

One Shot Learning Gesture Recognition with Kinect Sensor

Di Wu, Fan Zhu, Ling Shao

June 12, 2012



Dataset Description and Analysis

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

Our approach

Motion Descriptors

MSE

Others and Ending

Dataset Description and Analysis

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

- ▶ Availability of depth camera

Our approach

Motion Descriptors

MSE

Others and Ending

1. Preprocessing: Background Separation and Noise Reduction for Depth Images

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

Our approach

Motion Descriptors

MSE

Others and Ending

1. Preprocessing: Background Separation and Noise Reduction for Depth Images

- ▶ Background Separation: Otsu. A threshold selection method from gray-level histograms.

1. Preprocessing: Background Separation and Noise Reduction for Depth Images

- ▶ Background Separation: Otsu. A threshold selection method from gray-level histograms.
- ▶ Noise Reduction: 5×5 aperture median filter, morphological process: opening operation

2. Temporal Segmentation

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

Our approach

Motion Descriptors

MSE

Others and Ending

2. Temporal Segmentation

Approach: candidate cut—simple and effective:

Our approach

Motion Descriptors

MSE

Others and Ending

2. Temporal Segmentation

Approach: candidate cut—simple and effective:

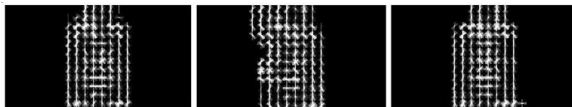


Figure: HOG descriptor for temporal segmentation

Implementation details:

2. Temporal Segmentation

Approach: candidate cut—simple and effective:

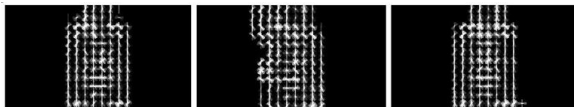


Figure: HOG descriptor for temporal segmentation

Implementation details:

how many similar frames should we search? $8 \times Q$

Our approach

Motion Descriptors

MSE

Others and Ending

2. Temporal Segmentation

Approach: candidate cut—simple and effective:

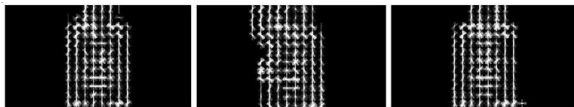


Figure: HOG descriptor for temporal segmentation

Implementation details:

how many similar frames should we search? $8 \times Q$

dimension for HOG $N \times N \times B$: $8 \times 8 \times 9$, $\mathcal{L}\mathcal{D}$ is 6.764% and
 $16 \times 16 \times 9$ is 5.235%.

2. Temporal Segmentation

Approach: candidate cut—simple and effective:

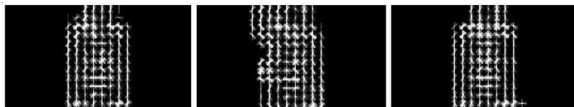


Figure: HOG descriptor for temporal segmentation

Implementation details:

how many similar frames should we search? $8 \times Q$

dimension for HOG $N \times N \times B$: $8 \times 8 \times 9$, $\mathcal{L}\mathcal{D}$ is 6.764% and
 $16 \times 16 \times 9$ is 5.235%.

Motion Descriptors

Cons for local method

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

Our approach

Motion Descriptors

MSE

Others and Ending

Motion Descriptors

Cons for local method

Our approach: Extended-MHI

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

Our approach

Motion Descriptors

MSE

Others and Ending

Motion Descriptors

Cons for local method

Our approach: Extended-MHI

Assume $I_t = (I_1, I_2, \dots, I_{nFrames}) \in \mathbb{R}^3$ is a gray scale image sequence and let $B_t = (B_1, B_2, \dots, B_{nFrames-1}) \in \mathbb{R}^3$ be a binary image sequence indicating regions of motion, which can be obtained from image differencing and thresholding:

Motion Descriptors

Cons for local method

Our approach: Extended-MHI

Assume $I_t = (I_1, I_2, \dots, I_{nFrames}) \in \mathbb{R}^3$ is a gray scale image sequence and let $B_t = (B_1, B_2, \dots, B_{nFrames-1}) \in \mathbb{R}^3$ be a binary image sequence indicating regions of motion, which can be obtained from image differencing and thresholding:

$$B_t = \begin{cases} 1 & \text{if } (I_{t+1} - I_t) > \textit{Threshold}, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Motion Descriptors

Cons for local method

Our approach: Extended-MHI

Assume $I_t = (I_1, I_2, \dots, I_{nFrames}) \in \mathbb{R}^3$ is a gray scale image sequence and let $B_t = (B_1, B_2, \dots, B_{nFrames-1}) \in \mathbb{R}^3$ be a binary image sequence indicating regions of motion, which can be obtained from image differencing and thresholding:

$$B_t = \begin{cases} 1 & \text{if } (I_{t+1} - I_t) > \text{Threshold}, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

where threshold is defined as:

$$\text{Threshold} = \sqrt{\sum_t^{nFrames} \sigma_t / (h \times w \times nFrames)} \quad (2)$$

where σ_t is the second moment (variance) of a single frame I_t ; $h, w, nFrames$ are the height, width and frame number of that video sequence.

