Thermset: A thermal-video dataset of elders at their bedrooms

Pablo Pusiol
pdp0109@famaf.unc.edu.ar

Federico Polacov
fdp0108@famaf.unc.edu.ar

Universidad Nacional de Córdoba - FaMAF
October 19, 2016

1 Introduction

Our interest is to automatically detect clinically relevant daily activities of seniors living independently or in nursing homes. To do that we will use thermal information and develop innovative computer vision technology. Modern computer vision is based on deep learning techniques and has proven to be very effective at automatizing visual recognition tasks such as detecting objects in images, tracking and detecting activities in videos, etc. A caveat of these techniques is the need of big amounts of labeled data for training their models. In practice, to achieve models that could reliably work in new environments (e.g. homes, people, etc.) would require of the training dataset to contain activity examples covering the full space of the manifold defined by all possible configurations of the target activities and environments.

While video cameras are traditionally used for daily activity monitoring, they need additional algorithms to overcome their inherent vulnerability to low light conditions, too much light, shadows, complex scenes. Far infrared sensors (thermal sensors) do not have any these issues because they create a crisp image based temperature data from a scene, so almost all the problems traditional RGB cameras face in processing and classifying images are avoided [1].

In this document, we introduce a new dataset: Thermset, an incremental dataset of thermal information recorded for different time intervals of (so far) 3 senior citizens at their bedrooms with automatic annotations as described in subsection 2.4. Each of the instances of the dataset (clips) corresponds to a sequence of frames where each frame corresponds to a matrix of temperature values captured by a far-infrarred sensor. The sensor captures for intervals of as little as 2 seconds or as long as 15 seconds. Until the date of writing this document, Thermset has about 120 hours of recordings, where typical examples of data reflect an elder sleeping, wandering, getting ready to bed, receiving assistance from a caregiver. The dataset also contains thermal data of the different bedrooms with no people as shown in Figure 1 Thermset is available through `http://forelderly.weebly.com/data`.

In section 2 we provide technical details of how Thermset is structured, characteristics of data, annotations and how data is being gathered. The rest of the document is structured as follows: In section 3 we provide qualitative information of the seniors recorded and environments and finally in section 4 we provide an example of how impactful Thermset can be to improve seniors wellbeing.

2 Technical details

2.1 Structure

Thermset is structured in groups. A group is a sequence of short clips with the same duration. A clip is composed by a sequence of temperature values matrixes expressed in Celsius degrees (from now on we will call each thermal matrix a frame). Each group has clips of a unique person recorded for a certain time interval with a certain configuration of our recording tools as we will describe in this section. There is a time gap between each clip and its duration depends on the recording configuration used as explained at paragraph 2.3.1.

Group names follow the structure

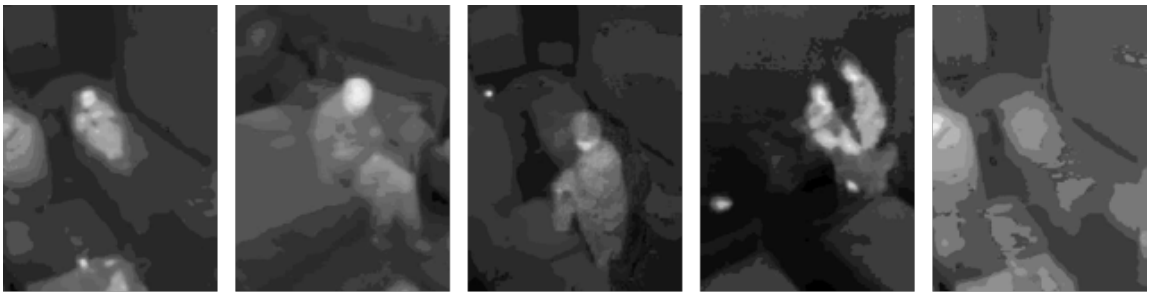

Figure 1: *Examples of thermal data extracted from Thermset. (Images where stretched for easy reading)*

Name_YYY, where Name refers to an elder's alias and **YYY** to the **YYY**-th group recorded of that elder. For instance, the group **Marge_001** is the 1st group with recordings of an elder whose alias is Marge.

