MS-KARD: A Benchmark for Multimodal Karate Action Recognition

Santosh Kumar Yadav
AcSIR, CSIR-CEERI, Pilani
santosh.yadav@pilani.bits-pilani.ac.in

Aditya Deshmukh
Birla Institute of Technology and Science
f20180246@pilani.bits-pilani.ac.in

Raghurama Varma Gonela
Birla Institute of Technology and Science
f20181120@pilani.bits-pilani.ac.in

Shreyas Bhat Kera
Birla Institute of Technology and Science
f20181119@pilani.bits-pilani.ac.in

Kamlesh Tiwari
Birla Institute of Technology and Science
kamlesh.tiwari@pilani.bits-pilani.ac.in

Hari Mohan Pandey
Bournemouth University
profharimohanpandey@gmail.com

Shaik Ali Akbar
AcSIR, CSIR-CEERI, Pilani
saakbar158@gmail.com

***Abstract*—Classifying complex human motion sequences is a major research challenge in the domain of human activity recognition. Currently, most popular datasets lack a specialized set of classes pertaining to similar action sequences (in terms of spatial trajectories). To recognize such complex action sequences with high inter-class similarity, such as those in karate, multiple streams are required. To fulfill this need, we propose MS-KARD, a Multi-Stream Karate Action Recognition Dataset that uses multiple vision perspectives, as well as sensor data - accelerometer and gyroscope. It includes 1518 video clips along with their corresponding sensor data. Each video was shot at 30fps and lasts around one minute, equating to a total of 2,814,930 frames and 5,623,734 sensor data samples. The dataset has been collected for 23 classes like Jodan Zuki, Oi Zuki, *etc*. The data acquisition setting involves the combination of 2 orthogonal web cameras and 3 wearable inertial sensors recording both vision and inertial data respectively. The aim of this dataset is to aid research that deals with recognizing human actions that have similar spatial trajectories. The paper describes statistics of the dataset, acquisition setting, and provides baseline performance figures using popular action recognizers. We propose an ensemble-based method, KarateNet, that performs decision-level fusion on the two input modalities (vision and sensor data) to classify actions. For the first stream, the RGB frames are extracted from the videos and passed into action recognition networks like Temporal Segment Network (TSN) and Temporal Shift Module (TSM). For the second stream, the sensor data is converted into a 2-D image and fed into a Convolutional Neural Network (CNN). The results reported were obtained on performing a fusion of the 2 streams. We also report results on ablations that use fusion with various input settings. The dataset and code will be made publicly available.**

***Index Terms*—Action recognition, Multimodal, Karate and martial arts, Sports and exercises, Deep learning, Vision and wearable**

I. Introduction

Human activity recognition (HAR) is a classification task in which, based on the sensory input, the machine understands and infers different activities performed by a subject. Out of the many sports and fitness activities, martial art serves as an excellent tool to promote physical and mental health. Karate, which has Japanese roots, is one of the most popular ancient martial arts. Practicing karate promotes improvement in mobility control [1] and greater intensity of health behaviors among individuals [2]. Karate is amongst the top 10 martial arts performed and has over 100 million practitioners around the globe [3]. It uses hand attacks more, and kicks are mainly used as backup so legs mostly stay grounded. Karate thus fits in the description wherein its move sequences have very high inter-class similarity and the data can be easily collected.

Building a dataset for karate has several challenges. Firstly, wearing an HMD [4], [5] or wearing many sensors on different parts of the body [6] is not very convenient to the practitioner in a realistic scenario, and hence a system with minimum obtrusiveness must be proposed. Another challenge includes gathering an adequate amount of quality data which has a major impact on the model's generalization performance. A good quality training dataset requires professional martial artists with several years of experience, to perform karate moves skillfully [7]. The multiple modalities recording data must be in complete synchronization in order to provide consistent information to the training models. Also, the fast and complex body movements of the karateka (practitioner of karate) make tracking and classification of karate moves a challenging task [8]. There is temporal correlation with the movement of an action. To capture this, we use sensor data from wearable devices.

We present MS-KARD, a Multi-Stream Karate Action Recognition Dataset that includes visual and sensor data for 23 karate moves. The dataset uses two web cameras and three inertial measurement units to record the motion data of the karateka and infer the performed karate move. We also propose KarateNet, a two-stream action recognition network, that uses a composite deep learning architecture, where the first sub-architecture processes visual cues by learning features from the video sequence using CNN-based popular action recognition networks to produce visual scores, and the second sub-architecture learns the features of inertial cues using 2D-

CNNs to produce inertial scores. Finally, the model performs decision level fusion of visual and inertial scores to produce final class scores. The major contributions of this paper are threefold:

- We propose a novel dataset, MS-KARD consisting of multi-stream data for 23 karate action with 2,814,930 frames and 5,623,734 sensor data samples for karate action recognition. To the best of our knowledge, it is the first of its kind where data has been recorded with 2 orthogonal RGB cameras and 3 wearable inertial sensors.
- We propose KarateNet, which uses deep learning architectures (like TSN [9], TSM [10], INM) trained on the vision and sensor streams of MS-KARD to classify the karate actions.
- We utilize various mechanisms at the data and decision levels to fuse the models in an attempt to improve results. We provide baselines and other ablation results using multiple fusion settings.

II. Related Works

This section provides details of related data acquisition methods as well as information about the relevant action recognition models.

A. Data Acquisition Techniques

Many researchers have used different data acquisition techniques to come up with smart systems for analyzing karate. Vision-based techniques are being focused in [4], [5], [7], [8], [18], [19]. A virtual reality training system for karate is presented in [4] and [5]. Wu et al. [4] proposed a training system based on 3D forecasting, using an RGB camera. Similarly, Petri et al. [5] used MoCap recordings of professional karate masters with a motion capture system, Vicon tracker with 12 cameras to animate virtual opponents. The practitioner wears an HMD to fight against the virtual opponent.

