25TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION

Vision-Based Multi-Modal Framework for Action Recognition

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

Djamila.Beddiar@oulu.fi

1 Center for Machine Vision and Signal analysis, Oulu, Finland
2 Laarbi Ben M'hidi University, Oum El Bouaghi, Algeria

Outline

Introduction

Human activity recognition (HAR)

Identification and categorization of recorded data into well-defined basic activity.

Activity Detection

Temporally localizing the movements of the person in the scene.

Activity Classification

Distinguishing the nature of a person movements using some spatial and temporal cues or any other meaningful features and assigning it to its corresponding class.

Introduction

Vision-based HAR has become a very active research topic in computer vision and image processing due to its wide application fields.

- automatic video surveillance,
- public security,
- virtual and augmented reality,
- health care and home monitoring,
- human-computer interaction and robot learning ...

Vision-based Human Activity Recognition

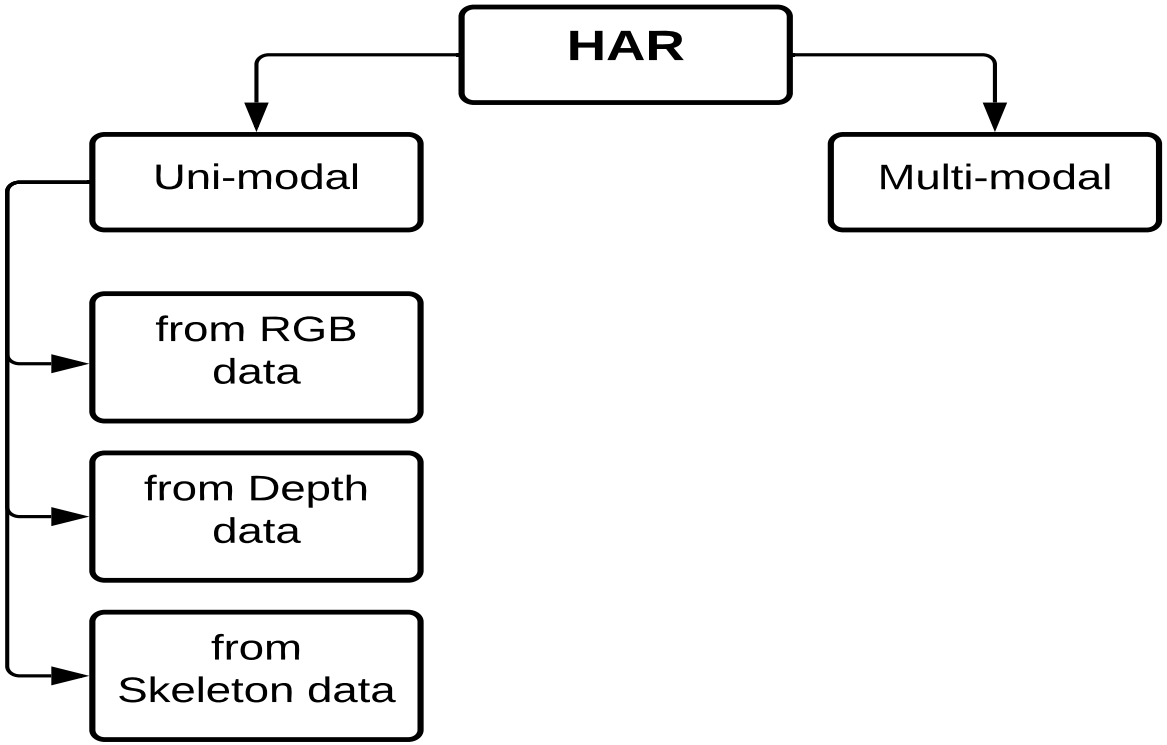


Figure 2.1: Human Activity Recognition according to the nature of the data

Multi-modal Human Activity Recognition I

1 Why data fusion?

- Limitations while discriminating complex activities due to environment conditions (Rahmani and Mian, 2016).
- To increase the robustness and the reliability of the recognition system while reducing single sensor effects (Nweke et al., 2019).
- It is essential to provide a complementary highly discriminative fusion of the modalities.
- Many fusion strategies: Feature-level fusion, through increasing feature-space, projecting on some external frame, or using correlation-like analysis (Nweke et al., 2019).
  - Canonical Correlation Analysis: learn from heterogeneous data and afford high linear correlation outputs.

