A Grid-based Representation for Human Action Recognition

Soufiane Lamghari, Guillaume-Alexandre Bilodeau and Nicolas Saunier

Polytechnique Montreal,
Montreal, Canada

January 2021

POLYTECHNIQUE MONTRÉAL
TECHNOLOGICAL UNIVERSITY

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

Human Action Recognition

- Consists in understanding actions performed by humans based on a sequence of visual observations.
- Various applications:
    - Smart video surveillance;
    - Sport video analysis;
    - Urban planning;
    - Autonomous robots.

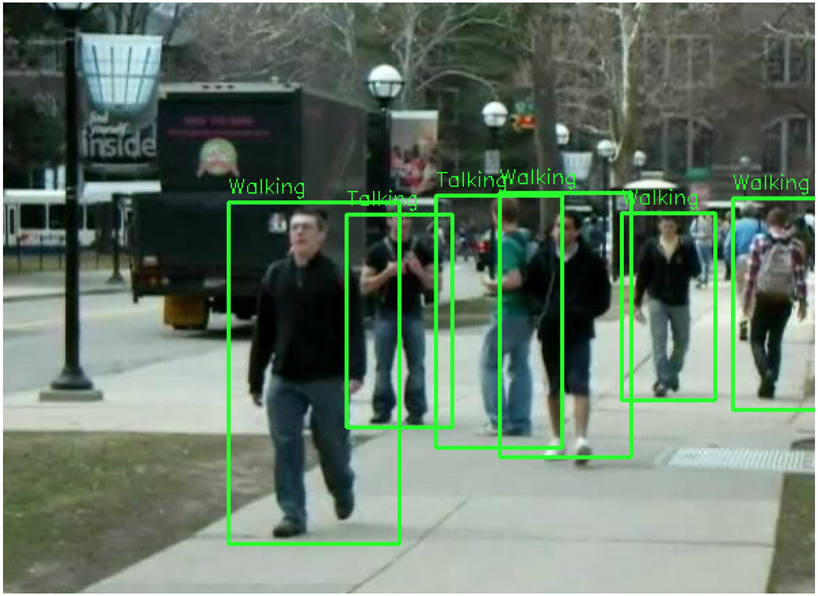


Figure 1: Sample frame from the Collective Activity dataset [1] with the ground-truth bounding boxes in green.

Human Action Recognition Challenges

- Human action recognition is challenging in realistic scenes due to:
  - Various types of elements and contexts;
  - Intra-class appearance variations;
  - Different motion speeds;
  - Occlusions.

Human Action Recognition Challenges

- Human action recognition is challenging in realistic scenes due to:
  - Various types of elements and contexts;
  - Intra-class appearance variations;
  - Different motion speeds;
  - Occlusions.
- Most of existing deep learning-based approaches do not properly model the temporal information and still represent actions by randomly learned features.

Proposed Approach

- GRAR: a novel pose-based approach for human action recognition.
- We consider an explicit attention mechanism that highlights the representative poses of the action.

