UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

HUMAN ACTION RECOGNITION BASED ON SPATIOTEMPORAL FEATURES FROM VIDEOS

MURILO VARGES DA SILVA

ADVISOR PROF. DR. APARECIDO NILCEU MARANA

São Carlos – SP

December/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

HUMAN ACTION RECOGNITION BASED ON SPATIOTEMPORAL FEATURES FROM VIDEOS

MURILO VARGES DA SILVA

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação, área de concentração: Processamento de Imagens e Sinais.

Advisor Prof. Dr. Aparecido Nilceu Marana

São Carlos – SP

December/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Murilo Varges da Silva, realizada em 22/12/2020.

Comissão Julgadora:

Prof. Dr. Aparecido Nilceu Marana (UNESP)

Prof. Dr. Ricardo Cerri (UFSCar)

Prof. Dr. João Paulo Papa (UFSCar)

Profa. Dra. Fátima de Lourdes dos Santos Nunes Marques (USP)

Prof. Dr. Clayton Reginaldo Pereira (UNESP)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.
O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

Este trabalho é dedicado aos meus queridos pais.

AGRADECIMENTOS

O presente trabalho é fruto de anos de esforço, superação e dedicação aos estudos, bem como de contribuições diretas ou indiretas de meus familiares, professores, amigos e colegas de profissão, sem os quais não seria possível atingir esses resultados.

Aos meus amados pais, Luiz e Luzinete (in memoriam), meus primeiros exemplos, professores e incentivadores. Obrigado por tudo, pelo carinho, educação, e por mostrarem que não existem limites para o crescimento na vida.

Às minhas irmãs Fabiana, Mônica e Priscila, pelo carinho, confiança e motivação que sempre me deram no decorrer desta trajetória. Agradeço por serem professoras e exemplos, se não fosse por vocês e a motivação transmitida provavelmente não teria ingressado nesta carreira tão desafiadora e importante de professor.

À minha esposa Priscila, por participar de toda esta trajetória desde o início lá em 2010 quando resolvi prestar o concurso para professor do IFSP, me motivando e acompanhando nas etapas do concurso. Agradeço pelo apoio incondicional, motivação e compreensão. Compartilho com você esta conquista.

À minha família e amigos, pelo grande apoio e compreensão nos momentos em que estive ausente, pelos quais tenho profunda admiração e gratidão por fazerem parte de minha vida.

Ao professor doutor Aparecido Nilceu Marana, mais conhecido como Nilceu, sem o qual não sei se seria possível realizar o sonho de concluir o mestrado e o doutorado, além de um excelente professor e orientador, agradeço por ser esta pessoa acolhedora, humana e sensível. Agradeço por me acolher em 2013 quando cursei a disciplina "Sistema Biométricos", no primeiro dia da disciplina conseguiu despertar em mim o interesse por essa área de pesquisa, confesso que talvez se não fosse por você teria desistido do mestrado quando estava iniciando as disciplinas como aluno especial.

Aos amigos do laboratório RECOGNA UNESP, em especial ao Clayton, Leandro, Douglas, Luís Sugi e Luiz Felix, pelo acolhimento nos períodos que passamos juntos em São Carlos e Bauru, pelas conversas, viagens, pelos bares da vida e auxílios oferecidos no decorrer da Pós-Graduação.

Ao Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), instituição que me orgulho de fazer parte do quadro de docentes desde 2011, que sempre incentivou a participação dos seus servidores em programas de qualificação e que não mediu esforços para flexibilização dos horários para atendimento dos compromissos do doutorado e que concedeu afastamento das atividades para conclusão do doutorado.

À Universidade Federal de São Carlos (UFSCar) e ao Programa de Pós-Graduação em Ciência da Computação (PPGCC), pela infraestrutura oferecida, por terem me dado oportunidade de cursar o Doutorado e pela qualidade do Programa de Pós-Graduação, agradeço a todos os professores que contribuíram para minha formação e aos servidores pelo suporte prestado neste período.

À Nvidia, pela doação das GPUs utilizadas nos estudos e experimentos e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) por financiar o programa de Pós-Graduação e pelos auxílios oferecidos no decorrer do Doutorado.

Não sabendo que era impossível, ele foi lá e fez.
(Jean Cocteau)

O sucesso é ir de fracasso em fracasso sem perder entusiasmo.
(Winston Churchill)

RESUMO

Atualmente, existe uma alta demanda para o desenvolvimento de novas técnicas de reconhecimento automático de padrões em vídeos, como por exemplo para o reconhecimento automático de ações humanas, demanda essa motivada pelos avanços nas tecnologias de produção, armazenamento, transmissão e compartilhamento de vídeos, tais avanços desencadearam a produção de um grande volume de vídeos que para serem úteis necessitam de tratamento automatizado. Dentre as principais aplicações do reconhecimento de ações humanas em vídeos, destacam-se: vigilância em locais públicos, detecção de quedas de idosos em suas residências, automação em lojas com sistema de *checkout* sem atendentes, detecção de ações de pedestres por parte de veículos autônomos, detecção de conteúdo inadequado postado na internet, como violência ou pornografia, etc. O reconhecimento automático de ações em vídeos é uma tarefa desafiadora, pois para se obter boas taxas de acurácia é necessário trabalhar com informações espaciais (por exemplo, formas encontradas em um único quadro do vídeo) e informações temporais (por exemplo, padrões de movimentos encontrados entre os quadros do vídeo). Nesta tese são propostos novos métodos para reconhecimento automático de ações humanas a partir de informações espaço-temporais extraídas de vídeos. Inicialmente, foram avaliadas diferentes arquiteturas de Redes Neurais de Convolução 3D (*3D CNN - Convolutional Neural Networks*) no contexto de detecção de pornografia em vídeos. Após, foram propostos novos métodos para o reconhecimento de ações humanas baseados em informações espaço-temporais extraídas de poses 2D. O uso de poses 2D se mostrou uma estratégia promissora, pois exige um custo computacional menor se comparado com técnicas que utilizam aprendizado de máquina em profundidade, além disso ao se utilizar poses 2D ao invés das imagens brutas pode-se preservar a privacidade das pessoas e dos ambientes onde as câmeras de vídeos estão instaladas. O método proposto, apresentou taxas de acurácia compatíveis com o estado-da-arte nas bases de dados públicas em que os experimentos foram realizados.

Palavras-chave: Reconhecimento de Ações Humanas, Poses em 2D, Classificação de Vídeo.

ABSTRACT

Currently, there is a high demand for the development of new techniques for automatic pattern recognition in videos, for example for the automatic recognition of human actions, this demand is motivated by the advances in the technologies of production, storage, transmission and sharing of videos, such advances triggered the production of a huge volume of videos that need to be automatically processed to be useful. Among the main applications, we can highlight: surveillance in public places, detection of falls of the elderly in their homes, automation in no-checkout-required stores, detection of pedestrian actions by self-driving car, detection of inappropriate content posted on the Internet like violence or pornography, etc. The automatic recognition of actions in videos is a challenging task because, in order to obtain good classification rates, it is necessary to work with spatial information (for example, shapes found in a single frame of the video) and temporal information (for example, movement patterns found throughout the frames in the video). In this thesis new methods are proposed for automatic recognition of human actions based on spatiotemporal features extracted from videos. Initially, different architectures of 3D Convolution Neural Networks (CNNs) were evaluated in the context of detecting pornography in videos. Afterwards, new methods were proposed for the recognition of human actions based on spatiotemporal information extracted from 2D poses. The use of 2D poses proved to be a promising strategy, as it requires a lower computational cost when compared to techniques that use deep learning. Besides, by using 2D poses, instead of raw images, one can preserve the privacy of people and places where the video cameras are installed. The proposed method has presented accuracy rates compatible with the state-of-the-art rates on the public databases in which the experiments were carried out.

Keywords: Human Action Recognition, 2D Poses, Video Classification.

LIST OF ABBREVIATIONS AND ACRONYMS

2D	Two-dimensional
3D	Three-dimensional
BoP	Bag of Poses
BoVW	Bag of Visual Words
BoW	Bag of Words
BRACIS	Brazilian Conference on Intelligent Systems
Caffe	Convolutional Architecture for Fast Feature Embedding
CIARP	Iberoamerican Congress on Pattern Recognition
CAPES	Coordenação de Aperfeiçomento de Pessoal de Nível Superior
CCTV	Closed-circuit television
CNN	Convolutional Neural Network
DT	Dense Trajectory
FPS	Frames per Second
FV	Fisher Vector
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HoF	Histogram of Optical Flow
HoG	Histogram of Oriented Gradient
I3D	Inflated 3D CNN
iDT	Improved Dense Trajectory

kNN	k-Nearest Neighbors
LOOCV	Leave-One-Out-Cross-Validation
LRCN	Long-term Recurrent Convolutional Network
LSTM	Long Short-Term Memory
MBH	Motion Boundary Histograms
MRI	Magnetic Resonance Imaging
OPF	Optimum-Path Forest
PAF	Part Association Field
PCA	Principal Component Analysis
PIF	Part Intensity Field
ResNet	Residual Network
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SIFT	Scale-Invariant Feature Transform
STIP	Space-time Interest Point
SVM	Support Vector Machine
t-SNE	t-Distributed Stochastic Neighbor Embedding
UFSCar	Universidade Federal de São Carlos
VGG	Visual Geometry Group

LIST OF FIGURES

List of Tables

CONTENTS

Chapter 1
INTRODUCTION

In this chapter, we present the problem investigated in the doctoral thesis, as well as the motivations, challenges, objectives, justification, contributions, research questions, and text organization.

1.1 Problem Characterization and Motivation

Currently, there is a high demand for the development of new techniques for automatic pattern recognition in videos, for example for the automatic recognition of human actions. This demand is motivated by the advances in the technologies of production, storage, transmission and sharing of videos, such advances triggered the production of a huge volume of videos that need to be automatically processed to be useful. Among the main applications, we can highlight: surveillance in public places (e.g., airports, hospitals, malls, etc.), detection of falls of the elderly in their homes, automation in no-checkout-required stores, detection of pedestrian actions by self-driving car, detection of inappropriate content posted on the Internet like violence or pornography, etc.

Human action recognition systems aim to identify actions and intentions of one or more individuals through a series of observations of each individual in a specific context. The recognition of human actions in a computer vision system can be considered the last step of a series of previous activities, such as image capture, segmentation, identification and tracking of objects of interest and their classification. An action is a sequence of movements of the human body and can involve several parts simultaneously. In the context of computer vision, the recognition of actions can be executed by matching a series of observations made in a sequence of frames of a video with a previously defined pattern and assigning a label to the action (CHENG et al., 2015).

Aggarwal & Ryoo (2011) defined four classes for human actions depending on their complexity:

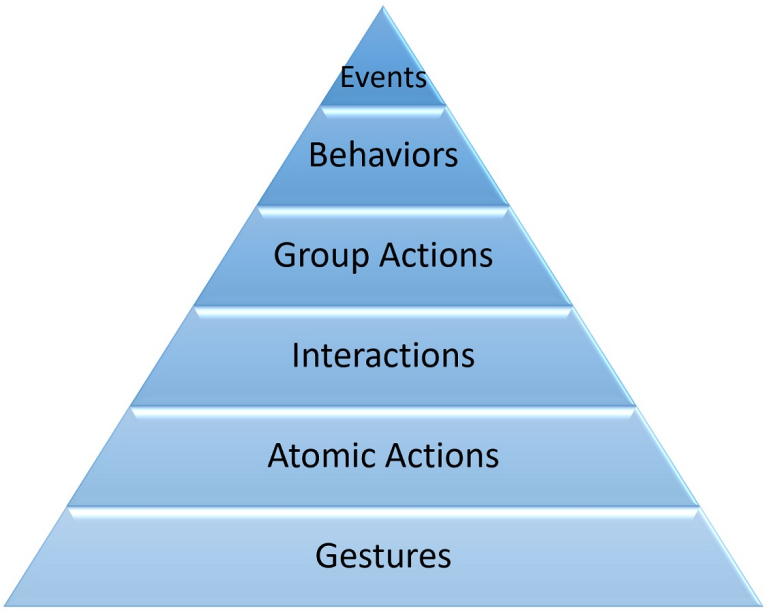
- **Gesture**: A sequence of movements associated with a meaning - made by limbs, head or

face. Example: stretching your arm, shaking your head, smiling, lifting your leg;

- **Action**: A sequence of multiple gestures performed by a person. Example: walking, waving, running and swimming;
- **Interaction**: A sequence of human actions involving at least two actors, one actor must be a human and the other can be a person or an object. Example: two people fighting, a person typing on a computer keyboard;
- **Group activity**: Activities performed by several people and/or objects. Example: a group of people marching, football game.

Vrigkas et al. (2015) categorized human activities into six classes depending on the complexity. In addition to the four classes defined by Aggarwal & Ryoo (2011), they also proposed the classification of actions in behaviors and events. Behaviors refer to physical actions that are associated with an individual's emotions, personality, and psychological state. Events are high-level activities that describe the social actions between individuals and indicate the intention or the social role of a person. Figure 1 presents the six categories defined by Vrigkas et al. (2015) organized by complexity, where the most straightforward human actions are classified as low-level (Gestures) and the most complex ones as high-level (Events).

Figure 1 – Classification of human actions proposed by Vrigkas et al. (2015).



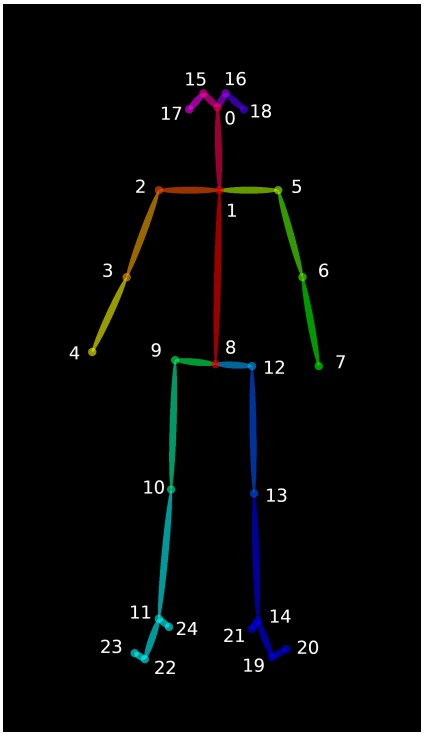
Source: Adapted from Vrigkas et al. (2015).

A common action recognition method usually contains two main components: action representation and action classification. The first one converts an action performed in a video into a feature vector, or a series of vectors. The second component infers a label to the action by

using the feature vector. Nowadays, deep networks, like convolutional neural networks (CNNs), can perform these two steps into a unified end-to-end trainable framework, which in general enhance the classification performance (KONG; FU, 2018). Then, we can group human action recognition techniques into two groups: handcrafted, for techniques where features are extracted according to a certain manually predefined algorithm based on the expert knowledge, and deep learning, for techniques where features are derived from an image dataset by a training procedure of a CNN, for instance, in order to fulfill a certain task (ANTIPOV et al., 2015).

Recently, new 2D pose detection techniques such as OpenPose (CAO et al., 2018) and PifPaf (KREISS et al., 2019) have emerged, featuring real-time processing and good people-detection rates. The 2D human pose detection is defined as the problem of localizing human body joints (elbows, shoulders, hips, wrists, etc) in images or videos. Figure 2 shows the 25 key points obtained from a 2D pose by using OpenPose framework.

Figure 2 – The skeleton and the 25 key points obtained from a 2D pose by the OpenPose framework.



Source: Cao et al. (2018)

Since 2D poses are good sources of information about the actions being performed across the video, and motivated by the success of 2D pose-based human action recognition (LV; NEVATIA, 2007; GALL et al., 2010; WANG et al., 2013), we propose new methods for human action recognition based on a new set of descriptors computed from 2D poses. These descriptors are based on angles formed by adjacent body parts and trajectories of body joints or body parts across the video frames.

1.2 Challenges

Automatic recognition of human actions from videos is a challenging task. Among the main difficulties, we can mention:

- **Human privacy concerns:** The indiscriminate use of closed-circuit television (CCTV) cameras for surveillance sometimes enable privacy intrusion. Since computer-vision based methods continue to improve, the engineers should assess the role of machines in people surveillance, and how automation can be used to help protect privacy. Many CCTV applications of extreme relevance do not have the need to identify people to fulfill their roles, such as applications for detecting the elderly fall in domestic and hospital environments. (Senior et al., 2005; CHEN et al., 2017);

- **Intra and inter-class variations:** For many actions, there is a variety of ways that humans can execute them (e.g., the walking action can have different speeds and stride sizes). This problem of variations can occur for other actions, especially for non-cyclical actions or those that are adapted because of the environment (e.g., avoid obstacles while walking). The human actions recognition approaches should be able to generalize the variations within a class and distinguish actions between classes. By increasing the number of classes of actions to be recognized, this task becomes even more challenging since the classes overlapping becomes greater (POPPE, 2010);

- **Diversity of action classes:** The set of actions that human beings can accomplish is vast, diverse and presents distinct levels of complexities, ranging from a simple gesture to complicated group interactions (CHAQUET et al., 2013);

- **Environmental variations:** The environment where the action is performed is an important source of variation in images. Therefore, the simple task of locating a person can be more difficult in cluttered or dynamic environments. Additionally, parts of a person's body may be hidden and light conditions can alter a person's appearance. The same action observed from different point-of-views can generate very different images (POPPE, 2010);

- **Processing time:** In video-recognition applications requiring real-time responses, a major challenge is the high computational cost required to perform automatic action recognition. In the past, an easy way to increase the performance of a computer was waiting for the evolution of the semiconductors, which resulted in an increase in the clock speed of the device and the speed of all the applications without being modified by the programmer. Unfortunately, these days are over. As the transistors become denser, they also leak more current and are therefore less energy efficient (LU; SHI, 2013);

- **Obtaining labeled data for training:** Although there are some public databases created to support the development of automatic techniques for recognizing human actions from

videos, not all of them present the challenges mentioned above. The task of labeling videos is very labor-intensive and challenging, so some techniques have been proposed to perform labeling automatically, for example, by using Internet searching results or video subtitles. Experiments showed significant differences between the labels chosen by humans and those chosen by automatic techniques applied on the existing databases (SHORTEN; KHOSHGOFTAAR, 2019).

1.3 Hypothesis and Research Questions

The recognition of human actions from videos is a challenging task, usually demanding high computational cost and, depending on the scenario of application, facing some problems regarding people's privacy. Based on that, the central hypothesis of this thesis is that it is possible to use 2D poses to recognize human actions in videos with low computational cost and with competitive accuracy rates (comparable to the accuracy rates obtained by techniques that use the raw video frames), while preserving people's privacy. Therefore, to validate the power of representation of 2D poses and the hypothesis of this thesis, our research was guided by some questions, considering the obstacles for the automatic recognition of actions in videos previously presented, our research questions are as follow:

1. What are the advantages and disadvantages of using deep learning to classify human actions in videos?
2. Can we use only 2D poses for the development of an automatic recognition method for human actions in videos?
3. How can we represent spatial and temporal information from 2D poses extracted from videos?
4. Is it possible to recognize human actions in videos while preserving the identity and the privacy of the people and places involved in the scenes?

1.4 Objectives

The general objective of this thesis is to propose and develop an approach for human action recognition based on 2D poses extracted from the video frames.

The specific objectives are:

- To review and assess some state-of-art methods that use deep learning to classify human action in videos;
- To present a new way of representing spatial and temporal information using 2D poses;

- To propose a human action recognition method that preserves the identity and the privacy of the people and places involved in the scenes.

1.5 Justification

The automatic recognition of human actions from videos is a task of paramount importance due to the wide range of relevant applications that can be employed. Thus, the development of robust methods that are concerned with the privacy issues previously raised is essential for the safe development of cities, as well as for the well-being of the people.

The use of 2D poses plays an important role in the task of human action recognition in videos. Firstly, we can highlight that 2D poses, unlike 3D poses, can be estimated from images captured from conventional cameras that are already spread out in many environments, such as houses, shopping centers, streets, airports, etc. Secondly, 2D poses can be used in environments that need to preserve human identity and privacy. As an example, we can mention an automatic system for monitoring possible falls of elderly people in their homes. Since this system needs to capture images of people within the privacy of their homes, many will not accept it. So by using 2D poses instead of the raw video frames, we can ensure that privacy and identity can be preserved. Buzzelli et al. (2020) proposed a vision-based system for monitoring elderly people at home that uses raw video frames and provides a web tool for monitoring and searching action in videos. However, this system can face some problems due to privacy concerns.

It is evident that the information about the environment can help the classification of some human actions. For example, He et al. (2016) proposed a method to recognize human actions using only the background. However, in order to preserve privacy of people and places, it is desirable to use only information from the two-dimensional poses of the individual in the scene. Figure 3 shows the original video frame with a person sitting on a couch and its correspondent 2D pose. It is clear that when using raw videos, the system can reveal the person's action, but also some private information, such as his/her identification, clothes, furniture and other valuables captured by the camera.

1.6 Contribution

Methods for automatic recognition of human actions from videos are very relevant because of the large number of applications. Thus, the evaluation of the use of 2D poses in the context of recognition of human actions can bring competitive results with the use of simple techniques.

The main contribution of this thesis is the development of an approach based on spatiotemporal features extracted from 2D poses for human action recognition. By using 2D poses, the proposed methods allow human action recognition using small labeled datasets in the

Figure 3 – Raw video frame and 2D pose extracted from raw video frame.

(a) Raw video Frame (b) 2D pose

Source: The author. Image from Buzzelli et al. (2020).

training stage, with conventional and affordable computers, while providing high accuracy rates and preserving people's privacy.

1.7 Document Organization

In addition to this introductory chapter, the structure of the document is as it follows:

- **Chapter 2 - 2D Pose Estimation**: This chapter presents an overview on 2D pose estimation and two state-of-the-art methods for 2D poses estimation;
- **Chapter 3 - Related Works**: This chapter presents relevant related work on human action recognition. Firstly, handcrafted methods for human action recognition are explained. Finally, some related works that use deep learning approaches are discussed;
- **Chapter 4 - Material and Methods**: This chapter introduces the proposed methods based on deep learning and 2D poses for the human action recognition in videos, as well as the material and metrics used in its proposals and for evaluations;
- **Chapter 5 - Spatiotemporal CNNs for Pornography Detection in Videos**: This chapter presents the initial results from this doctoral thesis, based on a study on Spatiotemporal Convolutional Neural Networks for classifying human actions in videos;
- **Chapter 6 - Human Action Recognition in Videos Based on Angles and Trajectories From 2D Poses**: This chapter presents a new method for human actions recognition in videos from new descriptors that can represent spatial and temporal information obtained from 2D poses;

- **Chapter 7 - Human Action Recognition in Videos Based on Spatiotemporal Features and Bag-of-Poses**: This chapter presents a new method of representing 2D poses. Instead of directly using the straight-line segments, the 2D pose is converted to the parameter space in which each segment is mapped to a point. Then, from the parameter space, spatiotemporal features are extracted and encoded using a Bag-of-Poses approach, then used for human action recognition in the video.
- **Chapter 8 - Conclusions**: This chapter presents the general conclusions of the thesis based on the studies carried out, the works developed and the results obtained, highlighting the main contributions, possible future works and the publications carried out during the doctorate.

