Charles Sturt University

++++

Three Dimensional Motion Trail Model for Gesture Recognition

Bin Liang, Lihong Zheng
Wagga Wagga NSW, Australia

22 December, 2013

OUTLINE

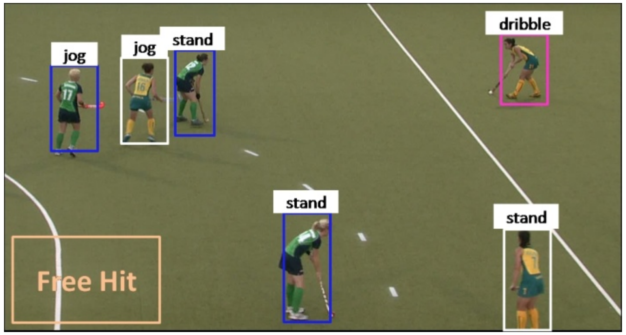
Charles Sturt University

Charles Sturt University

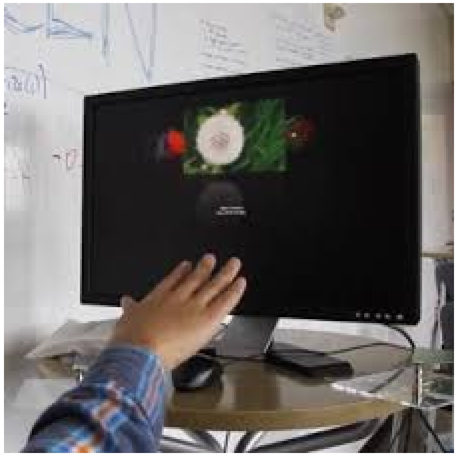
Introduction

Automatic Environmental Surveillance

Assisted Living



Sport Video Analysis



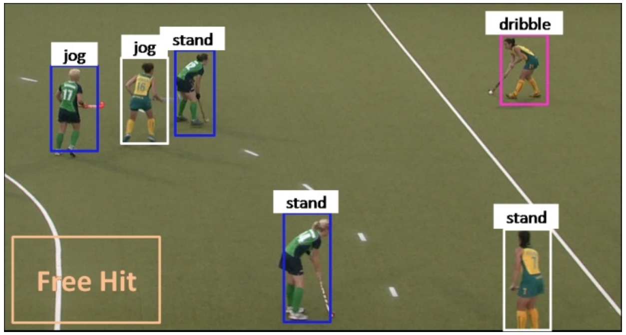
Human Computer Interaction

Introduction

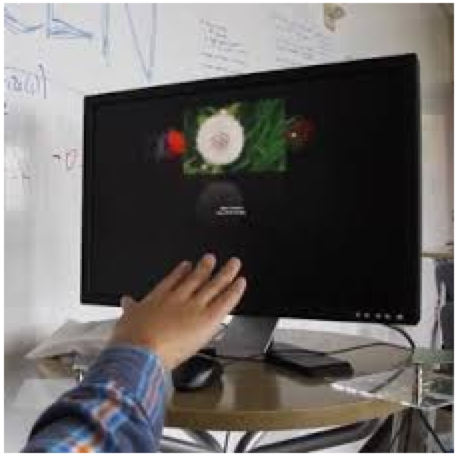
Automatic Environmental Surveillance

Assisted Living

Sensors ?



Sport Video Analysis



Human Computer Interaction

Introduction

Charles Sturt University

Wearable Sensors

- Extensive calibration
- Restricted natural movement

Vision Sensors

- Flexibility
- Markless
- Inexpensive
- Non-obtrusive

Introduction

Kinect

RGB images

Depth images

- Each pixel presents the distance between captured object and the camera.
- Motion ambiguities could be by passed.

