Task-oriented Computer Vision in 2D and 3D: from video text recognition to 3D human detection and tracking

Csaba Beleznai

Michael Rauter, Christian Zinner, Andreas Zweng, Andreas Zoufal, Julia Simon, Daniel Steininger, Markus Hofstätter und Andreas Kriechbaum

Csaba Beleznai
Senior Scientist
Center for Vision, Automation & Control
Autonomous Systems
AIT Austrian Institute of Technology GmbH
Vienna, Austria
Austrian Institute of Technology

Austrian Institute of Technology (AIT)

- Health & Environment
- Safety & Security
- Energy
- Mobility
- Foresight & Policy Development
- Seibersdorf Labor GmbH
- Nuclear Engineering Seibersdorf

Short intro – who are we in 20 seconds
Contents

- Motivate & stimulate
- Algorithms through applied examples

Optical flow driven motion analysis

Video text recognition

Left item detection

Queue length and waiting time estimation
A frequently asked question

Introduction
Why is Computer Vision difficult?
(from a Bayesian perspective)

- **Primary challenge in case of Vision Systems (incl. biological ones):**
  
<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Features</td>
</tr>
<tr>
<td>Mid</td>
<td>Groupings</td>
</tr>
<tr>
<td>High</td>
<td>Concepts</td>
</tr>
</tbody>
</table>

  **Visual analysis**

  - Complementary cues (depth, more views, more frames)
  - Complementary groupings (spatial, temporal across frames)
  - Complementary high-level information (user, learnt)

  **Motivation**

  - Prior knowledge
  - Parameters, off-line and incrementally learned information (from a Bayesian perspective)

  - Shadow
  - Texture
  - True boundary
Example: Crop detection

- Radial symmetry
- Near regular structure
Example 2: Seam tracking

- Weld seam segmentation and tracking

- Panoramic image generation
Motivation

- Challenges when developing Vision Systems:
  - Complexity ← Algorithmic, Systemic, Data
  - Non-linear search for a solution

- Introduction
  - Challenges when developing Vision Systems:
    - Complexity ← Algorithmic, Systemic, Data
    - Non-linear search for a solution
Research methodology

- Thematic areas and trends in Computer Vision also distributed *branch-and-bound*
- Balance: becoming a domain expert vs. being a „globalist“
- Researchers tend to favour certain paradigms - Learn to outline trends, look *upstream*
- Revisit old problems to see them under new light
- Specialize the general & Generalize the specific
- Factorize your know-how (code, topics, …) into components → sustainable, scalable
Real-time optical flow based particle advection
Optical flow driven advection

*Advection*: transport mechanism induced by a force field

\[ V_{x,i}, V_{y,i} \]

A particle trajectory induced by the OF field
Particle advection with FW-BW consistency

- A simple but powerful test

Forward:

Backward:

Consistency check: $\Delta \varepsilon < \beta \overline{\Delta x}$

$\overline{\Delta x}$ : mean offset
Pedestrian Flow Analysis

Public dataset: Grand Central Station, NYC: 720x480 pixels, 2000 particles, runs at 35 fps
Wide-area Flow Analysis

Other examples: wide area surveillance (small objects, nuisance, clutter)
OBJECT DELINEATION

OBJECT DETECTION PIPELINES

Spatial distribution of posterior probability
Score map (DPM, R-CNN, …)
Vote map
Occupancy map
back-projected similarity map
...

Delineated objects
Bounding boxes
More complex parametric representations
Instance segmentation

DPM: Deformable Part Models
R-CNN: Region-based Convolutional Neural Networks
Intermediate probabilistic representations → 2D distributions

Local grouping prior, structure-specific knowledge

generate consistent object hypotheses

Challenge:
- arbitrarily shaped distributions
- multiple nearby modes
- noise, clutter

Definitions:

\( \text{mode} = \text{location of maximum density} \)

computed using a kernel \( K \)

density estimation of variable \( x \)

\[
f(x) = \sum_a K(a - x)w(a)
\]
MOTIVATION – II.
Analysis of discrete two-dimensional distributions $\rightarrow$ object-based formalism
## RELATED STATE-OF-THE-ART

- **Clustering detections**
  - Weakly constrained structural prior
  - Non-maximum suppression
    - Rothe et al., 2014
    - Neubeck & Van Gool, 2006

- **Mean Shift, CAMShift**
  - Comaniciu & Meer, 2002
  - Bradski 1998

- **Detection by voting/segmentation/learning**
  - Implicit or explicit structural prior
  - Implicit Shape Model
    - Leibe et al. 2005
  - Markov Point Processes for object configurations
    - Verde, 2014
  - Structured random forests
    - Dollar & Zitnick 2013
    - Kontschieder et al. 2011
  - CNN’s for Non-Max. Suppression
    - Hosang et al. 2016
    - Wan et al. 2015
SHAPE LEARNING – CASE: COMPACT CLUSTERS

