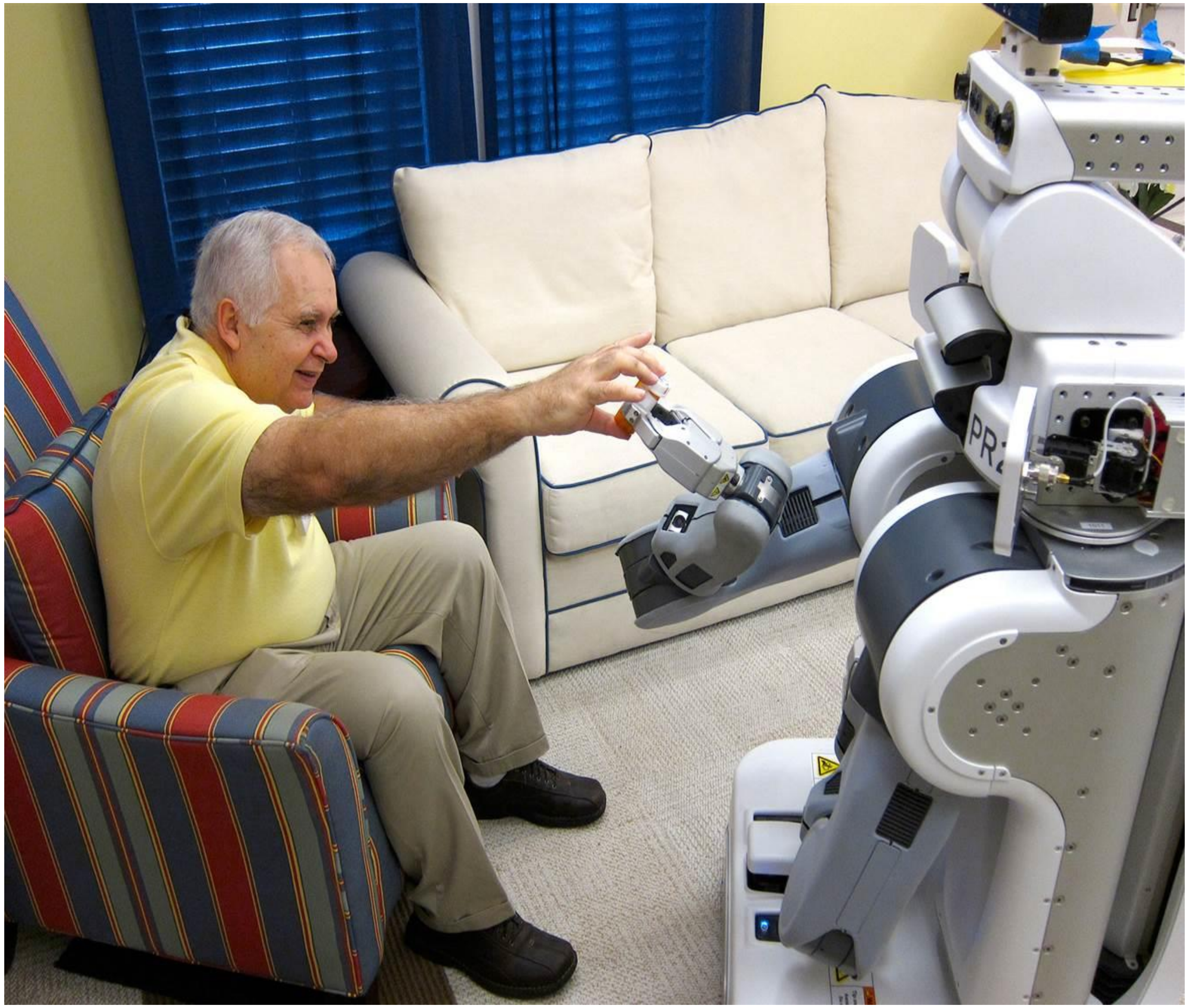
Project A4

Cross-modal joint sparse feature learning in robot tasks

Changshui Zhang
Dept. of Automation, Tsinghua University
Jianwei Zhang
Dept. of Informatics, Uni. of Hamburg

