Knowledge-driven Egocentric Multimodal Activity Recognition

YI HUANG, XIAOSHAN YANG, and JUNYU GAO, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences, China and Peng Cheng Laboratory, China
JITAO SANG, School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China and Peng Cheng Laboratory, China
CHANGSHENG XU, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences, China and Peng Cheng Laboratory, China

Recognizing activities from egocentric multimodal data collected by wearable cameras and sensors, is gaining interest, as multimodal methods always benefit from the complementarity of different modalities. However, since high-dimensional videos contain rich high-level semantic information while low-dimensional sensor signals describe simple motion patterns of the wearer, the large modality gap between the videos and the sensor signals raises a challenge for fusing the raw data. Moreover, the lack of large-scale egocentric multimodal datasets due to the cost of data collection and annotation processes makes another challenge for employing complex deep learning models. To jointly deal with the above two challenges, we propose a knowledge-driven multimodal activity recognition framework that exploits external knowledge to fuse multimodal data and reduce the dependence on large-scale training samples. Specifically, we design a dual-GCLSTM (Graph Convolutional LSTM) and a multi-layer GCN (Graph Convolutional Network) to collectively model the relations among activities and intermediate objects. The dual-GCLSTM is designed to fuse temporal multimodal features with top-down relation-aware guidance. In addition, we apply a co-attention mechanism to adaptively attend to the features of different modalities at different timesteps. The multi-layer GCN aims to learn relation-aware classifiers of activity categories. Experimental results on three publicly available egocentric multimodal datasets show the effectiveness of the proposed model.

This work was supported by National Key Research and Development Program of China (No. 2018AAA0100604), National Natural Science Foundation of China (No. 61720106006, 62072455, 61702511, 61751211, 61620106003, 61532009, U1836220, U1705262, 61872424) and Key Research Program of Frontier Sciences of CAS (QYZDJSSWJSC039). This work was also supported by Research Program of National Laboratory of Pattern Recognition (No. Z-2018007).

Authors' addresses: Y. Huang, X. Yang, and J. Gao, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences, NO. 95 Zhongguancun East Rd., Haidian District, Beijing, China, and Peng Cheng Laboratory, Shenzhen, China; emails: {yi.huang, xiaoshan.yang}@nlpr.ia.ac.cn, gaojunyu2015@ia.ac.cn; J. Sang, School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, 506 Room in No.9 Teaching Building of Beijing Jiaotong University, Haidian District, Beijing, China, and Peng Cheng Laboratory, Shenzhen, China; email: jtsang@bjtu.edu.cn; C. Xu, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences, NO. 95 Zhongguancun East Rd., Haidian District, Beijing, China, and Peng Cheng Laboratory, Shenzhen, China; email: csxu@nlpr.ia.ac.cn.

1551-6857/2020/12-ART133 $15.00
https://doi.org/10.1145/3409332

CCS Concepts: • **Computing methodologies** → **Activity recognition and understanding**; • **Theory of computation** → *Semantics and reasoning*;

Additional Key Words and Phrases: Egocentric videos, wearable sensors, graph neural networks

ACM Reference format:
Yi Huang, Xiaoshan Yang, Junyu Gao, Jitao Sang, and Changsheng Xu. 2020. Knowledge-driven Egocentric Multimodal Activity Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 4, Article 133 (December 2020), 21 pages.
https://doi.org/10.1145/3409332

---

1 INTRODUCTION

With the advancement of Mobile Internet and the Internet of Things, various wearable devices, such as mobile phones, portable cameras, and smart wristbands, are widely used in people's daily lives. The widespread use of these devices enables a low-cost and autonomous collection of large-scale multimodal data, such as personal photos, the egocentric (i.e., first-person view) videos, accelerometer data, heart rate, and GPS, which can record people's daily physical activities at any place and any time. Automated understanding and analysis of the collected raw data is extremely valuable in many applications such as healthcare monitoring [2, 68] and human-computer interaction [69]. For example, learning user behaviors facilitates to optimize the course of the day with regard to dietary control or sport.

Egocentric human activity recognition has been extensively investigated in the past decade with the help of various sensors. In pervasive computing area, the data streams collected from wearable sensors, such as accelerometer and gyroscope, are widely exploited for behavior analysis [18, 19, 40, 62, 66, 67]. However, the activity recognition performance is often limited by the drifts of sensors (e.g., mobile phone), which occurs during user's long operation time [8]. In computer vision area, the encouraging progress of deep learning–based research on activity recognition from exocentric (i.e., third-person view) videos [12, 24, 26, 63, 70] has inspired the research interest in egocentric videos (i.e., first-person view) [3, 42, 46]. However, compared with exocentric videos, the invisibility of the activity performer in the egocentric videos generates extra challenges to activity recognition task. In many tasks of the multimedia and computer vision fields, such as event detection [48, 55], micro-video understanding [31, 44, 65], and multiple social networks learning [17, 43], multimodal methods always benefit from the complementarity of different modalities. Thus, combining data streams collected from the wearable sensors and the egocentric cameras is useful to alleviate the sensor drift and wearer invisibility problems in recognizing activities using the single modality data. Although promising progress has been achieved, most of the existing methods only focus on directly concatenating multimodal features [2, 41, 56] or selectively using single modality features in different scenes [47]. They do not pay much attention to the challenge of the **large modality gap**: High-dimensional videos, consisting of spatial and temporal information, contain rich, high-level semantic information [59], while the low-dimensional sensor signal data only describe the simple motion pattern of the wearer. The simple fusion scheme may fail to exploit the complex semantic information in egocentric videos.

Another challenge in egocentric multimodal activity recognition is the **lack of large-scale annotated datasets**. Recently, third-view activity recognition has gained significant progress due to the availability of large-scale datasets, such as ActivityNet [5] and Kinetics [7]. However, the largest publicly available egocentric dataset including both video and sensor data, i.e., Stanford-ECM [41], has only 113 videos of 23 activity categories. Recently, Karpathy et al. [26] found that 3D CNN architectures are not able to properly learn motion features when there is non-existence of sufficiently large datasets. However, it is not an easy way to obtain a large-scale dataset from wearable sensors and cameras due to the high cost of data collection and class annotation. Alternatively, there is an imperative requirement of new egocentric activity recognition model that works well on the relatively small dataset.

Recently, knowledge graph has attracted notable attention as a flexible data structure for representing relationships of real-world entities. It has been increasingly applied in various fields including object recognition [36, 50] and video classification [15, 54]. Complex human actions have been shown to strongly relate to the objects in the context where human is embedded [10]. Actually, modeling semantic relations of intermediate objects and activities is also helpful for solving the challenge of the modality gap between the egocentric videos and the sensor data. Specifically, the relations among intermediate semantic features and activities to be recognized provide top-down guidance to adaptively find the most important features of different modalities to fuse for correct activity prediction. For example, egocentric videos always contain shaken and blurred shots due to the natural movements of the wearers. It is hard to find the effective visual feature to combine with the motion feature of the sensors in an bottom-up scheme. Instead, with the awareness of the high-level relations between objects and activities, the recognition model is more likely to make the correct decision to focus on the visual features of videos or movement patterns of sensor data. Moreover, Salakhutdinov et al. [50] demonstrate that different classifiers can share the implicit representations over the knowledge graph so the classifiers with few training samples can borrow statistical information from other classifiers with explicit relations (i.e., edges) in knowledge graph. Thus, the knowledge graphs are also suitable for solving the challenge of lack of large-scale annotated datasets in egocentric multimodal activity recognition.

Motivated by above observations, we propose a knowledge-driven multimodal activity recognition framework that exploits external knowledge to enhance the performance of activity recognition on the egocentric data. In this work, we mainly focus on the egocentric video and the accelerometer/gyroscope data. It is not difficult to extend it to other kinds of modalities. Figure 1 shows an overview of the proposed framework. Specifically, we first build a single-modality global prediction module that calculates the preliminary activity scores based on the motion feature of the sensor signal and the vision feature of the egocentric video, respectively. Then, we propose a knowledge-driven multimodal prediction module consisting of a dual-GCLSTM (Graph Convolutional LSTM) and a multi-layer GCN (Graph Convolutional Network), which collectively model the relations among activities and intermediate objects. The dual-GCLSTM is designed to model temporal multimodal features with the top-down relation-aware guidance. In addition, a co-attention mechanism is adopted to adaptively attend to different modality data at different timesteps. The multi-layer GCN aims to learn relation-aware classifiers of different activity categories. By leveraging external knowledge to model the relations among activities and intermediate objects, the classifiers can also reduce the dependence on large-scale training samples.

The main contributions of this article are summarized as follows:

(1) We propose a knowledge-driven egocentric multimodal activity recognition framework that can collectively leverage the external semantic context and relationship knowledge to augment the conventional recognition model.
(2) We propose a dual-GCLSTM to dynamically combine relation-aware multimodal features, which can alleviate the modality gap, and a multi-layer GCN to comprehensively learn relation-aware activity classifiers, which can reduce the dependency on large-scale training samples.
(3) We evaluate the proposed framework against several competitive existing methods. The extensive experiment results on three public datasets demonstrate the effectiveness of the proposed method.

