UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Coleta e Classificação de Dados de Atividade Humana

Felipe Aparecido Garcia

São Carlos – SP

Coleta e Classificação de Dados de Atividade Humana

Felipe Aparecido Garcia

Orientadora: **Profa. Dra. Roseli Aparecida Francelin Romero**

Monografia final de conclusão de curso apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação.
Área de Concentração: Sistemas Computacionais

USP – São Carlos
Novembro de 2019

Garcia, Felipe Aparecido
Coleta e Classificação de Dados de Atividade Humana / Felipe Aparecido Garcia. - São Carlos - SP, 2019.
50 p.; 29,7 cm.

Orientadora: Roseli Aparecida Francelin Romero.
Monografia (Graduação) - Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos - SP, 2019.

1. Dataset. 2. Aprendizado de Máquina. 3. Aprendizado Profundo. 4. HAR. 5. LSTM. 6. TCN. I. Romero, Roseli Aparecida Francelin. II. Instituto de Ciências Matemáticas e de Computação (ICMC/USP). III. Título.

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.
Em especial, aos pesquisadores do Instituto de Ciências Matemáticas e de Computação (ICMC).

AGRADECIMENTOS

Os agradecimentos principais são direcionados aos voluntários que tornaram a realização deste trabalho possível. Agradecimentos especiais são direcionados aqueles que me orientaram durante minha trajetória, como Roseli Aparecida Francelin Romero e Caetano Mazzoni Ranieri, assim como todos os professores da EESC e ICMC.

“As invenções são, sobretudo,
o resultado de um trabalho de teimoso.”
(Santos Dumont)

RESUMO

GARCIA, A. F.. **Coleta e Classificação de Dados de Atividade Humana**. 2019. 50 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Reconhecimento de atividade humana consiste em classificar atividades humanas com bases em dados de sensores, onde dados de diferentes fontes podem ser combinados para obter resultados mais precisos. Neste trabalho, foi criada uma base de dados multimodal de atividade humana, chamada de *Video-Inertial Human Activity Dataset* (VIHAD), com 10 atividades cotidianas realizadas por 6 voluntários. O banco de dados foi construído com dados de video RGB e de profundidade e com duas unidades inerciais, posicionadas, respectivamente, no pulso do braço dominante e no bolso frontal superior da perna direita do voluntário. Ademais, classificou-se a base de dados pública OPPORTUNITY e a base criada nesta obra, com técnicas de aprendizado profundo, usando arquiteturas baseadas em redes neurais convolucionais, capazes de extrair características relevantes dos dados, e também redes neurais recorrentes e redes temporais convolucionais, capazes modelar a dependência temporal intrínseca dos dados trabalhados. Com tais abordagens, foi possível montar uma base com 10 minutos de dados e classificar as bases trabalhadas, onde a combinação de dados de diferentes fontes gerou melhorias nos resultados.

Palavras-chave: Dataset, Aprendizado de Máquina, Aprendizado Profundo, HAR, LSTM, TCN.

ABSTRACT

GARCIA, A. F.. **Coleta e Classificação de Dados de Atividade Humana**. 2019. 50 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Human activity recognition consists of classifying human activities based on sensor data, where data from different sources can be combined to obtain more accurate results. In this work, a multimodal dataset of human activity was created, called Video-Inertial Human Activity Dataset (VIHAD), with 10 daily activities performed by 6 volunteers. The database was constructed with RGB and depth video data and two inertial units, positioned, respectively, on the wrist of the dominant arm and the upper front pocket of the right leg of the volunteer. In addition, the public database OPPORTUNITY and the database created in this work were classified using deep learning approaches using architectures based on convolutional neural networks, capable of extracting relevant features from the data, as well as recurrent neural networks and temporal convolutional networks, able to model sequences. With these approaches, it was possible to build a database with 10 minutes of data and classify the bases worked, where the combination of data from different sources generated improvements in the results.

Key-words: Dataset, Machine Learning, Deep Learning, HAR, LSTM, TCN.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE CÓDIGOS-FONTE

LISTA DE ABREVIATURAS E SIGLAS

CNN ..... *Convolutional Neural Networks*
CRC ..... *Collaborative Representation Classifier*
DNN ..... *Deep Neural Network*s
FN ..... Falsos Negativos
FP ..... Falsos Positivos
HAR ..... *Human Activity Recognition*
IMC ..... Índice de Massa Corporal
LSTM .... *Long Short-Term Memory*
MMR .... *MetaMotionR*
PCA ..... *Principal Component Analyzis*
ReLU .... *Rectified Linear Unit*
RNN ..... *Recurrent Neural Networks*
SVM ..... *Support Vector Machines*
SVM ..... *Support Vector Machines*
TCN ..... *Temporal Convolutional Networks*
TCN-FCN *Temporal Convolutional Network-Fully Convolutional Network*
VIHAD .. *Video-Inertial Human Activity Dataset*
VP ..... Verdadeiros Positivos

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Motivação e Contextualização

O reconhecimento de atividade humana, do inglês, *Human Activity Recognition* (HAR), é um campo amplo de pesquisa que possui grande importância por sua variabilidade de aplicações, como robótica, monitoramento, saúde, etc. (CHEN; JAFARI; KEHTARNAVAZ, 2015). Trata-se do conjunto de tarefas relacionadas a classificar interações entre indivíduos e ambiente com interpretação semântica à partir de dados de sensores. Um fator importante na evolução desse campo de pesquisa, é a evolução das tecnologias sensitivas, como sensores inerciais (vestíveis) e câmeras.

Os sensores inerciais se tornaram cada vez menores e mais presentes em equipamentos usados no cotiano, como *smartphones* e *smartwatches*. Por estarem presente em tais equipamentos, esses sensores possibilitam a coleta de informações de maneira não invasiva ao usuário, possuindo as seguintes vantagens: são pouco afetados pelo ambiente, podem obter melhor precisão da classificação de movimentos ao se utilizar múltiplos sensores distribuídos pelo corpo e garantem privacidade na coleta de informação (YANG *et al.*, 2015). Essas características tornam esses sensores preferíveis a câmeras de vídeo, que, a depender das circunstâncias, podem levar a questões mais severas quanto à privacidade.

Apesar de poderem representar uma rica fonte de informação sobre o movimento humano, sendo de grande interesse para o HAR, câmeras RGB possuem a desvantagem de serem afetadas pelo ambiente. Além disso, sozinhas, não fornecem diretamente dados de profundidade, limitando-se a representar a projeção das imagens em um plano. Por isso, câmeras de profundidade também são dignas de nota, uma vez que representam informações de profundidade da imagem com relativa invariância em relação à iluminação do ambiente (CHEN; JAFARI; KEHTARNAVAZ, 2015), ainda que não representem informações importantes como textura, brilho e cor. Dessa forma, tais câmeras se complementam, provendo dados úteis para o HAR.

Em geral, câmeras e sensores inerciais foram usados separadamente para classificação de atividade humana. Dessa forma, seu uso conjunto possui poucas referências na literatura (CHEN; JAFARI; KEHTARNAVAZ, 2015), mas, como mostrado em (CHEN; JAFARI; KEHTARNAVAZ, 2014), tal fusão de sensores pode melhorar a classificação dos modelos, pois pode agregar informações relevantes sobre os movimentos realizados.

Além disso, para classificar esses dados (tanto de sensores vestíveis como de câmeras), podem-se aplicar diferentes técnicas de aprendizado de máquina, como redes neurais, *K-Nearest Neighbors*, Modelos Ocultos de Markov, *Support Vector Machines* (SVM), etc. Dentre tais técnicas, as redes neurais possuem uma importância crescente na área, graças a artifícios como o *Deep Learning* e arquiteturas de redes neurais capazes de modelar a dependência temporal intrínseca desses dados.

Para lidar com a dependência temporal, Redes Neurais Recorrentes, *Recurrent Neural Networks* (RNN) são comumente utilizadas (HAYKIN *et al.*, 2009). Entretanto, Redes Convolucionais Temporais, *Temporal Convolutional Networks* (TCN) (LEA *et al.*, 2017), também são capazes de modelar tais dependências, sem o uso de conexões recorrentes. Ademais, extrair características relevantes dos dados pode gerar melhorias significativas nas classificações e Redes Neurais Convolucionais, *Convolutional Neural Networks* (CNN) podem realizar tal tarefa (KARIM *et al.*, 2018), podendo melhorar a performance de outros classificadores. Assim, arquiteturas que utilizam CNN para extrair características e RNN ou TCN para classificá-las podem obter bons resultados no reconhecimento de atividade humana (GARCIA; RANIERI; ROMERO, 2019).

