Action and Event Recognition
Using Depth Cameras

Bingbing Ni  Advanced Digital Sciences Center, Singapore
Pierre Moulin  University of Illinois at Urbana-Champaign
Visual Analytics Using Depth Camera

ADSC Illinois Faculty
Pierre Moulin

Research Scientists
Bingbing Ni
Jiwen Lu
Gang Wang

Research Engineers
Yong Pei
Venice Erin Liong
Vignesh Ramkrishnan

What are they doing?
Outline

- Overview of Kinect applications
- Kinect-based multi-modal visual analytics research at ADSC
  - Kinect-based tele-rehabilitation
  - Kinect-based action recognition
  - Kinect-based event detection
  - Research Highlight: the HARL contest
  - Research Highlight: fine-grained action detection
- Conclusions
The Kinect Camera

• Driving application: Games!

Microsoft’s Xbox game. Shotton et al., *Real time human pose recognition in parts from a single depth image*, CVPR 2011 Best Paper Award
Scientific and Engineering Applications of Kinect

- 3D scene structure
- Easy foreground segmentation
- 3D motion information
- Privacy
- Low cost
- Typically for indoor use
How Does the Kinect Depth Camera Work?

Projected speckle pattern

Shpunt et al, PrimeSense patent application
US 2008/0106746
How Does the Kinect Depth Camera Work? (Cont’d)
How Does the Kinect Depth Camera Work? (Cont’d)

Disparity is inversely proportional to depth.

\[ \text{disparity} = x - x' = \frac{b \times f}{z} \]
Outline

- Overview of Kinect applications
  - Kinect-based multi-modal visual analytics research at ADSC
    - Kinect-based tele-rehabilitation
    - Kinect-based action recognition
    - Kinect-based event detection
    - Research Highlight: the HARL contest
    - Research Highlight: fine-grained action detection
  - Conclusions
A Glimpse at Kinect Applications

- Low- and mid-level image processing applications
  - 3D reconstruction and modeling
  - Image enhancement
  - Video stabilization
  - Video segmentation
  - ...

- High-level vision applications
  - Foreground & human detection
  - 3D Human body/head pose identification
  - Gait analysis
  - Indoor spatial layout modeling
  - Interactive: game, control, surgery, rehabilitation, shopping etc.
  - ...

Kinect Application 1

• 3D scene reconstruction and modeling

Microsoft’s ‘KinectFusion’ creates a real-time, 3D model of an entire room

Izadi et al., *KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth*, ACM Symposium on User Interface Software and Technology, October 2011
Kinect Application 2

• Kinect-based video stabilization

Liu et al., *Video Stabilization with a Depth Camera*, CVPR 2012
Kinect Application 3

• Kinect-based video segmentation

Abramov et al., *Depth-supported real-time video segmentation with the Kinect*, International Workshop on Applications of Computer Vision, 2012
Kinect Application 4

- Kinect-based foreground detection

Salas and Tomasi, *People Detection using Color and Depth Images*, The Mexican Conference on Pattern Recognition, 2011
Kinect Application 5

• Kinect for surgery room

http://www.xbox.com/en-SG/Kinect/Kinect-Effect
http://www.zdnet.com/blog/health/xbox-kinect-helps-surgeons-in-the-operating-room/277
Kinect Application 6

• Kinect for interactions

Online shopping

http://www.youtube.com/watch?v=s0Fn6PyfJ0I&hl=en-GB&gl=SG

http://www.youtube.com/watch?v=L_cYKFdP1_0

Media content browsing
Kinect Application 7

- Kinect-based gait analysis

Kinect Application 8

- Kinect for painting & arts

http://www.kinecthacks.com/air-painting-via-kinect/
Kinect Application 9

• Kinect for 3D object scanning and model creation (known as: *KinectFusion*)

Kinect Fusion in action, taking the depth image from the Kinect camera with lots of missing data and within a few seconds producing a realistic smooth 3D reconstruction of a static scene by moving the Kinect sensor around. From this, a point cloud or a 3D mesh can be produced.

