Human Intention Prediction Using BIL-SCNN

Ma Tianqi Zhang Lin Diao Xiumin Ma Ou

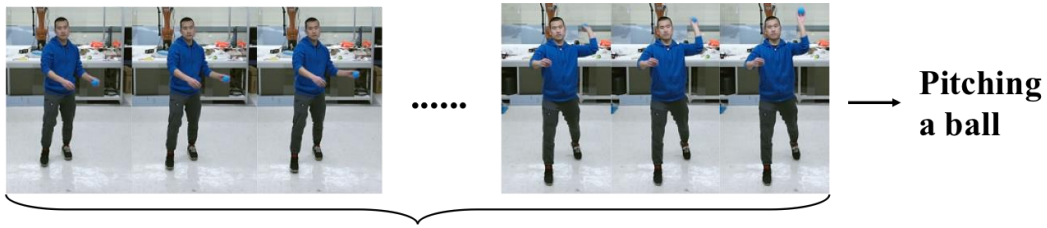
Active Vision and perception in Human(-Robot) Collaboration (AVHRC 2020)

Aug 31, 2020

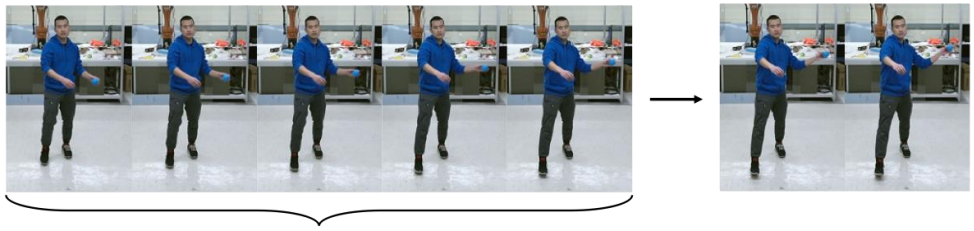
PROBLEM STATEMENT

SUMMARY

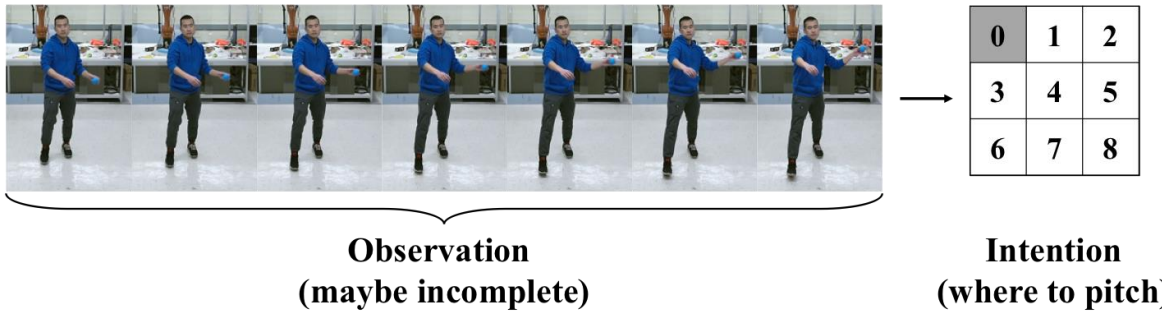
Action Recognition



Action Prediction



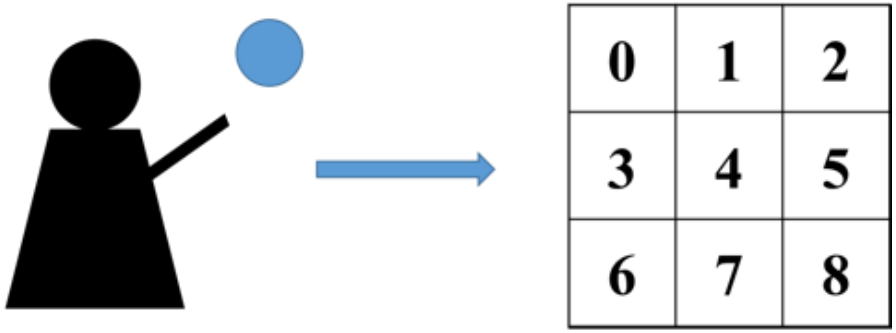
Intention Prediction



Tsinghua University

PROBLEM STATEMENT

DATASET



Ball pitching

9 blocks

- **Pitch to 9 target blocks in 3x3**
- **Record actions that pitch the ball to the intended block.**

