บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

งานวิจัยเรื่อง “เทคนิคสำหรับการควบคุมอินเทอร์เน็ตทีวีโดยคีเนคด้วยการผสมผสานระหว่างการรู้จำเสียงและการตรวจสอบการเคลื่อนไหว” ผู้วิจัยได้ศึกษาแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดดังนี้

2.1 คีเนค (Kinect)

2.2 Extensible Application Markup Language (XAML)

2.3 การตรวจจับการเคลื่อนไหวโดยใช้กล้อง Kinect และ Kinect SDK Beta 1

2.4 การรู้จำเสียง (SPEECH RECOGNITION)

2.5 งานวิจัยที่เกี่ยวข้อง

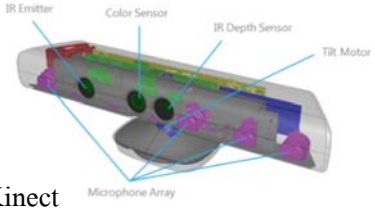
2.1 คีเนค (Kinect)

Michal Czerwonka(2010)ได้ให้ความหมายของคีเนคว่าเป็นอุปกรณ์รับรู้การเคลื่อนไหวที่ใช้การทำงานผสมผสานกันระหว่างฮาร์ดแวร์และซอฟท์แวร์ ฟังก์ชันหลักของ Kinect มีสองฟังก์ชันคือ สร้างภาพเคลื่อนไหวสามมิติของวัตถุในมุมมองที่กำหนด และแยกแยะมนุษย์ออกจากวัตถุเหล่านั้นได้ ในปัจจุบัน Kinect มีความละเอียดของภาพอยู่ที่ 640x480 pixel และสามารถทำงานได้ที่ 30 frames ต่อวินาทีสำหรับ Hardware ของ Kinect สามารถแบ่งออกเป็น 3 ส่วนประกอบสำคัญ ได้แก่

2.1.1 Color VGA video

2.1.2 Depth sensor

2.1.3 Multi-array microphone



ภาพที่ 2.1 สถาปัตยกรรมบนกล้อง Kinect

ที่มา: Dan Fernandez (2010)

2.1.1 Color VGA video camera

เป็นกล้องถ่ายภาพเคลื่อนไหวซึ่งช่วยในการจดจำใบหน้า (Face Recognition) และการตรวจจับลักษณะเด่นอื่นๆโดยใช้ความสามารถในการตรวจจับองค์ประกอบของสีทั้งสามสีอันได้แก่ สีแดง สีเขียว และสีน้ำเงิน

ภาพที่ 2.2 แสดงค่าสีที่ได้จากกล้องKinect

ที่มา: Dan Fernandez (2010)

2.1.2 Depth sensor

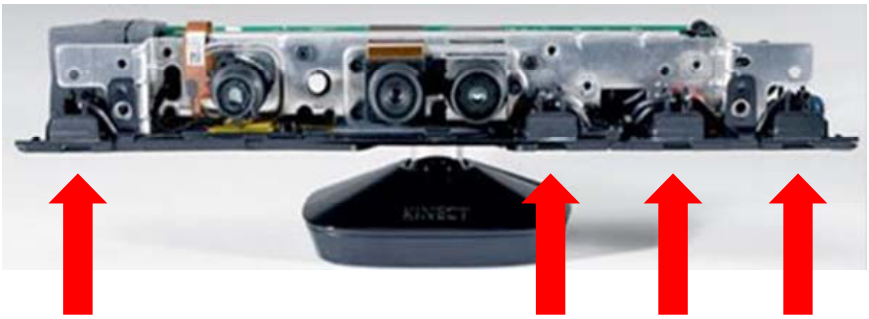
ประกอบด้วยการทำงานร่วมกันของต้นกำเนิดแสงอินฟราเรด (Infrared Projector) และตัวรับรู้แบบ monochrome CMOS(Complimentary Metal-Oxide Semiconductor) ซึ่งทำหน้าที่รับแสงอินฟราเรดที่ถูกสะท้อนกลับมาจากวัตถุ จากนั้นทำการวัดเวลาในการเดินทาง(Time of Flight) แสงอินฟราเรดนี้ใช้หลักการทำงานเช่นเดียวกับคลื่นโซนาร์คือ หากรู้ระยะเวลาที่แสงอินฟราเรดใช้ในการเดินทางไปกลับก็จะสามารถคำนวณระยะห่างระหว่างตัวรับรู้ความลึกกับวัตถุได้ โดยการทำงานด้วยความเร็วแสงหลายๆรอบทำให้สามารถระบุระยะห่างที่แน่นอนได้โดยไม่ต้องคำนึงถึงสภาพแสงสว่าง

ภาพที่ 2.3 แสงอินฟราเรดที่ใช้หาความลึกของภาพ

ที่มา: Dan Fernandez (2010)

2.1.3 Multi-array microphone

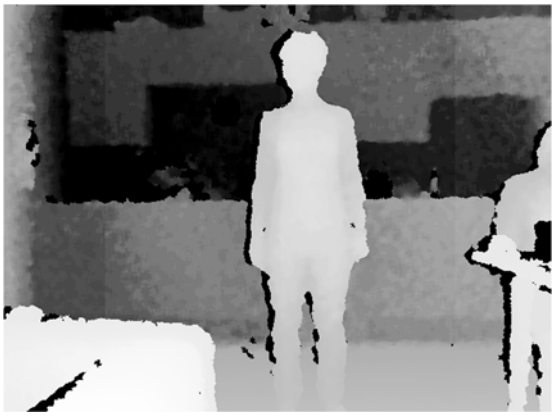
เป็นอาเรย์ของไมโครโฟน 4 ตัวซึ่งสามารถแยกแยะเสียงของผู้ใช้ออกจากเสียงรบกวนภายในห้องได้



