Abstract—Human Activity Recognition (HAR) has recently received remarkable attentions in numerous applications such as assisted living and remote monitoring. Existing solutions based on sensors and vision technologies have obtained achievements but still suffering from considerable limitations in the environmental requirement. Wireless signals like WiFi-based sensing has emerged as a new paradigm since it is convenient and not restricted in the environment. In this paper, a new WiFi-based and video-based neural network (WiNN) is proposed to improve the robustness of activity recognition where the synchronized video serves as the supplement for the wireless data. Moreover, a wireless-vision benchmark (WiVi) is collected for 9 class actions recognition in three different visual conditions, including the scenes without occlusion, with partial occlusion, and with full occlusion. Both machine learning method - support vector machine (SVM) as well as deep learning methods are used for the accuracy verification of the data set. Our results show that WiVi data set satisfies the primary demand and all three branches in proposed pipeline keep more than 80% of activity recognition accuracy over multiple action segmentation from 1s to 3s. In particular, WiNN is the most robust method in terms of all the actions on three action segmentation compared to the others.

I. INTRODUCTION

Human activity recognition has emerged as an important task recently in numerous applications, such as assisted living [1], human-computer interaction [2], health monitoring [3], surveillance [4], etc. In existing systems, the individual has to wear a device equipped with motion sensors such as a gyroscope and an accelerometer. The sensor data is processed locally on the wearable device or transmitted to a server for feature extraction, and then supervised learning algorithms are used for classification. This type of monitoring is known as active monitoring. The performance of such a system is shown to be around 90 percent for recognition of activities such as sleeping, sitting, standing, walking, and running. However, it is not always convenient to wear devices to monitor the activities in passive applications because of the additional burden and discomfort associated with wearing the device.

Camera-based systems play an important role in passive activity recognition. However, occlusion remains a fundamental challenge. The requirement for light is a major limitation for such systems. Furthermore, it often refers to privacy issues and is limited in many conditions. Unlike sensor-based and video-based solutions, WiFi sensing is not intrusive without the privacy issue and insensitive to lighting conditions. Therefore, WiFi sensing has obtained prevailing popularity in human action recognition [5].

Recent years have witnessed rapid growth of techniques using channel state information (CSI) of WiFi signals in sensing human bodies [6], [7]. Although WiFi CSI has obtained much progress in localizing people and tracking their motion, it is cumbersome for fine-grained human activities recognition because CSI is sensitive to environmental influences. Moreover, the vision information from the camera is a potential complement to assist CSI in some environments without any occlusions. These questions, give us an incentive to develop a new WiFi and video combined pipeline (WiNN) for human action recognition. The synchronous video and WiFi CSI with common supplement and mutual complementary makes WiNN adjustable and feasible for real conditions, especially complex visual conditions, including the scenes without occlusion, with partial occlusion, and with full occlusion, respectively.

To summarize, the contributions of this paper are three-fold.

• We first construct WiVi, wireless-vision activity data set, as a benchmark to evaluate the performance of existing activity recognition systems. We tested SVM, convolution neural network (CNN) as well as the proposed WiNN to verify the effectiveness of the WiVi dataset.

• We propose WiFi-based and video-based neural network as WiNN for activity recognition in scenes with partial occlusion and with full occlusion, which improves the robustness of activity recognition according to the synchronous video as a supplement and complement for WiFi CSI signals.

• We compare machine learning method SVM and deep learning methods CNN and WiNN to verify the quality of the WiVi data set. Particularly, WiNN achieves the most robust results over multiple action segmentation from 1s to 3s.

II. RELATED WORK

To detect human action, two major solutions have been paid attention, namely, camera-based methods [8], [9] and wireless signal based methods including radars (depth maps) [10], LiDARs (3D point clouds) [11], [12] and WiFi devices [13], [14].

A. Camera-based Methods

Camera-based solutions often generate 2D RGB or depth images or videos. It has got more research attention and been
prevailing in human recognition. Traditionally, camera-based methods approach detect human activities by using either a two-step framework [15]–[18] or a part-based framework [19], [20]. The two-step framework first detects all humans forming the bounding box, and then estimates the action from the bounding box independently. The part-based framework first detects body parts independently and then assembles the detected body parts to form multiple human poses to recognize the actions. Both frameworks have their advantages and disadvantages. Yet, camera-based recognition methods are prone to be influenced by the environment such as background, lighting and occlusion, and social limitations such as privacy concerns.