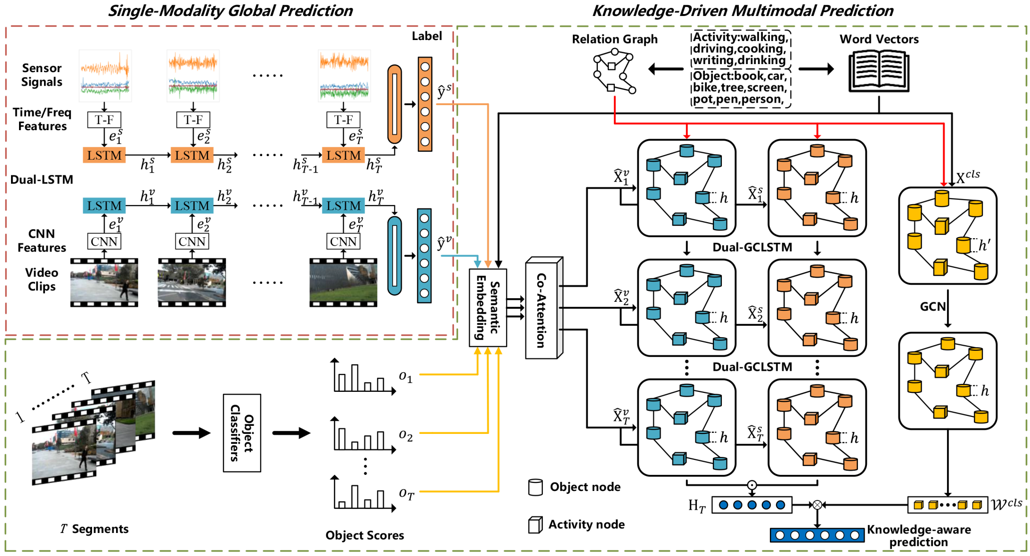


Fig. 1. An overview of the proposed knowledge-driven framework for multimodal activity recognition. In single-modality global prediction module, the dual-LSTM models the sequential features $\{\mathbf{e}_t^v\}_{t=1}^T$ and $\{\mathbf{e}_t^s\}_{t=1}^T$ to produce preliminary activity scores $\hat{\mathbf{y}}^v$ and $\hat{\mathbf{y}}^s$. In knowledge-driven multimodal prediction module, object classifiers first produce object scores $\{\mathbf{o}_t\}_{t=1}^T$ of each video clip. The object and activity scores of $\mathbf{o}_t$, $\hat{\mathbf{y}}^v$, and $\hat{\mathbf{y}}^s$ are used to multiply corresponding object or activity word-embeddings to obtain input node features $\mathbf{X}_t^v$ and $\mathbf{X}_t^s$ of graph. Then $\mathbf{X}_t^v$ and $\mathbf{X}_t^s$ are fed into co-attention module to produce weighted features $\hat{\mathbf{X}}_t^v$ and $\hat{\mathbf{X}}_t^s$. The Dual-GCLSTM is used to fuse multimodal data streams $\{\hat{\mathbf{X}}_t^v\}_{t=1}^T$ and $\{\hat{\mathbf{X}}_t^s\}_{t=1}^T$ and produce final concept features $\mathbf{H}_T$ with the consideration of temporal dynamic patterns of knowledge evolution. In addition, the GCN is used to model the static relations among concepts and learn the relation-aware classifiers $\mathcal{W}^{cls}$. The relation structure of the graph in all branches is constructed from external knowledge graph. The final knowledge-aware activity prediction is implemented via the learned $\mathbf{H}_T$ and $\mathcal{W}^{cls}$.

2 RELATED WORK

In this section, we review the most related work to our method in the following three aspects.

2.1 Egocentric Activity Recognition

In the past several years, activity recognition has been extended to egocentric cameras as well as wearable sensors and has attracted a lot of attention. We first review the methods based on the single modality data. Then, we introduce the state-of-the-art multimodal egocentric activity recognition methods.

For video-based egocentric activity recognition, traditional methods rely on handcrafted features such as objects [34, 46], hands [39, 61], and gaze [14, 29]. However, these methods strongly rely on prior knowledge in constrained environments, but have not addressed the datasets in natural environment. Recently, deep learning–based methods have achieved more successful performances in large-scale activity datasets. One research line of egocentric activity recognition methods directly follows the third-view activity recognition. Simonyan et al. [53] propose a two-stream Convolutional Neural Network over image frames and optical flows to explore spatial and temporal information. Song et al. [56] extend this method to egocentric video domain. Another research line focuses specifically on hand and object cues for activity recognition. Pirsiavash et al. [46] explore active object detection as an auxiliary task for activity recognition. Cai et al. [6] propose a structured approach where grasp types, object attributes, and their contextual relationships are analyzed together. Baradel et al. [1] extend this approach with considering both the spatial and temporal information.

Sensor-based activity recognition task has also been studied widely due to its low-power and lost-cost advantages. Bulling et al. [4] has shown that statistical features (e.g., time-domain and frequency-domain features) of sensor signals can achieve advanced performance in recognizing activities. In addition, CNN-based [18, 19] and RNN-based [40, 66] architectures have also achieved competitive performances.

More recently, there is an emerging tendency to incorporate video modality and sensor signal modality together to improve the accuracy of egocentric activity recognition. Hsieh et al. [22] leverage several mid-level representations in the surrounding of a subject as essential cues for inferring the activity class. The handcrafted features represent *what*, *where,* and *how* a subject is interacting with the context. However, this method only fuses the posterior probabilities based on each modality but ignores the interaction among the extracted mid-level concepts. Song et al. [56] propose a multi-stream feature extractor with a two-level multimodal fusion technique for egocentric multimodal activity recognition. Nakamura et al. [41] make activity recognition via concatenating video and sensor signal features and use the heart rate signal as self-supervised information to enhance the model performance. Bernal et al. [2] use an LSTM-based temporal fusion method on video and sensor data, where the correlations of two modalities are modeled into the hidden states of the LSTM. Possas et al. [47] propose a reinforcement learning framework to select the video or sensor signal modality for activity predicting in different scenes. This method reduces the computational consumption and increases the accuracy simultaneously. However, the above methods only directly combine multimodal features or selectively use single modality feature in different scenes, which cannot explicitly model the complementarity of videos and sensor signals.

2.2 Knowledge Graphs

Learning knowledge graphs and reasoning on graphs have recently been of interest to the vision and multimedia communities, such as object detection [50], image recognition [36], video retrieval [9], and visual question answering [25, 49]. Salakhutdinov et al. [50] demonstrate that different classifiers can share the implicit representations in the knowledge graph so classifiers with few training samples can borrow the statistical information from other classifiers with explicit relations. It is extremely helpful to deal with the unavailability of the large-scale annotated activity dataset. Marino et al. [36] propose a graph search neural network to exploit structured prior knowledge into image classification. Chen et al. [9] employ the knowledge graph on video retrieval task. They build the knowledge graph on user described queries and use Conditional Random Field to learn the node/edge predictors for semantic matching. Sadeghi et al. [49] introduce a visual knowledge extraction system to reason the entities in the context of a given relation phrase and exploit this system into question-answering task. To the best of our knowledge, there is no existing work that incorporates external knowledge to improve the egocentric activity recognition.

2.3 Graph Neural Networks

Graph Neural Network (GNN), a neural network employment for structured graphs, has gained much attention, as it can effectively employ local graph operation with learnable filters. Kipf et al. [28] use the GCN to handle a semi-supervised node classification problem. By propagating the information over explicit relation edges, the embedding features of nodes can be learned with only few labeled information. Lately, Wang et al. [64] regard graph nodes as classifiers to deal with the zero-shot problem. They model the semantic embeddings (represented as nodes) and the relations (represented as edges) between categories to learn the classifiers of unseen categories.

Gao et al. [16] propose that building knowledge graph with both the classifier and the pre-defined attribute can further improve generalization.

GNN has also been designed to model sequential structured data. Seo et al. [51] combine the GNN and RNN to build Graph Convolutional Recurrent Network (GCRN). It can identify the spatial structures on the graph via graph convolutional operations and capture temporal dependencies of the data via recurrent operations. Li et al. [30] propose a diffusion method that incorporates graph convolutional layer into GRU networks for spatiotemporal forecasting in traffic environments.

Different with the above methods, we employ GNN to enhance the prediction performance of the egocentric multimodal activity recognition with the external knowledge. We propose a knowledge-aware prediction module where a GCN is adopted to learn classifiers from semantic embeddings of activities and objects with explicit relations, and a dual-GCLSTM is adopted to model the dynamic patterns of the knowledge evolution in both the video and sensor signal streams.

3 METHODOLOGY

3.1 Overview

We give the details of our framework for Egocentric Multimodal Activity Recognition in this section. Formally, let $\mathcal{V}$ denote a video sequence and $\mathcal{S}$ denote a sequence of wearable sensor signals. We formulate this task as a sequential prediction problem where both $\mathcal{V}$ and $\mathcal{S}$ are segmented into T equal-length segments $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T\}$ and $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_T\}$. In addition, we have an object set $\mathcal{O}$ containing O objects that will be used as the intermediate features. The object set is important for building a top-down relation-aware guidance to fuse multimodal data to enhance the recognition capacity. Our target is to calculate the activity probability distribution over $\mathbf{y} = \{y_1, y_2, \ldots, y_C\}$ based on $\mathcal{V}$, $\mathcal{S}$, and $\mathcal{O}$. Here, C is the number of activity classes. The number of segments T is fixed for all videos and wearable sensor signals to perform sequential parallelization in our framework.