ภาพที่ 2.4 ไมโครโฟนบนตัวKinect

ที่มา: Dan Fernandez (2010)

หลักการในการตรวจจับการเคลื่อนไหวของKinect ทำให้การประมวลผลภาพทำได้ง่ายยิ่งขึ้นโดยอาศัยความลึกมาเป็นอีกหนึ่งในปัจจัยในการแยกแยะวัตถุ จากรูปจะพบว่าการแยกตัวคนออกจากพื้นหลังโดยใช้คุณสมบัติความแตกต่างของสีและพื้นผิวจะทำได้ยาก เพราะสีของพื้นหลังและสีของเสื้อมีความคล้ายคลึงกันมาก แต่จะเป็นเรื่องที่ง่ายมากถ้าใช้คุณสมบัติของความลึกเป็นเกณฑ์ในการแบ่ง



ภาพที่ 2.5 แสดงการแยกรูปร่างของวัตถุโดยใช้คุณสมบัติของความลึก

ที่มา: Dan Fernandez (2010)

โดยทางบริษัท Microsoft ได้เผยแพร่ชุดพัฒนาซอฟท์แวร์บน Windows Platform สำหรับ Kinect(The Kinect for Windows SDK beta) ซึ่งเป็นชุดเครื่องมือโปรแกรมมิ่งสำหรับนักพัฒนาแอพพลิเคชั่น ทำให้ผู้ที่สนใจในการพัฒนาสามารถเข้าถึงการใช้งานอุปกรณ์ Microsoft Kinect ได้อย่างง่ายดายด้วยการใช้งานเชื่อมต่อผ่านระบบปฏิบัติการWindows 7โดยชุดพัฒนาซอฟต์แวร์นี้มีลักษณะเด่นดังนี้

Raw sensor streams ทำให้สามารถเข้าถึงข้อมูลดิบจากตัวรับรู้ความลึก ตัวรับรู้สีของกล้อง และ Four-element Microphone Array

Skeletal tracking ทำให้สามารถติดตามโครงร่างกระดูกของมนุษย์หนึ่งหรือสองคนที่กำลังเคลื่อนที่ได้ ทำให้สามารถสร้างแอพพลิเคชั่นที่บังคับด้วยท่าทางได้

Advanced audio capabilities ทำให้สามารถประมวลผลเสียง กำจัดเสียงรบกวนที่ซับซ้อน กำจัดเสียงสะท้อน ระบุแหล่งที่มาของเสียง และสามารถบูรณาการร่วมกับ Windows speech recognition API ได้

Sample code and documentation ประกอบด้วยเอกสารเชิงเทคนิคมากกว่า 100 หน้า เอกสารตัวอย่างต่างๆ และ Built-in help files

Easy installation สามารถติดตั้งได้อย่างรวดเร็ว ไม่มีการตั้งค่าที่ซับซ้อน และขนาดของตัวติดตั้งน้อยกว่า 100MB

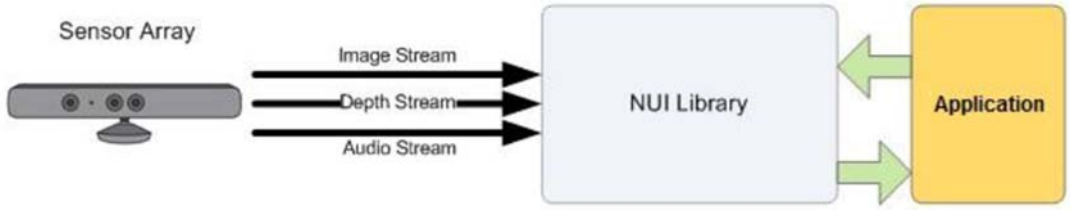
ตารางที่ 2.1 แสดงความสามารถพื้นฐานของกล้อง Kinect

Playable Ranges for the Kinect for Windows Sensor	
Sensor item	**Playable range**
Color and depth stream	4 to 11.5 feet (1.2 to 3.5 meters)
Skeletal tracking	4 to 11.5 feet (1.2 to 3.5 meters)
Viewing angle	43° vertical by 57° horizontal field of view
Mechanized tilt range (vertical)	±28°
Frame rate (depth and color stream)	30 frames per second (FPS)
Resolution, depth stream	QVGA (320 × 240)
Resolution, color stream	VGA (640 × 480)
Audio format (PCM)	16-kHz, 16-bit mono pulse code modulation

Audio input characteristics	A four-microphone array with 24-bit analog-to-digital converter (ADC) and Kinect-resident signal processing such as acoustic echo cancellation and noise suppression

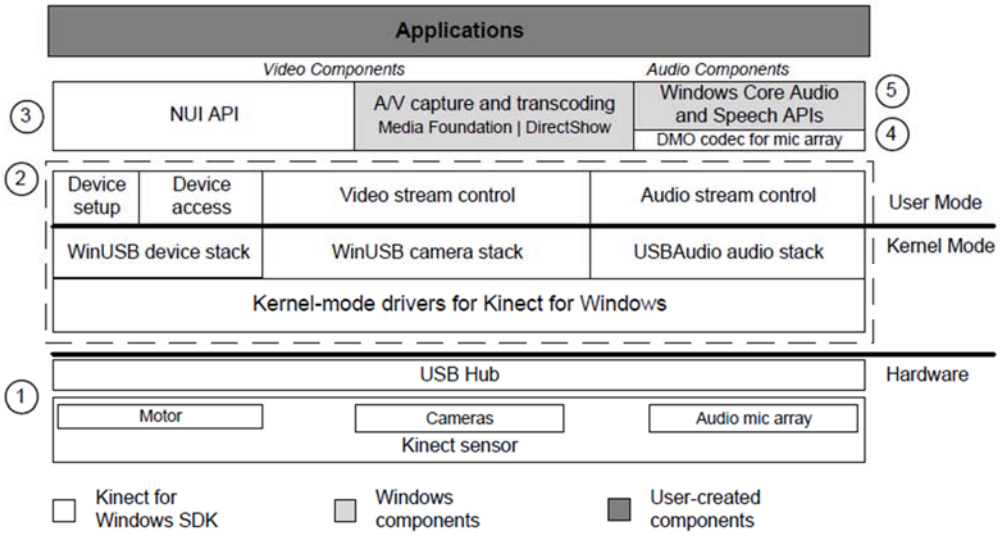
ที่มา: (http://msdn.microsoft.com/en-us/library/jj131023.aspx,2012)