Chapter 2
2D Pose Estimation

This chapter presents an overview on 2D pose estimation and two state-of-the-art methods for 2D poses estimation.

2.1 2D Pose Estimation

Estimating 2D poses is important for understanding human behavior in images and videos. The estimation of human 2D pose is related to the problem of localizing anatomical key points or "body parts" (CAO et al., 2018). Techniques used in 2D poses estimation are essential since 2D poses serve to solve various problems in many areas such as activity recognition, animation, gaming, and augmented reality. Current 2D pose estimating methods are mostly based on Convolutional Neural Networks (CNNs), surpassing traditional methods based on pictorial structures and deformable part models (TOSHEV; SZEGEDY, 2014).

Among the challenges in estimating 2D poses, we highlight the fact that images can have an unknown number of people, in different poses, scales and positions. The interactions that people may perform is another problem to deal with, such as contact (e.g., two people hugging), which can bring occlusions, thus making it difficult to detect and associate the parts of each body. Lastly, the computational cost tends to grow as the number of people increases in the image, making real-time performance a challenge (CAO et al., 2018).

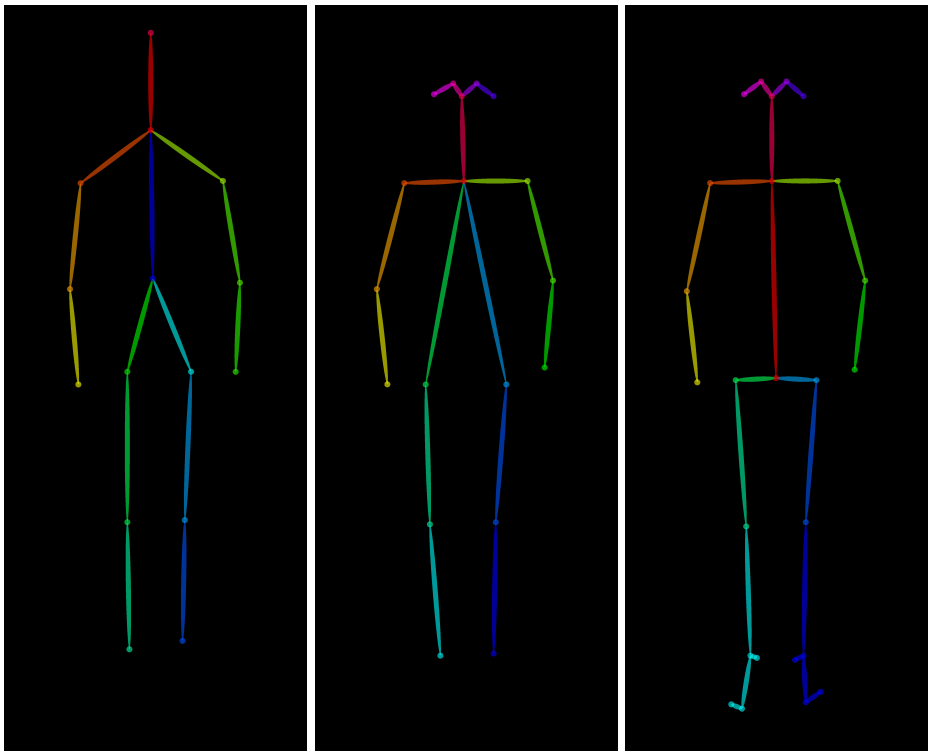
Toshev & Szegedy (2014) proposed a new method for human pose estimation based on Deep Neural Networks (DNNs). This method, called DeepPose, has leveraged a series of works that use deep learning for 2D pose estimation. In the last few years, 2D pose estimation methods have had significant improvements driven by the launch of the Microsoft Common Objects in Context (COCO) dataset (LIN et al., 2014), containing more than 200,000 images and 250,000 person instances labeled with joints keypoints.

Methods for human pose estimation can be categorized into two approaches: bottom-up and top-down. The first one estimates each body joint and then groups them to form a unique pose. The second one performs person detector and then estimates body joints within the detected

bounding boxes (TOSHEV; SZEGEDY, 2014).

With the release of new databases and techniques for estimating 2D poses from images and videos, several models to represent 2D poses have emerged. The most notable models are: MPI (INSAFUTDINOV et al., 2016), COCO (LIN et al., 2014) and BODY 25 (CAO et al., 2018). Figure 4 shows the main 2D body models.

Figure 4 – Main models for 2D pose representation.



(a) MPI (b) COCO (c) BODY 25

Source: The author.

Nowadays, methods that employ deep learning using a bottom-up approach in the task of estimating 2D poses stood out for images with scenes crowded of people, showing good accuracy and real-time processing. Among these methods, two are very promising for human action recognition tasks in videos, OpenPose (CAO et al., 2018) and PifPaf (KREISS et al., 2019).

2.2 OpenPose

OpenPose is an efficient method for multi-person 2D pose estimation with good performance on multiple public benchmarks. It uses a bottom-up approach of association scores via Part Association Fields (PAFs), a set of 2D vector fields that encode the location and orientation of limbs over the image domain. Cao et al. (2018) demonstrate that simultaneously inferring these bottom-up representations of detection and association encodes sufficient global

context for a greedy parse to achieve high-quality results at a fraction of the computational cost. Figure 5 presents an example of multi-person 2D pose estimation using OpenPose.

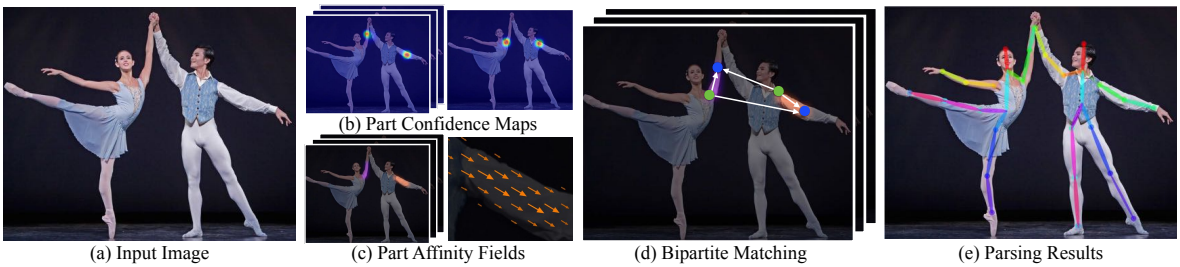
Figure 5 – OpenPose 2D pose estimation example.

(a) Original image (b) Multi-person 2D poses estimation

Source: The author. Image from COCO dataset (LIN et al., 2014).

In short, OpenPose receives as input a color image, as presented in Figure 6a, and as output produces the 2D body joints location and a confidence score for each person found in the image, as presented in Figure 6e. Initially, a CNN predicts a 2D confidence map for each body joint (Fig. 6b) and a 2D vector fields of part association fields (PAFs) for each limb, which represents the degree of association among body joints (Fig. 6c). In the end, the confidence maps and the PAFs are analyzed (Fig. 6d) to output the 2D pose, that includes 2D joints coordinates and a confidence score for all people detected in the image (CAO et al., 2018).

Figure 6 – OpenPose overall pipeline. (a) A color image input; (b) Confidence maps for joint detection; (c) PAFs for part association; (d) Set of matching to associate joints forming limbs; (e) The full-body 2D poses built for all people found in the image.



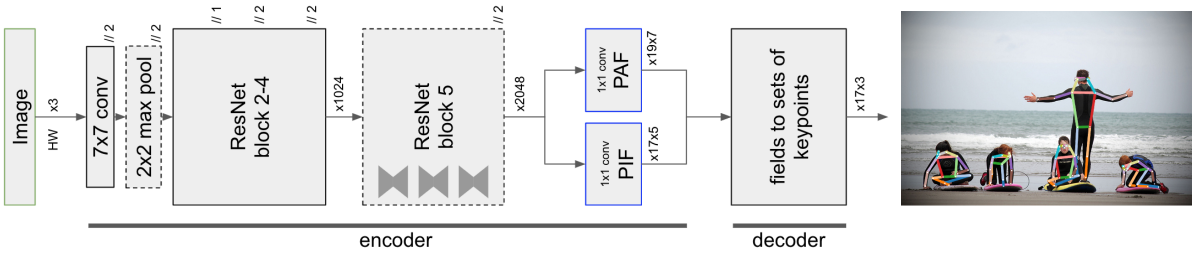
Source: Cao et al. (2018)

2.3 PifPaf

The goal of PifPaf method is estimating human poses in crowded images. The method tries to deal with challenges like low-resolution and partially occluded pedestrians. Techniques that use top-down approach usually fail when people are occluded by others where bound boxes overlap. In general, the bottom-up methods proposed before PifPaf are bounding-box free, but they still contain a coarse feature map localization. However, the PifPaf method is free of any grid-based constraint on the spatial localization of the joints and can detect multiples poses even when the occlusion happens. The main difference among PifPaf and OpenPose is that PifPaf go beyond scalar and vector fields to composite fields in the PAF module (KREISS et al., 2019).

Figure 7 presents the PifPaf architecture. PifPaf uses a shared Residual Network (ResNet) base with two head networks: one head network that has the goal of predicting the body joints (location, confidence, and size), which is called Part Intensity Field (PIF), and the other head network that outputs the associations between joints, called the Part Association Field (PAF), thus naming the method PifPaf.

Figure 7 – PifPaf Model Architecture. The input is a color image. The ResNet encoder output PIF and PAF fields. The decoder is a method that converts PIF and PAF fields into 2D poses estimation with 17 joints for each person found in the image. Each joint is represented by location and a confidence measure.



Source: Kreiss et al. (2019)

The first step is Part Intensity Fields (PIF) that detects and precisely locates human body parts. In order to do the fusion of a confidence map, a regression for key point detection is used. As a result of the PIF module, a joint structure is generated including the confidence measure, a vector that points to the closest body part and size of the joint. In Figure 8a for instance, is presented a confidence map for the left shoulders present in the image. In order to improve the location of confidence map, it is performed a fusion with the vectorial part of the PIF, as shown in Figure 8b, forming a confidence map with high-resolution.

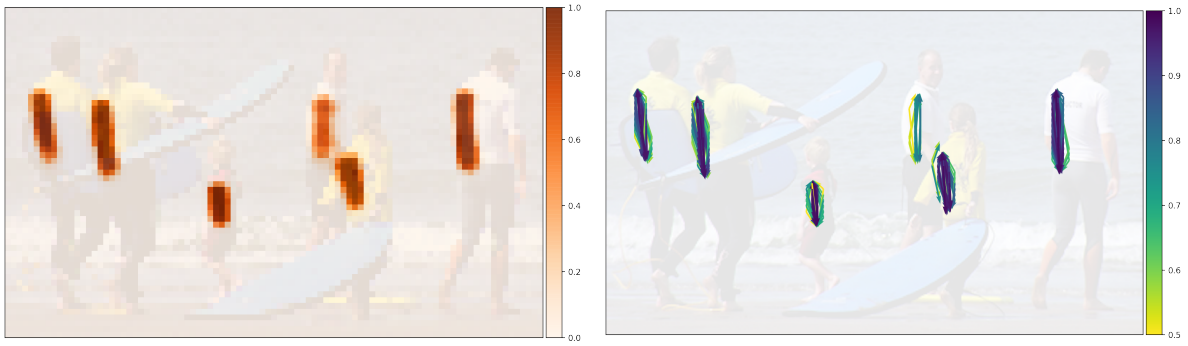
Figure 8 – PIF components for the left shoulder. (a) The confidence map; (b) The vector field; (c) The fused high-resolution components

(a) Original confidence map. (b) PIF vectorial part. (c) Fused confidence map.

Source: Kreiss et al. (2019)

The Part Association Fields (PAF) is the step to connect joint locations into 2D poses. For each joint, the PAF outputs: location, confidence score, two vectors to the two associated parts and two widths for the spatial precision. Figure 9 shows the PAF that associates left shoulder with left hips (KREISS et al., 2019).

Figure 9 – PAF that associates left shoulder with left hip. Each location of the feature map is the origin of two vectors which point to the shoulders and hips for association. (a) The confidence of associations; (b) The vector components.



(a) Confidence of associations. (b) Vector components.

Source: Kreiss et al. (2019)

Chapter 3
RELATED WORKS

In this chapter, we present some handcrafted and some deep learning-based methods related to our work.

3.1 Handcrafted Methods

Video understanding is a challenging task and during the last decade increased the interest in this research field. Many types of researches developed in this area are focused on extracting spatiotemporal features from videos. The most suitable methods that deal with handcrafted feature extraction from videos include those based on spatiotemporal interest points.

Laptev & Lindeberg (2003) were the first to propose the use of space-time interest points (STIPs) for automatic recognition of humans actions in videos. In order to detect spatiotemporal events, they used the idea of the Harris interest-point operators and detected local structures in space-time where the image values have significant local variations in both space and time. After that, they estimated the spatiotemporal extents of the detected events and extracted their scale-invariant spatiotemporal descriptors.

The concept of cuboids was presented by Dollar et al. (2005), where at each space-temporal interest point (local maxima of the response in the spatial and temporal domain), a cuboid is extracted, which contains the spatiotemporal windowed pixel values. Dalal et al. (2006) introduced a new motion-based descriptor for human detection in video. This descriptor is designed to capture the relative motion of different human limbs not suffering from background motions that can occur. They named this descriptor Motion Boundary Histograms (MBH). A 3-dimensional (3D) Scale-Invariant Feature Transform (SIFT) descriptor was presented by Scovanner et al. (2007) for video or 3D imagery such as Magnetic Resonance Imaging (MRI) data. They also show how this descriptor can better represent the 3D nature of video-data in the application of action recognition.

Klaser et al. (2008) presented a local descriptor for video sequences based on the success of Histograms of Oriented Gradients (HoG) descriptors for static images. In this method, it is

applied the key HoG concepts to 3D, so the video is treated as spatiotemporal volumes. This proposed descriptor is called Histograms of Oriented 3D Spatiotemporal Gradients (HoG3D).

All of these methods presented so far use different encoding schemes based on spatiotemporal interest points, histograms and pyramids to precompute the gradients on different temporal and spatial scales.

Another popular method is the improved Dense Trajectories (iDT) (WANG; SCHMID, 2013), that is an improvement of the previous version Dense Trajectories (DT) (WANG et al., 2011). This method is based on cuboid construction by using Bag-of-Visual-Words (BoVW) approach to do video classification, and works detecting interest points and obtaining their surroundings. Densely sampled points are tracked over some frames by using optical flow, so the spatial neighborhood at each position is appended to create a curving volume. The volumes are described using their Trajectory, Histograms of Oriented Gradients (HoG), Histograms of Optical Flow (HoF) and Motion Boundary Histograms (MBH), and then encoded in mid-level by using Fisher Vectors (FV) technique before the classification step, which is done using a linear Support Vector Machine (SVM) classifier. Even though iDT achieves excellent classification performance, it has an expensive computational cost and becomes intractable on datasets with a high number of videos. In some cases, the file with the descriptors computed from the video has a size larger than the original video file.

There are also methods that use body shape analysis, which has been the subject of some studies on the recognition of human actions in videos (JUNEJO; AGHBARI, 2012; ALCÂNTARA et al., 2013; RAJA et al., 2011; CHAARAOUI et al., 2013a; SINGH et al., 2010; CHEEMA et al., 2011; ALCANTARA et al., 2017; CHOU et al., 2018; SINGH; VISHWAKARMA, 2019). Those methods typically use silhouettes or poses for encoding human actions. In general, they use the centroids of the silhouettes and their representations are generated by using the distance value from the centroid to each silhouette point, using a radial scheme. The advantages of shape analysis are the simplicity and rich information in order to represent human actions. The disadvantage is that good poses and silhouettes can be difficult to acquire, relying mainly on background subtraction or frame difference, which may fail when parts of the body are occluded.

Some works proposed methods to recognize human actions using 3D skeleton information (VEMULAPALLI et al., 2014; PRESTI; CASCIA, 2016; AGAHIAN et al., 2020). Those methods usually represent poses using spatiotemporal features, for instance, skeleton coordinates and the motion pattern. The main limitation of using 3D skeleton information is the need to use 3D sensors like Microsoft Kinect. Usually, the 3D sensors are expensive and present some limitations, like working only in short distances and the difficulty to estimate 3D poses in crowded places.

3.2 Deep Learning Methods

Motivated by the excellent results achieved by approaches that use deep learning in still-image recognition tasks, driven by AlexNet (KRIZHEVSKY et al., 2012), by recent availability of powerful parallel computers (Cloud Computing, GPUs, CPU clusters), along with vast amounts of data available for training models, the interest in researches using Convolutional Neural Networks (CNNs) applied on automatic visual recognition tasks using videos has increased.

Karpathy et al. (2014) performed a study of multiple techniques for using CNN in the time domain to take advantage of the local spatiotemporal information. They analyze four strategies to use CNN to extract temporal information - **Single Frame**: it uses a single-frame baseline architecture to understand the contribution of static appearance to the classification accuracy; **Early Fusion**: it combines information across an entire time window immediately on the pixel level, implemented by modifying the filters on the first convolutional layer in the single-frame model by extending them to be of size $11 \times 11 \times 3 \times T$ pixels, where T is some temporal extent (they use $T = 10$); **Late fusion**: it uses two separate single-frame networks with shared parameters and then merges the two streams in the first fully connected layer; **Slow Fusion**: it is a balanced mix between the two approaches that slowly fuses temporal information throughout the network such that higher layers get access to progressively more global information in both spatial and temporal dimensions.

The concept of Two-Stream CNNs was introduced by Simonyan & Zisserman (2014a) and the authors achieved good results on action recognition. The main idea of Two-Stream approach is the incorporation of spatial and temporal networks. The spatial network extracts information from the appearance of individual frames and carries information about scenes and objects depicted in the video, while the temporal network extracts form of motion across the frames and conveys the movement of the camera and the objects.

Ji et al. (2013) proposed a 3D CNN for human action recognition. The idea is that the model can extracts features from both spatial and temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. Following the same line, Tran et al. (2015) proposed the use of segmented video volumes as inputs for a 3-convolution-layer 3D CNN to classify actions. Instead, this method takes full-video frames as inputs and does not rely on any preprocessing, thus easily scaling to a large number of videos.

In Tran et al. (2017), the authors assessed and discussed several forms of spatiotemporal convolutions for video analysis and studied their effects on action recognition. They showed that doing factorization of 3D convolution explicitly into two separate and successive operations, a 2D spatial convolution and a 1D temporal convolution, can add nonlinear rectification like ReLU between 2D and 1D convolution, thus doubling the number of nonlinearities compared to a 3D

CNN, but with the same amount of parameters to optimize, allowing the model to represent more complex functions. Further, the decomposition into two convolutions makes the optimization process easier, producing in practice less training and test losses. This method achieved results comparable or superior to the state-of-the-art on Sports1M, Kinetics, UCF101, and HMDB51 datasets.

Based on Two-Stream CNN proposes by Simonyan & Zisserman (2014a) that uses spatial and temporal networks, Carreira & Zisserman (2017a) introduced a new Two-Stream Inflated 3D CNN (I3D) based on 2D CNN inflation (filters and pooling kernels are expanded into 3D), making it possible to learn spatiotemporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters. The authors show that doing a pre-training on Kinetics dataset, I3D model can considerably improve upon the state-of-the-art in action classification.

Considering that the motion of body parts can characterize human actions over time, the use of Recurrent Neural Networks (RNNs) allows the description of human temporal dynamic behavior. Hochreiter & Schmidhuber (1997) proposed Long Short-Term Memory (LSTM) networks that are very successful extensions of the recurrent neural networks. LSTM networks utilize the gating mechanism over an internal memory cell to learn and describe a complex representation of long-term dependencies among the sequential input data, being proper for feature learning over a sequence of temporal data. Liu et al. (2016) presented a method to recognize human action using 3D skeleton data and LSTM networks with trust gates. The main focus of it is to use RNNs over the temporal domain for discovering the discriminative dynamics and body motion patterns for 3D action recognition.

Long-term Recurrent Convolutional Network (LRCN) model was introduced by Donahue et al. (2015). It blends a deep hierarchical visual feature extractor (CNN) with a model that can learn to understand and synthesize temporal dynamics for tasks involving sequential data (LSTM). Baradel et al. (2017) presented a method that uses Gated Recurrent Unit (GRU), that is an improvement to LSTM and was proposed by Cho et al. (2014). GRU can regulate the flow of information just like an LSTM does, but without using a memory unit and having fewer parameters and hence it may train a bit faster or need fewer data to generalize.

Chapter 4
MATERIAL AND METHODS

This chapter presents briefly the methods proposed in this thesis for human action recognition from videos, based on deep learning and 2D poses, as well as the datasets and the metrics used in the experiments. Complete descriptions of the proposed methods are presented in Chapters 5, 6 and 7.

4.1 Proposed Methods

At the beginning of our researches, in the order to answer the **Research Question 1**, presented in Section 1.3, which consists of evaluating techniques that use deep learning to classify human actions in videos, two distinct CNNs architectures that use spatiotemporal features, were chosen. Among the motivations for using spatiotemporal CNNs in the classification of human actions are the good results reported in the literature and the ease of using a unique end-to-end architecture, in which the extraction of features and classification of actions are carried out in a black-box.

The performances of these distinct CNN architectures were assessed on the dataset Pornography-800, proposed by Avila et al. (2013), and described in Section 4.2. This dataset was chosen due to the difficulty to classify pornography actions in videos. For instance, videos containing human actions performed by people wearing swimwear can be wrongly classified as pornography. In such cases, it is necessary to analyze not only the visual aspect of the scenes but also the temporal changes along the video frames, which may or may not characterize pornography. Chapter 5 describes the proposed methods based on CNNs architectures, which use spatiotemporal features, as well as the experimental results, discussions and conclusions.

