A System for Appearance-Based Probabilistic 3D Object Recognition and Its Applications

13/12/2007, Knowledge Engineering Group
University of Economics, Prague

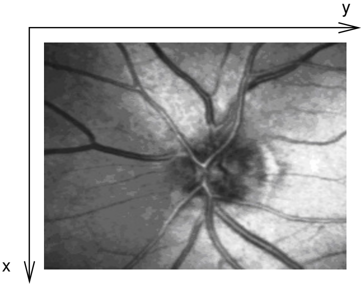
Marcin Grzegorzek

Multimedia & Vision Research Group
Queen Mary, University of London, UK

Overview

Overview

Digital Representation of Gray Level Images



$$\longrightarrow \begin{pmatrix} 039 & 041 & 041 & 043 & \dots & 057 \\ 041 & 043 & 043 & 043 & \dots & 066 \\ 041 & 051 & 053 & 055 & \dots & 066 \\ 041 & 051 & 053 & 056 & \dots & 066 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 054 & 052 & 042 & 042 & \dots & 095 \end{pmatrix}_{150 \times 200}$$

$$f(x, y) = \begin{pmatrix} f(0,0) & f(0,1) & \dots & f(0,199) \\ f(1,0) & f(1,1) & \dots & f(1,199) \\ \vdots & \vdots & & \vdots \\ f(149,0) & f(149,1) & \dots & f(149,199) \end{pmatrix} \quad ; \quad f(x, y) \in \{0, 1, 2, \dots, 255\}$$

Digital Representation of Color Images

$$
\boldsymbol{R} = \begin{pmatrix} 178 & 182 & 182 & \dots & 185 \\ 179 & 180 & 182 & \dots & 185 \\ 179 & 181 & 182 & \dots & 184 \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 032 & 034 & 037 & \dots & 111 \end{pmatrix}_{160 \times 240}
$$

$$
\boldsymbol{G} = \begin{pmatrix} 067 & 069 & 069 & \dots & 066 \\ 067 & 066 & 070 & \dots & 067 \\ 068 & 067 & 069 & \dots & 067 \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 101 & 104 & 104 & \dots & 098 \end{pmatrix}_{160 \times 240}
$$

$$
\boldsymbol{B} = \begin{pmatrix} 189 & 188 & 188 & \dots & 188 \\ 190 & 189 & 190 & \dots & 189 \\ 190 & 188 & 190 & \dots & 189 \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 198 & 200 & 200 & \dots & 065 \end{pmatrix}_{160 \times 240}
$$

Digital Video Representation

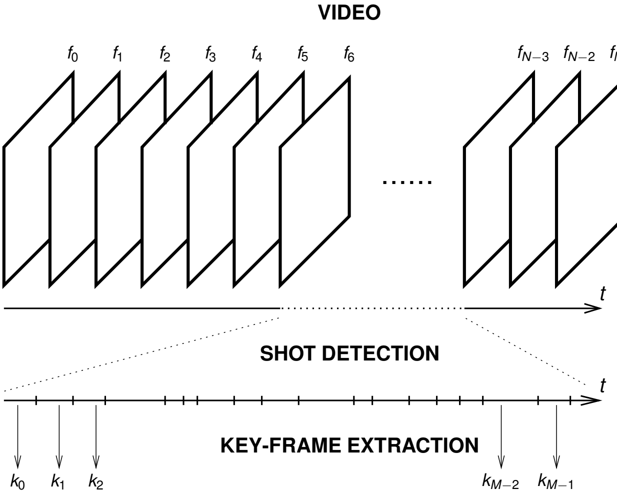
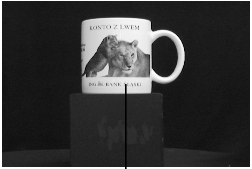


Image Acquisition - Stochastic Process



$f(120, 180) = 219$

$f(120, 180) = 210$

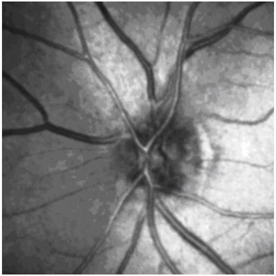
$f(120, 180) = 208$

$f(120, 180) = 204$

$f(120, 180) = 198$

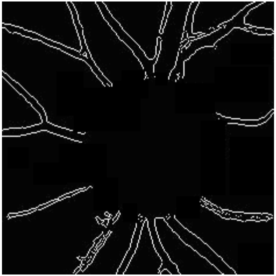
Digital Image Processing

Input Image $f = f(x, y)$

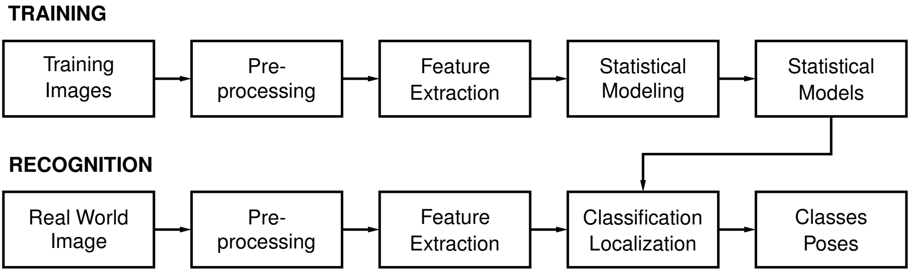


$$\mathrm{T}\{f\} = g$$

Output Image $g = g(x, y)$



Object Recognition Task



Classification – Which objects occur in the image?

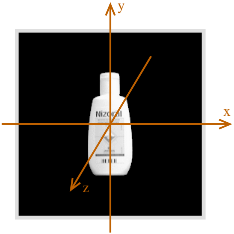
Localization – In which poses known objects occur in the image?

Recognition – Which objects and in which poses occur in the image?

Object Pose

$\boldsymbol{t}_{int} = (t_x, t_y)^{\mathrm{T}}$

$\phi_{int} = \phi_z$



$\boldsymbol{t} = (t_x, t_y, t_z)^{\mathrm{T}} = (0, 0, 100)^{\mathrm{T}}$
$\phi = (\phi_x, \phi_y, \phi_z)^{\mathrm{T}} = (0, 0, 0)^{\mathrm{T}}$

$\boldsymbol{t} = (50, 25, 100)^{\mathrm{T}}$
$\phi = (0, 0, 0)^{\mathrm{T}}$

$\boldsymbol{t} = (50, 25, 100)^{\mathrm{T}}$
$\phi = (0, 0, -30)^{\mathrm{T}}$

$t_{ext} = t_z$

$\phi_{ext} = (\phi_x, \phi_y)^{\mathrm{T}}$

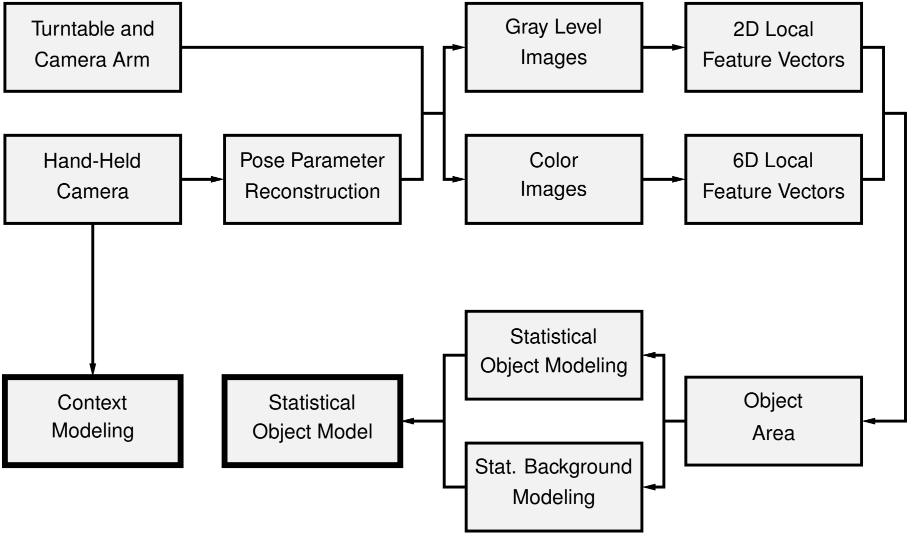
$\boldsymbol{t} = (0, 0, 100)^{\mathrm{T}}$
$\phi = (22.5, 0, 0)^{\mathrm{T}}$

$\boldsymbol{t} = (0, 0, 100)^{\mathrm{T}}$
$\phi = (0, 45, 0)^{\mathrm{T}}$

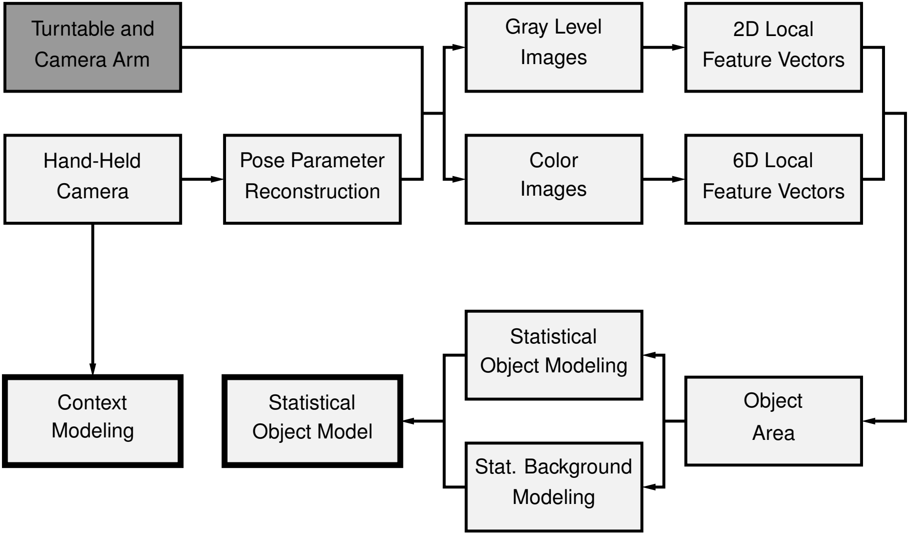
$\boldsymbol{t} = (0, 0, 80)^{\mathrm{T}}$
$\phi = (0, 0, 0)^{\mathrm{T}}$

