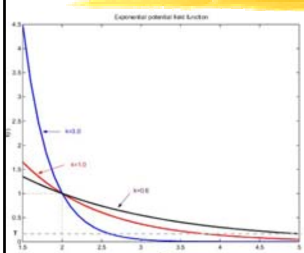
SPHERICAL METABALLS



TRIANGULAR METABALLS



POTENTIAL FIELD FUNCTION



Spherical:

$$f(r, r_i) = \exp(-k(r - d_0))$$

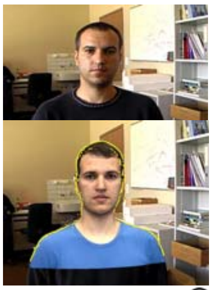
Triangular:

$$f(r, r_i) = \exp(-k(r - r_i))$$

$$F(x, y, z) = T - \sum_{i=1}^{N} f(r, r_i)$$

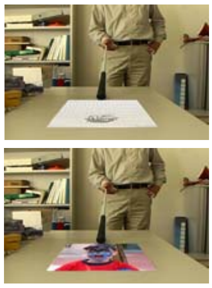
SILHOUETTES AND FEATURE POINTS

- Track feature points on the head.
- Track silhouettes on the shoulders.
- Deform neck according to both head movement and silhouette information.



TRACKING A DEFORMABLE PIECE OF PAPER

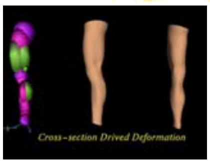
- Track feature points on page
- Fit page boundary
- Detect and use silhouettes when they appear.



FULL BODY MOTION CAPTURE

COMPLEX 3-D MOTION

IMPLICIT SURFACES IN COMPUTER GRAPHICS



J.F. Blinn. A Generalization of Algebraic Surface Drawing. *ACM Transactions on Graphics*, 1982.

M.P. Gascuel and A. Verroust and C. Puech. A Modeling System for Complex Deformable Bodies Suited to Animation and Collision Processing. *Journal of Visualization and Computer*, 1991.

D. Thalmann, J. Shen, and E. Chauvineau. Fast Realistic Human Body Deformations for Animation and VR Applications. In *Computer Graphics International*, June 1996.

The volumetric primitives melt like mercury drops

ELLIPSOIDAL METABALLS

- Each one defines a field.

$$d_i(\mathbf{x}) = \mathbf{x}^T \cdot \mathbf{Q}_i^T \cdot \mathbf{Q}_i \cdot \mathbf{x}$$

$$f_i(\mathbf{x}) = e^{-2d_i(\mathbf{x})}$$

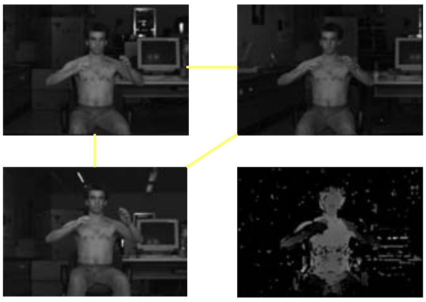
- The surface is an isosurface of their sums.

$$S = \{\mathbf{x} | F - T = 0\}, F(\mathbf{x}) = \sum_i^n f_i(\mathbf{x})$$

→ Algebraic distances of 3—D points to the surface can be computed without search and are differentiable.

→ Surface normals and curvatures can be computed both simply and exactly.

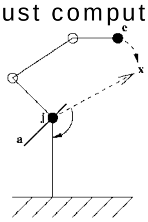
STEREO DATA



ROTATIONAL DERIVATIVES

To minimize: $F(\mathbf{x},\Theta) - T \;\rightarrow\; \min$

Must compute:

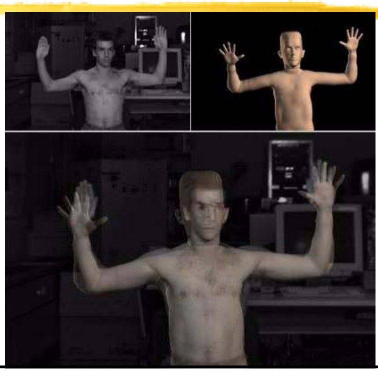


$$\frac{\partial}{\partial\theta} F(\mathbf{x},\Theta) = \frac{\partial}{\partial\theta} \sum_i^n f_i\left(d_i\left(\mathbf{x}\;,\Theta\right)\right)$$