Multi-modal Human Activity Recognition II

2 Why RGB, depth and skeleton data?

- Depth data, skeleton information and RGB images provide important complementary features;
- Depth data is more robust to illumination changes and scale variation but sensitive to occlusion;
- Depth cameras can overcome some inherent privacy and limitations issues related to traditional cameras;
- Skeleton information is more robust to occlusion effects;
- 3D locations and angles of joints are common features that can be used to build robust skeleton representations;
- RGB image provides fine-grained image segmentation;

Multi-modal Human Activity Recognition III

3 Why video representation?

- Efficient action representation is the key to yield robust and expressive features;
- Video as a spatial-temporal volume by stacking frames over a given sequence and action recognition is performed based on either spatial or temporal features or both;
- Rank pooling in videos (Fernando et al., 2015, 2016) allows us to capture the video-wide temporal evolution while preserving actions execution temporal ordering.

Dynamic RGB and Depth Images

Definition (Dynamic Image)

Dynamic image (DI) consists of a single image representation of a video sequence, capturing the temporal evolution of ongoing action (Fernando et al., 2015; Bilen et al., 2017).

- DIs focus on the motion instead of background pixels which are averaged away,
- DIs behave differently for actions of different speeds,
- DIs are reminiscent of some other imaging effects such as blur and panning.

- - - > Approximated Rank Pooling Method

Dynamic RGB and Depth Images

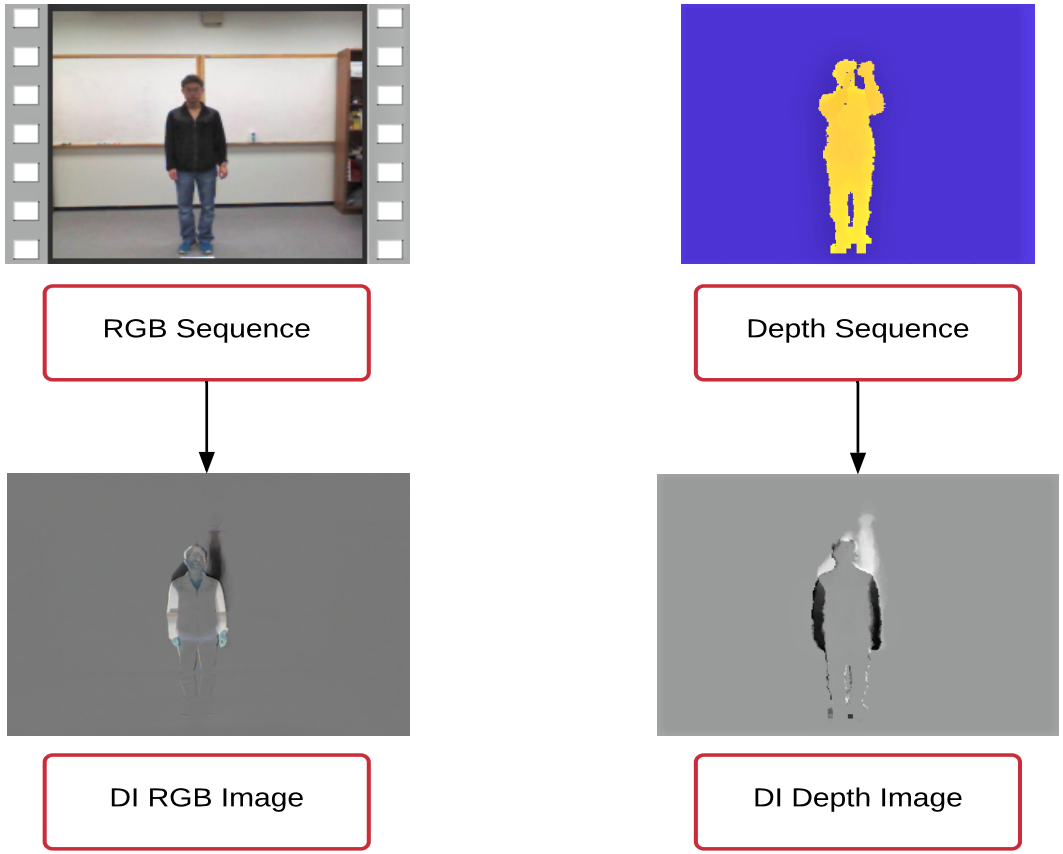
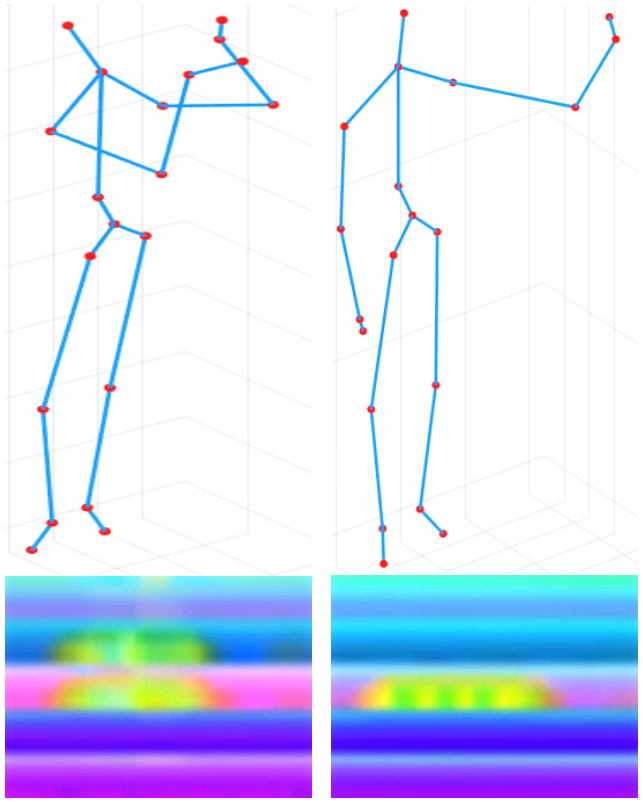


