1. Dataset Pre-processing and Samples

In total, we collect 6,257 raw votes from 274 participants, of which there are 141 males and 133 females. The male-female ratio is roughly 1:1. The age group of the participants is from 17 to 32 years old. Moreover, data privacy is under the supervision and protection of the Institutional Review Board (IRB) of our university.

Before applying the collected datasets to model training, we conducted pre-processing operations.

* We find that the interval between two adjacent votes from certain participants is too short (e.g., less than 5 minutes or even less than 1 minute), which can be considered as duplicate votes. Hence, we remove these duplicate votes from the datasets.
* We notice that the total number of +3 (hot) votes is deficient. There are only 31 “hot” votes received from 14 participants in our datasets. Thus, we consider that these “hot” votes cannot correctly reflect most of the participants' hot sensations and decide to remove them from both iTCM datasets. As a result, we will focus on the 6-point thermal comfort scale ranging from -3 to +2. However, we still share the datasets with “hot” votes.

Eventually, we obtain three datasets: one iTCM generic dataset and three iTCM personal datasets for thermal comfort modeling, as listed in Table 1.

Table 1 Information of modeling datasets

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Size (with hot votes)** | **Air Temperature Range** | **Relative Humidity Range** |
| iTCM generic dataset | 4293 (4314) | 19.6°C - 30.6°C | 37.3% - 83.6% |
| iTCM personal datasets | 345 (346) | 19.6°C - 29.9°C | 42.4% - 75.5% |
| 380 (385) |
| 341 (345) |

2. Privacy

In our dataset, we have participants’ age, gender, weight, and height. However, we do not contain the identification information (e.g., name, ID number, etc.). Each participant is anonymous and linked to a unique UUID.