Temporal Relations in Videos for Unsupervised Activity Analysis

Fabian Nater[1]
fnater@vision.ee.ethz.ch
Helmut Grabner[1]
grabner@vision.ee.ethz.ch
Luc Van Gool[1,2]
vangool@vision.ee.ethz.ch

[1]Computer Vision Laboratory
ETH Zurich, Switzerland
[2]ESAT-PSI / IBBT
K.U. Leuven, Belgium

Figure 1: In videos, each frame strongly correlates with its neighbors. Our approach exploits this fact and enables the segmentation of the video and the interpretation of unseen sequences.

Observing the different video sequences in Fig. 1, increments between frames are quite small compared to the changes throughout the whole sequence. For instance, the behavior of a tracked person (2^{nd} row) is composed of a certain repertoire of activities with transitions in between that are typically short in comparison. This can also be observed at larger scales, like day-night changes or seasonal changes (3^{rd} and 4^{th} row) and already suggests a hierarchical structure.

Contributions. In this work, we seek for an 'invariant characteristic' that can underpin the analysis of activities in an automatic manner. The contributions of our paper are twofold:

- We propose an unsupervised technique to segment the data into compact and meaningful activities. To this end, we explore the strong temporal relations in the video. The automatically discovered activities are efficiently represented and continuously refined in a hierarchical manner.
- Analysis and interpretation of unseen data is demonstrated as a result of the coarse to fine representation in the hierarchy that enables abnormal event detection. Anomalies can be spotted, such as the big tent in a street festival (3^{rd} row in Fig. 1).

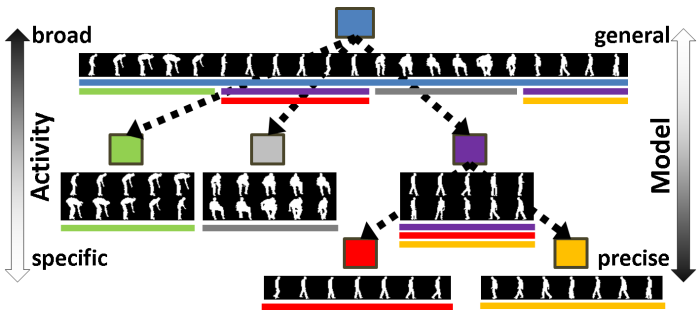

Figure 2: Overview of the proposed hierarchical model that automatically splits and represents the data in a coarse to fine manner. As an example, we consider indoor actions. At the top node, the entire video stream is taken into account, while at lower levels, more specific concepts, like picking up, or walking leftwards are found.

Concept. Due to the large variety of observations in a data stream, it often is difficult to build a single model which describes the data and its dynamic behavior precisely. In this work, we automatically split the data stream into meaningful subsequences. We call these subsequences *activities*. If they are consistent and have low complexity, they can be represented more easily and precisely. This principle is exploited by arranging the video data in a hierarchical manner as outlined in Fig. 2. In a long data-stream, some activities may be very distinct and can be segmented high up, while more subtle differences only appear deeper down.

In order to build up such a hierarchy, we use the strong link between temporally adjacent observations in time series data, such as videos. Hence, we characterize activities to have a certain duration, to be observed frequently, and to be interconnected by shorter transitions. In other words, *with high probability, neighboring frames share their activity label.*

Outline. In the paper, we detail our approach to unsupervised activity summarization that builds up the hierarchical model. We show that our technique is purely data-driven and feature-independent, but still extracts semantically interpretable hierarchies. A second part of the paper is dedicated to the application of the obtained model to unseen data, highlighting the usefulness of the established hierarchy for anomaly detection. To show the wide applicability of our approach in practice, we use four different datasets from recent works in visual surveillance, *i.e.*, human action segmentation [4], traffic analysis [2], monitoring of a public place in time lapse data [1], and human behavior analysis [3]. Due to the accurate modelling of activities from their temporal characteristics, our results are competitive with state-of-the-art in all cases.

[1] M. D. Breitenstein, H. Grabner, and L. Van Gool. Hunting Nessie – Real-Time Abnormality Detection from Webcams. In *ICCV WS on Visual Surveillance*, 2009.

[2] T. Hospedales, S. Gong, and T. Xiang. A Markov Clustering Topic Model for mining behaviour in video. In *Proc. ICCV*, 2009.

[3] F. Nater, H. Grabner, and L. Van Gool. Exploiting Simple Hierarchies for Unsupervised Human Behavior Analysis. In *Proc. CVPR*, 2010.

[4] P. Turaga, A. Veeraraghavan, and R. Chellappa. Unsupervised view and rate invariant clustering of video sequences. *CVIU*, 113(3):353–371, 2009.