Nicolò Bertozzi, Francesco Bianco Morghet
Machine Learning and Deep Learning
Master Degree in Data Science Engineering
Politecnico di Torino

First Person Action Recognition

Project Description

2^{nd} Semester | 10 July 2020

Table of Contents

Introduction

Related Works

Proposed Methods

Conclusions

Overview

Introduction

Related Works

Proposed Methods

Conclusions

Introduction 1/4

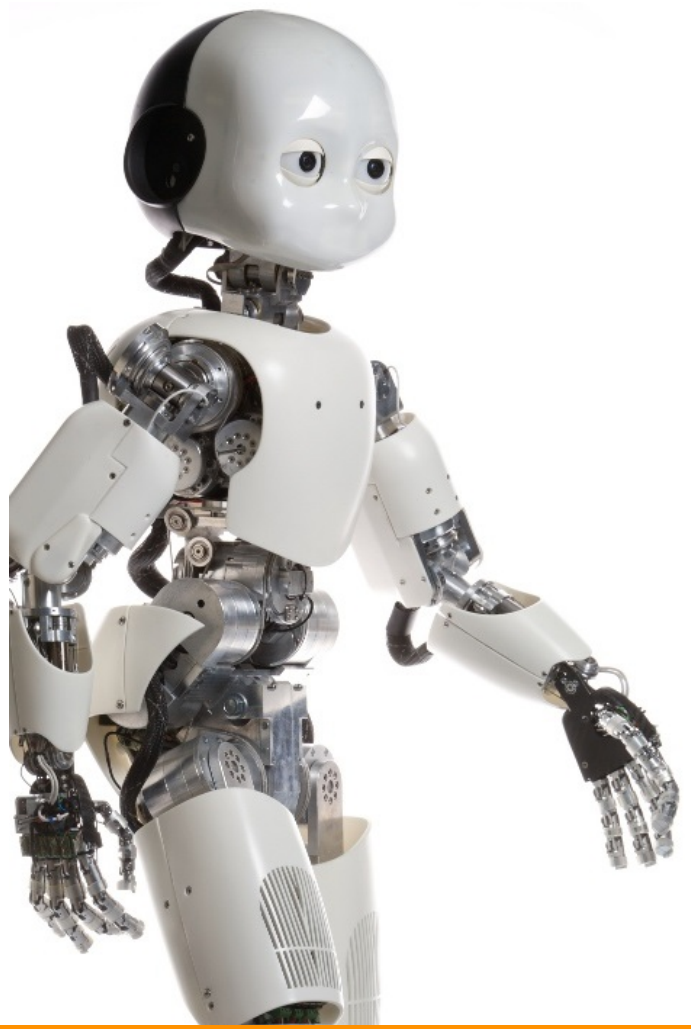
Goal:

- Record videos with the same cameraman's point of view;
- Recognize the actions performed by the subject;

Introduction 2/4

Interested Areas:

- Android intelligence;
- Autonomous driving;
- Surveillance;
- Loyalizing users' experience;



Introduction 3/4

Issues:

- Small datasets;
- Presence of **parts of the cameraman's body** in the video;
- The action **must be represented** by a *verb + noun*;

Introduction 4/4

Solutions:

- Sales of **wearable devices**;
- Incrementing chance of **having at hand a camera**;
- Incrementing number of **images taken every day** [**?**];
- Deeper neural networks;

Overview

Introduction

Related Works

Proposed Methods

Conclusions

Two Stream Approach 1/2

Main characteristics:

- Two **CNNs**: one to **extract features** from RGB images and one to **extract features** from flow images;
- **ConvLSTM** to take into account the **temporal dependencies**;
- Linear **classifier** to **join the networks**;

Two Stream Approach 2/2

Issue:

- The correlation and the **mutual influence** between motion and appearance information is **not taken into account**;

Solution:

- Implementing a single network accompanied by a **motion segmentation task**;

Attention Map

Features:

- **Focusing** the recognition on the **most important parts** of the video;
- **Discarding** the regions with **low importance**;
- The temporal flow information, i.e **the motion**, **is not included** in the mechanism;

Motion Segmentation Task

Features:

- Each **feature map** is forwarded to an **auxiliary branch** with a convolutional and a FC layer;
- **IDT** as ground truth: image which indicates if a **pixel is moving or not**, net to the camera motion;
- **Pixel-per-pixel** loss between the predicted motion map and the IDT;

Overview

Introduction

Related Works

Proposed Methods

Conclusions