Vision-Based Multi-Modal Framework for Action Recognition

Beddiar Djamila Romaissa[1,2], Oussalah Mourad[1], Nini Brahim[2]

1 Center for Machine Vision and Signal analysis, Oulu, Finland

2 Laarbi Ben M'hidi University, Oum El Bouaghi, Algeria

Introduction

- Major vision-based HAR works focus on using one single sensor modality to classify activities.
- However, many limitations while discriminating complex such as lighting, perspective changes and occlusions could influence on the recognition accuracy.
- To achieve good results and enable robust HAR systems, we exploit more than one modality.

Contributions

- Summarizing RGB and depth videos into dynamic images using an approximate rank pooling method.
- Encoding locations of skeleton joints into new representations.
- Transfer learning from pre-trained models for feature extraction and Feature fusion based on Canonical Correlation Analysis of RGB, depth and skeleton data.
- Classification of actions with bi-directional LSTM using the resulting features fusion vectors.

Proposed Multi-modal Method

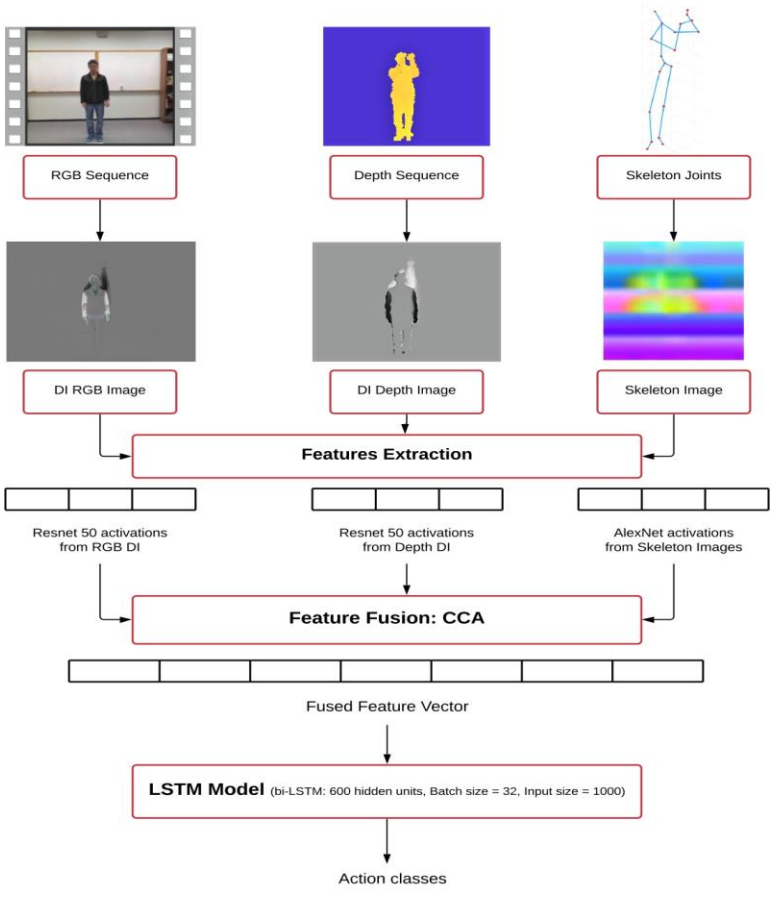

Figure1: general overview of our proposed vision-based multi-modal approach for HAR

Experiments & Results

Table1: Accuracy (%) of activity classification with LSTM of uni-modal features and features from our newly created images on the UTD-MHAD and NTU RGB+D datasets.

Uni-modal feature	UTD-MHAD	NTU RGB+D
RGB	51.35	39.85
Depth	37.45	45.90
Skeletal data	74.52	49.91
Dynamic RGB	72.28	41.53
Dynamic Depth	71.91	51.66
Skeleton images	87.43	50.81

Table2: Accuracy (%) of activity classification using fusion of multi-modal features extracted from our newly created images on the UTD-MHAD dataset and NTU RGB+D datasets.

Pairwise Fusion	UTD-MHAD	NTU RGB+D
DI RGB + DI Depth	85.39	60.42
DI RGB + Skeleton images	93.26	68.62
DI Depth + Skeleton images	**97.95**	**70.85**

By three Fusion	UTD-MHAD	NTU RGB+D
(DI RGB + DI Depth) + Skeleton images	**98.88**	**75.50**
(DI RGB + Skeleton images) + DI Depth	92.13	73.72
(DI Depth + Skeleton images) + DI RGB	93.26	72.64

Conclusion

- We proposed a vision-based multi-modality fusion approach for human activity recognition based on RGB dynamic images, depth dynamic images and skeleton images.
- Transfer learning inline with LSTM were used to extract significative features and classify actions.
- The experiments indicated that the proposed method yields in superior performance compared to baseline methods.

Acknowledgment

This work is partly supported by the Algerian Residential Training Program Abroad Outstanding National Program (PNE) that supported the first author stay at University of Oulu and European YougRes project (Ref. 823701), which are gratefully acknowledged.