Na bibliografia, existem diversas bases de dados públicas com dados de atividades humanas, com informações inerciais, de vídeo e multimodais (mais de uma categoria de sensor). Como exemplo, pode-se citar a OPPORTUNITY (CHAVARRIAGA *et al.*, 2013), PAMAP (REISS; STRICKER, 2012), UCF101 (SOOMRO; ZAMIR; SHAH, 2012) e UTD-MHAD (CHEN; JAFARI; KEHTARNAVAZ, 2015). Dentre tais bancos de dados, grande parte é constituída por apenas dados inerciais ou de vídeo, com poucos sendo multimodais. Adicionalmente, as obras de (CHEN; JAFARI; KEHTARNAVAZ, 2015) e (SONG *et al.*, 2016) mostraram que dados multimodais podem melhorar os resultados ao se classificar atividade humana.

1.2 Objetivos

Neste trabalho, pretende-se construir uma base de dados multimodal, com informações de sensores vestíveis inerciais, câmera RGB e de profundidade. Tais dados serão coletados de voluntários realizando tarefas cotidianas, em diferentes ambientes e com diferentes posicionamentos de sensores. A base construída será chamada de *Video-Inertial Human Activity Dataset* (VIHAD), para facilitar futuras referências.

Adicionalmente, a base construída será classificada a partir dos dados inerciais, com diferentes modelos de aprendizado de máquina, medindo a performance de cada modelo e comparando-os de acordo com tais métricas. As classificações serão baseadas em arquiteturas da literatura usando RNN, CNN e TCN, onde o principal objetivo da classificação é validar a possível melhoria alcançada na classificação ao se usar características extraídas pela CNN e, também, comparar RNN e TCN na modelagem de dependências temporais. Por fim, pretende-

se aplicar o modelo com a melhor performance em um ambiente de casa inteligente, onde o reconhecimento da atividade humana pode auxiliar a tomada de decisões e melhorar a interação homem-máquina.

1.3 Organização

O trabalho foi organizado da seguinte forma: o Capítulo 2 descreve alguns conceitos teóricos relevantes para melhor compreensão deste texto, o Capítulo 3 resume algumas obras similares à esta, que colaboraram para sua construção como um todo, o Capítulo 4 descreve a metologia, técnicas e tecnologias usadas para resolver o problema. Já o Capítulo 5 recapitula a resolução do problema e exibe os resultados obtidos, assim como as dificuldades encontradas. Por fim, o Capítulo 6 discute e conclui a obra.

Capítulo 2

REFERENCIAL TEÓRICO

Antes de entrar em detalhes sobre o projeto, é importante explicar alguns conceitos teóricos necessários para sua compreensão, como Redes Neurais Artificiais, Recorrentes e Convolucionais. Nesta seção, tais teorias serão sucintamente explicadas.

2.1 Neurônio Artificial

Um neurônio artificial (ou nó) é uma representação matemática de um neurônio humano. O modelo linear de neurônio possui entradas, denotadas por x_i, pesos w_i, que são usados para pondera-las, e um termo *bias*, b, que é somado às entradas já multiplicadas pelos pesos. Com as entradas ponderadas e o *bias* somados, aplica-se ao resultado uma função de ativação f, que determinará a resposta do neurônio (HAYKIN *et al.*, 2009). A Figura 1 mostra uma representação gráfica dessa estrutura.

2.2 Rede Neural Multicamadas

Redes Neurais Multicamadas são compostas por uma camada de entrada, uma camada de saída e uma ou mais camadas intermediárias (ocultas). Cada neurônio de cada camada pode

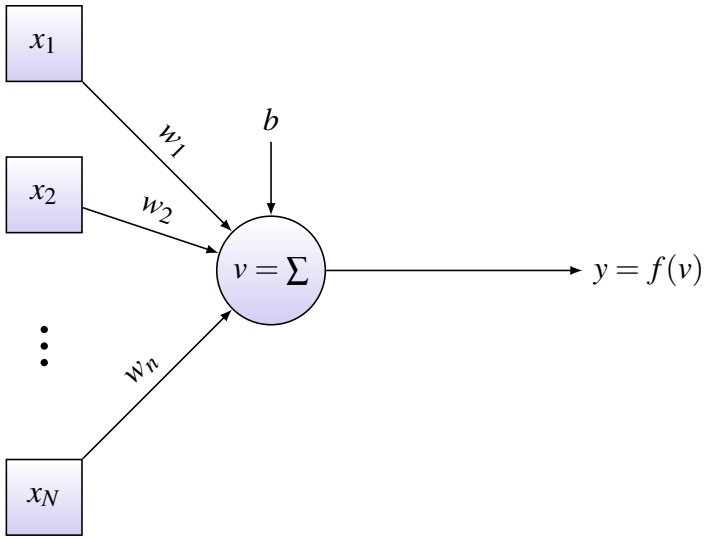


Figura 1 – Modelo gráfico de um neurônio artificial.

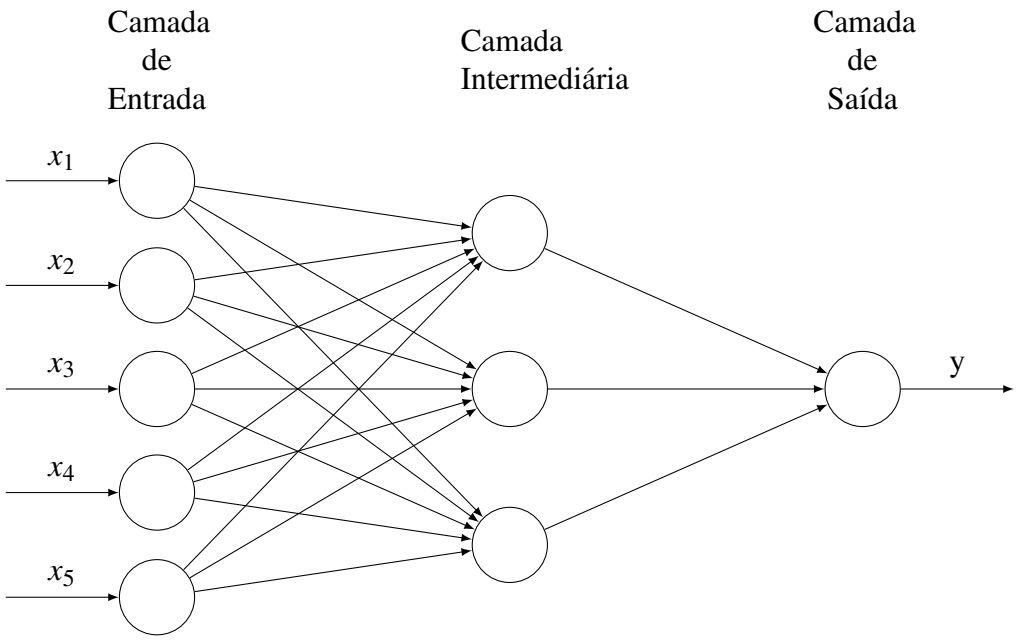


Figura 2 – Exemplo de Rede Neural Multicamadas.

receber como entradas as saídas de todos os nós da camada anterior (*Multi-Layer Perceptron* - MLP), processando-as e enviando as respostas para a próxima camada e assim sucessivamente, até a camada de saída. Assim, o sinal flui da camada de entrada até a camada de saída, que possui um número de neurônios de acordo com o número de classes que se deseja classificar e tem como resposta o resultado da função de ativação escolhida (HAYKIN *et al.*, 2009). A Figura 2 mostra um exemplo desta estrutura, contendo apenas uma camada intermediária

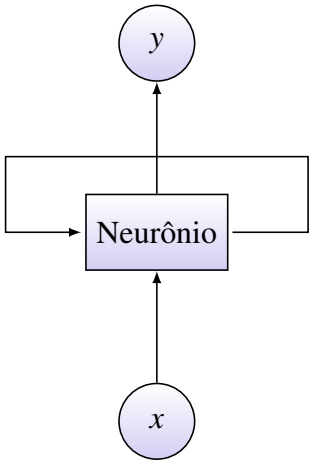


Figura 3 – Neurônio recorrente.

