Particle filtering for audio-visual and audio tracking

Heidi Christensen
(h.christensen@dcs.shef.ac.uk)

Speech and Hearing (SPandH) - University of Sheffield, United Kingdom

28. nov 2006, Pop meeting, Coimbra, Portugal
Particle filtering for audio-visual tracking

- Data fusion with particle filtering
- Hearing humanoid robots
- Smart meeting rooms
Outline

1. Particle filtering for audio-visual tracking
   - Data fusion with particle filtering
   - Hearing humanoid robots
   - Smart meeting rooms

2. Particle filtering for audio only tracking
   - Speech: enhancement and noise tracking
   - Music audio analysis: tempo and score tracking
   - Digital communication
   - Localization: microphone arrays; binaural microphones
Outline

1. Particle filtering for audio-visual tracking
   - Data fusion with particle filtering
   - Hearing humanoid robots
   - Smart meeting rooms

2. Particle filtering for audio only tracking
   - Speech: enhancement and noise tracking
   - Music audio analysis: tempo and score tracking
   - Digital communication
   - Localization: microphone arrays; binaural microphones

3. Current work
   - Localization using high-frequency regions
Outline

1. Particle filtering for audio-visual tracking
   - Data fusion with particle filtering
   - Hearing humanoid robots
   - Smart meeting rooms

2. Particle filtering for audio only tracking
   - Speech: enhancement and noise tracking
   - Music audio analysis: tempo and score tracking
   - Digital communication
   - Localization: microphone arrays; binaural microphones

3. Current work
   - Localization using high-frequency regions

4. Future work
   - Improving PF modelling
Outline

1. Particle filtering for audio-visual tracking
   - Data fusion with particle filtering
   - Hearing humanoid robots
   - Smart meeting rooms

2. Particle filtering for audio only tracking
   - Speech: enhancement and noise tracking
   - Music audio analysis: tempo and score tracking
   - Digital communication
   - Localization: microphone arrays; binaural microphones

3. Current work
   - Localization using high-frequency regions

4. Future work
   - Improving PF modelling

5. Current work - Yan-Chen
   - PF for auditory depth estimation
Outline

1 Particle filtering for audio-visual tracking
   - Data fusion with particle filtering
   - Hearing humanoid robots
   - Smart meeting rooms

2 Particle filtering for audio only tracking
   - Speech: enhancement and noise tracking
   - Music audio analysis: tempo and score tracking
   - Digital communication
   - Localization: microphone arrays; binaural microphones

3 Current work
   - Localization using high-frequency regions

4 Future work
   - Improving PF modelling

5 Current work - Yan-Chen
   - PF for auditory depth estimation
Advantages of PF for data fusion

- PF can propagate more general distributions (such as underlying audio and visual measurements)
- Allows the information from different measurement sources to be fused in a principled manner.
Data fusion for Visual Tracking with Particles

- A single (static) camera
- A stereo microphone pair

Audio–Visual Setup

Desktop teleconferencing task
Cues

- **Color** cues are persistent and robust, but prone to ambiguity.
- **Sound** and **motion** cues are intermittent, but very discriminant when present.

A likelihood model is constructed based on each cue.
Assume we have $M$ measurements sources

- Instantaneous measurement vector is $z = (z^1 \cdots z^M)$.
- Assume the measurements are conditionally independent ....
PF for multiple measurements

[Pe`rez04]

- Assume we have $M$ measurements sources
- Instantaneous measurement vector is $z = (z^1 \cdots z^M)$.
- Assume the measurements are conditionally independent ....

**Independent assumption**

“... can be justified in the light that any correlation that may exist between the color, motion, and sound of an object is likely to be weak”
Assume we have \( M \) measurements sources.

Instantaneous measurement vector is \( z = (z^1 \cdots z^M) \).

Assume the measurements are conditionally independent.