A4 – Aims and Strategies

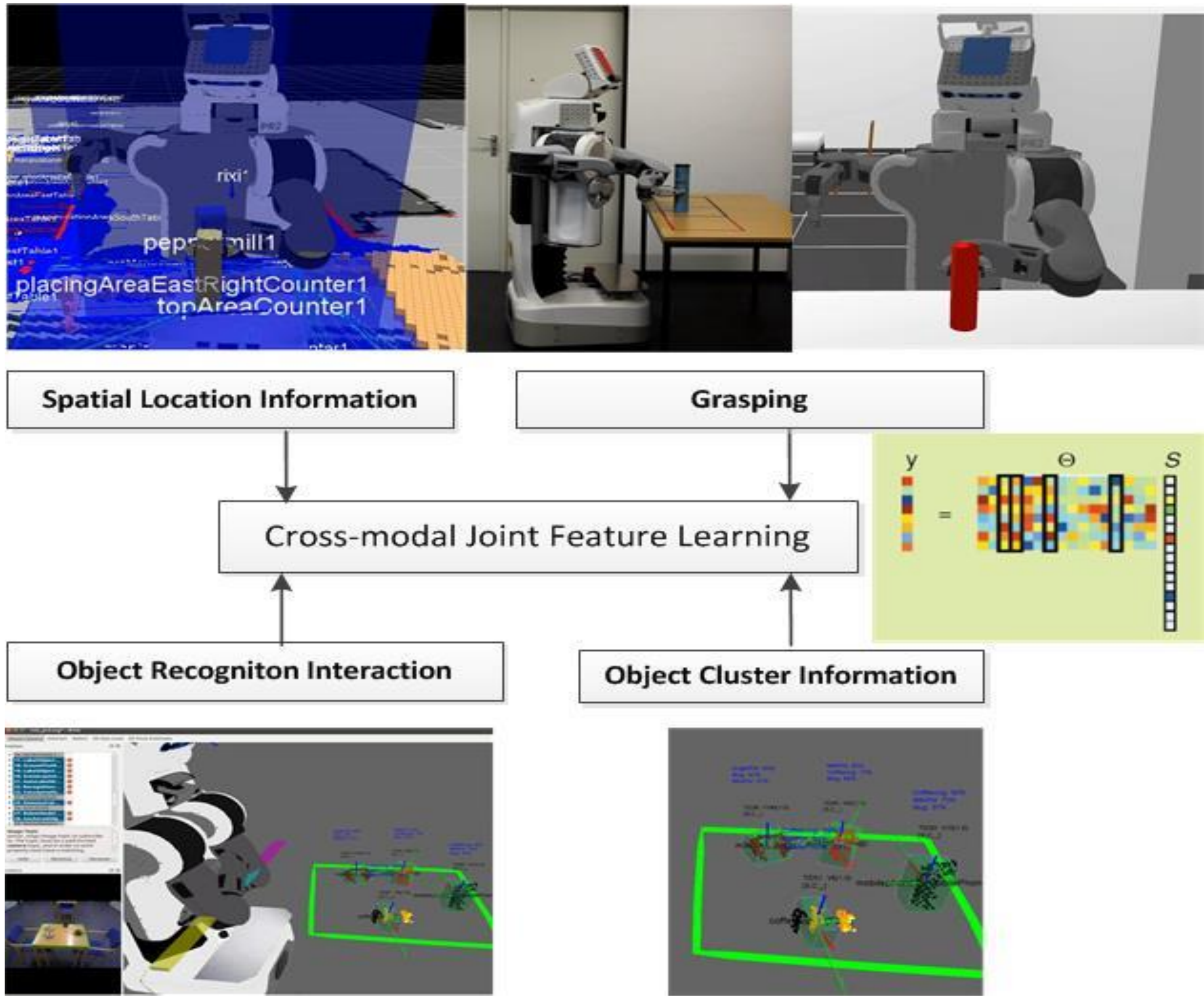
- Establishing a cross-modal feature learning method that is adaptive to the robot scenario using the modalities of image, sound and tactile signals as input
- Designing classifiers that are adaptive to new unknown categories
- Applying the trained model to the robot system and complete the task of assisting elderly people in taking medicine