Kinect Application 10

- Kinect for 3D body scanning and virtual fitting

http://www.styku.com
Kinect Application 11

- Kinect for 3D face tracking and recognition

Kinect Application 12

- Kinect for robot control

http://spectrum.ieee.org/automaton/robotics/diy/top-10-robotic-kinect-hacks
Kinect Application 13

- Kinect for consumer behavior capture

http://shopperception.com/
Outline

- Overview of Kinect applications

- Kinect-based multi-modal visual analytics research at ADSC
  - Kinect-based tele-rehabilitation
  - Kinect-based action recognition
  - Kinect-based event detection
  - Research Highlight: the HARL contest
  - Research Highlight: fine-grained action detection

- Conclusions
Goal: Effective and robust human action/event and activity analysis using consumer depth and color video cameras.

Research challenges
Effective and robust performance given complex background, changing viewpoints, occlusion, and poor illumination conditions.
Project roadmap

**Research Challenges**

**3D Human Motion Analysis**
- Infer 3D human pose (body positioning) in real time, accurately, and robustly

**Human Action/Event Analysis**
- Detect human atomic actions (e.g., wave hand, pick up cup) and abnormal events (e.g., drop spoon, fall down) accurately and robustly

**Human Activity Analysis**
- Detect and localize high-level human activity and behavior effectively

**Validation**
- Smart Office
- Rehabilitation
- Daily Activity Monitor

Low cost consumer depth + color camera
Outline

- Introduction to Kinect applications

- Kinect-based multi-modal visual analytics research at ADSC
  - Kinect-based tele-rehabilitation
  - Kinect-based action recognition
  - Kinect-based event detection
  - Research Highlight: the HARL contest
  - Research Highlight: fine-grained action detection

- Conclusions
Kinect-based Tele-Rehabilitation

- Fugl-Meyer upper body exercise protocol
- For patients with limb injuries
- Joint angle measurement
- Movement counting
- Incorrect movement/pose alarm
- Demo video: http://www.youtube.com/watch?v=PvuA3DTsXck

Figure 1. (a) (left) Pei performs a shoulder rotation while the assistant shows him the degree of rotation. (b)(middle) The screen reminds Pei how to do pendular exercises, while preparing to count his repetitions. (c)(right) The assistant’s screen shows a menu of assigned exercises, while reporting the degree of flexion for Pei’s elbows in the current exercise.
Outline

- Introduction to Kinect applications
- Kinect-based multi-modal visual analytics research at ADSC
  - Kinect-based tele-rehabilitation
  - Kinect-based action recognition
  - Kinect-based event detection
  - Research Highlight: the HARL contest
  - Research Highlight: fine-grained action detection
- Conclusions
Kinect-based Action/Activity Recognition
Application: daily activity monitoring for the elderly

- Go to bed
- Drink water
- Answer call

Daily activity recognition and summarization
Privacy (i.e., if depth only)
RGBD-HuDaAct Database Construction

• Device
  • Single Kinect, RGB + Depth, 640×480 pixels, 30 fps
  • Software: OpenNI platform

• Data Collection
  • Lab environment
  • 30 invited subjects, 5,000,000 frames (approx. 48 hours)
  • 1189 video samples, each spans about 30 – 150 seconds
  • 12 daily activities: make a phone call, mop the floor, enter the room etc.

To download this database: https://publish.illinois.edu/multimodalvisualanalytics/dataset/
RGBD-HuDaAct: Sample Images

- Make a phone call
- Mop the floor
- Enter the room
- Exit the room
- Go to bed
- Get up
- Eat meal
- Drink water
- Sit down
- Stand up
- Take off the jacket
- Put on the jacket
Activity Recognition
Feature Representation I - 3DMHIs

- Depth-Induced Motion History Images (DMHIs)
  Similarly to [1], each pixel intensity is a function of the motion recency in the depth channel at that location, where brighter value corresponds to more recent motion.
- Combine depth-induced f(forward)DMHIs and b(ackward)DMHIs with color channel MHIs, obtain 3DMHIs.
- Using Hu moments for feature representation (100 × 100 pixels)

Activity Recognition Results

Feature Representation I - 3DMHI

- **Experimental Settings**
  - Leave-one-subject-out (on RGBD-HuDaAct dataset)
  - SVM classifier using linear and RBF kernels, parameters set by cross-validation
  - Compare classification accuracies

<table>
<thead>
<tr>
<th>Kernel</th>
<th>MHI</th>
<th>fDMHI + bDMHI</th>
<th>3D-MHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>34.19</td>
<td>68.66</td>
<td>70.51</td>
</tr>
<tr>
<td>RBF</td>
<td>37.18</td>
<td>66.81</td>
<td>69.66</td>
</tr>
</tbody>
</table>