Overview

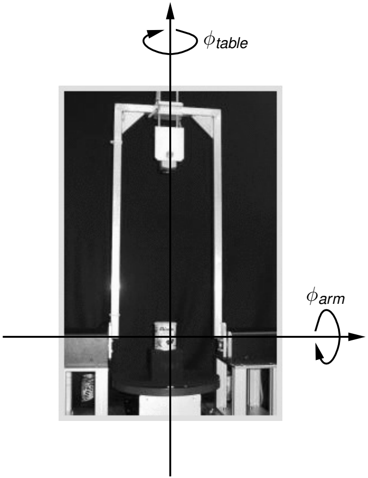
Training Phase

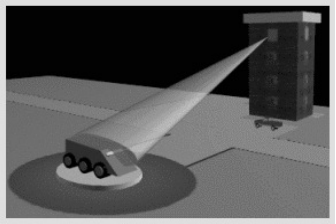


Turntable and Camera Arm



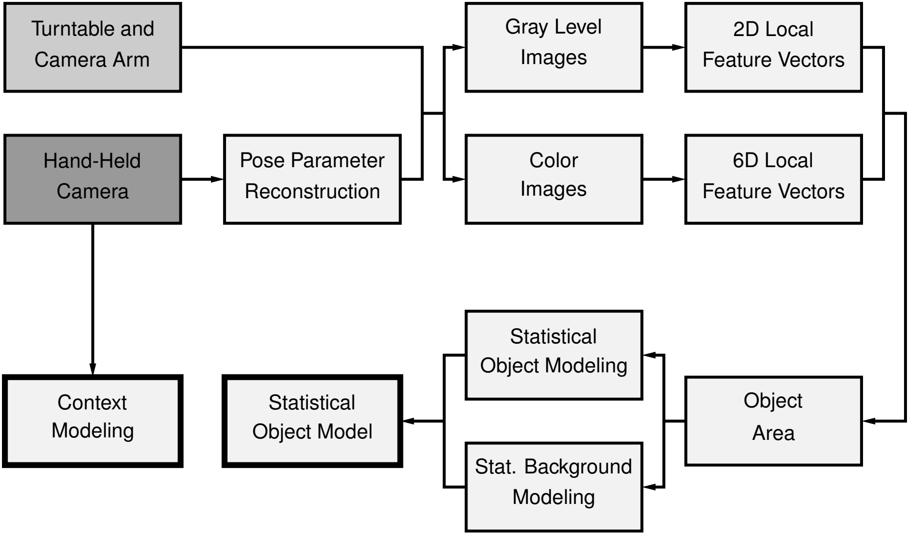
Turntable and Camera Arm





Object poses $(\phi_\rho, \boldsymbol{t}_\rho)$ for all N_ρ training images $\boldsymbol{f}_{\rho=1,\ldots,N_\rho}$ are known.

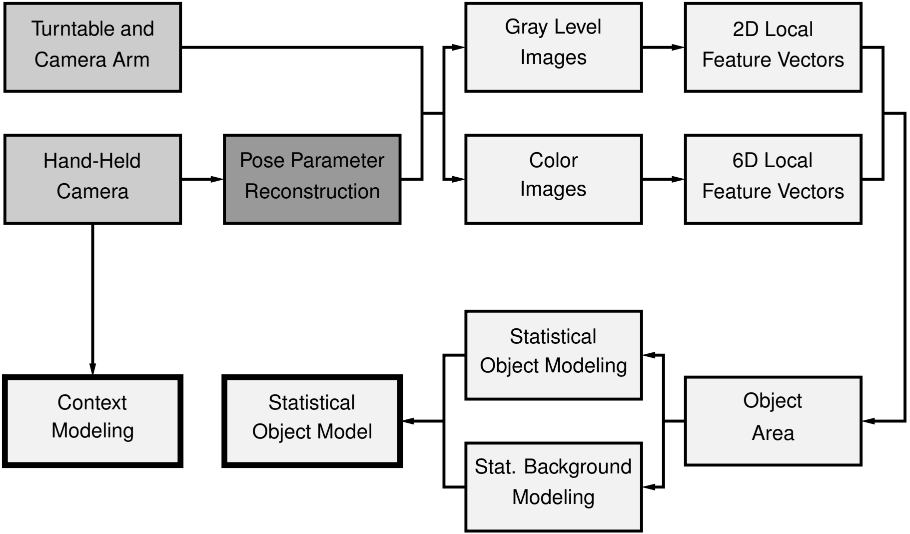
Hand-Held Camera



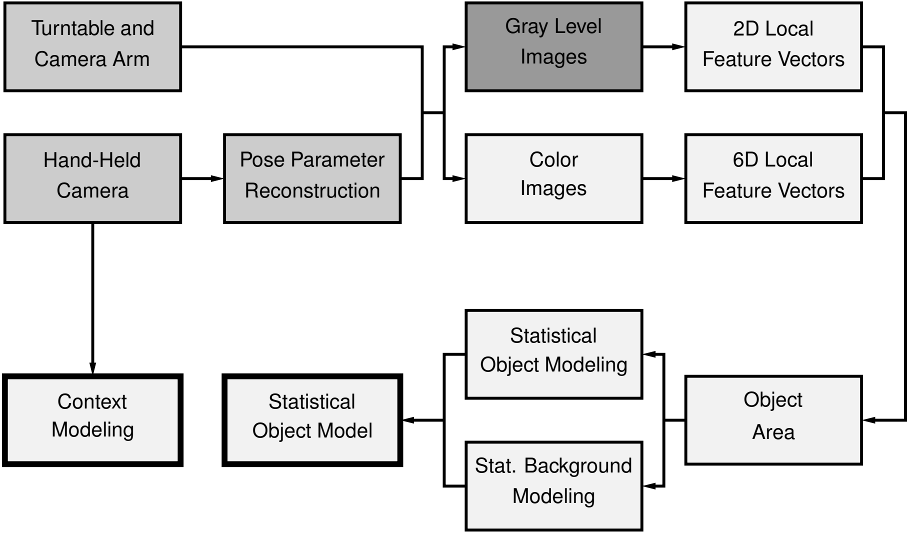
Hand-Held Camera

Object poses $(\phi_\rho, \boldsymbol{t}_\rho)$ for the training images $\boldsymbol{f}_\rho$ are unknown.