2.3 Redes Neurais Recorrentes

Em uma Rede Neural Recorrente (RNN), as saídas da rede são conectadas à entrada, formando um *loop* que permite a esta estrutura lidar com dados passados, ou seja, isso proporciona à rede uma memória (HOPFIELD, 1982). Essa característica das RNNs torna-as adequadas para lidar com dados temporais, pois estes, em geral, estão inseridos em contextos que devem ser levados em conta no reconhecimento. A Figura 3 exibe um exemplo de neurônio destas redes.

2.4 Redes Neurais Convolucionais

Redes Neurais Convolucionais (CNN) são redes neurais em que, ao invés da multiplicação de matriz comumente usada (pesos multiplicados pelas entradas), aplica-se a operação de convolução, dada na Equação 2.1 (GOODFELLOW; BENGIO; COURVILLE, 2016). Tais estruturas são, geralmente, seguidas por uma operação de *pooling*, que visam reduzir a quantidade de dados, e uma ou mais camadas totalmente conectadas.

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau \tag{2.1}$$

2.5 Redes Temporais Convolucionais

As Redes Temporais Convolucionais (TCN), são redes capazes de modelar dependências temporais, em um contexto limitado, usando convoluções causais dilatadas, onde o termo causal indica que só dependem de informações passadas e a dilatação é dada pelo fato do filtro pular

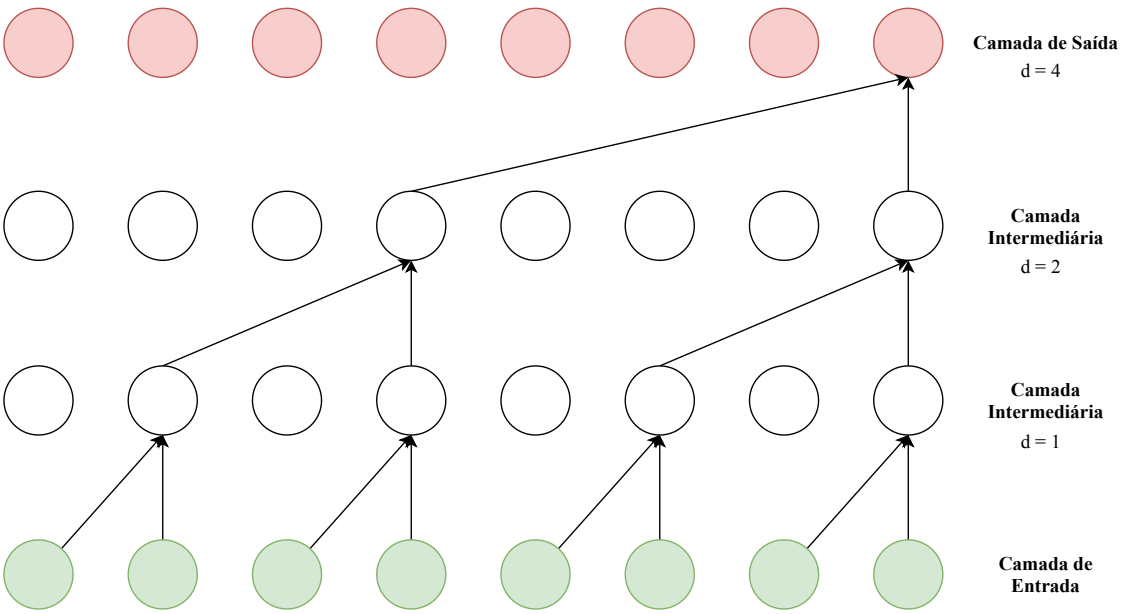


Figura 4 – Convolução causal dilatada.

valores de entrada, em um certo intervalo, permitindo que o mesmo seja aplicado à uma área maior que seu tamanho (OORD *et al.*, 2016).

A dilatação possibilita a implementação de grandes filtros receptivos, com o uso de poucas camadas, onde o intervalo de dilatação d, que indica o número de conexões omitidas, é um valor inteiro que, usualmente, é dobrado a cada camada. A Figura 4 exibe uma representação da convolução causal dilatada. Em suma, a convolução causal dilatada é dada pela Equação 2.2, onde k é o tamanho do filtro e o termo $s - d \cdot i$ indica a direção da convolução no passado.

$$F(s) = (x * f_d)(s) = \sum_{i=0}^{k-1} f(i) x_{s-d \cdot i} \tag{2.2}$$

2.6 Acurácia

Acurácia é uma medida que indica a precisão do modelo. Ela representa a porcentagem de Verdadeiros Positivos (VP) em relação à soma dos VP com os Falsos Positivos (FP), como mostra a Equação 2.3 (METZ, 1978).

$$Acuracia = \frac{VP}{VP + FP} \tag{2.3}$$

2.7 Recuo

O recuo indica a sensibilidade do modelo, sendo a porcentagem de VP em relação a soma dos VP com os Falsos Negativos (FN), como exibe a Equação 2.4 (METZ, 1978).

$$Recuo = \frac{VP}{VP + FN} \tag{2.4}$$

2.8 F1-Score

O F1-Score é uma medida que considera tanto a acurácia quanto o recuo, sendo, basicamente, a média harmônica entre as duas medições. Tal medida é apresentada na Equação 2.5 (METZ, 1978).

$$F1 - Score = 2 \frac{Acuracia \cdot Recuo}{Acuracia + Recuo} \tag{2.5}$$

Capítulo 3

TRABALHOS RELACIONADOS

Com o avanço de técnicas como o *Deep Learning*, tarefas como HAR se tornaram mais realizáveis, com seus padrões característicos sendo, em geral, mapeáveis pelas redes neurais artificiais. Todavia, essa técnica se beneficia de grandes volumes de dados, tornando interessante gerar o máximo de dados possíveis ao se construir uma base de dados com atividades humanas. Algumas bases de dados se destacam na área em questão, sendo algumas compostas exclusivamente por dados inerciais, outras por câmeras e algumas são multimodais (contém mais de um tipo de dado). Neste capítulo serão descritas algumas bases de dados da literatura, assim como técnicas que obtiveram sucesso em classificá-las.

Em (REISS; STRICKER, 2012), propôs-se o banco de dados público PAMAP, composto por informações de atividades cotidianas realizadas por voluntários, enquanto usavam um monitor cardíaco e três unidades inerciais, constituídas por dois acelerômetros, um giroscópio e um magnetômetro, sendo uma unidade posicionada no peito, uma no pulso do braço dominante e a última na lateral do quadril. Todas as unidades realizaram a coleta a uma taxa de 100 Hz. A coleta foi realizada com 9 voluntários, com 27,22 $\pm$ 3,31 anos e um Índice de Massa Corporal (IMC) de 25,11 $\pm$ 2,62 kg/m^2 realizando um protocolo de 12 atividades (deitar, sentar, andar, correr, andar de bicicleta, caminhada nórdica, passar roupa, usar aspirador de pó, pular corda, subir e descer escadas) e 6 opcionais (assistir TV, trabalhar no computador, dirigir carro, lavar roupa, limpar a casa e jogar futebol).

Já na obra de (CHAVARRIAGA *et al.*, 2013), criou-se uma base de dados pública criada usando 12 voluntários e 15 sistemas de sensores conectados, com 72 sensores de 10 modalidades, posicionados no ambiente, em objetos e no corpo dos voluntários. O sistema sensorial é composto por 24 acelerômetros e giroscópios *bluetooth*, 2 *Sun-SPOTs*, 2 *InertialCube3*, um sistema de localização *Ubisense* e um sensor de campo magnético. Os voluntários realizaram diversas atividades cotidianas gravadas em diferentes níveis de abstração de rótulos, realizadas de maneira natural, pois os roteiros utilizados foram de alto nível, deixando livre a interpretação de como realizar cada atividade. Para cada voluntário foram realizadas 6 tomadas de atividades, sendo 5 cotidianas e uma *drill*. A tomada cotidiana é composta pelas atividades: início (levantar da cadeira), *groom* (andar pelo ambiente e verificar o posicionamento dos objetos), relaxar, preparar café, beber café, preparar sanduíche, comer sanduíche, limpar e pausar. Por outro lado, a tomada *drill*, feita para gerar muitos dados de atividades, consiste em 20 repetições das seguintes tarefas: abrir e fechar geladeira, abrir e fechar lava-louças, abrir e fechar as 3 portas do armário, abrir e

fechar a porta 1, abrir e fechar a porta 2, ligar e desligar as luzes, limpar a mesa, beber em pé e beber sentado.