Proposed Method

Charles Sturt University

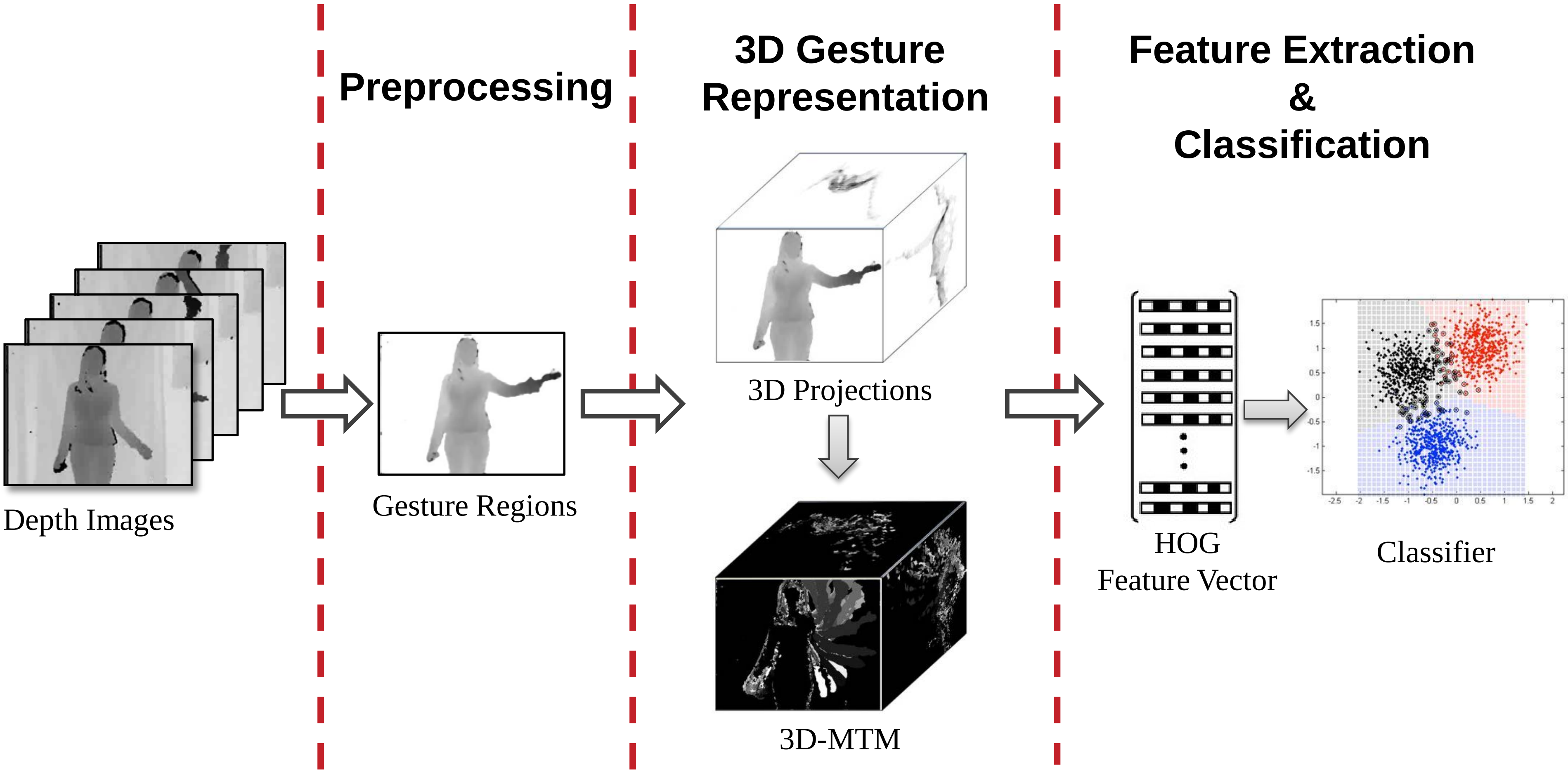
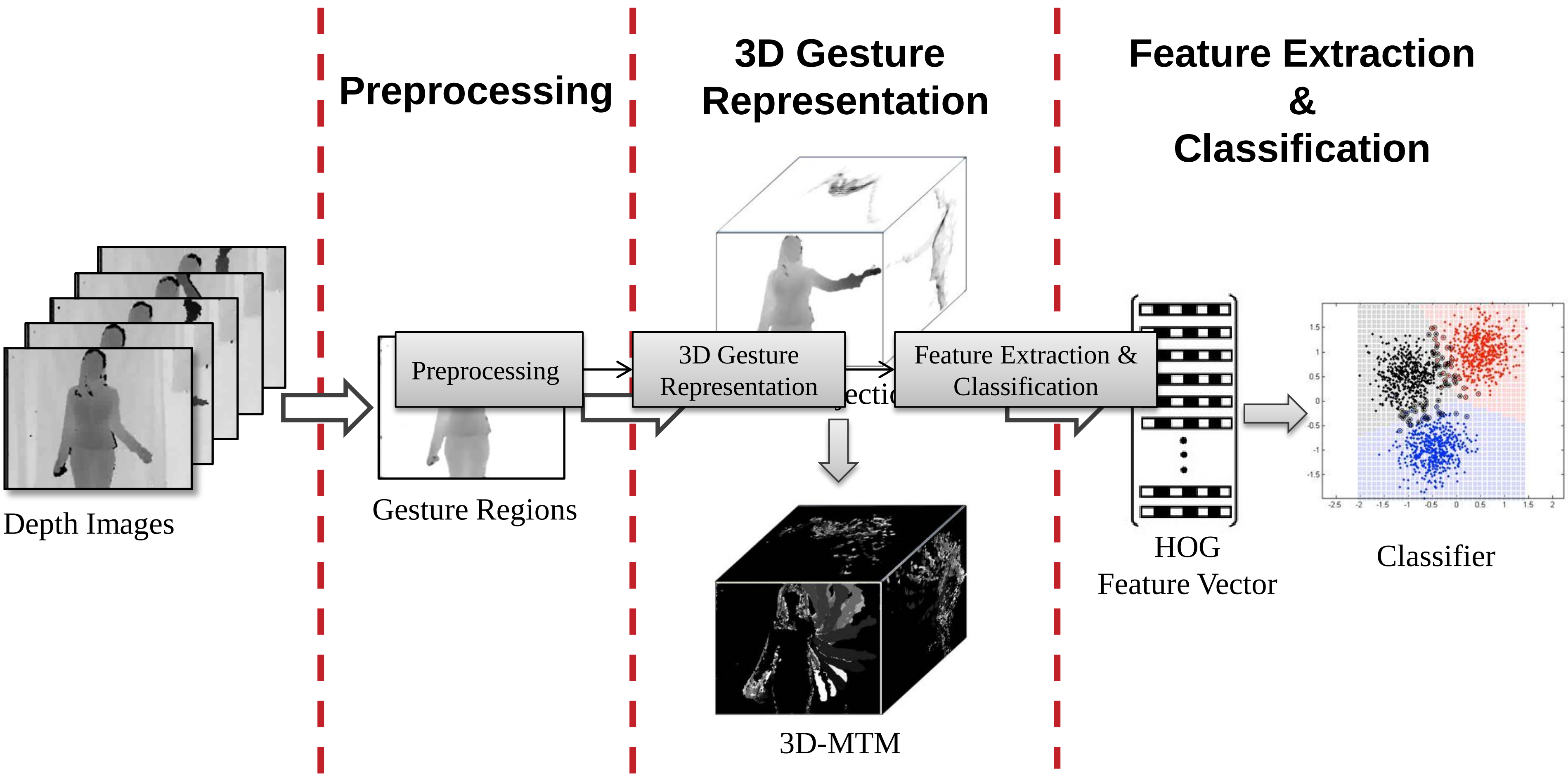