- **Connect the action outcome with the intention**
- **Label the action with the intention**

Predict the intention with the method of action recognition (A simplified problem)

Tsinghua University

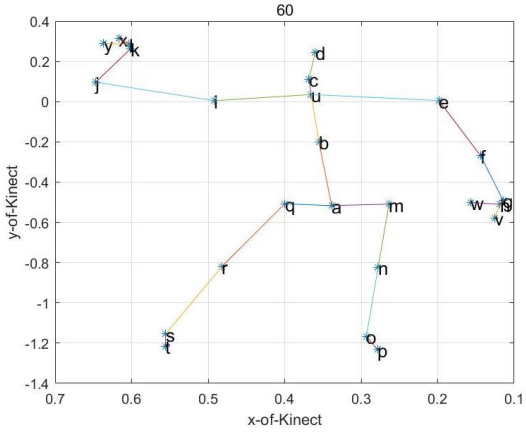
PROBLEM STATEMENT

DATASET

Kinect V2

- RGB images
- Depth images
- Joint positions

Input &
Starting frame

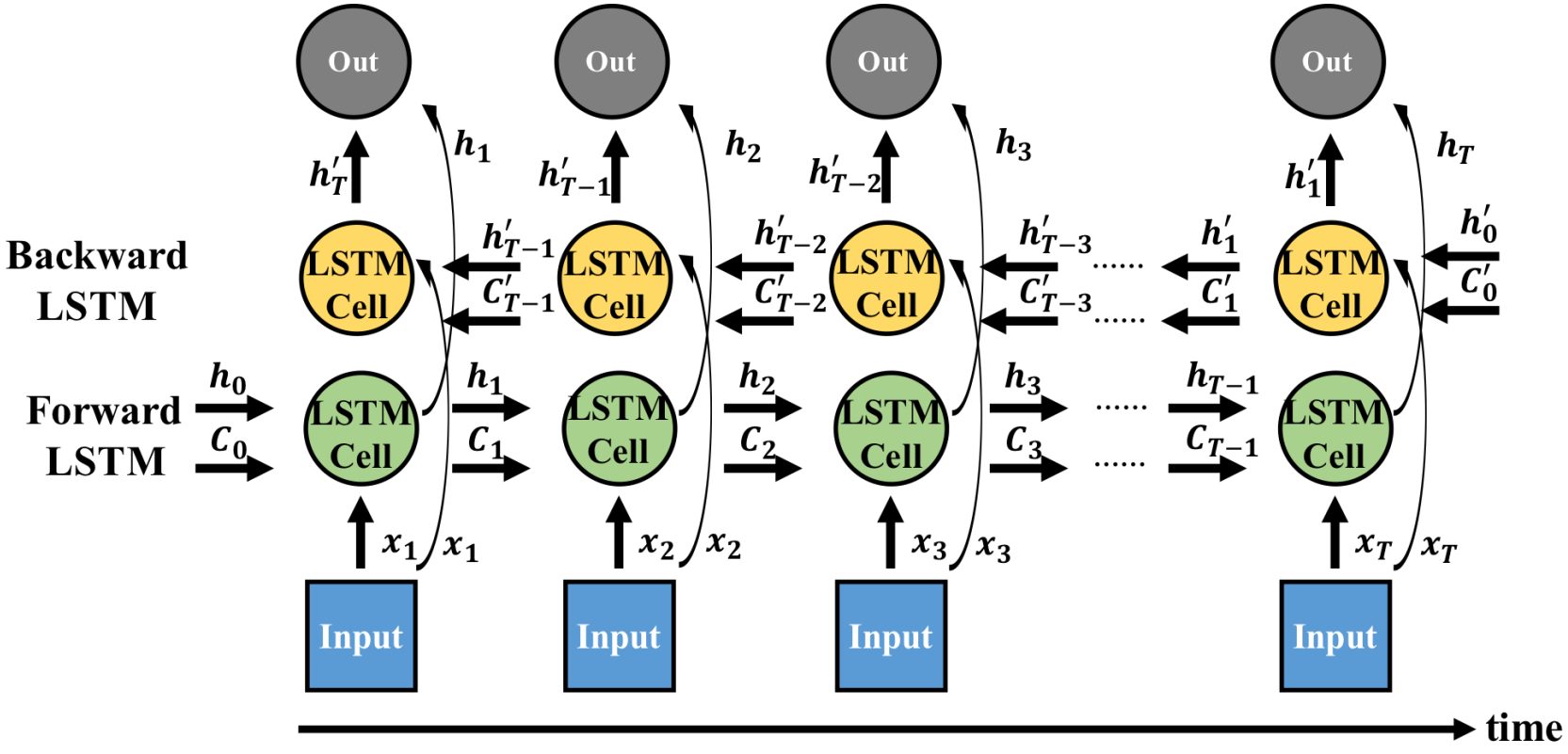


Starting frame

Tsinghua University