Em (KWAPISZ; WEISS; MOORE, 2011), foi feita a coleta de dados de acelerômetros de celulares posicionados nos bolsos frontais dos voluntários enquanto estes andavam, faziam *jogging*, subiam e desciam escadas, sentavam e ficavam em pé. Esses dados foram, então, pré-processados para que pudessem ser classificados por três ferramentas do WEKA, Árvores de Decisão, Regressão Logística e Redes Neurais Multicamadas. Com essas abordagens foram obtidas precisões acima de 90% na maioria das atividades, onde, no geral, as redes neurais foram mais precisas. Isso mostra que não são necessários diversos sensores distribuídos pelo corpo para atingir altas precisões em reconhecimento de atividade, sendo suficientes os acelerômetros de smartphones posicionados no bolso frontal do usuário.

Por outro lado, em (LI; ZHANG; LIU, 2010), propôs-se um banco de dados com 20 ações humanas, coletadas por um Kinect v1, usando mapas de profundidade. As ações foram focadas em movimentos usados em jogos, sendo, cada uma, realizada três vezes por todos os sete voluntários. Ademais, o artigo propõe um método para reconhecer atividade humana à partir de dados de profundidade, usando técnicas como *bag of 3D points* e *action graph*, mapeando o movimento nos mapas de profundidade usando uma baixa quantidade de pontos (aproximadamente 0,25 % dos pontos do mapa original) e alcançando baixas taxas de erro ao classificar tais pontos.

No trabalho de (SOOMRO; ZAMIR; SHAH, 2012), também foi apresentada uma base de dados de atividades humanas, a UCF101, constituído por 101 classes de ações humanas, somando mais de 13 mil vídeos com um total de 27 horas de dados. A base foi construída com clipes enviados por usuários e é uma extensão da base UCF50 (REDDY; SHAH, 2013), usando as mesmas 50 classes e 51 novas classes. As classes possuem cinco tipos: Interação Humano-Objeto, Apenas Movimento do Corpo, Interação Humano-Humano, Tocando Instrumentos Musicais e Esportes. Os vídeos foram obtidos do *YouTube* e todos possuem uma taxa de quadros e resolução constantes de 25 fps e 320×240, respectivamente. A obra também propõe uma metologia para classificar a base construída, usando *bag-of-words* e *Support Vector Machines* (SVM), alcançando uma precisão total de 44,5 %.

Em contrapartida, no artigo de (CHEN; JAFARI; KEHTARNAVAZ, 2015), foi criada uma base de dados multimodal, com informações de video (RGB e de profundidade) e inerciais. Para a coleta, usaram-se um Kinect v1 e um sensor inercial no laboratório ESSP da Universidade do Texas, sendo que a base possui 27 ações, realizadas por 8 voluntários, com cada um repetindo cada atividade 4 vezes. Nas gravações o Kinect foi posicionado em um tripé, para que capturasse todo o corpo do voluntário e o sensor inercial foi posicionado no pulso direito ou na coxa direita, dependendo da atividade realizada. Por fim, classificou-se o banco construído extraindo características estatísticas dos dados e aplicando *Principal Component Analyzis* (PCA), para selecionar as características relevantes. Com isso, usou-se um *Collaborative Representation*

Classifier (CRC) para classificar os dados de cada sensor individualmente e combinados, com a combinação melhorando em até 15 % a acurácia do modelo.

Já em (SONG *et al.*, 2016), propôs-se uma metodologia de classificação de dados egocêntricos (coletados da perspectiva do usuário) multimodais, usando aprendizagem profunda e *multi-stream*. Nessa abordagem, usaram 3 *streams* para os dados de vídeos, uma *single-frame*, outra para o fluxo óptico e outra para o fluxo óptico estabilizado. Cada *stream* alimentava uma CNN, sendo que as saídas das CNNs eram combinadas usando *Pooling* Médio e Máximo, além de uma camada *softmax*. Já para os dados de sensores, aplicou-se uma rede LSTM para cada sensor, combinando as saídas da mesma forma que para os vídeos. Ademais, criou-se uma base de dados egocêntrica, com dados de sensores inerciais e de video RGB, coletados composto por 20 atividades cotidianas realizadas por 10 voluntários, durante 15 segundos para cada atividade. As gravações foram realizadas tanto em ambientes fechados quando a céu aberto, permitindo diferentes variações de luminosidade.

No trabalho de (HAMMERLA; HALLORAN; PLÖTZ, 2016) foram utilizados diferentes métodos para HAR em cenários típicos, como gestos de manipulação, gestos repetitivos e atividades físicas, além de uma aplicação médica para a doença de Parkinson. No reconhecimento, foram utilizadas três abordagens diferentes, para comparar os resultados obtidos por cada uma e verificar a adequação de cada modelo para cada tarefa.

Essas abordagens são: *Deep Neural Network*s (DNN), CNN e RNN com *Long Short-Term Memory* (LSTM). Nos experimentos, cada modelo foi treinado de 30 a 300 épocas, usando três bancos de dados: OPPORTUNITY, PAMAP2 e Daphnet Gait (BACHLIN *et al.*, 2009). Esses bancos continham informações de movimentos rotuladas. Dentre esses modelos, o menos preciso nos resultados foi o DNN com 12% de erro no banco Opp enquanto o LSTM obteve a melhor performance sobre o mesmo banco, com 1% de erro, superando trabalhos anteriores na mesma base de dados.

A obra de Yang(YANG *et al.*, 2015) discute-se a vantagem das informações obtidas por sensores vestíveis sobre dados coletados externamente (por câmeras, por exemplo) para tarefas de HAR, pois tais sensores não são afetados por limitações do ambiente e configurações estacionárias, mantém a privacidade do usuário e podem ser mais preciso quando utilizam-se múltiplos sensores. Ademais, o trabalho estuda formas de melhor representar as séries temporais coletadas por sensores vestíveis, facilitando a reconhecimento de atividades, usando janelamento para segmentar o sinal e CNN para extrair características e classificar, comparando os resultados com outros classificadores. Os experimentos foram realizados nas bases OPPORTUNITY e *Hand Gesture* (BULLING; BLANKE; SCHIELE, 2014), com a CNN proposta obtendo os melhores resultados.

Por outro lado, o artigo de (RUEDA; FINK, 2018) busca novas formas de representar dados para aprimorar a tarefa de reconhecimento de atividade humana, de forma que não seja necessário extrair informações estatísticas dos dados, alcançando uma solução mais genérica.

Para isso, foram usadas três arquiteturas de redes neurais: CNN, *deepConvLSTM* e CNN-IMU. A CNN é composta de quatro convoluções temporais e duas camadas totalmente conectadas, com função de ativação ReLU e uma *softmax* classificadora. Já a *deepConvLSTM* também possui quatro convoluções temporais, porém, ao invés de duas camadas totalmente conectadas, usa camadas LSTM, extraindo as dependências temporais dos dados, seguida por uma *softmax*. Por fim, a CNN-IMU usa convoluções temporais e camadas totalmente conectadas para criar uma representação intermediária das leituras de cada IMU (*Inertial Measurement Unit*), fundi-as e aplica uma camada *softmax* classificadora. Para validar as arquiteturas, usaram as bases PAMAP2 e OPPORTUNITY, onde, dentre as arquiteturas citadas, a CNN-IMU obteve a melhor precisão na PAMAP2 e em gestos, enquanto a *deepConvLSTM* alcançou o melhor resultado no reconhecimento de atividades de locomoção.

Portanto, dados multimodais podem aprimorar a performance de classificadores no HAR, existindo poucas bases públicas de tal modalidade no campo de atividade humana. Além disso, *Deep Learning* é comumente usado para resolver tais problemas, com tais modelos, em geral, superando classificadores mais simples, pois algumas arquiteturas são capazes de extrair características relevantes dos dados e modelar a dependência temporal. Dessa forma, nesta obra construiu-se uma base de dados de atividade humana, classificando-a com arquiteturas de redes neurais baseadas na literatura.

Capítulo 4

MÉTODOS, TÉCNICAS E TECNOLOGIAS UTILIZADAS

O trabalho visou montar uma base de dados pública, com dados inerciais e de vídeo (RGB e de profundidade), coletados de voluntários realizando diversas atividades cotidianas, em diferentes ambientes. Além disso, também pretende-se estudar modelos para classificar dados de atividades humanas, a partir dos dados inerciais. Com isso, será explorada a base de dados OPPORTUNITY (CHAVARRIAGA *et al.*, 2013), além da base criada nesta obra, para comparar diferentes modelos de acordo com suas performances.