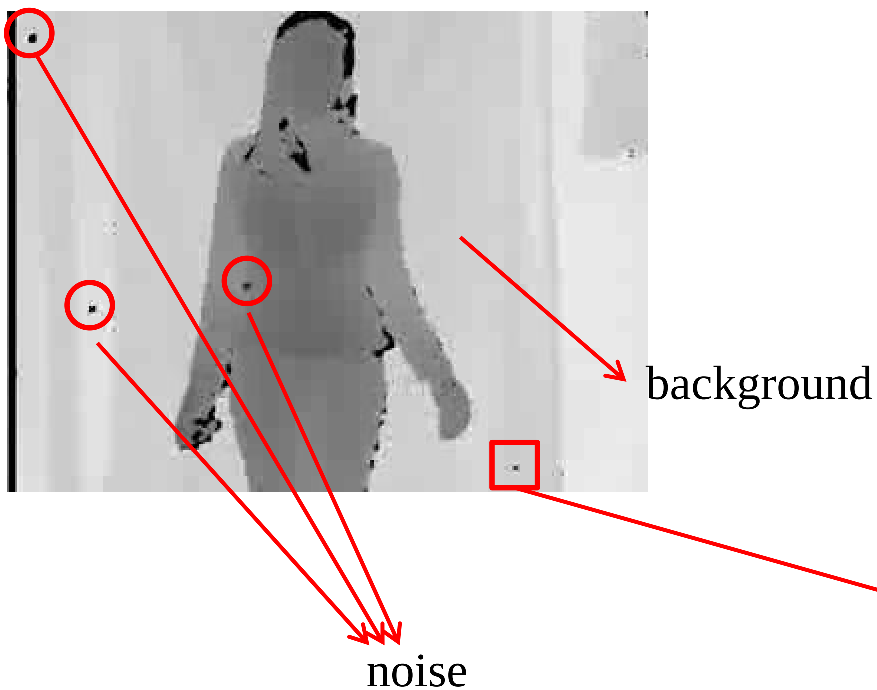
Fig. The general framework of the proposed method

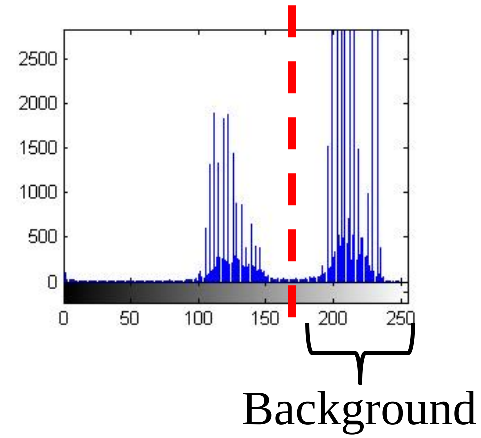
Proposed Method



Proposed Method

Preprocessing → 3D Gesture Representation → Feature Extraction & Classification





- Otsu's method[1]

101	69	0
56	255	87
123	96	157

0	56	69	87	96	101	123	157	255

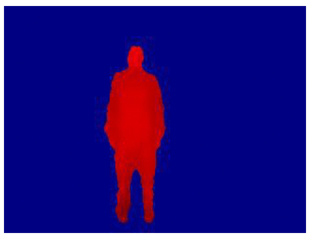
- Median Filter

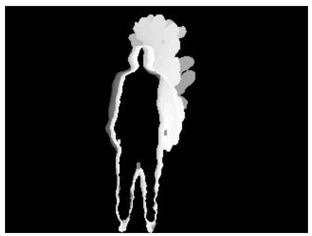
1. Otsu, Nobuyuki. "A threshold selection method from gray-level histograms."1975.

Proposed Method

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

- **Motion History Image (MHI)[1]**





$$\mathrm{H}(x,y,t;\tau)=\begin{cases}\tau & \text{if } \boxed{\Psi(x,y,t)=1,}\\ \max(0,\mathrm{H}(x,y,t-1;\tau)-\delta) & \text{otherwise.}\end{cases}$$

1. Bobick, Aaron F., and James W. Davis. “The recognition of human movement using temporal templates”, 2001.

Proposed Method

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

- **Motion History Image (MHI)[1]**

Limitations:

- fails to cover obstructed movements.
- information about static postures and repetitive movements is ignored.
- only encodes information along the *xoy*-plane

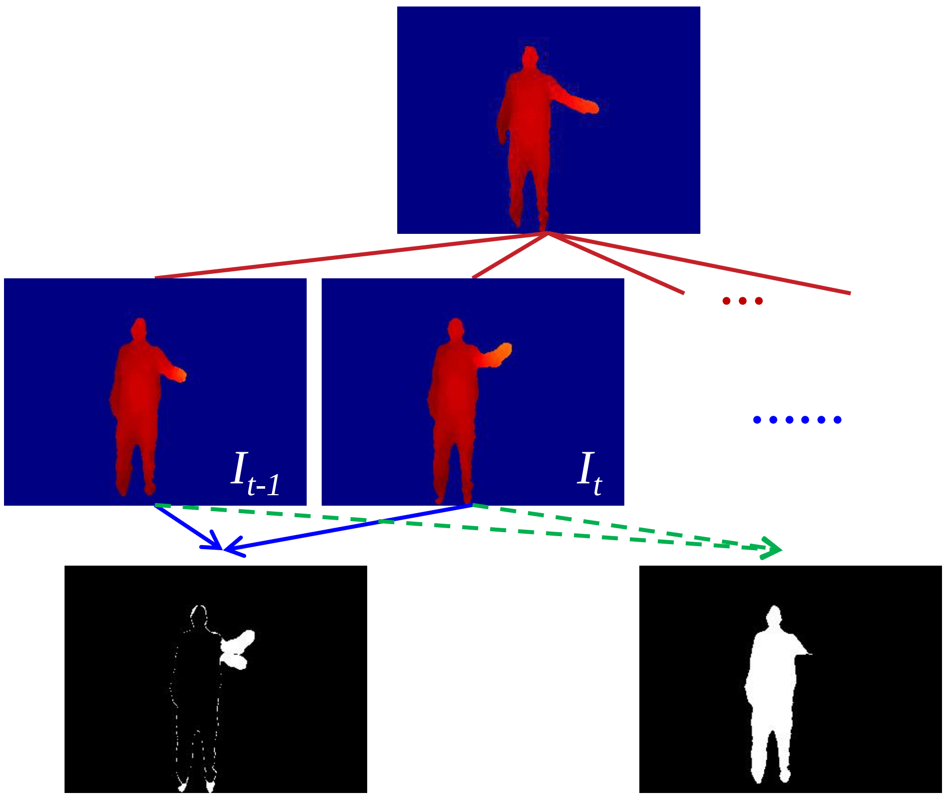
1. Bobick, Aaron F., and James W. Davis. “The recognition of human movement using temporal templates”, 2001.