Extended-MHI

- ▶ motion history image (MHI)

- ▶ motion history image (MHI)

$$\tilde{H}(t; \tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ \tilde{H}(t-1; \tau) - 1 & \text{otherwise.} \end{cases} \quad (3)$$

- ▶ motion history image (MHI)

$$\tilde{H}(t; \tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ \tilde{H}(t-1; \tau) - 1 & \text{otherwise.} \end{cases} \quad (3)$$

- ▶ gait energy information (GEI)

- ▶ motion history image (MHI)

$$\tilde{H}(t; \tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ \tilde{H}(t-1; \tau) - 1 & \text{otherwise.} \end{cases} \quad (3)$$

- ▶ gait energy information (GEI)

$$G = \frac{1}{\tau} \sum_{t=1}^{\tau} I_t \quad (4)$$

- ▶ motion history image (MHI)

$$\tilde{H}(t; \tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ \tilde{H}(t-1; \tau) - 1 & \text{otherwise.} \end{cases} \quad (3)$$

- ▶ gait energy information (GEI)

$$G = \frac{1}{\tau} \sum_{t=1}^{\tau} I_t \quad (4)$$

- ▶ Inversed recording (INV)

- ▶ motion history image (MHI)

$$\tilde{H}(t; \tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ \tilde{H}(t-1; \tau) - 1 & \text{otherwise.} \end{cases} \quad (3)$$

- ▶ gait energy information (GEI)

$$G = \frac{1}{\tau} \sum_{t=1}^{\tau} I_t \quad (4)$$

- ▶ Inversed recording (INV)

$$\tilde{I}(t; \tau) = \begin{cases} \tau & \text{if } B_t = 1, \\ \tilde{I}(t+1; \tau) - 1 & \text{otherwise.} \end{cases} \quad (5)$$

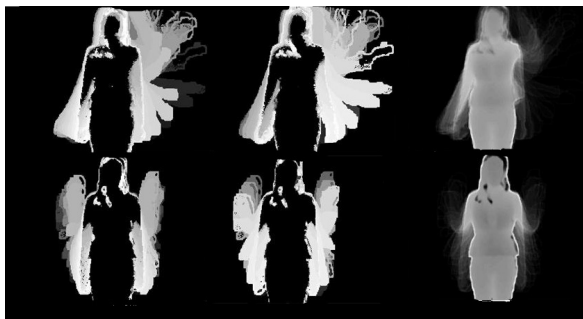


Figure: Illustration of the *MHI*, *INV* and *GEI* in two tokens (top row and bottom row). The projection images show that *MHI* emphasizes recent motion, ending frames whilst *INV* the beginning frames. *GEI* encodes the average gait information and is supplementary in repetitive actions where both *MHI* and *INV* are poor at representing.

Methods	<i>GEI</i>	<i>MHI</i>	<i>INV</i>	<i>Extended-MHI</i>
\mathcal{L}_D	0.2761	0.3010	0.3022	0.2600

Table: Performance comparison of three elements in *Extended-MHI*

Multiview Spectral Embedding

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

Our approach

Motion Descriptors

MSE

Others and Ending

1. learn the complementary nature of different views,

Multiview Spectral Embedding

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

Our approach

Motion Descriptors

MSE

Others and Ending

1. learn the complementary nature of different views,
2. search for a low dimensional representation and sufficiently smooth embedding over all views.

Multiview Spectral Embedding

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

Our approach

Motion Descriptors

MSE

Others and Ending

1. learn the complementary nature of different views,
2. search for a low dimensional representation and sufficiently smooth embedding over all views.
3. 4% improvement in $\mathcal{L}\mathcal{D}$

low-dimensional embedding

$$\underbrace{\operatorname{argmin}}_{Y, \alpha} \sum_{i=1}^m \alpha_i^r \operatorname{tr}(Y L^i Y^T) \quad (6)$$

$$\text{s.t. } \sum_i^m i = 1 \quad (7)$$

END

One Shot Learning
Gesture Recognition with
Kinect Sensor

Di Wu, Fan Zhu, Ling
Shao

Our approach

Motion Descriptors

MSE

Others and Ending