B. Wireless Signal based Methods

The use of passive radar and LiDARs sensors based non-cooperative detection is of increasing interest in many application areas, including defense, security, transport, and healthcare. However, these methods often require dedicated and specially designed hardware.

WiFi devices, which are cheaper and power-efficient than radars and LiDARs, invariant to illumination, and have little privacy concern compared to cameras. However, thus far the potential of WiFi-based methods has not been fully exploited. Mostly WiFi-based methods can only generate a coarse resolution of the recognition and thereby often are used for the perception such as indoor localization [21] and the rough classification [14], [22], [23]. To increase the robustness, some research adopts the WiFi arrays to simulates a 2D antenna array to improve the accuracy of recognition and localization [24]. The authors in [25] have made an attempt to use the WiFi CSI information combined with the videos for human pose interpretation. He adopted Alphapose to generate the 2D skeleton of human pose from the video as the supervised learning of a deep neural network first and then get the human pose estimation by using the WiFi CSI only. Inspired by this two branch idea, we propose a four-branch pipeline for human action recognition.

III. PROPOSED WiFi- AND VIDEO- BASED PIPELINE

A. WiFi-based and Video-based Neural Networks

Convolution neural network (CNN) is a kind of multi-layer neural network, which can extract depth features hierarchically from higher resolution and then convert them into more complex low-resolution features. The classical convolutional neural network consists of two trainable network structures: the convolutional layer, and the pooling layer. The convolutional layer and the pooling layer are used as feature extraction layers, where the convolutional layer provides filtering generation of convolution feature mapping, and the pooling layer produces more abstract and robust high-level features from the convolution feature. The output layer is the full connection layer, dedicated to sorting tasks.

In this work, we design CNN and WiNN based on convolutional neural networks as deep learning methods for the CSI classification. As shown in Fig. 1. The whole scheme is composed of three branches including WiNN (in purple lines), SVM (WiFi channel drawn in green lines), and CNN (WiFi channel with red lines). In addition, all of them can realize HAR based CSI signals. As seen in Fig. 1, the SVM and the CNN branches apply classification on WiFi CSI data only. By contrast, the WiNN branch uses the video as complementary for WiFi signals. WiNN contains two channels, namely the vision channel and the WiFi channel. The vision channel adopts the Alphapose [8], which takes the videos in the scene without occlusion as input in order to generate the human pose skeletons. The generated skeleton results further serve as the supervision for the WiFi channel learning. Then, the WiFi
channel trains the network by the CSI samples according to
the supervision. The trained WiFi channel is able to generate
the human skeleton based on the CSI data only without the
video. Based on the trained network, WiNN has the ability to
generate the human pose skeleton results from CSI samples
in occlusion conditions, and even dark scenes. Finally, the
generated skeletons act as the input of the SVM classifier for
action classification.

B. Channel State Information

CSI depicts how a signal propagates from the transmitter to
the receiver at a certain scale of subcarriers frequencies along
multiple paths. Generally, it can be modeled as follows:

\[ Y = H \times X + N \]

where \( X \) and \( Y \) correspond to the transmitted and received
signal vectors, respectively. \( H \) is the channel matrix repre-
senting the CSI values, and \( N \) is the additive white Gaussian
noise vector. In the frequency domain, CSI defines the
responses derived from orthogonal frequency-division multi-
plexing (OFDM) subcarriers, which can be defined as

\[ h = |h| e^{j \sin \theta} \]

where \(|h|\) and \(\theta\) are the amplitude and phase, \( h \) defines the
CSI values of each subcarrier. In this article, CSI contains
30 subcarriers to describe the multipath propagation affected
by the physical environment (e.g., reflection, diffraction, and
scattering). In particular, CSI amplitudes are further used for
analysis with respect to the fact that phase information are not
generally used for action recognition due to carrier frequency
offset (CFO) and sampling frequency offset (SFO) errors. We
activate 3 antennas both in transmitter and receiver, and then
each obtained CSI data packet contains up to \(3 \times 3 \times 30 = 270\)
raw features of amplitudes.

C. Constructing Wireless-vision Dataset (WiVi)

The data collection is conducted in one spacious office
apartment with five different rooms, i.e. kitchen, printer room,
private office, and the main office room as shown in Fig.
2. The office rooms have standard furniture: tables, chairs,
boards, etc. The interior walls of the building are 6-inch hollow
walls supported by steel frames with sheetrock on top. The
door and windows were closed to ensure stability in the test
environment.

