Kinect Fusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera

SHAHRAM IZADI, DAVID KIM, OTMAR HILLIGES, DAVID MOLYNEAUX, RICHARD NEWCOMBE, PUSHMEET KOHLI, JAMIE SHOTTON, STEVE HODGES, DUSTIN FREEMAN, ANDREW DAVIDSON, ANDREW FITZGIBBON

PROCEEDINGS OF THE 24TH ANNUAL ACM SYMPOSIUM ON USER INTERFACE SOFTWARE AND TECHNOLOGY. ACM, 2011.

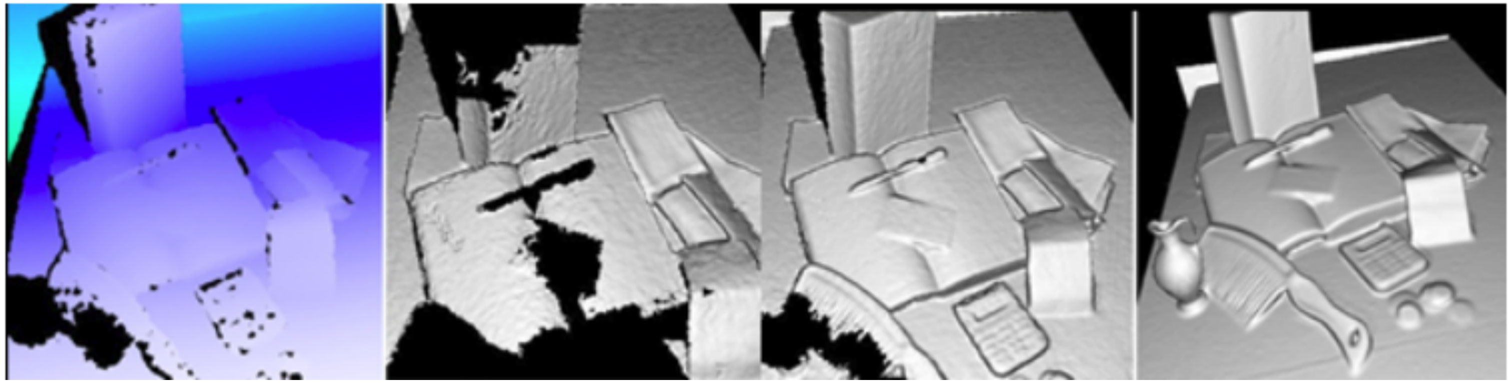
Overview

Difficult goal

3D reconstruction of an indoor scene

Use single depth camera

- Estimate pose of camera
- Compare depth map
- Update 3D reconstruction



Low-cost and real-time

Related Work:

- Active sensors
- Passive cameras
- Online Images
- Simultaneous Localization and Mapping (SLAM)

