Research & Development
–DRAFT–

Evaluation of current Approaches for Situation-Awareness in Autonomous Systems from Action Recognition in Video Data

Author:
Maximilian Schöbel

Advisors:
Prof. Dr. Erwin Prassler
Prof. Dr. Paul G. Plöger

January 12th, 2017

Contents

1 Introduction

META: Brief but concise review of the first two "W"s.

The operation of autonomous mobile systems in public, uncontrolled environments is despite active research still a difficult task.

Humans are perfectly able to act and move in unknown, crowded environments and even react succesfully to new situations because they are aware of their surroundings.

An important part of Situation Awareness is the knowledge of what actions are currently performed by persons in the vicinity of an agent.

Actions of interest are single-person actions, person-person interactions, person-object interactions and group activities.

Enabling situation-awareness in autonomous systems is an important goal, which has an impact on other problems in autonomous systems.

Possible applications:
Pedestrian movement predicition in robotics,
Risk and danger evaluation through video surveillance in public environements,
surveillance of children or the elderly in assisted living environments,
patient monitoring in hospitals,
video retrieval (content-based video indexing),
human-computer interaction.

Requirement: Automated recognition of high-level actions.

Situation awareness is an abstract concept, which includes lots of independent manifestations and involves multiple sensory inputs.

This work focuses on approaches that process time-sequential video data, because videocameras represent a cost-effective and widely used technology in many existing systems and are able to capture a lot of information.

1.1 Situation Awareness from video data

General Definition of Situation Awareness in the context of autonomous systems.

Placement of Action Recognition among other vision-based methods, i.e. Action Prediction, Anomaly Detection, Event and Action Detection, Person/Pedestrian Detection, Gesture Recognition.

Definitions of the above methods.

Simple case: Video contains the performance of a single human action which needs to be classified into one of several preknown classses.

General real-world case: System operates on a video stream and needs to perform continuous recognition of human actions, including detection of beginning and endings times of containing acions.

General Processing Pipeline for Action Recognition: Person Detection -> Tracking -> Action Detection -> Segmentation -> Action recognition.

Action Recognition: A part of Computer Vision research, it's goal is to automatically analyze human actions/actitvities from video-data.

Other sensory input than video possible

1.2 Survey Papers in Action Recognition (Related work)

Review of most important/recent review papers in Action Recognition with traditional and Deep Learning approaches.

1.2.1 A survey on vision-based human action recognition, Ronald Poppe (2010)

Definition of action: Uses the hierarchical classification of human motion in action primitives, actions and activities as given in Moeslund et al. (cite ??)

Action primitives are atomic movements at the limb-level.

Actions are possibly cyclic whole body movements and consist of multiple action primitives.

Activities consist of multiple actions whose subsequent execution make the movement interpretable.

Example: Action primitives: Left/right leg forward -> Action: Starting, Running, Jumping -> Activity: Jumping hurdles.

Scope: Gives a very good classification of conventional methods in human action recognition.

The discussion is split according to video representations and classification methods.

Challenges of the domain are described very well.

Deficits: No Deep Learning methods are discussed.

Datasets and benchmarks are only discussed briefly.

1.2.2 Human Activity Analysis: A Review – Aggarwal and Ryoo (2011)

Gives an approach-based taxonomy.

1.2.3 A survey on vision-based methods for action representation, segmentation and recognition – Weinland et al. (2011)

1.2.4 A survey of video datasets for human action and activity recognition – Chaquet et al. (2013)

1.2.5 A review of unsupervised feature learning and deep learning for time-series modeling – Längkvist et al. (2014)

1.2.6 Going Deeper into Action Recognition: A survey – Herath et al. (2016)

Definition of action:

1.3 Challenges in Action Recognition

Action Recognition is a classification-task.

Intra- and inter-class variances.

Background and recording settings.

Temporal variations.

Obtaining and labeling training data.

Difference to face/gate recognition: Generalize over person characteristics.

Main task of action recognition research: Overcome these challenges and built systems, that recognize actions robustly, even when performed by different persons in differently lighted environments at different speeds.

Main components: (i) A discriminative architecture that is able to recognise the general characteristics of different action classes while ignoring personal characteristics of different performers. (ii) Large datasets that provide this information by containing many different examples for each action class.

2 Conventional Methods in Action Recognition

META: Condensed overview and description of conventional Methods in action Recognition using the taxonomy of Aggarwal and Ryoo's fine survey paper. More detailed description of methods using local-features, since these have become the standard approach in action recognition after Aggarwal and Ryoo's overview.