1. Binary mask from **manual annotation** or from **synthetic data**

2. Sampling using an analysis window **discretized into a** \( n_i \times n_i \) **grid**

3. Building a **codebook of binary shapes** with a coarse-to-fine spatial resolution

\[
\mathcal{S} = \left\{ \{l_i\}_{i=1}^{3}, v, c \right\}
\]
EXAMPLE CODEBOOK – CASE: COMPACT CLUSTERS

FULL TREE

ZOOM LEVEL 1

ZOOM LEVEL 2

- off-the-mode structure
- mode-centered structure

- low
- mid
- high
SHAPE LEARNING – CASE: LINE STRUCTURES

Binary mask from **manually annotated** text lines

Spatial resolution of local structure

- **low**
- **mid**
- **high**

Line-centered samples

Off-the-line samples

Codebook: \( S = \{ \{ l_i \} \}_{i=1..3}, t, c \)
SHAPE DELINEATION – I.

Step 1: Fast Mode Seeking

Three integral images: \( I, I \cdot x \) and \( I \cdot y \)

Mode location:

\[
x' = \frac{\sum_a K''(a - x)ii_x(a)}{\sum_a K''(a - x)ii(a)}
\]

Step 2: Local density analysis

Density measure \( D_i \) for each resolution level \( i \) for the binary structure \( l_i \)

\[
D_i (l_i | I) = \frac{1}{A_F} \sum_{\{x, y \in C | l = 1\}} I(x, y) - \frac{1}{A_B} \sum_{\{x, y \in C | l = 0\}} I(x, y)
\]

Enumerating all binary shapes at each resolution level

\( \rightarrow \) Finding best matching entry:

\[
l_i^* = \arg \max_l D_i (l_i | I)
\]
SHAPE DELINEATION – II.

Recursive search for end points, starting from mode locations:

Relative line end locations define:
- Search direction
- Line end positions
Optical flow driven particle tracklets

The \(i\)th tracklet: \(T_i = \{x_t, y_t\}_{t=1..N}, \nu, w\)

\(\nu\) – velocity vector
\(w\) – weight (scalar)

Clustering directly performed in the discrete tracklet-domain
End-to-end video text recognition
Overview

- The End-to-End Video Recognition Process

**INPUT**

**Detection**

- Presence (y/n)

**Localization**

- Location (single frames) (x, y, w, h)

**Propagation**

- Location (frame span) (x, y, w, h)

**Segmentation**

- Binary image regions

**OUTPUT**

- Recognition, Propagation

- Text (e.g. in ASCII)

**Characterizing dynamic elements: running text**

**Evaluation:**

High accuracy at each stage is necessary

Very high recall throughout the chain

Increasing Precision toward the end of the chain
Algorithmic chain - Motivation

Main strategies for text detection:

What is text (when appearing in images)?

An oriented sequence of characters in close proximity, obeying a certain regularity (spatial offset, character type, color).

Sample text region + complex background
Algorithmic chain - Motivation

To detect → Representing text appearance:

- **Region based:**
  - Binary morphology (outdated technique: trying to find nearby characters and segmenting lines)
  - Statistics
    - Edge density, frequency, orientation (popular: HOG), …
    - Texture representation: filter banks, co-occurrence, …
  → Discriminative classifier → **relatively fast**, but some hard-to-discriminate cases (vegetation, dense regular patterns /grids, gravel/) + poor region segmentation

- **Analysis at character-level**
  - Requires a full or partial segmentation (a challenge itself) → character or stroke
    - Highly specific (stroke width is uniform, shape is very specific)
  → Segmentation → **rather slow**, but yields accurate segmentation

- **Analysis at grouped-character-level:** a sequence of similar characters is specific

- **Analysis at OCR-level:** comparison to a pre-trained alphanumerical set → highly specific (slow!!)
A typical algorithmic chain

Input image

Detector (Classifier)

Text present → no

Line Detector (Vessleness)

Text present → yes

Binary region mask

Stroke Width Transform (SWT*)

MSER*

Class-specific ER***

Grouping

OCR

Region-based analysis

Character-level analysis

Character-group-level analysis

OCR-level analysis

* SWT = Stroke Width Transform (Epshtein, 2010)
** MSER = Maximally Stable Extremal Regions (J. Matas, 2002) – Part of OpenCV
*** ER = Class-specific Extremal Regions (J. Matas, 2007)
Improved text detection – synthetic text generation
(Classification using Aggregated Channel Features)
Video segment from CNN
Experimental results - Case: Line structures (Text line segmentation)

Input image

Simple binarization

Binarization is very sensitive to employed threshold

Proposed scheme

Our scheme has no threshold, only local structural priors
Experimental results - Case: Text line segmentation
Convolutional Neural Network based OCR - Training

Generated single characters (0-9, A-Z, a-z): include spatial jitter, font variations

- role of jitter: characters can be recognized despite an offset at detection time
Convolutional Neural Network based OCR - Results

Analysis window is scanned along the textline, and likelihood ration \( \frac{\text{score}_1}{\text{score}_2} \) is plotted in the row (below textline) belonging to the maximum classification score.