METHOD

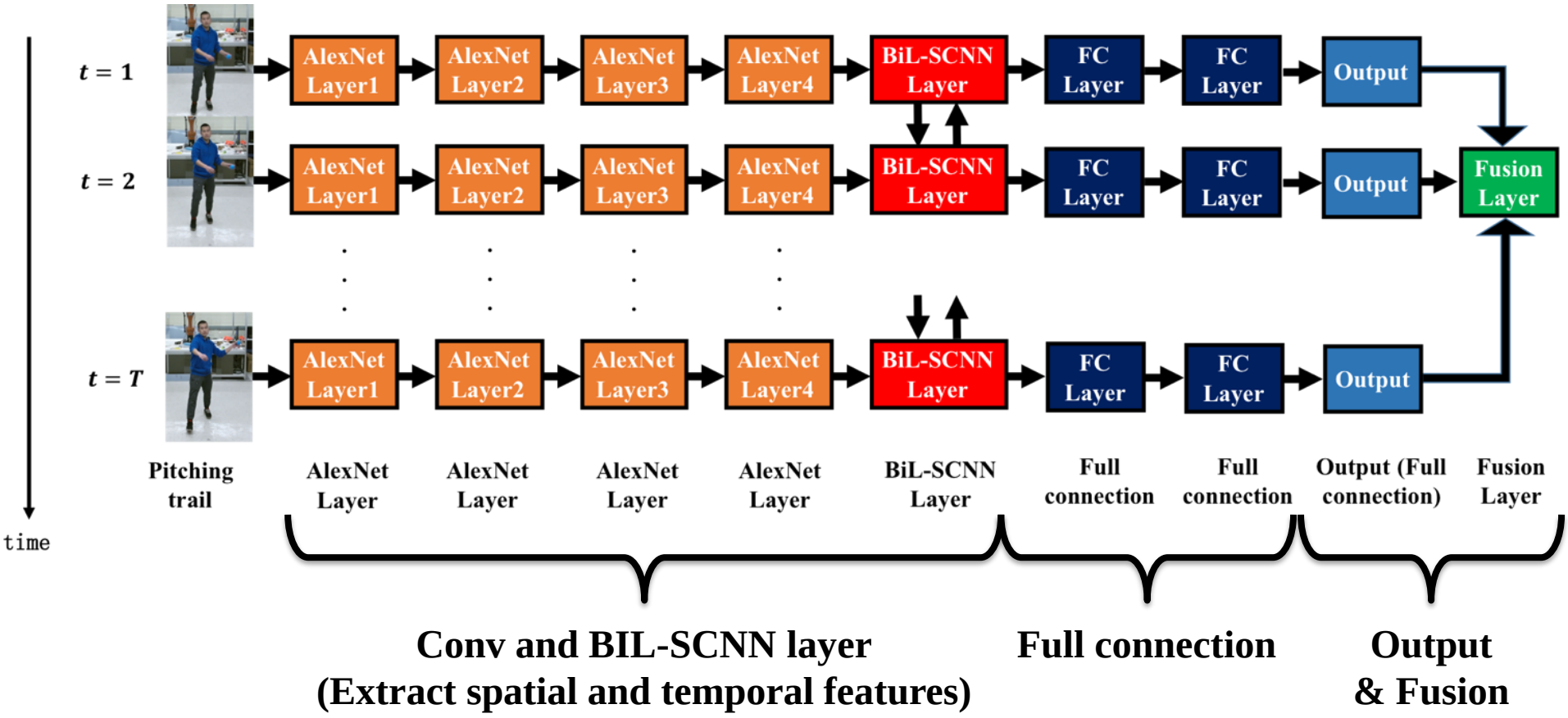
BIL-SCNN



- **Keep LSTM structure → Temporal features**
- **Replace the multiplication operation ($Wx + b$) with the convolution operation ($W * x + b$) → Spatial features**
- **Use bi-directional structure**

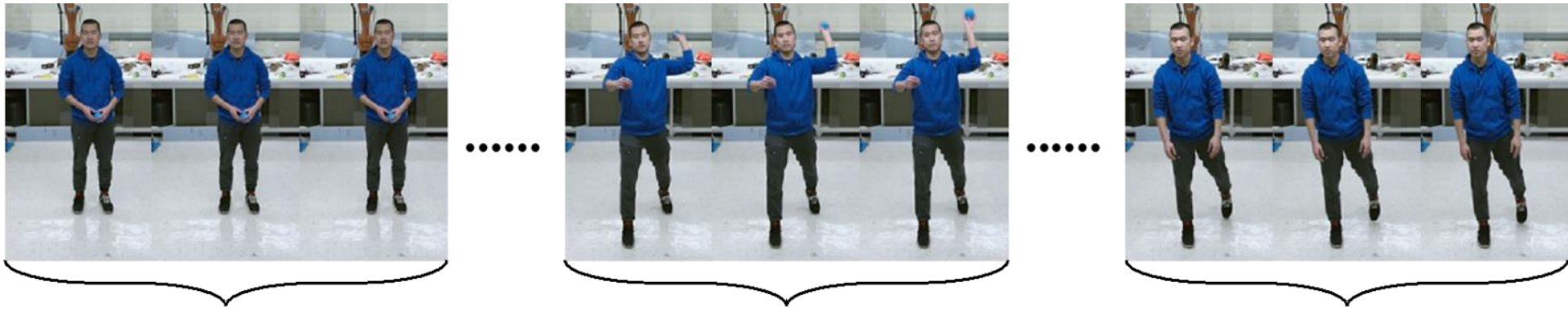
METHOD

PROPOSED NETWORK



METHOD

DETECT STARTING FRAMES



Useless frames (before pitching)

Useful frames (pitching)