4.1 Coleta de Dados

A base construída nessa obra (VIHAD), é composta por dados de vídeo RGB, video de profundidade e duas unidades inerciais vestíveis. As atividades realizadas pelos voluntários são cotidianas, realizadas de maneira natural, para facilitar a generalização de um modelo extraído desta base. Os dados foram coletados e armazenados em formato *csv*, com o seguinte formato: para uma atividade X, um voluntário Y e uma gravação Z, salva-se cada arquivo gerado por cada sensor em uma pasta cujo o nome indica o sensor e o nome do arquivo é constituído como *actXseqYptZ.csv*. Sendo que cada linha dos arquivos inerciais é construída como {*timestamp*, sensor1.X, sensor1.Y, sensor1.Z, ...}, onde *timestamp* indica o instante de tempo em segundos

Figura 5 – Microsoft Kinect v1.

Figura 6 – *MetaMotionR*.

desde 1 de janeiro de 1970 (UTC), usado para sincronizar os diferentes sensores. Para os vídeos, as linhas de cada arquivo é contruída por um *timestamp* e um *frame* (imagem coletada naquele instante).

4.1.1 Materiais

Para a coleta dos vídeos, usou um Microsoft Kinect v1 (mostrado na Figura 5), que possui uma câmera RGB e de profundidade, e, para os dados inerciais, utilizou-se um *smartphone* Motorola X4 (coletando dados de acelerômetro, magnetômetro e giroscópio, a uma taxa de aproximadamente 120 Hz) e um sensor vestível *MetaMotionR* (MMR) (coletando dados de acelerômetro, magnetômetro e giroscópio, a uma taxa de aproximadamente 100 Hz), ambos exibidos, respectivamente, nas Figuras 7 e 6.

As unidades inerciais escolhidas, possuem a vantagem de coletar informações de maneira não invasiva, dado que objetos como celulares e relógios, são comumente carregados no cotidiano. Além disso, o Kinect v1, também captura dados de maneira natural, pois não interfere no movimento do usuário. Portanto, com os materiais usados, é possível captar os movimentos realizados pelos voluntários naturalmente, com pouca interferência em seus movimentos.

4.1.2 Metodologia

Para a coleta, participaram 6 voluntários, sendo 5 homens e 1 mulher com $22{,}5 \pm 4{,}72$ anos, escolhidos com apenas a idade mínima de 18 anos como requisito, realizando cada tarefa duas vezes, durante cinco segundos para cada gravação. O MMR foi posicionado no braço

Figura 7 – Smartphone Motorola X4.

dominante do voluntário e o *smartphone* no bolso frontal direito da calça do voluntário. O Kinect v1 foi posicionado de forma a captar os movimentos de ângulos favoráveis, em geral, filmando a parte frontal do corpo dos voluntários. O posicionamento dos sensores inerciais pode ser visualizado na Figura 8. As gravações foram realizadas em quatro ambientes diferentes, duas cozinhas e duas salas de estar, com diferentes níveis de iluminação, sendo que cada voluntário participou da gravação em no máximo dois ambientes. No total, a base somou 10 minutos de dados, com número de tuplas variado para cada tipo de dado coletado (por conta da frequência de coleta de cada sensor).

Nesta base, gravaram-se as seguintes tarefas cotidianas, onde a numeração indica o ID da atividade no arquivo:

1. Andar
2. Assistir televisão
3. Usar laptop
4. Comer
5. Pegar objetos
6. Beber sentado

Figura 8 – Posicionamento dos sensores vestíveis.

7. Cozinhar
8. Lavar louça
9. Limpar superfície
10. Esfregar chão

Código-fonte 1: Pseudocódigo do *script* de coleta.

```
1     iniciar_conexao_com_sensores()
2     tempo_de_coleta = 0  // em segundos
3     tempo_inicial = horario_agora()
4     Enquanto tempo_de_coleta < 5:
5         coletar_dados_dos_sensores()
6         tempo_passado = horario_agora() - tempo_inicial
7         tempo_de_coleta = tempo_de_coleta + tempo_passado
8     Fim enquanto
9     finalizar_conexao_com_sensores()
10    salvar_dados_em_arquivos()
11
```

Para cada tarefa, deixou-se a interpretação da realização a critério do voluntário, para que as mesmas fossem realizadas da maneira mais natural possível e não fossem enviesadas. Além

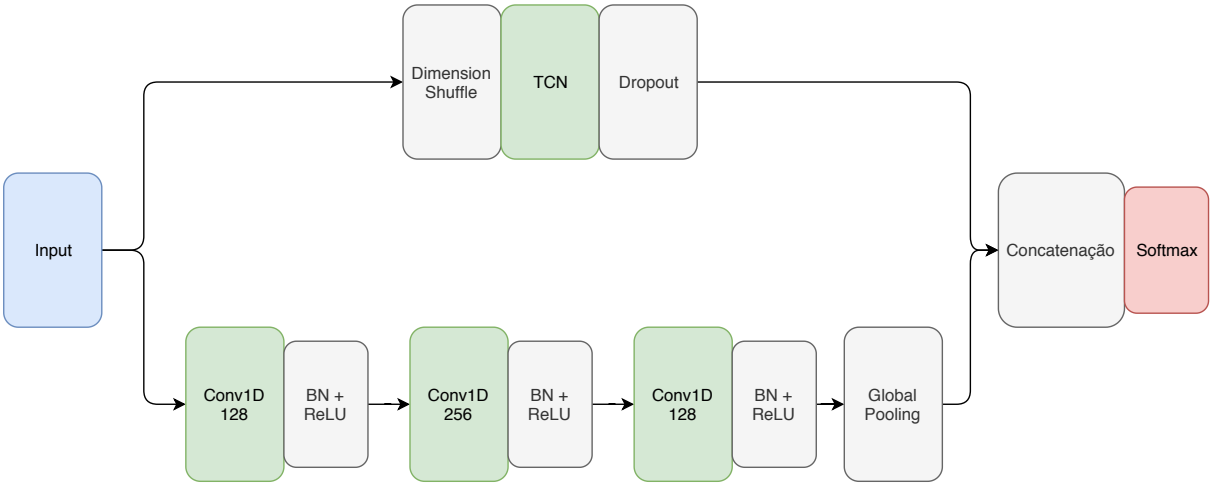


Figura 9 – Rede TCN-FCN.

disso, optou-se por coletar por cinco segundos por conta do alcance dos sensores em tarefas que envolviam deslocamento como, por exemplo, andar.

O *script* que inicia a coleta e sincroniza a aquisição de dados dos vídeos foi desenvolvido em *Python 3*, com auxílio da biblioteca *OpenKinect*[1], para lidar com o Kinect v1, com a biblioteca *MetaWear*[2], para lidar com o MMR e a comunicação com o *smartphone* foi realizada via *Socket* TCP. O Código-fonte 1 exibe um pseudocódigo do *script* de coleta.

4.2 Classificação das Bases

Reconhecimento de atividade humana, envolve dados temporais e, portanto, a dependência temporal é um fator importante nessa tarefa. Dessa forma, modelos que podem modelar dependências temporais, como RNN e TCN, são bons pontos de partida. Em especial, TCN, que pode ter uma memória maior e um menor tempo de treino (BAI; KOLTER; KOLTUN, 2018). Dessa forma, neste trabalho, foram usadas diferentes arquiteturas, compostas por TCN ou RNN com LSTM, para classificar as bases OPPORTUNITY e VIHAD, comparando a performance de cada modelo a fim de determinar a melhor abordagem para HAR.

