

Computer Vision 2 WS 2018/19

Part 17 – CNNs for Video Analysis I 15.01.2019

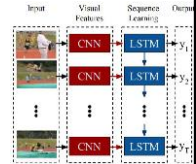
Prof. Dr. Bastian Leibe

RWTH Aachen University, Computer Vision Group
<http://www.vision.rwth-aachen.de>



Course Outline

- Single-Object Tracking
- Bayesian Filtering
- Multi-Object Tracking
- Visual Odometry
- Visual SLAM & 3D Reconstruction
 - Online SLAM methods
 - Full SLAM methods
- Deep Learning for Video Analysis
 - CNNs for video analysis
 - Optical flow
 - Video object segmentation



2 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis



Topics of This Lecture

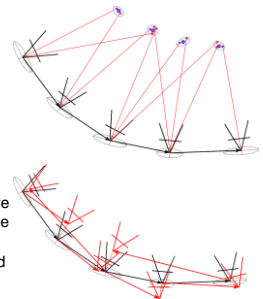
- Recap: Full SLAM methods
- CNNs for Video Analysis
 - Motivation
 - Example: Video classification
- CNN + RNN
 - RNN, LSTM
 - Example: Video captioning
- Matching and correspondence estimation
 - Metric learning
 - Correspondence networks

3 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis



Recap: Full SLAM Approaches

- **SLAM graph optimization:**
 - Joint optimization for poses and map elements from image observations of map elements and control inputs
- **Pose graph optimization:**
 - Optimization of poses from relative pose constraints deduced from the image observations
 - Map recovered from the optimized poses

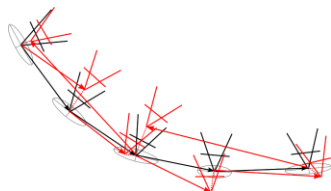


4 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis
Slide credit: Ivo Stuckler



Pose Graph Optimization

- Optimization of poses
 - From relative pose constraints deduced from the image observations
 - Map recovered from the optimized poses
- Deduce relative constraints between poses from image observations, e.g.,
 - 8-point algorithm
 - Direct image alignment



5 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis
Slide credit: Ivo Stuckler



Pose Graph Optimization Example

Dense Visual SLAM for RGB-D Cameras

Christian Kerl, Jürgen Sturm,
Daniel Cremers



Computer Vision and Pattern Recognition Group
Department of Computer Science
Technical University of Munich



Kerl et al., Dense Visual SLAM for RGB-D Cameras, IROS 2013

6 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis
Slide credit: Ivo Stuckler




Topics of This Lecture

- Recap: Full SLAM methods
- CNNs for Video Analysis**
 - Motivation
 - Example: Video classification
- CNN + RNN**
 - RNN, LSTM
 - Example: Video captioning
- Matching and correspondence estimation
 - Metric learning
 - Correspondence networks

7 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
RWTH AACHEN UNIVERSITY

Video Analysis with CNNs

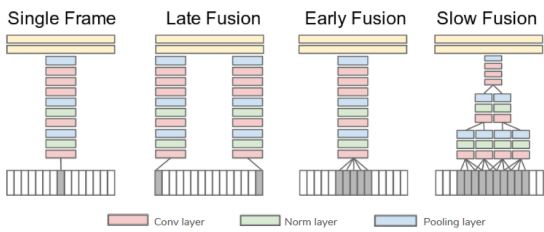


- Modeling perspective**
 - What **architecture** to use to best capture temporal patterns?
- Computational perspective**
 - Video processing is expensive!
 - How to reduce **computation cost** without sacrificing accuracy

8 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Fei-Fei Li
RWTH AACHEN UNIVERSITY

Large-Scale Video Classification with CNNs

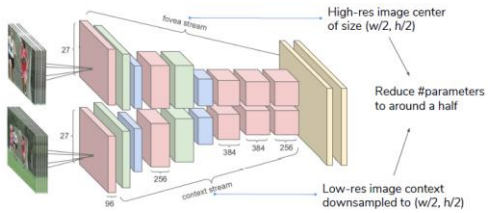
- Architecture**
 - Different ways to fuse features from multiple frames



9 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Fei-Fei Li
RWTH AACHEN UNIVERSITY

Large-Scale Video Classification with CNNs

- Computational cost**
 - Reduce spatial dimension to reduce model complexity
 - Multi-resolution: low-res context + high-res foveate



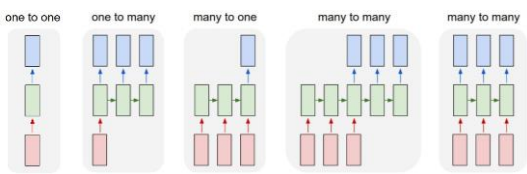
10 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
RWTH AACHEN UNIVERSITY

Topics of This Lecture

- Recap: Full SLAM methods
- CNNs for Video Analysis**
 - Motivation
 - Example: Video classification
- CNN + RNN**
 - RNN, LSTM
 - Example: Video captioning
- Matching and correspondence estimation
 - Metric learning
 - Correspondence networks

11 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
RWTH AACHEN UNIVERSITY

Recap: Recurrent Networks



- Feed-forward networks**
 - Simple neural network structure: 1-to-1 mapping of inputs to outputs
- Recurrent Neural Networks**
 - Generalize this to arbitrary mappings

12 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
RWTH AACHEN UNIVERSITY

Recap: RNNs

- RNNs are regular NNs whose hidden units have additional forward connections over time.
- You can unroll them to create a network that extends over time.
- When you do this, keep in mind that the weights for the hidden units are shared between temporal layers.