Bianco et al. [18] used a Kinect sensor to obtain the 3D image frames for 10 karate moves (5 blocking, 2 punching, and 3 kicking). Sotirios et al. [8] used an RGB camera to obtain 2D image frames for 5 basic kata sequences. Hachaj et al. [19] emphasized automatically recognizing karate sequences, using a combination of three Kinect sensors for data acquisition. The skeleton representation is performed by fusing the body joints obtained from each Kinect sensor and karate pose classification was performed using a gesture description language (GDL) script. The data was recorded for seven karate techniques (4 stances, 2 blocks, 1 kick). Further, the same authors, [7] proposed a Kinect-V2-based dataset containing 10 karate techniques (3 stances, 3 kicks, and 4 blocks).

Alternatively, wearable-based data acquisition techniques for karate analysis are being discussed in [20], and [6]. Pindari et al. [20] proposed a 'lexical-like' approach for movement classification, using five wearable inertial sensors from Xsens, where each inertial sensor is having an accelerometer, gyroscope, and magnetometer. The method was tested on the WARD and NIDA databases, which contains activities of daily living and three karate actions, namely, 'karate punch', 'karate front kick' and 'karate side kick'. Hachaj et al. [6] used seventeen wearable inertial-sensors by Shadow 2.0 wireless motion capture system to obtain data for Oyama and Shorin-Ryu karate techniques.

A hybrid data acquisition technique is used in [21]. They proposed an interactive learning system for karate that promotes game-based learning by inducing two-way interaction between the player and the computer. They used a combination of two sensor modalities, $i.e.$, a wireless wearable accelerometer and a Kinect sensor.

B. Related Datasets

In the domain of video action recognition, many existing popular datasets encompass a large, diverse set of activity classes, which are generally coarse-grained. This is evident when noting examples of classes in the widely researched HMDB51 [22] or UCF101 [23] datasets, which include generic class labels such as *kick*, *punch*, *hit*, *basketball*, $etc.$ Moving to more recent popular datasets such as Kinetics [24], the number of classes drastically increases (up to 700 classes), however, the class labels, such as *abseiling*, *exercising arm*, *swinging legs*, *wrestling*, *parkour*, $etc.$ are still generic in nature. Datasets like Something-Something [25], while fine-grained in nature, do not focus on Human activities. HAA500 [26], while being both human-centric and fine-grained, contains a broad list of classes in various domains. Our proposed dataset differs from these, in the sense that the classes are derived from a **specific domain**, $i.e.$, Karate. Such a specialized, fine-grained class set with high speeds of movement makes it a challenging task to distinguish between the actions performed. Further, the actions are **human-centric**. A necessity when considering a specialized domain such as Karate is to have atomic actions instead of generic class labels. An exemplar of this is the provision of disparate class labels for 'Zuki' actions (punching), such as Oi Zuki (lunge punch) which differs fundamentally from Heiko Zuki (parallel punch), thereby making the classes **atomic**.

In comparison to other present Karate datasets, MS-KARD uses a new system of information capture, $i.e.$, vision and sensor streams, as well as a quantitatively larger information pool in the form of 2,814,930 video frames and 5,623,734 sensor data samples. Table I lists the datasets which are broadly related to Karate, along with their characteristics, such as input modalities, number of subjects, number of frames, $etc.$ Although the datasets cannot be directly compared due to the differences in factors like input streams, these metrics demonstrate that the proposed dataset contains abundant data with a sufficient number of subjects (13), classes (23), samples (1518), and RGB frames (2,814,930) all collected from the relatively narrow domain of Karate. Further, 2 camera views were chosen to garner more visual information to be able to distinguish between classes by fusion schemes. Additionally, in practical scenarios, any trained model can be selected for use, either front view, side view, or both. As seen by the results of TSN and TSM, having 2 camera views and arriving at a decision level consensus, a higher accuracy can be obtained.

TABLE I
COMPREHENSIVE LIST OF KARATE-RELATED DATASETS AND THEIR CHARACTERISTICS.

Dataset	Modalities	# of Subjects	# of Karate Sequences	# of Samples	# of Frames
MS-KARD (ours), 2021	RGB, Sensors	13	23	1,518	2,8145,930
TUHAD [11], 2020	RGB, Depth, IR	10	8	1,936	99,982
iKarate [12], 2020	Skeleton	2	7	210	-
Blaszczyszyn et al. [13], 2019	Gait	26	1	-	-
Karate Kicks [14], 2018	MoCap	4	4	320	-
MADS [15], 2017	RGB, Depth	2	6	216	53,000
Hachaj et al. [16], 2017	MoCap	2	28	560	-
Hachaj et al. [7], 2015	MoCap	6	10	1236	-
Hachaj et al. [17], 2015	Skeleton	1	7	350	-
Bianco et al. [18], 2013	Skeleton	-	10	-	-