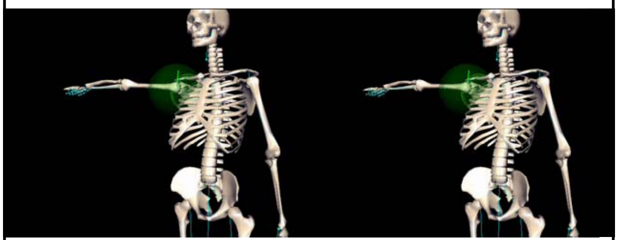
$$\frac{\partial}{\partial\theta} d(\mathbf{x},\Theta) = 2\;\mathbf{x}^T \cdot \mathbf{S}_\theta^T \mathbf{Q}_\theta^T \cdot \left[\frac{\partial}{\partial\theta}\mathbf{Q}_\theta \mathbf{S}_\theta\right] \cdot \mathbf{x}$$

$$\frac{\partial}{\partial\theta}\mathbf{Q}_\theta \mathbf{S}_\theta = \mathbf{Q}_\theta \mathbf{Rt}_{je} \vec{\mathbf{a}}_{\mathbf{j}} \times (\mathbf{x} - \mathbf{j})$$

COMPLEX 3-D MOTION

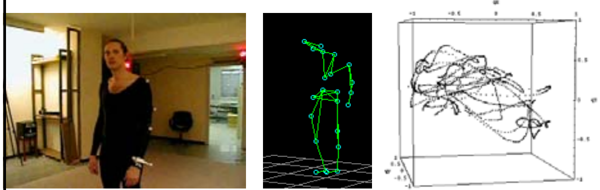


IMPOSING JOINT LIMITS



Without limits　　With limits

MEASURING JOINT LIMITS



1. Optical motion capture of allowable motions.
2. For each motion sequence, define a referential.
3. Compute rotations with respect to this referential.
4. Represent in quaternion space.

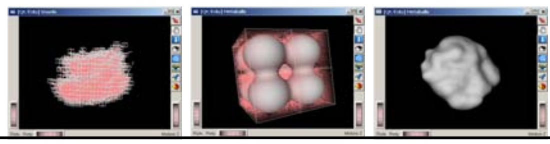
JOINT LIMITS AS IMPLICIT SURFACES

Defining an implicit surface:

$$f(P) = \sum_{i=1}^{n} f_i \qquad f_i(P) = \begin{cases} -k_i r + k_i e_i + 1 & \text{if } r \in [0, e_i] \\ \frac{1}{4}\left[k_i(r - e_i) - 2\right]^2 & \text{if } r \in [e_i, R_i] \\ 0 & \text{elsewhere} \end{cases}$$

Where $r = d(P, S_i)$ and $R_i = e_i + \frac{2}{k_i}$

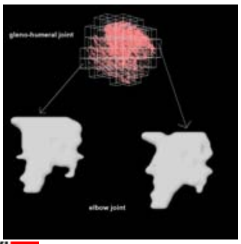
- Voxelizing the 3D quaternions.
- Placing a primitive in each voxel.
- Extracting the corresponding iso-surface.

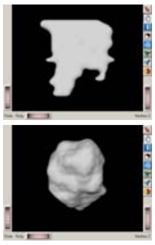


HIERARCHICAL JOINT LIMITS

For successive inter-dependent joints, create a hierarchy of joint limits:

- In each parent voxel, create a distinct implicit surface representing the joint limits of the child joint.
- Intermediate child implicit surfaces obtained by linear morphing between primitive centers and radii.

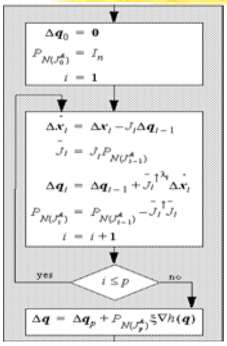




Morphed elbow joint limits.

Morphed G.H. joint limits.

HIERARCHICAL CONSTRAINTS



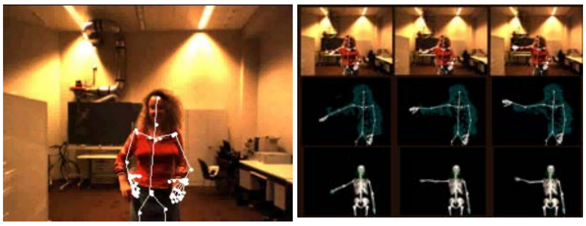
Initialization

Jacobian at level i

Updating the projector

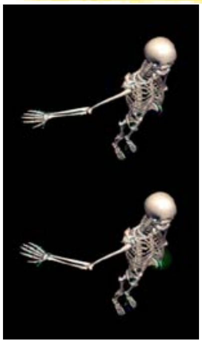
Optimization

UNCONSTRAINED TRACKING



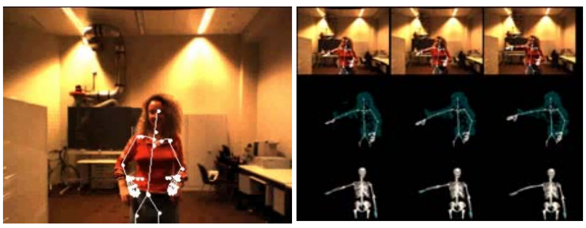
Projection roughly correct, but ..