**Independent assumption**

“... can be justified in the light that any correlation that may exist between the color, motion, and sound of an object is likely to be weak”

So, we make life easier and factorise the likelihood:

\[
p(z|x) = \prod_{m=1}^{M} p(z^m|x).
\]
Objective: obtaining \( \{x_k^{(i)}, w_k^{(i)}\}_{i=1:N} \) from \( \{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1:N} \)

**Generic Particle filter**

1. **Sampling**
   \[ x_k^{(i)} \sim \pi(x_k \mid x_0^{(i)}, z_{1:k}) \]

2. **Calculation of weights**
   \[ w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(z_k \mid x_k^{(i)}) \cdot p(x_k^{(i)} \mid x_{k-1}^{(i)})}{\pi(x_k^{(i)} \mid x_0^{(i)}, z_{1:k})} \]
   \( \sum_{i=1:N} w_k^{(i)} = 1 \)

3. **State estimation**
   \[ \mathbb{E} p(x_k \mid z_{1:k})[x_k] \approx \sum_{i=1}^N w_k^{(i)} x_k^{(i)} \]

4. **Resampling**
Assume that the state evolution and proposal distributions decompose as

\[ p(x_k|x_{k-1}) = \int p_M(x_k|x_{M-1}) \ldots p_1(x_1|x_{k-1}) \, dx^1 \ldots dx^{M-1} \]

\[ \pi(x_k|x_{0:k-1}, z) = \int \pi_M(x_k|x_{M-1}, z^M) \ldots \pi_1(x_1|x_{0:k-1}, z^1) \, dx^1 \ldots dx^{M-1} \]

where \( x^1 \ldots x^{M-1} \) are “auxiliary” state vectors.

This simply amount to splitting the original evolution model into \( M \) successive intermediary steps.
A generic importance sampling mechanism for data fusion is introduced.

**Approach ... in words ...**

- Having split into this abstract state space, kind of imagine going from previous state through intermittent $m$ modality (ie. ‘colour-cue’) state.
- Each layer / “auxiliary” state constitutes of a sampling of a filtered version of the proposal function, $\pi$.
- Basically add an extra loop in the algorithm for $m \in \{1 \ldots M\}$. 
Layered sampling approach to fuse $M$ modalities

[Pe'rez04]

Advantage of ‘layered sampling’ approach

- No obvious advantage over standard PF technique
- True benefit in cases where the measurement modalities differ in the level of information they provide about the state.
Layered sampling approach to fuse $M$ modalities

[Pérez04]

**Advantage of ‘layered sampling’ approach**

- No obvious advantage over standard PF technique
- True benefit in cases where the measurement modalities differ in the level of information they provide about the state.

**Algorithm**

- Layer modalities in order from coarse to fine; layered sampling approach will guide the search.
- Each state refining the search space.
Proposal function for the sound cue

Sound cue is based on Time Difference On Arrival (TDOA) calculations; Difference between left and right microphone.

Stereo sound signal

Generalized Cross–Correlation Function

Likelihood as a function of TDOA

Likelihood as a function of image horizontal position
Pérez et al. conclude:

- The combination of cues proved to be more robust than the cues individually.
- Layered sampling approach facilitates more efficient exploration of the state space.
- Event-based propagation models are essential for the detection of new objects.
Hearing Humanoid Robots
- tracking auditory events

Cog
MIT

SIG/SIG2
Kitano Symbiotic Systems Project, Japan

ASIMO
HONDA

Robcub
Univ. of Stockholm

[Irie95] [Kim06] [Nakadai06] [Gustavsson06]
Smart Meeting Rooms
- Task: Speaker Diarization ("who spoke when?")

Example of equipped meeting rooms

[Gatica-Perez03], [Lathoud03], [Juby03], [Wrigley05], [Anguerra06]
Smart Meeting Rooms
- Task: Speaker Diarization (“who spoke when?”)

Example of equipped meeting rooms

[Gatica-PerezXX], [Lathoud03], [Juby03], [Wrigley05], [Anguerra06]

PF formulation
- Observation: audio signal, video features (colour, contours etc.)
- Hidden state: physical position of person and speaking/not-speaking parameter.
Smart Meeting Rooms
- Task: Speaker Diarization ("who spoke when?")