Figure 3.1: Dynamic RGB and Dynamic Depth images from RGB and Depth images

Skeleton Images

For each video sequence:

1. We normalize the coordinates of the skeleton joints *(x,y,z)*,
2. Create RGB image using X as first channel, Y as second channel and Z as third channel.



(a) (b)

Figure 3.2: Some generated skeleton images from skeleton joints.

Feature Extraction with Pre-trained Models

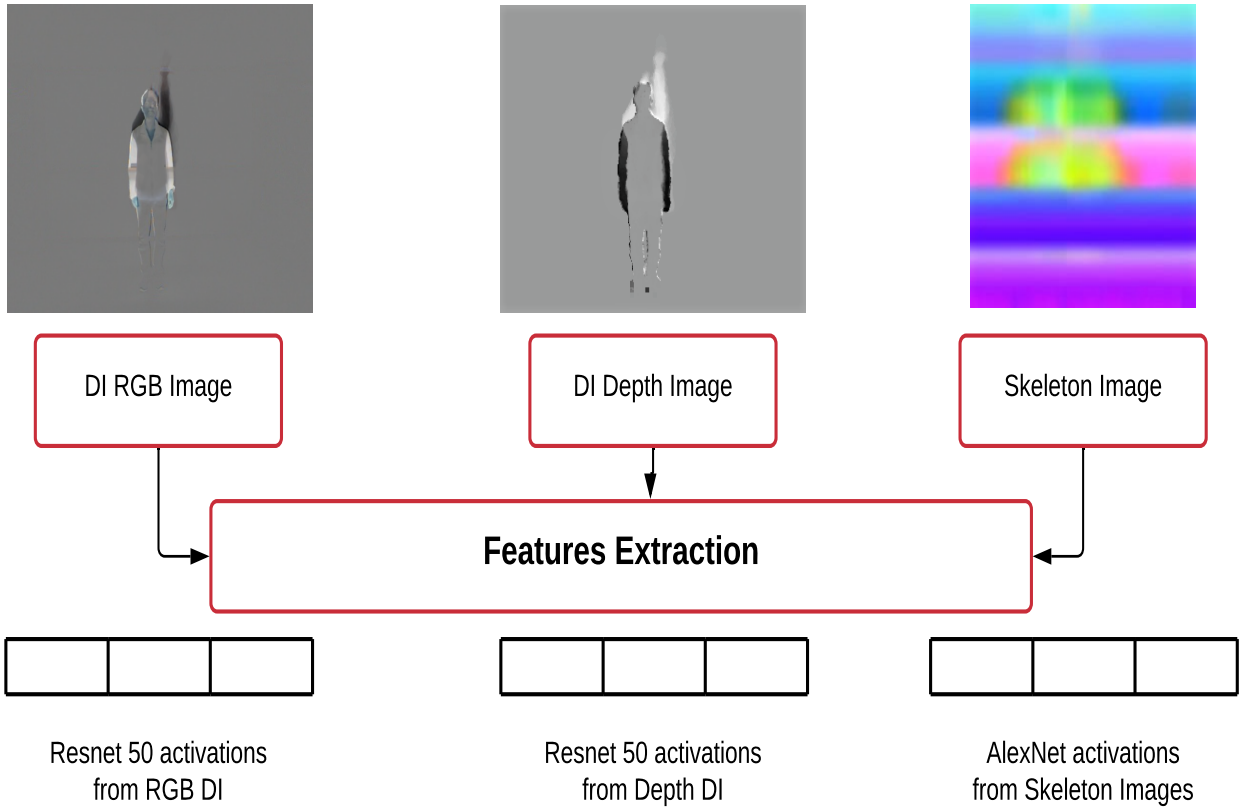


Figure 3.3: Feature extraction from DI RGB, DI Depth and Skeleton images using Resnet50 and Alexnet

Feature Fusion with Canonical Correlation Analysis

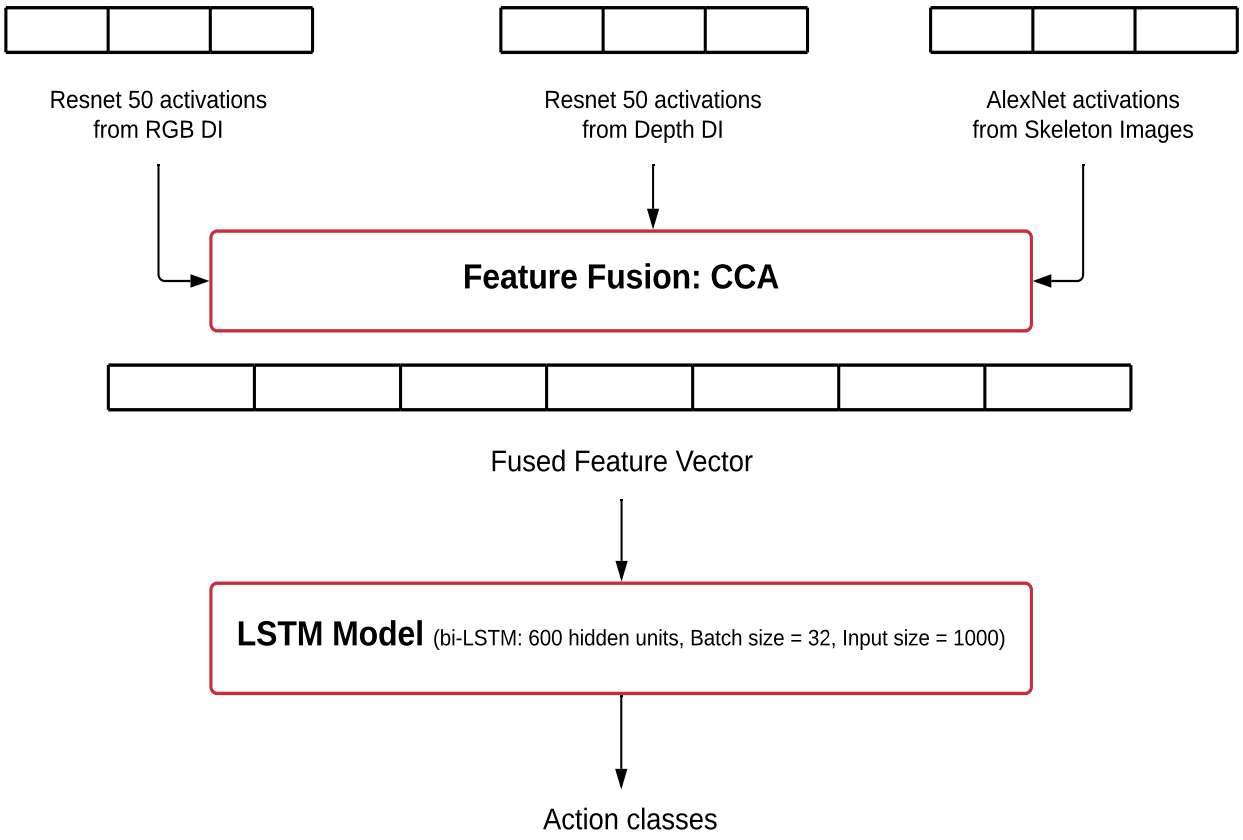


Figure 3.4: Canonical Correlation Analysis for feature fusion and classification with LSTM

Experimental Results

1. We calculate the performance accuracy of each single modality.
2. We compare activity classification from straightforward images towards newly created images (dynamic RGB, depth images and skeleton images).
3. We calculate the recognition accuracy for each pairwise fusion and for the three features fusion.
4. We also investigate the order of fusing features.
5. Finally, we compare the results of our method to the state-of-the-art on the publicly available UTD-MHAD and NTU-RGB datasets.