Kinect for Windows Architecture



ภาพที่ 2.6 แสดงสถาปัตยกรรมKinect SDK(1)

ที่มา: (http://msdn.microsoft.com/en-us/library/jj131023.aspx,2012)



1. Kinect hardware 2. Microsoft Kinect drivers 3. NUI API 4. Kinect Audio DMO 5. Windows 7 standard APIs

ภาพที่ 2.7 แสดงสถาปัตยกรรมKinect SDK(2)

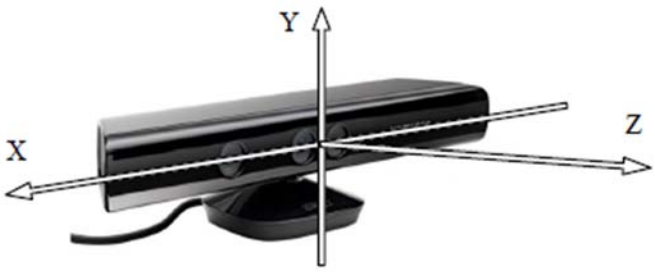
ที่มา: (http://msdn.microsoft.com/en-us/library/jj131023.aspx,2012)

2.2 Extensible Application Marksup Language (XAML)

เป็นภาษามาร์กอัปสำหรับกำหนดส่วนติดต่อผู้ใช้หรือ User Interface ในการพัฒนาโปรแกรม ใช้สำหรับกำหนดวัตถุ คุณลักษณะ ความสัมพันธ์ และการโต้ตอบของวัตถุ XAML นั้นถูกใช้เป็นหัวใจสำคัญในการสร้างส่วนติดต่อผู้ใช้งานด้วย Windows Presentation Foundation (WPF) (หรือ Avalon) ซึ่งเป็นไลบรารี่ส่วนสำหรับจัดการส่วนติดต่อผู้ใช้งานแบบใหม่ใน Microsoft .NET Framework 3.0 XAML เป็นภาษาสำหรับกำหนดส่วนติดต่อผู้ใช้ที่พัฒนามาจาก XML เช่นเดียวกับภาษา XML User Interface Languageออกแบบส่วนติดต่อผู้ใช้งาน ผ่านเครื่องมือพัฒนาอย่างเช่น Visual Studio หรือ XAMLPad

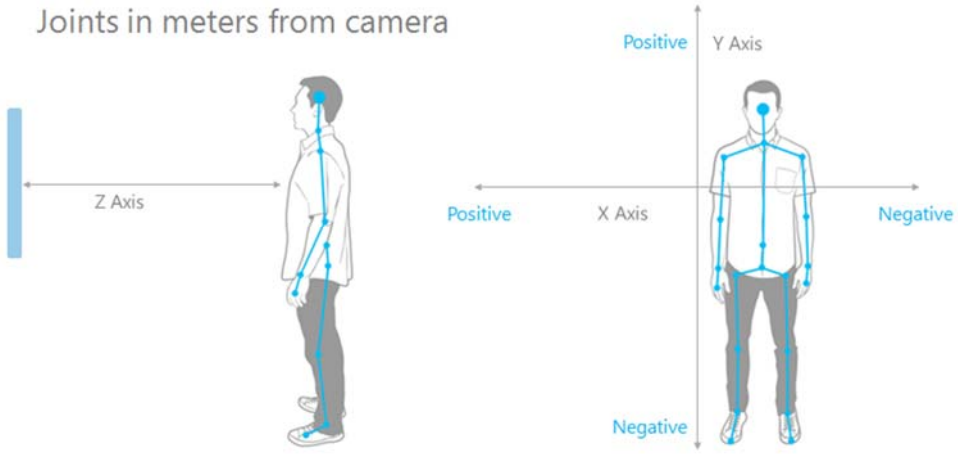
2.3 การตรวจจับการเคลื่อนไหวโดยใช้กล้อง Kinect และ Kinect SDK Beta 1

การตรวจจับการเคลื่อนไหวของร่างกายมนุษย์โดยใช้กล้อง Kinect และ Kinect SDK Beta 1 ข้อมูลตำแหน่งของข้อต่อที่ได้รับจาก Kinect SDK Beta 1 [4]จะอยู่ในรูปแบบจุดพิกัดสามมิติ (X, Y และ Z) โดยตำแหน่งของจุดกำเนิด (X = Y = Z = 0) จะเป็นตำแหน่งของกล้องที่ใช้ในการตรวจจับการเคลื่อนไหว และมีทิศทางของแกน X, Y และ Z ตามที่แสดงในภาพที่2.8ซึ่งแกน Z จะเป็นทิศทางที่กล้องตรวจจับ



ภาพที่ 2.8 แกน X,Y และ Z ของกล้อง Kinect เมื่อใช้ Kinect SDK Beta
ที่มา: นราวุฒิ พัฒโนทัย (2554: 252)

