One Shot Learning Gesture Recognition with Kinect Sensor

Di Wu, Fan Zhu, Ling Shao

June 12, 2012

The University Of Sheffield.

Contents

Dataset Description and Analysis

Dataset Description and Analysis

- Availability of depth camera

Dataset Description and Analysis

- Availability of depth camera

Figure: Noise in depth image

Dataset Description and Analysis

- Availability of depth camera

Figure: Noise in depth image

- Multiple gestures in testing set

Dataset Description and Analysis

- Availability of depth camera

Figure: Noise in depth image

- Multiple gestures in testing set
- One-shot-learning

Dataset Description and Analysis

- Availability of depth camera

Figure: Noise in depth image

- Multiple gestures in testing set
- One-shot-learning
- Depth RGB camera decision fusion

Dataset Description and Analysis

- Availability of depth camera

Figure: Noise in depth image

- Multiple gestures in testing set
- One-shot-learning
- Depth RGB camera decision fusion

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

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

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

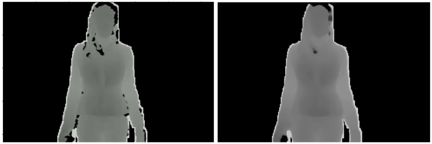


Figure: Background Segmentation and Denoising

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

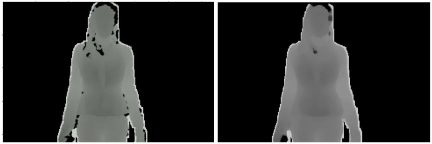


Figure: Background Segmentation and Denoising

- Improve the performance in terms of $\mathfrak{LD}$ as much as 9%

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

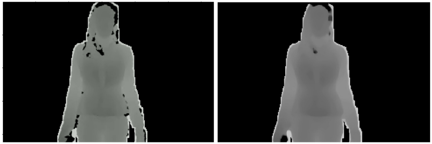


Figure: Background Segmentation and Denoising

- Improve the performance in terms of $\mathfrak{LD}$ as much as 9%

2.Temporal Segmentation

2.Temporal Segmentation

Approach: candidate cut–simple and effective:

2.Temporal Segmentation

Approach: candidate cut–simple and effective:

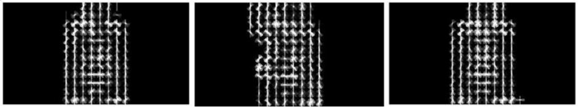


Figure: HOG descriptor for temporal segmentation

Implementation details:

2.Temporal Segmentation

Approach: candidate cut–simple and effective:

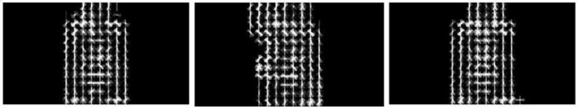


Figure: HOG descriptor for temporal segmentation

Implementation details:

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

2.Temporal Segmentation

Approach: candidate cut–simple and effective:

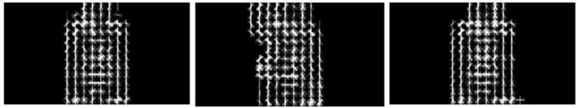


Figure: HOG descriptor for temporal segmentation

Implementation details:

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

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

2.Temporal Segmentation

Di Wu, Fan Zhu, Ling Shao

Approach: candidate cut–simple and effective:

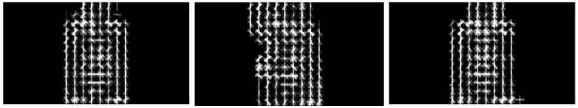


Figure: HOG descriptor for temporal segmentation

Implementation details:

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

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

Motion Descriptors

Cons for local method

Motion Descriptors

Cons for local method
Our approach: Extended-MHI

Motion Descriptors

Cons for local method

Our approach: Extended-MHI

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

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

Motion Descriptors

Cons for local method

Our approach: Extended-MHI

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

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

where threshold is defined as:

$$Threshold = \sqrt{\sum_{t}^{nFrames} \sigma_t / (h \times w \times nFrames)} \tag{2}$$

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)

Extended-MHI

- motion history image (MHI)

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

Extended-MHI

- motion history image (MHI)

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

- gait energy information (GEI)

Extended-MHI

- motion history image (MHI)

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

- gait energy information (GEI)

$$G = \frac{1}{\tau} \sum_{t=1}^{\tau} I_t \tag{4}$$

Extended-MHI

- motion history image (MHI)

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

- gait energy information (GEI)

$$G = \frac{1}{\tau} \sum_{t=1}^{\tau} I_t \tag{4}$$

- Inversed recording (INV)

Extended-MHI

- motion history image (MHI)