2.2 Data analytics

At the moment of writing this document, Thermset is composed by 9 groups of 3 different people. Two groups (Peter_001 and Fiona_001) were collected at each elder's home and the others in 2 different nursing homes. Until now:

- **Number of clips: 50574**
- **Total days recorded: 14**
- **Total hours of clips: 120 hs**
- **Number of people: 3**
- **Number of groups: 9**
- **Number of days per person:**
  - **Marge: 10**
  - **Fiona: 3**
  - **Peter: 2**
- **Hours of clips per person:**
  - **Marge: 41 hs**
  - **Fiona: 42 hs**
  - **Peter: 38 hs**

2.3 Recording tools

To record and create annotations we use Apple iOS devices and FLIR One sensors. The app used was Thermix[6] , available for free in the App Store. Thermix encodes videos in H.264 format and sends them to the cloud.

We relied on 3 different devices to run Thermix: an iPhone 5, an iPhone 5s and an iPod Touch 5th generation. The far-infrared sensors used were Dongle and SLED versions of the FLIR One [4][3]. We configure Thermix in the iPhone 5 with an SLED sensor and in the iPhone 5s and the iPod Touch with a dongle sensor.

2.3.1 Recording Configurations

Groups in Thermset were recorded using different configurations. Information of the configuration and device/sensor used for each group is available at `http://forelderly.weebly.com/data`. Below we describe the 4 aspects of each configuration.

Thermal resolution FLIR One SDK for iOS devices delivers a resolution of 240x320 for Dongle sensors and 120x160 for SLED sensors [5].

Temperature precision For R = *raw thermal data returned from the sensor* we construct a frame F, where $F(i, j)$ corresponds to the element i, j of the frame adjusted by precision.

- **Low precision:**

$$F(i,j) = \begin{cases} 2 * R(i,j), & \text{if } 1 \leq R(i,j) \leq 120 \\ 0, & \text{Otherwise} \end{cases}$$

Where $R(i, j)$ is the temperature expressed in Celsius degrees with no decimal places.

- **High precision:**

$$F(i,j) = \begin{cases} round(2 * R(i,j)), & \text{if } 1 \leq R(i,j) \leq 120 \\ 0, & \text{Otherwise} \end{cases}$$

  Where $R(i,j)$ is the temperature expressed in Celsius degrees with one decimal place. The *round* function returns the integral value nearest to its argument, rounding half-way cases away from zero.

Clips length Each group has clips of a certain duration that ranges between 2 and 15 seconds. Because Thermix excludes all frames during a sensor recalibration, the number of frames in a clip will be less if a recalibration occurs. We estimate the average fps (frame per second) depending on each device achieving 5fps, 3fps and 2fps for the iPhone5+SLED, iPhone5s+Dongle and iPod+Dongle respectively.

Dead intervals Thermix records and uploads small chunks of video to the cloud. There are two possible configurations: either Thermix saves chunks to the device disk continuously (continuous mode) or waits until a chunk is sent to start recording the next one (one-by-one mode).

- **Continuous mode:**
  Clips in groups recorded with continuous mode are separated by a small time window (between 2 and 5 seconds in average), that corresponds to the time Thermix takes to compose the chunk[1]. This means, Thermix did not wait for a video chunk to be sent before start recording the next one.

- **One-by-one mode:**
  Clips in groups recorded with one-by-one mode are separated by the time Thermix takes to compose the video chunk as in continuos mode plus the time the device takes to upload it to the cloud.

[1]The composition of the chunk consists in generating a video from thermal data using the encoding H.264 and annotating the first thermal frame using the method described in section 4

id	*class*
1	lying
2	sitting
3	standing
4	group
5	background

Table 1: *Classes and their corresponding ids.*

Figure 2: *Frame with annotation: 2016-09-06_03.53.33 1:0.998713 4:0.000638 2:0.000534 5:0.000074 3:0.000040*

2.4 Annotations

Thermset provides automatic annotations generated by Thermix for each clip. The annotation corresponds to the classification of the first frame of the clip using the method described at section 4. Classes are background, standing, sitting, lying and people. Annotations for each group are listed on file (named **GROUP_NAME**_annotations.txt) stored in each group folder. Annotations follow the structure:

clip_url $\{<class\_id{:}score>\}^5$,

where *class_id* corresponds to classes as detailed in Table 1, and *score* corresponds to the output of the classification for each *class_id*. An example of an annotation is shown in Figure 2.