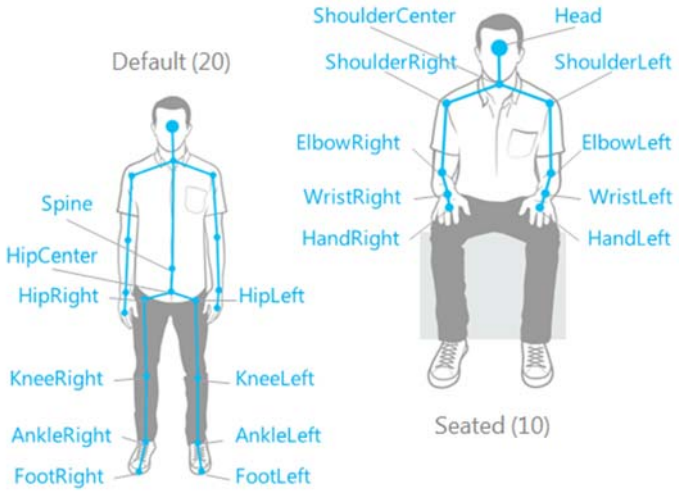
ค่า X จะเป็นค่าแสดงระยะทางที่ห่างออกไปจากจุดกำเนิดตามแนวนอนสัมพัทธ์กับกล้องที่ตรวจจับ จะมีค่าเป็นบวกเมื่อตำแหน่งที่ถูกตรวจจับได้อยู่ทางด้านขวาของกล้อง ค่า Y จะเป็นค่าแสดงระยะทางที่ห่างออกไปจากจุดกำเนิดตามแนวตั้งที่กล้องตรวจจับ จะมีค่าเป็นบวกเมื่อตำแหน่งที่ถูกตรวจจับได้อยู่สูงกว่าตำแหน่งของกล้อง และค่า Z จะเป็นระยะทางที่ห่างออกไปจากกล้องโดยจะมีค่า เป็นบวกเสมอ



ภาพที่ 2.9 แสดงแกนทางด้าน Positive และ Negative

ที่มา: (http://video.ch9.ms/teched/2012/eu/DEV330.pptx,2012)

ตำแหน่งของข้อต่อที่ได้รับจาก API จะสัมพันธ์กับตำแหน่งของร่างกายมนุษย์ 20 ตำแหน่งในท่ายืน และ 10 ตำแหน่งในท่านั่ง ตามที่แสดงในภาพที่2.10เมื่อผู้ใช้ยืนหันหน้าเข้าหากล้อง ตำแหน่งของข้อต่อข้างซ้ายและขวาจะสลับข้างกัน แตกต่างจากร่างกายจริง แต่ละตำแหน่งของข้อต่อจะมีค่า X, Y และ Z ในหน่วยมิลลิเมตร และมีฟิลด์ Confidence แสดงสถานะการตรวจจับตำแหน่งข้อต่อนั้นๆ โดยจะมีค่าระหว่าง 0 – 1 ซึ่ง 1 หมายถึง สามารถตรวจจับได้ และหากมีค่าน้อยกว่า 1 (เป็นตัวเลขทศนิยม) หมายถึง ไม่สามารถตรวจจับได้ ข้อมูลตำแหน่งของข้อต่อในส่วนนั้นจะเป็นค่าประมาณ ในบางครั้งอาจเป็นค่าที่ตรวจจับได้ล่าสุดก่อนที่จะตรวจจับไม่ได้ในเวลาต่อมา หรืออาจเป็นค่าที่ไม่มีความหมายก็ได้



ภาพที่ 2.10 ตำแหน่งของข้อต่อที่สัมพันธ์กับร่างกายมนุษย์

ที่มา: (http://video.ch9.ms/teched/2012/eu/DEV330.pptx,2012)

ในการตรวจจับการเคลื่อนไหว หากผู้ใช้อยู่ภายในมุมกล้อง ห่างจากกล้องประมาณ 1.2 – 3.5 เมตร (ระยะห่างอุดมคติตามที่ Microsoft แนะนำคือ ประมาณ 2.5 เมตร) ค่าตำแหน่งของข้อต่อที่ ถูกตรวจจับได้ จะมีค่าใกล้เคียงกับการวัดจริง เช่น ความยาวของข้อศอกไปจนถึงกลางฝ่ามือ เป็นต้น และฟิลด์ Confidence ของข้อต่อจะมีค่าเป็น 1 แต่หากมีบางข้อต่อที่กล้องไม่สามารถตรวจจับได้ หรือถูกบดบังไป เช่น ยืนมือไพล่หลัง เป็นต้น ฟิลด์ Confidence ของข้อต่อนั้นจะมีค่าน้อยกว่า 1

คำแนะนำเพิ่มเติมจาก Microsoft ในการตรวจจับการเคลื่อนไหวด้วย Kinect SDK

1. ผู้ใช้ไม่ควรสวมชุดที่หลวมโคร่งในขณะที่ทาการตรวจจับการเคลื่อนไหว เพราะจะทำให้ ไม่สามารถตรวจจับตำแหน่งของข้อต่อได้

2. การตรวจจับข้อต่อที่แขนจะมีความเสถียรน้อย หากแขนนั้นอยู่ใกล้ชิดกับส่วนของร่างกาย โดยเฉพาะอยู่ชิดกับลำตัว ถ้าแขนทั้งสองอยู่ชิดกับลำตัวหรืออยู่ใกล้กับส่วนอื่น อาจรวมแขนนั้นเป็นส่วนหนึ่งของร่างกายส่วนนั้นได้ และจะทำให้ไม่สามารถตรวจจับข้อต่อ ที่แขนได้

3. การตรวจจับข้อต่อที่ขาจะมีการตรวจจับที่ดีขึ้น หากผู้ใช้ไม่ยืนชิดขา

4. การเคลื่อนไหวอย่างรวดเร็วหรือการแสดงท่าทางที่ซับซ้อน อาจทำให้ไม่สามารถตรวจจับการเคลื่อนไหวได้ เช่น การทำท่าเตะฟุตบอลอย่างรวดเร็ว เป็นต้น

