Seeing the Objects Behind the Dots: Recognition in Videos from a Moving Camera

Bjorn Ommer, Theodor Mader, Joachim M. Buhmann
Present by: Alisa Kunapinun

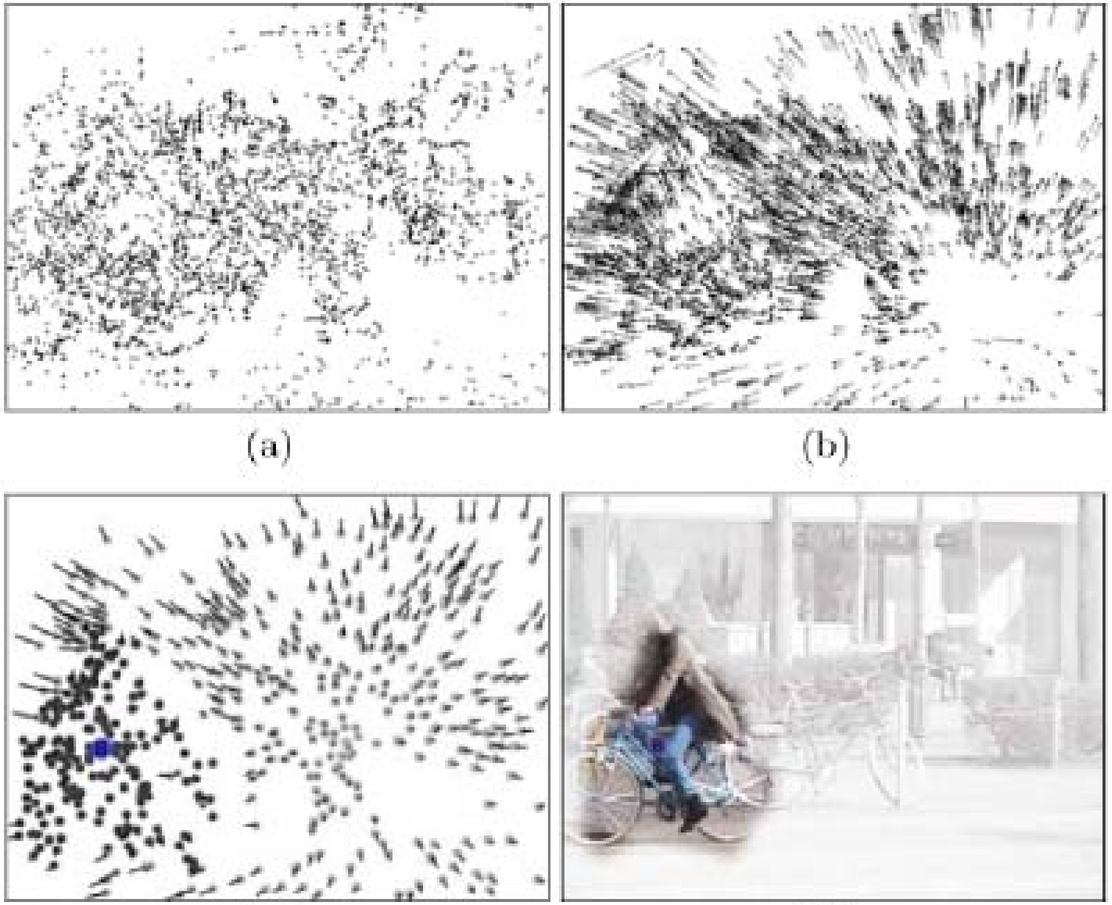
Abstract

- This paper also significantly contributes to the systems design aspect—showing how all of these subtasks can be combined in a computer vision system so that they mutually benefit from another.

Keywords

- Object Recognition
- Segmentation
- Tracking
- Videos Analysis
- Compositionality
- Visual Learning