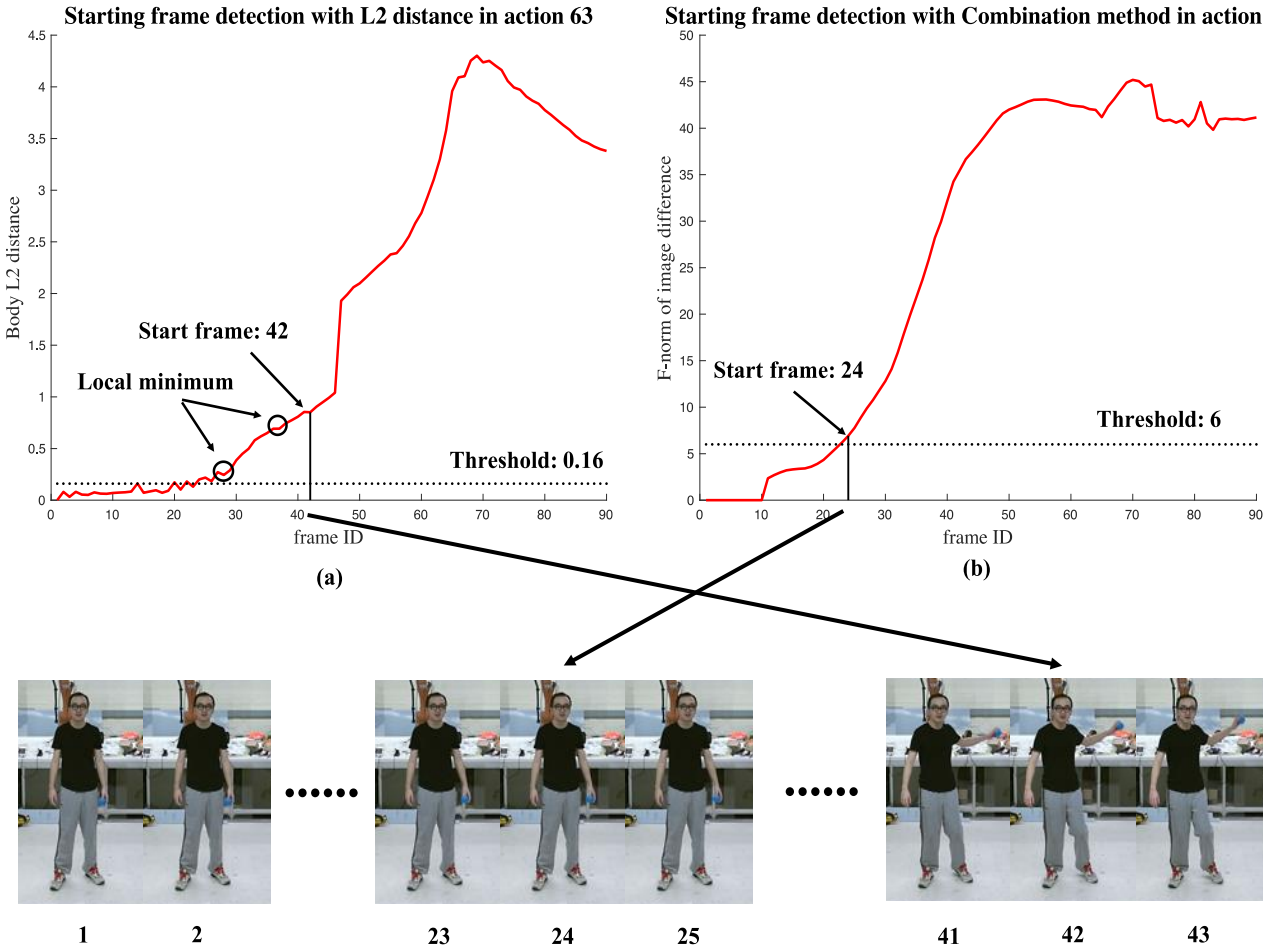
Useless frames (after pitching)

Where is the beginning of the action?

Tsinghua University

METHOD

DETECT STARTING FRAMES