SHE BROKE HER ARM



We prefer not to do that to our graduate students!

CONSTRAINED TRACKING

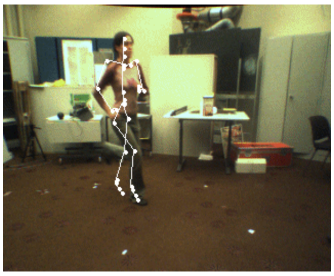


No more impossible postures.

LOW QUALITY STEREO DATA

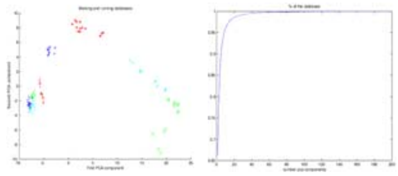
- High shutter speed to avoid motion blur
- Low resolution so that subject remains within capture volume

NO MOTION MODEL



Occlusions create problems!

WALKING AND RUNNING DATABASE

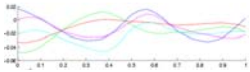


PCA decomposition of motion vectors

DETERMINISTIC MOTION MODEL

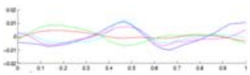
Motion model:

$$\Theta = \Theta_0 + \sum_{i=1}^{m} \alpha_i \Theta_i$$



State Vector:

$$\phi = \phi(\mu, \overrightarrow{\alpha^1}, \ldots, \overrightarrow{\alpha^T}) \text{ where } \overrightarrow{\alpha^i} = (\alpha_1^i, \ldots, \alpha_m^i)$$



Objective function:

$$F = \sum_{1 \leq t \leq T} F_t(G_t, \Theta(\mu_t, \alpha_i))$$

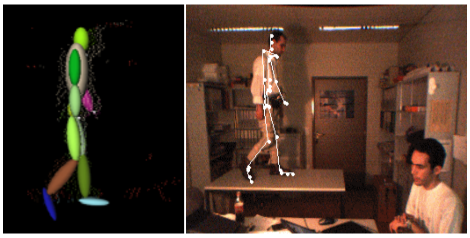
Global optimization:

$$\frac{\partial F}{\partial \alpha_i} = \sum_{j=1}^{ndof} \frac{\partial \theta_j}{\partial \alpha_i} \cdot \frac{\partial F}{\partial \theta_j}$$

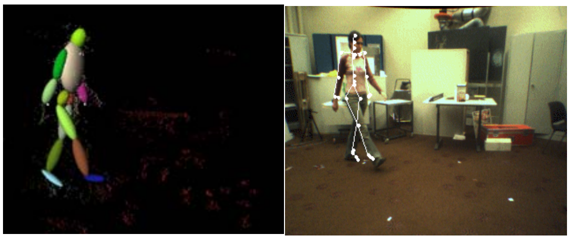
$$\frac{\partial F}{\partial \mu_t} = \sum_{j=1}^{ndof} \frac{\partial \theta_j}{\partial \mu_t} \cdot \frac{\partial F}{\partial \theta_j}$$

$$\frac{\partial \theta_j}{\partial \mu_t} = \sum_{i=1}^{m} \alpha_i \frac{\partial \Theta_{ij}}{\partial \mu_t}$$