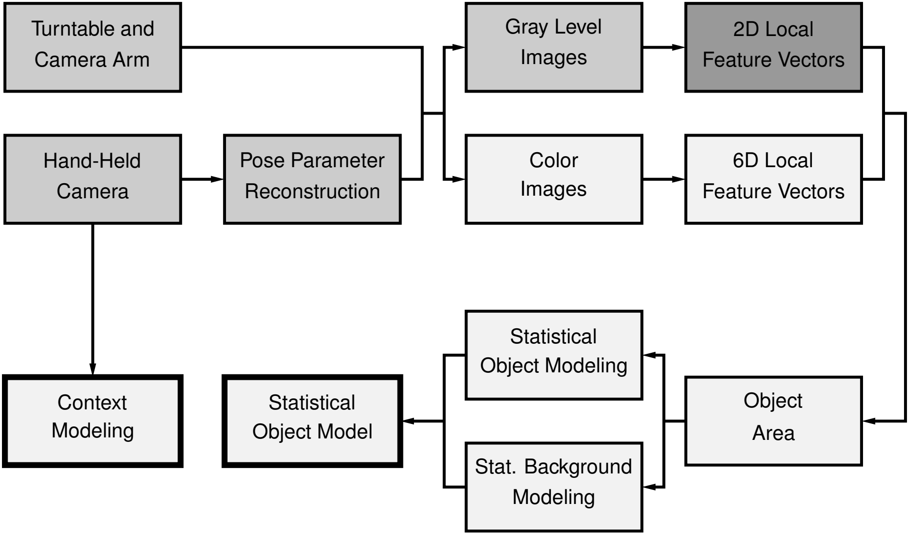
Pose Parameter Reconstruction



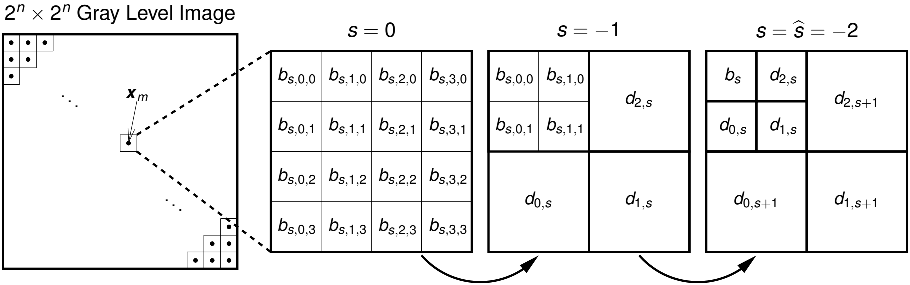
Gray Level Images



2D Local Feature Vectors

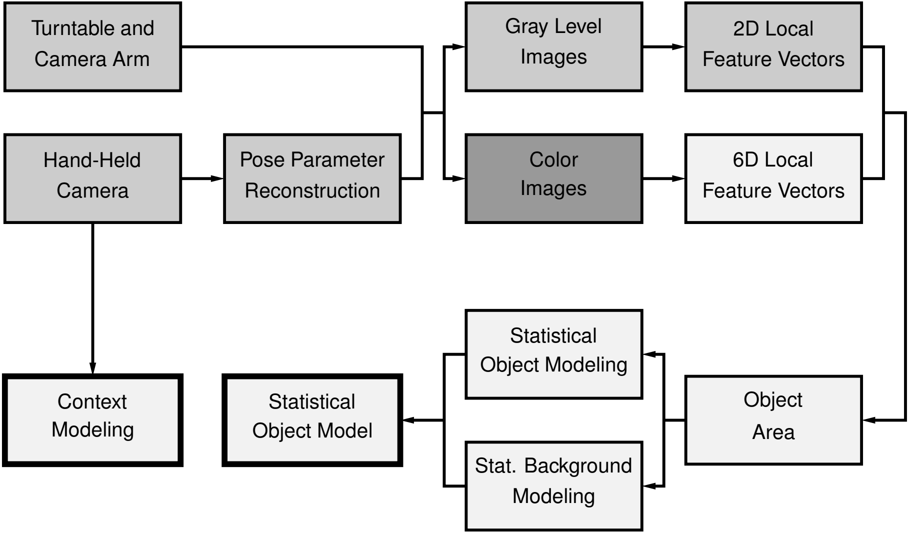


2D Feature Extraction with Wavelet Transform

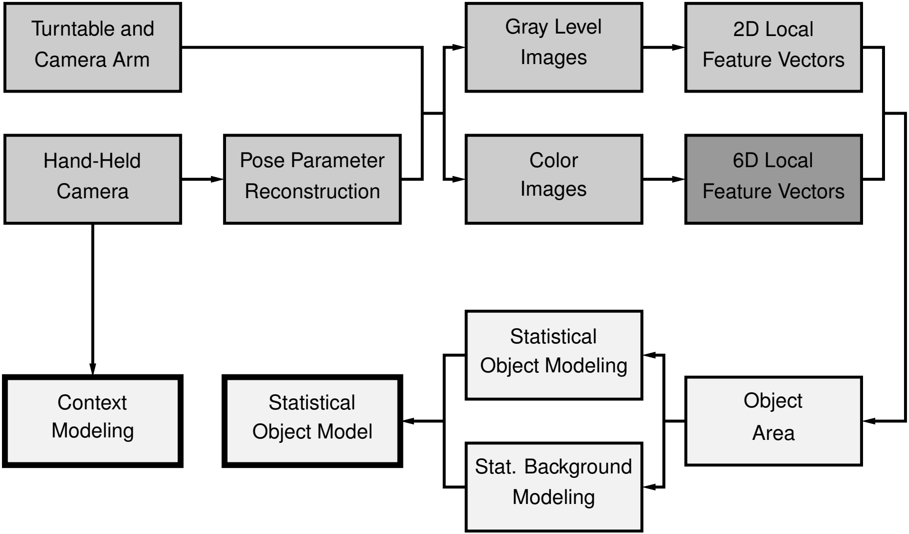


$$\boldsymbol{c}_m = \boldsymbol{c}(\boldsymbol{x}_m) = \begin{pmatrix} c_{m,1} \\ c_{m,2} \end{pmatrix} = \begin{pmatrix} \ln(2^s |b_s|) \\ \ln[2^s(|d_{0,s}| + |d_{1,s}| + |d_{2,s}|)] \end{pmatrix}$$

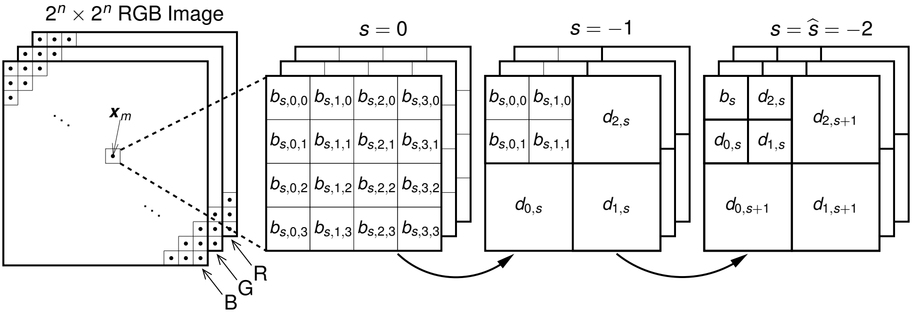
Color Images



6D Local Feature Vectors

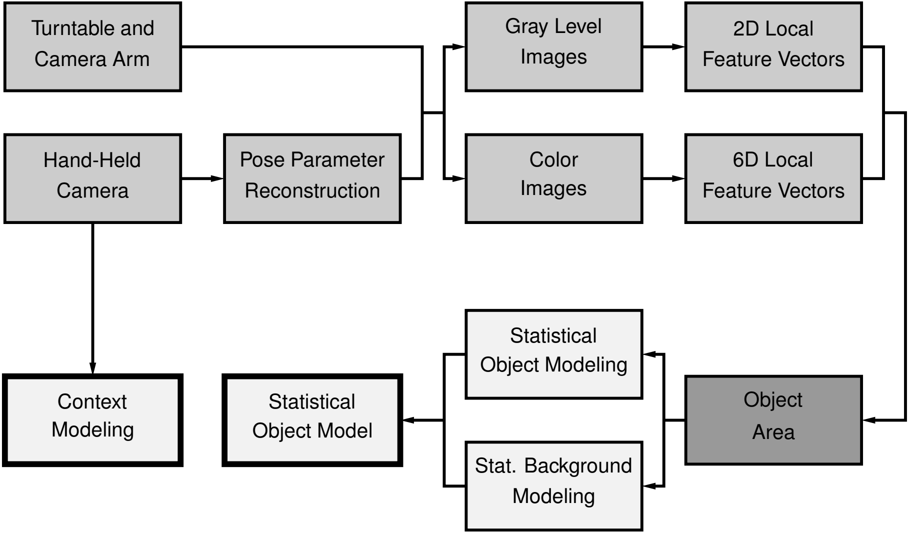


6D Feature Extraction with Wavelet Transform

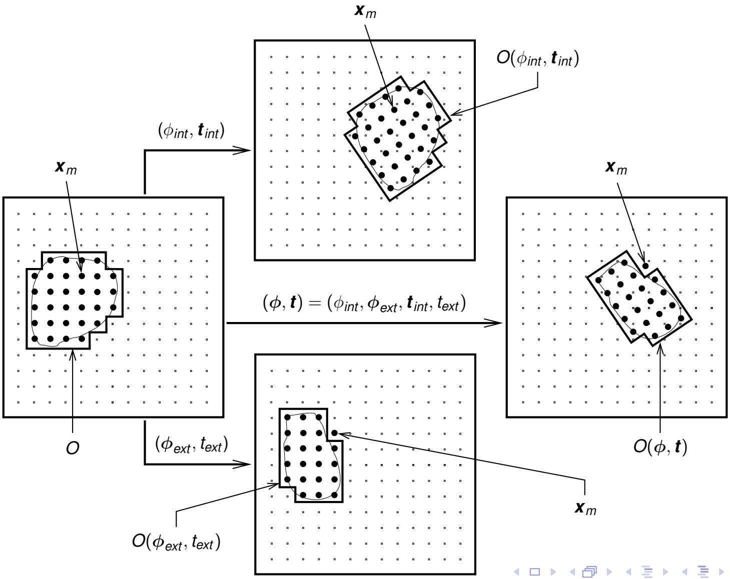


$$\mathbf{c}_m = \mathbf{c}(\mathbf{x}_m) = \begin{pmatrix} c_{m,r,1} \\ c_{m,r,2} \\ c_{m,g,1} \\ c_{m,g,2} \\ c_{m,b,1} \\ c_{m,b,2} \end{pmatrix}$$

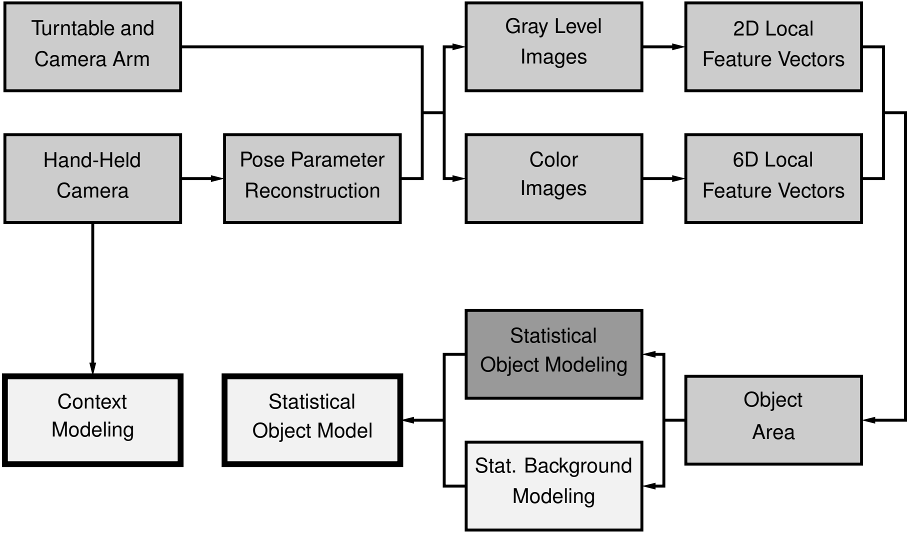
Object Area


Turntable and Camera Arm
Gray Level Images
2D Local Feature Vectors
Hand-Held Camera
Pose Parameter Reconstruction
Color Images
6D Local Feature Vectors
Statistical Object Modeling
Context Modeling
Statistical Object Model
Object Area
Stat. Background Modeling

Object Area $O = O(\phi, \boldsymbol{t})$



Statistical Object Modeling



Object Features as Normal Random Variables

- The elements $c_{m,q}$ of $\boldsymbol{c}_m \in C_O$ are considered as normal random variables