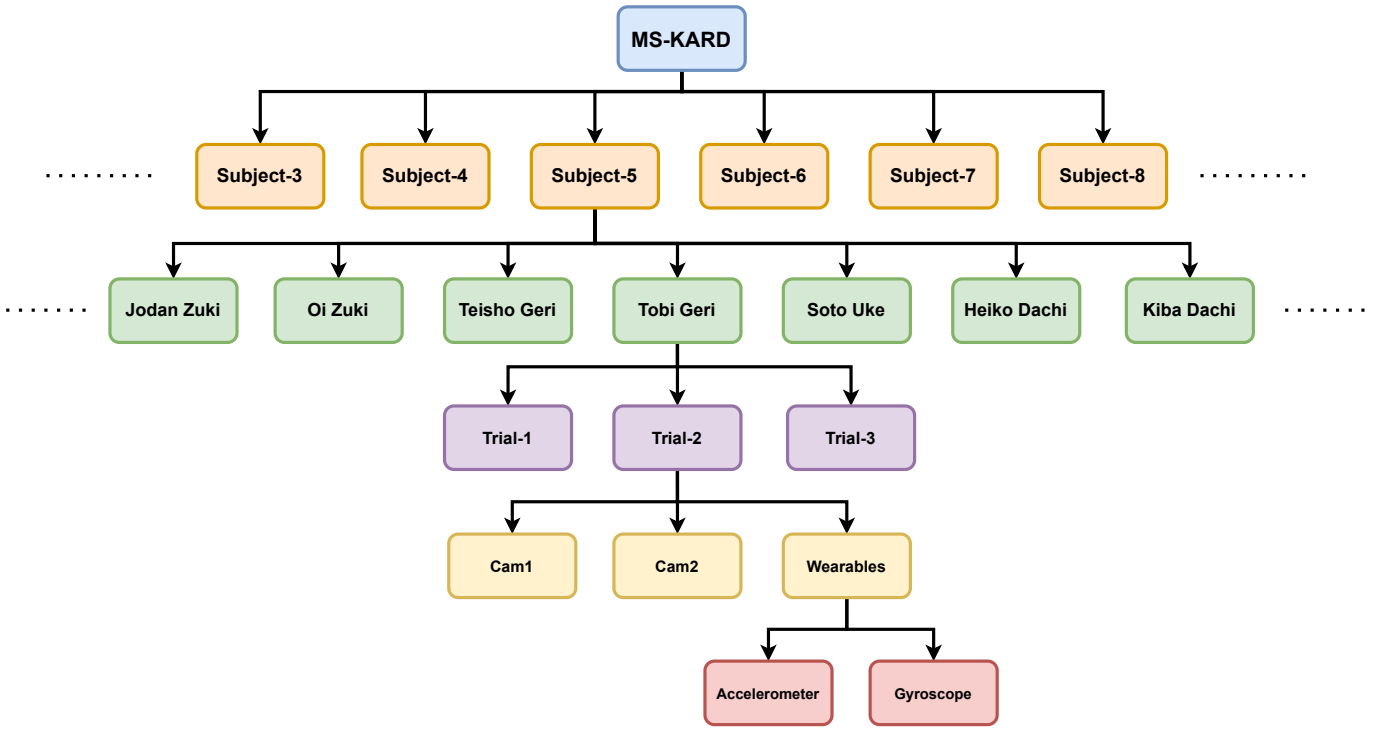

Fig. 1. Dataset Hierarchy of the proposed MS-KARD dataset.

III. DATASET BUILDING

This section presents the setup used to collect the dataset, followed by the technical specifications and dataset details.

A. Experimental Setup

The MS-KARD dataset is collected using a novel combination of two RGB web cameras and three wearable inertial sensors. The two cameras are placed orthogonal to each other, to capture the front and side views of the performer. Each wearable inertial sensor consists of a 3-axis accelerometer, gyroscope, and magnetometer. The dataset has been collected in a closed laboratory setup where the environment remains the same and the cameras were fixed at the same position throughout the data collection process.

B. Technical Specifications

Two Logitech cameras were used to record the videos. The data from both the cameras were recorded at a resolution of 1080p and a frame rate of 30 frames per second (fps). Both the cameras were fixed at a height of 4 feet from the ground and the performer always remains within the frame of both the cameras while performing the karate sequences, as presented in the left of Fig. 3. The practitioner wore the three wearable MbientLab Meta-Sensors on the left wrist, right wrist, and right leg. The data from the 3-axis accelerometer, gyroscope, and magnetometer are recorded at 100Hz, 100Hz, and 25Hz, respectively. The data from the wearable sensors are stored in CSV files containing the timestamp and the raw sensor values.

IV. DATASET HIERARCHY

The MS-KARD dataset was collected from 13 karateka performing various karate sequences using two orthogonally placed cameras and three sensors worn by the subject. The dataset is divided into the test, train and validation splits by subject. Subjects - 6,7,9 make up the test set, Subject - 13 makes up the validation set and the rest (Subjects - 1,2,3,4,5,8,10,11,12) make up the train set.

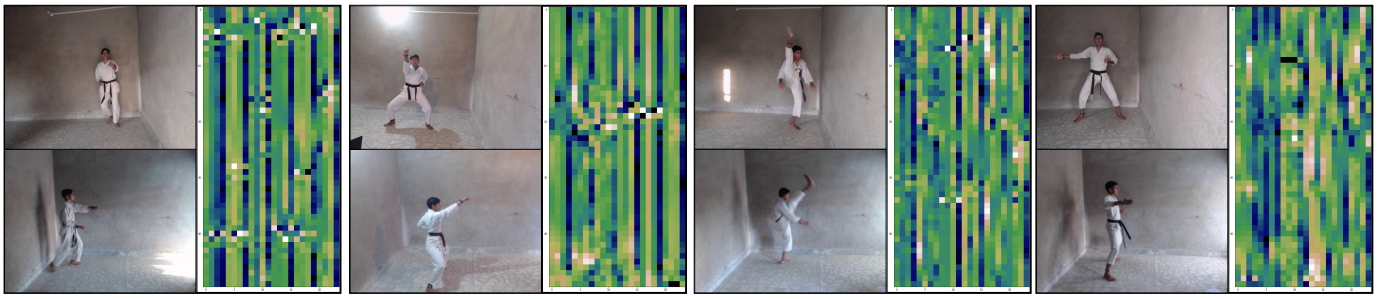

Fig. 2. Snapshots from MS-Kard displaying sample RGB (Front, Side) and Sensor Data.