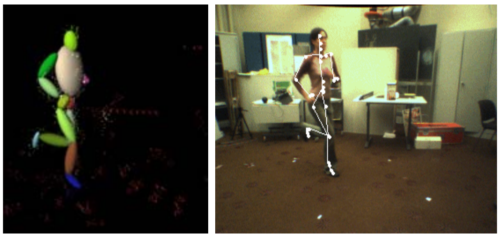
SLOW WALK



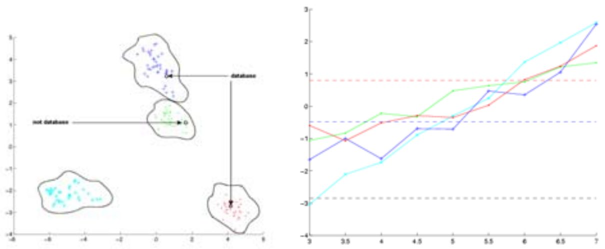
FASTER WALK



RUNNING

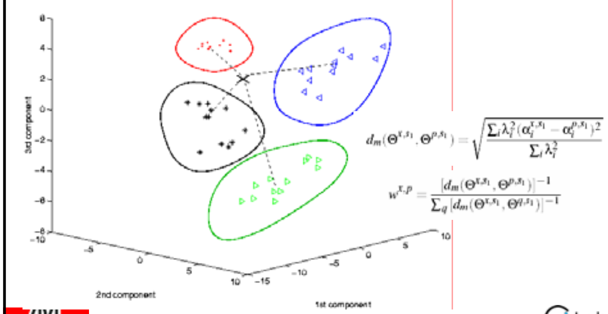


RECOGNITION



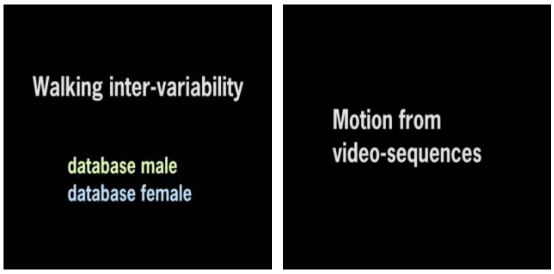
First two coefficients Fourth coefficient

ANIMATION

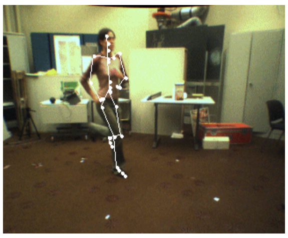


SYNTHESIZED RUNS

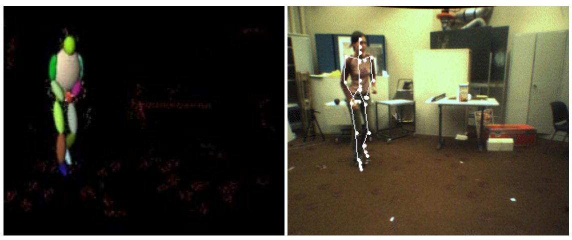
SYNTHESIZED WALKS



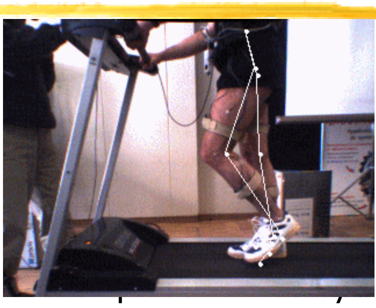
VARIABLE SPEED RUN



FROM WALKING TO RUNNING



MONOCULAR TRACKING



Correspondences and transfer function

MONOCULAR TRACKING

Correspondences and silhouettes

AUTOMATED GOLF COACH

FUTURE RESEARCH

More sophisticated

- Motion models
- Biomedical constraints

→Best possible compromise between anatomical "truth" and ease of use

→Accurate models from cheap sensors.

RELATED PUBLICATIONS

Real-time 3D tracking

- L. Vacchetti, V. Lepetit, and P. Fua. Stable real-time 3d tracking using online and offline information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004. In press.
- L. Vacchetti, V. Lepetit, and P. Fua. Combining Edge and Texture Information for Real-Time Accurate 3D Camera Tracking. In *International Symposium on Mixed and Augmented Reality*, Arlington, VA, November 2004.

Automated 3D detection

- V. Lepetit, J. Pilet, and P. Fua. Point Matching as a Classification Problem for Fast and Robust Object Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, Washington, DC, June 2004.

Face and Shoulder Modeling

- M. Dimitrijevic, S. Ilic, and P. Fua. Accurate Face Models from Uncalibrated and Ill-Lit Video Sequences. In *Conference on Computer Vision and Pattern Recognition*, Washington, DC, June 2004.
- S. Ilic and P. Fua. Generic Deformable Implicit Mesh Models for Automated Reconstruction. In *ICCV workshop on Higher-Level Knowledge in 3D Modelling and Motion Analysis*, Nice, October 2003, France.

Full Body Motion Capture

- R. Plaenkers and P. Fua. Articulated Soft Objects for Multi-View Shape and Motion Capture. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(10), 2003.
- L. Herda, R. Urtasun, and P. Fua. Hierarchical Implicit Surface Joint Limits to Constrain Video-Based Motion Capture. In *European Conference on Computer Vision, Prague*, Czech Republic, May 2004.
- R. Urtasun and P. Fua. 3--D Human Body Tracking using Deterministic Temporal Motion Models. In *European Conference on Computer Vision*, Prague, Czech Republic, May 2004.