Proposed Method

Charles Sturt University

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

- **2D Motion Trail Model (2D-MTM)**



Motion information

Static information

$$D_t = \begin{cases} I_1 & \text{if } t = 1 \\ \left| I_t - I_{t-1} \right| & \text{otherwise} \end{cases}$$

Motion update function:

$$\Psi_M(x, y, t) = \begin{cases} 1 & \text{if } D_t > \varsigma_M, \\ 0 & \text{otherwise.} \end{cases}$$

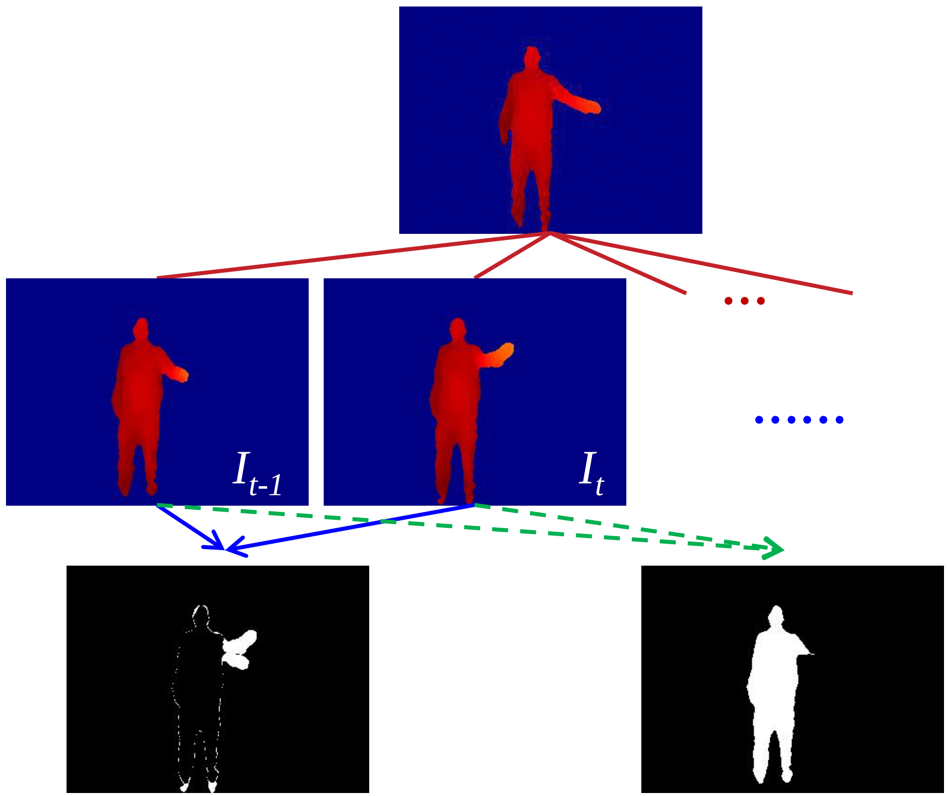
Static posture update function:

$$\Psi_S(x, y, t) = \begin{cases} 1 & \text{if } I_t - D_t > \varsigma_S, \\ 0 & \text{otherwise.} \end{cases}$$

Proposed Method

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

- **2D Motion Trail Model (2D-MTM)**



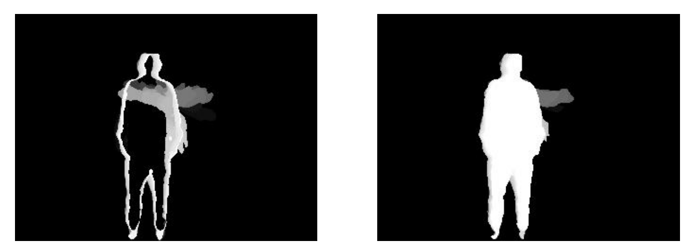
Motion information   Static information

Depth motion history image (D-MHI):

$$\mathrm{H_M}(x,y,t)=\begin{cases}T & \text{if } \Psi_\mathrm{M}(x,y,t)=1\\ \mathrm{H_M}(x,y,t-1)-1 & \text{otherwise}\end{cases}$$

Static posture history image (SHI):

$$\mathrm{H_S}(x,y,t)=\begin{cases}T & \text{if } \Psi_\mathrm{S}(x,y,t)=1\\ \mathrm{H_S}(x,y,t-1)-1 & \text{otherwise}\end{cases}$$