Design Goals

Interactive rates for camera tracking and reconstruction

- Direct feedback
- User interaction

No explicit feature detection

- Camera tracking avoids explicit detection step
- Works on depth maps

High-quality reconstruction of geometry

Design Goals

Dynamic interaction assumed

- user interaction is possible
- Dynamically changing scenes

Infrastructure-less

- Reconstruct arbitrary indoor spaces

Room scale

- Support room reconstructions and interaction

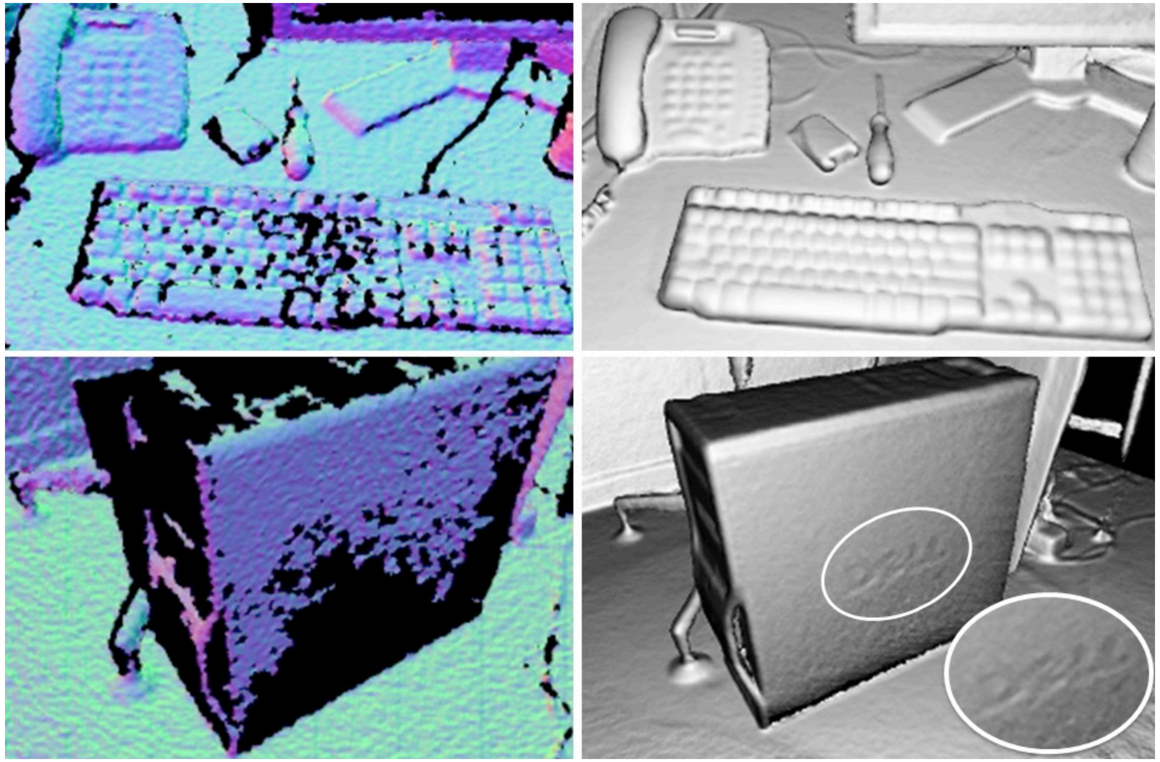
KinectFusion System

Construct 3D model of the scene:

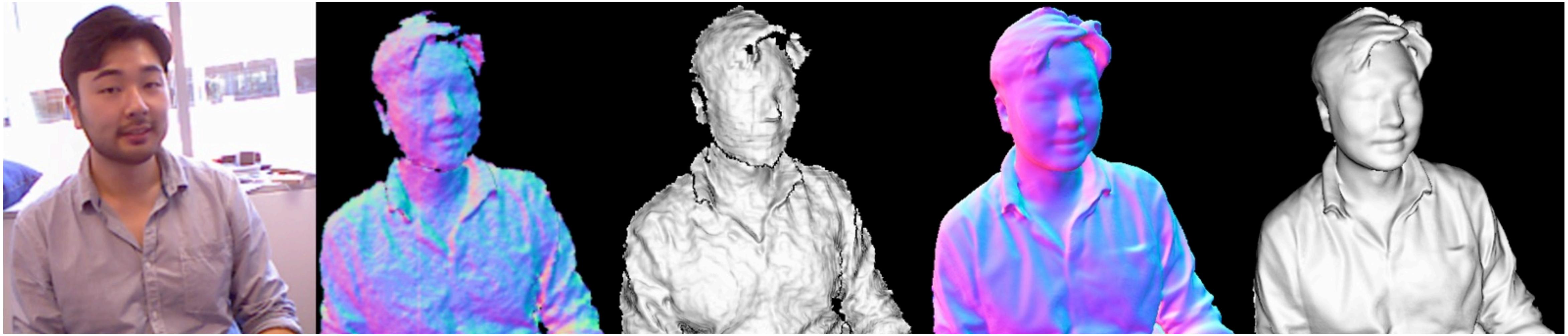
- Track 6DOF pose of camera
- Fuse live depth data into a 3D model