The framework of the proposed model is shown in Figure 1. We first create the single-modality global prediction module, which consists of a motion predictor and a vision predictor to produce the preliminary activity scores. Here, we use a dual-LSTM to extract the global motion feature based on sensor signals and the global vision feature based on videos, respectively. Then, we design a knowledge-driven multimodal prediction module to fuse two single-modality predictors with semantic features and knowledge graph. A dual-GCLSTM is adopted to produce the relation-aware multimodal features, and a multi-layer GCN is adopted to learn the final knowledge-aware classifiers. Specifically, the dual-GCLSTM is designed to model the dynamic knowledge evolution on the two modalities. At each timestep, the embedding vectors of the objects and activities are used as the node inputs of the GCLSTM. The relations (i.e., edges) among the objects and activities are built based on ConceptNet. In addition, a co-attention mechanism is adopted to adaptively attend to the features of different modalities at different timesteps. The multi-layer GCN is designed to learn relation-aware activity classifiers. In this branch, the input nodes of GCN are the word embeddings of all possible activities and objects, and the adopted relation graph is same as GCLSTM. Finally, the relation-aware classifiers are used to perform knowledge-aware prediction based on the relation-aware multimodal features.

3.2 Single-modality Global Prediction

The single-modality global prediction module is implemented by two preliminary predictors based on the video modality and the sensor signal modality, respectively.

3.2.1 Motion Predictor. We directly extract the Time-domain and Frequency-domain (T-F) features from the raw wearable sensor signals as in References [32, 41]. For time-domain features, we compute mean, standard deviation, skewness, kurtosis, percentiles (10th, 25th, 50th, 75th, 90th), the counts for each signal axis, as well as the correlation coefficients between each axis of sensor signal. For frequency-domain features, we use the spectral entropy, which is computed through short-time fourier transform (STFT) [45]. All time-domain and frequency-domain features of the sensor signal segment $\mathbf{s}_t$ are concatenated to get the final representation $\mathbf{e}_t^s \in \mathbb{R}^{d_s}$. The d_s is the dimension of the sensor signal feature. The resulting hand-crafted features are denoted as $\mathcal{E}_s = \{\mathbf{e}_1^s, \mathbf{e}_2^s, \ldots, \mathbf{e}_T^s\}$.

Then, we feed the sequential sensor signal features $\mathcal{E}_s$ into LSTM [21] networks, which are good at solving the problem of long-term dependencies compared with traditional RNNs, to capture the temporal structure of sensor signal modality:

$$\mathbf{h}_t^s = \mathrm{LSTM}_s\left(\mathbf{e}_t^s, \mathbf{h}_{t-1}^s\right), \tag{1}$$

where $\mathbf{h}_t^s$ is the hidden state in LSTM cell. To get the activity probability distribution, we employ a fully connected layer followed with a Softmax activation layer on the hidden state $\mathbf{h}_T^s$ of the final timestep:

$$\hat{\mathbf{y}}^s = \mathrm{Softmax}\left(\mathbf{W}^s \mathbf{h}_T^s + \mathbf{b}^s\right), \tag{2}$$

where $\mathbf{W}^s$ and $\mathbf{b}^s$ are trainable parameters, and $\hat{\mathbf{y}}^s \in \mathbb{R}^C$ is the probability distribution over all activity categories. The motion predictor is appropriate to capture the body movement–related patterns such as running, walking, and cycling. In other words, it will perform poorly for activities with limited movement.

3.2.2 Vision Predictor. Referring to the recently advanced work in video analysis domains such as video classification [15] and activity recognition [16, 41], we employ a pre-selected feature extraction layer of the deep CNNs Inception-V3 [58] pre-trained on ImageNet [11] to extract visual features over the frames sampled from the video modality. Then, we perform mean pooling over those frame features of the video segment $\mathbf{v}_t$ to compute the final representation $\mathbf{e}_t^v \in \mathbb{R}^{d_v}$. The resulting visual features of the video segments are denoted as $\mathcal{E}_v = \{\mathbf{e}_1^v, \mathbf{e}_2^v, \ldots, \mathbf{e}_T^v\}$.

Same as how to design the motion predictor, we feed the sequential features $\mathcal{E}_v$ into LSTM networks:

$$\mathbf{h}_t^v = \mathrm{LSTM}_v\left(\mathbf{e}_t^v, \mathbf{h}_{t-1}^v\right), \tag{3}$$

where $\mathbf{h}_t^v$ is the hidden state. Then, we employ a fully connected layer followed with a Softmax layer on the final hidden state $\mathbf{h}_T^v$ to get the activity probability distribution $\hat{\mathbf{y}}^v \in \mathbb{R}^C$:

$$\hat{\mathbf{y}}^v = \mathrm{Softmax}\left(\mathbf{W}^v \mathbf{h}_T^v + \mathbf{b}^v\right), \tag{4}$$

where $\mathbf{W}^v$ and $\mathbf{b}^v$ are trainable parameters.

3.3 Knowledge-driven Multimodal Prediction

In this section, we will introduce the knowledge-driven multimodal prediction module. The target of this module is to enhance the generalization capacity of the motion predictor and vision predictor in the manner of exploiting the explicit relations among activities and objects via graph neural network and knowledge graph. The object set $\mathcal{O}$ (containing O objects) is introduced as an intermediate to connect the low-level vision/motion features and the high-level activities to be predicted. The structured graph $\mathcal{G}$ describes the explicit relations (i.e., edges) among all the activities and objects (i.e., nodes).

We will first introduce the main building block of our knowledge-driven prediction module, i.e., graph convolutional layer, which is well known, as it can learn layer-wise propagation operations on graphs [28]. Graph convolutional operations are used for producing temporal dynamic knowledge-aware semantic features of multimodal data, as well as learning static knowledge-aware classifiers. To reduce the clutter in the following parts, we use a simplified notation to represent a one-layer graph convolutional operation as follows:

$$\mathbf{W} *_{\mathcal{G}} \mathbf{X} \overset{\Delta}{=} \mathbf{Z} = \hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X}\mathbf{W}, \tag{5}$$

where $\mathbf{X}$ is the input node features. $\mathbf{W} *_{\mathcal{G}}$ denotes a one-layer graph convolution operation based on relation graph $\mathcal{G}$ with trainable filter $\mathbf{W}$. Specifically, $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}_m$ is the adjacency matrix of the knowledge graph $\mathcal{G}$ with self-connections, $\mathbf{I}_m$ is the identity matrix. $\hat{\mathbf{D}}$ is the degree matrix of $\hat{\mathbf{A}}$ as $\hat{\mathbf{D}}_{ii} = \sum_j \hat{\mathbf{A}}_{ij}$. Note that the row vector $\mathbf{X}_i \in \mathbb{R}^k$ is the feature representation of the ith node. As a result, the input $\mathbf{X} \in \mathbb{R}^{m \times k}$ for this graph convolutional layer is transformed into a new feature space $\mathbf{Z} \in \mathbb{R}^{m \times c}$ with filter $\mathbf{W} \in \mathbb{R}^{k \times c}$.

In our work, the nodes of the graph are represented by $k-$dimensional word vectors pre-trained on the large-scale corpus with rich context information. We use the $\mathbf{X}^{cls} \in \mathbb{R}^{(C+O) \times k}$ to denote the word-embedding matrix of all activity and object nodes. The edges of the graph are created based on the extra public knowledge base, e.g., ConceptNet [57]. More details of building the graph are introduced in Section 4.2.

3.3.1 *Dynamic Knowledge-aware Semantic Feature.*

Our motivation is to fuse sensor signal and video in the manner of modeling the explicit relations among all activities and intermediate objects that vary dynamically over the time. To capture the knowledge evolution with a graph structure, we first build two-stream semantic embeddings as graph node inputs based on intermediate object scores and preliminary activity scores of two modalities. Then, we use the dual-GCLSTM to produce knowledge-aware features, which can model the temporal dynamic patterns of knowledge evolution in two-stream inputs.

Semantic Embedding: For a video segment $\mathbf{v}_t$, we employ Inception-V3 [58] model trained on ImageNet [11] to get the object scores of each frame using the output of the Softmax layer. Then, we obtain the object score vector $\mathbf{o}_t \in \mathbb{R}^O$ of the video segment through mean-pooling over all frames. At each timestep of the GCLSTM, the object score $\mathbf{o}_t$ and the preliminary predicted activity score $\hat{\mathbf{y}}^v$ (or $\hat{\mathbf{y}}^s$) is used as a weight vector to multiply with the initial graph input $\mathbf{X}^{cls}$. Specifically, the graph input $\mathbf{X}_t^v \in \mathbb{R}^{(C+O) \times k}$ of the video modality and the $\mathbf{X}_t^s \in \mathbb{R}^{(C+O) \times k}$ of the sensor signal modality at each timestep are computed as:

$$\mathbf{X}_t^v = \mathrm{diag}([\hat{\mathbf{y}}^v; \mathbf{o}_t])\mathbf{X}^{cls}, \quad \mathbf{X}_t^s = \mathrm{diag}([\hat{\mathbf{y}}^s; \mathbf{o}_t])\mathbf{X}^{cls}, \tag{6}$$

where $[;]$ denotes the concatenation operation of two vectors.