Subsequently, aiming to answer the **Research Questions 2, 3, and 4**, presented in Section 1.3, which refer to the use of 2D poses for the recognition of human actions in videos, we first proposed a method, described in details in Chapter 6, that uses 2D poses without any pre-processing. Next, we developed a new method, based on an original approach to encode 2D poses into the parameter space of the line segments, also proposed in this thesis, as a way of improving the classification of human actions. This last method is described in details in

Chapter 7.

The proposition of methods based on 2D poses for human actions recognition from videos was motivated by questions regarding the privacy of the place and people performing the actions, by concerns regarding the huge computational power and huge training data requirements imposed by the use of deep learning approaches, and also by the very positive results obtained by state-of-the-art methods available in literature for estimating in real time 2D human poses from videos.

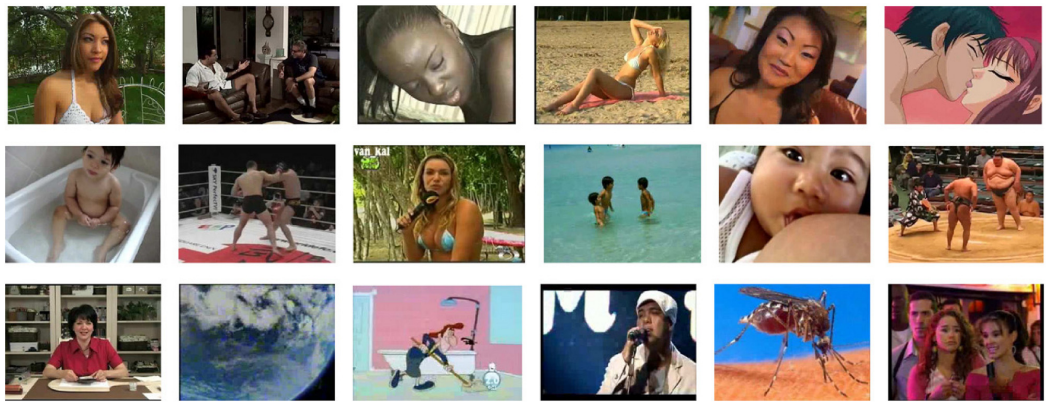
In order to assess the proposed methods based on 2D poses, we have used tree public datasets, KTH (LAPTEV et al., 2004), Weizmann (GORELICK et al., 2007) and Volleyball (IBRAHIM et al., 2016), since they are widely used as benchmarks by the scientific community in this area. These datasets are described in Section 4.2. The metrics used for the assessments are described in Section 4.3.

4.2 Video Datasets

In order to evaluate the proposed methods, the following four public datasets were selected, which contain videos with people performing certain actions: Pornography-800 (AVILA et al., 2013), KTH (LAPTEV et al., 2004), Weizmann (GORELICK et al., 2007) and Volleyball (IBRAHIM et al., 2016).

The Pornography-800 dataset (AVILA et al., 2013) was chosen to evaluate the 3D CNNs used in the present study. This dataset contains 800 videos, representing a total of 80 hours, which encompass 400 pornography videos and 400 non-pornography videos. Figure 10 shows some selected frames from a small sample of this dataset, illustrating the diversity and challenges posed.

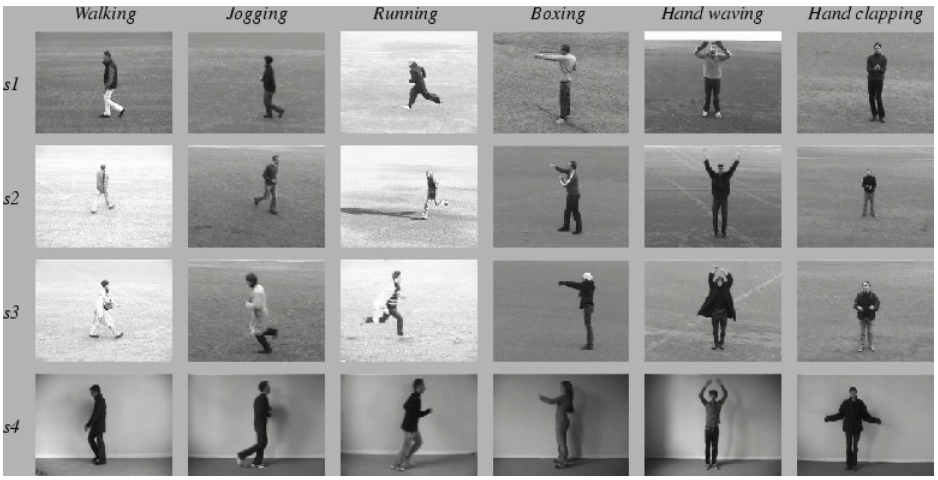
Figure 10 – Pornography-800 dataset. Top row: pornographic videos. Middle row: challenging cases of non-pornographic videos. Bottom row: easy cases of non-pornographic videos.



Source: Avila et al. (2013).

KTH dataset (LAPTEV et al., 2004) contains six classes of human actions (Walking, Jogging, Running, Boxing, Hand Waving and Hand Clapping) performed several times by 25 people in four different scenarios: outdoors (S1), outdoors with scale variation (S2), outdoors with different clothes (S3), and indoors (S4), as shown in Figure 11. The dataset contains 599 videos acquired in similar backgrounds with a static camera, summing up a total of 289,715 frames, 11,375.32 seconds, captured at 25 frames per second (FPS) and size of 160×120 pixels.

Figure 11 – Sample frames from the KTH dataset for six actions and four scenarios.

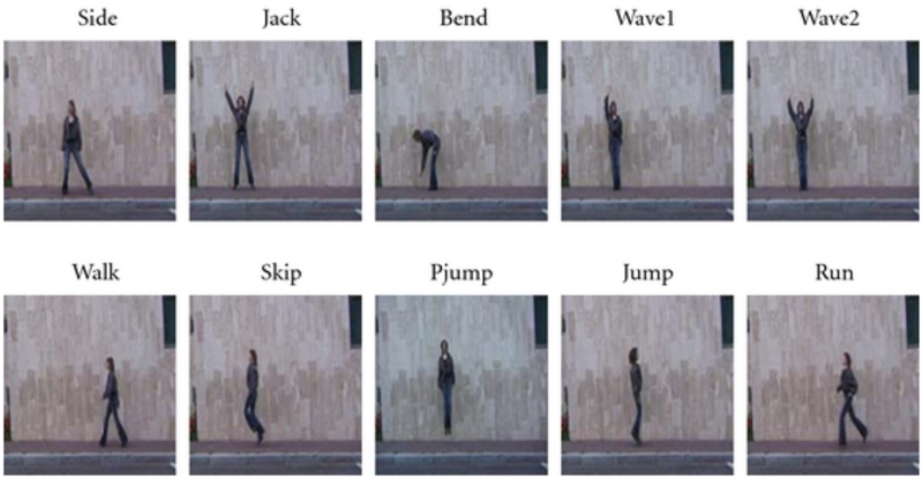


Source: Laptev et al. (2004).

Weizmann dataset (GORELICK et al., 2007) consists of 10 classes (Side, Jack, Bend, Wave1, Wave2, Walk, Skip, Pjump, Jump and Run), with nine actors performing each action, sometimes more than once, resulting in 93 videos. The dataset contains a total of 5,701 frames, 228.04 seconds, captured at 25 FPS and size of 180×144 pixels. All the actions occur on the same static background as shown in Figure 12.

The volleyball dataset (IBRAHIM et al., 2016), was recently available and it is one of few publicly available datasets for multi-person activity recognition that is relatively large-scale and contains labels for people locations, as well as their collective and individual actions. This dataset consists of 55 volleyball game videos, with resolution of 1920×1080 or 1280×720 pixels. The dataset contains 4830 labeled frames, where each player is annotated with a bounding box and labeled with one of the 9 individual actions (Waiting, Setting, Digging, Falling, Spiking, Blocking, Jumping, Moving and Standing) and the whole scene is assigned with one of the 8 collective activity labels (Right Set, Right Spike, Right Pass, Right Winpoint, Left Set, Left Spike, Left Pass and Left Winpoint), which define what part of the game is happening. For each annotated frame, there are multiple surrounding unannotated frames available (20 frames before the target frame, the target frame and 20 frames after the target frame). Figure 13 shows four selected frames from videos of this dataset illustrating four collective activities during a volleyball game (left pass, right pass, right spike and left winpoint), in which each player is inside a red bounding box and is labeled with his action. Figure 14 shows closer details of the

Figure 12 – Sample frames from the Weizmann dataset for ten actions.



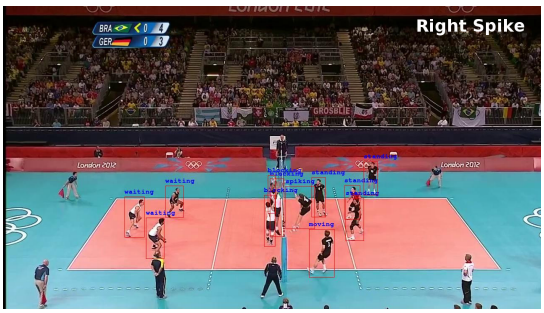
Source: Gorelick et al. (2007).

players' individual actions during a right spike.

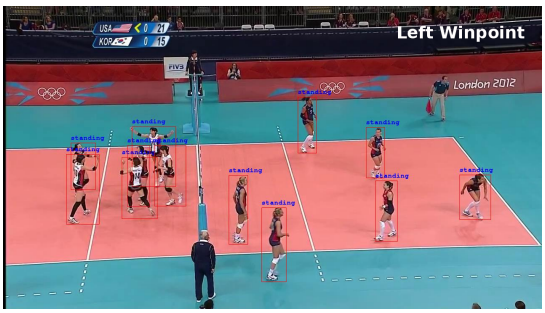
Figure 13 – Four frames from the Volleyball dataset, illustrating four collective activities during a volleyball game (left pass, right pass, right spike and left winpoint).

(a) Left pass

(b) Right pass



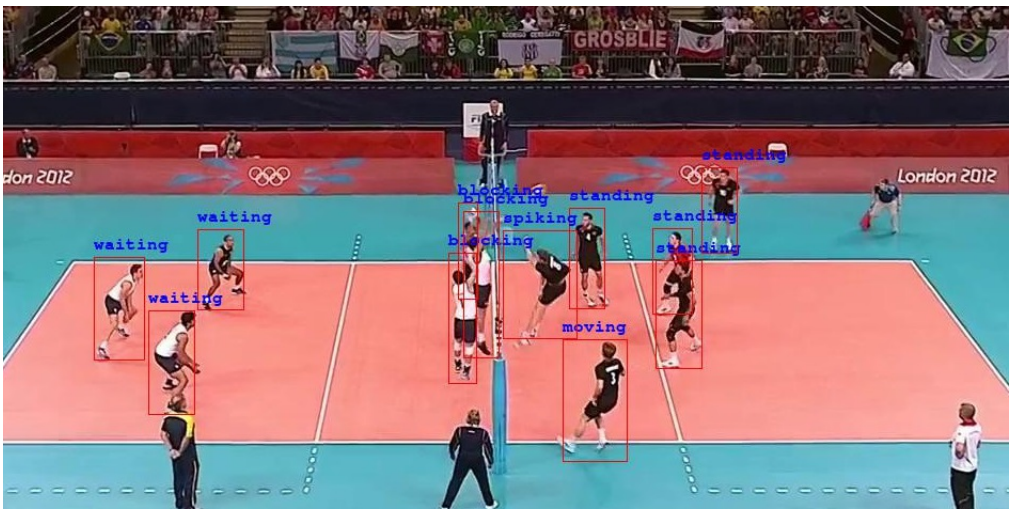
(c) Right spike



(d) Left winpoint

Source: The author. Images from Ibrahim et al. (2016).

Figure 14 – Closer details of the players' individual actions during a right spike in a frame from Volleyball dataset. Each player is inside a red bounding box and is labeled with his respective action in that frame.



Source: The author. Images from Ibrahim et al. (2016).

4.3 Metrics

In order to assess different methods objectively, we need to use a quantitative accuracy metric. Usually, for human action recognition, the standard is to use confusion matrices and recognition accuracy rates. From the confusion matrix it is possible to infer detailed results for each video class. The accuracy metric refers to the percentage of the correctly classified video in the test set (ZHANG et al., 2019). It is formulated in Equation 4.1:

$$Accuracy = \frac{\text{correct classified instances}}{\text{number total of instances}} \tag{4.1}$$

It is worth noting that there is a common practice in classification tasks that consists of dividing the database into p equal parts and performing cross-validation. This practice consists of dividing a dataset into p parts with approximately equal sizes, one part of the data being used for training and the other $p-1$ parts for testing. In this case, the accuracy is calculated for each part of the dataset division, P_i, as shown in Equation 4.2. To calculate the overall accuracy, we must compute the accuracy mean, using each part of the dataset division P_i, as shown in Equation 4.3.

$$ACC_{P_i} = \frac{\text{correct classified instances in the part } i}{\text{number total of instances in the part } i} \tag{4.2}$$

$$Accuracy = \frac{1}{p}\sum_{i=1}^{p} ACC_{P_i} \tag{4.3}$$

Besides evaluation metrics, we need to set up the validation protocol in order to compare

different models. Among the many cross-validation protocols, we highlight the most relevant for this work. Consider a dataset with n samples:

- **Predefined**: Usually used in large dataset, in this case the authors of dataset provide the samples that need to be used as train and test the model;
- **K-Fold**: This protocol has a single parameter, k, that refers to the number of groups that a given dataset is to be split into, at each k iteration, one part is taken as a test and the other $k - 1$ parts are joined into the training set;
- **Leave-One-Out**: Usually used in small databases, in this protocol, at each iteration a single sample was taken as the test sequence, while the other $n - 1$ samples were used to train the model, which was repeated for all video samples.

In order to maintain comparability with other literature methods, we choose the validation scheme adopted by the other studies being compared, and in each experiment described in the following chapters we mention the protocol used.

Chapter 5

SPATIOTEMPORAL CNNS FOR PORNOGRAPHY DETECTION IN VIDEOS

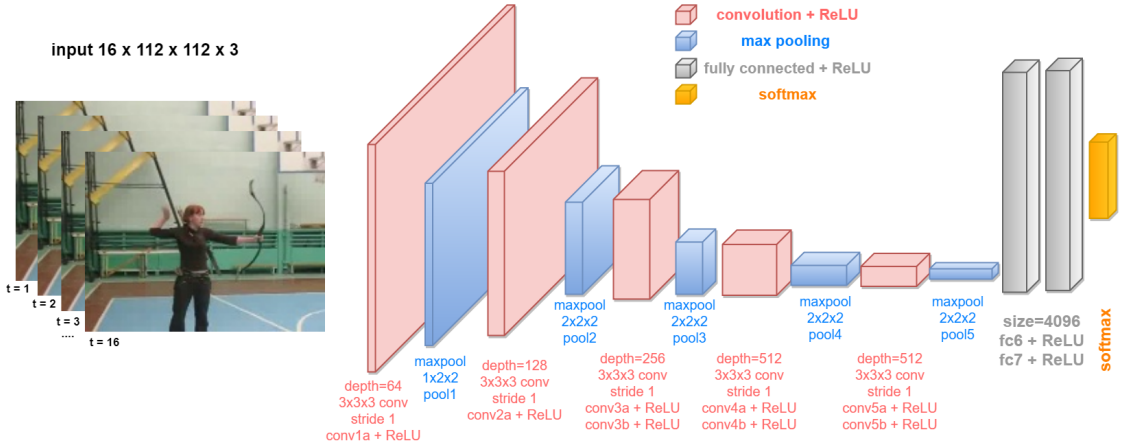
Motivated by the good results achieved recently by 3D Spatiotemporal Convolutional Neural Networks in the human action recognition from videos using large datasets such as Kinetics and Sports-1M, in this chapter, we assessed the power of representation and classification of these models in a pornography database. Although pornography does not fit into a typical problem of recognition of human actions, this problem shares some characteristics, such as the need to use space-time information to achieve good recognition rates, because, in the case of pornography, it is not enough only to evaluate spatial features such as the presence of human skin in the scenes, we also have to evaluate the movement pattern performed by people to characterize whether or not there is pornography in a video, this space-time characteristic is also very important for the recognition of human actions in videos. This chapter presents the initial results of this doctoral thesis, we evaluated the learning power of spatiotemporal-based Convolutional Neural Networks for the classification of videos containing pornography. Two spatiotemporal-based CNNs proposed in literature, VGG-C3D CNN (TRAN et al., 2015) and ResNet R(2+1)D CNN (TRAN et al., 2017), were used. To the best of our knowledge, this is the first study that uses 3D CNN to detect pornography in videos. The content of this chapter refers to the article **"Spatiotemporal CNNs for Pornography Detection in Videos"** published and presented at the 23th Iberoamerican Congress on Pattern Recognition (CIARP 2018), Madrid - Spain, November of 2018.

5.1 VGG-C3D CNN

Tran et al. (2015) realized that a homogeneous setting with convolution kernels of $3 \times 3 \times 3$ is the best option for 3D CNN (Convolutional Neural Networks). This is similar to the 2D CNNs proposed by Simonyan & Zisserman (2014b), which are also known as VGG (Visual Geometry Group). By using a dataset with a huge amount of data, it is possible to train a 3D CNN with $3 \times 3 \times 3$ kernels as deep as possible, due to the amount of memory available in current GPUs. The authors designed the 3D CNN to have 8 convolution layers, 5 pooling

layers, followed by two fully connected layers, and a softmax output layer. Fig. 15 shows the 3D Spatiotemporal CNN proposed by Tran et al. (2015) (called VGG-C3D in this thesis). The 3D convolution filters of VGG-C3D are of dimension $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$. In turn, the 3D pooling layers are $2 \times 2 \times 2$ with stride also of $2 \times 2 \times 2$, except for pool1 which presents kernel size of $1 \times 2 \times 2$ and stride $1 \times 2 \times 2$ with the intention of preserving the temporal information at the early phase. Each fully connected layer has 4,096 output units.

Figure 15 – VGG-C3D architecture based on VGG-11 (SIMONYAN; ZISSERMAN, 2014b) proposed by (TRAN et al., 2015).



Source: The author.

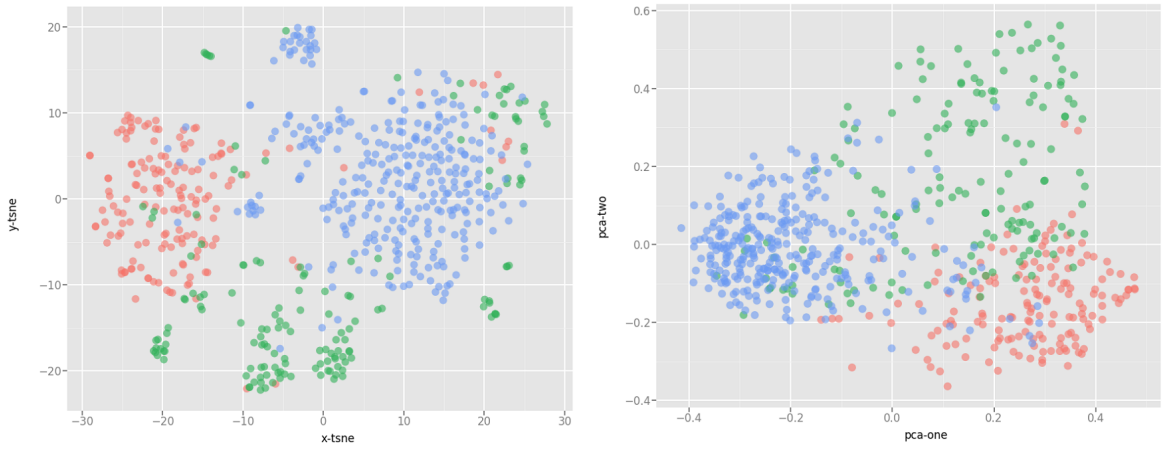
The model provided by the authors (TRAN et al., 2015), pre-trained on the Sports-1M dataset (KARPATHY et al., 2014) in train split, was used in the present study. Sports-1M was created by Google Research and Stanford Computer Science Department and contains 1,133,158 videos of 487 sports classes. Since Sports-1M has many long videos, five 2-seconds long clips were randomly extracted from every training video. The clips were then resized to have a frame size of 128×171. During the training phase, the clips were randomly cropped into $16 \times 112 \times 112$ crops for spatial and temporal jittering, and horizontally flipped with 50% probability. The training was done by Stochastic Gradient Descent (SGD) with a minibatch size of 30 examples. The initial learning rate was of 0.003 and was divided by 2 every 150K iterations. The optimization was stopped at 1.9M iterations (about 13 epochs).

After training, the VGG-C3D was used as a feature extractor. In order to extract features, a video needs to be split into clips with 16 frames in length. For the present study, clips with an 8-frame overlap between two consecutive clips were used. After that, the clips were submitted to the VGG-C3D to extract **fc6** activations. As each video may have an arbitrary number of clips, to generate only one descriptor for each video the **fc6** activations were averaged to form a 4,096-sized descriptor, followed by L2-normalization.

In order to evaluate the VGG-C3D features extracted from the Pornography-800 dataset (AVILA et al., 2013) through this transfer learning approach, **fc6** features were extracted from all clips and then projected to 2D space using the t-Distributed Stochastic Neighbor Embedding

(t-SNE) (MAATEN; HINTON, 2008b) (Fig. 16a) and Principal Component Analysis (PCA) (Fig. 16b). It is worth noting that no fine-tuning was conducted to verify if the model showed good generalization capability across the datasets. Fig. 16 shows that the pornography and non-pornography classes are very clearly clustered in their own subgroups, although videos from the pornography and difficult non-pornography classes of Pornography-800 dataset presented some overlapping. This visualization suggests that if we use a classification algorithm we could probably get good accuracy rates.

Figure 16 – Feature embedding visualizations of VGG-C3D on samples from Pornography-800 dataset: pornography (blue), easy non-pornography (red) and difficult non-pornography (green). (a) Using t-SNE and (b) Using PCA.



(a) (b)

Source: The author.