3 Qualitative description of the dataset

Recording environments Videos where recorded at several bedrooms in 2 different nursing home facilities and at 2 private homes. All videos where recorded during winter.

About the people Until the date of writing this document, Thermset has data of the following elders:

- **Marge:** 90+ years old, Female
  Diseases or disabilities: Hyperthyroidism. Had fractures due to falls
  Situation: Living at a healthcare facility. Shares the bedroom with another person. Takes naps daily. Sleeps with a diaper. Had a hip fracture after a fall. Moves around with walker and shes afraid of walking alone. Requires constant assistance for moving, entering, leaving bed, etc.

- **Fiona:** 75+ years old, Female
  Diseases or disabilities: Overweight, Hyperthyroidism
  Situation: Lives at her place, alone. She does not require assistance. Takes naps and watches TV while in bed.

- **Peter:** 80+ years old, Male
  Diseases or disabilities: Had fractures due to falls
  Situation: Lives at his place, alone. He does not require assistance, but due to a knee fracture has difficulties to move. Uses a walking stick. Takes naps and reads the newspaper while in bed.

4 Application: Pose estimation

In the context of daily activities detection, collecting a training set large enough to accurately detect clinically relevant activities is hard. It might require of hundreds of data feeds for several days, or years if we consider the environmental temperature fluctuations and the different appearances (e.g. clothes) that a person will have during the stations. In addition, it would require of several human annotators working 24/7 watching and extracting the target activities appearing in the streams. Likely, the number of labelers would be proportional to the number of data streams that need to be analyzed. In addition to the complexity of collecting a big labeled training dataset, the size of each labeled example could present a problem as well. Optimizing an utility function over long-term visual streams would require a massive amount of computing power and memory.

To overcome the limitations described above, instead of thinking in rigid models to detect all activities at once it is more convenient to design dynamic systems capable of learning incrementally new activities when new labeled data arrives. Daily living activities are hierarchically composed of sub-activities occurring in a variable period of time. In general, the deeper we move in the taxonomy of activities, the less training-data is needed for building reliable models to detect them. This observation is related to the lower dimensionality of their manifold in comparison with more complex activities. We are interested in detecting complex activities in incremental bottom-up steps. First, by detecting anchoring activities lying at the lower end of the human activities taxonomy. These anchor activities are persistent human features appearing as sub-activities of more complex ones. Enabling a robust detection of these anchors will enable our second stage of learning, which will use the anchors for detecting and analyzing complex long-term human activity. For the rest of this section we will focus in the first stage of our algorithm: defining and modeling the detection of anchoring activities.

Pose Anchors One of our anchor activities is the detection of pose. Human pose is a highly descriptive and persistent feature, capable of characterizing complex activities. For senior adults living independently, changes of pose in the particular contexts could be describing a high risk situation. For example, a person changes from "standing" to "lying in

the floor" could indicate the occurrence of a fall. In nursing-homes, detecting and alerting the "coming out of bed" could enable promptly caregiver's assistance.

The detection of pose anchors is connected to the monitoring of clinically relevant activities. Here we borrow some of the target daily living activities described in [7], where several of the activities presented there can be inferred directly from the human pose. For example:

1. **Falls:** corresponds to transitions from standing to lying or from sitting to lying -as long as lying is not happening in the bed area
2. **Front-door-Loitering:** corresponds to detecting a person standing in selected locations
3. **Day-Night-Reversals and night wanders:** corresponds to detecting a person standing at unusual night hours
4. **Sleep:** corresponds to detecting a person lying in bed, even when the person is covered by a blanket
5. **Immobility (bed or chair):** corresponds to long periods of time where the person is either sitting or lying (in bed)
6. **Restlessness:** corresponds to detecting and tracking pose changes for long periods of time where the person is perpetually agitated or in motion.

