A Grid-based Representation for Human Action Recognition

Soufiane Lamghari[§], Guillaume-Alexandre Bilodeau[§], Nicolas Saunier[‡]

[§]LITIV Lab., Department of Computer and Software Engineering, Polytechnique Montreal, Montreal, Canada
[‡]Department of Civil, Geological and Mining Engineering, Polytechnique Montreal, Montreal, Canada

ICPR 2020 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION Milan, Italy 10 | 15 January 2021

Abstract

In this work, we propose a novel method for human action recognition that encodes efficiently the most discriminative appearance information of an action with explicit attention on representative pose features, into a new compact grid representation. Our **GRAR** (Grid-based Representation for Action Recognition) method is tested on several benchmark datasets demonstrating that our model can accurately recognize human actions, despite intra-class appearance variations and occlusion challenges.

Introduction

Human Action Recognition (HAR)

- The goal is to **identify and classify human actions** from a sequence of visual observations that contains spatial and temporal information related to the human action;
- This work has various applications, including video surveillance, sport video analysis and urban planning.

Challenges

Realistic scenes usually contain:

- Various types of elements and contexts;
- Intra-class appearance variations;
- Different motion speeds;
- Occlusions.

Deep learning approaches

- The main focus of the majority of research studies in HAR is to extend Convolutional Neural Networks (CNNs) to explore the temporal information contained in videos.

Limitations

Most of existing approaches for HAR are unable to represent human actions in an efficient way:

- Do not properly model the **temporal** information with respect to the dynamic nature of human actions;
- Rely on some **irrelevant features** that do not properly identify the performed action.

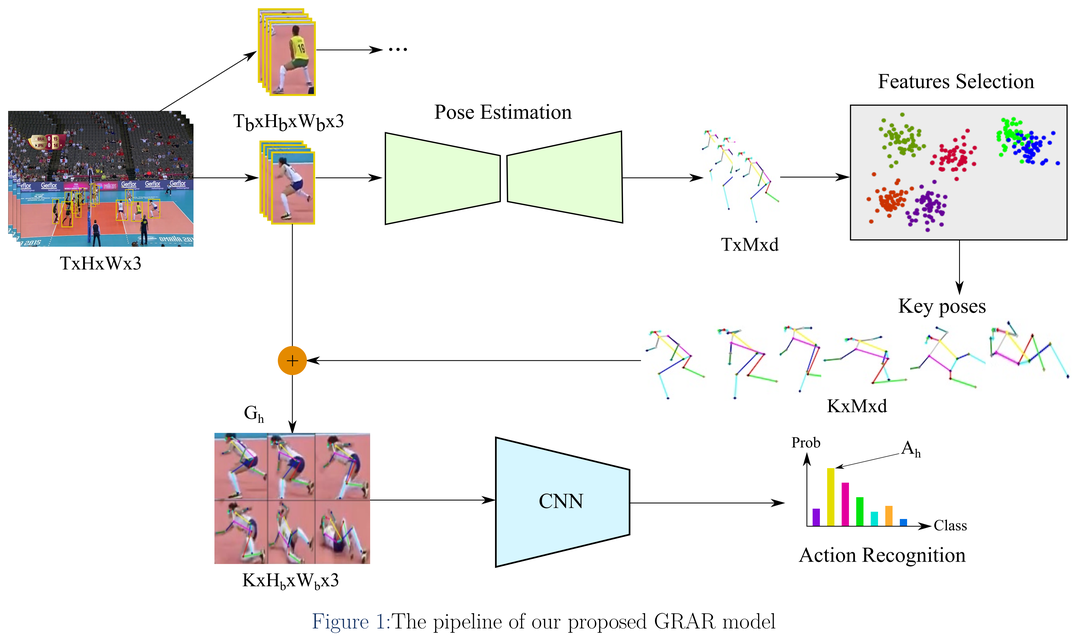

Figure 1:The pipeline of our proposed GRAR model

Proposed approach

We propose a novel **pose-based** approach for HAR that learns discriminative features of actions by integrating them into a **grid representation**.

- To restrict the analysis to only the most likely information related to the human action:
  $\Longrightarrow$ We only consider the **human region** in the scene for each frame, instead of the entire frame.
- To integrate the temporal discriminative features of the performed action:
  $\Longrightarrow$ We use **key poses** obtained by **clustering** the sequence of the 2D human keypoints, estimated at each time step.
- To obtain a relevant representation of actions:
  $\Longrightarrow$ We fuse valuable appearance features with representative poses into a **single grid image**, which effectively encodes an explicit attention guided by key poses.
- To recognize human actions:
  $\Longrightarrow$ We train a **CNN** on the obtained grid representations.

Experimental results

- We focus on **atomic actions** (e.g., running, dancing, jumping)

Table 1:Results on the Collective Activity dataset [31]

Method	Accuracy
Choi et al. [38]	70.9%
Tran et al. [43]	78.7%
Ibrahim et al. [4]	81.5%
Deng et al. [32]	81.2%
Shu et al. [27]	87.2%
Qi et al. [29]	89.1%
Zhang et al. [44]	83.8%
Lu et al. [45]	90.6%
Wu et al. [46]	91.0%
GRAR (Ours)	**91.5**%

Table 2:Results on the Collective Activity Extended dataset [38]

Method	Accuracy
Choi et al. [38]	82.0%
Tran et al. [43]	80.7%
Ibrahim et al. [4]	94.2%
Deng et al. [32]	90.2%
Qi et al. [29]	89.7%
Lu et al. [45]	91.2%
Zhang et al. [44]	96.2%
GRAR (Ours)	**97.4**%

Table 3:Results on the Volleyball dataset [4]

Method	Accuracy
Ibrahim et al. [4]	75.9%
Shu et al. [27]	69.0%
Bagautdinov et al. [28]	82.4%
Qi et al. [29]	81.9%
Biswas et al. [30]	76.6%
Wu et al. [46]	**83.1**%
GRAR (Ours)	82.9%

Conclusion

- Our model **generalizes well** to different scenes;
- We effectively deal with **action's periodicity** and **incorrect human poses estimation**;
- The attention-guided by pose successfully handles **intra-class action variations** and **occlusions** challenges;
- We exploit powerful CNN architectures designed for image classification tasks without requiring any **architectural changes**.

Acknowledgements

NSERC CRSNG

NVIDIA