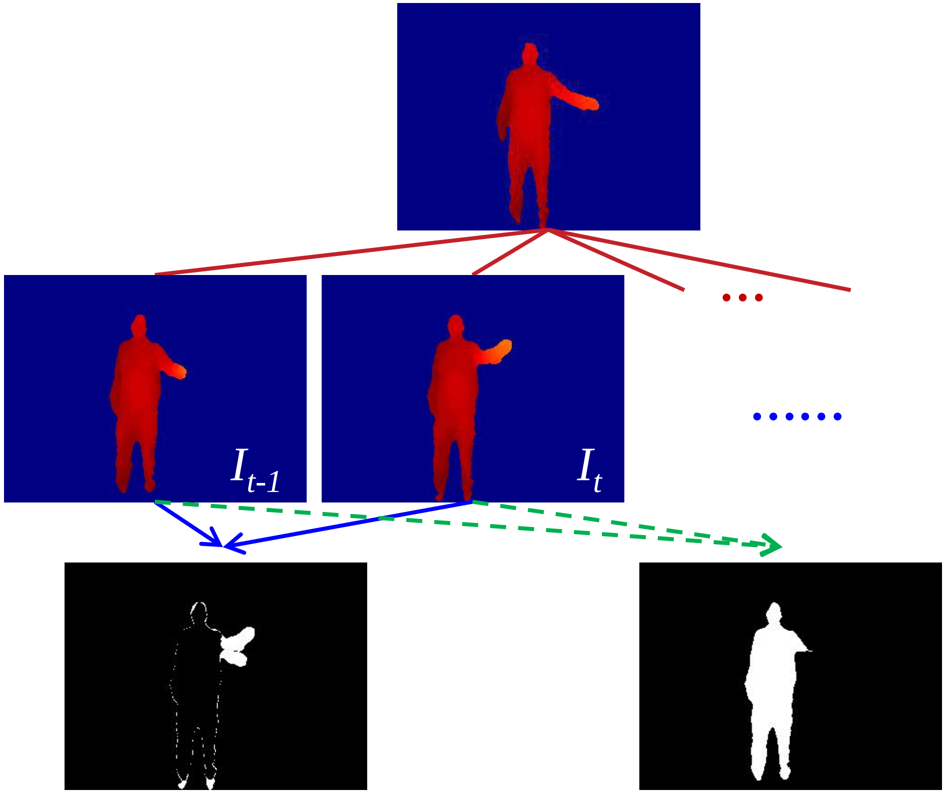


D-MHI   SHI

Proposed Method

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

- **2D Motion Trail Model (2D-MTM)**



Motion information

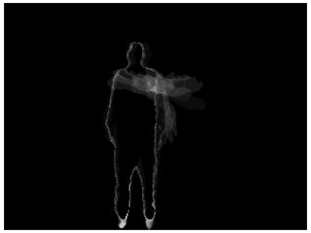
Static information

Average motion image (AMI):

$$A_M = \frac{1}{T}\sum_{t=1}^{T}\Psi_M(x,y,t)$$

Average static posture image (ASI):