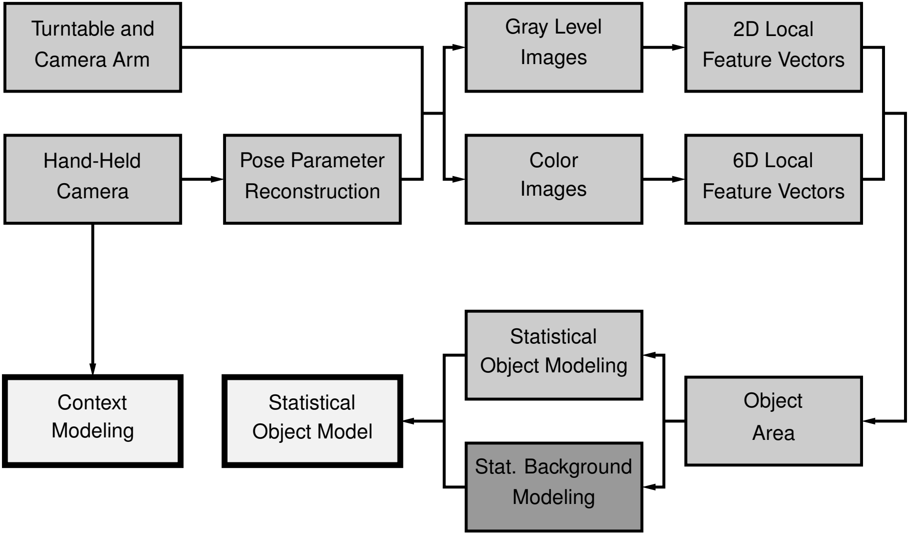
$$p(c_{m,q}|\mu_{m,q},\sigma_{m,q},\phi,\boldsymbol{t}) = \frac{1}{\sigma_{m,q}\sqrt{2\pi}} \exp\left(\frac{(c_{m,q}-\mu_{m,q})^2}{-2\sigma_{m,q}^2}\right)$$

- Assuming the statistical independence of the elements $c_{m,q}$

$$p(\boldsymbol{c}_m|\boldsymbol{\mu}_m,\boldsymbol{\sigma}_m,\phi,\boldsymbol{t}) = \prod_{q=1}^{N_q} p(c_{m,q}|\mu_{m,q},\sigma_{m,q},\phi,\boldsymbol{t})$$

$N_q = 2$ for gray level modeling $\qquad$ $N_q = 6$ for color modeling

Statistical Background Modeling



Background Features as Uniform Random Variables

- The elements $c_{m,q}$ of $\boldsymbol{c}_m \notin C_O$ are considered as uniform random variables

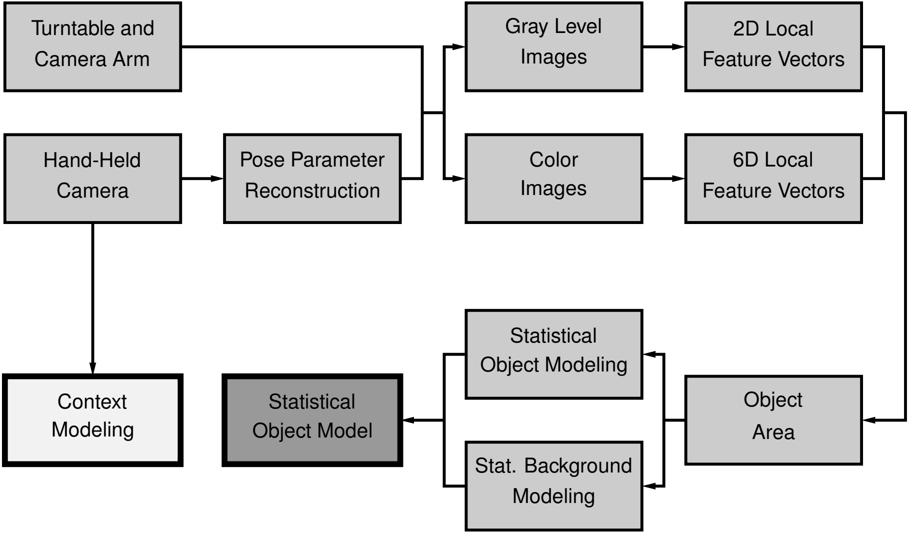
$$p(c_{m,q}) = \frac{1}{\max(c_{m,q}) - \min(c_{m,q})}$$

- Assuming the statistical independence of the elements $c_{m,q}$

$$p(\boldsymbol{c}_m) = \prod_{q=1}^{N_q} \frac{1}{\max(c_{m,q}) - \min(c_{m,q})} = p_b$$

$N_q = 2$ for gray level modeling $\qquad$ $N_q = 6$ for color modeling

Statistical Object Model



Statistical Object Model - Summary

For all $\Omega_{\kappa=1,\ldots,N_\Omega}$ the system creates statistical models $\mathcal{M}_{\kappa=1,\ldots,N_\Omega}$

$$\mathcal{M}_\kappa = \mathcal{M}_\kappa(\phi, \boldsymbol{t})$$

containing:

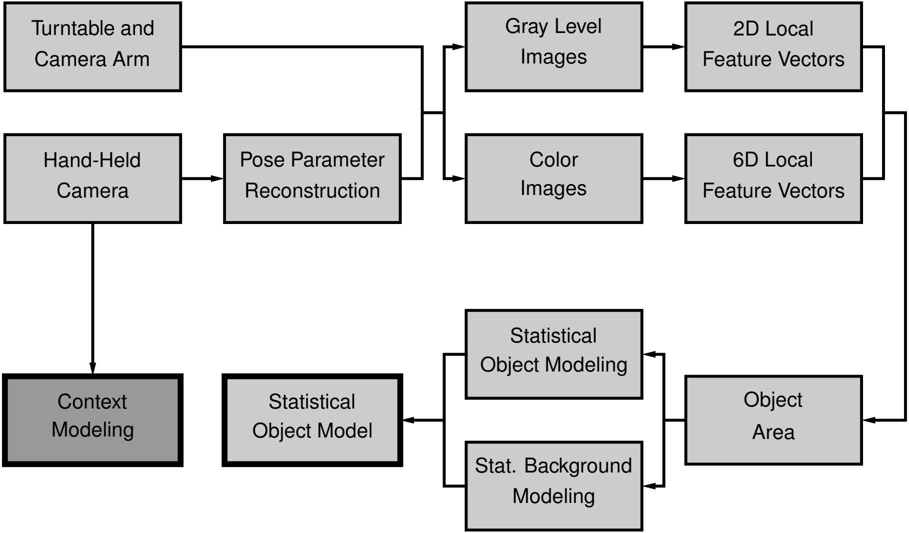
- Object area and set of object feature vectors

$$O_\kappa = O_\kappa(\phi, \boldsymbol{t}) \qquad C_{O_\kappa} = C_{O_\kappa}(\phi, \boldsymbol{t})$$

- Object density and background density

$$p(\boldsymbol{c}_m | \boldsymbol{\mu}_m, \boldsymbol{\sigma}_m, \phi, \boldsymbol{t}) \qquad p_b$$

Context Modeling



Motivation for Context Modeling

- Without context modeling it is assumed that

$$p(\Omega_1) = \cdots = p(\Omega_\kappa) = \cdots = p(\Omega_{N_\Omega})$$

- Considering context dependencies the a-priori probabilities $p(\Omega_\kappa)$ cannot be assumed to be equal and they have to be trained

$$p(\Omega_1) \neq \cdots \neq p(\Omega_\kappa) \neq \cdots \neq p(\Omega_{N_\Omega})$$

Training of A-Priori Probabilities $p(\Omega_\kappa)$

- Set of contexts is introduced

$$\Upsilon = \{\Upsilon_1, \Upsilon_2, \ldots, \Upsilon_\iota, \ldots, \Upsilon_{N_\Upsilon}\}$$

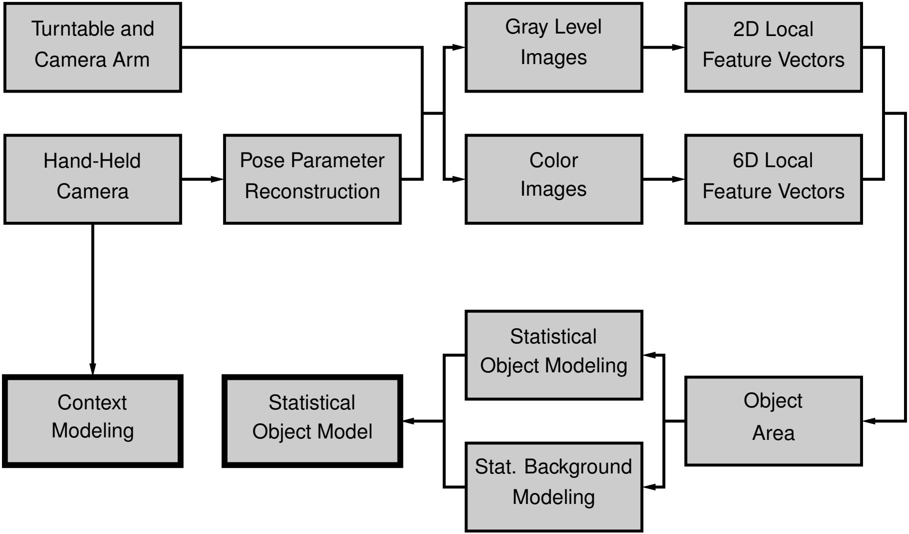
- N_ι images for each context Υ_ι are taken
- $N_{\iota,\kappa}$ denotes how often Ω_κ occurs in Υ_ι
- The a-priori probability for Ω_κ in Υ_ι is defined by

$$p_\iota(\Omega_\kappa) = \eta_\iota N_{\iota,\kappa}$$

- By η_ι a normalization factor is denoted so that

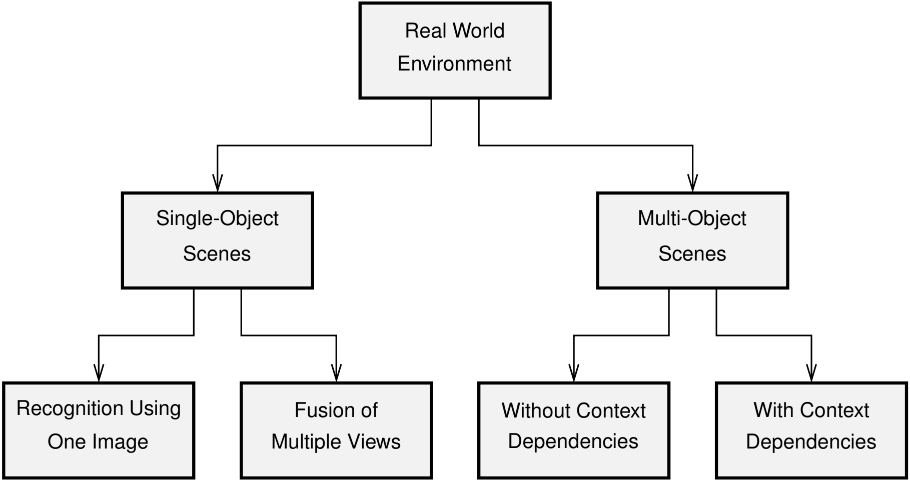
$$\eta_\iota \left[p_\iota(\Omega_1) + p_\iota(\Omega_2) + \ldots + p_\iota(\Omega_\kappa) + \ldots + p_\iota(\Omega_{N_\Omega})\right] = 1$$