GRAR Model Architecture

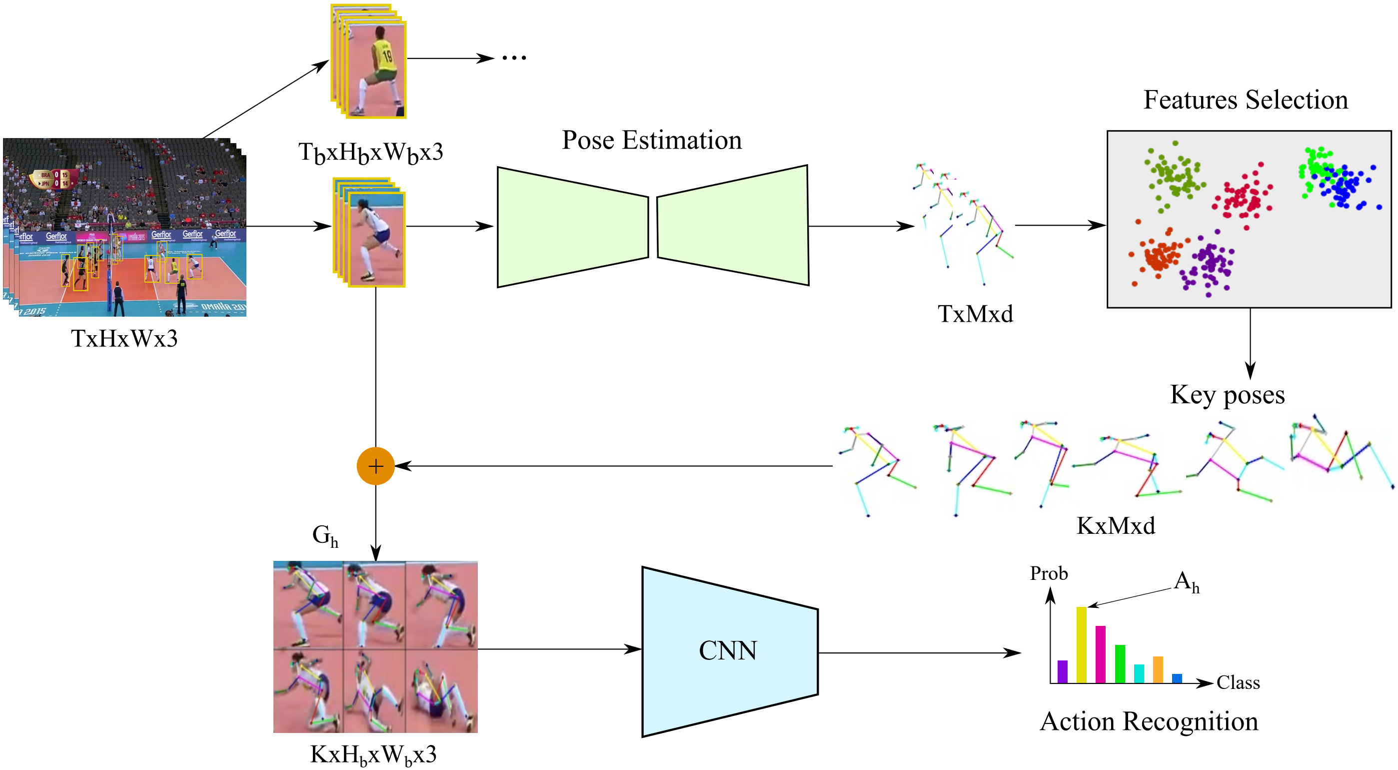
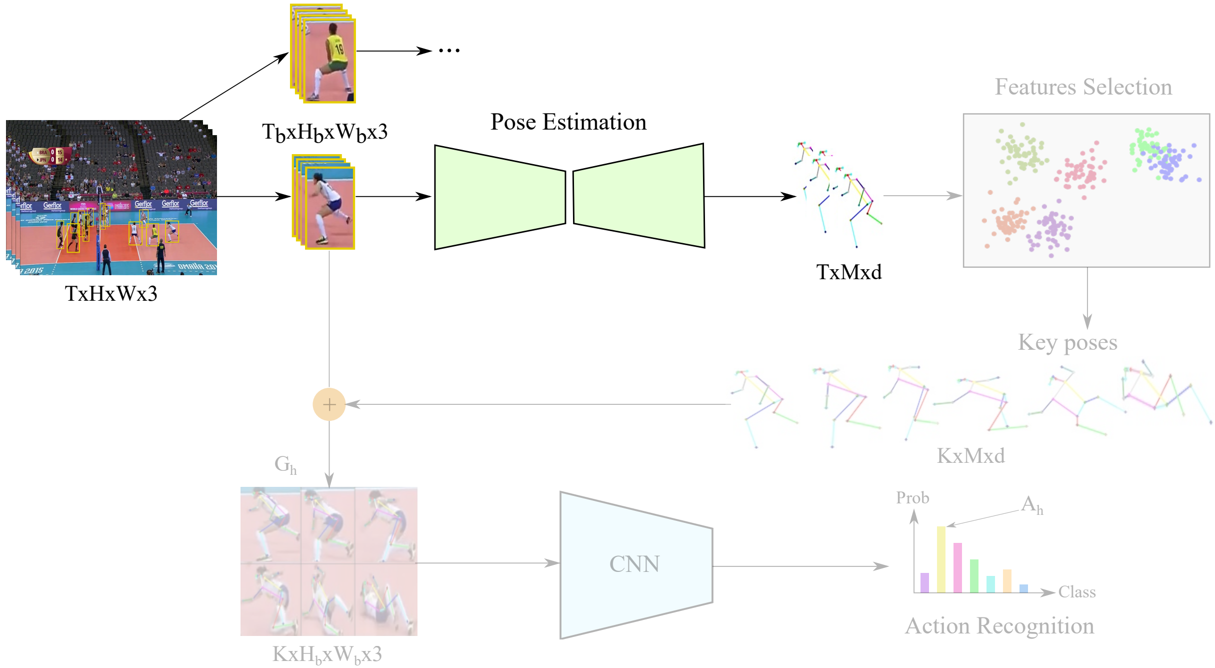


Figure 2: The pipeline of our proposed GRAR model.