User explores the space

- New views
- Reconstruction grows
- Image super-resolution



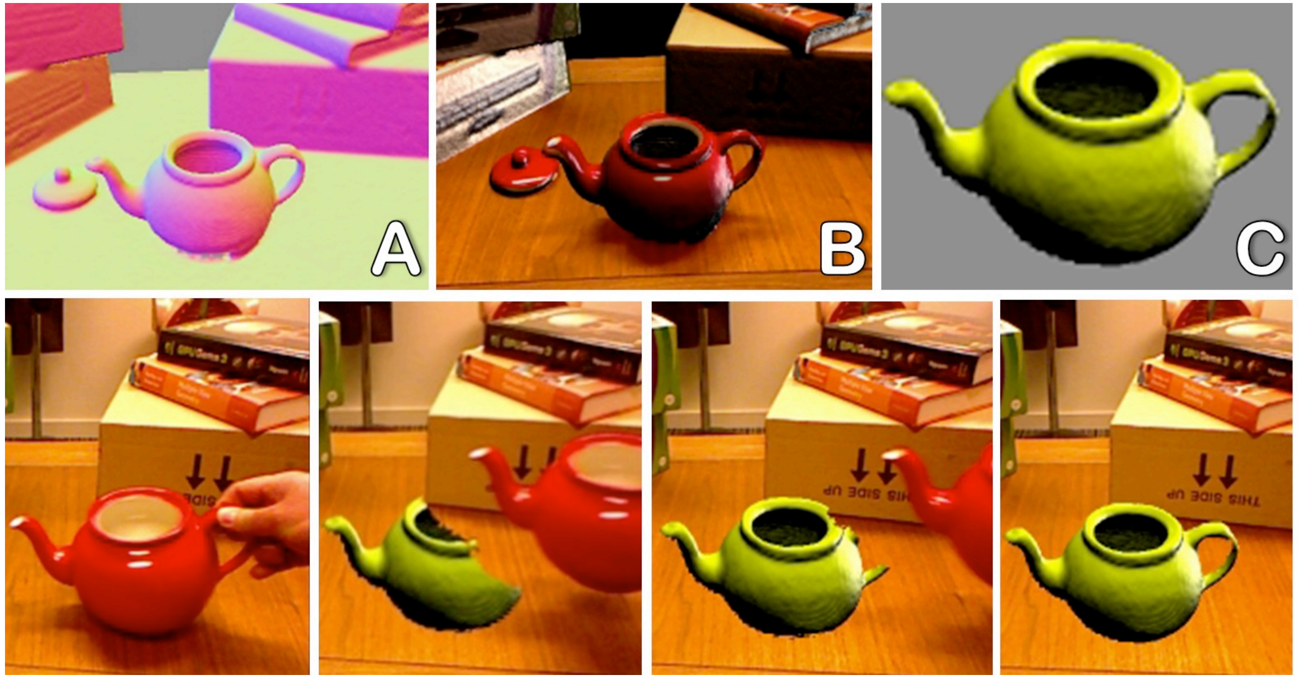
Examples



Object Segmentation

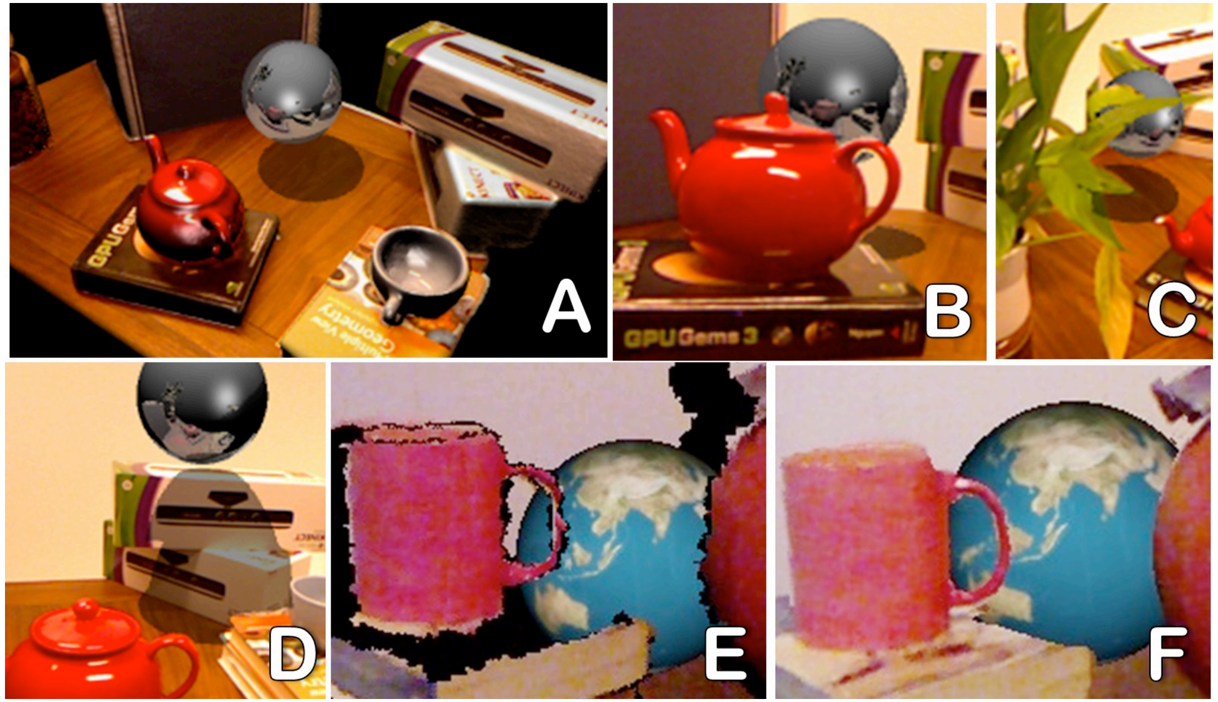
Scan specific physical object

- Monitor 3D reconstruction
- Observe changes over time
- Segment repositioned object