- The categories of medicines can be distinguished by different kinds of signals such as visual, audio or tactile information

- The cross-modal features can perform better than mono-modal ones since different categories' information are shared by the same kinds of features

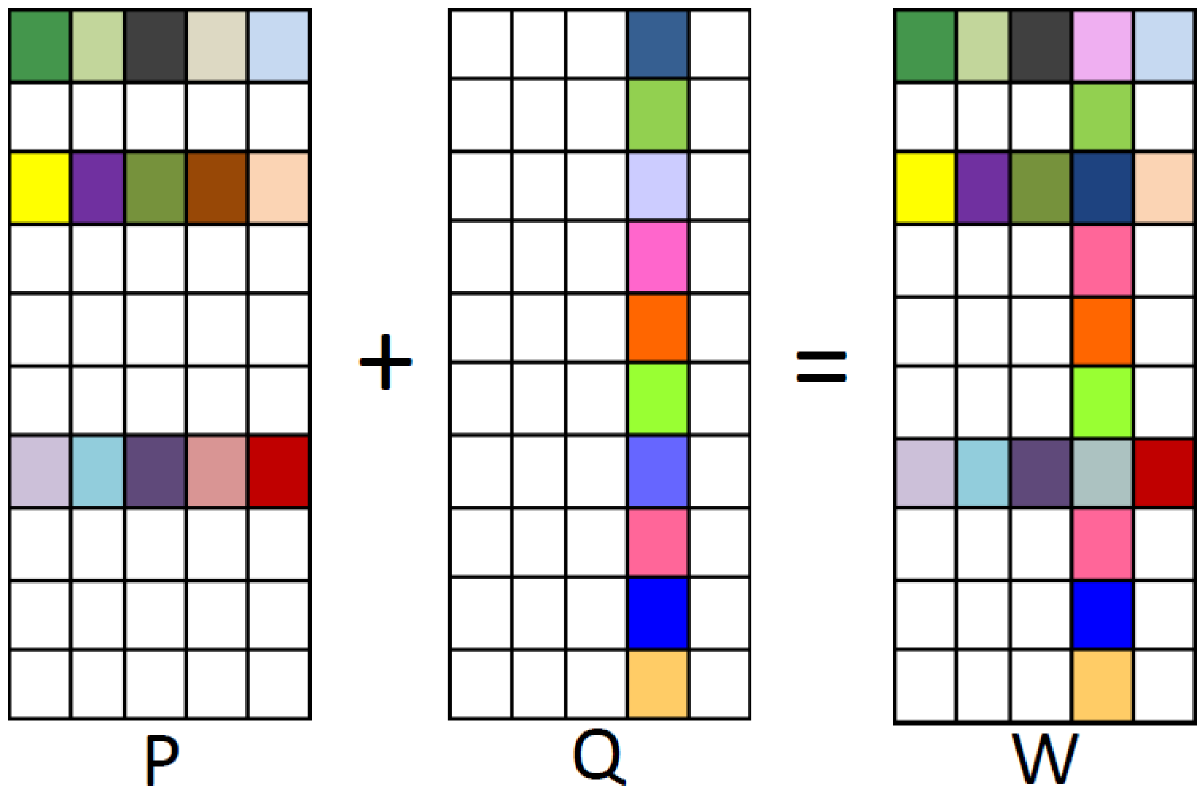
- The joint features of the samples are supposed to be sparse representations so that the learned joint features have the separability for different categories of samples