Multimodal Attention: To exploits the complementary and redundancy of two data modalities at each timestep t, a co-attention [33] operator is performed on $\mathbf{X}_t^v$ and $\mathbf{X}_t^s$ as follows:

$$\begin{aligned}
\mathbf{M}_t &= \tanh\left(\mathbf{X}_t^s \mathbf{W}_{ca}^b \mathbf{X}_t^{v\mathrm{T}}\right),\\
\mathbf{A}_t^v &= \tanh\left(\mathbf{W}_{ca}^v \mathbf{X}_t^{v\mathrm{T}} + \left(\mathbf{W}_{ca}^s \mathbf{X}_t^{s\mathrm{T}}\right)\mathbf{M}_t\right),\\
\mathbf{A}_t^s &= \tanh\left(\mathbf{W}_{ca}^s \mathbf{X}_t^{s\mathrm{T}} + \left(\mathbf{W}_{ca}^v \mathbf{X}_t^{v\mathrm{T}}\right)\mathbf{M}_t{}^{\mathrm{T}}\right),\\
\mathbf{a}_t^v &= \mathrm{Softmax}\left(\mathbf{w}_{ca}^v{}^{\mathrm{T}} \mathbf{A}_t^v\right),\\
\mathbf{a}_t^s &= \mathrm{Softmax}\left(\mathbf{w}_{ca}^s{}^{\mathrm{T}} \mathbf{A}_t^s\right),
\end{aligned} \tag{7}$$

where $\mathbf{a}_t^v, \mathbf{a}_t^s \in \mathbb{R}^{C+O}$ are normalized attention weights of the concept features in $\mathbf{X}_t^v$ and $\mathbf{X}_t^s$, respectively. $\mathbf{M}_t \in \mathbb{R}^{(C+O) \times (C+O)}$ is affinity matrix of $\mathbf{X}_t^v$ and $\mathbf{X}_t^s$, which is used for transforming sensor signal attention space to video attention space (vice versa for $\mathbf{M}_t{}^{\mathrm{T}}$). $\mathbf{W}_{ca}^b \in \mathbb{R}^{k \times k}$, $\mathbf{W}_{ca}^v$, $\mathbf{W}_{ca}^s \in \mathbb{R}^{h_c \times k}$, and $\mathbf{w}_{ca}^v, \mathbf{w}_{ca}^s \in \mathbb{R}^{h_c}$ are trainable parameters. h_c is the dimension of video attention space and sensor signal attention space. Based on the above attention scores, we can get weighted representation $\hat{\mathbf{X}}_t^v$ and $\hat{\mathbf{X}}_t^s$ of two data modalities as follows:

$$\hat{\mathbf{X}}_t^v = \operatorname{diag}\left(\mathbf{a}_t^v\right)\mathbf{X}_t^v, \quad \hat{\mathbf{X}}_t^s = \operatorname{diag}\left(\mathbf{a}_t^s\right)\mathbf{X}_t^s. \tag{8}$$

In the training phase, the adopted bottom-up attention will be constrained by the relation-aware guidance from the following dual-GCLSTM module.

Temporal Dynamic Modeling: We use graph convolutional LSTM (GCLSTM) in our framework to model the temporal dynamic patterns of knowledge evolution based on the graph inputs $\{\hat{\mathbf{X}}_t^v\}_{t=1}^T$ and $\{\hat{\mathbf{X}}_t^s\}_{t=1}^T$.

Since the video stream and sensor signal stream are temporally aligned, we adopt a dual-architecture with different parameters. For simplicity, we use $\hat{\mathbf{X}}_t$ to represent the graph input $\hat{\mathbf{X}}_t^v$ (or $\hat{\mathbf{X}}_t^s$). Next, we will briefly introduce the implementation details of one-branch of the dual-GCLSTM.

Given the input node features $\hat{\mathbf{X}}_t \in \mathbb{R}^{m\times k}$, hidden node states $\mathbf{H}_{t-1}$, and memory cell states $\mathbf{C}_{t-1}$ from last timestep, the updating units of GCLSTM in timestep t are calculated as follows:

$$\begin{aligned}
\mathbf{G}_t &= \tanh(\mathbf{W}_g *_{\mathcal{G}} \hat{\mathbf{X}}_t + \mathbf{R}_g *_{\mathcal{G}} \mathbf{H}_{t-1}),\\
\mathbf{I}_t &= \sigma(\mathbf{W}_i *_{\mathcal{G}} \hat{\mathbf{X}}_t + \mathbf{R}_i *_{\mathcal{G}} \mathbf{H}_{t-1}),\\
\mathbf{F}_t &= \sigma(\mathbf{W}_f *_{\mathcal{G}} \hat{\mathbf{X}}_t + \mathbf{R}_f *_{\mathcal{G}} \mathbf{H}_{t-1}),\\
\mathbf{C}_t &= \mathbf{G}_t^v \odot \mathbf{I}_t^v + \mathbf{C}_{t-1} \odot \mathbf{F}_t,\\
\mathbf{O}_t &= \sigma(\mathbf{W}_o *_{\mathcal{G}} \hat{\mathbf{X}}_t + \mathbf{R}_o *_{\mathcal{G}} \mathbf{H}_{t-1}),\\
\mathbf{H}_t &= \tanh(\mathbf{C}_t) \odot \mathbf{O}_t,
\end{aligned} \tag{9}$$

where $\mathbf{W}*_{\mathcal{G}}$ is the filter operation defined in Equation (5). $\mathbf{G}_t \in \mathbb{R}^{m\times h}$ is the cell input matrix. $\mathbf{I}_t$, $\mathbf{F}_t$, and $\mathbf{O}_t \in \mathbb{R}^{m\times k}$ are input gate, forget gate, and output gate matrixes, respectively. $\sigma(\cdot)$ is Sigmoid activation function and $\odot$ is element-wise product operation. All the gate parameters $\mathbf{W} \in \mathbb{R}^{h\times k}$ and $\mathbf{R} \in \mathbb{R}^{h\times h}$ are trainable. It is worth noting that k is the dimension of input node features and h is the dimension of hidden state features. By using the GCLSTM unit described in Equation (9), the number of model parameters is independent of the number of nodes in the graph $\mathcal{G}$.

We calculate the final feature matrix $\mathbf{H}_T^v$ and $\mathbf{H}_T^s$ at timestep T from both video and sensor signal data via GCLSTM defined in Equation (9) as follows:

$$\mathbf{H}_t^v = \mathrm{GCLSTM}_v\left(\hat{\mathbf{X}}_t^v, \mathbf{H}_{t-1}^v\right), \quad \mathbf{H}_t^s = \mathrm{GCLSTM}_s\left(\hat{\mathbf{X}}_t^s, \mathbf{H}_{t-1}^s\right), \tag{10}$$

where $\mathbf{H}_t^v \in \mathbb{R}^{(C+O)\times h}$ and $\mathbf{H}_t^s \in \mathbb{R}^{(C+O)\times h}$. Then, we fuse the final augmented features of videos and sensor signals by element-wise product operation as follows:

$$\mathbf{H}_T = \mathbf{H}_T^v \odot \mathbf{H}_T^s. \tag{11}$$

3.3.2 *Static Knowledge-aware Classifier.* For the activity classifiers, we also want to exploit the instance-agnostic concept relationships in graph $\mathcal{G}$, i.e., we use GCN to synthetically consider the semantic representations of activities and objects, and propagate information via the edge connections. Specifically, we use the initial node features $\mathbf{X}^{cls} \in \mathbb{R}^{(C+O)\times k}$ defined in Section 3.3 as input. Activity classifiers $\mathcal{W}^{cls}$ is produced via a two-layer GCN with following operations:

$$\mathcal{W}^{cls} = \phi\Big(\mathbf{W}_2 *_{\mathcal{G}} \phi\Big(\mathbf{W}_1 *_{\mathcal{G}} \mathbf{X}^{cls}\Big)\Big), \tag{12}$$

where ϕ is a nonlinear activation function. $\mathbf{W}_1 \in \mathbb{R}^{k\times h'}$ and $\mathbf{W}_2 \in \mathbb{R}^{h'\times h}$ are trainable parameters, h' and h are the feature dimensions of the two GCN layers. The output $\mathcal{W}^{cls} \in \mathbb{R}^{(C+O)\times h}$ is a matrix where each row can be used as a classifier.

In our setting, $\mathcal{W}^{cls}_{1:C}$ corresponds to activity classifiers and $\mathcal{W}^{cls}_{C+1:C+O}$ corresponds to object features. The proposed architecture collectively models all the relations among activities and objects. The classifiers $\mathcal{W}^{cls}_{1:C}$ learned from the explicit knowledge relations will be applied to predict activity probabilities. $\mathcal{W}^{cls}_{C+1:C+O}$, which will not be explicitly used for activity recognition, serves as a bridge between learned classifiers, and produced multimodal features. After obtaining the knowledge-aware classifiers $\mathcal{W}^{cls}_{1:C}$ and the semantic feature $\mathbf{H}_T$, we calculate probability of each activity as follows:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathrm{Softmax}(\mathbf{q}), \\ q_i &= \mathcal{W}^{cls}_i \sum_{o \in \mathcal{N}(i)} \mathbf{H}_{T,o}{}^{\mathrm{T}}, \end{aligned} \tag{13}$$

where q_i, the ith element of $\mathbf{q}$, is the classification score of the ith activity. $\mathcal{W}^{cls}_i \in \mathbb{R}^{1\times h}$ is the ith activity classifier. $\mathbf{H}_{T,o} \in \mathbb{R}^{1\times h}$ is the oth row of $\mathbf{H}_T$, which indicates the output semantic feature of the oth node in the graph $\mathcal{G}$. $\mathcal{N}(i)$ is the index set of one-hop neighbors of the ith node in the graph $\mathcal{G}$, which means that we focus strongly on related visual objects for specific activity prediction. In fact, using neighbors can avoid some interference visual objects in activity classification.

