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

Csaba Beleznai

Senior Scientist
Video- and Safety Technology
Safety & Security Department
AIT Austrian Institute of Technology GmbH
Vienna, Austria

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

why is computer vision
why is computer vision so difficult

Press Enter to search.
Why is Computer Vision difficult?
(from a Bayesian perspective)

- Primary challenge in case of Vision Systems (incl. biological ones):
  - Uncertainty/ambiguity

Motivation
Example: Crop detection

- Radial symmetry
- Near regular structure

Example for robust vision
Motivation

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

A deeper understanding towards the problem is developed during the search for a solution
Typical surveillance scenario:

**Who**: people, vehicle, objects, ...

**Where**: is their location, movement?

**What**: is the activity?

**When**: does an action occur?
Motivation

Visual Surveillance - Motivating example

Algorithmic units:
- Object detection and classification
  - Counting, Queue length, Density, Overcrowding
  - Abandoned objects
  - Intruders
- Tracking
  - Single objects
  - Video search
  - Flow
- Activity recognition
  - Near-field (articulation)
  - Far-field (motion path)

Typical surveillance scenario:
- **Who**: people, vehicle, objects, …
- **Where**: is their location, movement?
- **What**: is the activity?
- **When**: does an action occur?
Real-time optical flow based particle advection
Optical flow driven advection

*Advection*: transport mechanism induced by a force field

- **Dense optical flow field**
- A particle trajectory induced by the OF field
Particle advection with FW-BW consistency

- A simple but powerful test

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

$\bar{\Delta x}$ : 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)
End-to-end video text recognition
Overview

- The End-to-End Video Recognition Process

**INPUT**

Detection

Localization

Propagation

Segmentation

**OUTPUT**

Recognition, Propagation

Text

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!!)
Improved text detection – synthetic text generation
(Classification using Aggregated Channel Features)
Video segment from CNN
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 \((\text{score}_1/\text{score}_2)\) 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
Left-item detection using depth and intensity information

- Composite task:
  - Static object detection
  - Human detection and tracking
What is a static object?

- “non-human” foreground which keeps still over a certain period of time

- Two fundamentally different approaches:
  1. Background modeling (foreground regions becoming static)
     - +: simple, pixel-based
     - -: object removal, ghosts
  2. Tracking detected foreground regions
     - +: many adequate tracking approaches (blob-based, correlation-based)
     - -: crowd, occlusion → failure

Both techniques experience problems with illumination variations → motivation for depth-based sensing
A common approach

Temporal sub-sampling and combination procedure

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 \((x, y)\) image coordinates, \(d\) disparity \(d(x, y)\)
2.5D vs. 3D algorithmic approaches

2.5D == using disparity as an intensity image
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

Background model

Change detection

Ortho-transform

Object detection and validation in the ortho-map

Combination of proposals + Validation

Final candidates

Processing intensity and depth data

Stereo disparity

Ground plane estimation

Ortho-map generation
Quantitative evaluation

Detection results  Ground truth  Depth-based proposals  Motion-based proposals
Human/Object detection as clustering
A Frequently Occurring Task

Analysis of discrete two-dimensional distributions
Task definition

Intermediate probabilistic representations $\rightarrow$ 2D distributions

Local grouping $\rightarrow$ generate consistent object window hypotheses

prior, structure-specific knowledge

Challenge:

- arbitrarily shaped distributions
- multiple nearby modes
- noise, clutter
Related State-of-the-Art

- Weakly constrained structural prior:

  Non-maximum suppression
  Neubeck & Van Gool, 2006  R. Rothe et al., 2014

  Mean Shift, CAMShift
  Comaniciu & Meer, 2002  Bradski 1998

- Using structure information:

  Local structural elements such as bricks, shapelets
  Jin & Geman 2006

  Implicit Shape Model  B. Leibe et al. 2005

  Structured random forests  edge structure
  Dollar & Zitnick 2013  semantic label distribution within local patches
  Kontschieder et al. 2011
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

Codebook: $S = \left\{ \{l_i\}_{i=1}^{3}, v, c \right\}$
Example Codebook – Case: Compact clusters

FULL TREE

ZOOM LEVEL 1

ZOOM LEVEL 2

off-the-mode structure

mode-centered structure

Shape learning
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:** \( \mathcal{S} = \{ \{ l_i \}_{i=1}^{3}, t, c \} \)
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 \( D \) 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

Finding best matching entry:

\[
l_i^* = \arg \max_i D_i (l_i | I)
\]
Shape delineation – II.

Recursive search for end points, starting from mode locations:

Relative line end locations define:
- Search direction
- Line end positions
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>
Experimental results - Case: Line structures (Text line segmentation)

- **Input image**

- **Probability distribution for text**

- **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
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
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?

**DEFINITION**: Collective goal-oriented motion pattern of multiple humans exhibiting spatial and temporal coherence.
Visual queue analysis - Overview

- How can we detect (weak) correlation?

- Much data is necessary → Simulating crowding phenomena in Matlab
  - *Social force model* (Helbing 1998)

Queue analysis
Simulation tool → Creating infinite number of possible queueing zones

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

Straight line

Meander style

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.
Scene-aware heatmap
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

**Implementation details**

- **MATLAB:**
  - 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

- Moving Objects
- Advanced Background Model
- Moving Objects
- Static Objects

PC  GPU  FPGA / DSP

Innovations

Products
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
Thank you for your attention!

CSABA BELEZNAI
Senior Scientist
Digital Safety & Security Department
Video- and Security Technology

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