Room equipment

- Binaural Microphones
- Lapel Microphones
- Microphone Arrays (Note the swiss-style cheese board mounting ...)

Heidi Christensen (SPandH)

Particle Filtering for AV and A tracking
28. nov 2006, Coimbra
Outline

1. Particle filtering for audio-visual tracking
   - Data fusion with particle filtering
   - Hearing humanoid robots
   - Smart meeting rooms

2. Particle filtering for audio only tracking
   - Speech: enhancement and noise tracking
   - Music audio analysis: tempo and score tracking
   - Digital communication
   - Localization: microphone arrays; binaural microphones

3. Current work
   - Localization using high-frequency regions

4. Future work
   - Improving PF modelling

5. Current work - Yan-Chen
   - PF for auditory depth estimation
Speech enhancement [Vermaak02]

- Aim: learn/track model describing speech; then used to clean up speech.
- Speech signal is modelled as a time-varying autoregressive (TVAR) process.
- Parameters of model are stochastically evolving.

PF formulation

PF is used for online smoothed estimates of clean speech signal and model parameters.
- Observation: speech sample.
- Hidden state: TVAR model parameters.
Aim: extract information about the meter or tempo of a musical signal; "an algorithm which taps it’s hypothetical foot along to the music ...".

A low level task which humans can perform to many types of music.

Combined score and tempo tracking [Cemgil02]

“chicken-and-egg” problem: known score makes for easier tempo estimation, and vice-versa.

PF formulation

- Observation: onset observation
- Hidden state: tempo and score; two hidden variables.
Digital communication

Many problems can be represented as dynamic state-space models that involve nonlinear functions and non-Gaussian noise.
Good review: [Djurić02]

Blind equalization

- used when transmitting digital symbols over dispersive channel ("blind" refers to the unknown channel impulse response)
- objective is to estimate the transmitted symbols in presence of intersymbol interference.

PF formulation

- Observation: received signal.
- Hidden state: transmitted symbols and channel coefficients.
Reminder of simple localization system:

- Basilar Membrane Response
- Cross-Correlation

Advantages of using PF:
- Problem: multipath propagation of acoustic waves constitutes a major challenge.
- Advantage 1 of PF: location estimates are based on a series of past measurements; minimises spurious effects.
- Advantage 2 of PF: target dynamics are specified.

Chosen location estimate (cross-correlation or beamforming) is nonlinear and/or non-Gaussian.
Reminder of simple localization system:

Advantages of using PF

- Problem: multipath propagation of acoustic waves constitutes a major challenge.
Acoustic source localization and tracking

Reminder of simple localization system:

Advantages of using PF

- Problem: multipath propagation of acoustic waves constitutes a major challenge.
- Advantage 1 of PF: location estimates are based on a series of past measurements; minimises spurious effects.
Reminder of simple localization system:

Advantages of using PF

- Problem: multipath propagation of acoustic waves constitutes a major challenge.
- Advantage 1 of PF: location estimates are based on a series of past measurements; minimises spurious effects.
- Advantage 2 of PF: target dynamics are specified.
Reminder of simple localization system:

Advantages of using PF

- Problem: multipath propagation of acoustic waves constitutes a major challenge.

- Advantage 1 of PF: location estimates are based on a series of past measurements; minimises spurious effects.

- Advantage 2 of PF: target dynamics are specified.

- Chosen location estimate (cross-correlation or beamforming) is nonlinear and/or non-Gaussian.
Using a microphone array

Idea: to direct a beamformer in all possible directions and look for maximal output.
Using a microphone array

Idea: to direct a beamformer in all possible directions and look for maximal output.

PF formulation \cite{Legmann06} \cite{Ward02}

- Observation: location estimate.
- Hidden state: position of sound source (azimuth, distance, or 3D coordinates).
Particle filtering for location estimation

Two/Binaural microphones

Mimicking the human ’configuration’.
Particle filtering for location estimation

Two/Binaural microphones
Mimicking the human ’configuration’.