Process on this paper

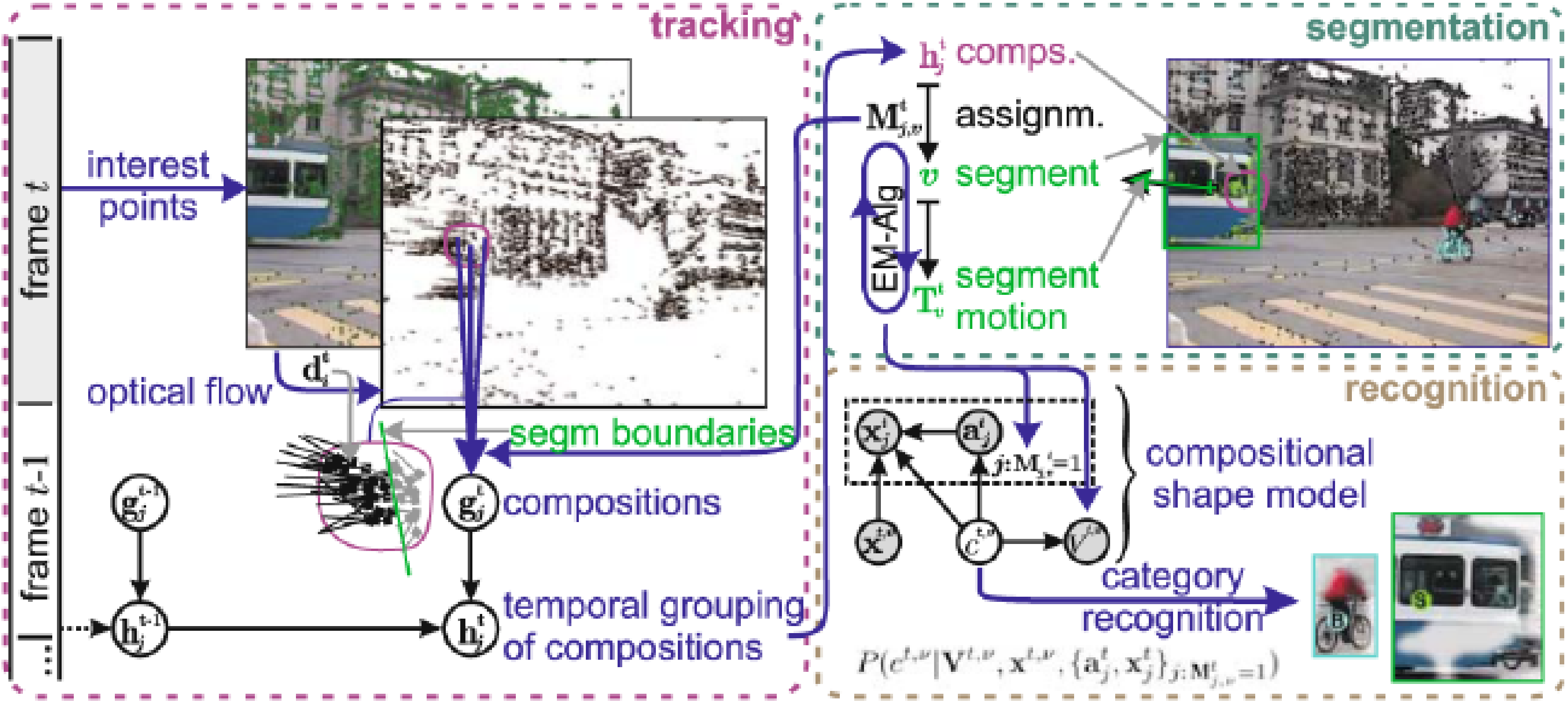


(a) (b) (c) (d)

Problem are investigated

- Category-level recognition
- Reducing supervision during learning
- Segmentation of videos from a moving camera
- Tracking without manual interaction
- Object models and shape representation

Process Pipeline



Region Tracking and Object Segmentation

- Tracking Object Regions
  - Compositions as Spatial Groupings of Parts
  - A composition represents all its constituent interest points
  - Tracking Compositions
  - Temporal Grouping of Composition

$$\Gamma^t(j) = \{i : \text{IP } i \text{ in neighborhood of } j\text{-th comp.}\}. \quad (1)$$

$$\mathbf{g}_j^t := \frac{1}{|\Gamma^t(j)|} \sum_{i \in \Gamma^t(j)} \mathbf{d}_t^i. \quad (2)$$

$$\mathbf{x}_j^{t+1} := \mathbf{x}_j^t + \frac{1}{|\Gamma^t(j)|} \sum_{i \in \Gamma^t(j)} \mathbf{d}_t^i. \quad (3)$$

$$\mathbf{h}_j^t = \eta \mathbf{g}_j^t + (1 - \eta) \mathbf{h}_j^{t-1}. \quad (4)$$

Joint Tracking and Segmentation of Objects Based on Floating Image Regions

- Problem: assemble the object regions into the different objects and into background.
- Solving: Expectation-Maximization approach (EM)
- After use EM-Algorithm: Using Segmentation to Refine Object Region Tracking Algorithm

