Frank Deinzer\textsuperscript{1}, Joachim Denzler\textsuperscript{2}

On Fusion of Multiple Views for Active Object Recognition

\textsuperscript{2}Lehrstuhl Digitale Bildverarbeitung
University Jena

\textsuperscript{1}Siemens Medical Solutions
Forchheim
Overview

- Motivation
- Context of the work
- Fusion of single views
  - Fusion using particle filters
- Experiments
- Conclusion
Motivation I

Typical problems

- Objects cannot be classified using a single view
- Objects cannot be localized using most of the view
Motivation I

Typical problems

- Objects cannot be classified using a single view

Main goal of active object recognition

- Select most appropriate view
- Fuse the collected views to get an improved result
Overall Framework

- Optimal sensor data acquisition for state estimation
Overall Framework

- Information theoretic approach

\[ p(x) \xrightarrow{\text{channel parameter } a} p(o | x, a) \xrightarrow{\text{observation } o} p(x | o, a) \]

**MMI:**
\[
a = \arg\max_{a} \int \int p(x) p(o | x, a) \log \left( \frac{p(o | x, a)}{p(o | a)} \right) dodx
\]

**BAYES:**
\[
p(x | o, a) = \frac{p(o | x, a) p(x)}{p(o | a)}
\]

Fusion is an important step for active viewpoint selection.

In general: improved classification and localization by fusion of $N$ views.
Back to the Problem

- Fusion is an important step for active viewpoint selection.
- In general: improved classification and localization by fusion of N views.
Modelling of the State

- Probabilistic modeling needs multivariate densities with
  - Continuous parameters of the pose \( q \)
Modelling of the State

- Probabilistic modeling needs multivariate densities with
  - Continuous parameters of the pose
  - Discrete class label

\[ p(q|f) \]
Representation of Densities

\[ p(\mathbf{q} \mid f) \]

view \( f \)
Representation of Densities

- Densities in general multimodal and difficult to describe in an analytical manner
- But in general: we can evaluate the density.

⇒ Represent such multimodal densities using particle sets
Condensation

Fusion of multiple views using Condensation algorithm

cup 1A
cup 2A
cup 1B
cup 2B
Condensation

Fusion of multiple views using Condensation algorithm

Central idea:

**Common** treatment of the problems *classification* and *localization*

⇒ sensible definition of particles

⇒ sensible prediction of particles over time (motion model used in the Condensation)
Condensation

- **Particle** consists of class and pose
  - 2 degrees of freedom of the camera movement
    \[
    s_i^n = \begin{pmatrix} \Omega \kappa & \text{pose}_i^n \end{pmatrix}^T = \begin{pmatrix} \Omega \kappa & \delta_i^n & \alpha_i^n \end{pmatrix}^T
    \]
  - 1 degree of freedom of the camera movement
    \[
    s_i^n = \begin{pmatrix} \Omega \kappa & \text{pose}_i^n \end{pmatrix}^T = \begin{pmatrix} \Omega \kappa & \delta_i^n \end{pmatrix}^T
    \]

- Particle \( s_1^0, s_2^0, \ldots, s_K^0 \) are initially (\( n=0 \)) uniformly distributed over classes and poses
Condensation

- **Particle** consists of class and pose
  - 2 degrees of freedom of the camera movement
    \[
    s^n = \begin{pmatrix} O & \text{pose}^n \end{pmatrix}^T = \begin{pmatrix} O & \delta^n & \alpha^n \end{pmatrix}^T
    \]
  - 1 degree of freedom of the camera movement

- Particles represent a mixed continuous/discrete approximation of the density over the state space \( q \)

- Particle \( s_1^0, s_2^0, \ldots, s_K^0 \) are initially ( \( n=0 \) ) uniformly distributed over classes and poses
Condensation

