Grenze International Journal of Engineering and Technology, Jan Issue

Robust Human Activity Recognition using Multimodal Feature-Level Fusion

Anil Kumar C J[1], Christo Abraham[2], Darshan M C[3], Freddy Dominic[4] and Anandakrishnan P S[5]
[1]Associate Professor, Department of CSE, ATME College of Engineering, Mysuru
[2-5]Department of CSE, ATME College of Engineering, Mysuru

Abstract— **Automated recognition of human activities or actions has great significance as it incorporates wide-ranging applications, including surveillance, robotics, and personal health monitoring. Over the past few years, many computer vision-based methods have been developed for recognizing human actions from RGB and depth camera videos. These methods include space-time trajectory, motion encoding, key poses extraction, space-time occupancy patterns, depth motion maps, and skeleton joints. However, these camera-based approaches are affected by background clutter and illumination changes and applicable to a limited field of view only. Wearable inertial sensors provide a viable solution to these challenges but are subjected to several limitations such as location and orientation sensitivity. Due to the complimentary trait of the data obtained from the camera and inertial sensors, the utilization of multiple sensing modalities for accurate recognition of human action is gradually increasing. This project presents a viable multimodal feature-level fusion approach for robust human action recognition, which utilizes data from multiple sensors, including RGB camera, depth sensor, and wearable inertial sensors. We extracted the computationally efficient features from the data obtained from RGB-D video camera and inertial body sensors. These features include densely extracted histogram of oriented gradient (HOG) features from RGB/depth videos and statistical signal attributes from wearable sensors data. The proposed human action recognition (HAR) framework is tested on a publicly available multimodal human action dataset UTD-MHAD consisting of 10 different human actions. Support Vector Machine and K-Nearest Neighbor classifiers are used for training and testing the proposed fusion model for HAR. The experimental results indicate that the proposed scheme achieves better recognition results as compared to the state of the art. The feature-level fusion of RGB and inertial sensors provide the overall best performance for the proposed system, with an accuracy rate of 97.6%.**

I. INTRODUCTION

Human action recognition (HAR) or activity recognition in an imperious area of research in signal and image processing. HAR mainly involves automatic detection, localization, recognition, and analysis of human actions from the data obtained from detection type of sensors, including RGB camera, depth sensor, range sensor, or internal sensor. Action detection involves determining the presence of the action of interest in a continuous data

Grenze ID:01.GIJET.9.1.618

from different types of sensors, including RGB camera, depth sensor, range sensor, or inertial sensor. Action detection involves determining the presence of the action of interest in a continuous data stream, whereas action localization estimates when and where an action of interest appears. The goal of action recognition or classification is to determine which action appears in the data. In the past few years, the research on HAR has gained significant popularity and is becoming increasingly vital in avariety of disciplines. Various sensor modalities have been utilized to monitor human beings and their activities. HAR approaches can generally be classified into two main categories depending upon the type of sensors used. These include vision-based HAR and inertial sensor-based HAR.

The advancement in image acquisition technology has made it possible to capture 3D action data using depth sensors. The depth images obtained for these sensors are insensitive to changes in illumination compared to conventional RGB images. Moreover, these depth images also provide a way to obtain 3D information of a person's skeleton to recognize human actions in a better way. Several feature extraction of a person's skeleton to recognize human actions in a better way to obtain 3D information of person's skeleton to recognize human actions in a better way. Several features extraction, description, and representation techniques have been developed for depth sensor-based HAR. These includes depth motion maps (DMMs), bag of 3D points, projection depth maps, space-time occupancy patterns, spatiotemporal depth cuboid, surface normal, and skeletal joints. Although vision-based HAR is continuously progressing, it is exposed to many hindrances such as camera position, a limited angle of view, subject disparities in carrying out different actions, occlusion, and background clutter. Furthermore, camera-based HAR systems require an extensive amount of hardware resources to run computationally complex computer vision algorithms. These limitations are addressed by low-cost, computationally efficient, and miniaturized inertial sensors.