ESCALATING TENSIONS

TURKEY TO RETURN PILOT’S BODY TO RUSSIA

0800 032 7000
Obtaining stereo depth information
Passive stereo based depth measurement

- 3D stereo-camera system developed by AIT
  - Area-based, local-optimizing, correlation-based stereo matching algorithm
  - Specialized variant of the Census Transform
  - Resolution: typically ~1 Mpixel
  - Run-time: ~ 14 fps (Core-i7, multithreaded, SSE-optimized)
  - Excellent “depth-quality-vs.-computational-costs” ratio
  - USB 2 interface

Advantage:
- Depth ordering of people
- Robustness against illumination, shadows,
- Enables scene analysis
Stereo camera characteristics

Trinocular setup:

- 3 baselines possible
- 3 stereo computations with results fused into one disparity image
Data characteristics

Intensity image

Disparity image

Planar surface in 3D space

$\mathbf{x}, \mathbf{y}$

$\mathbf{d}$

Disparity image

Intensity image

$\mathbf{y}$

$\mathbf{d}$
Queue length detection using depth and intensity information
Queue Length + Waiting Time estimation

What is waiting time in a queue?

Time measurement relating to last passenger in the queue

Why interesting?

Example: Announcement of waiting times (App) → customer satisfaction
Example: Infrastructure operator → load balancing
Experimental results - Case: Compact clusters

Human detection by **occupancy map** clustering:

Passive stereo depth sensing $\rightarrow$ depth data projected orthogonal to the ground plane

Occupancy map ($1246 \times 728$ pix.) clustering: **56 fps**, overall system (incl. stereo computation): **6 fps**
Experimental results - Case: Compact clusters

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Binarization</th>
<th>Mean Shift</th>
<th>Cam Shift</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall (R)</td>
<td>0.52</td>
<td>0.95</td>
<td>0.81</td>
<td>0.92</td>
</tr>
<tr>
<td>Precision (P)</td>
<td>0.86</td>
<td>0.76</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>F-measure (F)</td>
<td>0.65</td>
<td>0.84</td>
<td>0.85</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Queue analysis

- Challenging problem

\[ \text{Waiting time} = \frac{\text{Length}}{\text{Velocity}} \]

1. What is the shape and extent of the queue?
2. What is the velocity of the propagation?

- Shape
  - No predefined shape (context/situation-dependent and time-varying)

- Motion → not a pure translational pattern
  - Propagating stop-and-go behaviour with a noisy “background“
  - Signal-to-noise ratio depends on the observation distance

**DEFINITION**: Collective goal-oriented motion pattern of multiple humans exhibiting spatial and temporal coherence
How can we detect (weak) correlation?

Much data is necessary → Simulating crowding phenomena in Matlab

- Social force model (Helbing 1998)

Source: Parameswaran et al. Design and Validation of a System for People Queue Statistics Estimation, Video Analytics for Business Intelligence, 2012
Queue analysis

Simulation tool → Creating infinite number of possible queueing zones

Two simulated examples (time-accelerated view):
Queue analysis (length, dynamics)

Staged scenarios, 1280x1024 pixels, computational speed: 6 fps
Adaptive estimation of the spatial extent of the queueing zone

Estimated configuration (top-view)

Detection results

Left part of the image is intentionally blurred for protecting the privacy of by-standers, who were not part of the experimental setup.
Adaptive estimation of the spatial extent of the queueing zone
Scene-aware heatmap
Left Item Detection

Additional knowledge (compared to existing video analytics solutions):

- Stationary object (Geometry introduced to a scene)
- Object geometric properties (Volume, Size)
- Spatial location (on the ground)
Methodology

Input images

Processing intensity and depth data

Stereo disparity

Ground plane estimation

Ortho-map generation

Ortho-transform

Change detection

Background model

Combination of proposals + Validation

Final candidates

Object detection and validation in the ortho-map
Left Item Detection – Demos
Left Item Detection – Interesting cases

Object form

Transferred objects
Implementation details and strategy
Our development concept

- **MATLAB:**
  - Broad spectrum of algorithmic libraries,
  - Well-suited for image analysis,
  - Visualisation, debugging,
  - Rapid development → Method, Prototype, Demonstrator

- **C/C++**
  - Real-time capability
Our development concept

Advanced Methods
- Multi-Camera Tracking
- Automatic Calibration
- Soft Biometrics
- Person Detection and Tracking

Standard Methods
- Person Detection
- Moving Objects
- Advanced Background Model

Products
- Static Objects
- Moving Objects

Innovations
Thank you for your attention!

CSABA BELEZNAI
Senior Scientist
Center for Vision, Automation & Control
Autonomous Systems

AIT Austrian Institute of Technology GmbH
Donau-City-Straße 1 | 1220 Vienna | Austria
T +43(0) 664 825 1257 | F +43(0) 50550-4170
csaba.beleznai@ait.ac.at | http://www.ait.ac.at