PF formulation
- Observation : a function of the cross-correlation of left and right ear microphone
- Hidden state: position/location of target.
Outline

1. Particle filtering for audio-visual tracking
   - Data fusion with particle filtering
   - Hearing humanoid robots
   - Smart meeting rooms

2. Particle filtering for audio only tracking
   - Speech: enhancement and noise tracking
   - Music audio analysis: tempo and score tracking
   - Digital communication
   - Localization: microphone arrays; binaural microphones

3. Current work
   - Localization using high-frequency regions

4. Future work
   - Improving PF modelling

5. Current work - Yan-Chen
   - PF for auditory depth estimation
Localization using High-Frequency Regions

- high-frequency region (ambiguous)
- low-frequency region (blurry)
Localization using High-Frequency Regions

Problem:
- in theory primary contour should be biggest peak
- in practise secondary contours will dominate in many frames

Most common solution:
- Just ignore high-frequency regions ....
Localization using high-frequency regions

Secondary traces are important

- They carry localization information
Localization using high-frequency regions

Secondary traces are important

They carry localization information

and they provide disambiguating information in multiple speaker scenarios
Localization using high-frequency regions

Secondary traces are important

Solutions to high-frequency ambiguity:

- Just ignore high-frequency regions ....
- Sum along secondary contours as well (stencil method [Liu00]).
- Apply PF techniques and incorporate primary and secondary contours in model.
Sum high-frequency cross-correlogram channels and extract highest peak position.
PF on high-frequency regions

current observations….

- Sum high-frequency cross-correlogram channels and extract **highest** peak position.

Max peak position, channel 64

Cross–Correlation
PF on high-frequency regions

current observations....

HISTOGRAM OF MAX PEAK POSITIONS

Max peak occurred at secondary contour
Max peak occurred at primary contour
PF on high-frequency regions

Current system model ....

\[
\begin{align*}
p(x_0, x_0) &= P \\
p(x_0, x_{\{1,2,3\}}) &= (1-P)/3
\end{align*}
\]
Task: diarization

Example of meeting diarization annotation annotated with start and end point labels for each speaker.
Data and task

Data: “meeting data”

- Available binaural meeting recordings are too noisy to start with.
- Mixed meeting-like data from recordings of spatialised data; Speaker turn statistics taken from real recorded meetings.
Outline

1. Particle filtering for audio-visual tracking
   - Data fusion with particle filtering
   - Hearing humanoid robots
   - Smart meeting rooms

2. Particle filtering for audio only tracking
   - Speech: enhancement and noise tracking
   - Music audio analysis: tempo and score tracking
   - Digital communication
   - Localization: microphone arrays; binaural microphones

3. Current work
   - Localization using high-frequency regions

4. Future work
   - Improving PF modelling

5. Current work - Yan-Chen
   - PF for auditory depth estimation
PF on high-frequency regions
possible likelihood functions ...

We’re looking for $P(\text{obs}|\text{state})$. Derive from peak position histograms.
Improve observation

- use multiple peak positions; one for each channel.
Improve observation

- use multiple peak positions; one for each channel.
- use first, second and third highest peak position to capture more of curtain contours.
Improve observation

- use multiple peak positions; one for each channel.
- use first, second and third highest peak position to capture more of curtain contours.
- use cross-correlation peak value as well as peak position; energy level carries information.
Improve system model

System model
No duration minimum

Current speaker 0
Other speaker 2
Other speaker 1
Other speaker 3

System model
With duration minimum

Current speaker 0
Other speaker 3
Other speaker 2
Other speaker 1

Heidi Christensen (SPandH)

Particle Filtering for AV and A tracking
28. nov 2006, Coimbra
Outline

1. Particle filtering for audio-visual tracking
   - Data fusion with particle filtering
   - Hearing humanoid robots
   - Smart meeting rooms

2. Particle filtering for audio only tracking
   - Speech: enhancement and noise tracking
   - Music audio analysis: tempo and score tracking
   - Digital communication
   - Localization: microphone arrays; binaural microphones

3. Current work
   - Localization using high-frequency regions

4. Future work
   - Improving PF modelling

5. Current work - Yan-Chen
   - PF for auditory depth estimation
... powerpoint interlude ...