Assim, foram usadas as arquiteturas *Temporal Convolutional Network-Fully Convolutional Network* (TCN-FCN) e ConvTCN, propostas em (GARCIA; RANIERI; ROMERO, 2019), além da LSTM-FCN e ConvLSTM, propostas, respectivamente, em (KARIM *et al.*, 2018) e (RUEDA; FINK, 2018). Dessa forma, pode-se comparar a performance das camadas TCN e LSTM ao modelar dependências temporais. Ademais, como pretende-se trabalhar com dados brutos, extrair características relevantes pode gerar melhorias significativas nos resultados, assim, as camadas CNN, serão usadas para extrair tais características, que poderão ser classificadas pela TCN ou LSTM. Com tais modelos, será possível classificar as bases OPPORTUNITY e

[1] Disponível em: <https://github.com/OpenKinect/libfreenect>
[2] Disponível em: <https://github.com/mbientlab/MetaWear-SDK-Python>

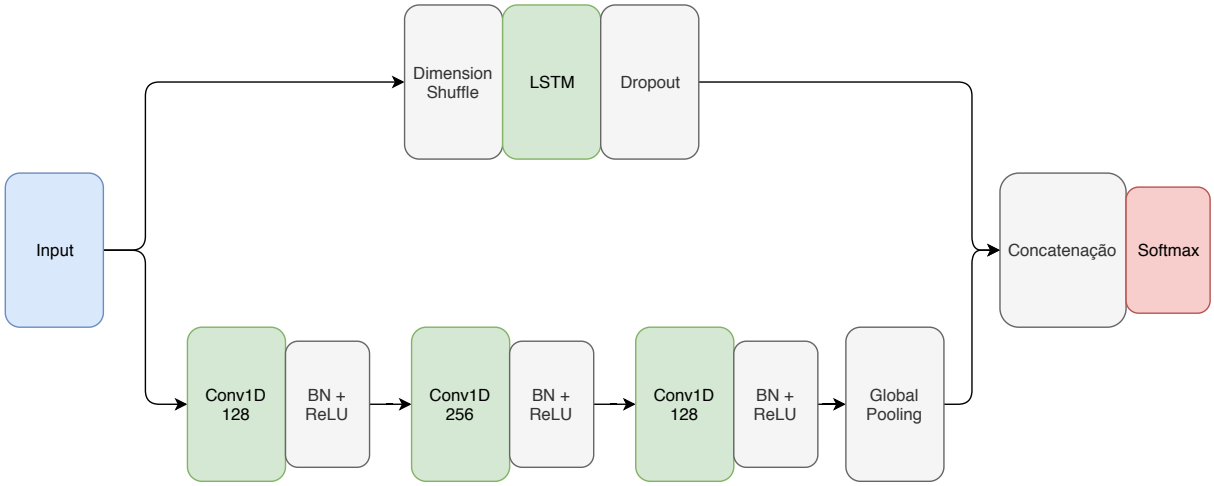


Figura 10 – Rede LSTM-FCN.

VIHAD, que serão pré-processadas com uma normalização L2 e janelamento temporal. Todo o desenvolvimento foi realizado em *Python 3*, com o auxílio da biblioteca *Keras* para a construção dos modelos.

A rede TCN-FCN é composta por dois blocos paralelos, o FCN e o TCN, onde ambos recebem a mesma entrada, simultaneamente e suas saídas são concatenadas, alimentando uma camada *softmax*. O bloco FCN possui três convoluções 1D encadeadas, com filtros de tamanho 128, 256 e 128, respectivamente, onde cada convolução é seguida por uma normalização em *batch* e uma função de ativação *Rectified Linear Unit* (ReLU). Após tais blocos convolucionais, aplica-se um *Pooling* Global. Já no bloco TCN, aplica-se à entrada um *Dimension Shuffle*, que, basicamente, transpões a dimensão do tempo com a dimensão das *features*. Em seguida, existe uma camada TCN, com filtro de tamanho 128, dilatações $d = \{1, 2, 4, 8, 16, 32\}$ e um *kernel* com tamanho 2. Após a camada TCN, aplica-se um *Droupout*, de 80%. A Figura 9 exibe uma representação desta rede. A rede LSTM-FCN, possui a mesma estrutura que a TCN-FCN, sendo que a única diferença é trocar a camanda TCN por uma camada LSTM com 128 células, como exibe a Figura 10. Os parâmetros para tais arquiteturas foram escolhidos baseando-se na literatura.

Por outro lado, a rede ConvTCN, é construída por uma convolução 1D na entrada, seguida por um *Max Pooling* e uma função ReLU, antecedendo outra camada convolucional constituída pelos mesmos componentes, ambas as convoluções possuem filtros de tamanho

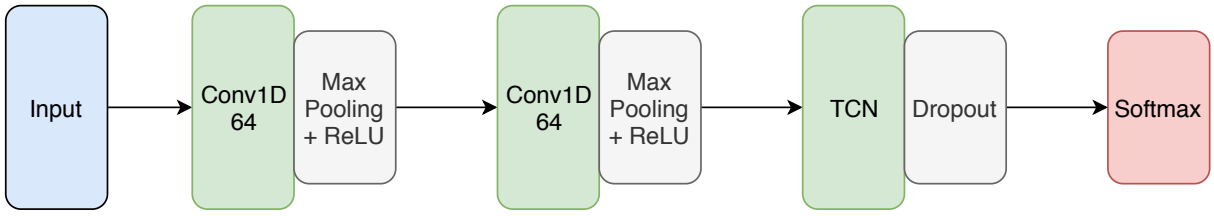


Figura 11 – Rede ConvTCN.

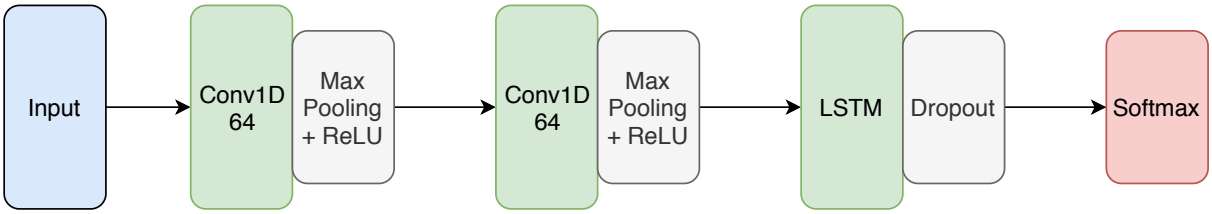


Figura 12 – Rede ConvLSTM.

64 e *kernel* de tamanho 2. Após as convoluções, existe uma camada TCN, com os mesmos parâmetros da camada TCN da rede TCN-FCN. A ConvLSTM, também possui a mesma estrutura de convoluções, se diferenciando por trocar a camada TCN por uma camada LSTM com 128 células. A rede ConvTCN pode ser visualizada na Figura 11, assim como a ConvLSTM na Figura 12. Da mesma forma, usaram-se os parâmetros da literatura para a construção das redes.

Com tais modelos, classificou-se a base OPPORTUNITY, usando 250 épocas de treino e validando a abordagem com *10-Fold Cross Validation*, deixando 80 % dos dados para treino e 20 % para teste. Além disso, essas arquiteturas também foram usadas para classificar a base VIHAD, treinando cada modelo por 50 épocas, por conta do menor volume de dados em relação à OPPORTUNITY, e validando com *6-Fold Cross Validation* (lembrando que a base foi construída com 6 voluntários) e *Leave-One-Out*, pelo baixo número de amostras disponíveis. Assim, na VIHAD, cada modelo foi treinado com dados de 5 voluntários e validados com os dados do voluntário restante, variando, em cada *Fold*, o voluntário usado para validação. Para ambas as validações, mediu-se a acurácia e *F1-Score* médios entre os *Folds*, comparando os valores obtidos em cada experimento para determinar o modelo que obteve maior sucesso em classificar cada base de dados.

Capítulo 5

DESENVOLVIMENTO

5.1 O Problema

O trabalho propôs uma base de dados pública multimodal, com dados de video RGB e de profundidade e de sensores inerciais, coletados em tarefas cotidianas realizadas por voluntários. Além disso, classificou-se a base OPPORTUNITY e base criada, chamada de VIHAD, com quatro diferentes arquiteturas baseadas na literatura, comparando a performance de camadas TCN e LSTM ao classificar dados temporais.

5.2 Atividades Realizadas

Para solucionar a problema, coletou-se dados inerciais e de vídeo de 6 voluntários, que realizaram 10 atividades durante cinco segundos cada, sendo que cada atividade foi gravada duas vezes. Enquanto tais voluntários vestiam uma pulseira MMR no pulso do braço dominante e um celular Motorola X4, posicionado no bolso frontal direito da coxa do usuário. Adicionalmente, os dados de vídeo RGB e de profundidade foram coletados usando um Kinect v1, com foco, em geral, na parte frontal do corpo do voluntário, de forma a capturar a maior quantidade de informações do movimento. Ao todo, coletou-se 10 minutos de dados.