TABLE II
MS-KARD CLASSES AND DESCRIPTIONS

Class	Description	Category
Jodan Zuki	Upper-Level Punch	Hand
Heiko Zuki	Parallel Punch	Hand
Oi Zuki	Lunge Punch	Hand
Shuto Uchi	Knife Hand Strike	Hand
Teisho Uchi	Palm Heel Strike	Hand
Ura-ken Uchi	Reverse Fist Strike	Hand
Mawashi Empi	Elbow Strike	Hand
Yoko Geri	Side Kick	Kick
Tobi Geri	Jumping Front Kick	Kick
Ushiro Mawashi Geri	Spinning Back Kick	Kick
Yoko Tobi Geri	Jumping Side Kick	Kick
Mae Geri (Kokomi)	Front Kick	Kick
Hiza Geri	Knee Strike	Kick
Gedan Barai Uke	Downward Block	Block
Mawashi Uke	Circular Block	Block
Soto Uke	Outside Block	Block
Ageuke	Upward Block	Block
Heiko Dachi	Natural Stance	Stance
Heisuko Dachi	Feet-Together Stance	Stance
Musubi Dachi	Attention Stance	Stance
Kiba Dachi	Horse Stance	Stance
Zenkutsu Dachi	Front Stance	Stance
Kosa Dachi	Cross Stance	Stance

The dataset hierarchy as shown in Fig. 1 can be summarized as follows:

- Each subject in the training and validation sets contains 23 action classes.
- Each action class contains 3 trials of a named karate sequence.
- Unlike the training and validation sets, the test set only comprises one trial per action class for each of the three subjects.
- Each trial, in any case, includes one front view video (cam_f), one side view video (cam_s), and inertial sensor readings (accelerometer and gyroscope).
- The names, descriptions, and categories of each class can be found in Table II.

A. Dataset Details

The MS-KARD dataset is collected with the help of 13 Goju-Ryu practitioners (subjects). The 23 karate techniques are composed of 6 kicking techniques (Yoko Geri, Tobi Geri, Ushiro mawashi Geri, Yoko Tobi Geri, Mae Geri (Kokomi), Hiza Geri), 6 basic stances (Heiko Dachi, Heisoku Dachi, Musubi Dachi, Kiba Dachi, Zenkutsu Dachi, Kosa Dachi), 7 hand techniques (Jodan Zuki, Heiko Zuki, Oi Zuki, Shuto Uchi, Teisho Uchi, Ura-ken Uchi, Mawashi Empi) and 4 blocking techniques (Gedan Barai uke, Mawashi uke, Soto uke, Age uke). The age of the participants varies from 15 to 25 years. All the classes were decided based on the extensive literature study and with the discussion of karate coaches. The data is recorded under the guidance of a karate coach (black belt). Every action is performed up to 3 times (trials) by a subject, for a duration of about one minute each. The final dataset contains 1564 minutes of video data from both RGB cameras, equating to a total of 2,814,930 frames and 5,623,734 sensor data samples. The split ratio we have used is 9:1:1 for train, test, and validation respectively. Subjects 6,7,9 belong to the test set, subject 13 belongs to the validation set, while the rest belong to the training set. Further details about the dataset hierarchy are provided in the Appendix.

V. Proposed Approach

Our main objective is to identify the user's action from the video and the sensor data collected. No sensing modality is perfect; no modality can completely describe the entire information about an activity. In this paper, we propose the KarateNet model which is constituted of two sub-architectures, each of which handles one of the two input streams, $i.e.$ RGB and Sensors. The model employs various methodologies of intra-stream and inter-stream fusion of its sub-architectures to reach a final classification.

A. Vision Stream

This stream makes use of popular action recognition models in order to handle the RGB video information. The models we have used include TSN [9] and TSM [10], which are both ResNet-based architectures. Input videos are first decoded, then a number of frames (T) are extracted from each video to be resized, normalized, and reshaped, thereby giving a vector of dimensions $T \times 3 \times H \times W$ (initial number of channels for RGB is 3, height is H and width is W). Following the convolution and pooling operations of the action recognizer, an average consensus produces a $1 \times C$ vector of confidence scores for each of the C classes. The datasets used for training can have videos of the front view (cam_f) or videos of the side view (cam_s).

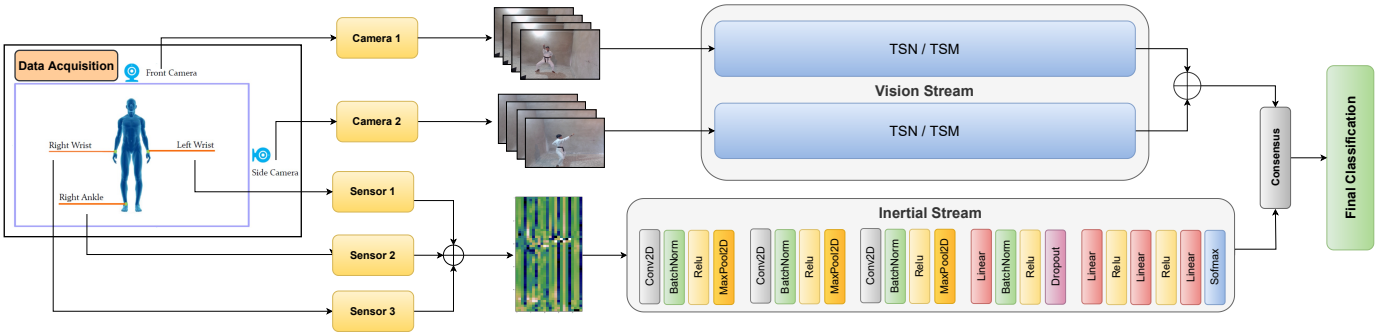

Fig. 3. The proposed KarateNet architecture, which uses two sub-architectures the work on the vision and inertial streams. Note: In the figure, data level decoupling is used as the fusion method for the vision stream.