$$A_S = \frac{1}{T}\sum_{t=1}^{T}\Psi_S(x,y,t)$$



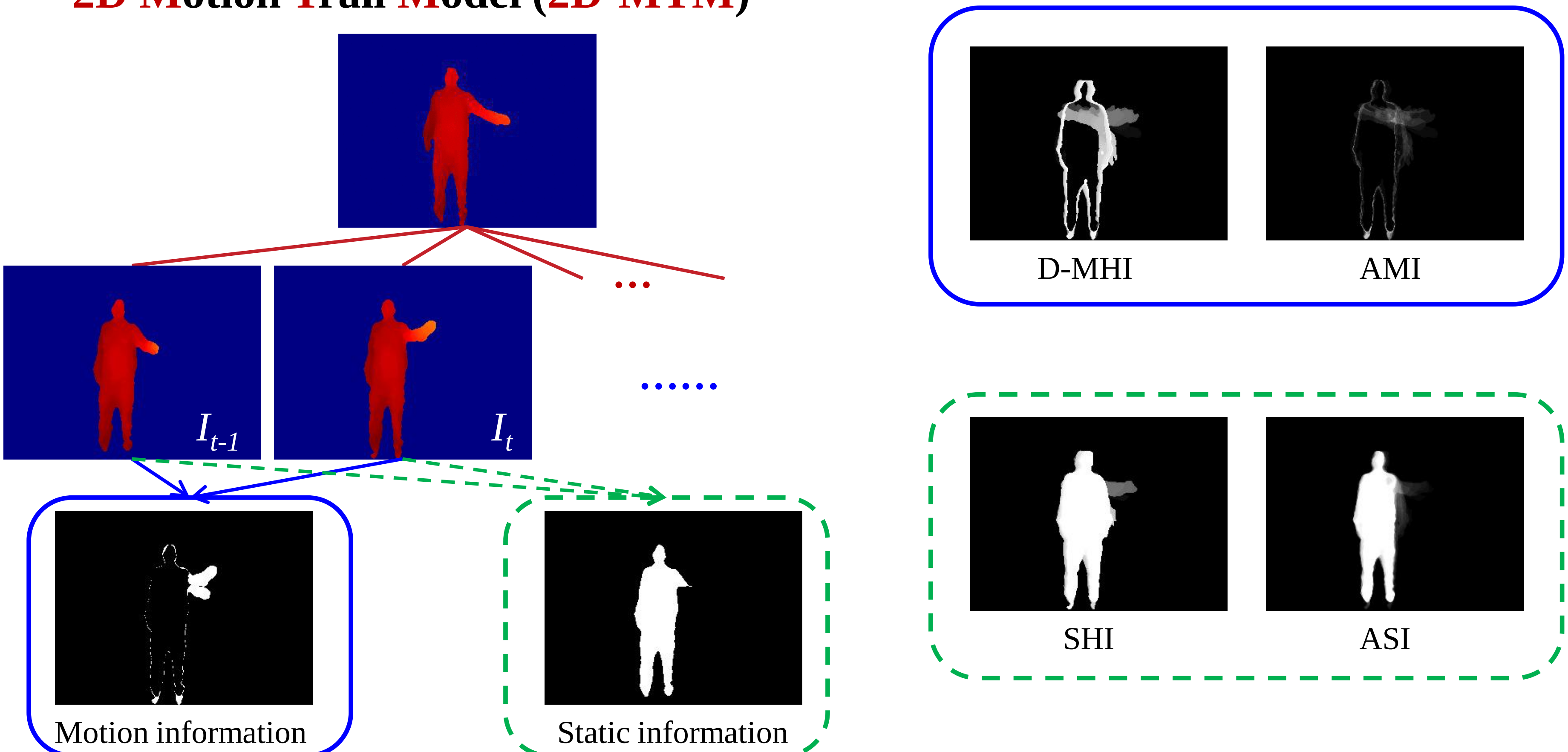
AMI

ASI

Proposed Method

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

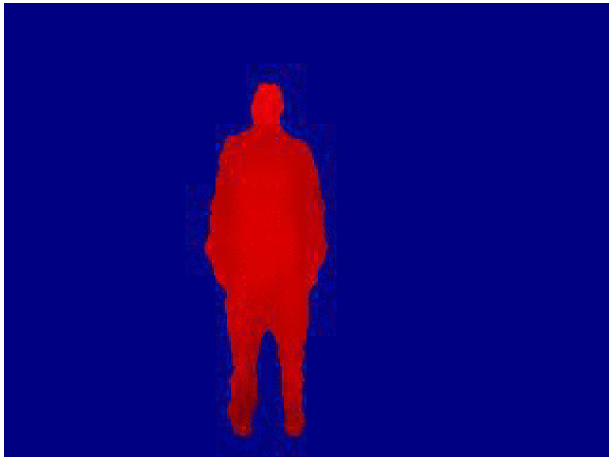
- **2D Motion Trail Model (2D-MTM)**



Proposed Method

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

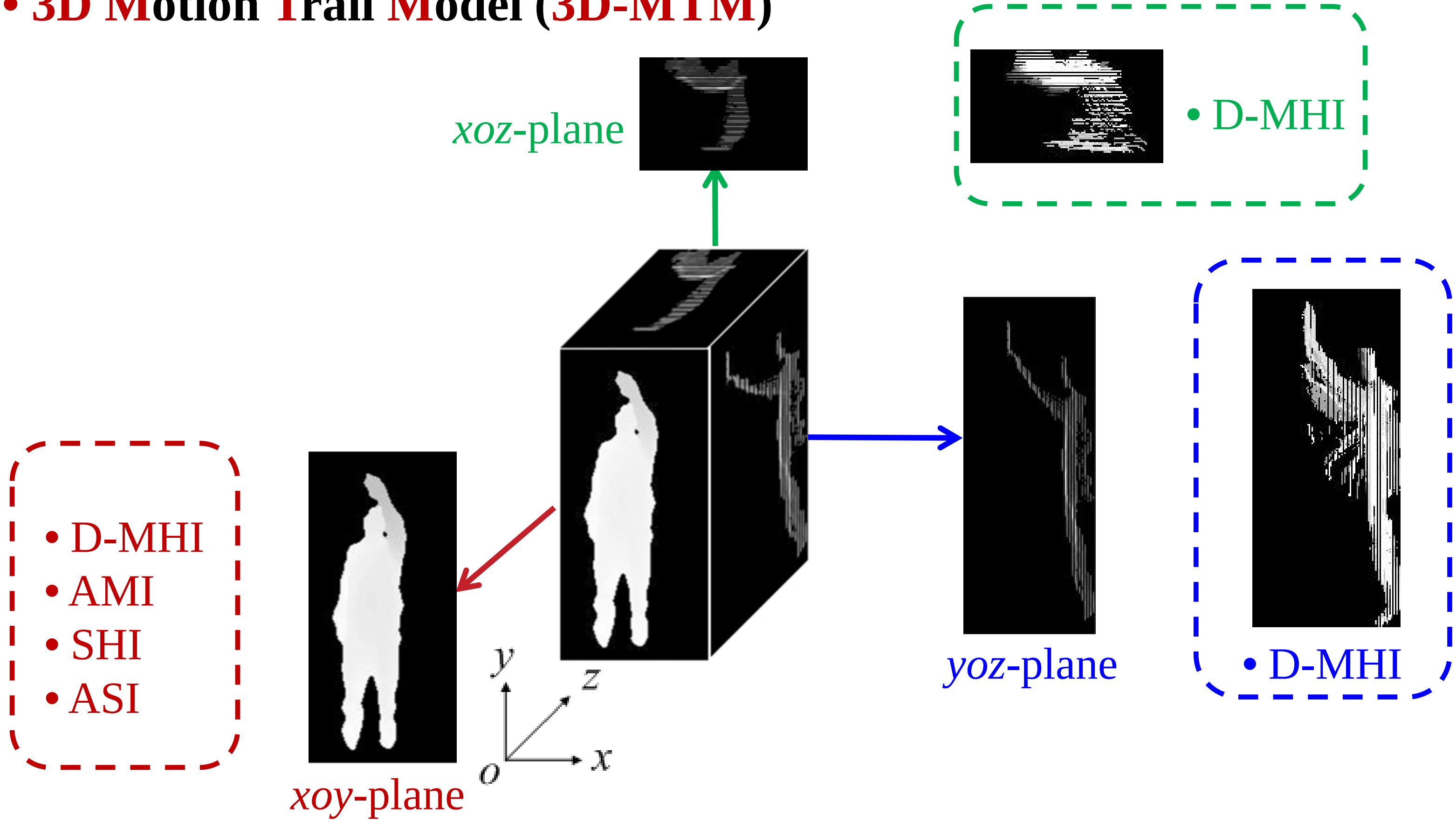
- **3D Motion Trail Model (3D-MTM)**



Proposed Method

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

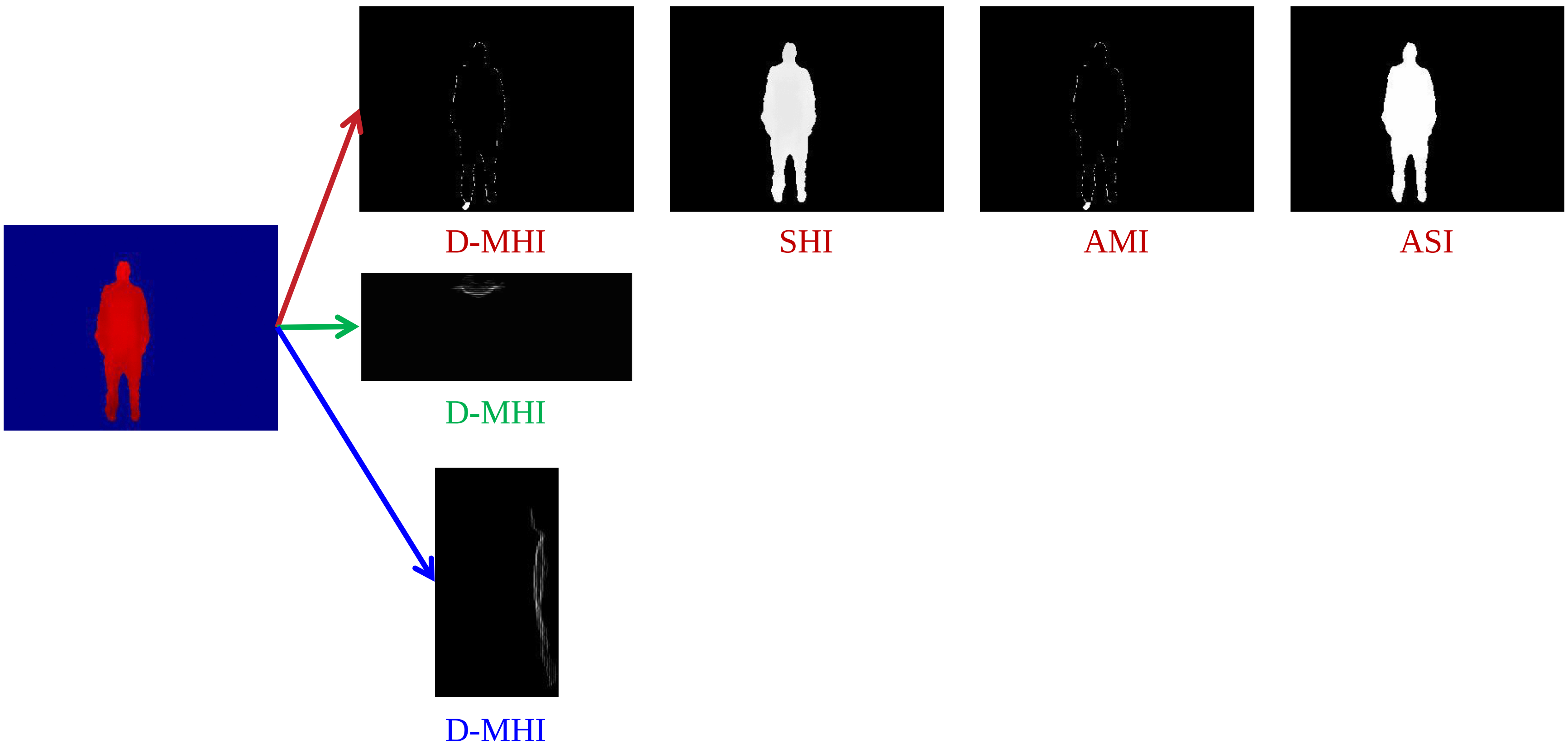
- **3D Motion Trail Model (3D-MTM)**



Proposed Method

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

- **3D Motion Trail Model (3D-MTM)**



Proposed Method

Charles Sturt University

Preprocessing → **3D Gesture Representation** → Feature Extraction & Classification

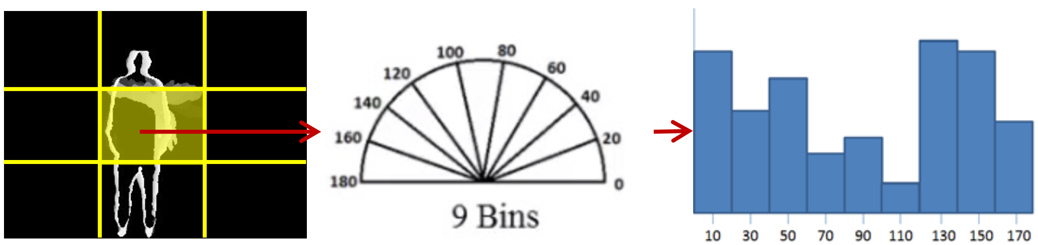
- 3D-MTM
  - is based on depth images
  - contains the information of motion history, static posture history, average motion and average static posture
  - represents movements from *xoy*-plane, *xoz*-plane and *yoz*-plane

- MHI
  - uses silhouette images
  - only records the most recent movements
  - only keeps motion information on a single plane

Proposed Method

Preprocessing → 3D Gesture Representation → **Feature Extraction & Classification**

- **HOG**



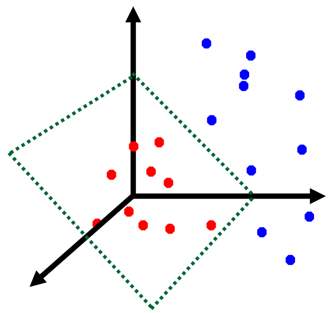
Dimension of the feature vector:

$N \times N \times B = 3 \times 3 \times 9$

Dimension of 3D-MTM representation:

$N \times N \times B \times \mathbf{6} = 3 \times 3 \times 9 \times 6 = \mathbf{486}$

- **Classification**



- SVM
- Maximum Correlation Coefficient

Experimental Results

- **Datasets**

Sample frames from the MSR Action3D dataset

- **MSR Action dataset**
  - 20 action types
  - Each action is performed by 10 subjects for 2 or 3 times
  - Resolution: 320 × 240