Training Phase Completed

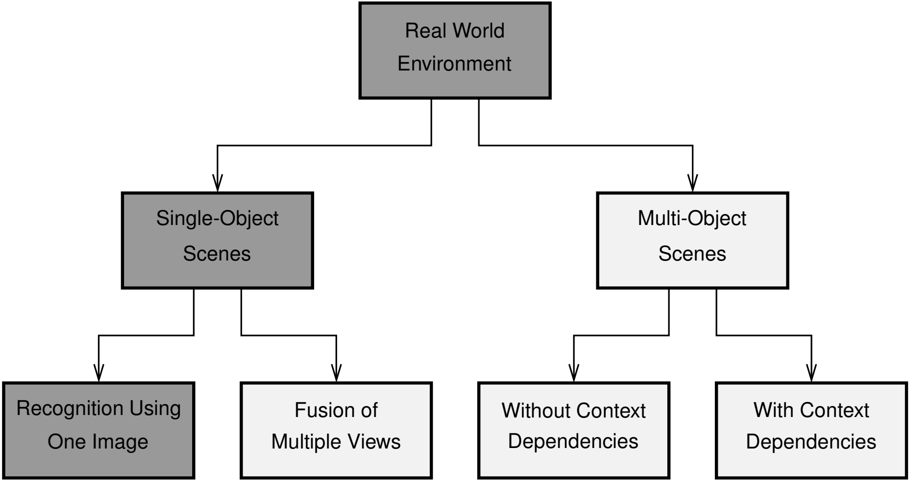


Overview

Recognition Phase



Single-Object, One Image



Maximum Likelihood Estimation

Assumptions

- Observation $\boldsymbol{f} \rightarrow C$
- A-priori probabilities for all $\Omega_{\kappa=1,\ldots,N_\Omega}$ are equal

$$p(\Omega_1) = \cdots = p(\Omega_\kappa) = \cdots = p(\Omega_{N_\Omega})$$

- All $(\phi, \boldsymbol{t})$ are also equiprobable

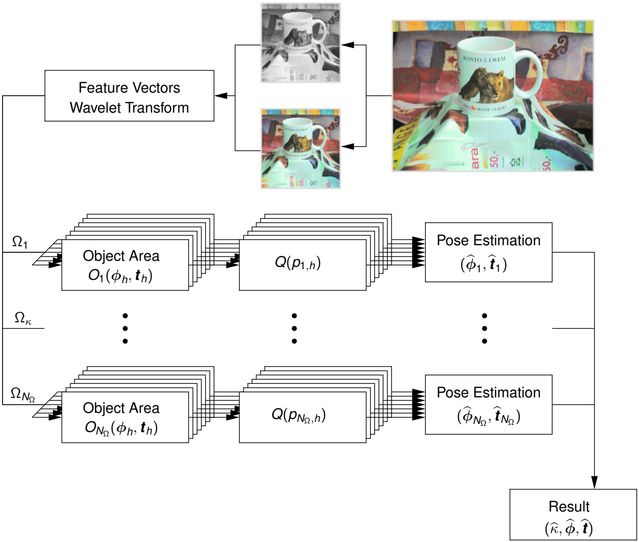
Classification

$$\widehat{\kappa} = \underset{\kappa}{\operatorname{argmax}}\, p(\Omega_\kappa | C) = \underset{\kappa}{\operatorname{argmax}} \frac{p(\Omega_\kappa)p(C|\Omega_\kappa)}{p(C)} = \underset{\kappa}{\operatorname{argmax}}\, p(C|\Omega_\kappa) = \underset{\kappa}{\operatorname{argmax}}\, p(C|\mathcal{M}_\kappa)$$

Recognition

$$(\widehat{\kappa}, \widehat{\phi}, \widehat{\boldsymbol{t}}) = \underset{(\kappa,\phi,\boldsymbol{t})}{\operatorname{argmax}}\, p(C|\mathcal{M}_\kappa(\phi, \boldsymbol{t}))$$

Classification and Localization Algorithm



Object Density Value

Problem

- How to compute $p_{\kappa,h}$ for given $\boldsymbol{f}$, $\mathcal{M}_\kappa$, and $(\phi_h, \boldsymbol{t}_h)$?

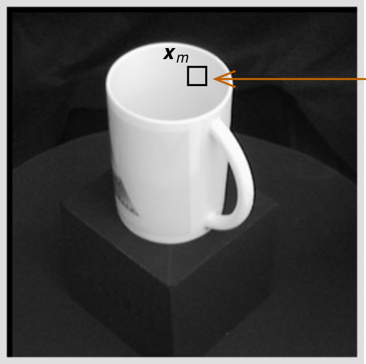
Solution

- Compute feature vectors in $\boldsymbol{f}$!
- Determine $O_\kappa(\phi_h, \boldsymbol{t}_h)$ using $\mathcal{M}_\kappa$!
- Hence, $C_{O_\kappa} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_M\}$ is known.
- Determine $(p_{\boldsymbol{c}_1}, p_{\boldsymbol{c}_2}, ..., p_{\boldsymbol{c}_M})$ using $\mathcal{M}_\kappa$!
- Compute $p_{\kappa,h}$ as follows

$$p_{\kappa,h} = \prod_{i=0}^{M} \max\{p_{\boldsymbol{c}_i}, p_b\} \quad !$$

Motivation for Background Density

Training Image of Ω_κ in $(\phi_h, \boldsymbol{t}_h)$

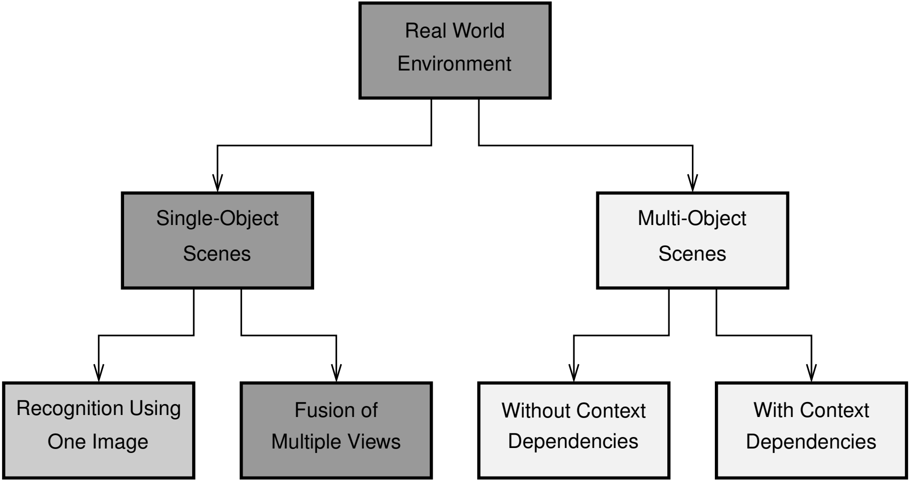


$\boldsymbol{\mu}_{\kappa,m} \neq \boldsymbol{c}_m$

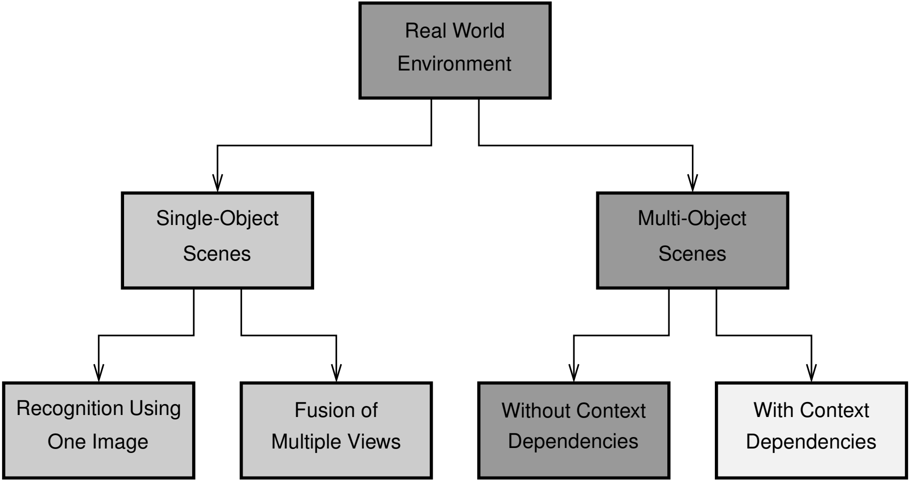
Test Image of Ω_κ in $(\phi_h, \boldsymbol{t}_h)$

$$p_{\boldsymbol{c}_m} = p(\boldsymbol{c}_m | \boldsymbol{\mu}_{\kappa,m}, \boldsymbol{\sigma}_{\kappa,m}, \phi_h, \boldsymbol{t}_h) \approx 0$$