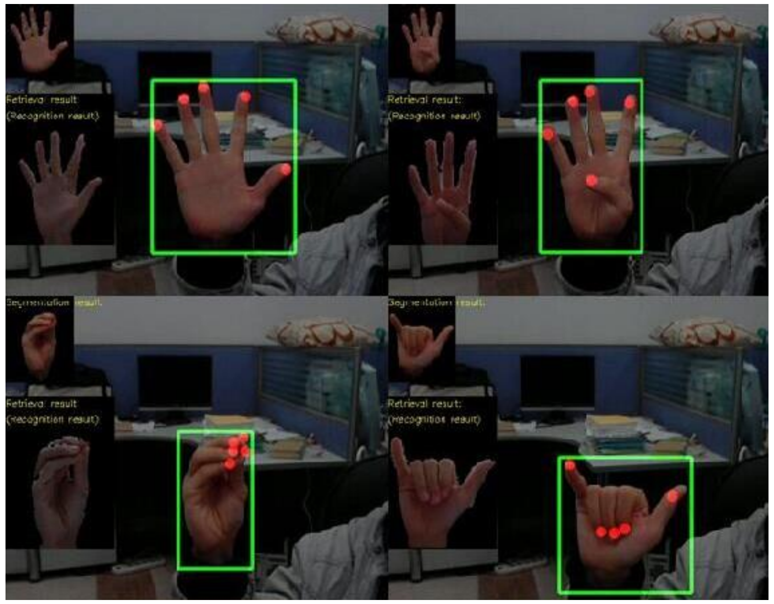


SFB TRR 169

A4 – Own Previous Work

- **Robust multi-task feature learning**: decomposes the weight matrix into two sparse matrices that denote the *shared features* among tasks and the *outlier tasks*.





- **Hand-pose recognition**: Retrieval-based template matching algorithm that estimates the pose of a hand in the video.

- **Traffic sign recognition**: Detection based on deep convolutional neural networks

A4 – Research Plan

- **Task 1: Acquisition and annotation of image, sound and tactile data**
- We will exploit the robot platform set up by project Z3/II-R and develop general cross-modal learning methods.