Geometry-Aware Augmented Reality

3D virtual world is overlaid onto the real world



Taking Physics Beyond the Surface

Simulate real-world physics.

Reaching into the Scene

User interaction

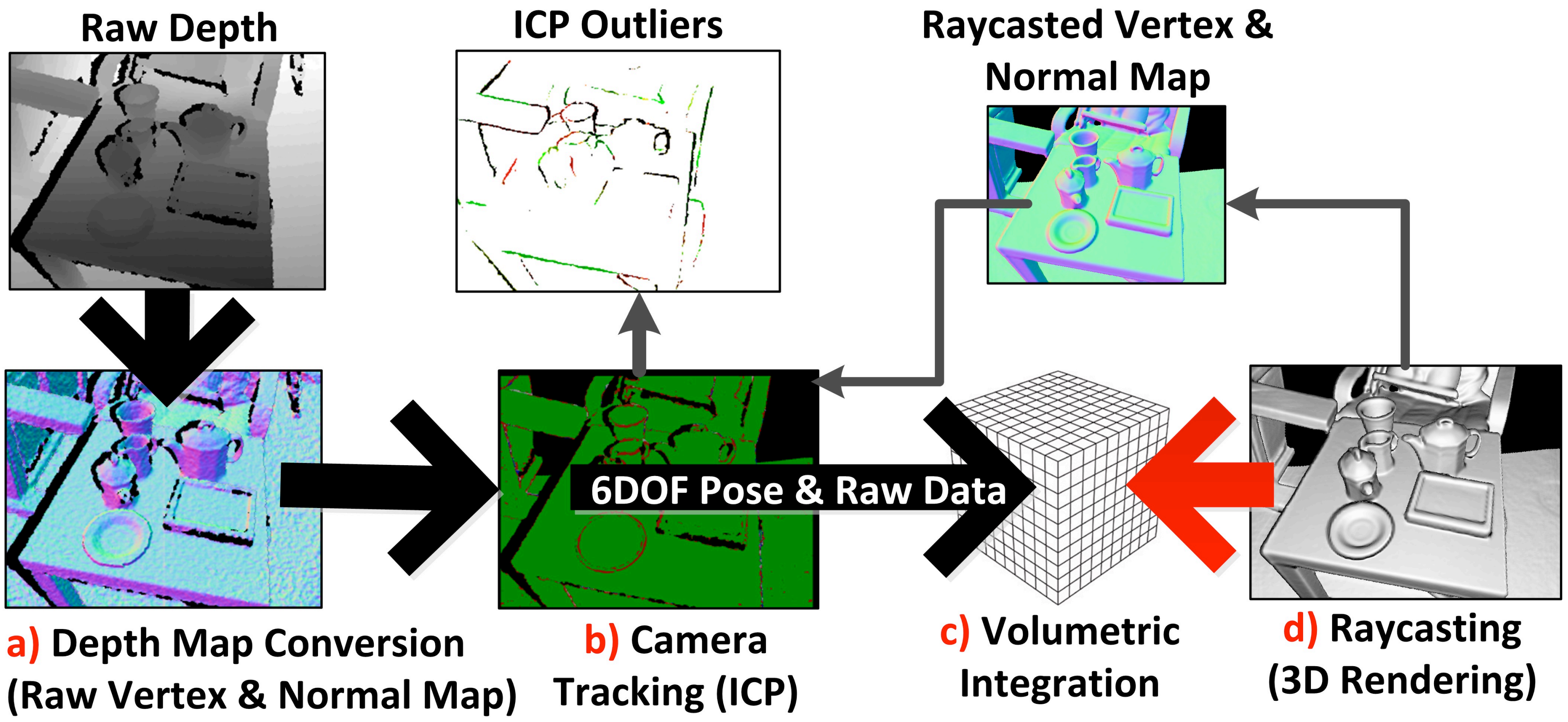
- Static scene -> dynamic scene
- Robust to transient and rapid scene motions
- Problems with prolonged interactions
  - User moves in front of the camera