B. Sensors Stream

The inertial sensor data, comprised of accelerometer (Acc) and gyroscope (Gyr) readings, is a time-series signal. Therefore an approach similar to that in speech recognition is used. A 12 column wide input image is formed by combining the 3-axis acceleration and/or angular velocity to get the overall acceleration and/or overall angular velocity for each of the 3 sensors at half the sampling frequency, thus a 50×12 image is obtained. The data is also normalized. While preparing data, a sliding window approach as a form of data augmentation is used. We create a sub-architecture called InertialNet to handle this sensor data.

InertialNet Model (INM): For the sensor data, a 2-D CNN is used. The input image is passed serially into 3 2D convolutional layers with 16, 32, and 64 filters. The output of each of these convolutional layers is batch-normalized and max-pooled with a 2×2 filter. The max-pooled output from the third layer is then flattened and passed through a series of fully connected layers with 256, 512, 128, and 23 units. After the first fully connected layer a 50% dropout is applied. ReLu activation is applied after the fully connected layers with 256, 512, and 128 units. The last fully connected layer is based on the softmax activation. The data-level fusion and the independent sensor models differ only in the first layer wherein, they accept different input dimensions.

C. Fusion Scheme

Fusion between constituent models in KarateNet is twofold: the fusion between models working on the same input modality and fusion between models working on different input modalities, $i.e.$ intra-stream fusion and inter-stream fusion respectively. Note that we denote base models with ψ, where $\psi \in \{TSN, TSM, INM\}$ and input modalities with ϕ, where $\phi \in \{cam_f, cam_s, Acc, Gyr\}$. Here, a singular model trained on a single modality would be represented by ψ_ϕ.

Intra-Stream Fusion: There are primarily two strategies to fuse predictions within a given stream based on how the input data is to be handled by the model, which are described as follows:

Data Level Fusion (DLF): To make models robust to recognizing highly similar inter-class actions (in terms of spatial trajectories and environment), we propose data-level fusion so as to generate a more challenging dataset for classifying such activities. Fusion of information at the data level is done between the two camera perspectives for the vision stream and between the accelerometer and gyroscope data for the sensor stream. DLF for the vision stream is done by jointly taking all the captured videos as a single dataset and training the models. Thus, the dataset that the action recognizer is trained on is a large collection of videos of differing perspectives, i.e. Cam_1 and Cam_2. To implement DLF for the sensor stream, the two independent images of the accelerometer and gyroscope signals are joined together to form a 50×24 image. Thus, multiple cues are presented simultaneously to the recognizer. In either case, only a single model needs to be trained for the respective stream, however, the diversity of the input datasets makes it challenging to train on. We denote a data level fused intra-stream model as $\psi_{DLF(\phi_1,\phi_2)}$.

Data Level Decoupling with Late Fusion (DLD): Rather than a single-step early fusion of the data, this second approach involves two steps. For the vision stream, the first step is to segregate the two camera inputs. This is done by initially treating the two views as distinct datasets - one with only cam_f video sequences, and the other with only cam_s video sequences. Thus for each model type, two different trained models are obtained. Each model, when tested on videos of its respective view, will give output probabilities pertaining to the confidence score of each class. Following this, the second step is to fuse the results using a consensus method between the scores. Similarly, for late fusion in the sensor stream, we treat the two sensor data inputs (Acc and Gyr) as two separate cues, and hence train 2 independent models. Let the two input models for the late fusion be denoted by ψ_{ϕ_1} and ψ_{ϕ_2}. For a given instance, let the predicted class probabilities be P and Q, where

$$P = \{(p_1, ..., p_C) \in \mathbb{N}^C : 0 \leq p_i \leq 1\} \quad (1)$$

$$Q = \{(q_1, ..., q_C) \in \mathbb{N}^C : 0 \leq q_i \leq 1\} \quad (2)$$

In this work we consider two methods for arriving at a consensus between the class scores, as follows:

Additive Consensus: This method is done by adding the

respective class scores between the two models, having an equivalent effect as average consensus, as the predicted class is that with the highest new class score. Thus, the new class score $N = \{n_1, ..., n_C\}$ is obtained where

$$n_i = p_i + q_i \; ; \quad 1 \leq i \leq C \tag{3}$$

Thus intra-stream data level decoupling with late additive consensus is denoted by
$DLD^{+}(\psi_{\phi_1}, \psi_{\phi_2})$.

Multiplicative Consensus: Consensus is achieved in a similar fashion to additive consensus, by using the product of the respective class scores instead. The new class score $N = \{n_1, ..., n_C\}$ is obtained where

$$n_i = p_i * q_i \; ; \quad 1 \leq i \leq C \tag{4}$$

Thus intra-stream data level decoupling with late multiplicative consensus is denoted by DLD$^{*}(\psi_{\phi_1}, \psi_{\phi_2})$.

Inter-Stream Fusion: Fusion between the two streams can take place in a similar manner as the intra-stream data level decoupling consensus methods. In this case, models that have undergone intra-stream fusion can be taken as inputs to this fusion mechanism. Therefore more than two models can simultaneously be plugged into the algorithm, which can be denoted by $ISF^{+}(\psi_{\phi_1}, ..., \psi_{\phi_n})$, $ISF^{*}(\psi_{\phi_1}, ..., \psi_{\phi_n})$ for additive and multiplicative consensus respectively. The computation can be done using Equations 3 (for additive consensus) or 4 (for multiplicative consensus), which can readily be extended to support n models $P_1, ..., P_n$.

