Monocular human upper body pose estimation for sign language analysis

Nicolas Burrus <burrus@montefiore.ulg.ac.be>

Groupe ULG - INTELSIG

4th Multitel Spring Workshop, 2 June 2009
Context

3D(Stereo)Media
- Part of the Wallonian “Marshall plan”
- Motion capture
- Animation of virtual characters

Signspeak
- European project
- Automatic sign language analysis
- ULg: Focus on feature extraction: hand motion, facial expressions
Objective

Upper Body Tracking

- Provide hand position and velocity
- Head position and arm configuration
- 2D tracking
Main Difficulties

Segmentation issues

- Motion blur
- Self occlusion
- Clothing variability, etc.

Nicolas Burrus
Monocular Upper Body Pose Estimation 4/21
Motivates a top-down approach

Too many ambiguities
- Separate tracking of parts difficult
- Joint tracking is more promising

→ Multi-part statistical models

Pictorial models
- Combine structural \textit{a priori} and likelihood
- Tree-shaped models allow fast inference

Main issues
1. Likelihood and \textit{a priori} models
2. Computational complexity
Motivates a top-down approach

**Too many ambiguities**
- Separate tracking of parts difficult
- Joint tracking is more promising

→ **Multi-part statistical models**

**Pictorial models**
- Combine structural *a priori* and likelihood
- Tree-shaped models allow fast inference

**Main issues**
1. Likelihood and *a priori* models
2. Computational complexity
Tree-shaped Bayesian Model

- Each part is a square or oriented rectangle (arms)
- Parameters are \((x, y, width, [height, angle])\)
Required probabilistic quantities

Notations
- Image $I$
- Pose $L = \{L_{head}, L_{tr}, L_{lUpArm}, L_{rUpArm}, L_{lLowArm}, L_{rLowArm}, L_{lHd}, L_{rHd}\}$

A posteriori probability of a configuration $L$

$$P(L|I) \propto P(I|L) \times P(L)$$

- $P(L)$: structural a priori
- $P(I|L)$: likelihood
### Structural *a priori*

#### Decomposition thanks to tree independence

\[
P(L) \propto P(L_{tr} | L_{head}) P(L_{IUpArm} | L_{tr}) P(L_{rUpArm} | L_{tr}) \\
\times P(L_{ILowArm} | L_{IUpArm}) P(L_{rLowArm} | L_{rUpArm}) \\
\times P(L_{IHd} | L_{ILowArm}) P(L_{rHd} | L_{rLowArm})
\]

#### Chosen models

- Gaussian for junctions distances
- Uniform for orientations

![Diagram of upper body model with Gaussian and Uniform distributions indicated]
Likelihood terms

Independence assumption between parts

- \( P(I|L) \propto \prod_{i=1}^{8} P(I|L_i) \)
- One color model per part (HS histogram)
- Histogram back-projection gives per pixel likelihood
- Can be thresholded

Hand example
Likelihood issues

Intuitively: proportional to the number of white pixels

- Depend upon the rectangle size
  - Less significant for small rectangles

- Depend upon the chosen threshold
  - Less significant if the threshold is low
**A contrario reasoning**

- Based on a perceptual principle (Helmholtz)

"The lower the probability for the proportion of white pixels to be high by accident, the most significant it is."

**Concretely**

- "Accident" $H_0 = \text{pixels i.i.d. in the image}$
- $P_{H_0}(N_w \geq N_w(L_i) \mid N, \rho_w) = B_{\geq}(N_w(L_i), N, \rho_w)$
- $N$: size of the rectangle
- $N_w$: number of white pixels in the rectangle
- $\rho_w$: number of white pixels in the image
### A contrario likelihood

#### Deducing the likelihood
- \( P_{H_0}(N_w \geq N_w(L_i) \mid N, p_w) \) quantifies the significance of the white pixel concentration
- The lower it is, the higher is the probability that the concentration is not due to chance, and thus to a part