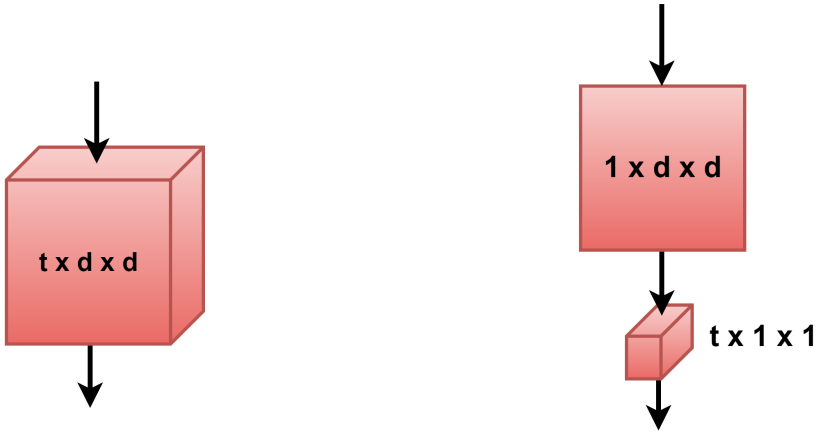
5.2 ResNet R(2+1)D CNN

Recent studies have indicated that replacing 3D convolutions by two operations, a 2D spatial convolution and a 1D temporal convolution, can improve the efficiency of 3D CNN models (TRAN et al., 2017; XIE et al., 2017). Tran et al. (2017) designed a new spatiotemporal convolutional block, R(2+1)D, that explicitly factorizes 3D convolution into two separate and successive operations, a 2D spatial convolution and a 1D temporal convolution. Using this architecture, we can add nonlinear rectification like ReLU between 2D and 1D convolution. This would double the number of nonlinearities compared to a 3D CNN, but with the same number of parameters to optimize, allowing the model to represent more complex functions. Moreover, the decomposition into two convolutions makes the optimization process easier, producing in practice less training loss and less test loss.

Another method proposed by Xie et al. (2017) showed that replacing 3D convolutions with spatiotemporal-separable 3D convolutions makes the model 1.5x more computationally efficient (in terms of FLOPS) than 3D convolutions.

Experiments performed by Tran et al. (2017) demonstrated that ResNets adopting homogeneous (2+1)D blocks in all layers, achieved state-of-the-art performance on both Kinetics (KAY et al., 2017) and Sports-1M datasets. Spatiotemporal decomposition can be applied to any 3D convolutional layer. An illustration of this decomposition is given in Fig. 17 for the simplified setting, where the input tensor contains a single channel.

Figure 17 – 3D convolution *vs* (2+1)D convolution. (a) Full 3D convolution using a filter of the size $t \times d \times d$, where t denotes the temporal extent and d is the spatial width and height. (b) A (2+1)D convolutional block, where a spatial 2D convolution is followed by a temporal 1D convolution.



(a) (b)

Source: Tran et al. (2017).

The architecture proposed by Tran et al. (2017) was applied in the present study. This relatively simple structure was based on deep residual networks, which have shown good performance. Table 1 presents details of R(2+1)D architecture.

Experiments were conducted using a model that had been pre-trained on Kinetics dataset. Since we used the R(2+1)D network with a softmax layer to perform classification, a transfer learning technique was applied to fine-tune the model on the Pornography-800 dataset. The R(2+1)D network used had 34 layers and videos frames were resized to 128×171, with each clip generated by randomly cropping 112×112 windows. A total of 32 consecutive frames were randomly sampled from each video applying temporal jittering during the process of fine-tuning.

Although Pornography-800 has only about 640 training videos in each split, epoch size was set at 2,560 for temporal jittering considering 4 clips for each training video per epoch. This setup was chosen to optimize the training time since the videos have different sizes.

Batch normalization was applied to all convolutional layers and mini-batch size was

Table 1 – R(2+1)D architecture (TRAN et al., 2017) used in the present study.

layer name	output size	34-layer
conv1	$L \times 5 \times 56$	$3 \times 7 \times 7, 64$, stride $1 \times 2 \times 2$
conv2	$L \times 56 \times 56$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 3$
conv3	$\frac{L}{2} \times 28 \times 28$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 4$
conv4	$\frac{L}{4} \times 14 \times 14$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 6$
conv5	$\frac{L}{8} \times 7 \times 7$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 3$
	$1 \times 1 \times 1$	spatiotemporal pooling, fc layer with softmax

set to 4 clips due to GPU memory limitations. The initial learning rate was set to 0.0001 and divided by 10 every 2 epochs, while the process of fine-tuning was conducted in 8 epochs. In the classification phase, the videos were split into 32-frame long clips. ResNet R(2+1)D CNN was used on clips with 16 frames that overlap between two consecutive clips to extract features and for softmax classification. Each video can have an arbitrary number of clips, so average pooling on softmax probabilities was conducted to aggregate predictions over clips to obtain video-level prediction.

5.3 Experiments and Results

The Pornography-800 dataset (AVILA et al., 2013) was chosen to evaluate the 3D CNNs used in the present study. The VGG-C3D network evaluated in the present study was developed using Caffe (Convolutional Architecture for Fast Feature Embedding) (JIA et al., 2014) and the ResNet R(2+1)D architecture was developed using Caffe2[1]. All experiments were run on a computer with an Intel Xeon E5-2630 v3 2.40GHz processor, 32 GB RAM and a NVIDIA Titan XP GPU with 12 GB of memory. The results presented are the mean value obtained from the 5 splits of the Pornography-800 dataset using 5-fold-cross-validation protocol (640 videos in training set and 160 in the test set on each fold, which is the same protocol proposed by Avila et al. (2013)).

Table 2 (two first rows) shows the accuracy of both approaches: VGG-C3D with a Linear SVM classifier and ResNet R(2+1)D CNN with softmax classifier. The VGG-C3D architecture with a linear SVM classifier achieved a better performance, with accuracy of 95.1%, while with

[1] https://caffe2.ai/

the ResNet R(2+1)D architecture using the softmax classifier achieved accuracy of 91.8%.

The results reported in the literature by others state-of-the-art method are also presented in Table 2. It is possible to observe that the CNN-based methods, including our proposed method based in 3D-CNNs, outperform all methods based on Bag-of-Visual-Words (BoVW).

Table 2 – Results achieved by VGG-C3D and ResNet R(2+1)D on the Pornography-800 dataset and results obtained by state-of-the-art methods for pornography video detection on the Pornography-800 dataset.

Approach	**Reference**	**Year**	**Accuracy(%)**
VGG-C3D + Linear SVM	our work	2018	95.1% ± 1.7
ResNet R(2+1)D CNN + Softmax	our work	2018	91.8% ± 2.1
BoVW-Based	Avila et al. (2011)	2011	87.1% ± 2.0
	Valle et al. (2011)	2011	91.9% ± NA
	Souza et al. (2012)	2012	91.0% ± NA
	Avila et al. (2013)	2013	89.5% ± 1.0
	Caetano et al. (2014)	2014	90.9% ± 1.0
	Caetano et al. (2016)	2016	92.4% ± 2.0
	Moreira et al. (2016)	2016	95.0 % ± 1.3
2D CNN RGB	Moustafa (2015)	2015	94.1% ± 2.0
	Perez et al. (2017)	2017	97.0% ± 2.0
2D CNN OF	Perez et al. (2017)	2017	95.8% ± 2.0
Two Stream CNN	Perez et al. (2017)	2017	**97.9% ± 1.5**

5.4 Conclusions

The experimental results obtained on the Pornography-800 dataset showed that the spatiotemporal CNNs, VGG-C3D and ResNet R(2+1)D, proposed in this thesis for pornography detection in videos, performed better than all methods based on Bag-of-Visual-Words (BoVW) assessed in this study. Moreover, these spatiotemporal CNNs were competitive with other CNN-based approaches assessed in this study. To the best of our knowledge, this is the first study to use 3D CNN to detect pornography in videos.

With recent creation and availability of large video databases, along with the evolution of GPUs, we believe that 3D CNNs will be able to achieve even better results in video understanding tasks (including human action recognition from videos), similar to what happened with the launch of AlexNet. A strong evidence of this is that a 3D CNN model, I3D proposed by Carreira & Zisserman (2017b), has recently reached the best result in the Kinetics database.

Chapter 6

HUMAN ACTION RECOGNITION IN VIDEOS BASED ON ANGLES AND TRAJECTORIES FROM 2D POSES

Aiming to develop techniques for human action recognition from videos with low cost and conventional computer resources, with a relatively small set of training data and preserving the privacy of people and places involved in the actions, we focused our efforts in proposing and developing new methods and new descriptors for human action recognition in videos based on spatiotemporal information obtained from 2D poses. This chapter presents our first method developed to this purpose, and refers to the article **"Human Action Recognition using 2D Poses"**, published and presented at the 8th Brazilian Conference on Intelligent Systems (BRACIS 2019), Salvador - Brazil, October of 2019.

6.1 Proposed Method

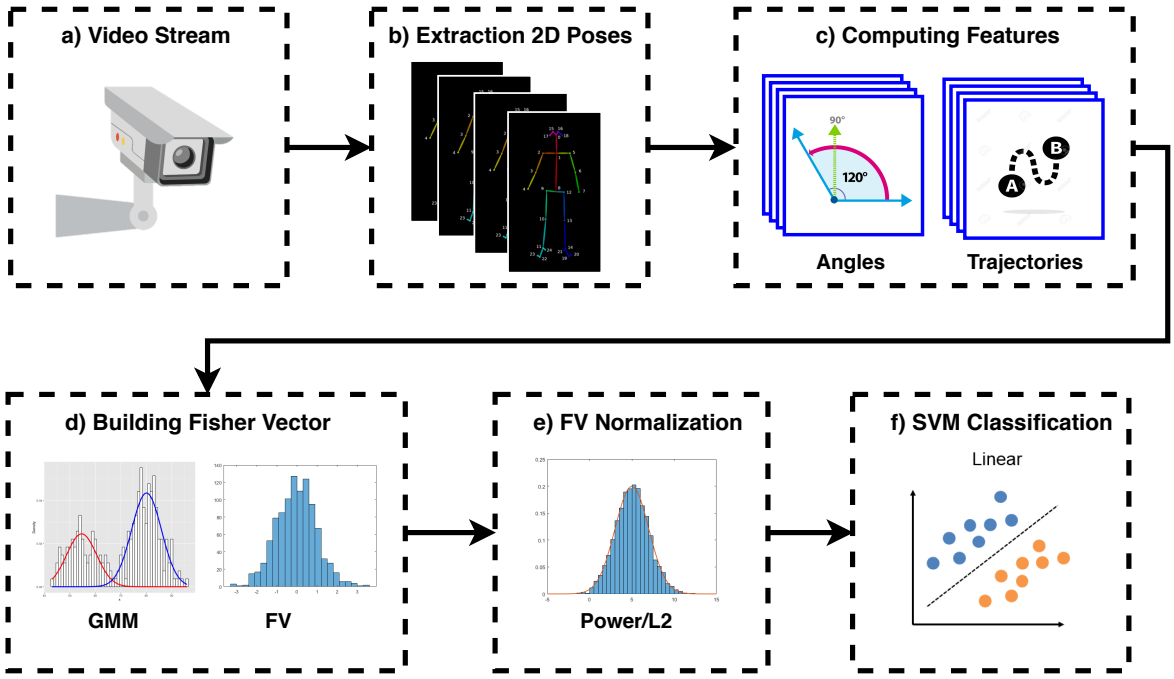
Our first proposed method extracts features from 2D human poses obtained from videos by using the OpenPose framework (CAO et al., 2018)[1]. From the 25 key points obtained from each video frame, 15 key points (0 to 14) (see Figure 2 in Chapter 1) are used in our work to calculate the pose descriptors, which are based on the angles between two adjacent straight line segments defined by three key points, and on the trajectories of all key points along L frames of the video. The main steps of our method, proposed to compute the pose descriptors and to classify the pose as a predefined action class, are shown in Figure 18.

6.1.1 Skeleton Angles

In our method, we use as descriptors 14 angles calculated from some specific parts of the 2D human skeleton generated by the OpenPose framework, as shown in Table 3.

[1] It is noticed that there are other techniques for detecting 2D poses (PifPaf, for instance), and those techniques can easily replace OpenPose in our method.

Figure 18 – Illustration of the main steps of our human action recognition method.



Source: The author.

Each skeleton part is represented by a vector $\vec{v}$ defined by two key points, $p_i = (x_i, y_i)$ and $p_j = (x_j, y_j)$, according to equation 6.1:

$$\vec{v} = (x_j - x_i, y_j - y_i) \tag{6.1}$$

Then, the angle θ formed by two adjacent skeleton parts, represented by the vectors $\vec{v}$ and $\vec{u}$, is calculated according to equation 6.2:

$$\theta = \arccos\left((\vec{v} * \vec{u})/(||\vec{v}||||\vec{u}||)\right), \tag{6.2}$$

where $||\vec{v}||$ means the length of vector $\vec{v}$ and the operator $*$ is the dot product of two vectors.

Therefore, for each frame, a feature vector with 14 angles formed by adjacent body parts is generated:

$$Angles = (\theta_1, \theta_2, ..., \theta_{14}) \tag{6.3}$$

Appendix A presents in detail an example of how the angles feature vector is calculated from a 2D pose.

6.1.2 Key Joint Points Trajectories

Motivated by Wang et al.(WANG et al., 2011), who used a descriptor of trajectories for densely sampled points of interest, we used a trajectory descriptor for the key points of the

Table 3 – The 14 angles calculated between adjacent parts of the human skeleton.

Angle	**Part 1**	**Part 2**
01	Main Body	Left Shoulder
02	Main Body	Right Shoulder
03	Main Body	Left Hip
04	Main Body	Right Hip
05	Left Shoulder	Neck
06	Right Shoulder	Neck
07	Left Forearm	Left Arm
08	Left Arm	Left Shoulder
09	Right Forearm	Right Arm
10	Right Arm	Right Shoulder
11	Left Thigh	Left Hip
12	Left Thigh	Left Leg
13	Right Thigh	Right Hip
14	Right Thigh	Right Leg

skeleton detected on the video. The structure of the trajectory of one key point P that defines a skeleton joint describes the motion pattern of such skeleton joint.

Given a trajectory of length L, we encode its shape in a sequence:

$$T = (\Delta P_t, ..., \Delta P_{t+L-1}) \tag{6.4}$$

of displacement vectors $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$.

The resulting vector is normalized by the sum of the magnitudes of the displacement vectors:

$$T' = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} ||\Delta P_j||} \tag{6.5}$$

Equations 6.4 and 6.5 are calculated for all 15 points (0 to 14 Fig. 2) and L frames to form a feature vector of trajectories:

$$Trajectories = (T'_1, T'_2, ..., T'_{15}) \tag{6.6}$$

6.1.3 Feature Encoder Fisher Vector

We use the Fisher Vector (FV) to encode low-level features (Angles and Trajectories) in mid-level features. The FV can be used as a generic framework which combines the benefits of generative and discriminative approaches. In the context of image/video classification, FV has

shown to extend the popular Bag-of-Visual-Words (BoVW) by going beyond statistical counting (PERRONNIN et al., 2010).

While BoVW encodes the zero-order statistics of the distribution of descriptors by counting the number of occurrences of visual-codewords, the FV extends the BoVW by encoding the average first and second order differences between the descriptors and visual-codewords.

Fisher Vector (PERRONNIN et al., 2010) encodes both first and second order statistics between the 2D skeleton descriptors and a Gaussian Mixture Model (GMM). We set the number of Gaussians to $K = 20$ and sample all features from the training set to estimate the GMM. Each video is, then, represented by a $K + 2DK$ dimensional Fisher Vector for each descriptor type (Angles and Trajectories), where D is the descriptor dimension, similar to (KRAPAC et al., 2011). Finally, we apply power and L2 normalization to the Fisher Vector, as in (PERRONNIN et al., 2010). Aiming to combine both descriptors, we concatenate their normalized Fisher Vectors. Finally, a linear SVM[2] is used for classification.

6.2 Experiments and Results

During the experiments, an Intel XEON(R) CPU E5620 @2.40GHZ with 16 cores, 40GB of RAM and TITAN XP GPU was used. The 2D pose extraction was performed using OpenPose (CAO et al., 2018) which was coded in C++ with Caffe framework (JIA et al., 2014) and ran on TITAN XP GPU. The feature extraction described in Section 6.1 was coded in Python, while the Fisher Vector encoder and the classification was written in Python using some functions from Scikit-learn (PEDREGOSA et al., 2011) without parallel computing. The code for all steps performed is available in GitHub[3].

Two public datasets were used to evaluate our method: KTH (LAPTEV et al., 2004) and Weizmann (GORELICK et al., 2007), both presented in Chapter 4.

6.2.1 Features Embedding

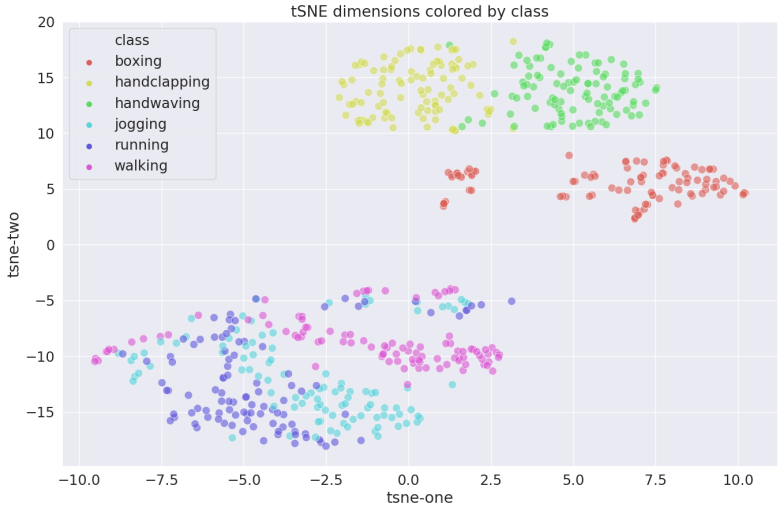
In order to assess the representation power of the features extracted from the two datasets (KTH and Weizmann), the features were extracted from all videos and then projected into a 2D space using the t-SNE (MAATEN; HINTON, 2008a) and PCA (JOLLIFFE, 2011). To plot the 2D representation, we use the Fisher Vector obtained from the concatenation of Angles and Trajectories, as described in Section 6.1.3.

Figures 19 and 20 show the feature embedding from KTH dataset. Figure 19 presents the t-SNE and Figure 20 presents the PCA. One can see that the KTH classes are clearly clustered in

[2] Other classifiers were used: Radial SVM, Polynomial SVM, k-Nearest Neighbors (kNN), Decision Tree, Gaussian Naive Bayes and Optimum-Path Forest (OPF), however the linear SVM presented the best result.

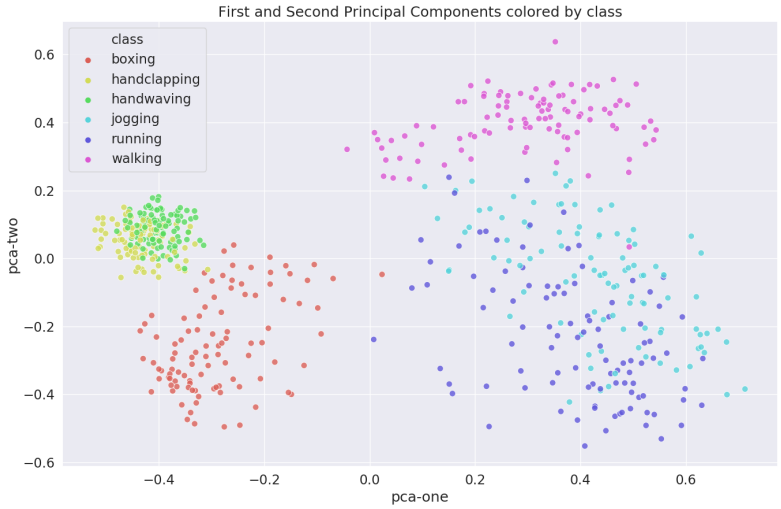
[3] https://github.com/murilovarges/HumanActionRecognition2DPoses

Figure 19 – t-SNE features embedding visualizations of the KTH dataset (Perplexity=100).



Source: The author.

Figure 20 – PCA features embedding visualizations of the KTH dataset.



Source: The author.

their own subgroups, although videos from Jogging and Running classes of the KTH dataset presented some overlapping.

The Weizmann dataset contains fewer samples (93) compared to KTH (599), thus making it easier to classify, as shown in Figures 21 and 22. Figure 21 presents the t-SNE and Figure 22 presents the PCA.

Figure 21 – t-SNE features embedding visualizations of the Weizmann dataset (Perplexity=100).

Source: The author.

Figure 22 – PCA features embedding visualizations of the Weizmann dataset.

Source: The author.

6.2.2 Classification

In order to evaluate and compare the proposed method, we used the Leave-One-Out-Cross-Validation (LOOCV) protocol as the validation technique. In LOOCV at each iteration, a single sample is taken as the test sequence, while the other $n - 1$ samples are used to train the model, which is repeated for all video samples.

In order to compute the trajectory descriptor (Subsection 6.1.2), we need to set L (trajectory length) and W (sampling step size) parameters. Based on some experiments, we

achieved the best results setting $L = 20$ and $W = 10$ (for Weizmann dataset $W = 1$).

Table 4 shows the results of the two descriptors presented in this work, their concatenation for the KTH and Weizmann datasets, and a comparison with state-the-art-methods. It is worth noting that only methods that used the LOOCV protocol were presented to avoid the divergences that other protocols can cause in the accuracy of each technique. The results show, for both datasets, that using the fusion between Angles and Trajectories, we can achieve better results.

Table 4 – Accuracy rates (%) for KTH and Weizmann datasets.