3.4 Joint Learning and Inference

Suppose that we have a training set $\mathcal{D} = \{\mathcal{V}_n, \mathcal{S}_n, \mathbf{y}_n\}_{n=1}^{N}$ containing N videos, their associated sensor signals, and activity labels. Since the construction of knowledge-driven prediction module relies on the single-modality predictors, we train all the predictors together with following losses.

For the motion predictor, cross-entropy loss is typically used over all data samples to constrain the model parameters:

$$L_s = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} \mathbf{y}_{n,i} \log\left(\hat{\mathbf{y}}^{s}_{n,i}\right). \tag{14}$$

Similarly, the vision predictor can be optimized as follows:

$$L_v = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} \mathbf{y}_{n,i} \log\left(\hat{\mathbf{y}}^{v}_{n,i}\right). \tag{15}$$

And the knowledge-aware predictor can also be optimized as follows:

$$L_k = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} \mathbf{y}_{n,i} \log(\hat{\mathbf{y}}_{n,i}). \tag{16}$$

In above formulations, $\mathbf{y}_{n,i}$ is the ith value (0 or 1) of the ground-truth label $\mathbf{y}_n$. The $\hat{\mathbf{y}}^{s}_{n,i}$, $\hat{\mathbf{y}}^{v}_{n,i}$, and $\hat{\mathbf{y}}_{n,i}$ are the probabilities of the ith activity calculated by the motion predictor in Equation (2), the vision predictor in Equation (4), and the knowledge-aware classifier in Equation (13), respectively.

We comprehensively train all the predictors together by minimizing the following loss function:

$$L = \alpha\left(\frac{L_s + L_v}{2}\right) + (1-\alpha)L_k, \tag{17}$$

where α is the hype-parameter to balance the single-modality global prediction module and the knowledge-driven multimodal prediction module. We equalize the contribution of L_s and L_v, which denotes that motion predictor and vision predictor carry the same significance compared with the knowledge-aware predictor.

During the inference stage, we fuse the motion predictor, vision predictor, and knowledge-aware predictor together to calculate the enhanced recognition result of the i^{th} activity as follows:

$$P(i) = \beta \left(\frac{\hat{\mathbf{y}}_{s,i} + \hat{\mathbf{y}}_{v,i}}{2} \right) + (1 - \beta)\hat{\mathbf{y}}_i, \tag{18}$$

where β is used to balance the contribution from each predictor.

4 EXPERIMENTS

4.1 Datasets

4.1.1 Multimodal Data. Multimodal dataset [56] is an early publicized dataset for egocentric activity recognition. Egocentric videos and sensor signals are recorded by the *Google Glass* in a synchronized manner. It is a small-scale dataset that contains 20 life-logging activities recorded by different subjects. Each activity has 10 sequences and each sequence has a duration of 15 seconds. We use full 200 videos with three-axis accelerations and three-axis gyroscopes as sensor signals in our experiments. We split the Multimodal dataset into a training set with 140 samples and a testing set with 60 samples.

4.1.2 Stanford ECM. This new dataset is proposed by Nakamura et al. [41]. It comprises 31 hours of egocentric videos augmented with heart rate and acceleration data. There are totally 23 activity classes in the Stanford ECM. Both the videos and the acceleration signals are collected by a mobile phone, which is placed in the chest pocket of subjects. A wrist sensor provides the corresponding heart rate data. Nakamura et al. use the heart rates as self-supervised information for activity recognition. In the experiments, we only use three-axis accelerations as sensor signals. Different from the Multimodal Data, Stanford ECM is collected in a more natural way, where the videos have different time durations from 2 minutes to 51 minutes, and each video may contain multiple activities. To conduct activity recognition task on this dataset, we split each video into multiple instances so each of them has a single activity label. Finally, we get 559 instances and split them into 373 instances for training and 186 instances for testing.

4.1.3 DataEgo. Possas et al. [47] propose an egocentric multimodal dataset named DataEgo. It contains 20 activities performed in different conditions and by different subjects. Each recording has a five-minute video that contains a flow of four to six activities, and the whole dataset contains four hours of continuous activities. Besides the video, each recording also contains accelerometer and gyroscope signals. Same as the Stanford ECM, we split each video into multiple instances with individual activity labels. Finally, we get 264 instances. We split them into 176 instances for training and 88 instances for testing.

4.2 Implementation Details

4.2.1 Knowledge Graph Building. In our experiments, we use the ConceptNet-v5.7 [57] to build our knowledge graph. ConceptNet connects words and phrases of natural language with labeled edges. Following previous work [13], we employ the English sub-graph of ConceptNet, which has about 1.5M nodes. We follow the previous methods [13, 36] to build the adjacency matrix **A** over activity and object nodes by retrieving the relation weight between two concept nodes from ConceptNet. In practice, the edges of the graph are frozen during the training step as References [16, 36], since fine tuning **A** is computationally burdensome and will lose generalization capacity, as it changes the intrinsic knowledge structures of **A**. Moreover, we impose some constraints on the fully connected adjacency matrix to make it sparse. The reasons are as follows: (1) Fully connected **A** will burden the computations. (2) For each activity, we should only attend to a subset of concept nodes. To this end, we select top K nodes with the highest edge weights for each activity as follows:

$$\mathcal{N}(i) = \mathtt{topK}(\mathbf{a}_i), \tag{19}$$

where $\mathtt{topK}(\mathbf{x})$ returns the indices of the K largest values of the input vector $\mathbf{x}$, and $\mathbf{a}_i$ denotes the ith row of the adjacency matrix $\mathbf{A}$.

Here, we present the computational complexity analysis for building the knowledge graph. The relation weight between two concept nodes is retrieved by the API of ConceptNet [57]. For the C activity nodes and O object nodes, we need to do $(C + O)(C + O - 1)/2$ operations of retrieval. Since each operation of the relation retrieval takes a fixed time τ (about 1–2 milliseconds on an ordinary PC), the total time of building the graph is $(C + O)(C + O - 1)\tau/2$. Once the knowledge graph is constructed, the graph structure is fixed for all samples in training and prediction stage and will not change the computational complexity of the proposed framework.

Following Reference [37], node representations of the graph are computed by the skip-gram network of Word2Vec pre-trained on the meta-data of YFCC100M dataset [60]. The reason of choosing the YFCC100M dataset is that its meta-data have semantic context information that closely match activity recognition. The trained model produces a 500-dimensional representation for each word. In the aforementioned three activity datasets, the labels of activities and objects always contain two or above words. To represent each node in a fixed length, we average all word vectors as in Reference [37].

4.2.2 Model Details. We set the length of the video sequence or the sensor signal sequence to 10 ($T = 10$). We choose the 2,048-dimensional outputs ($d_v = 2{,}048$) of the *pool*-3 layer in Inception-V3 [58] as video frame features. The dimensions of the hidden states in the motion predictor LSTM_s and the vision predictor LSTM_v are set as $h_s = 64$ and $h_v = 256$. To recognize key objects contained in the video data, we consider the 1,000 categories in ImageNet [11]. Specifically, the object score vector is computed through the last layer in the Inception-V3 model. The dimension h_c of multimodality attention space is set to 256. The dimension h of hidden features in GCLSTM is set to 256. The output dimension of the first GCN layer is set as $h' = 512$, while the second GCN layer is set as $h'' = h = 256$ for element-wise feature classification. We use PReLU, which returns $\max(\rho x, x)$ as activation function in GCN, since it can increase the convergence speed of neural network compared with ReLU [20]. Recently, it is also popularly used in graph neural networks [38, 52]. The coefficient ρ is initialized to 0.2. α and β are set by using the grid-search approach in range [0, 1]. The final values are 0.7 for α and 0.3 for β on the MultiModal and Stanford ECM datasets, and 0.8 for α and 0.2 for β on the DataEgo dataset. We employ Adam [27] gradient descent optimizer with a learning rate of 0.001 for end-to-end training on one NVIDIA RTX GPU. The batch size is set to 32. The overall framework is implemented on TensorFlow.

4.3 Performance Comparison

4.3.1 Comparison to Other Models. We use Accuracy and AUC-PR (i.e., area under the curve of precision-recall) metrics to compare our model on the egocentric multimodal activity recognition task with recent state-of-the-art methods. **LSTM (motion)** [56] only uses the sensor signal modality to predict activities. **LRCN** [47] adopts a video-based method, where CNNs are used for vision feature extraction and RNNs for temporal dependency modeling. **ERCN** [41] uses an early fusion scheme that is combined with LSTM for activity recognition. **TFusion** [2] uses a hierarchical multimodal data fusion scheme for egocentric activity recognition, where the first layer consists of two single-modality LSTMs, and the second layer attempts to explicitly capture the temporal sequence behavior and correlations of multi-modality.