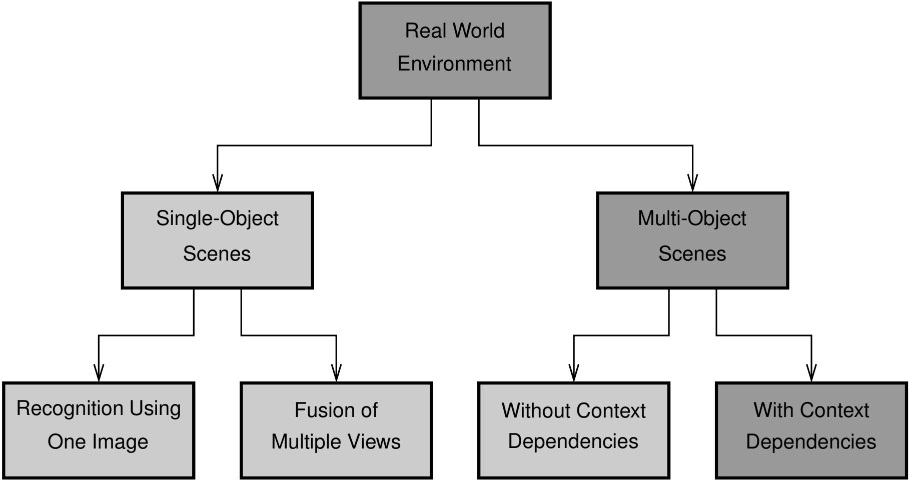
Single-Object, Multiple Views



Multi-Object Scenes without Context



Multi-Object Scenes with Context



Recognition Task for Multi-Object Scenes

Given

- multi-object image $\boldsymbol{f}$

Expected Result

first object	$(\kappa_1, \widehat{\boldsymbol{\phi}}_{\kappa_1}, \widehat{\boldsymbol{t}}_{\kappa_1})$
second object	$(\kappa_2, \widehat{\boldsymbol{\phi}}_{\kappa_2}, \widehat{\boldsymbol{t}}_{\kappa_2})$
$\vdots$	
last object	$(\kappa_{\widehat{i}}, \widehat{\boldsymbol{\phi}}_{\kappa_{\widehat{i}}}, \widehat{\boldsymbol{t}}_{\kappa_{\widehat{i}}})$

Unknown

- object classes and poses $(\kappa_i, \widehat{\boldsymbol{\phi}}_{\kappa_i}, \widehat{\boldsymbol{t}}_{\kappa_i})$
- number of objects $\widehat{i}$
- context $\Upsilon_{\widehat{\iota}}$

Multi-Object Scenes with Context

First Object

- recognition with ML estimation $(\kappa_1, \widehat{\phi}_{\kappa_1}, \widehat{\boldsymbol{t}}_{\kappa_1})$
- context $\Upsilon_{\widehat{\iota}}$ determined with

$$\widehat{\iota} = \underset{\iota}{\operatorname{argmax}}\, p_\iota(\Omega_{\kappa_1})$$

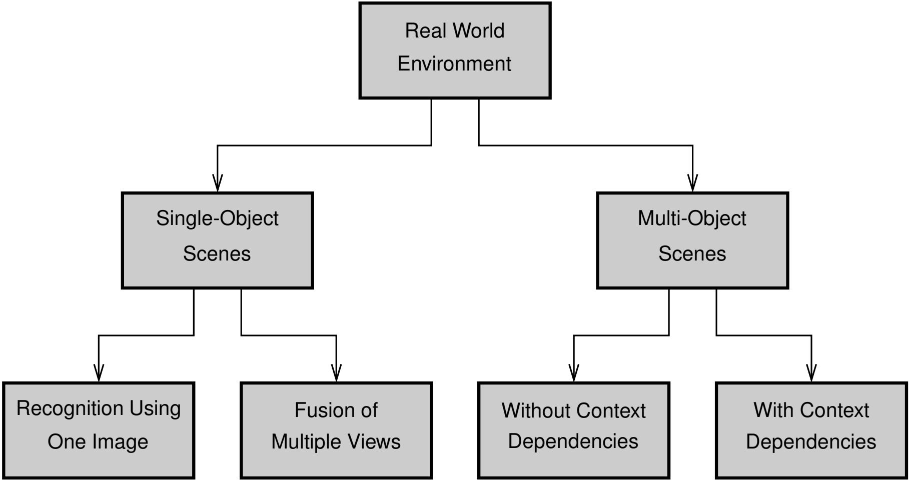
Ranking of Remaining Objects

- object densities weighted with the trained a-priori probabilities

Last Object

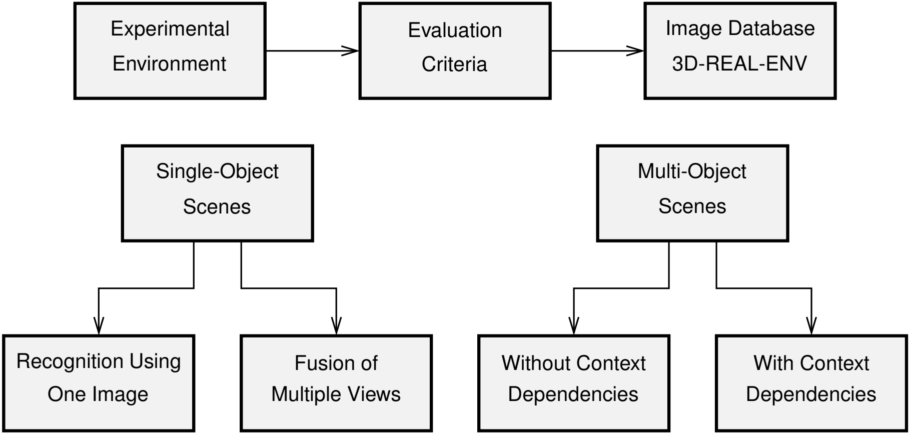
- determined by the highest distance in the ranking

Recognition Phase Completed

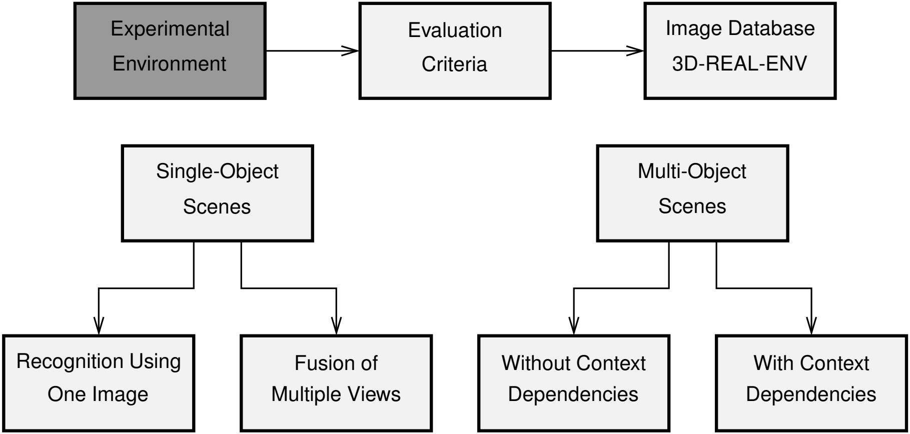


Overview

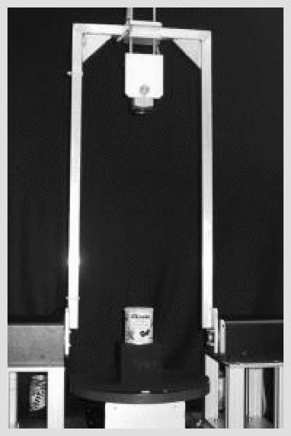
Experiments and Results



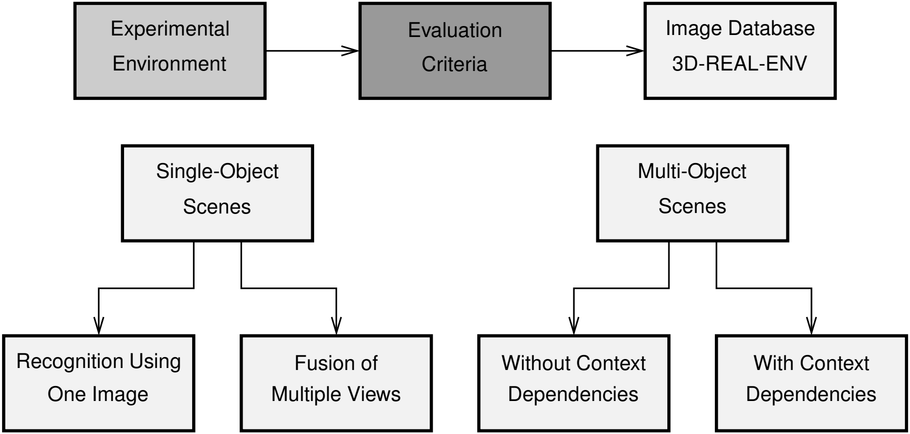
Experimental Environment



Experimental Environment



Evaluation Criteria



Evaluation Criteria

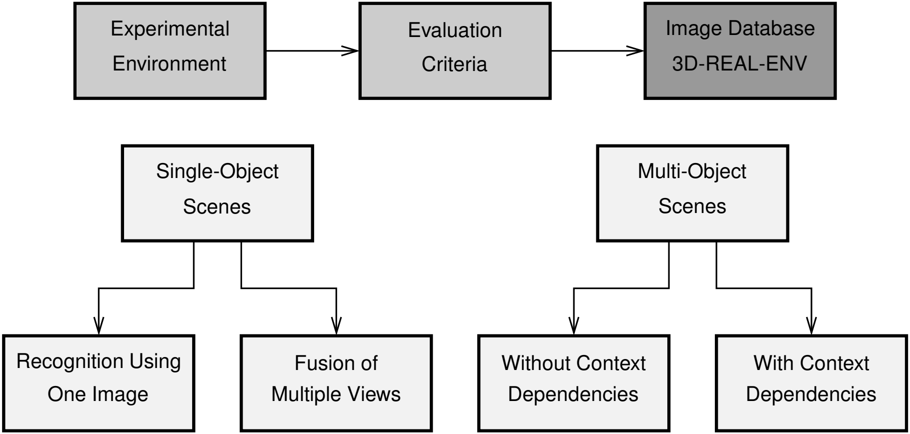
- **Classification** is correct or not
- **Localization** is correct if