Method	**Year**	**Dataset**	
		KTH	**Weizmann**
FV (Angles + Trajectories)	2019	**95.33**	**97.85**
FV (Angles)	2019	94.32	87.10
FV (Trajectories)	2019	78.96	76.34
Zhang & Tao (2012)	2012	93.50	93.87
Junejo & Aghbari (2012)	2012	-	88.60
Chaaraoui et al. (2013b)	2013	-	90.32
Guo et al. (2013)	2013	98.50	100
Ravanbakhsh et al. (2015)	2015	95.60	-
Doumanoglou et al. (2016)	2016	88.70	-
Alcantara et al. (2017)	2017	92.20	100
Almeida et al. (2017)	2017	96.80	-
Carmona & Climent (2018)	2018	97.50	98.80
Chou et al. (2018)	2018	90.58	95.56
Singh & Vishwakarma (2019)	2019	94.50	97.66

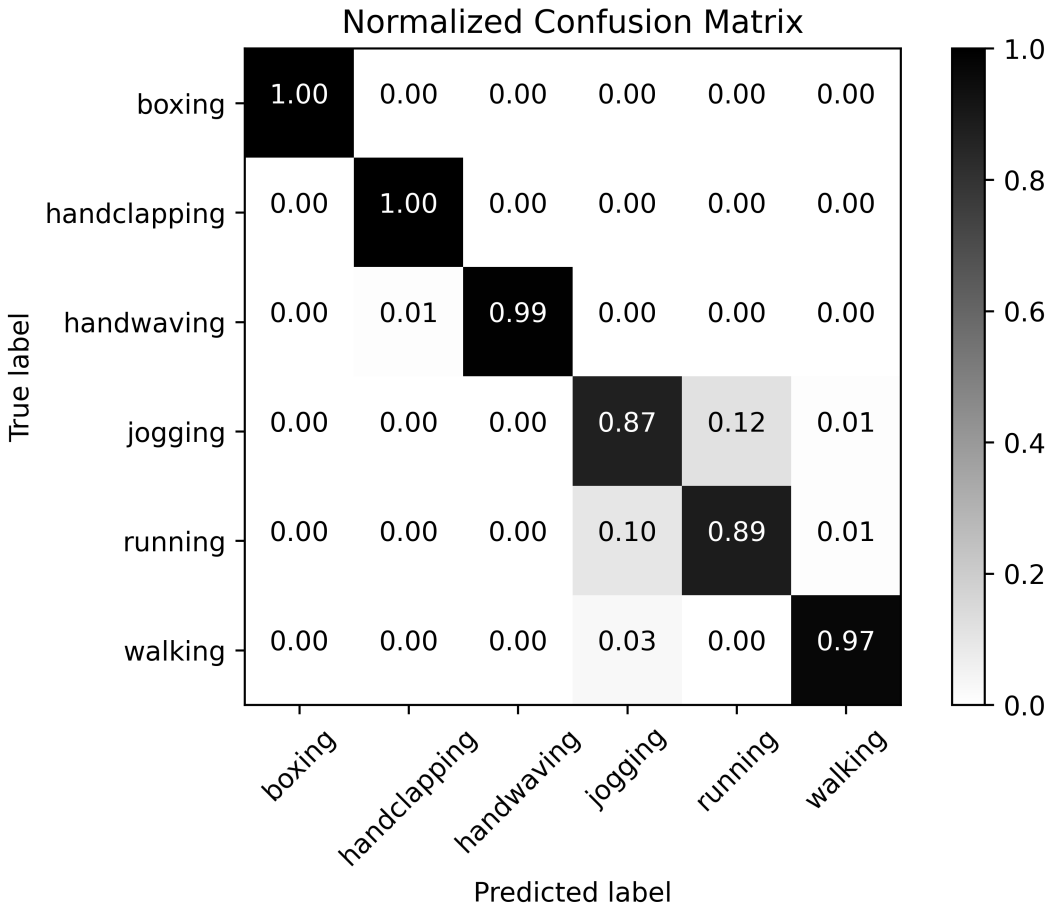
Figure 23 shows the confusion matrix for the KTH dataset by using the fusion of Angles and Trajectories that achieved 95.33% of accuracy. It is possible to notice that the errors occur mainly between the classes Running and Jogging, which have the same spatial pattern and a small temporal difference (speed of the action).

The KTH is a challenging dataset to our method since some classes (Walking, Jogging, Running) present the same spatial pattern, then the approach needs to accurately represent the movement pattern to separate those classes in the classification phase. Despite of it, our method achieved excellent classification accuracy compared to state-of-the-art methods.

The confusion matrix for the Weizmann dataset is shown in Figure 24. The results presented are for the fusion of Angles and Trajectories that achieved 97.85% of accuracy.

The Weizmann dataset contains small video sequences. Thus, we can use only a few

Figure 23 – Confusion Matrix for FV (Angles + Trajectories descriptors) with 95.33% of accuracy on KTH Dataset.



Source: The author.

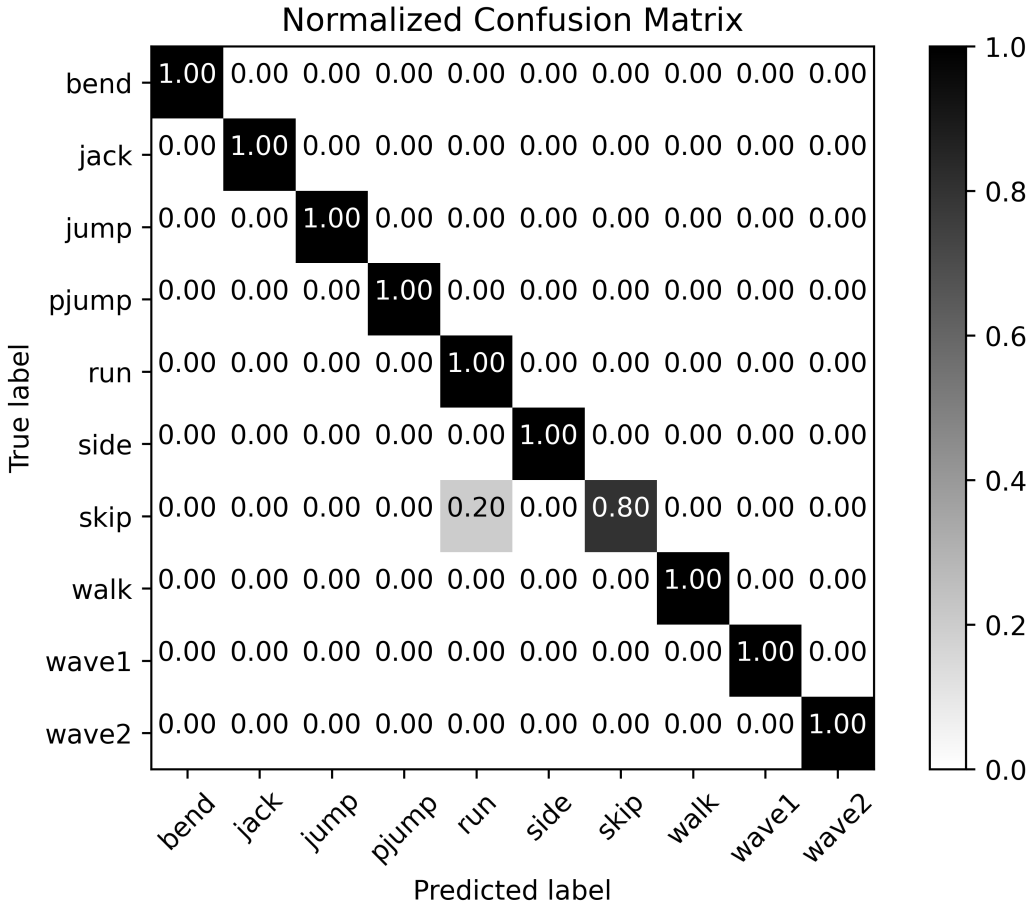
videos for training. However, our method was effective to represent samples and provided excellent classification accuracy compared to state-of-the-art methods.

6.3 Conclusion

Human actions recognition in video is a challenging problem. Consequently, the development of a robust method that deals well with any possible action and environment is also a challenge.

This chapter presented a new method to recognize human action in videos by combining information of angles formed by adjacent human skeleton parts and trajectories of skeleton key points (that define skeleton parts) across frames. Our descriptors are easier and lighter to compute comparing to other state-of-the-art methods. Our method is based on 2D poses and there are nowadays some methods that achieve good results in the extraction of 2D poses with real-time processing speed, such as OpenPose (CAO et al., 2018) and PifPaf (KREISS et al., 2019). Thus,

Figure 24 – Confusion Matrix for FV (Angles + Trajectories descriptors) with 97.85% of accuracy on Weizmann Dataset.



Source: The author.

due to the simplicity of our technique, it can be used in applications that require real-time processing with low computational costs. The results obtained are competitive compared to more sophisticated and complex state-of-the-art methods, like those that rely on dense trajectories (WANG et al., 2013).

Chapter 7

HUMAN ACTION RECOGNITION IN VIDEOS BASED ON SPATIOTEMPORAL FEATURES AND BAG-OF-POSES

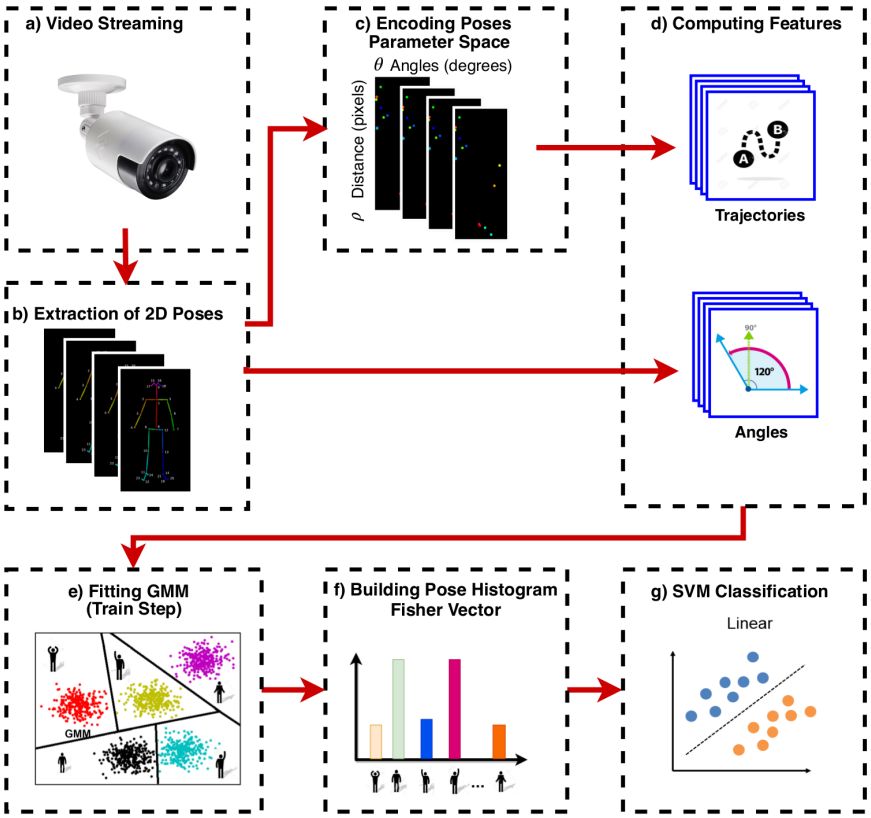
This chapter presents a new method of representing 2D poses. Instead of directly using the straight-line segments, as we did in the method proposed in Chapter 6, the 2D poses are mapped to the straight line parameter space, where each component part of the skeleton corresponds to a point. Then, from the parameter space, spatiotemporal features are extracted and encoded using a Bag-of-Poses approach. Experiments on three well-known public datasets showed that the proposed method using 2D poses encoded into the parameter space can improve the recognition rates, obtaining competitive accuracy rates compared to state-of-the-art methods. The content of this chapter refers to the content of the article **"Human Action Recognition in Videos Based on Spatiotemporal Features and Bag-of-Poses"** published in the Applied Soft Computing Journal, as well as some new results obtained with experiments carried out on an additional dataset (Volleyball dataset).

7.1 Proposed Method

The overall pipeline of our proposed method for action recognition in video based on spatiotemporal features extracted from 2D pose and encoded into the (θ, ρ) parameter space is shown in Figure 25.

The proposed method extracts features from 2D human poses obtained from videos by using the OpenPose framework (CAO et al., 2018). We use the pose model "BODY 25", which represents a body pose with 25 key joint points. Out of the 25 key joint points obtained from each video frame, only 15 are used in our study to calculate spatiotemporal pose descriptors. They were chosen because they are related to the limbs and trunk of the human body. The discarded joint points are related to elements from the face and feet that do not contain relevant information for the recognition of human actions.

Figure 25 – Overall pipeline of proposed method. (a) Our method takes the entire video streaming as input to class prediction, (b) 2D human poses extraction frame-by-frame. (c) 2D poses encoding into the (θ, ρ) parameter space, (d) Computing features (Angles and Trajectories), (e) Fitting GMMs, (f) Computing Fisher Vector, and (g) Video classification using a linear SVM classifier.



Source: The author.

In short, as shown in Figure 25 and detailed in subsections 7.1.1 to 7.1.3, the proposed method consists of the following steps:

(a) A video is obtained and submitted to analysis;

(b) 2D human poses are extracted from each frame of the input video;

(c) 2D human poses are encoded into the parameter space (θ, ρ);

(d) Features are computed from 2D poses (Angles) and parameter space (Trajectories);

(e) GMMs that represent the pattern of 2D poses are fitted (only in the training stage);

(f) The input video is represented by a codebook using a Fisher Vector (Bag-of-Poses);

(g) Finally, the video is classified using an SVM linear classifier.

7.1.1 Features from 2D Poses

The first set of features used by our method, Angles, are computed from the 2D human poses obtained from videos by using the OpenPose framework (CAO et al., 2018) without doing any preprocessing, as presented in Chapter 6 (Subsection 6.1.1 - Skeleton Angles).

7.1.2 Features from the Parameter Space

The second set of features used by our method, Trajectories, is obtained from 2D poses encoded into the straight line (θ, ρ) parameter space. After encoding each of the 14 skeleton parts (straight line segments) into the parameter space, the temporal features are then computed.

7.1.2.1 Encoding 2D Poses into the Parameter Space

In general, a straight line can be represented using a general equation:

$$y = mx + c \tag{7.1}$$

where, m is the gradient (slope) and (0, c) are the coordinates of the y-intercept. Thus, each straight line can be represented as a point (m, c) in the parameter space. However, the vertical lines represent a problem, since the slope parameter m would rise to unbounded values. So, for computational reasons, we use the Hesse normal form given by Equation:

$$\rho = x\cos\theta + y\sin\theta \tag{7.2}$$

where ρ is the distance of the straight line to the origin of the coordinate system, and θ is the angle between the x-axis and the perpendicular to the straight line. Figure 26 shows an example of a straight line segment represented using Hesse normal form (red dotted line represents the left forearm of the skeleton).

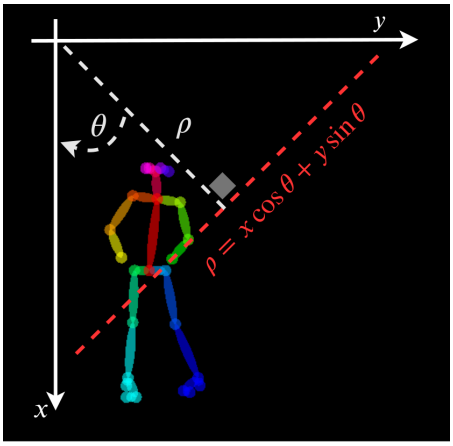
In our method, since each skeleton part is represented by two joint points, $p_i = (x_i, y_i)$ and $p_j = (x_j, y_j)$, in order to convert those body parts to the parameter space, we need to find two parameters, θ and ρ. By using this parameterization, every point (x, y) on the straight line will satisfy Equation 7.2.

Through Equation 7.2, we can find the parameter θ (angle) of the straight line defined by two points $p_i = (x_i, y_i)$ and $p_j = (x_j, y_j)$ by using Equation 7.3:

$$\theta = \arctan\left(\frac{(x_j - x_i)}{(y_i - y_j)}\right) \tag{7.3}$$

By using the parameter θ (angle) found with Equation 7.3, we can find the parameter ρ (distance) by using θ and coordinates (x, y) of point p_i or p_j in Equation 7.2. Applying those

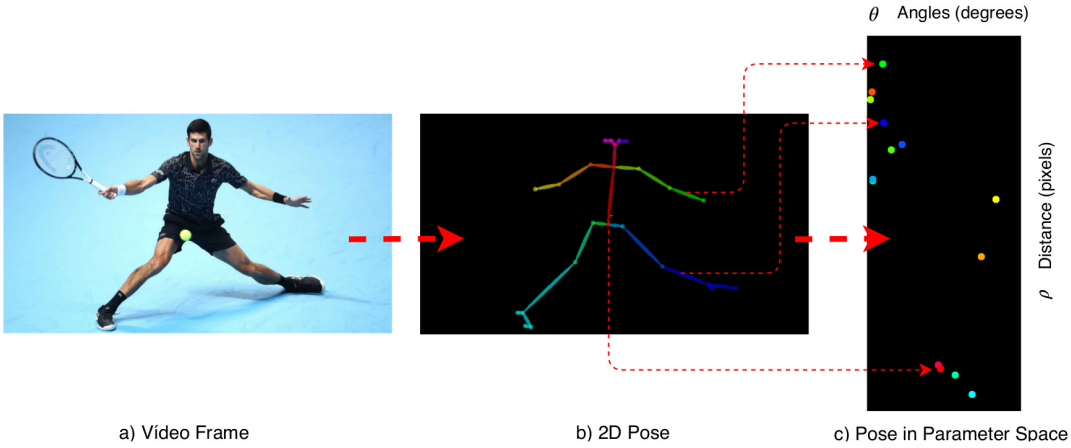
Figure 26 – Example of a straight line represented using Hesse normal form.



Source: The author.

equations to the straight lines that represent each body part, we can find their respective θ and ρ values in the parameter space. In this way, Equations 7.3 and 7.2 are applied to the 14 body parts to encode a 2D pose into the parameter space. After that, a min–max normalization (ADEYEMO et al., 2018) is applied to the feature vector. Figure 27 shows the process used to represent 14 body parts into the parameter space.

Figure 27 – Example of a 2D pose encoded into the (θ, ρ) parameter space. a) Original video frame, b) 2D pose extracted from the video frame and c) Pose encoded into the straight line parameter space.



Source: The author.

The motivation for encoding 2D poses into the parameter space is that in the parameter space, each human body limb becomes a point, so the features extracted from this new space are able to better represent the movement pattern of the human skeleton (e.g., trajectory calculation over frames). A video frame sequence is presented in Figure 28, where the 2D pose is extracted and, then, this pose is encoded into the parameter space. Figure 28 shows that the 2D pose changes reflect in the movement of the points in the parameter space.

Figure 28 – Example of 2D pose extraction and encoding into the (θ, ρ) parameter space in a video frame sequence. The first column presents original video frames, the second column presents 2D poses extracted, and the third column presents 2D poses encoded into the straight line parameter space.

Source: The author.

7.1.2.2 Body Parts in Parameter Space Trajectories

By using 2D poses encoded into the parameter space it is possible to compute the trajectory descriptor. The proposed trajectory descriptor is similar to the presented previously in Chapter 6, where we computed trajectories for human joint key points. In this method, since each point in parameter space represents a limb or body part, the descriptor can represent more information about the pattern of the human body motion over time. The structure of the trajectory of one point $P = (\rho, \theta)$ that defines a limb or body part in parameter space describes the motion pattern of such a body part in the image-space.

Given a trajectory of length L, we encode its shape in a sequence:

$$T = (\Delta P_t, ..., \Delta P_{t+L-1}) \tag{7.4}$$

of displacement vectors $\Delta Pt = (P_{t+1} - P_t) = (\rho_{t+1} - \rho_t, \theta_{t+1} - \theta_t)$.

The resulting vector is normalized by the sum of the magnitudes of the displacement vectors:

$$T' = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} ||\Delta P_j||} \tag{7.5}$$

The equations 7.4 and 7.5 are repeated for 14 points that represent the 14 body parts (limbs) in the parameter space and L frames to form a feature vector of trajectories:

$$Trajectories = (T'_1, T'_2, ..., T'_{14}) \tag{7.6}$$

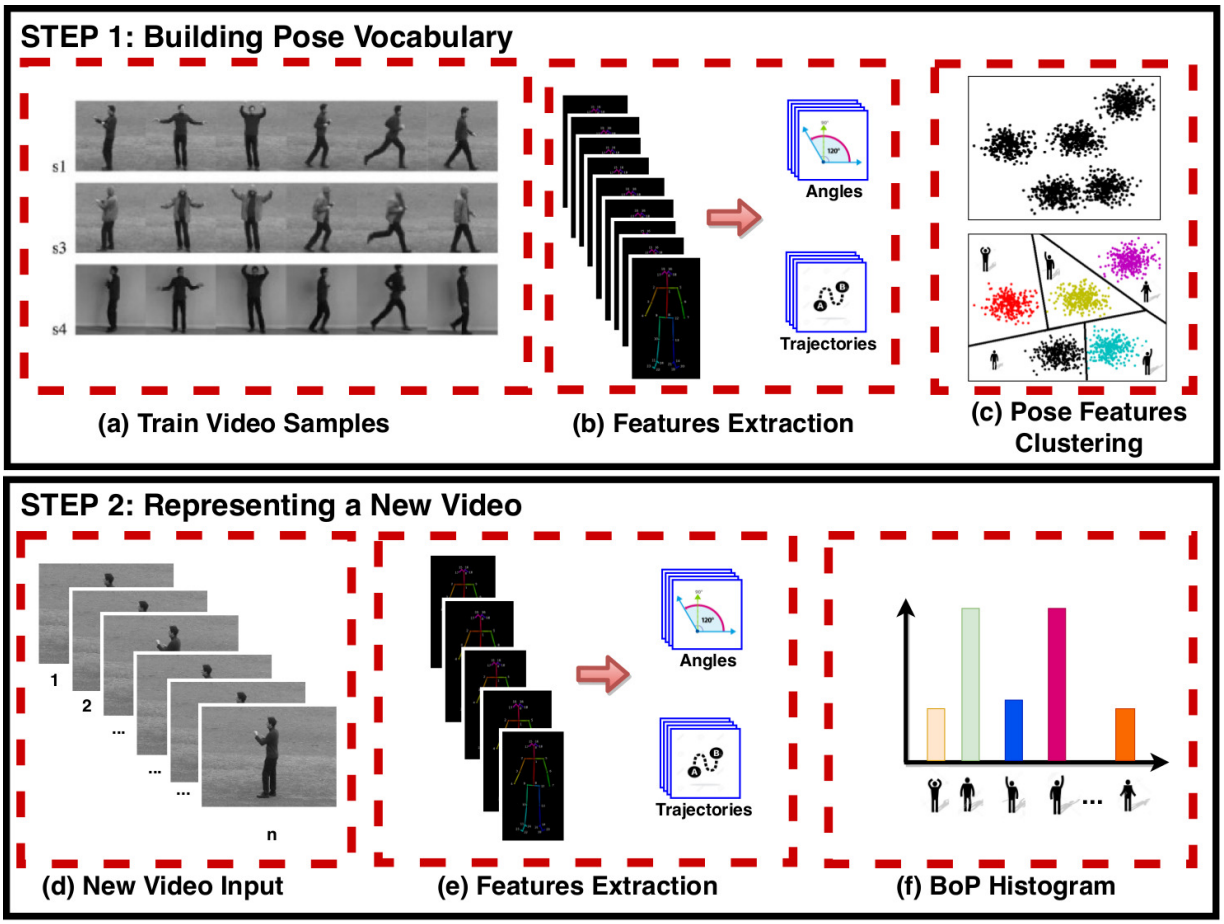
7.1.3 Bag-of-Poses for Mid-Level Feature Encoding

Some works introduced the Bag-of-Poses (BoP) framework for action recognition using skeleton information (SEIDENARI et al., 2013; AGAHIAN et al., 2019). Those works use the assumption that any action can be represented by a set of predefined spatiotemporal poses. In our method, in order to encode low-level spatiotemporal features computed from 2D poses across the videos frames in mid-level features, we also use the Bag-of-Poses framework. Figure 29 presents the main steps required to encode features using BoP.