- Class confusion matrix

RGBD-HuDaAct: 12 daily action classes + 1 background action class
Activity Recognition
Feature Representation II - DLMC-STIPs

• Depth-Layered Multi-Channel STIPs (DLMC-STIPs)
  Basic idea is related to space partitioning. The entire space-time video volume is divided into \( x - y - t \) sub-volumes, and STIPs [2] are spatially pooled within each \( x-y-t \) sub-volume.

\[ h = [h_1, h_2, \ldots, h_m] \]

Activity Recognition
Feature Representation II - DLMC-STIPs

Color images
STIPs
Depth maps

Visual Words Vocabulary

Depth-layered channel 1
Visual Word ID

Depth-layered channel 2
Visual Word ID

Depth-layered channel 3
Visual Word ID

Multi-channel Histogram
Activity Recognition Results
Feature Representation I - DLMC-STIPs

- Leave-one-subject-out
- SVM classifier using $\chi^2$ distance kernel, parameters set by cross-validation
- Different code book sizes
- Different number of depth layers
- Compare classification accuracies

<table>
<thead>
<tr>
<th>Setting</th>
<th>$K = 128$</th>
<th>$K = 256$</th>
<th>$K = 512$</th>
</tr>
</thead>
<tbody>
<tr>
<td>STIPs ($\chi^2$)</td>
<td>68.95</td>
<td>76.78</td>
<td>79.77</td>
</tr>
<tr>
<td>DLMC-STIPs ($\chi^2$, $M = 2$)</td>
<td>72.43</td>
<td>77.10</td>
<td>79.91</td>
</tr>
<tr>
<td>DLMC-STIPs ($\chi^2$, $M = 4$)</td>
<td>74.22</td>
<td>77.91</td>
<td>79.23</td>
</tr>
<tr>
<td>DLMC-STIPs ($\chi^2$, $M = 8$)</td>
<td>76.64</td>
<td>79.49</td>
<td>79.49</td>
</tr>
<tr>
<td>DLMC-STIPs (SPM)</td>
<td>77.64</td>
<td>81.05</td>
<td>81.48</td>
</tr>
</tbody>
</table>
Outline

- Introduction to Kinect applications
- Kinect-based multi-modal visual analytics research at ADSC
  - Kinect-based tele-rehabilitation
  - Kinect-based action recognition
  - Kinect-based event detection
  - Research Highlight: the HARL contest
  - Research Highlight: fine-grained action detection
- Conclusions
Kinect-based Event Detection

Application: get-up event detection for hospital fall prevention

- A vision system can help to detect the event patient gets up from bed in a non-intrusive way. An alarm can be sent to the nurse for assistance. Potential fall can be avoided.

- The depth camera (Kinect) provides 3-D motion sensing 24/7. Fusing depth and color information improves detection performance. Privacy can also be preserved.
Using domain knowledge, we identify a Region of Interest (ROI) around the bed area. We divide the ROI into 8 blocks of equal size. From each block, extract different features including shape (Histogram of Oriented Gradients) and motion (Histogram of Optic Flows, Motion History Images). Use Multiple-kernel SVM classifier.
Experiment

- Collect 240 video samples (40 get-up events) from 4 subjects in the hospital ward. Testing scheme is leave-one-subject out.
- Compare the detection accuracy, ROC using different feature channels and their combination.
- Compare with state-of-the-art methods: STIP and dense trajectory method [4].

### Recognition accuracy of event detector using different features

<table>
<thead>
<tr>
<th>Feature</th>
<th>All</th>
<th>MHI RGB</th>
<th>MHI Depth</th>
<th>HOF RGB</th>
<th>HOF Depth</th>
<th>HOG RGB</th>
<th>HOG Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>98.76</td>
<td>98.35</td>
<td>99.17</td>
<td>95.87</td>
<td>97.52</td>
<td>88.84</td>
<td>86.78</td>
</tr>
</tbody>
</table>

### Comparison with state-of-the-art color-based methods

<table>
<thead>
<tr>
<th>Method</th>
<th>STIP</th>
<th>Dense Trajectory</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>75.43</td>
<td>85.96</td>
<td>99.17</td>
</tr>
</tbody>
</table>

Outline

- Introduction to Kinect applications

- Kinect-based multi-modal visual analytics research at ADSC
  - Kinect-based tele-rehabilitation
  - Kinect-based action recognition
  - Kinect-based event detection
  - Research Highlight: the HARL contest
  - Research Highlight: fine-grained action detection

- Conclusions
HARL-ICPR 2012 Challenge

Multi-Level Depth + Image Fusion for Human Activity Recognition and Localization

Objective: to not classify activities, but also to detect and to localize them; focus on complex human behavior involving several people in the video at the same time, on actions involving several interacting people and on human-object interactions.