$$\Delta t_x \leq 10\text{P} \qquad \Delta \phi_x \leq 15^\circ$$

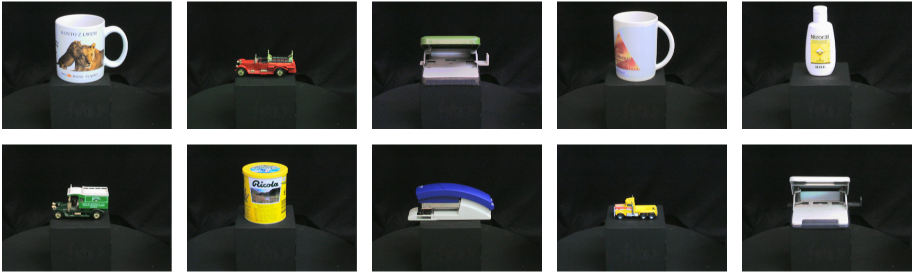
$$\Delta t_y \leq 10\text{P} \qquad \Delta \phi_y \leq 15^\circ$$

$$\Delta t_z \leq 10\% \qquad \Delta \phi_z \leq 10^\circ$$

Image Database 3D-REAL-ENV



Training Images



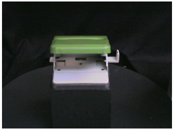
1680 Training Viewpoints, 33600 Images

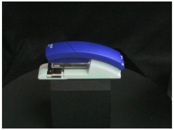
$$\phi_{x,\rho} = (0.0°, 4.5°, 9.0°, \ldots, 85.5°, 90.0°)$$

$$\phi_{y,\rho} = (0.0°, 4.5°, 9.0°, \ldots, 351.0°, 355.5°)$$

Test Images HomBack





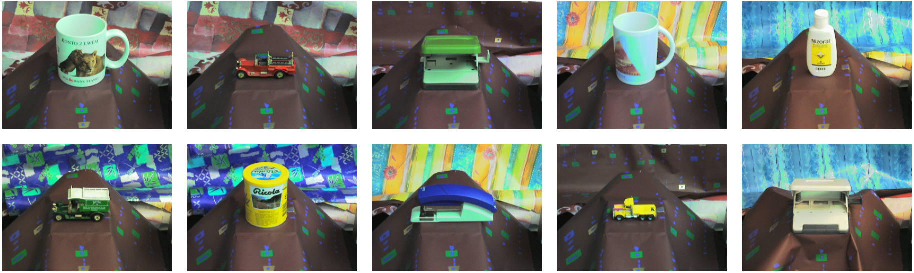



288 Test Viewpoints, 2880 Images

$$\phi_{x,\tau} = (0.00°, 11.25°, 22.50°, \ldots, 78.75°, 90.00°)$$

$$\phi_{y,\tau} = (0.00°, 11.25°, 22.50°, \ldots, 337.50°, 348.25°)$$

Test Images LessHetBack

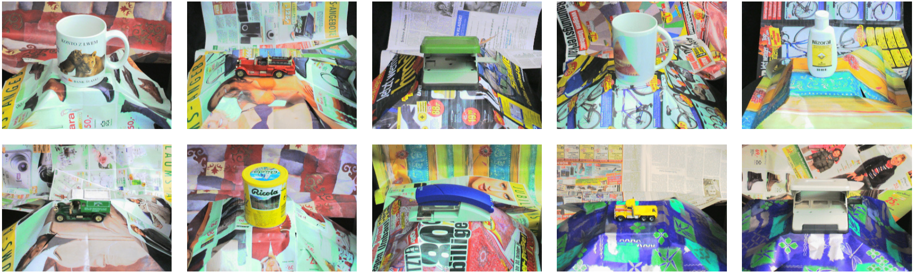


288 Test Viewpoints, 2880 Images

$$\phi_{x,\tau} = (0.00^\circ, 11.25^\circ, 22.50^\circ, \ldots, 78.75^\circ, 90.00^\circ)$$

$$\phi_{y,\tau} = (0.00^\circ, 11.25^\circ, 22.50^\circ, \ldots, 337.50^\circ, 348.25^\circ)$$

Test Images MoreHetBack

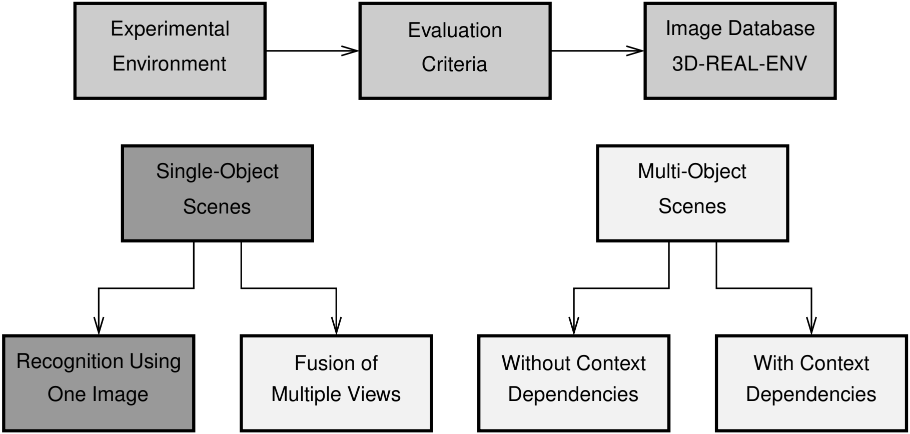


288 Test Viewpoints, 2880 Images

$$\phi_{x,\tau} = (0.00°, 11.25°, 22.50°, \ldots, 78.75°, 90.00°)$$

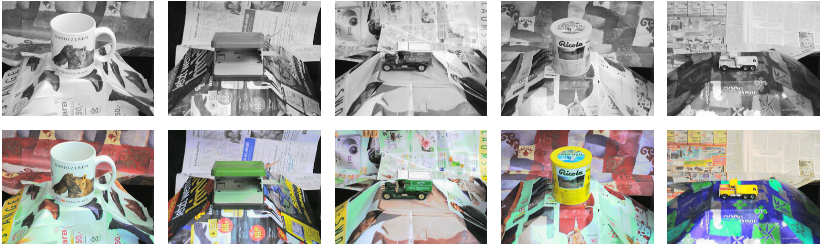
$$\phi_{y,\tau} = (0.00°, 11.25°, 22.50°, \ldots, 337.50°, 348.75°)$$

Single-Object, One Image

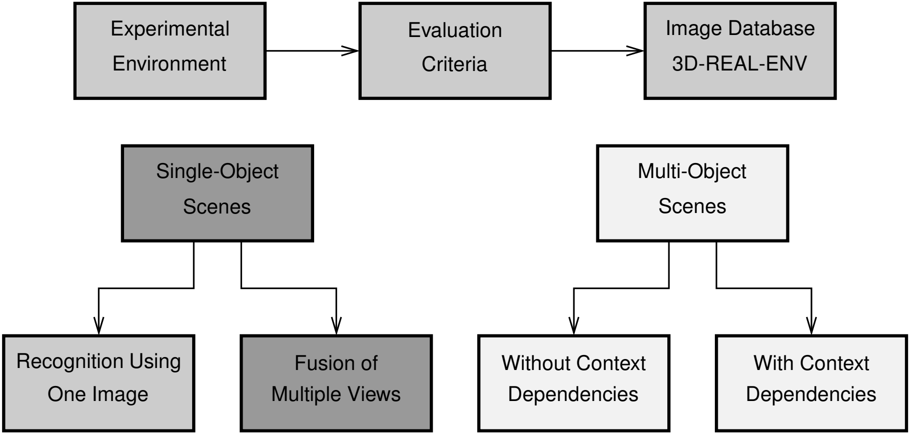


Classification and Localization Rates

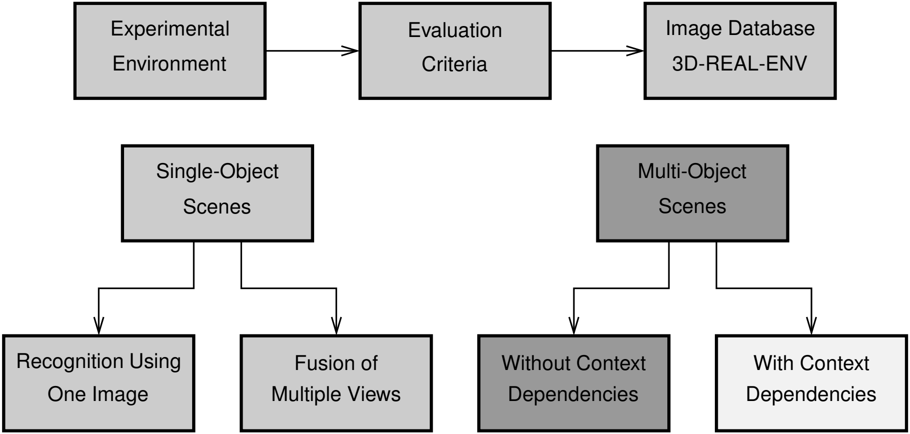
Distance of Training Views 4.5°	Classification			Localization		
	Hom. Back.	**Less Het. Back.**	**More Het. Back.**	**Hom. Back.**	**Less Het. Back.**	**More Het. Back.**
Gray Level	100%	92.2%	54.1%	99.1%	80.9%	69.0%
Color	100%	88.0%	82.3%	98.5%	77.8%	73.6%