We used the Fisher Vector (FV) to encode low-level features in mid-level features. The FV can be used as a generic framework which combines the benefits of generative and discriminative approaches. In the context of image/video classification, FV has shown to extend the popular Bag-of-Visual-Words (BoVW) by going beyond statistical counting (PERRONNIN et al., 2010).

While BoVW encodes the zero-order statistics of the distribution of descriptors by counting the number of occurrences of visual-codewords, the FV extends the BoVW by encoding the average first- and second-order differences between the descriptors and visual-codewords.

Figure 29 – Illustration of the Bag-of-Poses (BoP) steps of our human action recognition method using 2D poses.



Source: The author.

In our method, Fisher Vector (PERRONNIN et al., 2010) encodes both first- and second-order statistics between the 2D skeleton descriptors and a Gaussian Mixture Model (GMM). To build a GMM, we need to set the number of Gaussians (K parameter). As in our method each Gaussian represent a cluster of similar skeleton poses or skeleton motion pattern, for spatial and temporal descriptors respectively, we performed a search of the best K parameter testing values between 10 and 30 with a step size of 5. In our experiments, setting the number of Gaussians to 20 presented the best accuracy for spatial and temporal descriptors.

Table 5 shows some studies that used BoVW or BoP approach to perform human action recognition in videos, in this table is presented the K parameter used in each study. For example Wang et al. (2013) used dense trajectories features and the BoVW approach with the $K = 4000$, so in our method, we decrease the code book size by 200 times still keeping good performance. Agahian et al. (2019) used 3D poses obtained from depth sensors and used a BoP approach setting the number of key poses between 100 and 160 to obtain the best performance. So, in our method, we decrease the code book size from 4 to 8 times. It is worth remembering that this parameter is very sensitive to the dataset used and to the number of classes of human actions to

be classified, so this parameter must be adjusted for each database (AGAHIAN et al., 2019).

Table 5 – Number of key poses (K parameter) used by other methods.

Method	**K Parameter**
Wang & Schmid (2013)	4000
Agahian et al. (2019)	100:160
Carmona & Climent (2018)	256

We set the number of Gaussians to 20 ($K = 20$) and sampled all features from the training set to estimate the GMM. Each video was represented by a $K + 2DK$ dimensional Fisher Vector for each descriptor type (Angles and Trajectories), where D is the descriptor dimension, similarly to described by Krapac et al. (2011). A power and L2 normalization was applied to the Fisher Vector, as used by Perronnin et al. (2010) and similar to the approach proposed by Wang & Schmid (2013), in order to increase the precision, we initialized the GMM 10 times and keep the result with the lowest classification error.

In order to fuse different types of descriptors, we concatenate their normalized Fisher Vectors. Lastly, a linear SVM is trained and used for video classification.

7.2 Experiments and Results

In our experiments, an Intel XEON(R) CPU E5620 @2.40GHZ with 16 cores, 40GB of RAM and TITAN XP GPU was used. The 2D pose extraction was performed using OpenPose (CAO et al., 2018), which was coded in C++ with the Caffe framework (JIA et al., 2014) and used GPU. The 2D pose encoding into the parameter space and the feature extraction described in Section 7.1 was coded in Python. The Fisher Vector encoder and the classification were written in Python using some functions from Scikit-learn (PEDREGOSA et al., 2011) without parallel computing. The codes for all proposed steps are available in GitHub[1].

Three public datasets were used to evaluate our method: KTH (LAPTEV et al., 2004) and Weizmann (GORELICK et al., 2007), which contain only one person per video, and Volleyball (IBRAHIM et al., 2016), which contains multiple persons per video, as previously presented in Chapter 4.

7.2.1 Features Embedding

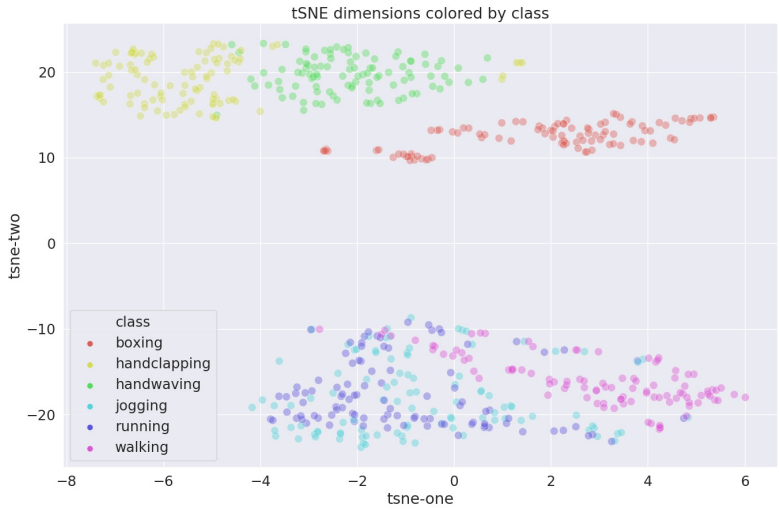
Aiming to evaluate the power of representation of the proposed features, we used, initially, two datasets (KTH and Weizmann). First, the features were extracted from all videos and then projected to 2D space using the t-Distributed Stochastic Neighbor Embedding

[1] https://github.com/murilovarges/HARBoP/

(t-SNE) (MAATEN; HINTON, 2008a) and Principal Components Analysis (PCA) (JOLLIFFE, 2011). The 2D representation uses the Fisher Vector acquired from the concatenation of all features for KTH and Weizmann datasets, as described in Section 7.1.3.

Figures 30 and 31 show the feature embedding from KTH dataset. Figure 30 presents the t-SNE and Figure 31 presents the PCA. One can see that the KTH classes are clearly clustered in their own subgroups, although videos from Jogging and Running classes of the KTH dataset presented some overlapping.

Figure 30 – t-SNE features embedding visualizations of the KTH dataset (Perplexity=100).



Source: The author.

Figure 31 – PCA features embedding visualizations of the KTH dataset.

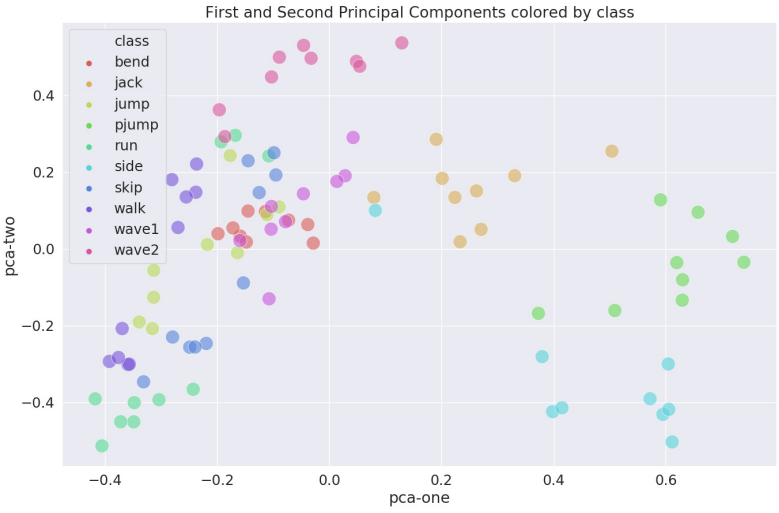
Source: The author.

The Weizmann dataset is a small dataset and contains ten classes, four more than KTH. However, classes are easier to classify, as shown in Figures 32 and 33. Figure 32 presents the t-SNE and Figure 33 presents the PCA visualization.

Figure 32 – t-SNE features embedding visualizations of the Weizmann dataset (Perplexity=100).

Source: The author.

Figure 33 – PCA features embedding visualizations of the Weizmann dataset.



Source: The author.

7.2.2 Classification of KTH and Weizmann Datasets

In order to evaluate and compare the proposed method, we used the Leave-One-Out-Cross-Validation (LOOCV) protocol as the validation technique. In

LOOCV, at each iteration, a single video sample is taken as the test sequence, while the other $n - 1$ video samples are used for training the model, which is repeated for all video samples.

To compute the trajectory descriptor (Subsections 7.1.2.2), we need to set L (trajectory length) and W (sampling step size) parameters. Based on some previous works (WANG; SCHMID, 2013; WANG et al., 2013; CARMONA; CLIMENT, 2018) and experiments performed, we achieved the best results setting L = 20 for both datasets, and W = 10 for the KTH dataset and W = 1 for Weizmann dataset. Because Weizmann contains few videos, with short video sequences and setting a small value to W helps to increase the training data.

Table 6 shows the results using the two descriptors presented in this chapter and their combination for the KTH and Weizmann datasets. The results show that, for both datasets, by encoding 2D poses in the space parameter we can improve the performance of classification, wherein by only using trajectory features we can improve accuracy from 78.96% to 93.66% in the KTH dataset and from 76.34% to 84.95% in the Weizmann dataset compared to the method proposed in Chapter 6.

Another important aspect perceived is that the fusion between Angles and Trajectories features, again improved the results for the KTH dataset achieving the result of 97.16% of accuracy. This improvement suggests that the spatial and temporal features are complementary.

Table 6 – Accuracy rates (%) for KTH and Weizmann datasets of our proposed method.

Method	**Dataset**	
	KTH	**Weizmann**
BoP (Angles + Trajectories)	**97.16**	**97.85**
BoP (Angles)	94.32	87.10
BoP (Trajectories)	93.66	84.95

Table 7 shows the best results obtained by our methods (Chapters 6 and 7) for the KTH and Weizmann datasets and a comparison with other methods available in literature. It is worth noting that only methods that used the LOOCV protocol were presented to avoid the divergences that other protocols can cause in the accuracy of each technique. The results show that for KTH and Weizmann datasets the proposed methods achieved good results compared to state-of-the-art methods. Although some methods present a slight superior accuracy, they usually use a more complex set of features.

Figure 34 shows the confusion matrix for the KTH dataset by using the fusion of Angles and Trajectories features that achieved 97.16% of accuracy. It is possible to observe that the errors occur mainly between the classes Running and Jogging that have the same spatial pattern and a small temporal difference (speed of the action).

The KTH is a challenging dataset to our method since some classes (Walking, Jogging,

Table 7 – Accuracy rates (%) for KTH and Weizmann datasets.

Method	**Approach**	**Year**	**Dataset**	
			KTH	**Weizmann**
BoP (Angles + Trajectories) (Chapter 7)	Pose	2020	97.16	97.85
FV (Angles + Trajectories) (Chapter 6)	Pose	2019	95.33	97.85
Zhang & Tao (2012)	Local Features	2012	93.50	93.87
Junejo & Aghbari (2012)	Pose	2012	-	88.60
Ji et al. (2013)	3D CNN	2013	90.20	-
Chaaraoui et al. (2013b)	Pose	2013	-	90.32
Guo et al. (2013)	Local Features	2013	98.50	100
Ravanbakhsh et al. (2015)	2D CNN	2015	95.60	-
Doumanoglou et al. (2016)	Local Features	2016	88.70	-
Alcantara et al. (2017)	Shape Analysis	2017	92.20	100
Almeida et al. (2017)	Local Features	2017	96.80	-
Carmona & Climent (2018)	Local Features	2018	97.50	98.80
Chou et al. (2018)	Silhouette	2018	90.58	95.56
Singh & Vishwakarma (2019)	Silhouette	2019	94.50	97.66
Moreira et al. (2020)	Shape Analysis	2020	97.50	98.00

Running) present the same spatial pattern, and thus the approach needs to accurately represent the movement pattern to separate them in the classification phase. Even so, our method achieved good accuracy compared to state-of-the-art methods as presented in Table 7.

The confusion matrix for the Weizmann dataset is shown in Figure 35. The results presented are for the fusion of Angles and Trajectories features that achieved 97.85% of accuracy.

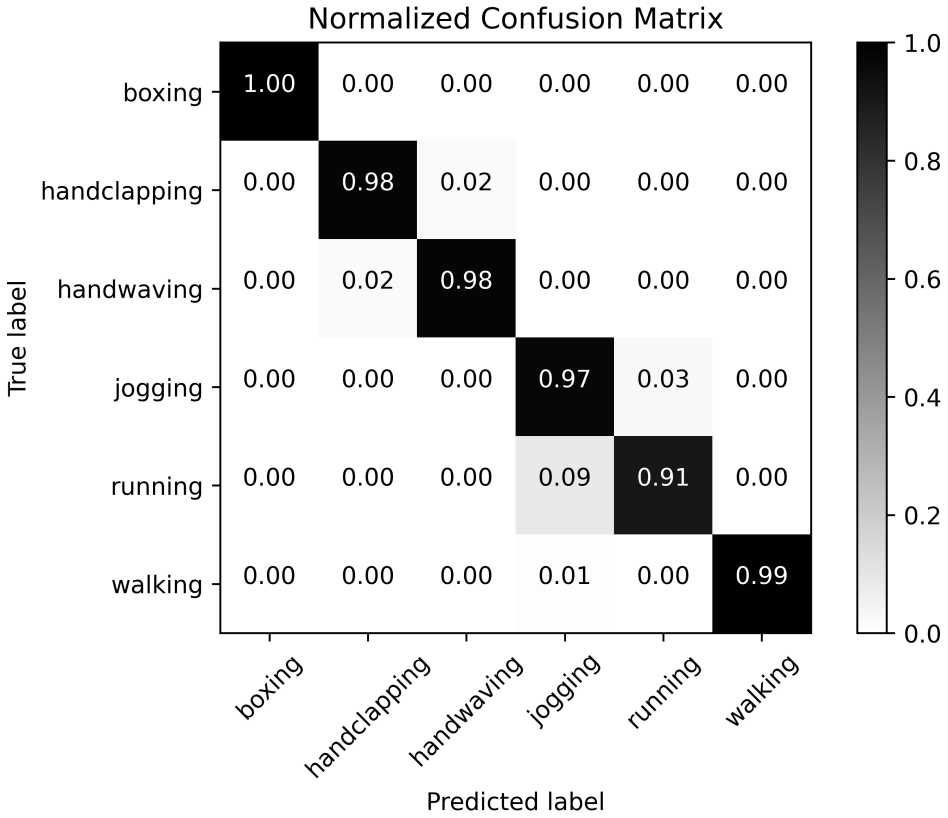
The Weizmann dataset contains few videos and small video sequences, thus we can use only a small number of videos for training. However, our method was effective in representing samples and provided excellent classification results compared to the other methods.

7.2.3 Classification Volleyball Dataset

With the purpose of assessing the proposed method in an uncontrolled environment with multiple-people, occlusions, viewpoint changes, and dynamic actions performed by people, we tested the proposed method in the Volleyball Dataset. The protocol used was based on the predefined dataset train-test split, which was performed by the authors at video level, rather than at frame level so that it makes the evaluation of models more convincing (IBRAHIM et al., 2016).

In the Volleyball Dataset, each clip contains 41 frames, but only the 21st frame is annotated with the players' bounding boxes and their individual actions, so to perform the classification of each players action we track the subjects in a scene and get the ground truth bounding boxes of unannotated frames, we used the tracker proposed by Danelljan et al. (2014)

Figure 34 – Confusion Matrix for BoP (Angles + Trajectories descriptors) with 97.16% of accuracy on KTH Dataset.



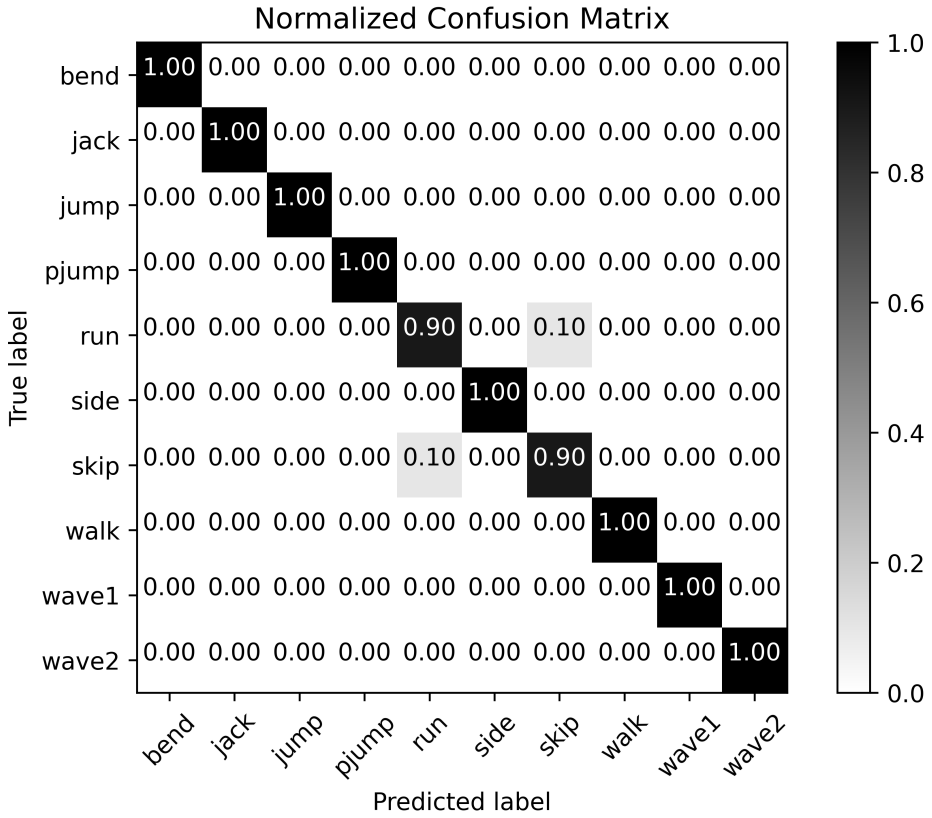
Source: The author.

and implemented in the Dlib Library (KING, 2009). Although for each clip of the Volleyball dataset we have 41 images (20 frames before the target frame, the target frame, and 20 frames after the target frame), the authors alert for the fact that the scenes change quite rapidly in volleyball, hence there is no guarantee that frames of the same clip may belong to the annotated frame most of the time.

Our technique requires a larger number of frames to be able to extract with quality the spatial and temporal information of the action performed. But in this case, we have a limited amount of frames per clip and when increasing the number of frames we would have more information to classify the action but with a great chance of occurring more than one action in the same clip for the same player causing in this case classification errors.

The Volleyball Dataset was created with the goal of classifying group activities and not human actions, thus, there is a great imbalance between the amount of sample of each class of human action. To minimize this problem we chose to remove the samples with the label “Jumping” from the database because they represent only 0.61% of the total samples, we also chose to remove the samples with the label “Standing” because they represent 68.68% of the samples. In other words, we removed the samples from the classes with the smallest and

Figure 35 – Confusion Matrix for BoP (Angles + Trajectories descriptors) with 97.85% of accuracy on Weizmann Dataset.



Source: The author.

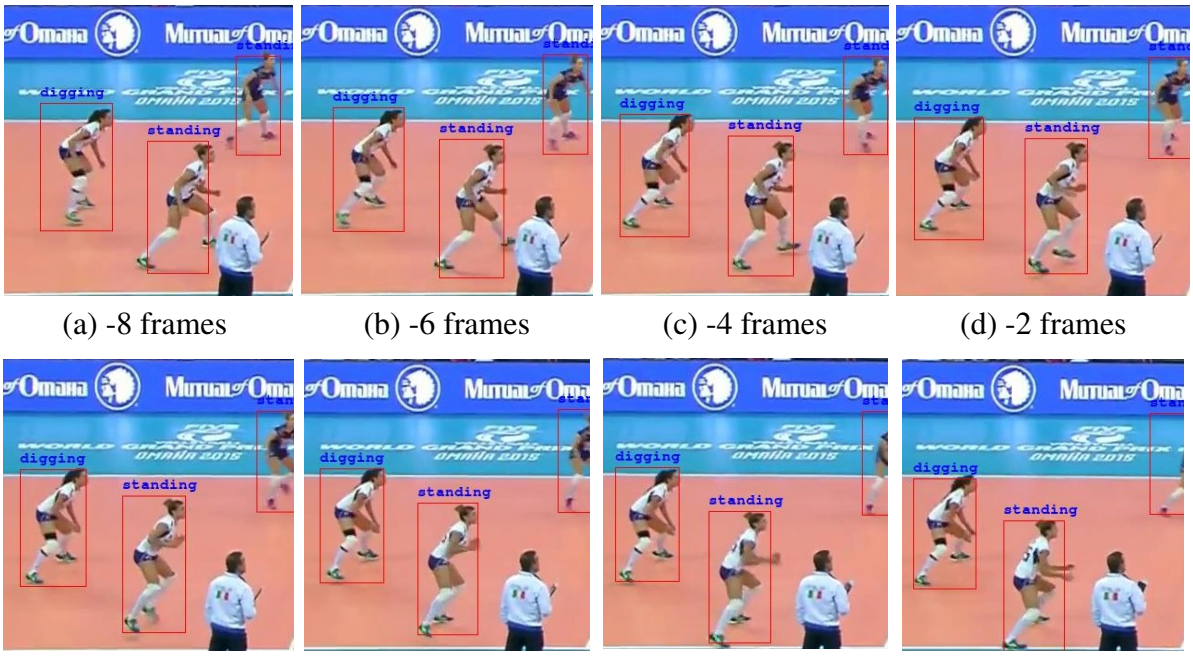
largest number of clips and made a balance in the other classes of the database through an under-sampling. Another problem presented in the dataset are the ambiguities and incorrect labels. Performing a qualitative analysis, it was possible to notice a large number of samples with inconsistent labels for the "Digging" class, which was also not considered in the experiments. Figure 36 shows an example where two players apparently perform the same action in the game, however one was labeled with the "Digging" class and the other with the "Standing" class.