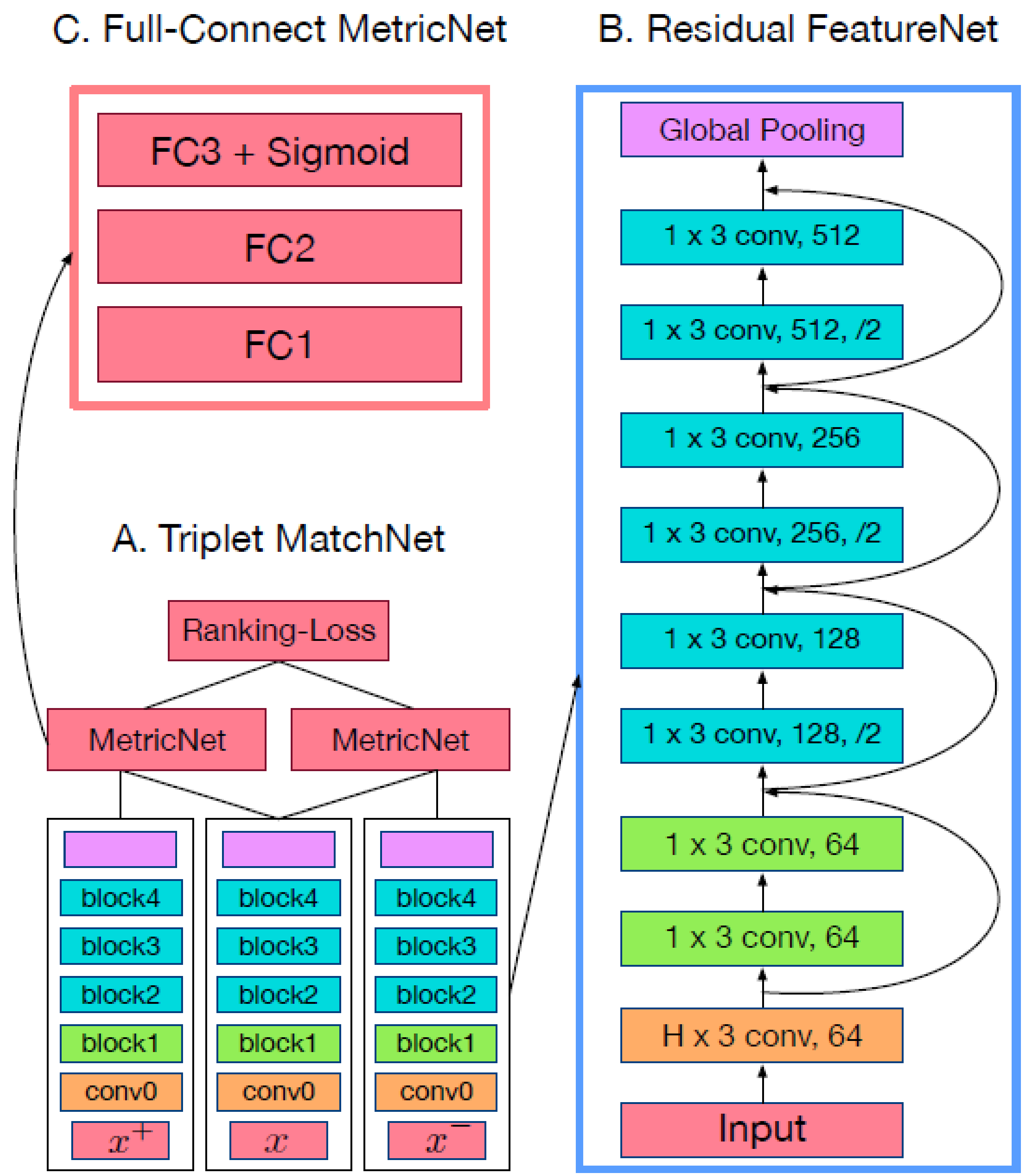
- **Task 2: Algorithms for multi-modal joint feature learning**
- We will mainly focus on multi-task sparse learning algorithms that enable the robot to process different tasks using cross-modal features.

- **Task 3: Classifiers suitable for new object categories**
- In realistic circumstances, the robot may encounter unknown medicines, which requires the robot to recognize objects from new categories.

- **Task 4: Application of the models and methods to a robot system**
- Based on previous tasks, the models will be applied to the robot-assisted elderly medicine task in order to verify our approach.

A4 – First Experiments

- Deep model for content-based similarity learning: Triplet MatchNet.
- Since we have lots of works in the field of CV, we explore audio feature extraction.
- The model is made up of three main parts: feature extraction layers, metric calculation layers and a rank-based loss layer.