A monotonically increasing period of the distance

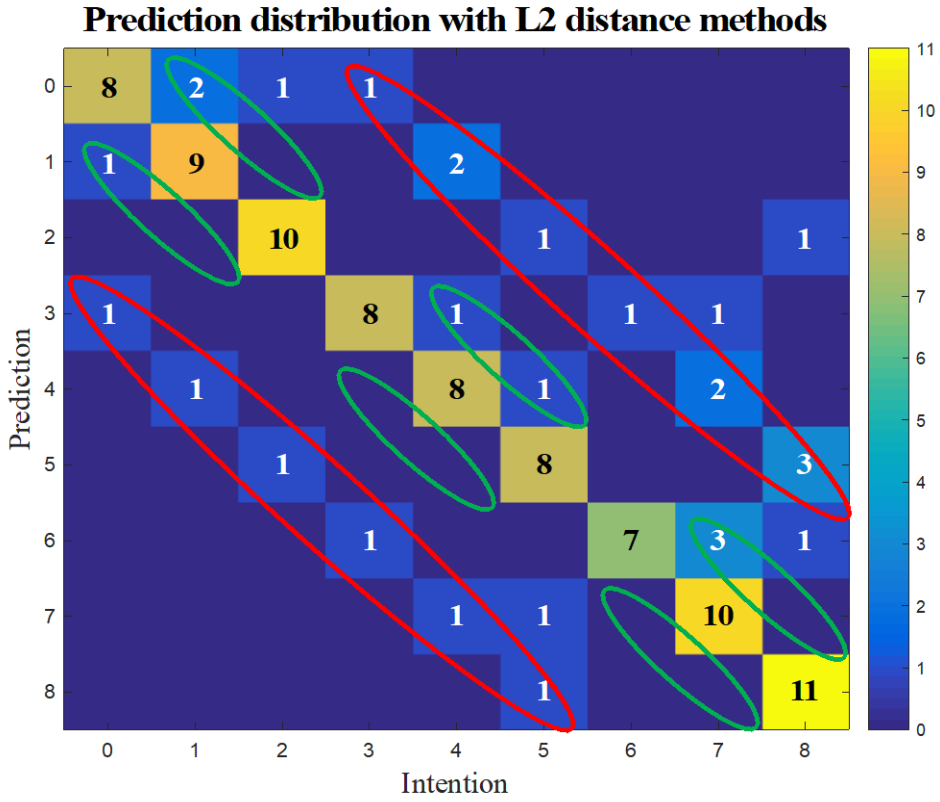
1) Joint positions:
L2 distance of joint positions increases during the action.