จากการใช้งานกล้อง Kinect และ Kinect SDKในการตรวจจับการเคลื่อนไหวของร่างกายมนุษย์ ในเบื้องต้น ผู้วิจัยพบว่า ตำแหน่งที่ดีที่สุดในการตั้งกล้องตรวจจับการเคลื่อนไหวคือการตั้งกล้อง ให้เผชิญหน้ากับผู้ใช้ และไม่มีสิ่งที่บดบังการตรวจจับ เพราะส่วนของร่างกายเกือบทุกๆ ส่วน จะสามารถตรวจจับได้ตลอดเวลา และหากมีสิ่งของที่อยู่ใกล้กับผู้ใช้ เช่น ผู้ใช้นั่งอยู่บนเก้าอี้ เป็นต้น Kinect SDKอาจรวมสิ่งของที่อยู่ใกล้นั้นเป็นส่วนหนึ่งของร่างกายผู้ใช้ และระบุตำแหน่งของข้อต่อ ที่ผิดพลาดได้ การระบุส่วนของร่างกายด้านซ้ายและขวาอาจสามารถสลับกันได้ เนื่องจากKinect SDKไม่สามารถวิเคราะห์ภาพโครงร่างมนุษย์ที่กล้องตรวจจับได้ว่าเป็นด้านหน้าหรือด้านหลังของร่างกาย หากผู้ใช้หันหลังให้กับกล้องตั้งแต่เริ่มการตรวจจับการเคลื่อนไหว ตำแหน่งของข้อต่อที่ควรจะเป็นด้านซ้ายก็จะสลับไปเป็นด้านขวาได้ นอกจากนี้ หากมีสิ่งของที่มีรูปร่างคล้ายกับโครงร่างมนุษย์ เช่น พัดลมตั้งพื้น เป็นต้น Kinect SDKอาจทำการตรวจจับและให้ค่าตำแหน่งได้

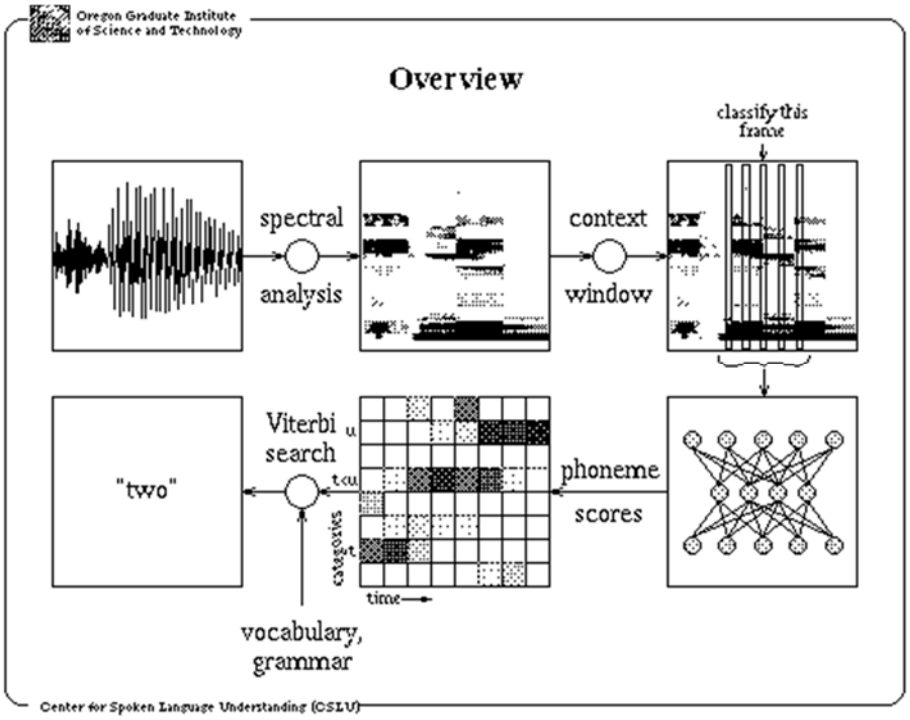
2.4 การรู้จำเสียง (SpeechRecognition)

Parinya (2001) กล่าวว่าSpeech recognition หมายถึง การทำให้คอมพิวเตอร์สามารถที่จะฟังคำพูดและตัดสินใจได้ว่าคำพูดนั้นเป็นคำว่าอะไรหรืออีกความหมายหนึ่งคือการนำ File Audio ที่บรรจุเสียงพูดนำมาแปลงเป็น Text ได้

Automatic Speech Recognition (ASR) เป็นเทคโนโลยีที่เกิดขึ้นเพื่อที่จะให้คอมพิวเตอร์สามารถแยกแยะคำพูดต่างๆที่มนุษย์สามารถพูดใส่อุปกรณ์ไมโครโฟนหรือเครื่องโทรศัพท์หรืออื่นๆเท่าที่จะเป็นไปได้ " Holy grail " of ASR Search เป็นโปรแกรมที่อนุญาตให้คอมพิวเตอร์เข้าใจคำศัพท์ทุกคำอย่างถูกต้อง 100% ซึ่งมีความสามารถเข้าใจถึงคำพูดได้ดีไม่ว่าจะเป็นคำพูดของใครก็ตาม เป็นอิสระจากขนาดของกลุ่มคำศัพท์ , ความดัง , ลักษณะของผู้พูด และการออกเสียง หรือเงื่อนไขของช่องทางต่างๆที่เป็นไปได้

เทคโนโลยีที่เป็นส่วนสำคัญที่ใช้ทำ ASR ถูกเรียกว่า " Hidden Markov Model " หรือ "HMM" เทคโนโลยีชนิดนี้สามารถที่จะเข้าใจถึงคำพูดโดยการประมาณการถึงความเป็นไปได้ของแต่ละหน่วยที่เป็นพื้นฐานของเสียงที่อยู่ติดๆกัน ซึ่งแต่ละเสียงจะมีขอบเขตของตัวสัญญาณ คำแต่ละคำในกลุ่มของคำศัพท์เหล่านั้น ต่างก็มีลักษณะเฉพาะที่มีความแตกต่างกัน โดยสังเกตจากส่วนประกอบของหน่วยที่เป็นพื้นฐานของเสียง