#### Final likelihood
- \[ P(I \mid L_i) \propto 1.0 - \left[ P_{H_0}(N_w \geq N_w(L_i) \mid N, p_w) \right]^\alpha \]
- \( \alpha \): quantifies how the non-accidentality increases the confidence that the part is actually there
- Can be learned
### Inference

#### Objective
- We can compute $P(L|I)$ for a given pose $L$
- How to we find the most probable one?
- The number of possible poses is too large to test them all

#### Classical solution 1: coarse discretization
- Efficient inference algorithms in trees (Belief Propagation)
- Needs to be very coarse to remain efficient

#### Classical solution 2: non-parametric belief propagation (NBP)
- Approximate all quantities by particle filters
- Accurate sampling of candidates
- Time-consuming (several minutes per frame)
### Inference

#### Objective
- We can compute $P(L|I)$ for a given pose $L$
- How to we find the most probable one?
- The number of possible poses is too large to test them all

#### Classical solution 1: coarse discretization
- Efficient inference algorithms in trees (Belief Propagation)
- Needs to be very coarse to remain efficient

#### Classical solution 2: non-parametric belief propagation (NBP)
- Approximate all quantities by particle filters
- Accurate sampling of candidates
- Time-consuming (several minutes per frame)
## Inference

### Objective
- We can compute $P(L|I)$ for a given pose $L$
- How to we find the most probable one?
- The number of possible poses is too large to test them all

### Classical solution 1: coarse discretization
- Efficient inference algorithms in trees (Belief Propagation)
- Needs to be very coarse to remain efficient

### Classical solution 2: non-parametric belief propagation (NBP)
- Approximate all quantities by particle filters
- Accurate sampling of candidates
- Time-consuming (several minutes per frame)
Discretization by importance sampling

**Overall idea**

- Use some proposal distribution $q$ to sample candidates
- Assign them a weight $\frac{p}{q}$ (importance sampling)
- Find the best pose using classical BP (max-product)

→ **Less accurate but faster than sampling from posterior**

**Different kind of proposals can be used**

- Detection-based, e.g. gaussian around a detector output
- Temporal, e.g. gaussian around the position predicted by a constant velocity model
- Structural: e.g. sample a position from a parent candidate using the *a priori* model
- Can be combined, e.g. into a mixture of gaussians
Discretization by importance sampling

Overall idea
- Use some proposal distribution \( q \) to sample candidates
- Assign them a weight \( \frac{p}{q} \) (importance sampling)
- Find the best pose using classical BP (max-product)

→ **Less accurate but faster than sampling from posterior**

Different kind of proposals can be used
- Detection-based, e.g. gaussian around a detector output
- Temporal, e.g. gaussian around the position predicted by a constant velocity model
- Structural: e.g. sample a position from a parent candidate using the *a priori* model
- Can be combined, e.g. into a mixture of gaussians
Example of structural proposal

1. Draw a parent candidate according to their weights
2. Sample a child candidate according to the *a priori* model
Preliminary results

Settings

- Color models estimated from rough manual segmentation in one frame
- No temporal term
- Detection proposal for the head
- Structural proposals for other parts
- About 200 sampled candidates per part
- Only the left parts are shown

Dataset

- NGT Corpus of sign language
- Mostly static backgrounds
Examples of frames correctly estimated
Examples of frames incorrectly estimated
Conclusion

First results are encouraging

- Less than one second per frame
- Able to find the right pose on “easy” frames

Two originalities

- *A contrario* likelihoods
  - Combine quantities in a principled way
  - Can use multiple thresholds to increase robustness
- Discretization by importance sampling + BP
  - Focus on promising regions
  - Can integrate various heuristics
  - Inference remains efficient
## Perspectives

### Quantitative evaluation
- Need for a labelled database
- Comparison with existing approaches

### Improve the model
- Temporal terms
- Contour-based terms
- Learn parameters
- Handle self-occlusions explicitly to improve likelihood
- Automatic color model initialization
- Language analysis predictions
Questions?