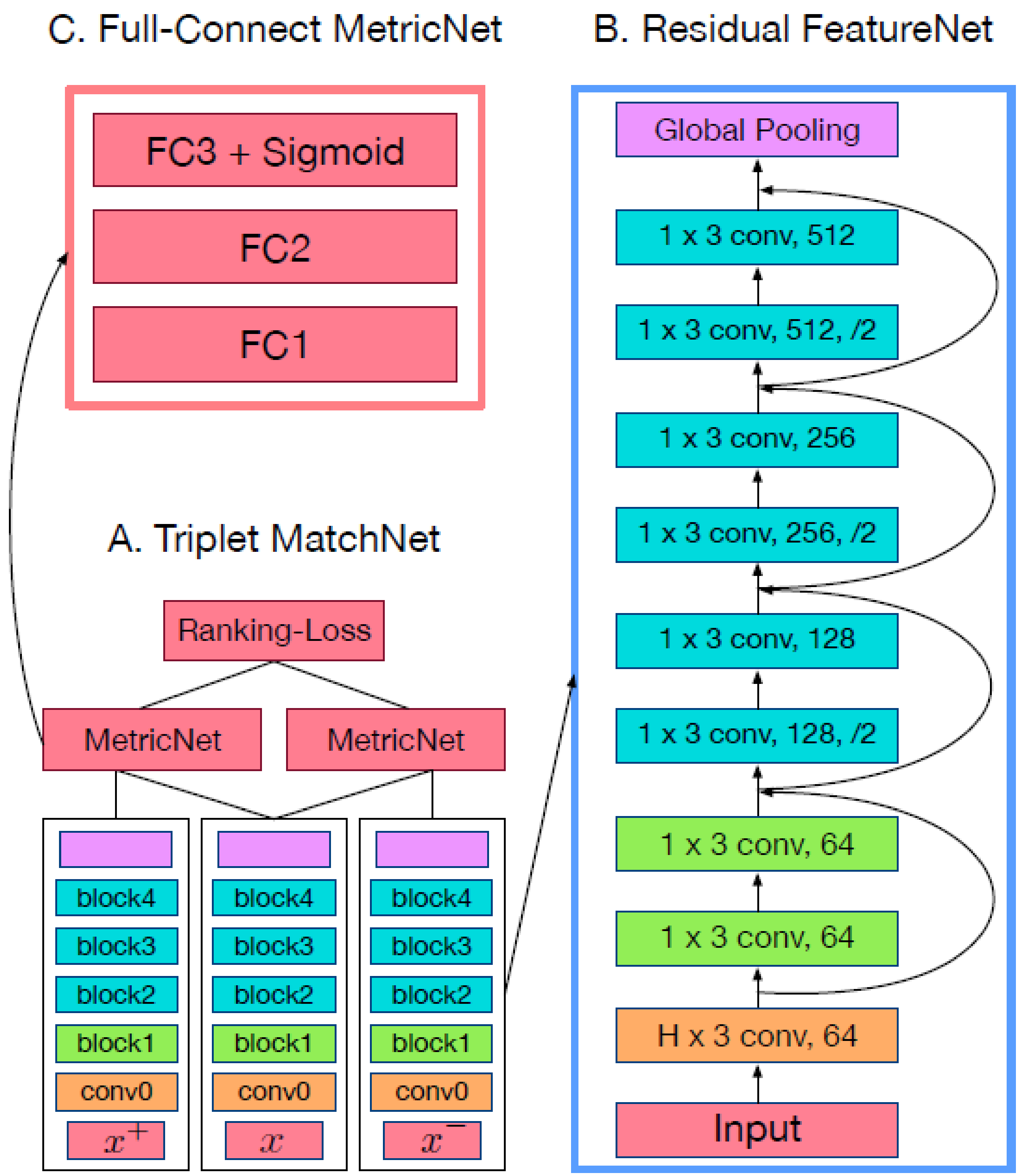
- Relative similarity:
  "(A, B) are more similar than (A, C) "

- Given a query data x and its corresponding positive/negative data x^+/x^-, we feed them to the model:

$$d^+ = f_W(x, x^+) = G(F(x), F(x^+))$$
$$d^- = f_W(x, x^-) = G(F(x), F(x^-))$$

- The design a continuous version of partial-order that describing the ranking-loss:

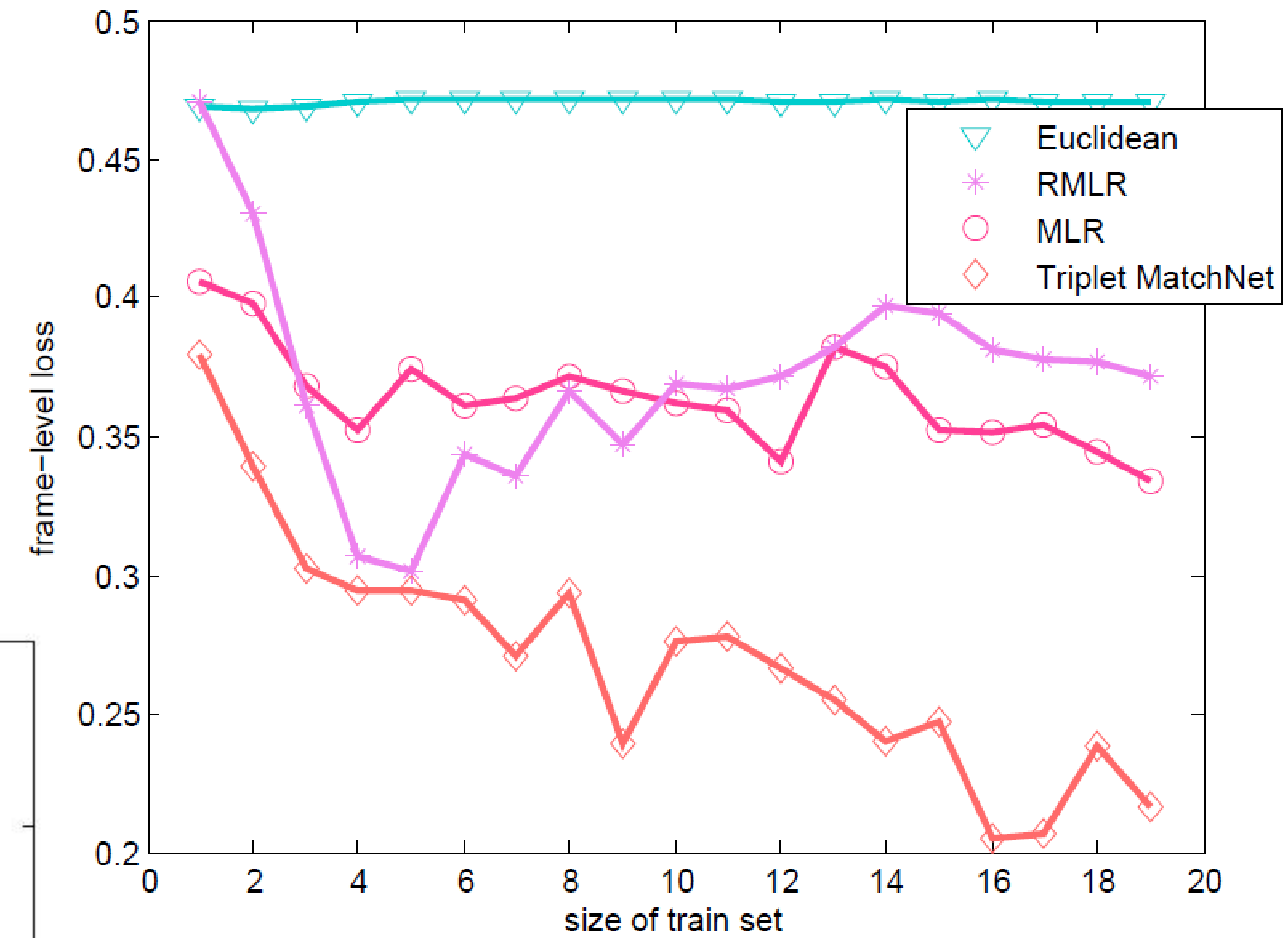
$$\hat{\psi}(x) = \sum_{x^- \in \chi^-} max\{0, d^+_{max} - f_W(x, x^-)\}$$