Procedure ที่เกี่ยวกับการค้นหาที่มีชื่อเรียกว่า " Viterbi Search " ถูกใช้เพื่อกำหนดถึงความต่อเนื่องของหน่วยพื้นฐานของเสียงด้วยความเป็นไปได้สูงสุด เทคนิคชนิดนี้ถูกจำกัดเฉพาะมองหาจากความต่อเนื่องของหน่วยพื้นฐานของเสียง ซึ่งตรงกันกับความคำต่างๆในกลุ่มของคำศัพท์ที่มี และความต่อเนื่องของหน่วยพื้นฐานของเสียงเหล่านั้นด้วยความเป็นไปได้สูงสุดทั้งหมดถูกระบุด้วยคำศัพท์ที่ถูกพูดออกมา ในมาตรฐานของ HMMs ความเป็นไปได้จะถูกคำนวณโดยใช้ " a Gaussian Mixture Model " : ซึ่งอยู่ในขอบเขต HMM/ANN ซึ่งค่าเหล่านี้จะถูกคำนวณโดย Artificial neural network (ANN)



ภาพที่ 2.11 รวมการทำงานของ Speech Recognition

ที่มา: (http://vclass.mgt.psu.ac.th/~parinya/project2001/speech_recognition.html,2001)

แบบจำลองฮิดเด็นมาร์คอฟ สามารถโมเดลคำหรือหน่วยเสียงก็ได้ แต่เหตุผลของการที่จะโมเดลว่าให้เป็นคำหรือหน่วยเสียง จะขึ้นอยู่กับการใช้งานของระบบการรู้จำเสียงเช่น ถ้าระบบรู้จำเสียงมีข้อจำกัดคือให้รู้จำคำศัพท์ที่น้อยและรู้จำคำพูดที่ไม่ติดต่อกัน การโมเดลเสียงให้เป็นคำเสียงจะดีกว่าการโมเดลหน่วยเสียงเนื่องจากการโมเดลเป็นคำจะให้ความแม่นยำมากกว่าและวิธีการของระบบรู้จำคำเสียงจะง่ายกว่าการโมเดลหน่วยเสียง การโมเดลหน่วยเสียงจะต้องการโมเดลแต่ละหน่วยเสียงและจะต้องรู้จำหน่วยเสียงนั้นๆก่อน เมื่อรู้จำหน่วยเสียงได้แล้วก็จะต้องมีวิธีการสร้างคำจากหน่วยเสียงที่ได้รู้จำมา ถ้ามีการรู้จำหน่วยเสียงหนึ่งผิดไปก็จะทำให้คำคำนั้นรู้จำผิดได้ แต่สำหรับการโมเดลคำเสียง ความแม่นยำในการรู้จำจะขึ้นโดยตรงกับโมเดลคำเสียงที่ได้ฝึกฝนมาและไม่ต้องการวิธีสร้างคำจากหน่วยเสียง แต่ถ้าสำหรับระบบที่มีคำศัพท์มาก (มากกว่า 1000 คำ) และรู้จำเสียงแบบพูดติดต่อกันระบบรู้จำที่มีโมเดลหน่วยเสียงจะมีประสิทธิภาพมากกว่าเพราะจำนวนโมเดลของหน่วยเสียงที่ต้องการจะมีน้อยกว่าโมเดลของคำเสียงและทำให้ระบบรู้จำเสียงใช้เวลาในการค้นหาคำเสียงที่มีความน่าจะเป็นมากที่สุดมากกว่า อีกประการหนึ่งที่สำคัญก็คือการรู้จำคำพูดแบบติดต่อกัน มีความเป็นไปได้สูงที่คำหนึ่งคำจะเปลี่ยนไปเมื่อใช้กับคำๆอื่น ถ้าระบบมีการฝึกฝน

ที่ไม่ดีก็จะทำให้ระบบรู้จำมีความแม่นยำต่ำ แต่สำหรับหน่วยเสียงจะมีการเปลี่ยนแปลงน้อยกว่าเมื่อใช้กับคำคำอื่นจึงทำให้มีความแม่นยำมากกว่า

ตามทฤษฎีแบบจำลองฮิดเด็นมาร์คอฟ คือ การเก็บรวบรวมของสถานะหลายๆ สถานะที่ถูกเชื่อมโยงโดยการเปลี่ยนสถานะในที่นี้คือ ลักษณะของสัญญาณเสียงหรือเครื่องหมายที่หมายถึงสัญญาณเสียงนั้น ณ เวลาหนึ่งๆ ส่วนการเปลี่ยนสถานะคือ การเลื่อนหรือเปลี่ยนแปลงของสัญญาณเสียงจากเวลาหนึ่งไปหาอีกเวลาหนึ่ง แบบจำลองฮิดเด็นมาร์คอฟ มีค่าความน่าจะเป็นอยู่สองชนิดคือ ค่าความน่าจะเป็นที่การเปลี่ยนสถานะหนึ่งที่จะเกิดขึ้น (transition probability) และ ค่าความน่าจะเป็นของผลลัพธ์ (output symbol) เมื่อมีการเปลี่ยนสถานะหนึ่งเกิดขึ้นผลลัพธ์ที่ว่านี้จะนำไปใช้ในการตัดสินใจว่าเสียงที่ได้ยินนั้นเป็นเสียงอะไร (output probability)
ปัญหาของแบบจำลองฮิดเด็นมาร์คอฟมีอยู่สามข้อด้วยกันคือ