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Image source: Andrei Karpathy

Extension: Long Short-Term Memory (LSTM)

- LSTMs
 - Inspired by the design of memory cells
 - Each module has 4 layers, interacting in a special way.
 - Effect: LSTMs can learn longer dependencies (~100 steps) than RNNs

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Image source: Christopher Olah, <http://olshch.github.io/posts/2015-08-11-understanding-LSTMs/>

Recap: RNNs for Text Generation

- RNN for text generation

10,001D class scores (Softmax over 10k words and a special <END> token)
 $y_i = W_{hy} h_i$

Hidden layer (e.g., 500D vectors)
 $h_i = \max\{0, W_{xh} x_i + W_{hh} h_{i-1}\}$

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li
Image source: Andrei Karpathy

Recap: RNNs for Text Generation

- Training this on a lot of sentences would give us a language model.
- I.e., a way to predict
 $p(\text{next word} | \text{previous words})$

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li

Recap: RNNs for Text Generation

- Training this on a lot of sentences would give us a language model.
- I.e., a way to predict
 $p(\text{next word} | \text{previous words})$

sample!

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li

Recap: RNNs for Text Generation

- Training this on a lot of sentences would give us a language model.
- I.e., a way to predict
 $p(\text{next word} | \text{previous words})$

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li

Recap: RNNs for Text Generation

- Training this on a lot of sentences would give us a language model.
- I.e., a way to predict $p(\text{next word} \mid \text{previous words})$

Visual Computing Institute | Prof. Dr. Bastian
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li

Recap: RNNs for Text Generation

- Training this on a lot of sentences would give us a language model.
- I.e., a way to predict $p(\text{next word} \mid \text{previous words})$

Visual Computing Institute | Prof. Dr. Bastian
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li

Recap: RNNs for Text Generation

- Training this on a lot of sentences would give us a language model.
- I.e., a way to predict $p(\text{next word} \mid \text{previous words})$

Visual Computing Institute | Prof. Dr. Bastian
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li

Recap: RNNs for Text Generation

- Training this on a lot of sentences would give us a language model.
- I.e., a way to predict $p(\text{next word} \mid \text{previous words})$

Visual Computing Institute | Prof. Dr. Bastian
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li

Recap: RNNs for Text Generation

- Training this on a lot of sentences would give us a language model.
- I.e., a way to predict $p(\text{next word} \mid \text{previous words})$

Visual Computing Institute | Prof. Dr. Bastian
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li

Recap: RNNs for Text Generation

- Training this on a lot of sentences would give us a language model.
- I.e., a way to predict $p(\text{next word} \mid \text{previous words})$

Visual Computing Institute | Prof. Dr. Bastian
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Andrei Karpathy, Fei-Fei Li

Applications: Image Tagging

- Simple combination of CNN and RNN
 - Use CNN to define initial state h_0 of an RNN.
 - Use RNN to produce text description of the image.

25 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide adapted from Andrei Karpathy

Applications: Image Tagging

- Setup
 - Train on corpus of images with textual descriptions
 - E.g. Microsoft CoCo
 - 120k images
 - 5 sentences each

26 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide adapted from Andrei Karpathy

Results: Image Tagging

a group of people standing around a room with remotes
logprob: -9.17

a young boy is holding a baseball bat
logprob: -7.61

a cow is standing in the middle of a street
logprob: -8.54

Spectacular results!

27 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide adapted from Andrei Karpathy

Results: Image Tagging

a baby laying on a bed with a stuffed bear
logprob: -8.66

a young boy is holding a baseball bat
logprob: -7.65

a cat is sitting on a couch with a remote control
logprob: -11.65

- Wrong, but one can still see why those results were selected...

28 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide adapted from Andrei Karpathy

Application: Video to Text Description

Our LSTM network is connected to a CNN for RGB frames or a CNN for optical flow images.