Wearable inertial sensors enable dealing with a much broader field of view and changing illumination conditions as compared to RGB and depth sensors. They are attached directly on the human body or entrenched into outfits, smartphones, footwear, and wrist watches to track human activities. They generate 3D acceleration and rotation signals conforming to human action. Hence, like depth sensors, the inertial sensors also track 3D action data entailing 3-axis acceleration in case of an accelerometer and 3-axis angular velocity in case of a gyroscope. Many researchers utilized smartphones, smart watches, and wearable inertial sensors, incorporating an accelerometer and gyroscope, for human activity recognition. With the growth of deep learning applications in vision-based action recognition systems, we witnessed the utilization of deep learning for sensor-based activity recognition. However, with the continuous evolvement in pulling down the power consumption of wearable sensors, deep learning based approaches are becoming futile for unobtrusive human activity monitoring. Moreover, sensor-based activity recognition approaches have certain other limitations as well. For instance, sensor readings are sensitive to their orientation and location on the body. Also, wearing or placing these sensors on the bodies creates inconvenience for the users to carry out their tasks in a natural way. Table 1 provides the pros and cons regarding the use of different sensing modalities (i.e., RGB camera, depth camera, and inertial sensors) for HAR.

II. METHODOLOGY

A conventional HAR system typically makes use of a single sensor modality, i.e., either a vision-based sensing modality or a wearable inertial sensor. However, under realistic operational settings, no sensor modality alone can handle varying conditions that may take place in real time. The RGB and depth images from an RGB-D camera and 3D inertial signals from a wearable sensor offer complementary information. For instance, vision-based sensors provide global motion features whereas inertial signals give 3D information about local body movement. Hence, by fusing data from two complementary sensing modalities, the performance of HAR systems can be improved. Few existing studies utilized the fusion of depth and inertial sensors, aiming to increases the accuracy of action recognition and their results revealed significant improvement in recognition. Some authors also worked on using deep learning for multiple sensing modalities for robust action recognition. For depth camera, CNN based features are extracted, whereas, for the inertial sensors, CNN and LSTM networks are used. Therefore, in this project work, we use a multimodal HAR framework that utilizes the combination of multiple sensing modalities for action classification.

The fusion of multiple sensors can be performed at base level (descriptor-level), feature-level (representation-level), or decision-level (score-level). Each fusion type has its own merits and demerits, and the selection of the fusion method is generally dependent on the type of features and descriptors. Existing studies for multimodal HAR mostly focus on the decision-level due to its independence on the type, length, a numerical scale of different features and descriptors. Existing studies for multimodal HAR mostly focus on the decision- level

TABLE 1. PROS AND CONS OF DIFFERENT MODALITIES OF HAR

	RGB Cameras	Depth Sensors	Inertial Sensor
Pros	• Economical, easy-to-use and readily available • Offer color and texture information	• Inexpensive and widely available • Impermeable to variation in lighting and illumination settings • Insensitive to color and texture changes • Provide 3D action information	• Very cheap and easily available • Impermeable to change in lighting and illumination settings • Provide a high sampling rate • Provide 3D action data • Can work in an unconstrained environment
Cons	• Limited viewing angle • Sensitive to variation in lighting and illumination • Affected by camera adjustment • Require computationally expensive processing algorithms	• Can work only in a constrained field of view • Sensitive to different types of noise • Depth information is sensitive to objects varying reflection properties • Lack of color and rich texture information	• Position and orientation sensitive • Sensor drift • Insufficient onboard power • Use multiple sensors for recording full body motion • Wearing these sensors can create inconvenience for the users in performing their tasks

fusion due to its independence on the type, length, and numerical scale of different features extracted from multiple sensing modalities. Moreover, decision-level fusion does notrequire any post-processing of the extracted features and reduce the dimensions of the final feature vector for classification decision relating to each sensing modality, which are then combined using some soft rule to make the final decision. Hence, for n different sensing modalities, the decision-level fusion requires n classifiers to be trained and tested independently on each sensing modality. For any multimodal HAR system, the acquisition of concurrent data from multiple sources is necessary to collect a sufficient amount of information for making improved decision about human actions.However, with the decision-level fusion, it is not possible to combine multimodal data at an earlier stage to produce adequate information for recognizing human actions. In contrast, the feature-level fusion helps to collect concurrent features from multiple sensors and integrate them to generate sufficient information for making a strong decision. Moreover, it provides the best results in the case when the features extracted from different sensing modalities have the same dimensions and numerical scale. Therefore, in this study, we focused on the feature-level fusion of multiple sensing modalities for robust HAR. We extracted time domain features for inertial sensor data, whereas, to obtain the best results for feature-level fusion, we used densely extracted Histogram of Oriented Gradients (HOG) as features for both RGB and depth video data. The features extracted from multiple sensors are then fused and used to train the machine learning algorithm for action classification.