To compute the trajectory descriptor (Subsections 7.1.2.2), we need to set L (trajectory length) and W (sampling step size) parameters, similar to experiments performed in the Weizmann dataset we achieved the best results setting $L = 20$ and $W = 1$.

Table 8 shows the results using the two descriptors presented in this chapter and their combination for the Volleyball dataset. By using only Spatial Angles features we achieved an accuracy rate of 67.12%, using Temporal Trajectories features we achieved an accuracy of 55.70%, and fusing Angles and Trajectories we improve the accuracy to 70.71%. Those results, likewise KTH and Weizmann results, suggest that the spatial (Angles) and temporal (Trajectories) features are complementary and help to improve the classification rates.

Figure 37 shows the confusion matrix for the Volleyball dataset by using the fusion of

Figure 36 – Sequence of cropped frames from the Volleyball dataset (IBRAHIM et al., 2016) in which it is possible to observe that the dataset has some ambiguities. For example, there are two players that are apparently performing the same action, however one was labeled "Digging" and the other as "Standing".



(a) -8 frames (b) -6 frames (c) -4 frames (d) -2 frames

(e) Annotated frame (f) +2 frames (g) +4 frames (h) +6 frames

Source: The author.

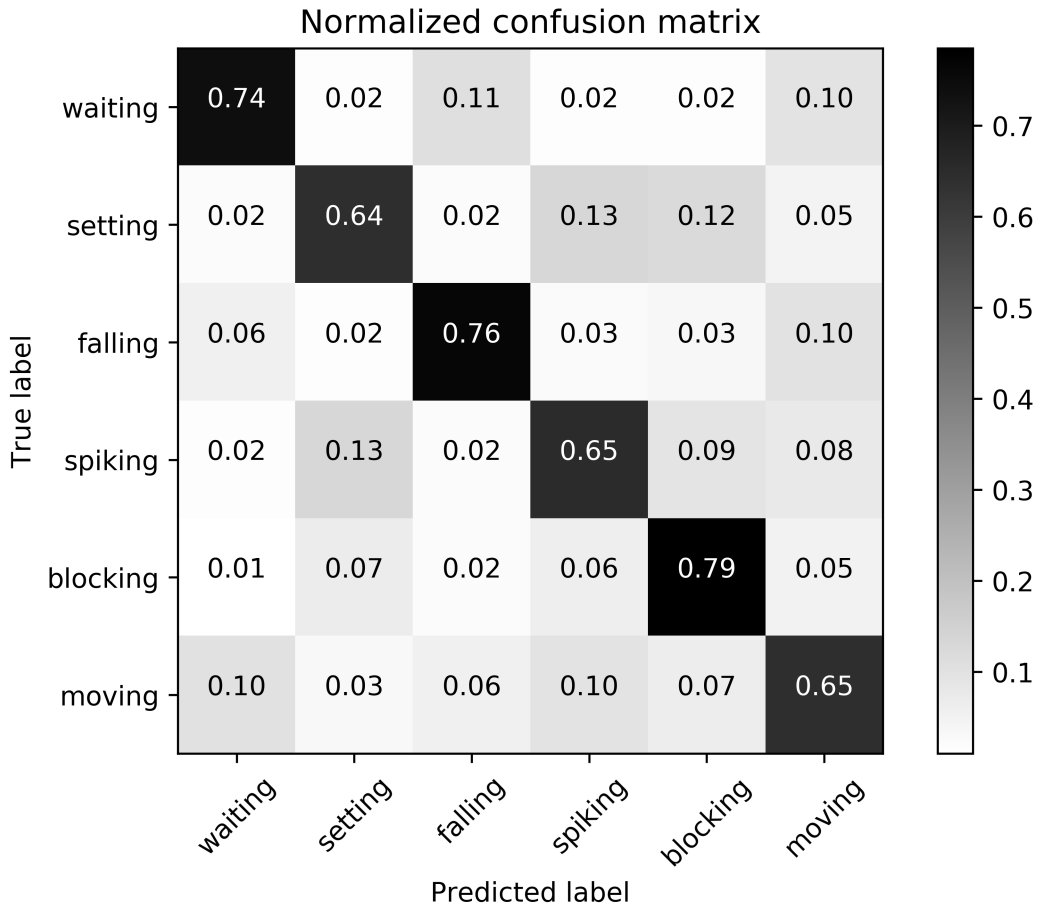
Table 8 – Accuracy rates (%) for Volleyball dataset of our proposed methods.

Method	**Accuracy**
BoP (Angles + Trajectories)	**70.71**
BoP (Angles)	67.12
BoP (Trajectories)	55.70

Angles and Trajectories features that achieved 70.71% of accuracy. It is possible to observe that the method present a stable classification rate between the classes, varying from 64% to 79% of accuracy, a small variation between the classification rates of each class indicates that the method was able to represent spatiotemporal patterns well so as not to privilege any specific class.

The results reported in the literature for individual actions classification in Volleyball dataset by state-of-the-art methods are presented in Table 9. However, since the reported methods used different clip sizes, we can't do a direct comparison. The presented methods use a small number of frames for each clip (only 10 frames for each clip). On the other hand, since our method needs a minimum number of frames to extract spatiotemporal features, thus we used all 41 frames.

Figure 37 – Confusion Matrix for BoP (Angles + Trajectories descriptors) with 70.71% of accuracy on Volleyball Dataset.



Source: The author.

Table 9 – Accuracy rates (%) for individual actions classification in Volleyball dataset by state-of-the-art methods.

Method	**Accuracy**
Shu et al. (2017)	69.10
Bagautdinov et al. (2017)	81.80
Wu et al. (2019)	83.10
Gavrilyuk et al. (2020)	85.90

As the Volleyball dataset present a lot of challenges (multiple-people in the scene, dynamic actions, occlusions, camera motion, and short video sequences), we consider that our method achieved good results.

7.3 Conclusion

In this chapter, we presented a new method for human action recognition using raw 2D poses encoded into the (θ, ρ) parameter space. We addressed the problem by using the power and simplicity of parameter space to represent straight lines. Our method is based on 2D poses and, nowadays, there are methods that achieve good results in the extraction of 2D poses with real-time processing speed and in unconstrained video sequences, such as OpenPose (CAO et al., 2018) and PifPaf (KREISS et al., 2019). Our method can recognize human actions in video sequences in real-time, with excellent accuracy, achieving state-of-the-art performance.

Furthermore, using poses instead of using the raw images may be a good choice for applications where it is important to preserve privacy, such as the automatic systems for monitoring possible falls of elderly people in their homes, where it is usually necessary to monitor the elderly 24 hours a day and in all rooms of the residence.

By encoding 2D poses into the parameter space, it was possible to extract useful spatial information (e.g., shapes found in a single frame) and temporal information (e.g., movements found across frames). Those features show promising performance using only a little amount of low-level features, in contrast to the iDT (WANG; SCHMID, 2013) method, for instance, which uses a lot of points densely sampled and is more computationally expensive.

Chapter 8
CONCLUSION

Through this work, it was found that the automatic recognition of human actions from videos is a current and very relevant topic of research and technological development, both for the monitoring of people in public areas, where in general there is great circulation, as well as inside residences or other private areas, in which preserving the privacy of environments and people is crucial.

Recent researches have pointed to a number of methods for recognizing actions in videos using techniques based on deep learning, however these techniques require an huge amount of labeled data for training. In addition, in order to work with deep learning in videos it is necessary to use computers with very high processing power. In some studies, researchers have even used clusters with more than 60 GPUs (CARREIRA; ZISSERMAN, 2017a). Such a necessity makes the use of these techniques practically impossible for applications that do not have high processing power and need a real-time response. As preliminary studies in this thesis, we carried out experiments using deep neural networks already trained in large datasets, such as Kinetics (KAY et al., 2017) and Sports-1M (KARPATHY et al., 2014), as described in Chapter 5. In this case, we carried out a transfer learning step and some changes in the classification phase, obtaining very competitive results in detecting pornography in videos (accuracy of 95.1% in the Pornography 800 dataset). However, in addition to requiring extensive training data sets and computers with large processing and storage capacities, techniques based on deep learning use raw images as input. Therefore, they are not the best option in scenarios where preserving the privacy of people and places is the main concern.

Recent researches also have shown that it is possible to carry out 2D pose estimation from video, in real-time, in complex scenarios, and with a number of people in the scenes. OpenPose (CAO et al., 2018) and PifPaf (KREISS et al., 2019) are good examples of methods that have such characteristics. They return 2D poses of people present in video frames and discard all the surrounding information, which makes them very good solutions for action recognition from videos, with preservation of the privacy of places and individuals.

Motivated by these new developments, in this thesis, we proposed new methods to extract

spatiotemporal features from 2D poses. In addition to being lighter for processing, 2D poses also help to solve the privacy problem, since using only the poses, it is not necessary to have access to the information of the environment where the images were collected or information that can identify the person who is performing a certain action, thus ensuring privacy. Also, the use of 2D poses, in contrast to the use of raw images and deep neural networks, can be a good choice for applications where the volume of data available for training the models is limited.

The methods for recognizing human actions in videos based on 2D poses presented in Chapters 6 and 7 showed that with the use of only the pose information it is possible to perform the classification of human actions with good accuracy. Another important aspect is the approach to encode 2D poses into the parameter space, such a technique converts each line segment that represents a limb or part of the human body into a point in the parameter space, such encoding has shown to be effective in extracting features and improving the performance of the proposed method.

Based on the studies and conclusions presented, we can answer the research questions of this thesis presented in Chapter 1:

1. What are the advantages and disadvantages when using deep learning to classify human actions in videos?

   Answer: The positive points of using deep learning to classify human actions is the possibility of using transfer learning from models already trained in large databases, and generally have state-of-the-art results. As negative points we can mention the need to use many GPUs for the development of new neural network architectures, the need for a lot of data for training and the use of raw images, which can bring privacy problems to human action recognition applications.

2. Can we use only 2D poses for the development of an automatic recognition method for human actions in videos?

   Answer: Based on the proposed methods and experiments presented in Chapters 6 and 7 we can state that it is possible to use only information extracted from 2D poses to perform the task of classifying human actions from videos, the experiments showed good classification rates, compared with others state-of-the-art methods that are not based on 2D poses.

3. How can we extract spatial and temporal information of 2D poses from videos?

   Answer: In this thesis, in Chapters 6 and 7, new descriptors were proposed based on 2D poses in order to capture spatial and temporal information about human action. Regarding spatial information, a descriptor based on angles formed by adjacent parts or limbs of the human body has been proposed. Regarding temporal information, a descriptor was proposed based on the trajectories of the parts and limbs of the human body through the

video frames. Such descriptors proved to be very effective, presenting good results and improving the results when fusing the spatial and temporal features.

4. Is it possible to recognize human actions in videos while preserving the identity of the people involved in the scenes?

   Answer: The concern about privacy in computer vision applications is of paramount importance. In the recognition of human actions, it is not different, not all applications of recognition of human actions is allowed to identify the actors of a specific action, in this case, the use of 2D poses can be a good alternative since only with the information of the pose, it is possible to preserve the privacy of both the environment being controlled and the people involved in the actions.

8.1 Thesis Contributions

In short, among the main contributions of this thesis can be highlighted the following points:

- Evaluations of deep 3D Convolutional Neural Networks (3D CNNs) for pornography detection in videos, evaluations so far, to best of our knowledge, not performed in the literature;
- Proposal of new descriptors based only on 2D poses (descriptors based on spatial and temporal information extracted from the pattern of the poses in each frame (angles) and the pattern of movement of the human body through the frames (trajectories));
- Proposal of a new way of encoding poses in the parameter space, where each straight line segment that represents a limb or part of the human body is represented by a point in the parameter space. This point is nothing more than the parameters (θ, ρ) that represent the line that passes through the part of the body. Such encoding proved to be efficient to improve the extraction of features (trajectories) and consequently the recognition rates;
- Proposal of a new framework for the classification of human actions based on Bag-of-Poses (BoP) and Fisher Vectors (FV). Since human actions generally do not have a fixed duration in videos, it is necessary a framework that can handle videos containing arbitrary sizes. This work presents a new framework based on Bag-of-Poses that explores the benefits of using Fisher Vectors instead of the conventional Bag-of-Visual-Words (BoVW) dictionary approach;
- Proposal of new methods that can be used for applications where people's privacy is a requirement to be met. Since the proposed methods do not need to use raw images, but only 2D poses, we are able to guarantee that people's identity will be preserved, as well as information about the environment where the images were collected;

- Publication of two repositories with the source codes for extracting 2D poses, extracting features, and for classifying the human actions from videos based on the methods proposed in this thesis, more information in Appendix B.

8.2 Future Work

For future work, we point out some possible directions:

- Use of PifPaf in the methods proposed in this thesis and comparison with OpenPose;
- Evaluation of methods for 3D human pose estimation from 2D poses or 2D images (PARK et al., 2016; CHEN; RAMANAN, 2017; ZHOU et al., 2017; DROVER et al., 2018) and assessment of the performance of 3D pose estimation in human action recognition from videos;
- Application of the proposed methods for detecting falls in the elderly.
- Combination of information from 2D poses with deep neural networks as a means of taking advantage of the benefits of deep networks with regard to accuracy and the benefits of 2D poses in preserving people's privacy;

8.3 Publications

The following papers were published during the development of this thesis:

1. Murilo Varges da Silva and Aparecido Nilceu Marana. **Spatiotemporal CNNs for Pornography Detection in Videos**. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2018. Lecture Notes in Computer Science, vol 11401. Springer, Cham. ISBN 978-3-030-13468-6. Paper presented in 23th Iberoamerican Congress on Pattern Recognition (CIARP), Madrid, Spain, November 2018. (SILVA; MARANA, 2018)
2. Murilo Varges da Silva and Aparecido Nilceu Marana. **Human Action Recognition using 2D Poses**, 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), Salvador, Brazil, 2019, pp. 747-752, doi: 10.1109/BRACIS.2019.00134. Paper presented in 8th Brazilian Conference on Intelligent Systems (BRACIS), Salvador, Brazil, October 2019. (SILVA; MARANA, 2019)
3. Murilo Varges da Silva and Aparecido Nilceu Marana. **Human Action Recognition in Videos Based on Spatiotemporal Features and Bag-of-Poses.** Applied Soft Computing. Volume 95, October 2020, 106513. ISSN 1568-4946. doi: 10.1016/j.asoc.2020.106513. (SILVA; MARANA, 2020)

BIBLIOGRAPHY

ADEYEMO, A.; WIMMER, H.; POWELL, L. Effects of normalization techniques on logistic regression in data science. In: *Proceedings of the Conference on Information Systems Applied Research ISSN*. [S.l.: s.n.], 2018. v. 2167, p. 1508. Cited on page 58.

AGAHIAN, S.; NEGIN, F.; KÖSE, C. Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition. *The Visual Computer*, Springer, v. 35, n. 4, p. 591–607, 2019. Cited 3 times on pages 60, 61, and 62.

AGAHIAN, S.; NEGIN, F.; KÖSE, C. An efficient human action recognition framework with pose-based spatiotemporal features. *Engineering Science and Technology, an International Journal*, Elsevier, v. 23, n. 1, p. 196–203, 2020. Cited on page 31.

AGGARWAL, J.; RYOO, M. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, ACM, New York, NY, USA, v. 43, n. 3, p. 16:1–16:43, abr. 2011. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/1922649.1922653>. Cited 2 times on pages 17 and 18.

ALCÂNTARA, M. F. de; MOREIRA, T. P.; PEDRINI, H. Motion silhouette-based real time action recognition. In: RUIZ-SHULCLOPER, J.; BAJA, G. Sanniti di (Ed.). *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 471–478. ISBN 978-3-642-41827-3. Cited on page 31.

ALCANTARA, M. F. de; MOREIRA, T. P.; PEDRINI, H.; FLÓREZ-REVUELTA, F. Action identification using a descriptor with autonomous fragments in a multilevel prediction scheme. *Signal, image and video processing*, Springer, v. 11, n. 2, p. 325–332, 2017. Cited 3 times on pages 31, 52, and 66.

ALMEIDA, R.; BUSTOS, B.; PATROCÍNIO, Z. K. G. do; GUIMARÃES, S. J. F. Human action classification using an extended bow formalism. In: SPRINGER. *International Conference on Image Analysis and Processing*. [S.l.], 2017. p. 185–196. Cited 2 times on pages 52 and 66.

ANTIPOV, G.; BERRANI, S.-A.; RUCHAUD, N.; DUGELAY, J.-L. Learned vs. hand-crafted features for pedestrian gender recognition. In: *Proceedings of the 23rd ACM international conference on Multimedia*. [S.l.: s.n.], 2015. p. 1263–1266. Cited on page 19.

AVILA, S.; THOME, N.; CORD, M.; VALLE, E.; ARAúJO, A. de A. Bossa: Extended bow formalism for image classification. In: *18th IEEE ICIP*. [S.l.: s.n.], 2011. p. 2909–2912. ISSN 1522-4880. Cited on page 45.

AVILA, S.; THOME, N.; CORD, M.; VALLE, E.; ARAúJO, A. de A. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, v. 117, n. 5, p. 453–465, 2013. ISSN 1077-3142. Cited 5 times on pages 34, 35, 41, 44, and 45.

BAGAUTDINOV, T.; ALAHI, A.; FLEURET, F.; FUA, P.; SAVARESE, S. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 4315–4324. Cited on page 70.

BARADEL, F.; WOLF, C.; MILLE, J. Human action recognition: Pose-based attention draws focus to hands. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2017. p. 604–613. Cited on page 33.

BUZZELLI, M.; ALBÉ, A.; CIOCCA, G. A vision-based system for monitoring elderly people at home. *Applied Sciences*, Multidisciplinary Digital Publishing Institute, v. 10, n. 1, p. 374, 2020. Cited 2 times on pages 22 and 23.

CAETANO, C.; AVILA, S.; GUIMARãES, S.; ARAúJO, A. d. A. Pornography detection using bossanova video descriptor. In: *2014 22nd (EUSIPCO)*. [S.l.: s.n.], 2014. p. 1681–1685. ISSN 2219-5491. Cited on page 45.

CAETANO, C.; AVILA, S. E. F. de; SCHWARTZ, W. R.; GUIMARÃES, S. J. F.; ARAÚJO, A. de A. A mid-level video representation based on binary descriptors: A case study for pornography detection. *CoRR*, abs/1605.03804, 2016. Cited on page 45.

CAO, Z.; HIDALGO, G.; SIMON, T.; WEI, S.-E.; SHEIKH, Y. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: *arXiv preprint arXiv:1812.08008*. [S.l.: s.n.], 2018. Cited 12 times on pages 19, 25, 26, 27, 46, 49, 53, 55, 57, 62, 71, and 72.

CARMONA, J. M.; CLIMENT, J. Human action recognition by means of subtensor projections and dense trajectories. *Pattern Recognition*, v. 81, p. 443–455, 2018. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320318301493>. Cited 4 times on pages 52, 62, 65, and 66.

CARREIRA, J.; ZISSERMAN, A. Quo vadis, action recognition? a new model and the kinetics dataset. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2017. p. 6299–6308. Cited 2 times on pages 33 and 72.

CARREIRA, J.; ZISSERMAN, A. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. Cited on page 45.

CHAARAOUI, A. A.; CLIMENT-PÉREZ, P.; FLÓREZ-REVUELTA, F. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, Elsevier, v. 34, n. 15, p. 1799–1807, 2013. Cited on page 31.

CHAARAOUI, A. A.; CLIMENT-PÉREZ, P.; FLÓREZ-REVUELTA, F. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, Elsevier, v. 34, n. 15, p. 1799–1807, 2013. Cited 2 times on pages 52 and 66.

CHAQUET, J. M.; CARMONA, E. J.; FERNáNDEZ-CABALLERO, A. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, Elsevier Science Inc., New York, NY, USA, v. 117, n. 6, p. 633–659, jun. 2013. ISSN 1077-3142. Disponível em: <http://dx.doi.org/10.1016/j.cviu.2013.01.013>. Cited on page 20.

CHEEMA, S.; EWEIWI, A.; THURAU, C.; BAUCKHAGE, C. Action recognition by learning discriminative key poses. In: IEEE. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. [S.l.], 2011. p. 1302–1309. Cited on page 31.

CHEN, C.-H.; RAMANAN, D. 3d human pose estimation= 2d pose estimation+ matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2017. p. 7035–7043. Cited on page 75.

CHEN, T.-Y.; BIGLARI-ABHARI, M.; KEVIN, I.; WANG, A. K. Trusting the computer in computer vision: A privacy-affirming framework. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2017. p. 56–63. Cited on page 20.

CHENG, G.; WAN, Y.; SAUDAGAR, A. N.; NAMUDURI, K.; BUCKLES, B. P. Advances in human action recognition: A survey. *CoRR*, abs/1501.05964, 2015. Disponível em: <http://arxiv.org/abs/1501.05964>. Cited on page 17.

CHO, K.; MERRIËNBOER, B. V.; GULCEHRE, C.; BAHDANAU, D.; BOUGARES, F.; SCHWENK, H.; BENGIO, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. Cited on page 33.

CHOU, K.-P.; PRASAD, M.; WU, D.; SHARMA, N.; LI, D.-L.; LIN, Y.-F.; BLUMENSTEIN, M.; LIN, W.-C.; LIN, C.-T. Robust feature-based automated multi-view human action recognition system. *IEEE Access*, IEEE, v. 6, p. 15283–15296, 2018. Cited 3 times on pages 31, 52, and 66.

DALAL, N.; TRIGGS, B.; SCHMID, C. Human detection using oriented histograms of flow and appearance. In: LEONARDIS, A.; BISCHOF, H.; PINZ, A. (Ed.). *Computer Vision – ECCV 2006*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 428–441. ISBN 978-3-540-33835-2. Cited on page 30.

DANELLJAN, M.; HÄGER, G.; KHAN, F.; FELSBERG, M. Accurate scale estimation for robust visual tracking. In: BMVA PRESS. *British Machine Vision Conference, Nottingham, September 1-5, 2014*. [S.l.], 2014. Cited on page 66.

DOLLAR, P.; RABAUD, V.; COTTRELL, G.; BELONGIE, S. Behavior recognition via sparse spatio-temporal features. In: *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. [S.l.: s.n.], 2005. p. 65–72. Cited on page 30.

DONAHUE, J.; HENDRICKS, L. A.; GUADARRAMA, S.; ROHRBACH, M.; VENUGOPALAN, S.; SAENKO, K.; DARRELL, T. Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 2625–2634. Cited on page 33.