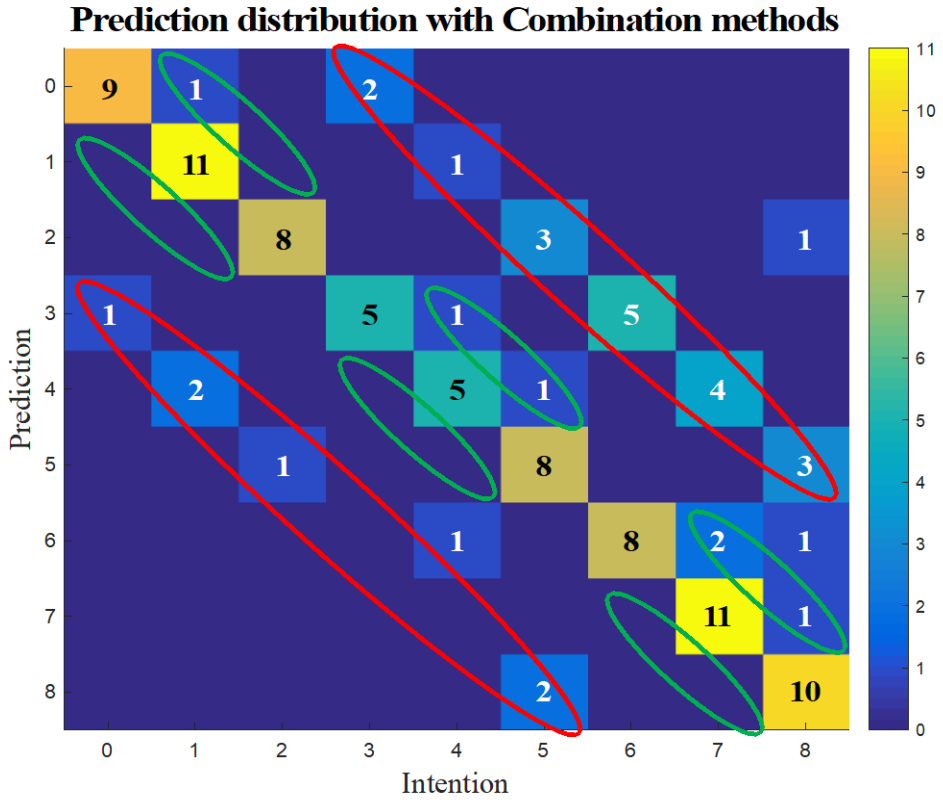
2) Joint + images:
differences between images increase more smoothly. Joint position of right hand is also used to more accurately locate the starting frame

