Sensor Fusion and Cross Modal Transfer in Human Action Recognition

Abhi Kamboj, Minh Do
Department of Electrical and Computer Engineering, College of Engineering, University of Illinois at Urbana-Champaign

INTRODUCTION

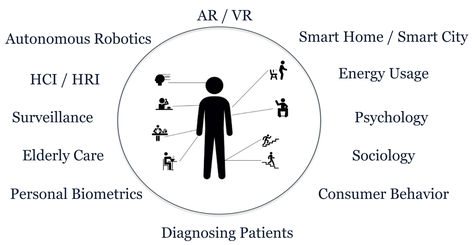
Despite living in a multi-sensory world, most AI models are limited to textual and visual interpretations of human motion and behavior. Inertial measurement units (IMUs) provide a salient signal to understand human motion; however, in practice, they have been understudied due to numerous difficulties including the uninterpretability of the data to humans. In fact, full situational awareness of human motion could best be understood through a combination of visual and physical motion sensors.

We investigate a method to merge, and transfer learned knowledge between IMU data and RGB videos for Human Action Recognition (HAR), i.e., sensor fusion and cross-modal transfer learning. Using techniques from multimodal representation learning and feature-level sensor fusion, we create a system that trains and transfers knowledge when only one modality is present and fuses knowledge from both modalities when both are present. This novel sensor fusion and cross-modal transfer (SF-CMT) system performs zero-shot HAR when transferring across modalities. Understanding human actions using IMUs (common in watches, phones, and earbuds) is a fundamental step in identifying subtle variations in motions, potentially indicating underlying health conditions or aging-related changes.

MOTIVATION

Human Action Recognition

The human action recognition or detection problem involves a model classifying or describing a brief human motion. Below are some applications of such a system [1].



Benefits of Multi-Modal Models

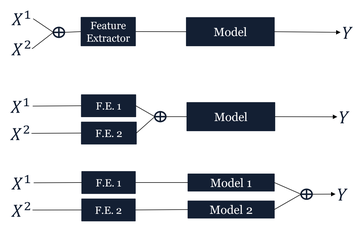
Different sensors provide various strengths and weakness that a model can leverage to perform better in more situations. The table below demonstrates how the strengths and weakness of IMU and RGB camera modalities complement each other.

IMU	RGB
Small form factor	Contactless ambient sensor
Privacy preserving	Visual identification abilities
Low dimensional data	High dimensional detail
Insensitive to lighting	Provides color information
Motion based sensor	Stationary sensor

BACKGROUND

Sensor Fusion [2]

X^1 = Modality 1 Space
X^2 = Modality 2 Space
Y = Label Space
$\oplus$ = Sum or Average



Cross-Modal Transfer Learning

Problem Setup:

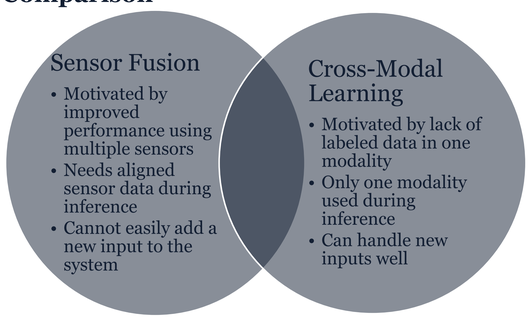
$$\mathcal{D}^S = \{(x_i^S, y_i^S)\}_{i=0}^{M}, x_i \in X^S, y_i^S \in Y^S$$
$$\mathcal{D}^T = \{(x_i^T, y_i^T)\}_{i=0}^{N}, x_i \in X^T, y_i^T \in Y^T$$

$\mathcal{D}$ = Dataset
S = Source
T = Target
X = Input Space
Y = Label Space

Transfer Learning [3]:
We have a model $f: X^S \rightarrow Y^S$, and we want to transfer its knowledge to construct a model $g: X^T \rightarrow Y^T$

- Inductive Transfer: ($X^S = X^T$ or $X^S \neq X^T$) and $Y^S \neq Y^T$
- Transductive Transfer: $X^S \neq X^T$ and $Y^S = Y^T$
  - Domain Adaptation: A method of transudative transfer that can include adapting across domains in the same modality, e.g. different body positions of an IMU [4].
  - Cross Modal Transfer: A method of transductive transfer where the input modalities are different
    - Instance-based transfer: Learn a mapping between each modality's input space
    - Feature-based transfer: Align intermediate representation and translate between them

Comparison



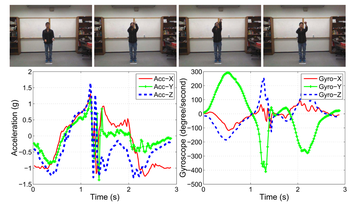
METHOD

Problem Statement

State of the art works leverage either sensor fusion, or cross modal learning, not both. Sensor fusion models fail to train or perform when one modality is missing. Cross-modal transfer cannot leverage multiple modalities during inference when multiple are present.