1. Evaluation Problem: แบบจำลองสามารถสร้างหรือเข้าใจเสียงที่ได้ยินดีแค่ไหน
2. Decoding Problem: ลำดับของสถานะในแบบจำลองสร้างเสียงที่ได้ยินดีแค่ไหน
3. Learning Problem: ตัวแปร (parameter) ของแบบจำลองควรเป็นอย่างไรเพื่อที่จะทำให้มีความผิดพลาดในการสร้างเสียงหรือเข้าใจเสียงที่ได้ยินน้อยที่สุด

ถ้าปัญหา Evaluation Problem ถูกแก้ไข ก็จะมีทางที่ทำให้แบบจำลองมีลักษณะเดียวกันกับเสียงที่ได้ยิน ซึ่งสามารถใช้ในการรู้จำเสียงพูดที่ไม่ติดต่อกันได้ ถ้าปัญหา Decoding Problem ถูกแก้ไข ก็จะสามารถหาลำดับของสถานะที่ดีที่สุดในการทำความเข้าใจหรือเปรียบเทียบกับเสียงที่ได้ยิน เพื่อที่จะนำมาใช้ในการรู้จำเสียงที่ติดต่อกันได้ และที่สำคัญที่สุดคือถ้าปัญหา Learning Problem ถูกแก้ไข ก็จะทำให้มีระบบที่สามารถเรียนรู้ได้แบบอัตโนมัติจากค่าตัวแปร (Parameter) ต่างๆโดยการฝึกฝนจากชุดข้อมูลที่ป้อนเข้าระบบ

2.5 Microsoft Speech Application Programming Interface (SAPI)

โปรแกรมสำหรับออกเสียงสนทนาหรือ SAPIคือ API ซึ่งพัฒนาโดย Microsoft รองรับการรู้จำเสียงสนทนา และการสังเคราะห์เสียงในโปรแกรมสำหรับ Windows เวอร์ชันปัจจุบันที่ถูกปล่อยออกมาซึ่งถูกนำไปประกอบกับส่วนต่างๆของ Speech SDKหรือแต่ละส่วนของระบบปฏิบัติการ Windows เอง ประกอบด้วย Microsoft Office, Microsoft Agent และ Microsoft Speech Server ในเวอร์ชันทั่วไปของ API ถูกออกแบบมาเพื่อให้นักพัฒนาสามารถเขียนโปรแกรมประยุกต์เพื่อรู้จำเสียงสนทนาและสังเคราะห์เสียงโดยใช้มาตรฐานเดียวกัน สามารถใช้ภาษาในการพัฒนาได้หลากหลาย นอกจากนี้ยังมีความเป็นไปได้สำหรับบริษัทอื่นๆที่จะพัฒนา Text-to-Speech หรือ Speech-to-Textของตัวเองโดยพัฒนาจาก SAPI โดยหลักแล้วสามารถใช้จนกว่าจะพัฒนาได้ดี

ยิ่งกว่าหรือใช้แทนของ Microsoft ได้ โดยทั่วไป Speech API สามารถเผยแพร่ได้อย่างอิสระซึ่งสามารถส่งไปพร้อมกับโปรแกรมประยุกต์ของ Windows ที่ใช้เทคโนโลยีการพูด แต่ไม่ใช่ทุกเวอร์ชันของการรู้จำเสียงสนทนาและการสังเคราะห์เสียงจะไม่เสียค่าใช้จ่าย มีอยู่ 2 ตระกูลหลักด้วยกันของ Microsoft Speech API โดย SAPI 1-4 จะคล้ายกับตัวอื่นๆแต่จะมีเพิ่มคุณลักษณะพิเศษเข้ามาในเวอร์ชันที่ใหม่กว่า SAPI เวอร์ชัน 5 ถูกสร้างในปี 2000และมีเวอร์ชันย่อยออกมาหลังจากนั้นเป็นระยะๆ

2.5.1 SAPI (Speech Application Programming Interface)

แบ่งได้ 2 ประเภท คือ Text-to-Speech และ Speech-to-Text

Text-to-Speech เป็นส่วนที่สร้างเสียงพูดของมนุษย์ ด้วยการผสมเสียงจากตัวอักษรที่ประกอบเป็นคำหรือประโยค เพื่อให้เครื่องคอมพิวเตอร์สามารถสร้างเสียงเลียนแบบเสียงมนุษย์และความหมายที่มนุษย์สามารถเข้าใจ

Speech-to-Text เป็นที่ใช้ในการวิเคราะห์จำนวนเสียง และควบคุมการทำงานคอมพิวเตอร์ด้วยเสียง โดยจะพยายามทำความเข้าใจในข้อมูลเสียงที่ได้รับ โดยจะทำการตรวจสอบโครงสร้างและรูปแบบของเสียงว่าถูกต้องตามที่กำหนดหรือไม่ และดำเนินการตามเงื่อนไขของโปรแกรมที่กำหนดไว้

2.5.2 SAPI 5.1

ในเวอร์ชันนี้ถูกปล่อยมาหลังจากปี 2001เป็นส่วนหนึ่งของ Speech SDK 5.1 ซึ่งได้เพิ่ม Automation-compliant interfaces เข้ามารองรับการใช้งานโดย Visual Basic ภาษาสคริปต์ได้แก่ JScript และ managed code เวอร์ชันนี้ถูกปล่อยมาพร้อมกับ Windows XP ได้พัฒนาเวอร์ชัน 6 ออกมาใน Office 2003 และ Windows XP Tablet PC Edition

2.5.3 สถาปัตยกรรมพื้นฐาน (SAPI)