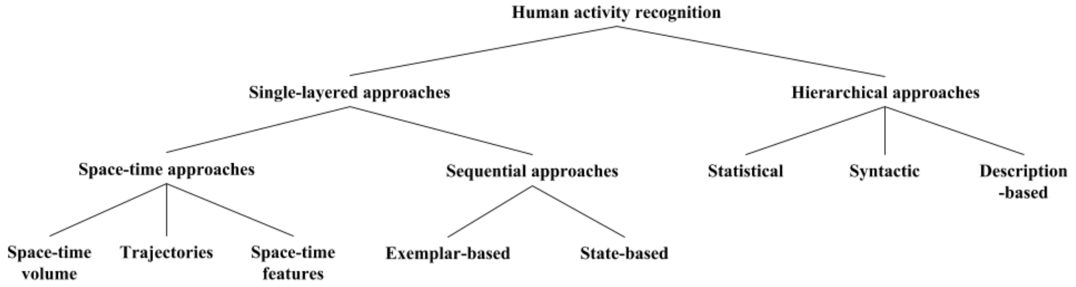

Abb. 1: Approach-based taxonomy for conventional methods in human activity recognition as given by Aggarwal and Ryoo[1]

3 Main components in action recognition using local features: Feature Extraction, Representation Building, Classification.

Methods for feature extraction: Interest point detectors or dense sampling.

Space-time interest point detectors: Harris3D[2], Cuboids[3], Hessian Detector[4]

Descriptors for 3D volumes around previously detected space-time interest points: Histogram of Gradient HOG[5], Histogram of Optical Flow (HOF)[6], 3D Histogram of Gradient (HOG3D)[7], Extended SURF (ESURF)[4]

3 Deep Learning Methods in Action Recognition

Review of approaches that use Deep Learning.

3.1 Spatio-Temporal Networks

I.e. convolutional methods.

3.2 Multiple Stream Networks

The most successfull architecture at action recognition. They are equally powerful as the improved dense trajectories approach. cite TDD

3.3 Generative Models

Restricted Boltzmann Machine

3.4 Temporal Coherency Networks

4 Datasets and Benchmarks in Action Recognition

4.1 Review of Datasets for Human Action Classification

Review of the most important currently existing datasets, focus on newest ones (since 2013)

Reference dataset survey paper.

4.2 Data Augmentation

4.3 Alternative Benchmarks for Action Recognition Algorithms

4.4 Inter-Dataset Approaches

5 Evaluation

What do we need, what do we have, what is best suited so far?

References

[1] Jake K. Aggarwal and Michael S. Ryoo. "Human Activity Analysis: A Review". In: *ACM Computing Surveys (CSUR)* 43.3 (2011). 01121, p. 16. URL: `http://dl.acm.org/citation.cfm?id=1922653` (visited on 05/23/2016).

[2] Ivan Laptev. "On Space-Time Interest Points". In: *International Journal of Computer Vision* 64 (2-3 2005). 02614, pp. 107–123. URL: `http://link.springer.com/article/10.1007/s11263-005-1838-7` (visited on 06/01/2016).

[3] Piotr Dollár et al. "Behavior Recognition via Sparse Spatio-Temporal Features". In: *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on.* 02076. IEEE, 2005, pp. 65–72. URL: `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1570899` (visited on 05/16/2016).

[4] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector". In: *European Conference on Computer Vision.* 00647. Springer, 2008, pp. 650–663. URL: `http://link.springer.com/chapter/10.1007/978-3-540-88688-4_48` (visited on 10/18/2016).

[5] Navneet Dalal and Bill Triggs. "Histograms of Oriented Gradients for Human Detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).* Vol. 1. 15268. IEEE, 2005, pp. 886–893. URL: `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1467360` (visited on 07/18/2016).

[6] Ivan Laptev et al. "Learning Realistic Human Actions from Movies". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* 02233. IEEE, 2008, pp. 1–8. URL: `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4587756` (visited on 05/25/2016).

[7] Alexander Klaser, Marcin Marsza\lek, and Cordelia Schmid. "A Spatio-Temporal Descriptor Based on 3d-Gradients". In: *BMVC 2008-19th British Machine Vision Conference.* 00929. British Machine Vision Association, 2008, pp. 275–1. URL: `https://hal.inria.fr/inria-00514853/` (visited on 10/18/2016).