Special GPU-based pipeline

- Geometry of background scene
- Geometry of the foreground user

Determine interactions

System pipeline



Camera Tracking

Iterative Closest Point (ICP)

- Projective data association
- Find correspondences between oriented points

$$\arg\min \sum_{\substack{\mathbf{u} \\ \mathbf{D}_i(\mathbf{u})>0}} ||(\mathbf{T}^{\mathrm{rel}}\mathbf{v}_i(\mathbf{u}) - \mathbf{v}^{\mathrm{g}}_{i-1}(\mathbf{u})) \cdot \mathbf{n}^{\mathrm{g}}_{i-1}(\mathbf{u})||^2$$

Output: relative transformation matrix that minimizes the point-to-plane error metric

Dense tracking

Listing 1 Projective point-plane data association.

for each image pixel $\mathbf{u} \in$ depth map $\mathbf{D_i}$ **in parallel do**
  if $\mathrm{D}_i(\mathbf{u}) > 0$ **then**
    $\mathbf{v}_{i-1} \leftarrow \mathbf{T}^{-1}_{i-1}\mathbf{v}^{\mathrm{g}}_{i-1}$
    $\mathbf{p} \leftarrow$ perspective project vertex $\mathbf{v}_{i-1}$
    if $\mathbf{p} \in$ vertex map $\mathbf{V}_i$ **then**
      $\mathbf{v} \leftarrow \mathbf{T}_{i-1}\mathbf{V}_i(\mathbf{p})$
      $\mathbf{n} \leftarrow \mathbf{R}_{i-1}\mathbf{N}_i(\mathbf{p})$
      if $||\mathbf{v} - \mathbf{v}^{\mathrm{g}}_{i-1}|| <$ distance threshold and $\mathbf{n} \cdot \mathbf{n}^{\mathrm{g}}_{i-1} <$ normal threshold **then**
        point correspondence found

D: Depth map
T: global camera pose
V: vertex map
N: Normal map
R: Rotation matrix

Volumetric Representation

3D volume with fixed resolution

Integrate 3D vertices into voxels using Signed Distance Function (SDF)

- Surface defined by the zero-crossing

Truncated Signed Distance Function (TSDF)

3D voxel grid is allocated on the GPU as aligned linear memory

Listing 2 Projective TSDF integration leveraging coalesced memory access.

for each voxel $\mathbf{g}$ in x,y volume slice **in parallel do**
  while sweeping from front slice to back **do**
    $\mathbf{v}^{\mathrm{g}} \leftarrow$ convert $\mathbf{g}$ from grid to global 3D position
    $\mathbf{v} \leftarrow \mathbf{T}_i^{-1}\mathbf{v}^{\mathrm{g}}$
    $\mathbf{p} \leftarrow$ perspective project vertex $\mathbf{v}$
    if $\mathbf{v}$ in camera view frustum **then**
      $\mathbf{sdf}_i \leftarrow ||\mathbf{t}_i - \mathbf{v}^{\mathrm{g}}|| - \mathbf{D}_i(\mathbf{p})$
      if $(\mathbf{sdf}_i > 0)$ **then**
        $\mathbf{tsdf}_i \leftarrow min(1, \mathbf{sdf}_i /$ max truncation$)$
      else
        $\mathbf{tsdf}_i \leftarrow max(-1, \mathbf{sdf}_i /$ min truncation$)$
      $\mathbf{w}_i \leftarrow$ min(max weight, $\mathbf{w}_{i-1} + 1$)
      $\mathbf{tsdf}^{\mathrm{avg}} \leftarrow (\mathbf{tsdf}_{i-1}\mathbf{w}_{i-1} + \mathbf{tsdf}_i\mathbf{w}_i)/\mathbf{w}_i$
      store $\mathbf{w}_i$ and $\mathbf{tsdf}^{\mathrm{avg}}$ at voxel $\mathbf{g}$

Summary

3D reconstruction and camera pose estimation using single depth camera

Features:

- Novel GPU pipeline – real time
- Low–cost object scanning
- Physics based interaction
- Dynamic content

Future work

- Reconstruction of larger scenes
- More details in the reconstruction
- Open new research topics

References

1. S. Izadi et al., “KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera,” in Proceedings of the 24th annual ACM symposium on User interface software and technology, 2011, pp. 559– 568.

2. https://msdn.microsoft.com/en-us/library/dn188670.aspx