The experimental hardware consists of two desktop com-
puters as transmitter and receiver operating in IEEE 802.11n
monitor mode at 5.4 GHz. Both desktops are equipped with
off-the-shelf Intel 5300 card. They run Ubuntu 14.04 with a
modified Intel driver. During the experiment, the transmitter
desktop and the receiver desktop are fixed located on the
left side of the office, while the volunteer persons perform
activities in different areas to simulate the occlusion occasions.

During the data capturing, wireless and vision data is
synchronously collected. WiFi signals are collected by the
receiver at a sampling rate of 100 packets per second. The
receiving desktop computer calculates CSI data for ev-
ery packet received from the transmitter. We employ the
CSI extraction tool (introduced in https://github.com/dhalperi/
linux-80211n-csitool-supplementary) for CSI signals record-
ning and CSI packets extraction. The CSI measurement from
each data packet contains up to \(3 \times 3 \times 30 = 270\) raw features.

To keep pace with the CSI recording, a deep camera
D435i (available at https://www.intelrealsense.com/
depth-camera-d435i/) is attached to our receiver desktop at
the same location of the wireless card. To synchronize the images and wireless data, we record the video at 20 FPS. That is, every five CSI samples are corresponding to one frame in the video, as shown in Fig. 4. Fig. 4a is the CSI sequence and Fig. 4b is the corresponding frame recorded in videos.

We recruited 2 volunteers to join our evaluation, including 1 male and 1 female. We implemented 9 activities, i.e. falling down, throwing, pushing, kicking, punching, jumping, drinking, phone talking, and seating.

In reality, such kind of activities is complex. To simulate as many situations as possible, each action is repeatedly performed by the volunteer and lasts about 20-30 seconds. During the time, the activities are repeated in different motion details, such as varied directions, speed, intensity, strength, and etc., to ensure the diversity of the actions. In addition, we also simulate the occlusion occasions, as shown in Fig. 3, including three scenarios, namely, the scenes without occlusion (as shown in Fig 3a), with partially occlusion (as seen in Fig. 3b) and with fully occlusion (Fig. 3c).

D. Pre-processing of CSI

The gathered CSI is easily affected by the surrounding electromagnetic noise. An example for dynamic falling-down action is demonstrated in Fig.4a. A relatively static phone-talking action is shown in Fig.5a. Both of them are from the first subcarrier between the first transmitter antenna and the first receiver antenna of the female person in the scene without occlusion. Fig.4a and Fig.5a show that the CSI sequence both behaves periodicity changes of signals when the person performs an action. Meanwhile, the fluctuations of the dynamic and the static action curves are obviously different. Therefore, the CSI has the potential to distinguish different actions. However, both of the two action CSI sequences are very noisy signals. To address the issue, we apply three preprocessing steps to remove the noises in both time and frequency domains.

- Outliers removing. As seen in the first rows of Fig. 6 as well as Fig. 7, the raw CSI inevitably contains some outliers, that values are extremely high or low. To mitigate the performance of CSI, a median filter is applied to get a smooth CSI sequence, where the sliding window is 40, as illustrated in the second row of Fig. 6 as well as Fig. 7.
- Smoothing. After the above operation, there still exists slight fluctuations, the median filter is adopted to smooth the CSI sequences with a sliding window of 40. The third rows of Fig. 6 and Fig. 7 draw the smooth results.
- Low pass filtering. Butterworth filter with fifth order and 10Hz pass-band is implemented to filtered out high-frequency noise, as shown in the fourth rows Fig. 6 and Fig. 7.

E. Action Segmentation and Data Augmentation

A prerequisite for further CSI processing such as classification is action segmentation, as the CSI sequences are recorded along time. Ever since the advent of CSI, tremendous efforts have been made to develop robust and accurate action segmentation, such as the classical envelope extraction method [26] by decomposing data into intrinsic mode functions for action separation. However, this method is only viable supposing the scenarios is stationary and not suitable in most actual cases. In addition, there are differences varying from person to person causing the variance in the movement. Therefore the empirical design of a special action segmentation method is cumbersome for other cases. How to find an effective way of action segmentation remains a fundamental challenge. To this end, we propose a way for action segmentation by time in this work. That is because each CSI sequence records the same activities, it is assumed to be separated into different movements by an equal time interval.