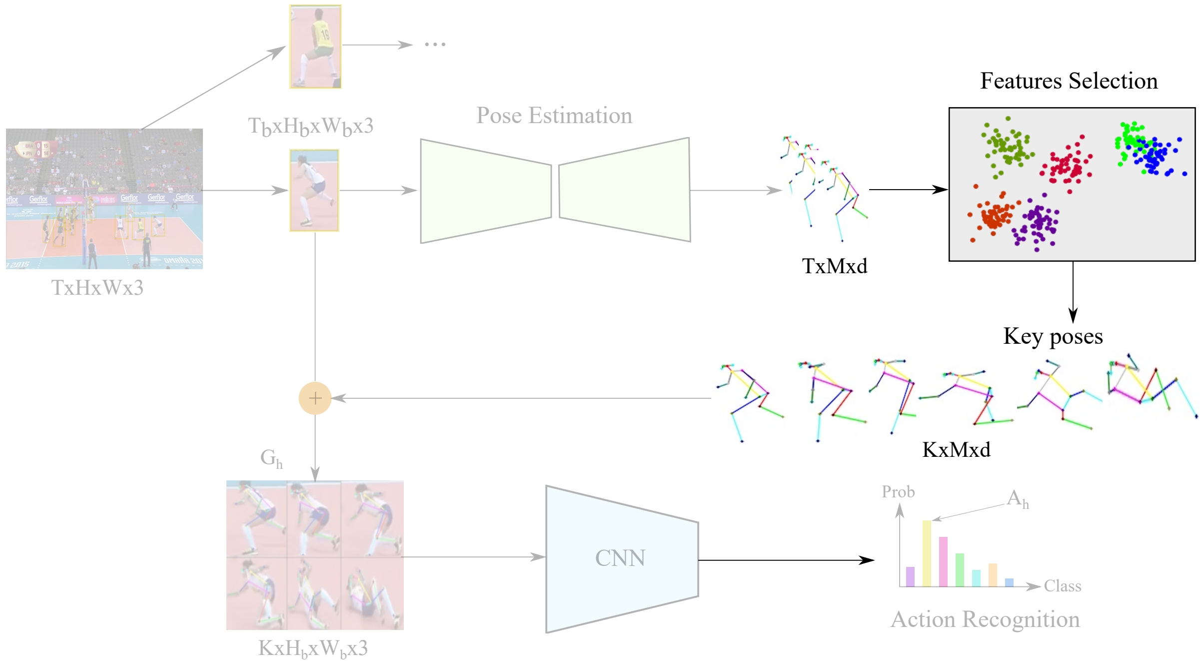
Human Pose Estimation



- Human actions are highly correlated with their corresponding poses:
  - 2D human keypoints.
  - HRNet.