As shown in Table 1, the results of the LSTM (motion) method on three datasets show that it performs poorly, as the daily and natural environment activities are hard to be recognized via

Table 1. Performance Comparison of Accuracy (%) and AUC-PR (%) on Three Datasets

Model	Multimodal Data		Stanford ECM		DataEgo	
	Accuracy	AUC-PR	Accuracy	AUC-PR	Accuracy	AUC-PR
LSTM (motion) [56]	48.33	58.13	33.69	30.00	44.32	55.57
LRCN [47]	53.33	60.22	67.91	74.49	68.18	78.32
ERCN [41]	63.33	70.96	68.35	76.22	71.59	81.19
TFusion [2]	67.80	71.99	68.98	75.79	72.73	81.63
Ours	**71.67**	**77.24**	**73.80**	**79.94**	**76.14**	**83.78**

Table 2. Ablation Studies of the Proposed Method on Three Datasets

				Multimodal Data		Stanford ECM		DataEgo	
MP	VP	CA	KA	Accuracy	AUC-PR	Accuracy	AUC-PR	Accuracy	AUC-PR
✓			✓	60.00	61.72	36.36	35.06	53.41	62.09
	✓		✓	58.33	64.35	68.40	75.11	69.32	78.80
✓	✓		✓	61.67	73.77	67.38	75.65	71.59	81.27
✓	✓	✓	✓	**71.67**	**77.24**	**73.80**	**79.94**	**76.14**	**83.78**

*Notations: MP, VP, and KA denote the motion predictor, vision predictor, and knowledge-aware predictor, respectively. CA denotes co-attention mechanism. ✓ denotes whether the the component is employed in our model for end-to-end training.

sensor signals. The results of the LRCN method show that, compared with the sensor signal, the visual data have richer information, which is beneficial for daily activity recognition. According to the results of the ERCN and TFusion methods, fusing the data from two modalities can further improve the recognition performances. These results demonstrate that the visual content can well complement the sensor signal. However, the TFusion only has small improvements than ERCN, since the data scale of all three egocentric multimodal datasets limits the performance improvement of complicated deep model. Moreover,the feature fusion operation of multimodal data in References [2, 41] can neither model the explicit interaction of two modalities nor use the external knowledge to enhance the recognition capacity. In comparison, the proposed model performs favorably on all three datasets.

4.3.2 Ablation Studies. In this part, we evaluate the effectiveness of each component of the proposed model. Our model is composed of two key modules: the single-modality global prediction module and the knowledge-driven multimodal prediction module. For simplicity, we use the abbreviations: motion predictor (MP), vision predictor (VP), knowledge-aware predictor (KA), and co-attention mechanism (CA). As shown in Table 1 and Table 2, MP+KA achieves much better performances than LSTM (motion) [56], as MP+KA uses explicit concept relations among visual objects for activity recognition, while LSTM (motion) only uses sensor signal information. Moreover, the VP+KA, which consists of the vision predictor and the knowledge-aware predictor, also achieves better performances than LRCN [47]. It shows that external knowledge can improve recognition performances even when only visual data are used. Aforementioned ablation studies have not synthetically considered the information of two modalities. When we combine the MP, VP, and KA together, the results in Table 2 show that MP+VP+KA outperforms both MP+KA and VP+KA on MultiModal and DataEgo datasets. In contrast, the MP+VP+KA performs worse than the MP+KA in the Accuracy metric on the Stanford ECM dataset. It is mainly because, on Multimodal and DataEgo datasets, the sensor signals consist of accelerations and gyroscopes, while the

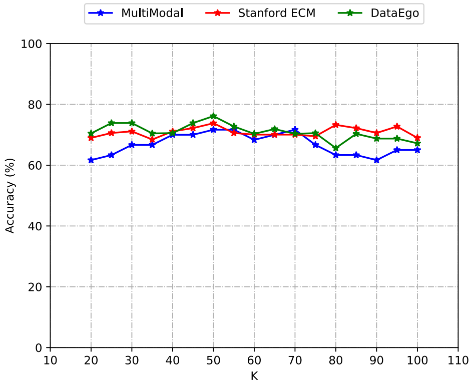


Fig. 2. Influence of the graph sparsity. The K determines how many local neighboring nodes in the graph will be connected with a given activity node. K changes from 20 to 100 with the step of 5.

Stanford ECM only has acceleration signals. Thus, fewer motion features can be extracted on the Stanford-ECM dataset and the MP+KA has much lower accuracy than the VP+KA. With the big performance gap between MP and VP, it is difficult to boost the performance by the late fusion of them. Finally, with the co-attention module, which dynamically pays different attentions to each data modality at different timesteps with the top-down relation guidance, the MP+VP+CA+KA has significant improvements on all the three datasets.

4.4 Further Remarks

4.4.1 Influence of the Graph Sparsity. The number K illustrated in Section 4.2 determines how many local neighboring nodes in the graph will be connected with a given activity node. A small number of K produces a sparse knowledge graph, which will reduce the burden of computation and memory usage, but loses much relation knowledge. As shown in Figure 2, when $K = 50$, our method obtains the highest performance compared with other values of K on all three datasets. With the increase of K, the recognition performance decreases, because considering too many object nodes will bring noises into the model. In contrast, if K is too small, insufficient relation knowledge will also influence the model performances.

4.4.2 Impact of the Semantic Embedding. In our work, besides external knowledge graph structure that is used to describe the relations between concepts, another important prior knowledge is the node's embedding feature, whose semantic information will propagate in the graph with regard to the connected edges. To illustrate the effectiveness of the adopted semantic embedding scheme, we use randomly initialized vectors to replace the pre-trained embedding vectors. As shown in Figure 3, our model with random embedding vectors performs worse than the w/o KA model, which denotes a directly concatenated fusion model without use of external knowledge. This shows that wrong semantic information is harmful for the generalization ability of the proposed model. When we set the randomly initialized node's embedding vectors trainable, it obtains better performances. However, on Multimodal dataset, our model with randomly initialized and trainable embedding vectors still performs worse than the w/o KA model. This is due to the fact that such models cannot learn robust embedding vectors from the limited context of the small-scale Multimodal dataset. On the contrary, when using the embedding vectors pre-trained from large-scale corpus, the proposed model has a significant improvement on the Multimodal dataset and also obtains best performances on the Stanford ECM and DataEgo datasets. To this end, we can

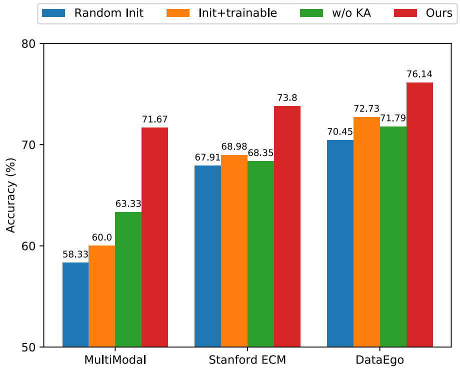


Fig. 3. The impact of different node embedding features of the graph. **Random Init** denotes the proposed model with random initialized embedding vectors. **Init+trainable** has random initialized and trainable embedding vectors. **w/o KA** is a directly concatenated fusion model without knowledge-aware predictor.

Table 3. Recognition Accuracies (%) of Individual Predictors in Proposed Model on Three Datasets

	Multimodal Data	Stanford ECM	DataEgo
LSTM (motion) [56]	48.33	33.69	44.32
LRCN [47]	53.33	68.91	68.18
MP	53.33	34.76	45.45
VP	63.33	68.98	70.45
KA	70.00	73.26	72.73
Ours	**71.67**	**73.80**	**76.14**

*Notations: The MP, VP, and KA, respectively, denote the jointly trained motion predictor, vision predictor, and knowledge-aware predictor in the proposed model.

conclude that using the embedding vectors pre-trained on external large-scale corpus is helpful for knowledge propagation in our task, especially on the small-scale dataset.

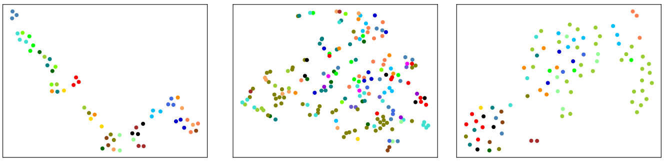
4.4.3 Performances of the Individual Predictors. We employ a late fusion scheme for knowledge-driven activity recognition in the prediction phase, as shown in Equation (18). In Table 3, we show the performance of individual predictors in the proposed model. Note that this setting is different from the ablation study in Section 4.3.2, where the proposed model is trained in an end-to-end from with/without specific modules. But here all three predictors are trained individually and we just employ late fusion scheme to fuse the results of them. As shown, the fused prediction model achieves further improvements on all three datasets than the individual knowledge-aware predictor. Moreover, the motion predictor and vision predictor in our model also have better performances than the LSTM(motion) and LRCN, respectively. It shows that the joint training scheme also has positive impact on the capacity of the single-modality based model, because the external knowledge information has been propagated back to both single-modality predictors during the training process.

4.4.4 Performances on Few-shot Classes. The Stanford ECM is an extremely unbalanced dataset with 23 activity classes, where some classes only have 2 samples and some others have above 300 samples. To explore the performance of the proposed method on few-shot classes, we show the recognition performance on classes that have less than 25 samples in Table 4. As shown, our

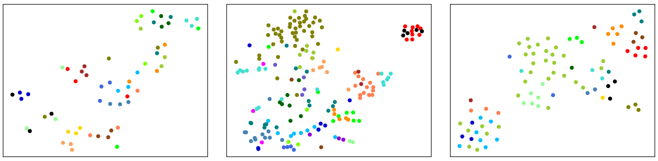
Table 4. Recognition Accuracies (%) of Few-shot Classes on the Stanford ECM Dataset

	<10	<15	<20	<25	Over All
ERCN [41]	41.67	49.09	51.14	53.54	68.35
TFusion [2]	41.67	43.64	54.55	54.55	68.98
Ours	**58.33**	**61.82**	**62.50**	**62.63**	**73.80**
Improvement	**+16.66**	**+12.73**	**+7.95**	**+8.08**	**+4.82**

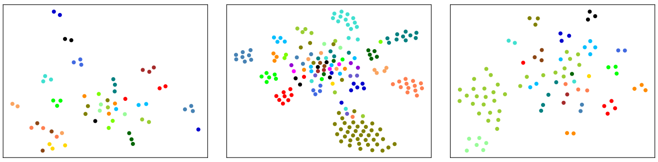
The performances of classes with $< \{10, 15, 20, 25\}$ training samples are illustrated. Over all denotes all classes in dataset.



