数据结构说明

唐朝 唐朝

1. Directory layout |目录结构

```
[task]
├── episodes.csv
├── [modality 0]
│   ├── [data file]
│   ├── [data file]
│   ├── [data file]
│   └── ...
├── [modality 1]
│   ├── [data file]
│   └── ...
└── ...
```

The above diagram shows the directory layout for the data of a single task. Every task directory contains an episodes.csv file and multiple modality directories.

如上展示的是一个task的数据文件结构。每个task根目录包含1个episodes.parquet和若干个modality文件夹。

2. File definitions | 文件定义：

2.1 episodes.csv

Contains metadata and statistics for a task. episodes.csv has the following columns:

记录了一个task的各个episode的metadata和statistics。episodes.parquet包括了如下column项:

episode_id: *string* unique identifier for episode (IETF RFC9562 UUIDv1)

period: *int* total period of episode in seconds (rounded to an integer)

num_modalities: *int* number of modalities of data associated with the episode

metadata: *JSON* relevant metadata regarding the episode

modalities: *JSON* information about the different modalities of data collected (explained in detail later)

The **modalities** JSON contains the name and statistics of each modality data, eg.

对于**modalities**项JSON中包含各个modality的名字和相关统计数据,如下:

```
{
"color":{"frames": 370, "resolution":"1280*800", "sensor_model":XXXX}
"mocap":{"frames": 641, "sensor_model":XXXX},
"text":{"wordlength": 25}
}
```

2.2 modality directories and data modality |文件夹和data

Modality directories contain the data collected for each episode, stored as data files. Each data file has the same name as the **episode_id** of an episode. For instance, a file containing RGB image data for episode 0020770101132254 would be named 0020770101132254.mp4.

Every modality directory contains data of a single modality.For instance, color image data would be stored as compress MP4 files and text data as plain text files. However, data of the same modality could be stored in different formats or have different dimensions because they are collected differently.

Modality文件夹包含了各个episode收集到数据文件。每个数据文件的文件名和它对应的episode的**episode_id**是一样的。如一个包含了RGB图像数据的文件如果对应了episode 0020770101132254就会被命名为0020770101132254.mp4。

每个modality文件夹包含了一种modality的数据。如彩色图像数据会是MP4格式，文字数据会是plain text格式。有时，同一modality的数据可能因为采集方式不同而拥有不同的格式或者维度。