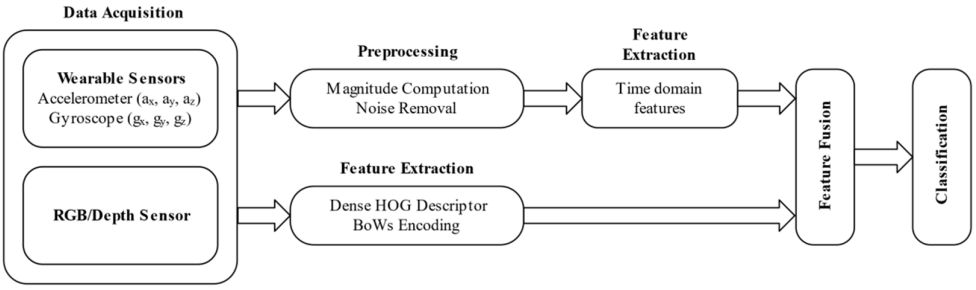

Figure 1. Block diagram of the proposed HAR method

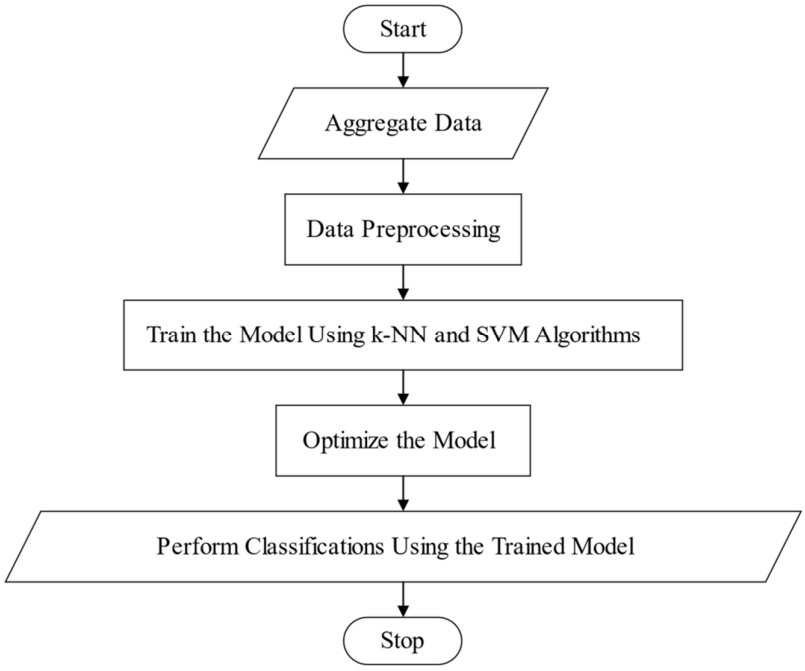

Figure 2. Flow Chart

III. CONCLUSION

In this paper, a feature-level fusion method has been proposed for human action recognition, which utilizes data from two differing sensing modalities: vision and inertial. The proposed system merges the features extracted from individual sensing modalities to recognize an action using a supervised machine learning approach. The detailed experimental results indicate the robustness of our proposed method regrading classifying human actions as compared to the setting where each sensor modality is used individually. Also, the feature-level fusion of time domain features computed from inertial sensors and densely extracted HOG features from depth/RGB videos reduces the computational complexity and improves the recognition accuracy of the system as compared to state-of-the-art deep CNN methods. Regarding classifier performance, K-NN classifier provides better results for the proposed HAR system as compared to SVM classifier. The proposed HAR methods also have some limitations. For example, its works with pre-segmented actions, which do not exist in practice. Moreover, it does not incorporate Multiview HAR, and the orientation of the person whose action is being recognized remains the same with respect to the camera. In the future, we plan to extend the proposed HAR method to address these

limitations. Furthermore, we aim to investigate the specific applications of the proposed fusion framework using RGB-D camera and wearable inertial sensors.

REFERENCES

[1] M. Ehatisham-Ul-Haq et al., "Robust Human Activity Recognition Using Multimodal Feature-Level Fusion," in IEEE Access, vol. 7, pp. 60736-60751, 2019, doi: 10.1109/ACCESS.2019.2913393.
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8701429&isnumber=8600701

[2] C. Chen, R. Jafari and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 2712-2716, doi: 10.1109/ICASSP.2016.7472170.
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7472170&isnumber=7471614

[3] C. Chen, R. Jafari and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 168-172, doi: 10.1109/ICIP.2015.7350781.
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7350781&isnumber=7350743