(a) Multimodal-Sensor (b) Stanford ECM-Sensor (c) DataEgo-Sensor



(d) Multimodal-Video (e) Stanford ECM-Video (f) DataEgo-Video



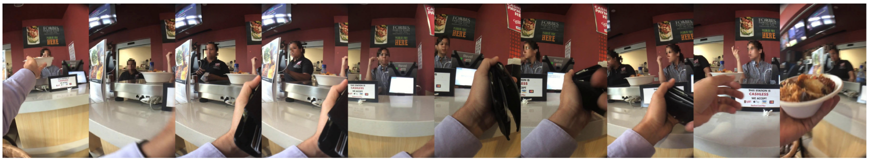
(g) Multimodal-Fused (h) Stanford ECM-Fused (i) DataEgo-Fused

Fig. 4. The t-SNE scatter plots of feature representations on test set. Representations in (a–f) are trained on single modality without external knowledge, while (g–i) fuse two modality with external knowledge.

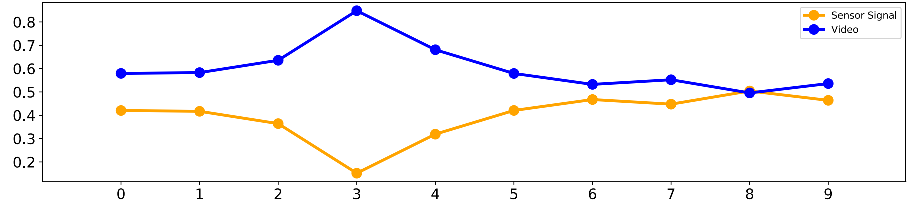
model obtains much better performances than other egocentric multimodal methods. Moreover, the largest performance improvement 16.66% is achieved on activities classes that have less than 10 samples. These results demonstrate that the proposed knowledge-aware predictor can reduce the dependence on large-scale training samples.

4.4.5 Visualization. In Figure 4, we show t-SNE visualization [35] of several important features of the proposed model on all three datasets, i.e., Multimodal, Stanford ECM, and DataEgo. As shown in Figures 4(a-f), the sensor features are harder to be discriminated into different classes than visual features. It is because the video modality has richer information than the sensor signal modality. Moreover, As shown in Figures 4(g–i), the fused features that augmented by the external knowledge become more discriminative.

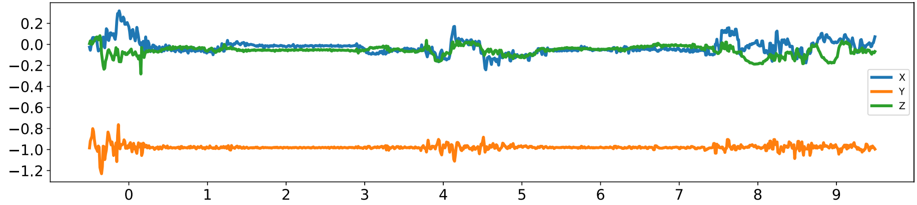
Figure 5 shows the attention scores of an example instance of *shopping* in the testing set of the Stanford ECM dataset. Specifically, Figure 5(a) shows the video stream, Figure 5(b) shows the



(a) Video Stream



(b) Time-Varying Attention Scores



(c) Sensor Signal Stream

Fig. 5. A qualitative result of *shopping* activity on the testing set of the Stanford ECM dataset. (a) and (c) show inputs of the video and sensor signal modalities. (b) shows the attention scores obtained by the proposed method at different timesteps.

time-varying attention scores computed by the proposed model, and Figure 5(c) shows the sensor signal stream. As shown, the video stream always obtains higher attention score than sensor signal modality in most of time, since the video data have more information for recognition. As shown in Figure 5(b) and Figure 5(c), when sensor signals vary drastically (i.e., in 0, 4, and 8 timesteps), this modality gets relatively higher attention scores compared with other timesteps. When the signals are flat, attention scores will gradually decrease, because the drastic motion pattern of the sensor modality provides more discriminative features for activity recognition. The above discussions show the effectiveness of the adopted multimodal co-attention scheme.

5 CONCLUSION

In this work, we present a knowledge-driven multimodal activity recognition model on the egocentric video and the sensor data. The proposed model consists of a single-modality global prediction module and a knowledge-driven multimodal prediction module. The single modality module produces the preliminary activity scores that will be augmented by a knowledge-driven module. Graph neural networks are adopted to exploit the relations among intermediate objects and activities. The proposed approach improves the performances of egocentric multimodal activity recognition on three public datasets. Moreover, the incorporated external knowledge successfully reduces the dependence on large-scale training samples. Our future work will focus on event segmentation and egocentric video summary using knowledge graph [23], which can reduce the recording and analyzing cost of daily living activities.

REFERENCES

[1] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. 2018. Object level visual reasoning in videos. In *Proceedings of the 15th European Conference on Computer Vision (ECCV'18)*. Springer, 105–121. DOI: https://doi.org/10.1007/978-3-030-01261-8_7

[2] Edgar A. Bernal, Xitong Yang, Qun Li, Jayant Kumar, Sriganesh Madhvanath, Palghat Ramesh, and Raja Bala. 2017. Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors. *IEEE Trans. Multimedia* 20, 1 (Jan. 2017), 107–118. DOI: https://doi.org/10.1109/TMM.2017.2726187

[3] Alejandro Betancourt, Pietro Morerio, Carlo S. Regazzoni, and Matthias Rauterberg. 2015. The evolution of first person vision methods: A survey. *IEEE Trans. Circ. Syst. Vid. Technol.* 25, 5 (May 2015), 744–760. DOI: https://doi.org/10.1109/TCSVT.2015.2409731

[4] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.* 46, 3 (Jan. 2014), 33. DOI: https://doi.org/10.1145/2499621

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 961–970. DOI: https://doi.org/10.1109/CVPR.2015.7298698

[6] Minjie Cai, Kris M. Kitani, and Yoichi Sato. 2016. Understanding hand-object manipulation with grasp types and object attributes. In *Proceedings of the Robotics: Science and Systems Conference*, Vol. 3.

[7] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 6299–6308. DOI: https://doi.org/10.1109/CVPR.2017.502

[8] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2017. A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools. Applic.* 76, 3 (Feb. 2017), 4405–4425. DOI: https://doi.org/10.1007/s11042-015-3177-1

[9] Yuting Chen, Joseph Wang, Yannan Bai, Gregory Castañón, and Venkatesh Saligrama. 2018. Probabilistic semantic retrieval for surveillance videos with activity graphs. *IEEE Trans. Multimedia* 21, 3 (Mar. 2018), 704–716. DOI: https://doi.org/10.1109/TMM.2018.2865860

[10] Maria Cornacchia, Koray Ozcan, Yu Zheng, and Senem Velipasalar. 2017. A survey on activity detection and classification using wearable sensors. *IEEE Sensors J.* 17, 2 (Jan. 2017), 386–403. DOI: https://doi.org/10.1109/JSEN.2016.2628346

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*. 248–255. DOI: https://doi.org/10.1109/CVPR.2009.5206848

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 2625–2634. DOI: https://doi.org/10.1109/CVPR.2015.7298878

[13] Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Chandrasekhar. 2017. Object detection meets knowledge graphs. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. 1661–1667. DOI: https://doi.org/10.24963/ijcai.2017/230

[14] Alireza Fathi, Yin Li, and James M. Rehg. 2012. Learning to recognize daily actions using gaze. In *Proceedings of the 12th European Conference on Computer Vision (ECCV'12)*. Springer, 314–327. DOI: https://doi.org/10.1007/978-3-642-33718-5_23

[15] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2018. Watch, think and attend: End-to-end video classification via dynamic knowledge evolution modeling. In *Proceedings of the 26th ACM International Conference on Multimedia (MM'18)*. ACM, 690–699. DOI: https://doi.org/10.1145/3240508.3240566

[16] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2019. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19)*, Vol. 33. DOI: https://doi.org/10.1609/aaai.v33i01.33018303

[17] Weili Guan, Xuemeng Song, Tian Gan, Junyu Lin, Xiaojun Chang, and Liqiang Nie. 2019. Cooperation learning from multiple social networks: Consistent and complementary perspectives. *IEEE Trans. Cybern.* (2019). DOI: https://doi.org/10.1109/TCYB.2019.2951207

[18] Sojeong Ha and Seungjin Choi. 2016. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'16)*. 381–388. DOI: https://doi.org/10.1109/IJCNN.2016.7727224

[19] Sojeong Ha, Jeong-Min Yun, and Seungjin Choi. 2015. Multi-modal convolutional neural networks for activity recognition. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC'15)*. 3017–3022. DOI: https://doi.org/10.1109/SMC.2015.525

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*. 1026–1034. `DOI`: https://doi.org/10.1109/ICCV.2015.123

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. `DOI`: https://doi.org/10.1162/neco.1997.9.8.1735

[22] Peng-Ju Hsieh, Yen-Liang Lin, Yu-Hsiu Chen, and Winston Hsu. 2016. Egocentric activity recognition by leveraging multiple mid-level representations. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'16)*. 1–6. `DOI`: https://doi.org/10.1109/ICME.2016.7552937

[23] Fairouz Hussein and Massimo Piccardi. 2017. V-JAUNE: A framework for joint action recognition and video summarization. *ACM Trans. Multimedia Comput. Commun. Appl.* 13, 2 (May 2017), 20. `DOI`: https://doi.org/10.1145/3063532

[24] Ahmad Babaeian Jelodar, David Paulius, and Yu Sun. 2019. Long activity video understanding using functional object-oriented network. *IEEE Trans. Multimedia* 21, 7 (July 2019), 1813–1824. `DOI`: https://doi.org/10.1109/TMM.2018.2885228

[25] Weike Jin, Zhou Zhao, Yimeng Li, Jie Li, Jun Xiao, and Yueting Zhuang. 2019. Video question answering via knowledge-based progressive spatial-temporal attention network. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 2s (Aug. 2019), 1–22. `DOI`: https://doi.org/10.1145/3321505

[26] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. 1725–1732. `DOI`: https://doi.org/10.1109/CVPR.2014.223

[27] Diederik P. Kingma and Jimmy Ba. 2013. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR'15)*. Retrieved from http://arxiv.org/abs/1412.6980.