Sample frames from the ChaLearn gesture dataset

- **ChaLearn gesture dataset**
  - Used for one-shot learning challenge
  - Performed on the first 10 data batches
  - Each test video contains 1 to 5 gestures

Experimental Results

Charles Sturt University

MSR Action3D dataset

	Bag-of-3D-Points[1]	Proposed Method
T1-AS1	89.5%	**96.0%**
T1-AS2	89.0%	**94.9%**
T1-AS3	96.3%	**97.3%**
T2-AS1	93.4%	**100.0%**
T2-AS2	92.9%	**97.5%**
T2-AS3	96.3%	**100.0%**
CST-AS1	72.9%	**73.7%**
CST-AS2	71.9%	**81.5%**
CST-AS3	79.2%	**81.6%**

Tab. Recognition accuracy comparison of different subsets that are Test One (T1), Test Two (T2), and Cross Subject Test (CST)

1. Li, Wanqing, Zhengyou Zhang, and Zicheng Liu. "Action recognition based on a bag of 3d points". 2010

Experimental Results

Charles Sturt University

One-shot learning ChaLearn gesture dataset

Method	Average error rate
Baseline	62.8%
Dynamic Time Warping	43.1%
Principle Motion	37.4%
MHI	37.6%
2D-MTM (ours)	**24.4%**
3D-MTM (ours)	**21.7%**

Tab. Performance comparison

Experimental Results

One-shot learning ChaLearn gesture dataset

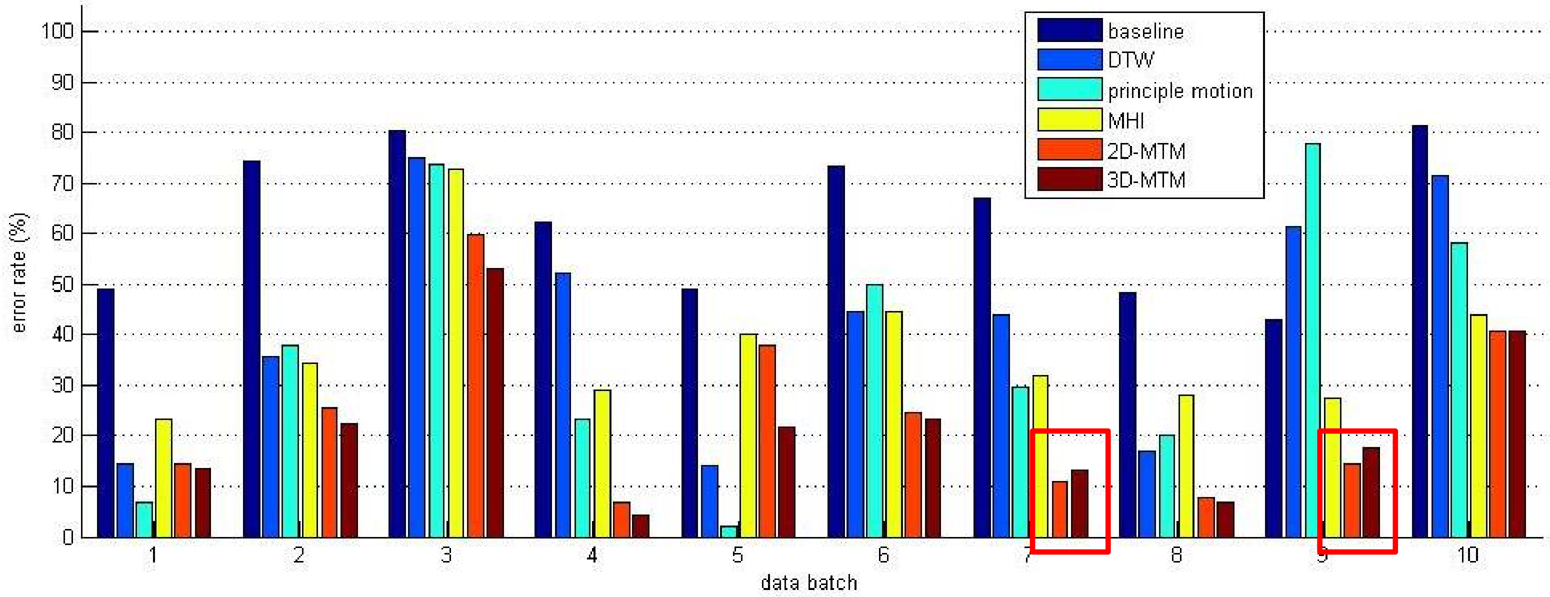
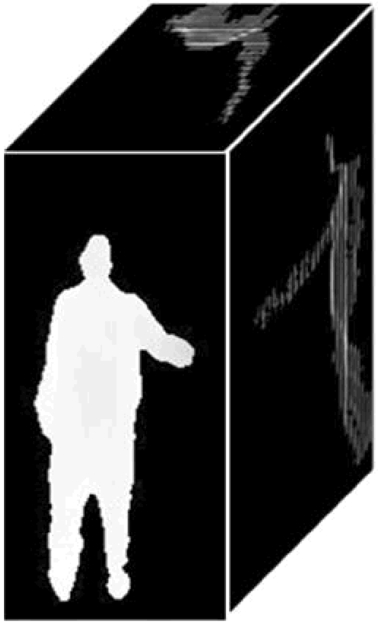


Fig. Results of the 10 data batches

Conclusions

Charles Sturt University

- A 2D motion trail model (**2D-MTM**) is proposed to represent the motion information and static posture information of a gesture sequence.
- A novel 3D model (**3D-MTM**) is extensively proposed by projecting depth images onto other two planes, and it is shown to be robust to model gestures in 3D space.
- The proposed method based on 3D-MTM achieves competitive performance on MSR Action3D dataset and ChaLearn dataset..
- The future work will focus on gesture recognition on variant subjects to improve recognition performance in the Cross Subject Test.



Thank you