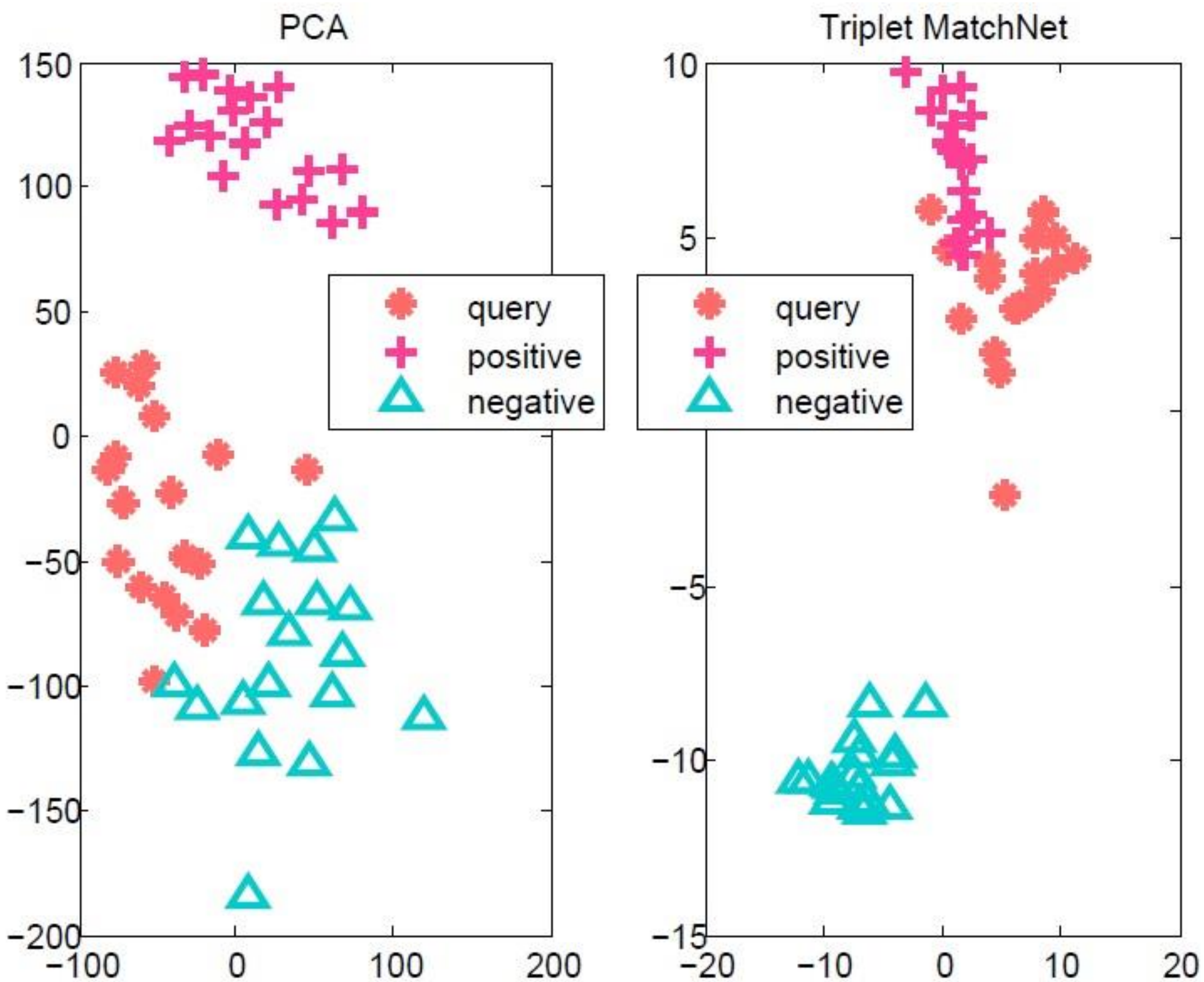


SFB TRR 169

A4 – Experimental results

- Training curve shows generalization ability of our model compared to traditional ones.





- Features extracted by our trained model shows favourable separability.

- **Rank-based** training strategy is effective in modeling content-based audio separation task.

- Flexible **end-to-end** framework.

- **Sparse representation** by residual network.

- **Current progress**
- We show that deep sparse representation for audio data is feasible.

- **Multi-modal feature learning**
- Image features and tactile features

- **Classifiers adaptive to unknown categories**
- Dictionary-based searching algorithm
- Learning the categories' boundaries