Flow images

CNN - Action pretrained

29 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Source: Subhasthi Venugopalan, ICCV14

Video-to-Text Results

Correct descriptions.	Relevant but incorrect descriptions.	Irrelevant descriptions.
<p>SZVT: A man is doing stunts on his bike.</p>	<p>SZVT: A small bus is running into a building.</p>	<p>SZVT: A man is pouring liquid in a pan.</p>
<p>SZVT: A herd of zebras are walking in a field.</p>	<p>SZVT: A man is cutting a piece of a pair of paper.</p>	<p>SZVT: A polar bear is walking on a hill.</p>
<p>SZVT: A young woman is doing her hair.</p>	<p>SZVT: A cat is trying to get a small board.</p>	<p>SZVT: A man is doing a pencil.</p>
<p>SZVT: A man is shooting a gun at a target.</p>	<p>SZVT: A man is spreading butter on a tortilla.</p>	<p>SZVT: A black clip to walking through a path.</p>

30 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Source: Subhasthi Venugopalan, ICCV14

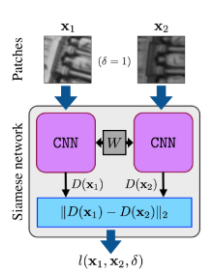
Topics of This Lecture

- Recap: Full SLAM methods
- CNNs for Video Analysis
 - Motivation
 - Example: Video classification
- CNN + RNN
 - RNN, LSTM
 - Example: Video captioning
- Matching and correspondence estimation
 - Metric learning
 - Correspondence networks

31 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
RWTH AACHEN UNIVERSITY

Learning Similarity Functions

- Siamese Network
 - Present the two stimuli to two identical copies of a network (with shared parameters)
 - Train them to output similar values if the inputs are (semantically) similar.
- Used for many matching tasks
 - Face identification
 - Stereo estimation
 - Optical flow
 - ...

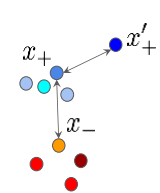


32 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
RWTH AACHEN UNIVERSITY

Metric Learning: Contrastive Loss

- Mapping an image to a metric embedding space
 - Metric space: distance relationship = class membership

$$\|f(x) - f(x_+)\| \rightarrow 0$$

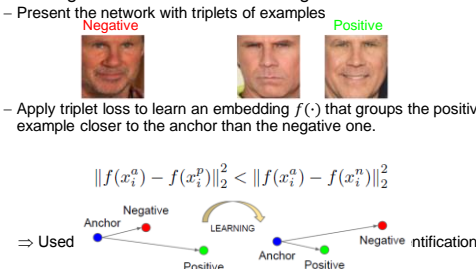
$$\|f(x) - f(x_-)\| \geq m$$


Yi et al., LIFT: Learned Invariant Feature Transform, ECCV 16

33 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Christopher Choy
RWTH AACHEN UNIVERSITY

Metric Learning: Triplet Loss

- Learning a discriminative embedding
 - Present the network with triplets of examples
- Apply triplet loss to learn an embedding $f(\cdot)$ that groups the positive example closer to the anchor than the negative one.




\Rightarrow Used for Negative identification

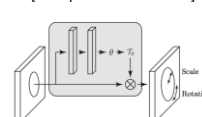
34 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
RWTH AACHEN UNIVERSITY

Patch Normalization with Spatial Transformer Nets

- Patch Normalization
 - Key component of local feature matching
 - Finding the scale and rotation
 - Invariant to perspective transformation
- Spatial Transformer Network
 - Adaptively apply transformation



[SIFT patch normalization]

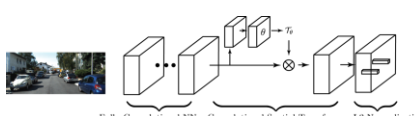


[Spatial Transformer Network]

35 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Christopher Choy
RWTH AACHEN UNIVERSITY

Universal Correspondence Network

- Computing a patch descriptor



36 Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 - CNNs for Video Analysis
Slide credit: Christopher Choy
RWTH AACHEN UNIVERSITY

Universal Correspondence Network

- Siamese architecture for matching patches

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis
Slide credit: Christopher Choy

Universal Correspondence Network

- UCN Training

- Contrastive loss

$$\|f(x_+) - f(x'_+)\| \rightarrow 0$$

$$\|f(x_-) - f(x'_-)\| > m$$

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis
Slide credit: Christopher Choy

Semantic Correspondences with UCN

Ground truth UCN VGG Conv4

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis
Slide credit: Christopher Choy

Exact Correspondences with UCN (Disparity Estimation)

C. Choy, J.Y. Gwak, S. Savarese, M. Chandraker, *Universal Correspondence Network*, NIPS'16

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis
Slide credit: Christopher Choy

References and Further Reading

- RNNs
 - R. Pascanu, T. Mikolov, Y. Bengio, [On the difficulty of training recurrent neural networks](#), JMLR, Vol. 28, 2013.
 - A. Karpathy, [The Unreasonable Effectiveness of Recurrent Neural Networks](#), blog post, May 2015.
- LSTM
 - S. Hochreiter, J. Schmidhuber, [Long short-term memory](#), Neural Computation, Vol. 9(8): 1735–1780, 1997.
 - A. Graves, [Generating Sequences With Recurrent Neural Networks](#), ArXiv 1308.0850v5, 2014.
 - C. Olah, [Understanding LSTM Networks](#), blog post, August 2015.

Visual Computing Institute | Prof. Dr. Bastian Leibe
Computer Vision 2
Part 17 – CNNs for Video Analysis