VI. Experimental Results

We provide experimental results on the MS-KARD dataset using the proposed KarateNet method with its various settings for fusion schemes. The specification for model training and fusion schemes can be found in the Appendix.

A. Experimental Settings

Model Training: All the models are trained in an offline setting from scratch and independent of each other.

Temporal Shift Module (TSM) and Temporal Segment Network (TSN): We obtained the base models for TSN [9] and TSM [10] from the MMAction2 repository [27]. The ResNet architectures used for both the TSN and TSM derived backbones had a depth of 50. The input spatial resolution provided to the backbone is 224 × 224. For both TSN and TSM, the value of the temporal resolution, $i.e.$, frames sampled was kept at 8. The optimizer chosen was SGD and the dropout ratios were set to 0.4 for TSN and 0.5 for TSM. The loss used was Cross-Entropy Loss. The learning rate was set to 0.000625.

InertialNet Model (INM): For the sensors stream, the SGDM optimization algorithm is used to train the model, and one cycle learning rate scheduling policy is used. The momentum value is set to 0.9 and the maximum learning rate is set to 0.2, 0.25, and 0.15 for the accelerometer, gyroscope, and the data level fusion models. Weight decay and gradient clipping are also used. The loss used was Cross-Entropy Loss.

Fusion: During inter-stream fusion, the class scores received from the sensor stream varied by a large margin in terms of the standard deviation of the values, compared to that of the vision stream. In order to reduce any untoward effect/inordinate bias towards the sensor stream, the class scores were softmaxed, which brought them to a more comparable range, while still maintaining the previous relative order of confidence scores.

B. Baseline Results

Several observations can be made using the evaluations of individual base models of TSN, TSM, and INM, when independently trained on a singular input modality, $i.e.$ cam_f, cam_s, Acc, Gyr. The different model configurations, as well as their results, can be seen in the first two rows of Tables III, IV, V. For the vision stream, we note that TSM outperforms TSN on either input view. This is expected as TSM involves a shifting mechanism to facilitate information exchanged among neighboring frames and thus improves temporal understanding. This demonstrates the temporal sensitivity of the fine-grained actions performed in the MS-KARD dataset. Further, for TSN there is an evident disparity between the performance on front view videos in comparison to side view videos for Top-1 accuracy, although the Top-3 accuracies are comparable. This may be because the correct class predictions are predicted with a confidence slightly lower than the requisite amount. The intra-stream and inter-stream fusions would mediate these confidence scores and boost accuracy. Results for the sensor stream indicate that the InertialNet Model is able to achieve 5.88% higher accuracy when trained on Acc than Gyr. These results serve as baselines in order to evaluate the effectiveness of the different fusion schemes.

C. Data Level Fusion

This first method of intra-stream fusion is evaluated using independently trained models on the fused datasets of cam_f and cam_s for the vision stream models (TSN and TSM) and Acc and Gyr for the sensor stream model (INM). The results of data-level fusion can be seen in the third row of Tables III, IV, V. Notably, each DLF scheme requires only one model to train, however, the accuracies of all these models are unable to provide greater accuracies than both of the corresponding data decoupled models. For example, TSM$_{DLF(cam_f,cam_s)}$ is 11.59% and 10.14% lower than TSM$_{cam_f}$ and TSM$_{cam_s}$ respectively and TSN$_{DLF(cam_f,cam_s)}$ is 2.9% and 21.74% lower than TSN$_{cam_f}$ and TSN$_{cam_s}$ respectively. This drop in performance is likely due to the model's inability to generalize over the front and side view videos of classes in the fused dataset.

D. Data Level Decoupling

Data level decoupling with late fusion is implemented using a consensus methodology on two models trained on differing input modalities of the same stream, for example,
DLD^{+}(TSN$_{cam_f}$, TSN$_{cam_s}$). Other settings, as well as each of their results, are shown in the last two rows of Tables III,

TABLE III
RESULTS USING TSN BASED METHODS

Model	Top-1	Top-3
TSN_{cam_f}	44.93%	81.16%
TSN_{cam_s}	63.77%	86.96%
$\text{TSN}_{DLF(cam_f,cam_s)}$	42.03%	73.12%
$\text{DLD}^{+}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s})$	66.67%	84.06%
$\text{DLD}^{*}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s})$	63.77%	85.51%

TABLE IV
RESULTS USING TSM BASED METHODS

Model	Top-1	Top-3
TSM_{cam_f}	73.91%	88.41%
TSM_{cam_s}	72.46%	92.75%
$\text{TSM}_{DLF(cam_f,cam_s)}$	62.32%	82.61%
$\text{DLD}^{+}(\text{TSM}_{cam_f}, \text{TSM}_{cam_s})$	75.36%	92.75%
$\text{DLD}^{*}(\text{TSM}_{cam_f}, \text{TSM}_{cam_s})$	75.36%	92.75%

TABLE V
RESULTS USING INM BASED METHODS

Model	Top-1	Top-3
INM_{Acc}	50.00%	73.53%
INM_{Gyr}	44.12%	61.76%
$\text{INM}_{DLF(Acc,Gyr)}$	45.59%	73.53%
$\text{DLD}^{+}(\text{INM}_{Acc},\text{INM}_{Gyr})$	57.35%	76.47%
$\text{DLD}^{*}(\text{INM}_{Acc},\text{INM}_{Gyr})$	51.47%	75.00%

TABLE VI
RESULTS USING INTER-STREAM FUSION

Model	Top-1	Top-3
$\text{ISF}^{+}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s}, \text{INM}_{Acc}, \text{INM}_{Gyr})$	66.18%	86.76%
$\text{ISF}^{*}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s}, \text{INM}_{Acc}, \text{INM}_{Gyr})$	73.53%	88.24%
$\text{ISF}^{+}(\text{TSM}_{cam_f}, \text{TSM}_{cam_s}, \text{INM}_{Acc}, \text{INM}_{Gyr})$	76.47%	94.12%
$\text{ISF}^{*}(\text{TSM}_{cam_f}, \text{TSM}_{cam_s}, \text{INM}_{Acc}, \text{INM}_{Gyr})$	76.47%	94.12%

IV, V. The primary observation is that this method of fusion produces results that are either equivalent or superior to both the individually decoupled constituent models. This is apparent in all cases; for example $\text{DLD}^{+}(\text{INM}_{Acc}, \text{INM}_{Gyr})$ produces an accuracy of 7.35% and 13.23% higher than the constituent models INM_{Acc} and INM_{Gyr}. The improvement is seen for both the additive and multiplicative consensus methods.