Na classificação, usaram-se arquiteturas de redes neurais propostas na literatura: TCN-FCN, LSTM-FCN, ConvTCN e ConvLSTM. Com tais modelos, foi possível realizar o treinamento e validação nas bases OPPORTUNITY e VIHAD, focando, no segundo caso, apenas nos dados inerciais. Além disso, com tais redes, é possível comparar a performance das camadas TCN com a LSTM, verificando a aptidão de cada abordagem para a tarefa de classificação de atividade humana.

5.3 Resultados

Na classificação da base OPPORTUNITY, obtiveram-se os resultados exibidos na Tabela 1, onde a rede TCN-FCN superou as demais em acurácia e F1-Score, porém com um F1-Score muito similar ao da rede LSTM-FCN. Exibindo que, para este caso, a TCN melhorou o modelo, porém não tão significativamente, dado que ambos obtiveram performances muito similares. Já entre as redes ConvLSTM e ConvTCN, a rede ConvLSTM alcançou os melhores resultados.

Tabela 1 – Resultados para a base OPPORTUNITY, com os melhores exibidos me negrito.

Arquitetura	OPPORTUNITY	
	Acurácia (%)	F1 Score (%)
LSTM-FCN	88,11	88,09
TCN-FCN	**88,15**	**88,16**
ConvLSTM	85,19	84,27
ConvTCN	85,12	83,29

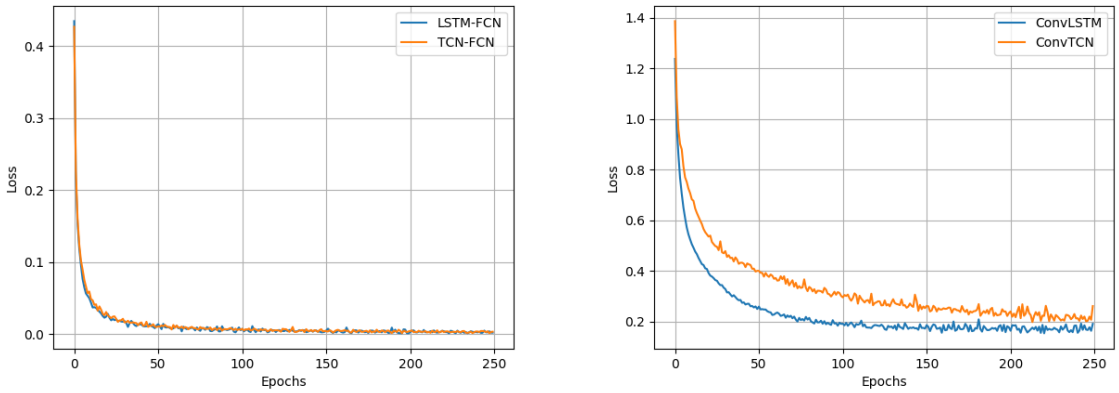
Entretanto, a acurácia de ambas também foi próxima. Assim, nota-se que redes TCN também são capazes de modelar dependências temporais, obtendo resultados similares ou até superiores em relação a LSTM.

Todavia, na base VIHAD, os resultados foram apresentados na Tabela 2, onde separou-se a classificação por sensores usados. Dessa forma, os modelos foram validados usando somente dados do MMR ou do *Smartphone* como entrada ou usando a combinação de ambos os sensores. Assim, a rede TCN-FCN superou as demais ao considerar apenas o MMR, sendo que a ConvTCN alcançou o segundo melhor resultado, com valores similares aos da LSTM-FCN e a rede ConvLSTM obteve os menores valores. Entretanto, usando apenas o *Smartphone*, a rede LSTM-FCN atingiu a maior acurácia e F1-Score, com, novamente, a ConvLSTM com a pior performance e a LSTM-FCN e ConvTCN com valores próximos, no segundo lugar. Finalmente, combinando-se os sensores, a rede TCN-FCN obteve os maiores resultados, superando, inclusive, os valores alcançados usando as unidades inerciais individualmente. Ademais, a rede LSTM-FCN obteve a segunda melhor performance, exibindo, também, uma melhoria em relação aos resultados da mesma rede sem combinar os sensores. Por outro lado, a arquitetura ConvLSTM obteve resultados semelhantes ao se utilizar apenas o MMR, com os piores resultados, assim como a ConvTCN, que apesar de superar a ConvLSTM, também não apresentou melhoria significativa com a combinação dos sensores.

Portanto, observamos que combinar os sensores inerciais pode melhorar consideravelmente os resultados, apesar de algumas arquiteturas se beneficiarem da adição de informações com mais intensidade que outras. Adicionalmente, a camada TCN alcançou melhores resultados em relação a LSTM na maioria das abordagens, mostrando o potencial de tal abordagem. Dessa forma, ao analisar dados temporais, o uso de TCN pode ser um bom ponto de partida, principalmente ao se considerar seu possível menor tempo de treino.

Tabela 2 – Resultados para a base VIHAD, com os melhores exibidos me negrito.

Arquitetura	MMR		*Smartphone*		MMR+*Smartphone*	
	Acurácia (%)	F1 Score (%)	Acurácia (%)	F1 Score (%)	Acurácia (%)	F1 Score (%)
LSTM-FCN	75,83	73,17	**80,16**	**79,28**	84,16	84,28
TCN-FCN	**76,01**	**74,34**	74,51	74,22	**85,83**	**85,51**
ConvLSTM	67,50	67,61	59,16	56,50	67,52	67,13
ConvTCN	75,88	73,21	76,41	75,94	76,84	76,16



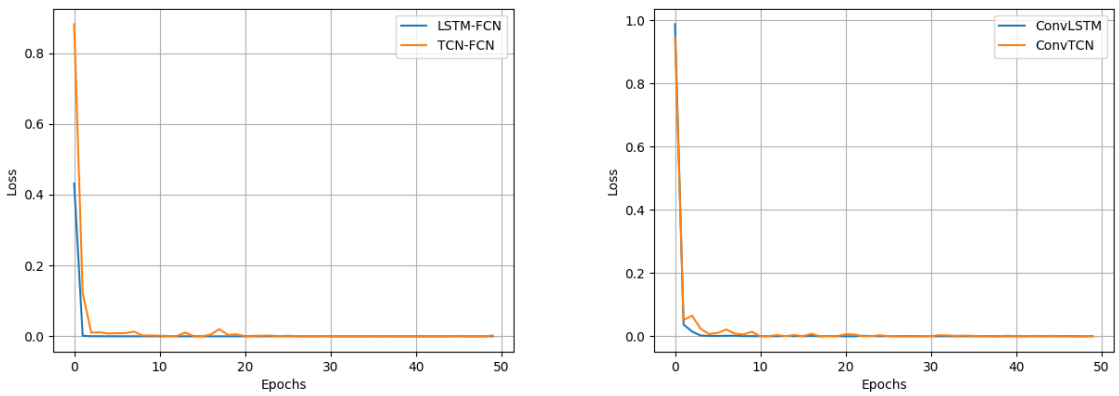
(a) TCN-FCN e LSTM-FCN (b) ConvTCN e ConvLSTM

Figura 13 – Perda por época para o primeiro *Fold* no treino na base OPPORTUNITY.

Por outro lado, ao se analisar a perda por época no treinamento de cada modelo nas bases usadas, verifica-se na Figura 13a que, na OPPORTUNITY, as redes TCN-FCN e LSTM-FCN atingiram a praticamente a mesma perda no fim do treino, com curvas de convergência muito similares, porém, na Figura 13b, nota-se que a ConvTCN levou mais épocas para convergir, atingindo uma perda maior no fim do treinamento, em relação à ConvLSTM. Já na base VIHAD, é possível visualizar que, pelas Figuras 14a e 14b, que as redes com TCN convergiram um pouco mais devagar, com uma perda similar à da LSTM no fim do treino.

5.4 Dificuldades e Limitações

Dentro os problemas que surgiram neste trabalho, destacam-se a dificuldade em gravar e sincronizar dados de quatro fontes diferentes, simultaneamente. Tal problema, foi solucionado iniciando a gravação de todos os sensores ao mesmo tempo, gravando o instante de tempo em



(a) TCN-FCN e LSTM-FCN (b) ConvTCN e ConvLSTM

Figura 14 – Perda por época para o primeiro *Fold* no treino na base VIHAD, com os sensores combinados.

segundos para cada ponto salvo e parando a gravação simultaneamente. Assim, os arquivos possuíam aproximadamente a mesma duração, podendo ser sincronizados usando o instante de tempo salvo. Todavia, o posicionamento dos sensores durante as atividades também foi uma dificuldade encontrada, resolvida com revisão bibliográfica e planejamento para coleta não invasiva. Adicionalmente,propor atividades para a base construída também foi uma dificuldade, solucionada baseando-se em atividade comumente citadas em artigos da bibliografia. Por fim, um outro problema encontrado na construção da base de dados foi reunir voluntários para a coleta, que poderia ter sido mais interessante com mais participantes.