EM-Algorithm

$\textsc{CompSegmentation}(\{\mathbf{h}_j^t\}_j, \{\mathbf{T}_\nu^{t-1}\}_{\nu=1,\ldots,K})$

Initialization: $\forall \nu : \mathbf{T}_\nu^t \leftarrow \mathbf{T}_\nu^{t-1}$
repeat
  E-Step: $\triangleright$ *update assignments:*
    $\mathbf{M}_{j,\nu}^t \leftarrow \mathbf{1}\left\{\nu = \operatorname{argmin}_{\widehat{\nu}} \mathcal{R}\left(\mathbf{T}_{\widehat{\nu}}^t, \mathbf{h}_j^t, \mathbf{x}_j^t\right)\right\}$
  M-Step: $\triangleright$ *update segments:*
    for $\nu = 1, \ldots, K$
    do Solve with Levenberg-Marquardt
      (start with $\widehat{\mathbf{T}}_\nu^t \leftarrow \mathbf{T}_\nu^t$):
      $\mathbf{T}_\nu^t \leftarrow \operatorname{argmin}_{\widehat{\alpha},\widehat{s},\widehat{\delta}_x,\widehat{\delta}_y} \sum_j \mathbf{M}_{j,\nu}^t \mathcal{R}\left(\widehat{\mathbf{T}}_\nu^t, \mathbf{h}_j^t, \mathbf{x}_j^t\right)$
until convergence of $\mathbf{M}_{j,\nu}^t$
return $\mathbf{M}^t$, $\{\mathbf{T}_\nu^t\}_{\nu=1,\ldots,K}$

Using Segmentation to Refine Object Region Tracking Algorithm

$\textsc{CompositionTracking}(\{\mathbf{h}_j^{t-1}, \mathbf{x}_j^t\}_j, \{\mathbf{T}_\nu^{t-1}\}_{\nu=1,\dots,K})$

Detect interest points i in frame t

for all compositions j $\quad \triangleright$ *update comps with IP flow:*

do $\Gamma^t(j) \leftarrow \{i : \|\mathbf{x}_j^t - \bar{\mathbf{x}}_i^t\| \leq w\}$

$\quad \mathbf{g}_j^t \leftarrow \frac{1}{|\Gamma^t(j)|} \sum_{i \in \Gamma^t(j)} \mathbf{d}_t^i$

$\quad \mathbf{h}_j^t \leftarrow \eta \mathbf{g}_j^t + (1-\eta)\mathbf{h}_j^{t-1}$

$\mathbf{M}^t, \{\mathbf{T}_\nu^t\}_\nu \leftarrow \textsc{CompSegmentation}(\{\mathbf{h}_j^t\}_j, \{\mathbf{T}_\nu^{t-1}\}_\nu)$

for all compositions j $\triangleright$ *update comps with segmentat.:*

do $\Gamma^t(j) \leftarrow \{i : i \in \Gamma^t(j) \wedge$

$\qquad 1 = \mathbf{M}^t_{j, \operatorname{argmin}_{\tilde{\nu}} \mathcal{R}(\mathbf{T}_\nu^t, \mathbf{d}_t^i, \bar{\mathbf{x}}_i^t)}$

$\quad \mathbf{g}_j^t \leftarrow \frac{1}{|\Gamma^t(j)|} \sum_{i \in \Gamma^t(j)} \mathbf{d}_t^i$

$\quad \mathbf{h}_j^t \leftarrow \eta \mathbf{g}_j^t + (1-\eta)\mathbf{h}_j^{t-1}$

$\quad \mathbf{x}_j^{t+1} \leftarrow \mathbf{x}_j^t + \frac{1}{|\Gamma^t(j)|} \sum_{i \in \Gamma^t(j)} \mathbf{d}_t^i$

return $\{\mathbf{h}_j^t, \mathbf{x}_j^{t+1}\}_j, \{\mathbf{T}_\nu^t\}_\nu$

Object Representations for Category-Level Recognition

- Compositional, Appearance-Based Model: Use multi-class SVM
- Recognition Using the Motion of Dot Patterns: Use SVM
- Global Shape and Local Appearance Combined
- Processing Pipeline for Training: Use SVM

Output


(a)


(b)


(c)


(d)