Tsinghua University

RESULTS

ACCURACY ON FULL ACTION



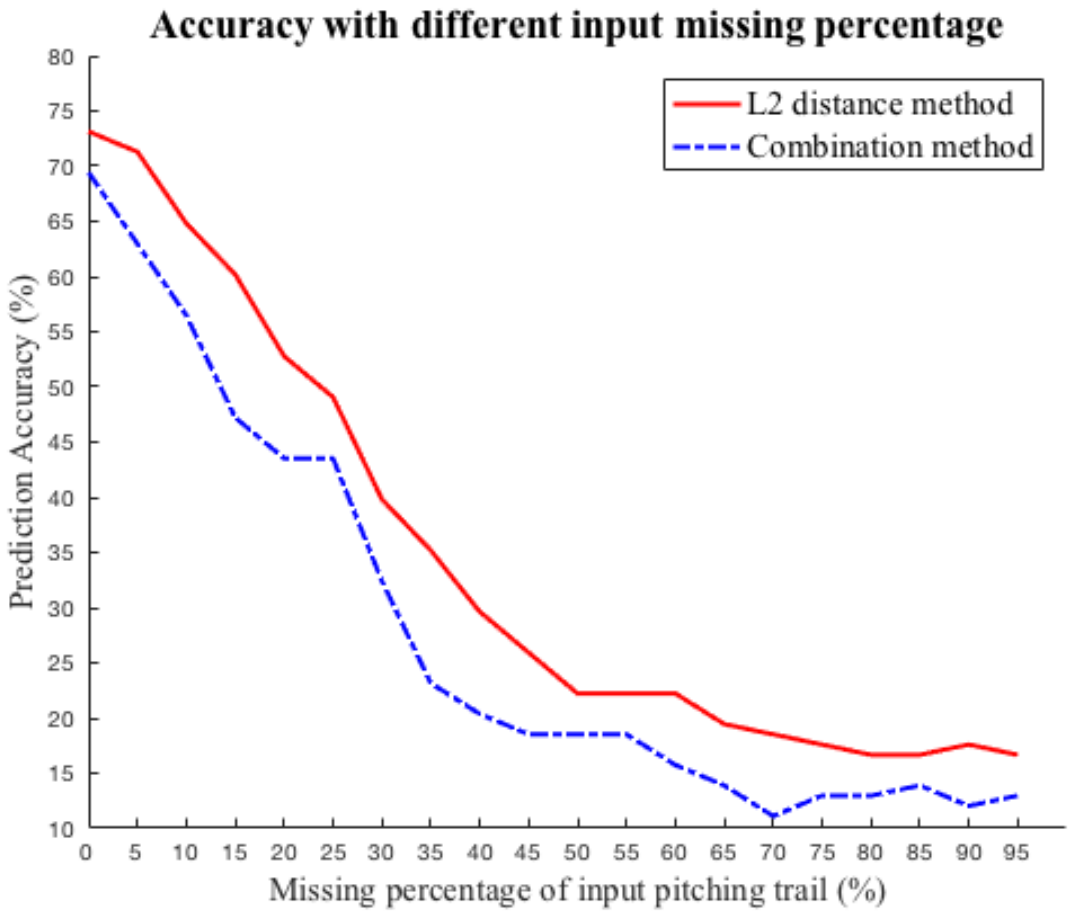


- **Accuracy: 73.15% (almost the same as two-stream methods in previous work)**
- **Mis-prediction mainly focuses on the neighborhood block.**
- **Vertical neighborhood errors > horizontal neighborhood errors**
  (Red circle) (Green circle)

清華大學 Tsinghua University

RESULTS

ACCURACY ON INCOMPLETE ACTION



- **As the missing percentage increases, the accuracy decreases.**
- **Over 50% accuracy at about 20-25% of missing input.**
- **The network loses the ability of intention prediction when the input is reduced to about 5 frames.**

CONCLUSION AND FUTURE WORK

- **Enlarge dataset: 1) Actions that outcomes differ from intentions. It can help us analyze the connection between intentions and actions. 2) Actions that from multi-view sensors. It may decrease the neighborhood errors.**

- **Explore more network architectures and implementation with Convolutional LSTM that may reach higher accuracy.**

Ma Tianqi

Ph.D. candidate

Tsinghua University

mtq19@mails.tsinghua.edu.cn

Zhang Lin

Sr. Research Associate

University of Cincinnati

zhang3l7@ucmail.uc.edu

Diao Xiumin

Assistant Professor

Purdue University

diaox@purdue.edu

Ma Ou

Chair Professor

University of Cincinnati

maou@ucmail.uc.edu