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

- gait energy information (GEI)

$$G = \frac{1}{\tau} \sum_{t=1}^{\tau} I_t \tag{4}$$

- Inversed recording (INV)

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

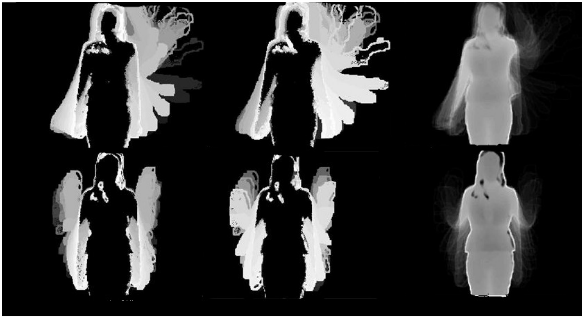


Figure: Illustration of the *MHI*, *INV* and *GEI* in two tokens (top row and bottom row). The projection images show that *MHI* emphasizes recent motion, ending frames whilst *INV* the beginning frames. *GEI* encodes the average gait information and is supplementary in repetitive actions where both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

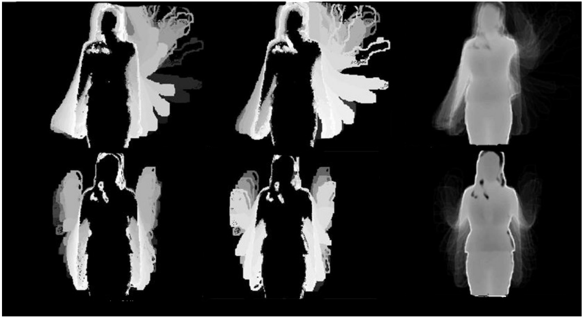


Figure: Illustration of the *MHI*, *INV* and *GEI* in two tokens (top row and bottom row). The projection images show that *MHI* emphasizes recent motion, ending frames whilst *INV* the beginning frames. *GEI* encodes the average gait information and is supplementary in repetitive actions where both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

Methods	*GEI*	*MHI*	*INV*	*Extended-MHI*
$\mathfrak{LD}$	0.2761	0.3010	0.3022	**0.2600**

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

Multiview Spectral Embedding

1. learn the complementary nature of different views,

Multiview Spectral Embedding

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

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 $\mathfrak{LD}$

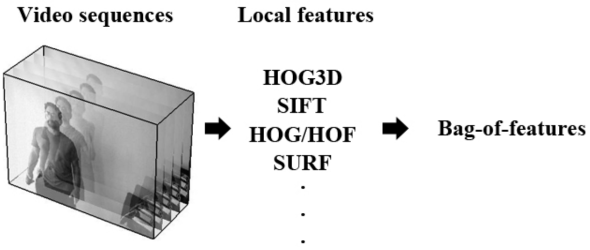
low-dimensional embedding

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

$$\text{s.t.} \sum_{i}^{m} i = 1 \tag{7}$$

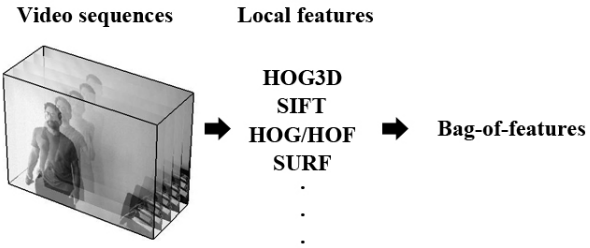
Local approach

- Local features+Bag-of-Words



Local approach

- Local features+Bag-of-Words



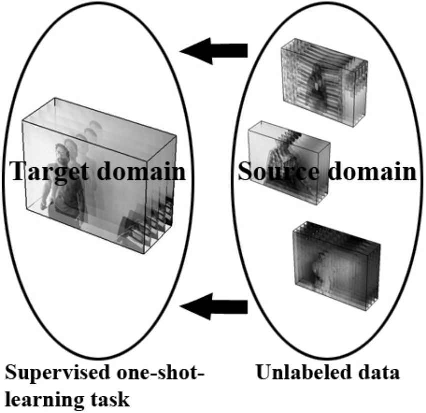
- Fail!!!

Local approach

- Why the local approach not working?
  - The local patterns used for training the codebook are not sufficient.
  - The Bag-of-Words histogram representation is not discriminative.

Transfer learning approach

- To incorporate more information into the supervised one-shot-learning task



Transfer learning approach

- Different from those action categories running or kicking, the given gesture categories are uniquely designed so that it is difficult to find related data that contains similar gestures as the given ones.
- One possible approach is to utilise the unlabelled data from the developmentvalidation batches to generate a more informative codebook or generate a higher level representation of the input.

END