Dataset: captured by Kinect (gray + depth images); indoor office scenario; camera is moving; activities include: talk on the phone, enter/leave room, drop bag, pass object, pick up/put an object, shake hands, discuss, type on keyboard, unlock door successfully/unsuccesfully (10 classes)

Contest website: http://liris.cnrs.fr/harl2012/
HARL Challenges

Inter-class Ambiguity

Intra-class Variation

Scale Variation

Occlusion
**Methodology**

Multi-Level Depth & Image Fusion for Activity Detection

**HARL D1**: Depth + Grayscale

**Feature Extraction Level**

- With “depth”: More Accurate Detection

**Context Encoding Level**

- With “depth”: Direct in 3D, More Accurate

**Scene Modeling Level**

- With “depth”: 3D scene structural information

Integrate above three levels using Bayesian Network for more accurate activity detection
Feature Extraction Level:
Robust Human Key Pose/Object Detection

- Extracted HoG features from cropped human/object samples
- For human: apply K-means clustering to get 25 clusters, i.e., key poses
- Train HoG-SVM detector for each key pose
- Three object models: door, document box, mailbox

Robust Human Key Pose/Object Detection (Cont’d)

• Using depth based constraints to filter out false detections by HOG-SVM methods

• Significantly improves detection accuracy

Depth-based Constraint A

\[ r_l \leq r(x) = \frac{\text{Area}(x)}{d_m(x)} \leq r_u \]

Depth-based Constraint B

\[ d_m(l_b_x) > d_m(x), \quad d_m(x) < d_m(r_b_x) \]

• \( x \) – detection; \( d_m() \) – median depth value; \( r_l, r_u \) – lower and upper bounds
Human + Human Tracked Sequences

Relative 3D distance of two tracklets $f_d$

Relative 3D velocity $f_v$

Relative temporal ordering $f_o$

- All distance/velocity measurements are in X, Y, Z, t coordinates. This removes 2D projection ambiguity
- $f_d$ and $f_v$: discretize into several values
- $f_o$: discretize into 3 states: precede, overlap, and succeed
Scene Level: Depth Based Scene Modeling

• Extract surface normals using depth image
• Project onto four 2D directions: up, down, left, right
• Representation: histogram of directions + 4 centers of gravity
• Linear SVM for classification into 5 scene types

4 scene examples: different color means different directions, circles indicate centers of gravity
Results

Action localization performance

Action recognition using Bayesian Network which integrate the above mentioned three components: 1) feature extraction; 2) contextual modeling; and 3) scene modeling.

The evaluation metric is based on four criteria:

- "Recall_Temp": the fraction of the ground truth temporal length that is correctly found;
- "Prec_Temp": the fraction of the detected temporal length that is covered by ground truth;
- "Recall_Space": the portion of the ground truth bounding box space that is covered by the detected action;
- "Prec_Space": the portion of the detected bounding box space that is covered by the ground truth action.

See the evaluation metric page for details: http://liris.cnrs.fr/harl2012/evaluation.html

<table>
<thead>
<tr>
<th>Team</th>
<th>Dataset</th>
<th>Recall_Temp</th>
<th>Prec_Temp</th>
<th>Recall_Space</th>
<th>Prec_Space</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADSC-NUS-UIUC</td>
<td>D1</td>
<td>0.27</td>
<td>0.37</td>
<td>0.29</td>
<td>0.37</td>
<td>0.33</td>
</tr>
<tr>
<td>TATA-ISI</td>
<td>D1</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>VPULABUAM</td>
<td>D2</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>IACAS</td>
<td>D2</td>
<td>0.03</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Examples of detected actions:

Last three examples: false detections

<table>
<thead>
<tr>
<th>DI</th>
<th>Discussion of two or several people</th>
</tr>
</thead>
<tbody>
<tr>
<td>BO</td>
<td>An item is picked up or put down (into/from a box, drawer, desk etc.)</td>
</tr>
<tr>
<td>ET</td>
<td>A person tries to enter an office unsuccessfully</td>
</tr>
<tr>
<td>UB</td>
<td>A person leaves baggage unattended (drop and leave)</td>
</tr>
<tr>
<td>KB</td>
<td>A person types on a keyboard</td>
</tr>
<tr>
<td>TE</td>
<td>A person talks on a telephone</td>
</tr>
<tr>
<td>GI</td>
<td>A person gives an item to a second person</td>
</tr>
<tr>
<td>LO</td>
<td>A person unlocks an office and then enters it</td>
</tr>
<tr>
<td>EN</td>
<td>A person enters or leaves an office</td>
</tr>
<tr>
<td>LS</td>
<td>Handshaking of two people</td>
</tr>
<tr>
<td>HS</td>
<td>Handshaking of two people</td>
</tr>
</tbody>
</table>
Outline

- Introduction to Kinect applications
- Kinect-based multi-modal visual analytics research at ADSC
  - Kinect-based tele-rehabilitation
  - Kinect-based action recognition
  - Kinect-based event detection
  - Research Highlight: the HARL contest
  - Research Highlight: fine-grained action detection
- Conclusions
Outline

- Introduction to Kinect applications
- Kinect-based multi-modal visual analytics research at ADSC
  - Kinect-based tele-rehabilitation
  - Kinect-based action recognition
  - Kinect-based event detection
  - Research Highlight: the HARL contest
  - Research Highlight: fine-grained action detection
- Conclusions
Fine-grained Action Detection

- Fine-grained action detection has potential application in assisted living
- It is a difficult task due to frequent and subtle interaction between hand/object

![Breaking](image1)
![Cutting](image2)
![Baking](image3)
![Mixing](image4)
Coarse-to-Fine Search for Action Detection

Methodology

- Track hand and object jointly using RGB-Depth data
- Infer the “interaction status”: what is the object being manipulated and where is the position of interaction
- Use the inferred “interaction status” to retrieve relevant kitchen action sequences from the training database
- Parse the action labels from the relevant training videos towards the testing video sequence
Experimental Results

- Example frames with tracked bounding boxes for various objects (ICPR 2012 kitchen action dataset)
- Our joint hand/object tracking (solid rectangle) is better than separate hand/object tracking (dashed rectangle)
Experimental Results

- Detection performance (mean F-score) on the ICPR 2012 kitchen action dataset (KSCGR)

<table>
<thead>
<tr>
<th>Method</th>
<th>Best@KSCGR</th>
<th>DenseTraj [48]</th>
<th>Ours (w/global pooling)</th>
<th>Ours (w/interaction pooling)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean F-score</td>
<td>74.0</td>
<td>68.8</td>
<td>77.4</td>
<td>80.5</td>
</tr>
</tbody>
</table>

- Detection performance (precision, recall and average precision) on the Max-Planck-Institute for Informatics (MPII) kitchen action dataset

<table>
<thead>
<tr>
<th>Approach</th>
<th>Prec.</th>
<th>Recall</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best@MPII</td>
<td>19.8</td>
<td>40.2</td>
<td>45.0</td>
</tr>
<tr>
<td>DenseTraj [48]</td>
<td>22.3</td>
<td>44.7</td>
<td>49.6</td>
</tr>
<tr>
<td>Ours (w/global pooling)</td>
<td>25.0</td>
<td>44.1</td>
<td>47.7</td>
</tr>
<tr>
<td>Ours (w/interaction pooling)</td>
<td>30.8</td>
<td>51.3</td>
<td>57.8</td>
</tr>
</tbody>
</table>

- Our method outperforms the state-of-the-art dense trajectory based method
- Interaction centered feature pooling is more discriminative than global feature pooling as it screens out irrelevant motion information
- Interaction status based candidate sequence retrieval narrows down the entire search space, making final action detection performance more accurate
- Coarse-to-fine search scheme for action detection is effective
Conclusions

• From depth and color image sequences to multi-modal visual analytics
• Quality metrics for activity recognition tasks
• New features for depth images and for fusion
• Machine learning framework
• New applications