After the action segmentation, to make CSI actions set more robust to the noise, data augmentation is performed on the
Fig. 4. Raw CSI packets and synchronous frames of female person for falling down in the scene without occlusion. The data is CSI in the first subcarrier between the first transmitter antenna and the first receiver antenna. According to frame images, the complete action is denoted for the CSI sequence.

Fig. 5. Raw CSI packets and synchronous frames of female person for phone talking in the scene without occlusion. The data is CSI in the first subcarrier between the first transmitter antenna and the first receiver antenna.

action sets. Inspired by the data augmentation method in [4], we propose a method including two steps also.

- Averaging an action into CSI samples. We first uniformly split the sequence into segments as the individual actions. Then we extract 10 CSI packages with equal intervals in a segment derived from the action segmentation step, and then iteratively select every 10 CSI packages with a stride of size 1. At last, these 10 CSI packages are averaged into one CSI sample separately for further processing.
- Data augmentation for the training data. With respect to the limited number of CSI sequences, it tends to cause the over-fitting problem. To address this obstacle, the data augmentation described in [4] is applied with $k = 3, 5, 7, 9$ to augment the dataset.

IV. IMPLEMENTATION AND EVALUATION

A. Experiment Setup

1) Datasets: As described in subsection B of section IV, we use 1, 2, 3 seconds to segment the CSI sequences to get the actions dataset WiVi. Then, we average the action segmentation sets into CSI samples and augment the CSI samples to form the training datasets WiVi. At last, the dataset for three action segmentation are shown in Table I. The dataset was split into a training and a testing sets with an 70 - 30% split respectively. All the test results on our dataset use only WiFi signals without vision-based input.

<table>
<thead>
<tr>
<th>Segment time</th>
<th>Sets</th>
<th>training instances</th>
<th>testing instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1s</td>
<td>24938</td>
<td>10707</td>
<td></td>
</tr>
<tr>
<td>2s</td>
<td>24398</td>
<td>10483</td>
<td></td>
</tr>
<tr>
<td>3s</td>
<td>23840</td>
<td>10280</td>
<td></td>
</tr>
</tbody>
</table>

TABLE I
THE DATASET FOR THREE ACTION SEGMENTATION

In this section, we employed both the traditional machine learning and deep learning methods to verify the efficiency of the dataset, namely, SVM [27], CNN [27], and WiNN. The comparisons of three methods give the quantitative evaluation to demonstrate the efficiency of the dataset WiVi, while WiNN [28] generates the skeleton results for the visual inspection of the WiVi dataset.

B. Parameter setup

1) SVM: SVM adopts the Radial basis function (RBF) as the kernel with a gamma of 0.143. The penalty is set as 100, and the pyramid level and probability threshold both define as 0.

2) CNN: CNN is consist of 4 basic blocks of ResNet [29] with 18 convolutional layers (ResNet18). At first, we
upsample each CSI sample from $30 \times 3 \times 3$ into $6 \times 224 \times 224$ to be compatible with the input of ResNet. After this, the upsampled CSI samples are put forward to ResNet18 modules for classification. It was trained using the Adam optimizer with a 0.001 learning rate and 0.01 decay rate for 5 epochs.

3) WiNN: WiNN also adopts ResNet16 where the same upsampling processing as CNN to generate the required size $6 \times 224 \times 224$. Then, four ResNet modules are stacked together to extract the features of size $300 \times 18 \times 18$ from the upsampled CSI data. After this, two convolutional layers are constructed to decode the features into $2 \times 18 \times 18$ as the skeleton features. The output skeleton features are further taken into the SVM classifier for final action classification. The parameters of SVM is the same as the SVM branch.