Além disso, na classificação, não houve tempo hábil para compor os modelos com dados multimodais, usando também a câmera como entrada e validando sua contribuição na performance do modelo, assim como criar um modelo para cada sensor, validando a contribuição individual e combinada dos mesmos.

Capítulo 6

CONCLUSÃO

Nesta obra, foi apresentada uma nova base de dados multimodal, apelidada de VIHAD, construída por dados inerciais, coletados por um sensor MMR e um *smartphone* Moto X4, posicionados, respectivamente, no pulso do braço dominante e na bolso superior frontal da calça do usuário. Além de dados de vídeo RGB e de profundida, obtidos com o auxílio de um Kinect v1, com posicionamento variado de acordo com a atividade realizada pelo participante. Ao todo, a base possui 10 atividades: andar, assistir televisão, usar laptop, comer, pegar objetos, beber sentado, cozinhar, lavar louça, limpar superfície e esfregar chão. Participaram da construção da VIHAD 6 voluntários, todos realizando todas as atividades propostas duas vezes, sendo orientados a decidirem como realizar cada tarefa, para evitar enviesar a coleta.

Adicionalmente, usaram-se quatro arquiteturas de redes neurais para classificar a base de dados pública OPPORTUNITY e a VIHAD. Tais redes foram: TCN-FCN, LSTM-FCN, ConvTCN e ConvLSTM, baseadas em arquiteturas da literatura. Com tais modelos, foi possível comparar a performance de redes TCN e LSTM, com a TCN superando a LSTM na maioria dos experimentos realizados. Dessa forma, foi possível obter uma acurácia de 88,15 % na classificação da OPPORTUNITY, com a TCN-FCN e 85,51 % na VIHAD, também com a TCN-FCN. Assim, nota-se que, ao trabalhar com dados que possuem dependência temporal, redes TCN podem ser interessantes como ponto de partida, pois podem obter melhores resultados que redes recorrentes, podem possuir um menor tempo de treinamento e também, em alguns casos, uma memória maior que a LSTM (BAI; KOLTER; KOLTUN, 2018).

Como trabalhos futuros, pretende-se expandir a base de dados com mais voluntários e disponibiliza-la publicamente, assim como classificá-la usando os dados de vídeo e combinando os sensores, verificando a possível melhoria alcançada ao se utilizar informações multimodais.

REFERÊNCIAS

BACHLIN, M.; ROGGEN, D.; TROSTER, G.; PLOTNIK, M.; INBAR, N.; MEIDAN, I.; HERMAN, T.; BROZGOL, M.; SHAVIV, E.; GILADI, N. *et al.* Potentials of enhanced context awareness in wearable assistants for parkinson's disease patients with the freezing of gait syndrome. In: IEEE. **Wearable Computers, 2009. ISWC'09. International Symposium on**. [S.l.], 2009. p. 123–130. Citado na página 33.

BAI, S.; KOLTER, J. Z.; KOLTUN, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. **arXiv preprint arXiv:1803.01271**, 2018. Citado 2 vezes nas páginas 39 e 47.

BULLING, A.; BLANKE, U.; SCHIELE, B. A tutorial on human activity recognition using body-worn inertial sensors. **ACM Computing Surveys (CSUR)**, ACM, v. 46, n. 3, p. 33, 2014. Citado na página 33.

CHAVARRIAGA, R.; SAGHA, H.; CALATRONI, A.; DIGUMARTI, S. T.; TRÖSTER, G.; MILLÁN, J. d. R.; ROGGEN, D. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. **Pattern Recognition Letters**, Elsevier, v. 34, n. 15, p. 2033–2042, 2013. Citado 3 vezes nas páginas 24, 31 e 35.

CHEN, C.; JAFARI, R.; KEHTARNAVAZ, N. Improving human action recognition using fusion of depth camera and inertial sensors. **IEEE Transactions on Human-Machine Systems**, IEEE, v. 45, n. 1, p. 51–61, 2014. Citado na página 23.

____. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: IEEE. **2015 IEEE International conference on image processing (ICIP)**. [S.l.], 2015. p. 168–172. Citado 3 vezes nas páginas 23, 24 e 32.

GARCIA, F.; RANIERI, C.; ROMERO, R. A. F. Temporal approaches for human activity recognition using inertial sensor. In: **SBR-LARS 2019** (). [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 24 e 39.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>. Citado na página 29.

HAMMERLA, N. Y.; HALLORAN, S.; PLÖTZ, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. **arXiv preprint arXiv:1604.08880**, 2016. Citado na página 33.

HAYKIN, S. S. *et al.* **Neural networks and learning machines/Simon Haykin.** [S.l.]: New York: Prentice Hall,, 2009. Citado 3 vezes nas páginas 24, 27 e 28.

HOPFIELD, J. J. Neural networks and physical systems with emergent collective computational abilities. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 79, n. 8, p. 2554–2558, 1982. Citado na página 29.

KARIM, F.; MAJUMDAR, S.; DARABI, H.; CHEN, S. Lstm fully convolutional networks for time series classification. **IEEE Access**, IEEE, v. 6, p. 1662–1669, 2018. Citado 2 vezes nas páginas 24 e 39.

KWAPISZ, J. R.; WEISS, G. M.; MOORE, S. A. Activity recognition using cell phone accelerometers. **ACM SigKDD Explorations Newsletter**, ACM, v. 12, n. 2, p. 74–82, 2011. Citado na página 32.

LEA, C.; FLYNN, M. D.; VIDAL, R.; REITER, A.; HAGER, G. D. Temporal convolutional networks for action segmentation and detection. In: **proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2017. p. 156–165. Citado na página 24.

LI, W.; ZHANG, Z.; LIU, Z. Action recognition based on a bag of 3d points. In: IEEE. **2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops**. [S.l.], 2010. p. 9–14. Citado na página 32.

METZ, C. E. Basic principles of roc analysis. In: ELSEVIER. **Seminars in nuclear medicine**. [S.l.], 1978. v. 8, n. 4, p. 283–298. Citado na página 30.

OORD, A. V. D.; DIELEMAN, S.; ZEN, H.; SIMONYAN, K.; VINYALS, O.; GRAVES, A.; KALCHBRENNER, N.; SENIOR, A. W.; KAVUKCUOGLU, K. Wavenet: A generative model for raw audio. **SSW**, v. 125, 2016. Citado na página 30.

REDDY, K. K.; SHAH, M. Recognizing 50 human action categories of web videos. **Machine Vision and Applications**, Springer, v. 24, n. 5, p. 971–981, 2013. Citado na página 32.

REISS, A.; STRICKER, D. Introducing a new benchmarked dataset for activity monitoring. In: IEEE. **2012 16th International Symposium on Wearable Computers**. [S.l.], 2012. p. 108–109. Citado 2 vezes nas páginas 24 e 31.

RUEDA, F. M.; FINK, G. A. Learning attribute representation for human activity recognition. In: IEEE. **2018 24th International Conference on Pattern Recognition (ICPR)**. [S.l.], 2018. p. 523–528. Citado 2 vezes nas páginas 33 e 39.

SONG, S.; CHANDRASEKHAR, V.; MANDAL, B.; LI, L.; LIM, J.-H.; BABU, G. S.; SAN, P. P.; CHEUNG, N.-M. Multimodal multi-stream deep learning for egocentric activity recognition. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops**. [S.l.: s.n.], 2016. p. 24–31. Citado 2 vezes nas páginas 24 e 33.

SOOMRO, K.; ZAMIR, A. R.; SHAH, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. **arXiv preprint arXiv:1212.0402**, 2012. Citado 2 vezes nas páginas 24 e 32.

YANG, J.; NGUYEN, M. N.; SAN, P. P.; LI, X. L.; KRISHNASWAMY, S. Deep convolutional neural networks on multichannel time series for human activity recognition. In: **Twenty-Fourth International Joint Conference on Artificial Intelligence**. [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 23 e 33.