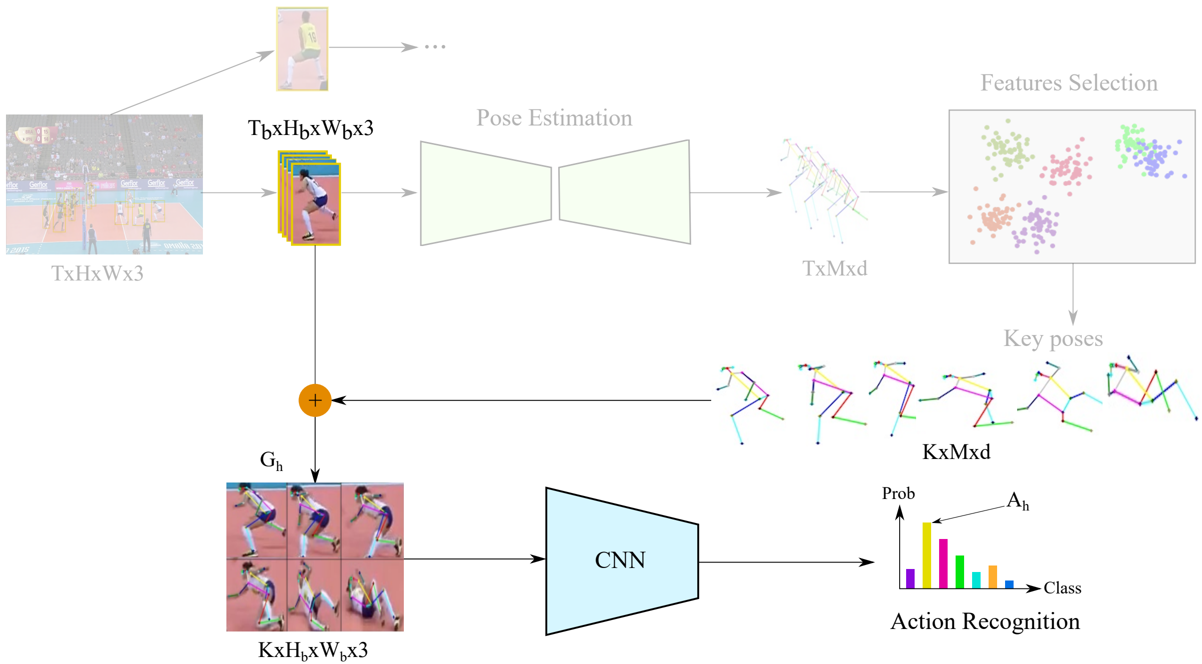
    + Bounding box refinement.

Relevant Features Selection



- To integrate the temporal representative information of the performed action:
  - Key poses.
  - Clustering.

Grid Representation Learning



- To represent actions given discriminative temporal RGB and pose information:
  - Grid representation.
  - Pre-trained CNN.

Datasets

- Collective Activity (CA) dataset [1]
  - 5 action categories (talking, crossing, queuing, waiting and walking).
- Collective Activity Extended (CAE) dataset [2]
  - 6 action categories (talking, crossing, queuing, waiting, jogging and dancing).
- Volleyball dataset [3]
  - 9 individual actions (moving, spiking, waiting, blocking, jumping, setting, falling, digging and standing).

Results on the CA and CAE datasets

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

Method	Accuracy
Choi et al. [2]	70.9%
Tran et al. [4]	78.7%
Ibrahim et al. [3]	81.5%
Deng et al. [5]	81.2%
Shu et al. [6]	87.2%
Qi et al. [7]	89.1%
Zhang et al. [8]	83.8%
Lu et al. [9]	90.6%
Wu et al. [10]	91.0%
GRAR (Ours)	**91.5**%

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

Method	Accuracy
Choi et al. [2]	82.0%
Tran et al. [4]	80.7%
Ibrahim et al. [3]	94.2%
Deng et al. [5]	90.2%
Qi et al. [7]	89.7%
Lu et al. [9]	91.2%
Zhang et al. [8]	96.2%
GRAR (Ours)	**97.4**%

Results on the Volleyball dataset

Table 3: Results on the Volleyball dataset [3]

Method	Accuracy
Ibrahim et al. [3]	75.9%
Shu et al. [6]	69.0%
Bagautdinov et al. [11]	82.4%
Qi et al. [7]	81.9%
Biswas et al. [12]	76.6%
Wu et al. [10]	**83.1**%
GRAR (Ours)	82.9%

Ablation study

Table 4: Impact of different modules on the accuracy of GRAR based on the CAE dataset.

Model Variants	Accuracy
Random	89.2%
Key poses only (K-Pose)	80.5%
Key Frame (K-RGB)	92.3%
Key Frame+Box enhancement (K-RGB+EB)	92.9%
Key Frame+Box enhancement+Pose Attention (K-RGB+EB+PA)	**95.2**%

Conclusion

- We presented GRAR, a novel pose-based model for human action recognition.
- 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.

Acknowledgement

We thank the National Sciences and Engineering Research Council of Canada (NSERC) for their support, and NVIDIA Corporation for their donation of a Titan Xp GPU.

References

- W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pp. 1282–1289, IEEE, 2009.
- W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *CVPR 2011*, pp. 3273–3280, IEEE, 2011.
- M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1971–1980, 2016.
- K. N. Tran, A. Bedagkar-Gala, I. A. Kakadiaris, and S. K. Shah, "Social cues in group formation and local interactions for collective activity analysis.," in *VISAPP (1)*, pp. 539–548, 2013.

Thank You!