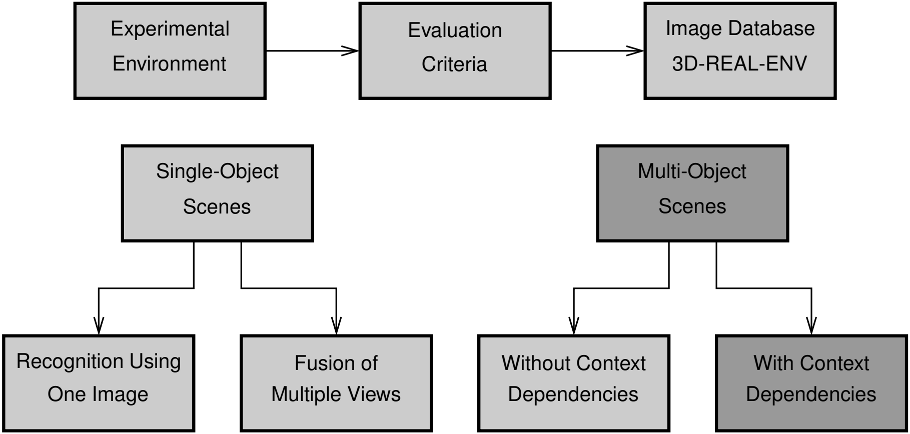
Single-Object, Multiple Views



Multi-Object, Without Context



Multi-Object, With Context



Test Images and Recognition Rates

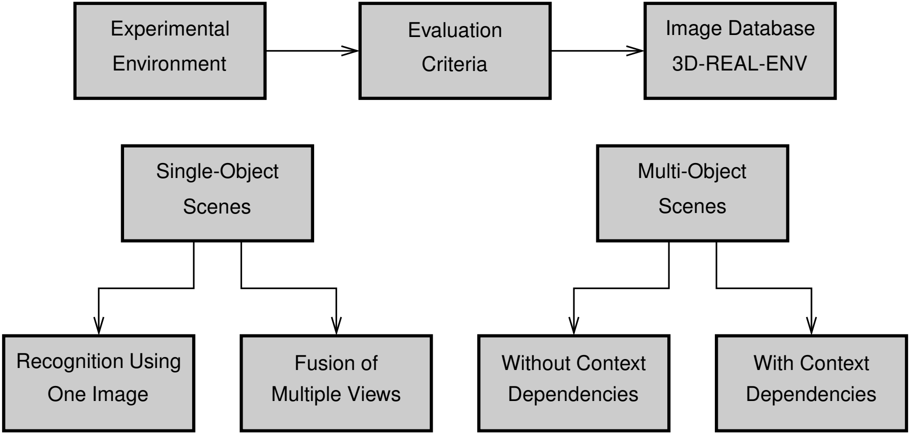
1080 HomBack

1080 LessHetBack

1080 MoreHetBack

3D-REAL-ENV Image Database	Without Context Modeling			With Context Modeling		
	Hom	LessHet	MoreHet	Hom	LessHet	MoreHet
Classification	100%	91.9%	62.9%	100%	97.0%	87.5%
Localization	99.7%	81.7%	58.1%	99.7%	81.7%	58.1%

Evaluation Completed



Overview

Museum Artifact Classification

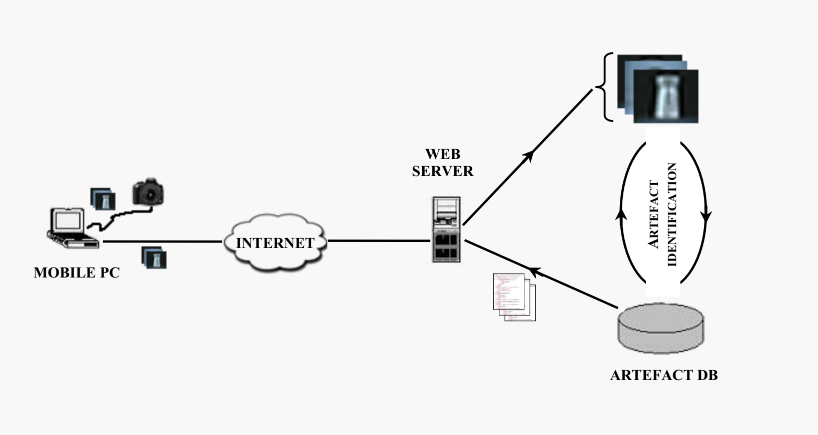
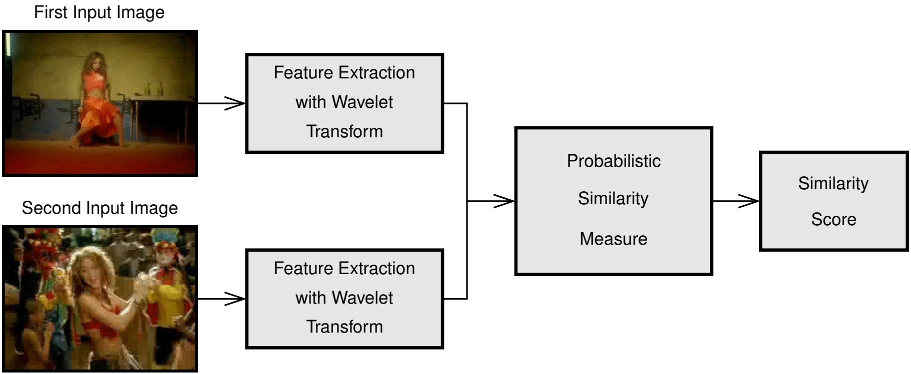


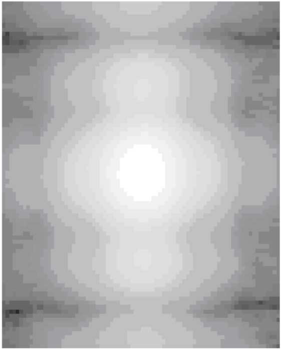
Image Similarity Measure



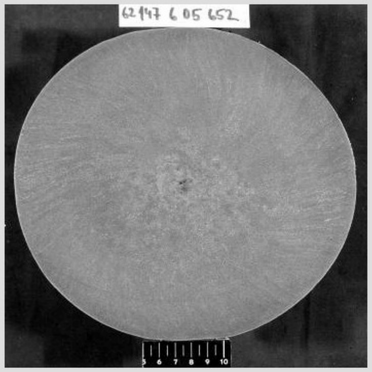
Applications

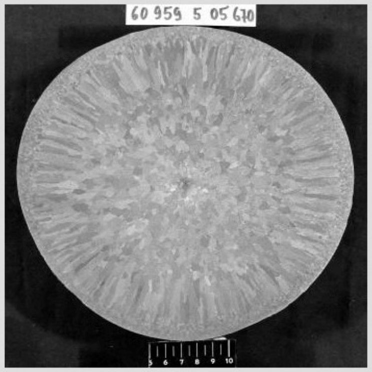
- Similarity Measure for Antenna Patterns
- Quality Scoring of Metallography Images
- Content Management and Retrieval System

Antenna Patterns

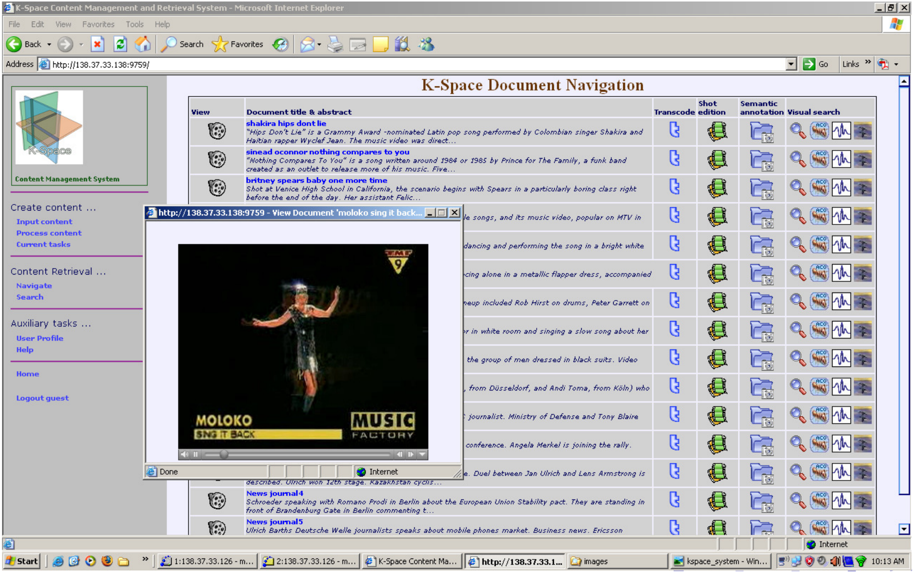


Metallography Images





Video Management and Retrieval System



Overview

Final Remarks & Future Work

Final Remarks

- Color Modeling (Classification Rate: 54.1% $\rightarrow$ 82.3%)
- Context Modeling (Classification Rate: 62.9% $\rightarrow$ 87.5%)
- 3D-REAL-ENV Image Database is Becoming Popular
- System is Applicable for Real Life Tasks

Future Work

- Museum Artifact Classification
- Quality Scoring of Metallography Images
- Image Classification for Video Retrieval
- Appearance-Based and Shape-Based Object Recognition

Summary

1. **Fundamental Concept**
2. **Statistical Modeling**
3. **Classification and Localization**
4. **Experiments and Results**
5. **System Applications**
6. **Conclusion**