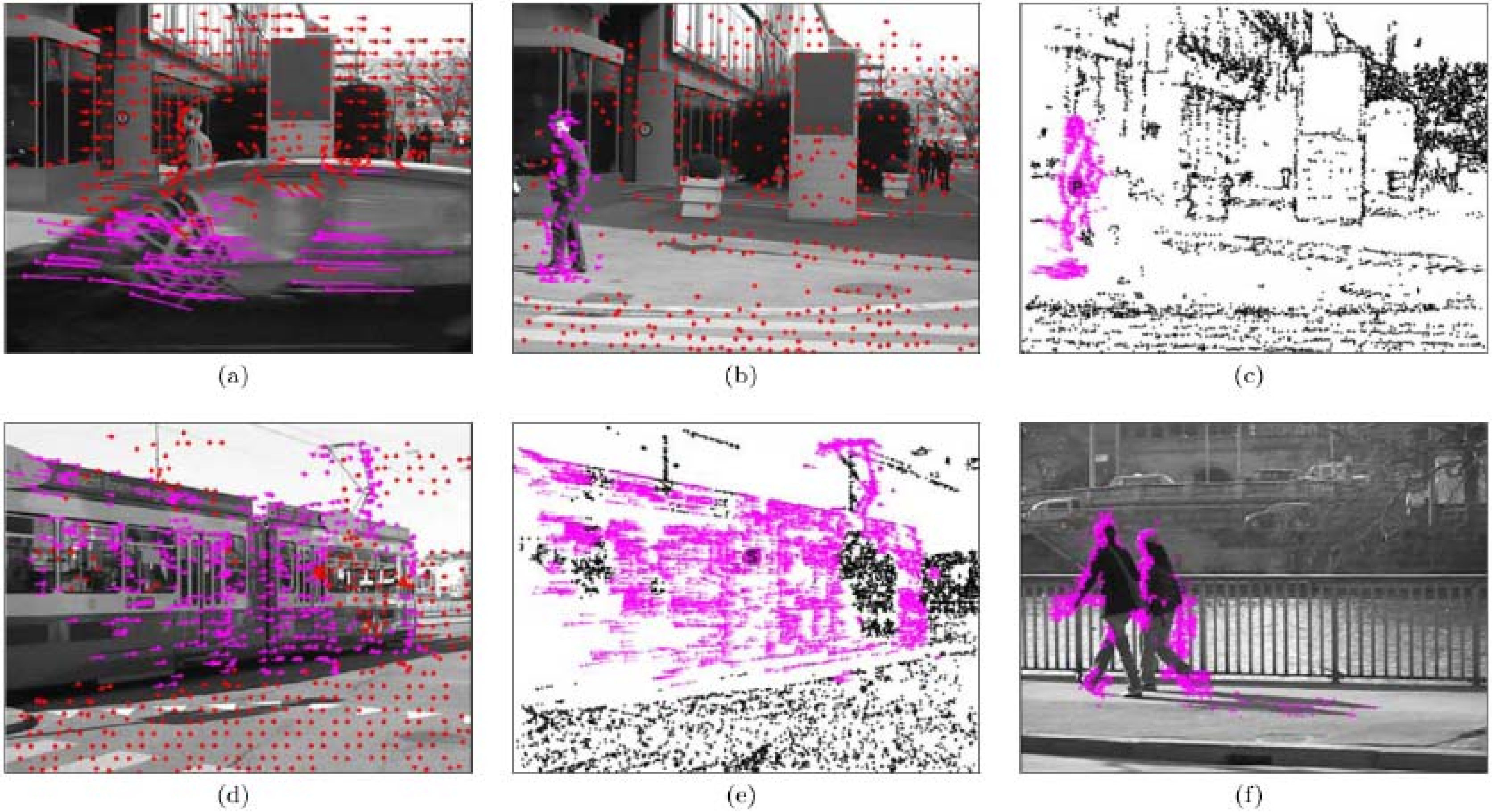

(e)


(f)


(g)


(h)

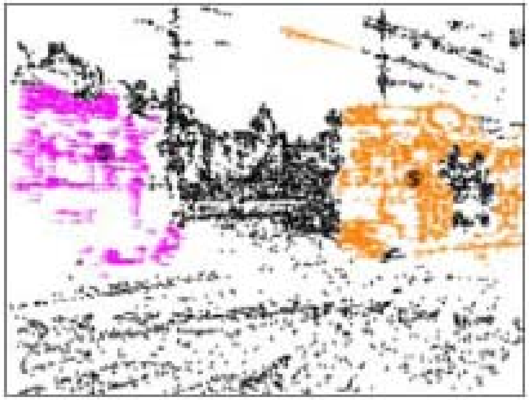
Process Output



(a) (b) (c) (d) (e) (f)