Further, the accuracies far surpass those of the DLF method for all streams; an instance of this point can be seen by $\text{DLD}^{+}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s})$, whose accuracy exceeds that of $\text{TSN}_{DLF(cam_f,cam_s)}$ by a large margin of 24.64%. This shows that late consensus of decoupled models is a better approach to fusion rather than fusing the input modalities at the data level.

Another pertinent result is that for intra-stream late fusion, the consensus method that generally procures greater accuracies on the MS-KARD dataset is based on the additive scheme rather than multiplicative. Results for TSM are equal, but for TSN and INM, there is a loss of 2.9% and 5.88% respectively when using multiplicative over additive consensus. This may be attributed to the fact that within the stream, the models produce class probabilities of similar scales. Thus, the addition of class probabilities between these comparable scales provides an apposite estimate of an average agreement between the two models. On the other hand, the multiplicative consensus strategies may fail in cases where predictions with high confidence scores are reduced to sub-optimal values because of 'disagreement' between the two models, and thereby lose relevance while making the final prediction; this would not happen in additive consensus as the values would necessarily increase.

E. Inter-Stream Fusion

Having arrived at a fusion within the stream level, these models can be further fused through inter-stream strategies as described earlier. The two overall settings use either one of DLD fused TSN or DLD fused TSM from the vision stream in conjunction with the DLD fused INM from the sensor stream. Further, each setting can be implemented using additive or multiplicative consensus. The results of the evaluations using this fusion scheme can be found in Table VI . Clearly, inter-stream fusion provides an improvement over any of its constituent models; an example of this can be seen by $\text{ISF}^{+}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s}, \text{INM}_{Acc}, \text{INM}_{Gyr})$ being 21.25%, 2.42%, 16.18%, and 22.06% higher than TSN_{cam_f}, TSN_{cam_s}, INM_{Acc} and INM_{Gyr}, respectively. However, in comparison with DLD models, it is evident that improving results using the additive consensus version of inter-stream fusion is challenging - it improved accuracy by 1.11% for the TSM based model, while there was a drop of 0.49% for the TSN based model. On the other hand, the multiplicative consensus fared better for the TSN-based model, with an improvement of 9.76% from its vision stream DLD counterpart. The rise in accuracy using this fusion scheme may be imputed to the fact that the class probabilities from the differing streams (vision and sensor) may have varied scales, which is expected since they arise from inherently different base models ($eg.$ - TSN vs INM). In such a situation, the product of the scores gives a fairer balance as the output scores are largely independent of the input scales, which is not the case for additive consensus. This supports the findings of [28].

The greatest accuracy obtained is 76.47% Top-1 and 94.12% Top-3 through inter-stream fusion between the vision stream using TSM and the sensor stream using INM. A complete table with all the acquired accuracies of various models has been presented in Table VII for ease of reference.

VII. CONCLUSION

Most popular datasets lack a specialized set of classes with similar spatial trajectories. To recognize such complex action sequences with high inter-class similarity, like those in karate, datasets, and models that use multiple streams are required. In this paper, we presented MS-KARD, a Multi-Stream Karate Activity Recognition Dataset as well as a fusion-based action recognition network, KarateNet, to address this need. KarateNet uses a combination of inertial cues from 3 inertial measurement units (IMUs) and visual cues from 2 web cameras. The network uses a deep learning architecture,

TABLE VII
COMPLETE LIST OF ABLATION STUDY RESULTS OBTAINED USING VARIOUS MODELS ON OUR PROPOSED MS-KARD DATASET.

Model	Top-1	Top-3
TSM_{cam_f}	73.91%	88.41%
TSM_{cam_s}	72.46%	92.75%
$\text{TSM}_{DLF(cam_f,cam_s)}$	62.32%	82.61%
$\text{DLD}^{+}(\text{TSM}_{cam_f}, \text{TSM}_{cam_s})$	75.36%	92.75%
$\text{DLD}^{*}(\text{TSM}_{cam_f}, \text{TSM}_{cam_s})$	75.36%	92.75%
TSN_{cam_f}	44.93%	81.16%
TSN_{cam_s}	63.77%	86.96%
$\text{TSN}_{DLF(cam_f,cam_s)}$	42.03%	73.12%
$\text{DLD}^{+}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s})$	66.67%	84.06%
$\text{DLD}^{*}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s})$	63.77%	85.51%
INM_{Acc}	50.00%	73.53%
INM_{Gyr}	44.12%	61.76%
$\text{INM}_{DLF(Acc,Gyr)}$	45.59%	73.53%
$\text{DLD}^{+}(\text{INM}_{Acc},\text{INM}_{Gyr})$	57.35%	76.47%
$\text{DLD}^{*}(\text{INM}_{Acc},\text{INM}_{Gyr})$	51.47%	75.00%
$\text{ISF}^{+}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s}, \text{INM}_{Acc}, \text{INM}_{Gyr})$	66.18%	86.76%
$\text{ISF}^{*}(\text{TSN}_{cam_f}, \text{TSN}_{cam_s}, \text{INM}_{Acc}, \text{INM}_{Gyr})$	73.53%	88.24%
$\text{ISF}^{+}(\text{TSM}_{cam_f}, \text{TSM}_{cam_s}, \text{INM}_{Acc}, \text{INM}_{Gyr})$	**76.47%**	**94.12%**
$\text{ISF}^{*}(\text{TSM}_{cam_f}, \text{TSM}_{cam_s}, \text{INM}_{Acc}, \text{INM}_{Gyr})$	**76.47%**	**94.12%**