Dataset

The University of Texas Dallas Multimodal Human Action Dataset (UTD-MHAD) consists of 27 actions performed by 8 subjects 4 times each, which is about 860 data sequences recorded in 4 modalities [5] . We only use the RGB and IMU modalities.

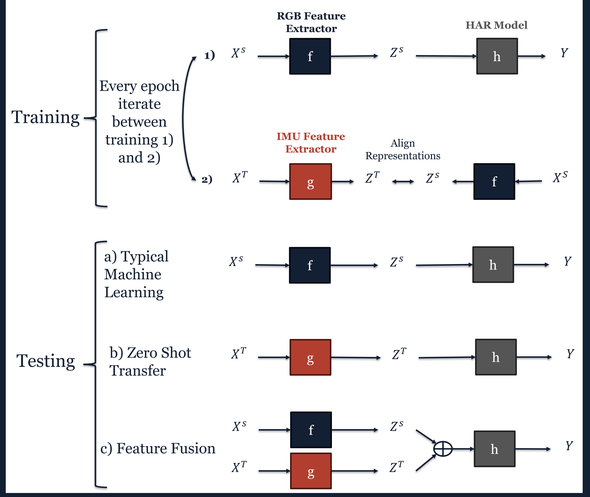


Experiment Steps

1) 45% of data (X^{RGB}, Y) was used to train an RGB HAR model with a cross entropy loss function
2) 45% of data (X^{RGB}, X^{IMU}) was used to align representations with a contrastive loss, as in CLIP [6]
3) 10% of data was used thrice, for three different tests
   a) Regular model evaluation: RGB (X^{RGB}, Y)
   b) Zero shot transfer: IMU (X^{IMU}, Y)
   c) Feature fusion: IMU+RGB (X^{RGB}, X^{IMU}, Y)

Model

X^S = Source input (RGB data)
Z^T = Source feature
X^T = Target input (IMU data)
Z^T = Target feature
Y = Task Output (Action class)
f = RGB Feature Extractor (Resnet18, 1 3D conv, 1 fc)
h = HAR Model (2 fc)
g = IMU Feature Extractor (1 1D conv, 1 fc)



RESULTS

Our proposed model was trained on labeled RGB data and tested on RGB data, IMU data and Both. Testing on IMU data is zero shot transfer, since the system has not seen labeled IMU data during training.

HAR Accuracy on UTD-MHAD Dataset [5]

Model	RGB	IMU	Both
Early [7]	82 %	1.7 %	77 %
Middle [7] [8]	45 %	5.2 %	5.2 %
Late [7]	**94 %**	5.2 %	5.2 %
ImageBind [9]	40 %	3.4 %	33%
SF-CMT (Ours)	91 %	**66 %**	**84 %**

The SF-CMT model gives a 13x improvement for zero-shot cross-modality transfer to IMU and a slight improvement when using both modalities, compared to conventional sensor fusion methods.

CONCLUSIONS

- Sensor fusion methods assume a fixed set of inputs but leverage multiple modalities to perform well.
- Cross-modal transfer methods only use one modality at a time, but they can leverage data even when one modality is absent.
- Our Sensor Fusion and Cross-Modal Transfer (SF-CMT) model uses feature space representation learning to transfer knowledge across modalities, and sensor fusion to combine related features during inference for more robust performance depending on the availability of data during test time.

SOURCES

[1] Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. International Journal of Computer Vision, 130(5), 1366-1401.
[2] Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine, 34(6), 96-108.
[3] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.
[4] Wang, J., Zheng, V. W., Chen, Y., & Huang, M. (2018, July). Deep transfer learning for cross-domain activity recognition. In proceedings of the 3rd International Conference on Crowd Science and Engineering (pp. 1-8).
[5] Chen, C., Jafari, R., & Kehtarnavaz, N. (2015, September). 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) (pp. 168-172). IEEE.
[6] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.
[7] Wei, H., Jafari, R., & Kehtarnavaz, N. (2019). Fusion of video and inertial sensing for deep learning–based human action recognition. Sensors, 19(17), 3680.
[8] Islam, M. M., & Iqbal, T. (2020, October). Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 10285-10292). IEEE.
[9] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15180-15190).

ILLINOIS