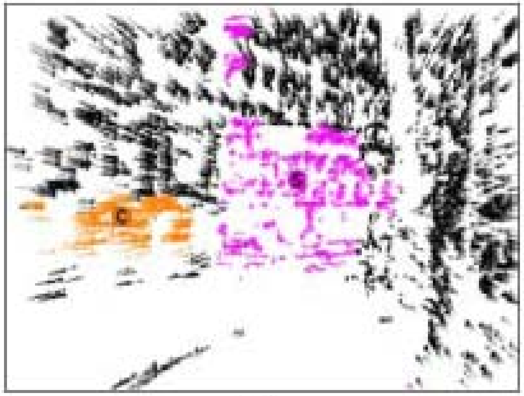
Process Output(2)

(a)



(b)

(c)



(d)

Experiments

- Recognition Performance on Videos with Substantial Camera Motion
  - use 10-fold cross-validation and train on 16 randomly
  - Object models are learn on a randomly drawn subset of 15 frames per train video

Experiment Output

- Baseline Performance of Appearance w/o Compositions and Shape—Bag-of-Parts
  - 53.0 . 5.6% of all frames correctly.
- Compositional Segmentation and Recognition w/o Shape Model
  - 64.9 . 5.4% per frame.

Experiment Output(2)

- Comparing the Different Algorithm

Object model	Per frame	Per video
Dataset of Ommer and Buhmann *(2007)* *(car, bicycle, pedestrian, streetcar)*:		
Approach of Ommer and Buhmann (2007)	74.3 ± 4.3	87.4 ± 5.8
Compositional motion (13)	52.6 ± 1.1	68.2 ± 3.4
Appearance-only: bag-of-parts	53.0 ± 5.6	58.9 ± 6.5
Segment. w/o shape: bag-of-comps	64.9 ± 5.4	78.9 ± 5.8
Shape: $P(c^{t,\nu}\|\mathbf{V}^{t,\nu})$ (11)	74.4 ± 5.3	88.4 ± 5.2
Compositional appear + location (12)	79.6 ± 5.5	90.7 ± 5.3
Combined shape + appear (14)	81.4 ± 2.9	94.5 ± 4.9
Dataset (Ommer and Buhmann 2007) *plus additional category "cow" from* (Magee and Boyle 2002):		
Compositional appearance (12)	76.5 ± 2.4	88.4 ± 2.3

Experiment Output(3)

- Comparing Different Object Models

True classes →	Bicycle	Car	Pedest	Streetcar
Bicycle	**74.3**	3.2	13.7	2.9
Car	7.8	**84.1**	4.2	5.9
Pedestrian	13.3	2.5	**80.0**	3.9
Streetcar	4.7	10.2	2.2	**87.3**

Experiment Output(4)

- Computational Demands
  - recognizes objects in videos of 768x576 pixel
  - using the combined shape and appearance model at the order of 1 fps on a 3 GHz Pentium 4 desktop PC.

Processing step	Comp. demand
Tracking and segmentation, Algorithm 2:	
IPs i, flow $\mathbf{d}_t^i$ (Algorithm 2, line 1)	27.7%
Updating comps (Algorithm 2, line 2–5)	5.2%
EM estimation Algorithm 1, i.e. (Algorithm 1, line 6)	4.9%
Updating comps with segm. (Algorithm 2, line 7–11)	0.3%
Feature extraction and recognition:	
Computing loc feat hists to represent $\mathbf{a}_j^t$ (Sect. 3.2)	36.5%
Computing all individual probs in (14)	12.3%
Eval. GM of Fig. 5, i.e. calc. product in (14)	0.09%
Video stream ops, writing of results, etc.	12.9%

Experiment Output(5)

- Action Recognition using KTH

True classes →	Box	Hclp	Hwav	Jog	Run	Walk
Boxing	**84.5**	0.0	5.5	0.0	0.0	0.0
Hand clapping	1.0	**87.0**	16.5	0.0	0.0	0.0
Hand waving	12.5	13.0	**75.5**	0.0	0.0	0.0
Jogging	0.0	0.0	0.5	**93.0**	0.0	0.0
Running	2.0	0.0	0.0	3.0	**92.3**	5.0
Walking	0.0	0.0	2.0	4.0	7.7	**95.0**