DOUMANOGLOU, A.; VRETOS, N.; DARAS, P. Action recognition from videos using sparse trajectories. *IET Conference Proceedings*, Institution of Engineering and Technology, p. 10 (5 .)–10 (5 .)(1), jan. 2016. Disponível em: <https://digital-library.theiet.org/content/conferences/10.1049/ic.2016.0078>. Cited 2 times on pages 52 and 66.

DROVER, D.; CHEN, C.-H.; AGRAWAL, A.; TYAGI, A.; HUYNH, C. P. Can 3d pose be learned from 2d projections alone? In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 0–0. Cited on page 75.

GALL, J.; YAO, A.; GOOL, L. V. 2d action recognition serves 3d human pose estimation. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2010. p. 425–438. Cited on page 19.

GAVRILYUK, K.; SANFORD, R.; JAVAN, M.; SNOEK, C. G. Actor-transformers for group activity recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2020. p. 839–848. Cited on page 70.

GORELICK, L.; BLANK, M.; SHECHTMAN, E.; IRANI, M.; BASRI, R. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, v. 29, n. 12, p. 2247–2253, dez. 2007. Cited 5 times on pages 35, 36, 37, 49, and 62.

GUO, K.; ISHWAR, P.; KONRAD, J. Action recognition from video using feature covariance matrices. *IEEE Transactions on Image Processing*, IEEE, v. 22, n. 6, p. 2479–2494, 2013. Cited 2 times on pages 52 and 66.

HE, Y.; SHIRAKABE, S.; SATOH, Y.; KATAOKA, H. Human action recognition without human. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2016. p. 11–17. Cited on page 22.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. Cited on page 33.

IBRAHIM, M. S.; MURALIDHARAN, S.; DENG, Z.; VAHDAT, A.; MORI, G. A hierarchical deep temporal model for group activity recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. Cited 8 times on pages 13, 35, 36, 37, 38, 62, 66, and 69.

INSAFUTDINOV, E.; PISHCHULIN, L.; ANDRES, B.; ANDRILUKA, M.; SCHIELE, B. *DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model*. 2016. Cited on page 26.

JI, S.; XU, W.; YANG, M.; YU, K. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 35, n. 1, p. 221–231, jan. 2013. ISSN 0162-8828. Cited 2 times on pages 32 and 66.

JIA, Y.; SHELHAMER, E.; DONAHUE, J.; KARAYEV, S.; LONG, J.; GIRSHICK, R.; GUADARRAMA, S.; DARRELL, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. Cited 3 times on pages 44, 49, and 62.

JOLLIFFE, I. *Principal component analysis*. [S.l.]: Springer, 2011. Cited 2 times on pages 49 and 63.

JUNEJO, I. N.; AGHBARI, Z. A. Using sax representation for human action recognition. *Journal of Visual Communication and Image Representation*, Elsevier, v. 23, n. 6, p. 853–861, 2012. Cited 3 times on pages 31, 52, and 66.

KARPATHY, A.; TODERICI, G.; SHETTY, S.; LEUNG, T.; SUKTHANKAR, R.; FEI-FEI, L. Large-scale video classification with convolutional neural networks. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1725–1732. Cited 3 times on pages 32, 41, and 72.

KAY, W.; CARREIRA, J.; SIMONYAN, K.; ZHANG, B.; HILLIER, C.; VIJAYANARASIMHAN, S.; VIOLA, F.; GREEN, T.; BACK, T.; NATSEV, P.; SULEYMAN, M.; ZISSERMAN, A. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. Cited 2 times on pages 43 and 72.

KING, D. E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, JMLR.org, v. 10, p. 1755–1758, dez. 2009. ISSN 1532-4435. Cited on page 67.

KLASER, A.; MARSZALEK, M.; SCHMID, C. A Spatio-Temporal Descriptor Based on 3D-Gradients. In: EVERINGHAM, M.; NEEDHAM, C.; FRAILE, R. (Ed.). *BMVC 2008 - 19th British Machine Vision Conference*. Leeds, United Kingdom: British Machine Vision Association, 2008. p. 275:1–10. Cited on page 30.

KONG, Y.; FU, Y. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*, 2018. Cited on page 19.

KRAPAC, J.; VERBEEK, J.; JURIE, F. Modeling spatial layout with fisher vectors for image categorization. In: *2011 International Conference on Computer Vision*. [S.l.: s.n.], 2011. p. 1487–1494. ISSN 2380-7504. Cited 3 times on pages 49, 62, and 86.

KREISS, S.; BERTONI, L.; ALAHI, A. Pifpaf: Composite fields for human pose estimation. *CoRR*, abs/1903.06593, 2019. Disponível em: <http://arxiv.org/abs/1903.06593>. Cited 7 times on pages 19, 26, 28, 29, 53, 71, and 72.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. USA: Curran Associates Inc., 2012. (NIPS'12), p. 1097–1105. Cited on page 32.

LAPTEV, I.; CAPUTO, B. et al. Recognizing human actions: a local SVM approach. In: IEEE. *ICPR*. [S.l.], 2004. p. 32–36. Cited 4 times on pages 35, 36, 49, and 62.

LAPTEV, I.; LINDEBERG, T. Space-time interest points. In: *Proceedings Ninth IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2003. p. 432–439 vol.1. Cited on page 30.

LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P.; ZITNICK, C. L. Microsoft coco: Common objects in context. In: SPRINGER. *European conference on computer vision*. [S.l.], 2014. p. 740–755. Cited 3 times on pages 25, 26, and 27.

LIU, J.; SHAHROUDY, A.; XU, D.; WANG, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2016. p. 816–833. Cited on page 33.

LU, Z. M.; SHI, Y. Fast video shot boundary detection based on svd and pattern matching. *IEEE Transactions on Image Processing*, v. 22, n. 12, p. 5136–5145, dez. 2013. ISSN 1057-7149. Cited on page 20.

LV, F.; NEVATIA, R. Single view human action recognition using key pose matching and viterbi path searching. In: IEEE. *2007 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.], 2007. p. 1–8. Cited on page 19.

MAATEN, L. v. d.; HINTON, G. Visualizing data using t-sne. *Journal of machine learning research*, v. 9, n. Nov, p. 2579–2605, 2008. Cited 2 times on pages 49 and 63.

MAATEN, L. van der; HINTON, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, v. 9, p. 2579–2605, 2008. Cited on page 42.

MOREIRA, D.; AVILA, S.; PEREZ, M.; MORAES, D.; TESTONI, V.; VALLE, E.; GOLDENSTEIN, S.; ROCHA, A. Pornography classification: The hidden clues in video space–time. *Forensic Science International*, v. 268, p. 46–61, 2016. ISSN 0379-0738. Cited on page 45.

MOREIRA, T. P.; MENOTTI, D.; PEDRINI, H. Video action recognition based on visual rhythm representation. *Journal of Visual Communication and Image Representation*, Elsevier, p. 102771, 2020. Cited on page 66.

MOUSTAFA, M. Applying deep learning to classify pornographic images and videos. *CoRR*, abs/1511.08899, 2015. Cited on page 45.

PARK, S.; HWANG, J.; KWAK, N. 3d human pose estimation using convolutional neural networks with 2d pose information. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2016. p. 156–169. Cited on page 75.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Cited 2 times on pages 49 and 62.

PEREZ, M.; AVILA, S.; MOREIRA, D.; MORAES, D.; TESTONI, V.; VALLE, E.; GOLDENSTEIN, S.; ROCHA, A. Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, v. 230, p. 279–293, 2017. ISSN 0925-2312. Cited on page 45.

PERRONNIN, F.; SÁNCHEZ, J.; MENSINK, T. Improving the fisher kernel for large-scale image classification. In: SPRINGER. *European conference on computer vision*. [S.l.], 2010. p. 143–156. Cited 4 times on pages 49, 60, 61, and 62.

POPPE, R. A survey on vision-based human action recognition. *Image Vision Comput.*, Butterworth-Heinemann, Newton, MA, USA, v. 28, n. 6, p. 976–990, jun. 2010. ISSN 0262-8856. Disponível em: <http://dx.doi.org/10.1016/j.imavis.2009.11.014>. Cited on page 20.

PRESTI, L. L.; CASCIA, M. L. 3d skeleton-based human action classification: A survey. *Pattern Recognition*, Elsevier, v. 53, p. 130–147, 2016. Cited on page 31.

RAJA, K.; LAPTEV, I.; PÉREZ, P.; OISEL, L. Joint pose estimation and action recognition in image graphs. In: IEEE. *2011 18th IEEE International Conference on Image Processing*. [S.l.], 2011. p. 25–28. Cited on page 31.

RAVANBAKHSH, M.; MOUSAVI, H.; RASTEGARI, M.; MURINO, V.; DAVIS, L. S. Action recognition with image based cnn features. *arXiv preprint arXiv:1512.03980*, 2015. Cited 2 times on pages 52 and 66.

SCOVANNER, P.; ALI, S.; SHAH, M. A 3-dimensional sift descriptor and its application to action recognition. In: *Proceedings of the 15th ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2007. (MM '07), p. 357–360. ISBN 978-1-59593-702-5. Cited on page 30.

SEIDENARI, L.; VARANO, V.; BERRETTI, S.; BIMBO, A.; PALA, P. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2013. p. 479–485. Cited on page 60.

Senior, A.; Pankanti, S.; Hampapur, A.; Brown, L.; Ying-Li Tian; Ekin, A.; Connell, J.; Chiao Fe Shu; Lu, M. Enabling video privacy through computer vision. *IEEE Security Privacy*, v. 3, n. 3, p. 50–57, 2005. Cited on page 20.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data*, Springer, v. 6, n. 1, p. 60, 2019. Cited on page 21.

SHU, T.; TODOROVIC, S.; ZHU, S.-C. Cern: confidence-energy recurrent network for group activity recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 5523–5531. Cited on page 70.

SILVA, M. V. da; MARANA, A. N. Spatiotemporal cnns for pornography detection in videos. In: VERA-RODRIGUEZ, R.; FIERREZ, J.; MORALES, A. (Ed.). *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Cham: Springer International Publishing, 2018. p. 547–555. ISBN 978-3-030-13469-3. Cited on page 75.

SILVA, M. V. da; MARANA, A. N. Human action recognition using 2d poses. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2019. p. 747–752. ISSN 2643-6256. Cited on page 75.

SILVA, M. V. da; MARANA, A. N. Human action recognition in videos based on spatiotemporal features and bag-of-poses. *Applied Soft Computing*, Elsevier, v. 95, p. 106513, 2020. Cited on page 75.

SIMONYAN, K.; ZISSERMAN, A. Two-stream convolutional networks for action recognition in videos. In: GHAHRAMANI, Z.; WELLING, M.; CORTES, C.; LAWRENCE, N. D.; WEINBERGER, K. Q. (Ed.). *Advances in Neural Information Processing Systems 27*. [S.l.]: Curran Associates, Inc., 2014. p. 568–576. Cited 2 times on pages 32 and 33.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. Cited 3 times on pages 11, 40, and 41.

SINGH, S.; VELASTIN, S. A.; RAGHEB, H. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: IEEE. *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. [S.l.], 2010. p. 48–55. Cited on page 31.

SINGH, T.; VISHWAKARMA, D. K. A hybrid framework for action recognition in low-quality video sequences. *arXiv preprint arXiv:1903.04090*, 2019. Cited 3 times on pages 31, 52, and 66.

SOUZA, F. D. M. de; VALLE, E.; CÁMARA-CHÁVEZ, G.; ARAÚJO, A. An evaluation on color invariant based local spatiotemporal features for action recognition. In: *IEEE SIBGRAPI*. [S.l.: s.n.], 2012. Cited on page 45.

TOSHEV, A.; SZEGEDY, C. Deeppose: Human pose estimation via deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2014. p. 1653–1660. Cited 2 times on pages 25 and 26.

TRAN, D.; BOURDEV, L.; FERGUS, R.; TORRESANI, L.; PALURI, M. Learning spatiotemporal features with 3d convolutional networks. In: *IEEE ICCV*. Washington, DC, USA: [s.n.], 2015. p. 4489–4497. ISBN 978-1-4673-8391-2. Cited 4 times on pages 11, 32, 40, and 41.

TRAN, D.; WANG, H.; TORRESANI, L.; RAY, J.; LECUN, Y.; PALURI, M. A closer look at spatiotemporal convolutions for action recognition. *CoRR*, abs/1711.11248, 2017. Cited 6 times on pages 14, 32, 40, 42, 43, and 44.

VALLE, E.; AVILA, S. E. F. de; JR., A. da L.; SOUZA, F. D. M. de; COELHO, M. de M.; ARAÚJO, A. de A. Content-based filtering for video sharing social networks. *CoRR*, abs/1101.2427, 2011. Cited on page 45.

VEMULAPALLI, R.; ARRATE, F.; CHELLAPPA, R. Human action recognition by representing 3d skeletons as points in a lie group. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2014. p. 588–595. Cited on page 31.

VRIGKAS, M.; NIKOU, C.; KAKADIARIS, I. A. A review of human activity recognition methods. *Frontiers in Robotics and AI*, v. 2, p. 28, 2015. ISSN 2296-9144. Disponível em: <http://journal.frontiersin.org/article/10.3389/frobt.2015.00028>. Cited 2 times on pages 11 and 18.

WANG, C.; WANG, Y.; YUILLE, A. L. An approach to pose-based action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2013. p. 915–922. Cited on page 19.

WANG, H.; KLÄSER, A.; SCHMID, C.; LIU, C.-L. Action Recognition by Dense Trajectories. In: *IEEE Conference on Computer Vision & Pattern Recognition*. Colorado Springs, United States: [s.n.], 2011. p. 3169–3176. Disponível em: <http://hal.inria.fr/inria-00583818/en>. Cited 2 times on pages 31 and 47.

WANG, H.; KLÄSER, A.; SCHMID, C.; LIU, C.-L. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, Springer, v. 103, n. 1, p. 60–79, 2013. Cited 3 times on pages 54, 61, and 65.

WANG, H.; SCHMID, C. Action recognition with improved trajectories. In: *2013 IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2013. p. 3551–3558. ISSN 1550-5499. Cited 4 times on pages 31, 62, 65, and 71.

WU, J.; WANG, L.; WANG, L.; GUO, J.; WU, G. Learning actor relation graphs for group activity recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 9964–9974. Cited on page 70.

XIE, S.; SUN, C.; HUANG, J.; TU, Z.; MURPHY, K. Rethinking spatiotemporal feature learning for video understanding. *CoRR*, abs/1712.04851, 2017. Cited 2 times on pages 42 and 43.

ZHANG, H.-B.; ZHANG, Y.-X.; ZHONG, B.; LEI, Q.; YANG, L.; DU, J.-X.; CHEN, D.-S. A comprehensive survey of vision-based human action recognition methods. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 19, n. 5, p. 1005, 2019. Cited on page 38.

ZHANG, Z.; TAO, D. Slow feature analysis for human action recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, v. 34, n. 03, p. 436–450, mar. 2012. ISSN 0162-8828. Cited 2 times on pages 52 and 66.

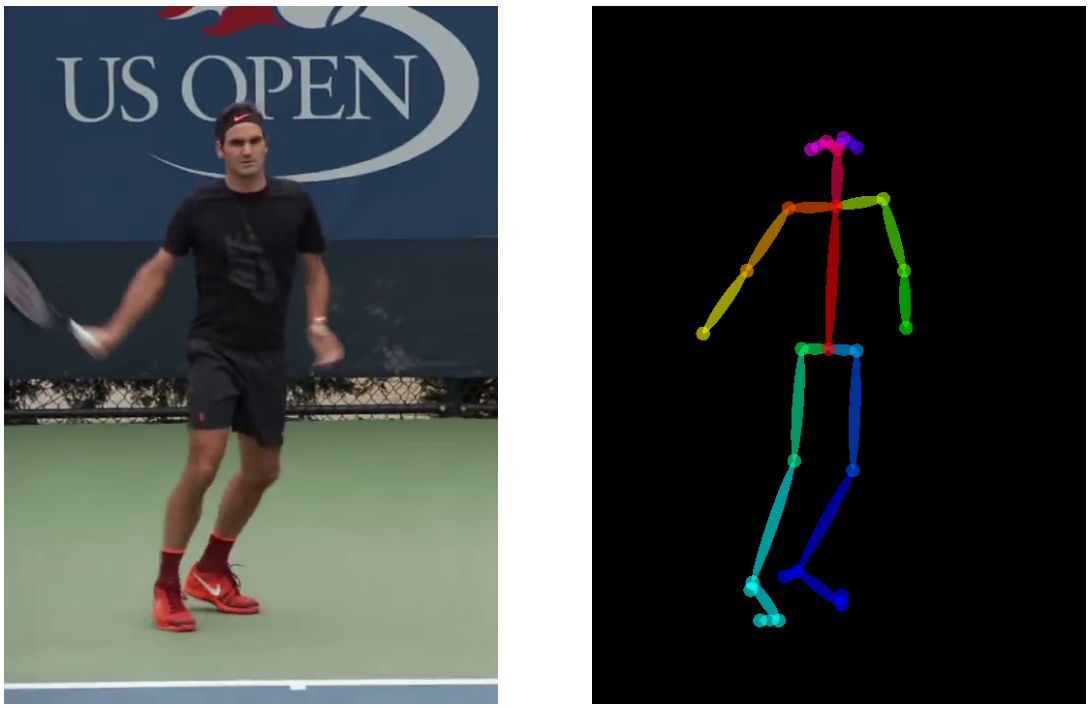
ZHOU, X.; HUANG, Q.; SUN, X.; XUE, X.; WEI, Y. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2017. p. 398–407. Cited on page 75.

APPENDIX A
ANGLES FEATURE

In this appendix, it is presented in detail how the angle measurement features are computed from 2D poses and how the low-level features are encoded at an intermediate level using Fisher Vector.

Firstly, from a video frame that contains a person, the 2D pose is extracted, as shown in the Figure 38.

Figure 38 – Example of an image with a person and the 2D pose estimation.



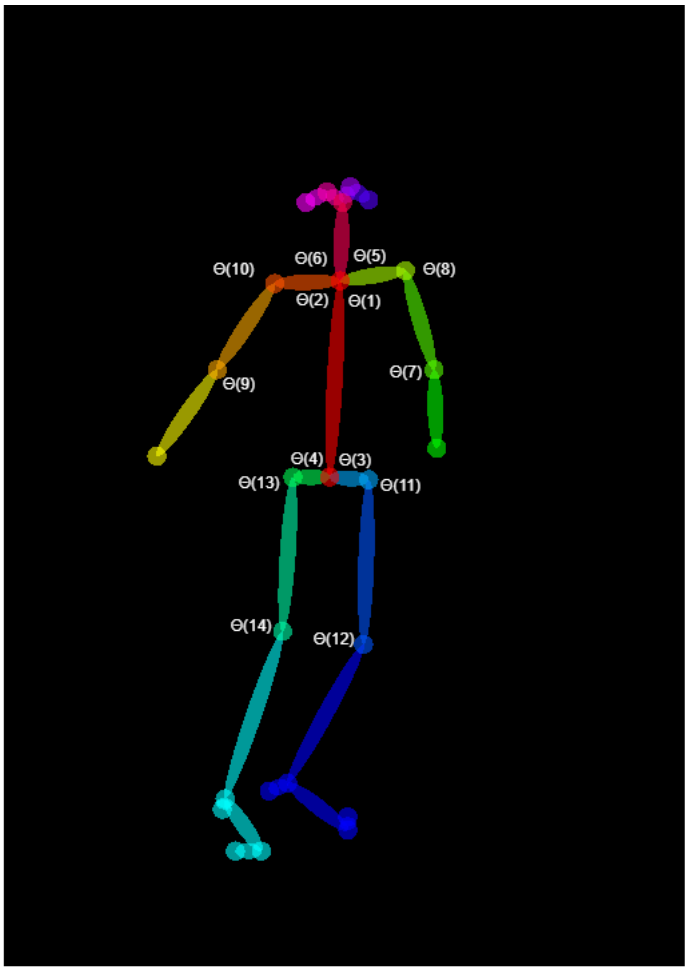
(a) Original frame (b) 2D pose

Source: The author.

From the 2D pose, using the Equation 6.2 (presented in Chapter 6) the fourteen angles are computed, forming a feature vector as presented in Figure 39.

Therefore, to describe a human action in a video, we compute the fourteen angles for

Figure 39 – The 2D pose and the fourteen angles calculated from 2D pose. Top the 2D pose with angles location and bottom the feature vector with computed angles.



Angle	Θ(1)	Θ(2)	Θ(3)	Θ(4)	Θ(5)	Θ(6)	Θ(7)	Θ(8)	Θ(9)	Θ(10)	Θ(11)	Θ(12)	Θ(13)	Θ(14)
Value	102	84	90	93	78	94	166	96	179	126	92	153	93	164

Source: The author.

each 2D pose found in each frame and use the Fisher Vector encoder to represent the human action. As presented in Chapter 6, Subsection 6.1.3, Fisher Vector uses a Gaussian Mixture Model (GMM), so we set the number of Gaussians to $K = 20$ and sample all features from the training set to estimate the GMM. Each video is represented by a $K + 2DK$ dimensional Fisher Vector, where D is the descriptor dimension, in this example $D = 14$. Thus, each clip can be represented by a vector of size 580 ($20 + 2 * 14 * 20 = 580$). Using Fisher Vector, in addition to encoding first and second order statistics, we also have the possibility to encode clips of arbitrary sizes. The details of computation of Fisher Vector is described in Krapac et al. (2011) and the implementation can be found in Github[1].

[1] https://gist.github.com/danoneata/9927923

APPENDIX B
SOURCE CODES

During the development of this thesis, the source codes of the published papers were made available on GitHub.

- **Human Action Recognition Using 2D Poses**:

  In this repository, it is available the source code for the paper presented in BRACIS conference. For more information, see Chapter 6 or the paper *"Human Action Recognition Using 2D Poses"* listed in Chapter 8. The experiments were run in the KTH and Weizmann databases.

  <https://github.com/murilovarges/HumanActionRecognition2DPoses>

- **Human Action Recognition in Videos Based on Spatiotemporal Features and Bag-of-Poses**:

  In this repository, it is available the source code for the paper published in Applied Soft Computer Journal. For more information, see the Chapter 7 or the paper *"Human Action Recognition in Videos Based on Spatiotemporal Features and Bag-of-Poses"* listed in Chapter 8. The experiments were run in the KTH and Weizmann databases.

  <https://github.com/murilovarges/HARBoP>