[28] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR'17)*. Retrieved from https://openreview.net/forum?id=SJU4ayYgl.

[29] Shiro Kumano, Kazuhiro Otsuka, Ryo Ishii, and Junji Yamato. 2016. Collective first-person vision for automatic gaze analysis in multiparty conversations. *IEEE Trans. Multimedia* 19, 1 (Jan. 2016), 107–122. `DOI`: https://doi.org/10.1109/TMM.2016.2608002

[30] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*. Retrieved from https://openreview.net/forum?id=SJiHXGWAZ.

[31] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. 2017. Towards micro-video understanding by joint sequential-sparse modeling. In *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*. 970–978. `DOI`: https://doi.org/10.1145/3123266.3123341

[32] Shaopeng Liu, Robert Gao, and Patty Freedson. 2012. Computational methods for estimating energy expenditure in human physical activities. *Med. Sci. Sports Exer.* 44, 11 (2012), 2138–46. `DOI`: https://doi.org/10.1249/MSS.0b013e31825e825a

[33] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems (NeurIPS'16)*. Curran Associates, Inc., 289–297. Retrieved from https://papers.nips.cc/paper/6202-hierarchical-question-image-co-attention-for-visual-question-answering.

[34] Minghuang Ma, Haoqi Fan, and Kris M. Kitani. 2016. Going deeper into first-person activity recognition. In *Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 1894–1903. `DOI`: https://doi.org/10.1109/CVPR.2016.209

[35] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9 (Nov. 2008), 2579–2605. Retrieved from http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[36] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The more you know: Using knowledge graphs for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 20–28. `DOI`: https://doi.org/10.1109/CVPR.2017.10

[37] Pascal Mettes and Cees G. M. Snoek. 2017. Spatial-aware object embeddings for zero-shot localization and classification of actions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 4443–4452. `DOI`: https://doi.org/10.1109/ICCV.2017.476

[38] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. 2020. Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 14424–14432.

[39] Pietro Morerio, Lucio Marcenaro, and Carlo S. Regazzoni. 2013. Hand detection in first person vision. In *Proceedings of the 16th International Conference on Information Fusion (FUSION'13)*. IEEE, 1502–1507.

[40] Abdulmajid Murad and Jae-Young Pyun. 2017. Deep recurrent neural networks for human activity recognition. *Sensors* 17, 11 (Nov. 2017), 2556. DOI: https://doi.org/10.3390/s17112556

[41] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. 2017. Jointly learning energy expenditures and activities using egocentric multimodal signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 6817–6826. DOI: https://doi.org/10.1109/CVPR.2017.721

[42] Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, Francisco Florez-Revuelta, et al. 2016. Recognition of activities of daily living with egocentric vision: A review. *Sensors* 16, 1 (Jan. 2016), 72. DOI: https://doi.org/10.3390/s16010072

[43] Liqiang Nie, Xuemeng Song, and Tat-Seng Chua. 2016. Learning from multiple social networks. *Synt. Lect. Inf. Conc., Retr., Serv.* 8, 2 (2016), 1–118. DOI: https://doi.org/10.2200/S00714ED1V01Y201603ICR048

[44] Liqiang Nie, Xiang Wang, Jianglong Zhang, Xiangnan He, Hanwang Zhang, Richang Hong, and Qi Tian. 2017. Enhancing micro-video understanding by harnessing external sounds. In *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*. 1192–1200. DOI: https://doi.org/10.1145/3123266.3123313

[45] Alan V. Oppenheim. 1999. *Discrete-time Signal Processing*. Pearson Education India.

[46] Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. 2847–2854. DOI: https://doi.org/10.1109/CVPR.2012.6248010

[47] Rafael Possas, Sheila Pinto Caceres, and Fabio Ramos. 2018. Egocentric activity recognition on a budget. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 5967–5976. DOI: https://doi.org/10.1109/CVPR.2018.00625

[48] Shengsheng Qian, Tianzhu Zhang, and Changsheng Xu. 2018. Online multimodal multiexpert learning for social event tracking. *IEEE Trans. Multimedia* 20, 10 (Oct. 2018), 2733–2748. DOI: https://doi.org/10.1109/TMM.2018.2815785

[49] Fereshteh Sadeghi, Santosh K. Kumar Divvala, and Ali Farhadi. 2015. VisKE: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 1456–1464. DOI: https://doi.org/10.1109/CVPR.2015.7298752

[50] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. 2011. Learning to share visual appearance for multiclass object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. 1481–1488. DOI: https://doi.org/10.1109/CVPR.2011.5995720

[51] Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. 2018. Structured sequence modeling with graph convolutional recurrent networks. In *Proceedings of the 25th International Conference on Neural Information Processing (ICONIP'18)*. Springer, 362–373. DOI: https://doi.org/10.1007/978-3-030-04167-0_33

[52] Zhijuan Shen, Jun Cheng, Xiping Hu, and Qian Dong. 2019. Emotion recognition based on multi-view body gestures. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'19)*. IEEE, 3317–3321. DOI: https://doi.org/10.1109/ICIP.2019.8803460

[53] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 28th International Conference on Advances in Neural Information Processing Systems (NeurIPS'14)*. 568–576. Retrieved from https://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.

[54] Mohammad Soltanian and Shahrokh Ghaemmaghami. 2019. Hierarchical concept score postprocessing and concept-wise normalization in CNN-based video event recognition. *IEEE Trans. Multimedia* 21, 1 (Jan. 2019), 157–172. DOI: https://doi.org/10.1109/TMM.2018.2844101

[55] Hao Song, Xinxiao Wu, Wennan Yu, and Yunde Jia. 2018. Extracting key segments of videos for event detection by learning from web sources. *IEEE Trans. Multimedia* 20, 5 (May 2018), 1088–1100. DOI: https://doi.org/10.1109/TMM.2017.2763322

[56] Sibo Song, Vijay Chandrasekhar, Bappaditya Mandal, Liyuan Li, Joo-Hwee Lim, Giduthuri Sateesh Babu, Phyo Phyo San, and Ngai-Man Cheung. 2016. Multimodal multi-stream deep learning for egocentric activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'16)*. 378–385. DOI: https://doi.org/10.1109/CVPRW.2016.54

[57] Robert Speer and Catherine Havasi. 2013. ConceptNet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*. Springer, Berlin, 161–176. DOI: https://doi.org/10.1007/978-3-642-35085-6_6

[58] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 2818–2826. DOI: https://doi.org/10.1109/CVPR.2016.308

[59] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[60] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (Jan. 2016), 64–73. DOI: https://doi.org/10.1145/2812802

[61] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2013. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* 103, 1 (2013), 60–79. DOI: https://doi.org/10.1007/s11263-012-0594-8

[62] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recog. Lett.* 119 (Mar. 2019), 3–11. DOI: https://doi.org/10.1016/j.patrec.2018.02.010

[63] Lei Wang, Xu Zhao, Yunfei Si, Liangliang Cao, and Yuncai Liu. 2017. Context-associative hierarchical memory model for human activity recognition and prediction. *IEEE Trans. Multimedia* 19, 3 (Mar. 2017), 646–659. DOI: https://doi.org/10.1109/TMM.2016.2617079

[64] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 6857–6866. DOI: https://doi.org/10.1109/CVPR.2018.00717

[65] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Trans. Image Proc.* 29 (2019), 1–14. DOI: https://doi.org/10.1109/TIP.2019.2923608

[66] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. DeepSense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*. International World Wide Web Conferences Steering Committee, 351–360. DOI: https://doi.org/10.1145/3038912.3052577

[67] Jun Ye, Hao Hu, Guo-Jun Qi, and Kien A. Hua. 2017. A temporal order modeling approach to human action recognition from multimodal sensor data. *ACM Trans. Multimedia Comput. Commun. Appl.* 13, 2 (Mar. 2017), 14. DOI: https://doi.org/10.1145/3038917

[68] Xingliang Yuan, Xinyu Wang, Cong Wang, Jian Weng, and Kui Ren. 2016. Enabling secure and fast indexing for privacy-assured healthcare monitoring via compressive sensing. *IEEE Trans. Multimedia* 18, 10 (Oct. 2016), 2002–2014. DOI: https://doi.org/10.1109/TMM.2016.2602758

[69] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. 2018. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'18)*. 5628–5635. DOI: https://doi.org/10.1109/ICRA.2018.8461249

[70] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. 2018. Towards universal representation for unseen action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 9436–9445. DOI: https://doi.org/10.1109/CVPR.2018.00983

Received January 2020; revised June 2020; accepted July 2020