Datasets

UTD-MHAD

- Multi-modal dataset (Chen et al., 2015),
- Four data modalities: RGB, depth, skeleton and inertial signals,
- 861 video sequences,
- Microsoft Kinect sensor and wearable inertial sensor,
- 27 actions by 8 subjects (4 times).
  - Training: subjects 1, 3, 5, 7
  - Testing: subjects 2, 4, 6, 8

NTU RGB+D

- Large-scale multi-modal dataset (Shahroudy et al., 2016),
- 56880 videos,
- 60 action classes of 40 subjects (twice),
- Three Microsoft Kinect v2 sensors.
  - Training: cameras 2 and 3
  - Testing: camera 1

Results of classification with uni-modal features

Table 1: Accuracy (%) of activity classification with LSTM of uni-modal features and features extracted (using pre-trained models) from our newly created image representations 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

Results of classification with multi-modal features

Table 2: Accuracy (%) of activity classification using fusion of multi-modal features extracted (using pre-trained models) from our newly created image representations on the UTD-MHAD dataset and NTU RGB+D dataset respectively (DI refers to dynamic images).

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

Comparison with state-of-the-art methods

Table 3: Comparison of the proposed method with previous methods on UTD-MHAD Dataset.

Method	Accuracy %
Decision Fusion Using LOGP (Bulbul et al., 2015)	88.40
Depth + inertial data fusion + CRC (Chen et al., 2015)	79.10
5-CNN fusion of skeleton images (Khaire et al., 2018)	95.38
fusion with CCA and KELM (Imran and Raman, 2020)	97.91
DI RGB + DI Depth + Skeleton images + LSTM (Ours)	**98.88**

Comparison with state-of-the-art methods

Table 4: Comparison of the proposed method with previous methods on NTU RGB+D Dataset.

Method	Accuracy %
Deep RNN (Shahroudy et al., 2016)	64.09%
Deep LSTM (Shahroudy et al., 2016)	67.29%
Joint trajectory maps + CNN (Wang et al., 2016)	75.20%
Part-aware LSTM (Shahroudy et al., 2016)	70.20%
DI RGB + DI Depth + Skeleton images + LSTM (Ours)	**75.50%**

Conclusion

- A vision-based multi-modality fusion approach for human activity recognition.
- RGB dynamic images, depth dynamic images and skeleton images are constructed.
- Automatic features are extracted from the newly constructed visual images using pre-trained models.
- CCA (feature fusion strategy) was employed to select highly discriminative features.
- The resulting feature fusion vectors are then fed to a bi-directional LSTM network to recognize and classify activities.

Conclusion

- Our results can achieve high recognition accuracy and outperform the state-of-the-art results for both datasets.
- Will explore other fusion schemes, integrate some data augmentation methods and use more fine-gained optimization of the LSTM parameters.

Thank You.

References I

Bilen, H., Fernando, B., Gavves, E., and Vedaldi, A. (2017). Action recognition with dynamic image networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2799–2813.

Bulbul, M. F., Jiang, Y., and Ma, J. (2015). Dmms-based multiple features fusion for human action recognition. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 6(4):23–39.

Chen, C., Jafari, R., and Kehtarnavaz, N. (2015). Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE.

References II

Fernando, B., Gavves, E., Oramas, J., Ghodrati, A., and Tuytelaars, T. (2016). Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):773–787.

Fernando, B., Gavves, E., Oramas, J. M., Ghodrati, A., and Tuytelaars, T. (2015). Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387.

Imran, J. and Raman, B. (2020). Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):189–208.

References III

Khaire, P., Imran, J., and Kumar, P. (2018). Human activity recognition by fusion of rgb, depth, and skeletal data. In *Proceedings of 2nd International Conference on Computer Vision & Image Processing*, pages 409–421. Springer.

Nweke, H. F., Teh, Y. W., Mujtaba, G., and Al-Garadi, M. A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion*, 46:147–170.

Rahmani, H. and Mian, A. (2016). 3d action recognition from novel viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1506–1515.

References IV

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019.

Wang, P., Li, Z., Hou, Y., and Li, W. (2016). Action recognition based on joint trajectory maps using convolutional neural networks. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 102–106.