4.1 Model

We built our model as a Fully Convolutional Network derived Matt Zeiler's model [8], which takes an input of 225x255x3 thermal image is convolved through 5 layers. The last layer is fully connected, taking features from the top convolutional layer as input in vector. The final layer is a 5-way softmax function, being 5 the number of classes: *lying*, *sitting*, *standing*, *group* and *background*. All filters and feature maps are square in shape.

4.2 Dataset

The dataset used to train this model was labelled manually and contains 25326 frames with approximately 5055 frames by class, where 8362 were selected from Thermset clips and 16964 where collected from a controlled environment.

We randomly split 80%/20% of data into training and testing datasets. Since our training units are video clips, we were careful of not including any frame of a testing video in the training set (and vice versa)

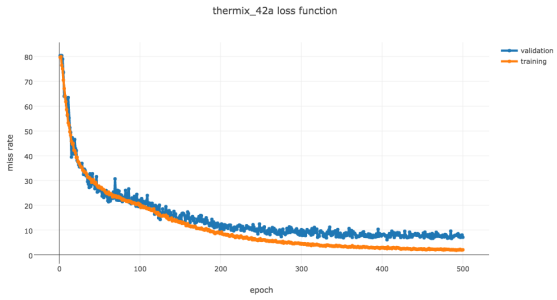

Figure 3: *Cross entropy loss for logits*

4.3 Evaluation

Training was done for 500 epochs with a Dropout rate of 0.5. Results were obtained running the algorithm in the testing dataset, showing a generalization error of 7.09%. Our loss function is cross entropy for logits. (See Figure 3).

Thermix provides a publicly available account where this model is being used live for presence detection for 3 people living in a nursing home, this helps nursing home staff to monitor activity in those scenes. Access can be requested via `http://forelderly.weebly.com/data`.

5 Discussion and further work

5.1 Extending Thermset

Annotations Thermset provides automatic annotations created using the procedure explained at section 4. To train the model, we labelled 25326 frames. Creating different ground truths, such as localization of people would allow a better understanding of the scene.

5

Size Thermix provides a plug-and-play interface to easily gather thermal data. So far we collected 120 hours during 14 of 3 seniors. At the moment of writing this document, Thermix is set up in three different locations and we have plans of augmenting the dataset with more seniors in different locations.

Exploiting Thermset Thermset provides hours of recordings of seniors being aided, sitting, wandering, people that felt to the ground, nurses, etc. Such data is highly valuable for training new algorithms aimed to help the elderly live better and benchmarking existing algorithms in real life situations.

Acknowledgement The authors would like to like to thank all the nurses and doctors at the healthcare facilities from where we collected the data. Thank also Professor Jorge Sanchez from National University of Córdoba for his helpful remarks.

References

[1] Hong Chengl, Zicheng Liu, Yang Zhaol, Guo Yel *Real world activity summary for senior home monitoring.* 2011.

[2] FLIR One SDK, `http://developer.flir.com/sdk-documentation/`, accessed August 2016.

[3] FLIR One SLED Tech Specs, `http://www.flir.com/flirone/press/FLIRONE_Fast_Facts_Tech_Specs.pdf`, accessed August 2016.

[4] FLIR One Dongle Tech Specs, `http://www.flirmedia.com/flir-instruments/industrial/datasheets/flir-one-datasheet.html`, accessed August 2016.

[5] FLIR One SDK Documentation, `https://developer.flir.com/wp-content/uploads/2015/06/documentation-FLIR-One-SDK-iOS-DOCSET.zip`, accessed August 2016.

[6] Thermix for FLIR One, `https://appstore.com/thermixforflirone/`, accessed August 2016.

[7] Stanford PAC Seniorcare website, `http://vision.stanford.edu/pac/seniorcare/`, accessed August 2016.

[8] Matthew D. Zeiler and Rob Fergus *Visualizing and Understanding Convolutional Networks*, ECCV 2014