- **Prediction** of the particles
  - class remains constant
  - Pose is predicted according to the *motion* of the camera and *noise* is added
  - 2 degrees of freedom

\[
\mathbf{s}_i^n = \mathbf{s}_i^{n-1} + \begin{pmatrix} 0 \\ \Delta\phi_t \\ \Delta\phi_a \end{pmatrix} \quad \text{with} \quad \Delta\phi_t \sim N(\phi_t, \sigma_t) \quad \Delta\phi_a \sim N(\phi_a, \sigma_a)
\]

- 1 degree of freedom

\[
\mathbf{s}_i^n = \mathbf{s}_i^{n-1} + \begin{pmatrix} 0 \\ \Delta\phi \end{pmatrix} \quad \text{with} \quad \Delta\phi \sim N(\phi, \sigma)
\]
Condensation

- **Prediction** of the particles
  - class remains constant
  - Pose is predicted according to the *motion* of the camera and *noise* is added

  in principle known motion between two views but some uncertainty remains

  Treated by the motion model of the condensation algorithm

  Uncertainty given from „outside“

\[
\mathbf{s}_i^n = \mathbf{s}_i^{n-1} + \begin{pmatrix} 0 \\ \Delta\phi \end{pmatrix} \quad \text{with} \quad \Delta\phi \sim N(\phi, \sigma)
\]
Condensation

- Judgment $p_i^n$ of particles of the particle set

$$S^n = \{(s_1^n, p_1^n), \ldots, (s_K^n, p_K^n)\}$$

can be easily done during classification by

$$p_i^n = p(f_n \mid \Omega^*_i, \text{Lage}_i^n)$$

- Generate a new particle set $S^n$ by sampling from the particle set $S^{n-1}$

- Parameters within the Condensation
  - Size of the particle set
  - Noise model in the motion model
  - Sampling method
Condensation

- Judgment $p_i^n$ of particles of the particle set

$$S^n = \left\{ \left( \mathbf{s}_1^n, p_1^n \right), \ldots, \left( \mathbf{s}_K^n, p_K^n \right) \right\}$$

can be easily done during classification by

$$p_i^n = p \left( f_n \mid \Omega_\kappa, \text{Lage}_i^n \right)$$

- Parameters within the Condensation
  - Size of the particle set
  - Noise model in the motion model
  - Sampling method

Judgment of the particles represent the quality of classification and localization results
Condensation – Examples

Objects that cannot be classified using one view

- cup 1A
- cup 2A
- cup 1B
- cup 2B
Condensation – Example

Object cannot be localized for most of the views

Tasse 1  Tasse 2  Tasse 3  Tasse 4
Experiments

- Evaluation using 25 sequences with 10 random views per object

<table>
<thead>
<tr>
<th>Number of fused views</th>
<th>N=1</th>
<th>N=2</th>
<th>N=5</th>
<th>N=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rate</td>
<td>35 %</td>
<td>59 %</td>
<td>69 %</td>
<td>71 %</td>
</tr>
</tbody>
</table>
Experiments

Influence of the number $K$ of particles on the recognition rate
Experiments

Influence of the number $K$ of particles on the recognition rate

**Computation time** (1GHz Athlon) for one fusion step

- 1.8 seconds for $K=43400$
- 0.15 seconds for $K=3500$
  $\Rightarrow$ Fusion of 7 views per second

**Time for acquiring image** due to the mechanics (robot arm) much higher
Experiments

Influence of the noise model (modeling the uncertainty in the motion of the camera)

\[
\sigma_t = 1.8^\circ, \sigma_a = 1.5^\circ \\
\sigma_t = 7.2^\circ, \sigma_a = 6^\circ \\
\sigma_t = 0.36^\circ, \sigma_a = 0.3^\circ \\
\sigma_t = 0.02^\circ, \sigma_a = 0.03^\circ
\]
Experiments

Results for localization error

![Graph showing localization error over N values]
Conclusion

- Fusion of views using Condensation
  - Common treatment of classification and localization
  - Scaleable (number of particles) up to real time capability
  - Improved recognition rate

- Not tackled in the talk (since fusion was the topic)
  - How to get the next best viewpoint (or \( n \) next best viewpoints)