<table>
<thead>
<tr>
<th>Action</th>
<th>Method</th>
<th>SVM</th>
<th>CNN</th>
<th>WiNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PA</td>
<td>OA</td>
<td>PA</td>
<td>OA</td>
</tr>
<tr>
<td>1s falling</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>throwing</td>
<td>0.89</td>
<td>0.99</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>pushing</td>
<td>1</td>
<td>0.91</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>kicking</td>
<td>0.98</td>
<td>0.97</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>punching</td>
<td>1</td>
<td>0.95</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>jumping</td>
<td>0.91</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>phonetalk</td>
<td>1</td>
<td>1</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>seating</td>
<td>0.8</td>
<td>1</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>drinking</td>
<td>0.96</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2s falling</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>throwing</td>
<td>0.92</td>
<td>0.97</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>pushing</td>
<td>0.9</td>
<td>0.85</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>kicking</td>
<td>0.94</td>
<td>0.99</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>punching</td>
<td>0.93</td>
<td>0.94</td>
<td>0.74</td>
<td>0.91</td>
</tr>
<tr>
<td>jumping</td>
<td>0.94</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>phonetalk</td>
<td>0.91</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>seating</td>
<td>0.88</td>
<td>0.89</td>
<td></td>
<td>0.81</td>
</tr>
<tr>
<td>drinking</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3s falling</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>throwing</td>
<td>0.89</td>
<td>0.8</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>pushing</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>kicking</td>
<td>0.95</td>
<td>1</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>punching</td>
<td>0.95</td>
<td>0.97</td>
<td>0.87</td>
<td>0.95</td>
</tr>
<tr>
<td>jumping</td>
<td>0.95</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>phonetalk</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>seating</td>
<td>0.93</td>
<td>0.94</td>
<td></td>
<td>0.89</td>
</tr>
<tr>
<td>drinking</td>
<td>1</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE II

<table>
<thead>
<tr>
<th>Action</th>
<th>Method</th>
<th>SVM</th>
<th>CNN</th>
<th>WiNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PA</td>
<td>OA</td>
<td>PA</td>
<td>OA</td>
</tr>
<tr>
<td>1s falling</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>throwing</td>
<td>0.89</td>
<td>0.99</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>pushing</td>
<td>1</td>
<td>0.91</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>kicking</td>
<td>0.98</td>
<td>0.97</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>punching</td>
<td>1</td>
<td>0.95</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>jumping</td>
<td>0.91</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>phonetalk</td>
<td>1</td>
<td>1</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>seating</td>
<td>0.8</td>
<td>1</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>drinking</td>
<td>0.96</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### C. Quantity Comparison

The subsection shows the activity recognition performance using several classification algorithms on the WiVi dataset. We analyse activity data of all volunteers. On one hand, we demonstrate the the validity of WiVi data set; on the other, we evaluate the performance of SVM, CNN and WiNN respectively in terms of three action segmentation. As illustrated in Table.II, PA is product accuracy and OA represents overall accuracy. Fig.8 shows the corresponding confusion matrices for three action segmentation by SVM, CNN, and WiNN.

As illustrated in Table.II, although our data set is really challenging in terms of huge differences from scenes, persons as well as activity performances, all three methods still distinguish 9 actions well and keep more than 80% of activity recognition accuracy. The results demonstrated the robustness of the data set WiVi.

By considering each actions, our method WiNN achieves most robust results compared to compared with SVM and CNN. For instance, WiNN obtains the most actions with full mark of classification accuracy regard to all action segmentation. To further evaluate our method intuitively , we report the results on three action segmentation in Fig.8.

We study the impact of action segmentation on performance of activity recognition by using three classification algorithms. It shows that the most suitable action segmentation for CSI sequences varies according to the classification model. For SVM, it achieved the most accurate OA for 3s as all 9 activities classes obtain the satisfied PA to some degree. For
CNN, it witnesses the highest accuracy of 98% in 1s action segmentation. Compared with SVM and CNN, WiNN achieves more robust results for 3s with respect to the PAs of 9 actions.

D. Visually Interpretation

The section analyses and elaborates the robustness of the data set intuitively, we randomly selected twelve CSI samples and corresponding frames from each activity. To have a

intuitive interpretation, we only show two scenarios, namely without occlusion scene and with partial occlusion, to produce the skeleton results by WiNN in Fig. 9. As shown in the figure, though the WiVi data set is diverse, it yields comparable results to videos when there is no occlusion, and even better results when there exists partial occlusion.
V. CONCLUSION

In this paper, we proposed a WiNN scheme to tackle the occlusion problem of activity recognition. To evaluate the performance of existing methods, we collected WiVi dataset as a benchmark in three scenes, including without occlusion, with partial occlusion, and with full occlusion. Three methods, namely, machine learning method SVM, deep learning methods CNN as well as the proposed WiNN were used to verify the effectiveness of the WiVi data set. Experimental results showed that WiVi dataset satisfied primary demand, and WiNN achieved the most robust results compared to other two methods regard to multiple action segmentation from 1s to 3s. In future, it is promising to extend the diversity of WiVi dataset with respect to activity, location, and speed. Specifically, we will investigate a deep learning pipeline to realize the automatic selection of the branch for recognition according to the input data types.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Yixuan Zou and Zhong Yang as the volunteers of the experiments.

REFERENCES