trained on the MS-KARD dataset composed of 23 karate moves for final predictions/inferences. KarateNet consists of 2 sub-architectures. The first computes the score of visual cues by using popular action recognizers. The second computes the scores of inertial cues by first learning features by mapping inertial signals to an image, and then passing them to a 2D CNN. We also devised a consensus scheme describing data-level fusion, data-level decoupling with late fusion and discussed which fusion scheme is suitable for inter-stream and intra-stream fusion.

REFERENCES

[1] M. Błaszczyszyn, A. Szczesna, M. Pawlyta, M. Marszałek, and D. Karczmit, "Kinematic analysis of mae-geri kicks in beginner and advanced kyokushin karate athletes," *International journal of environmental research and public health*, vol. 16, no. 17, p. 3155, 2019.

[2] K. Kotarska, L. Nowak, M. Szark-Eckardt, and M. A. Nowak, "Intensity of health behaviors in people who practice combat sports and martial arts," *International journal of environmental research and public health*, vol. 16, no. 14, p. 2463, 2019.

[3] W. K. Federation, "Wkf claims 100 million karate practitioners," 2002.

[4] E. Wu and H. Koike, "Futurepose-mixed reality martial arts training using real-time 3d human pose forecasting with a rgb camera," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1384–1392.

[5] K. Petri, P. Emmermacher, M. Danneberg, S. Masik, F. Eckardt, S. Weichelt, N. Bandow, and K. Witte, "Training using virtual reality improves response behavior in karate kumite," *Sports Engineering*, vol. 22, no. 1, p. 2, 2019.

[6] T. Hachaj, M. Piekarczyk, and M. R. Ogiela, "Human actions analysis: templates generation, matching and visualization applied to motion capture of highly-skilled karate athletes," *Sensors*, vol. 17, no. 11, p. 2590, 2017.

[7] T. Hachaj, M. R. Ogiela, and K. Koptyra, "Application of assistive computer vision methods to oyama karate techniques recognition," *Symmetry*, vol. 7, no. 4, pp. 1670–1698, 2015.

[8] S. Stasinopoulos and P. Maragos, "Human action recognition using histographic methods and hidden markov models for visual martial arts applications," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 745–748.

[9] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[10] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[11] J. Lee and H. Jung, "Tuhad: Taekwondo unit technique human action dataset with key frame-based cnn action recognition," *Sensors*, vol. 20, no. 17, p. 4871, 2020.

[12] B. Emad, O. Atef, Y. Shams, A. El-Kerdany, N. Shorim, A. Nabil, and A. Atia, "ikarate: Karate kata guidance system," *Procedia computer science*, vol. 175, pp. 149–156, 2020.

[13] M. Błaszczyszyn, A. Szczesna, M. Pawlyta, M. Marszałek, and D. Karczmit, "Kinematic analysis of mae-geri kicks in beginner and advanced kyokushin karate athletes," Aug 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6747486/

[14] T. Hachaj and M. R. Ogiela, "Classification of karate kicks with hidden markov models classifier and angle-based features," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2018, pp. 1–5.

[15] W. Zhang, Z. Liu, L. Zhou, H. Leung, and A. B. Chan, "Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3d human pose estimation," *Image and Vision Computing*, vol. 61, pp. 22–39, 2017.

[16] T. Hachaj, M. Piekarczyk, and M. R. Ogiela, "Human actions analysis: Templates generation, matching and visualization applied to motion capture of highly-skilled karate athletes," *Sensors*, vol. 17, no. 11, 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/11/2590

[17] T. Hachaj, M. R. Ogiela, and K. Koptyra, "Application of assistive computer vision methods to oyama karate techniques recognition," Sep 2015. [Online]. Available: https://www.mdpi.com/2073-8994/7/4/1670/htm

[18] S. Bianco and F. Tisato, "Karate moves recognition from skeletal motion," in *Three-Dimensional Image Processing (3DIP) and Applications 2013*, vol. 8650. International Society for Optics and Photonics, 2013, p. 86500K.

[19] T. Hachaj, M. R. Ogiela, and M. Piekarczyk, "Real-time recognition of selected karate techniques using gdl approach," in *Image Processing and Communications Challenges 5*. Springer, 2014, pp. 99–106.

[20] S. Pinardi and R. Bisiani, "Movement recognition with intelligent multisensor analysis, a lexical approach." in *Intelligent Environments (Workshops)*, 2010, pp. 170–177.

[21] C. Chye, M. Sakamoto, and T. Nakajima, "An exergame for encouraging martial arts," in *International Conference on Human-Computer Interaction*. Springer, 2014, pp. 221–232.

[22] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[23] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[25] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[26] J. Chung, C.-h. Wuu, H.-r. Yang, Y.-W. Tai, and C.-K. Tang, "Haa500: Human-centric atomic action dataset with curated videos," *arXiv preprint arXiv:2009.05224*, 2020.

[27] M. Contributors, "Openmmlab's next generation video understanding toolbox and benchmark," https://github.com/open-mmlab/mmaction2, 2020.

[28] D. M. Tax, M. Van Breukelen, R. P. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?" *Pattern recognition*, vol. 33, no. 9, pp. 1475–1485, 2000.