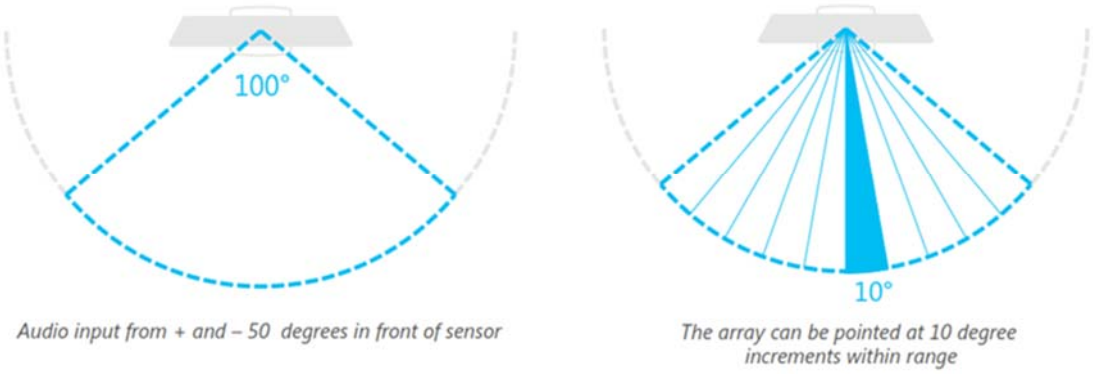
Speech APIเป็นที่รู้จักอย่างกว้างขวางในลักษณะของอินเทอร์เฟส (Interface) ซึ่งถูกจัดให้อยู่ระหว่างแอพพลิเคชั่น (Application) และSpeech engine ใน SAPI 1 - 4, แอพพลิเคชั่นสามารถสื่อสารกับเอนจิ้นได้โดยตรง มีinterface definitionที่แอพพลิเคชั่นและเอนจิ้นได้ทำการปรับเปลี่ยนให้สามารถทำงานด้วยได้ และแอพพลิเคชั่นก็ยังสามารถใช้Objectระดับสูงได้โดยง่าย โดยที่ไม่ต้องเรียกใช้เมธอด (Method) จากเอนจิ้น

แต่ใน API เวอร์ชัน 5 นั้น แอพพลิเคชั่นและเอนจิ้นไม่ต้องสื่อสารกันโดยตรงอีกต่อไป แต่ติดต่อกับรันไทม์(sapi.dll)แทน มี API Implement ที่กระทำโดยตัวรันไทม์นี้และอีกตัวถูกตั้งขึ้นเพื่อเป็น interface ของเอนจิ้นในเวอร์ชัน 5คำสั่งของแอพพลิเคชั่นเรียกผ่าน API (ตัวอย่างเช่น การจดจำแกรมมาหรือการเตรียมข้อความเพื่อนำไปสังเคราะห์) sapi.dll รันไทม์ ตัวนี้จะทำการแปลง

คำสั่งและประมวลผล เมื่อมีการเรียกใช้เอนจิ้นเกิดขึ้นผ่านเอนจิ้น อินเตอร์เฟส (เช่น การใช้แกรมมาจากไฟล์ที่ได้กระทำเสร็จแล้ว แต่ข้อมูลแกรมมาถูกส่งไปยัง recognition engine เพื่อใช้ในขั้นตอนการ recognition) recognition และ synthesis engine ทำการสร้าง event ในขณะที่ประมวลผล(เช่น การออกเสียงถูกนำไปต่อกันเพื่อสร้างเสียงสังเคราะห์ขึ้นมา) ทั้งหมดนี้เกิดขึ้นในทิศทางที่ตรงกันข้าม จากเอนจิ้น ผ่านรันไทม์ dll และไปสู่ event sink ที่อยู่ในแอพพลิเคชั่น

จากคำจำกัดความของ API และ dll รันไทม์แล้ว ส่วนประกอบอื่นๆถูกเพิ่มเข้าไปเพื่อทำให้ toolkit นี้สมบูรณ์ ส่วนประกอบต่อไปนี้มีอยู่ในเกือบทุกเวอร์ชันของ Speech SDK

1. API definition files – อยู่ใน MIDLพัฒนาโดยภาษา C หรือ C++
2. Runtime components – เช่น sapi.dll.
3. Control Panel applet – มีไว้สำหรับเลือก default speech recognizer และ synthesizer.
4. Text-To-Speech engines - รองรับหลายๆภาษา
5. Speech Recognition engines - รองรับหลายๆภาษา
6. Redistributable components อนุญาตให้ผู้พัฒนานำไปใช้ในโค้ดและแอพพลิเคชั่นที่สร้างขึ้นได้
7. Sample application code.
8. Sample engines – คำสั่งที่จำเป็นของเอนจิ้น แต่ไม่ใช่การประมวลผลเสียงอ่านทีแท้จริง



ภาพที่ 2.12 แสดงมุมองศาและทิศทางในการรับเสียงจาก Kinectบน SAPI 5.1

ที่มา: (http://video.ch9.ms/teched/2012/eu/DEV330.pptx,2012)

2.6 งานวิจัยที่เกี่ยวข้อง

การตรวจจับการเคลื่อนไหวของร่างกายมนุษย์ (Human Motion Capture) เป็นกระบวนการการบันทึกการเคลื่อนไหวของร่างกายมนุษย์ให้อยู่ในรูปแบบดิจิตอล การตรวจจับการเคลื่อนไหวของร่างกายสามารถแบ่งออกได้เป็น 2 รูปแบบหลัก ได้แก่ การตรวจจับการเคลื่อนไหวโดยการทำเครื่องหมายตามตำแหน่งต่างๆ บนร่างกาย (Marker Motion Capture) และการตรวจจับการเคลื่อนไหวโดยปราศจากการทำเครื่องหมายตามตำแหน่งต่างๆ บนร่างกาย (Markerless Motion Capture)

การตรวจจับการเคลื่อนไหวโดยการทำเครื่องหมายตามตำแหน่งต่างๆ บนร่างกาย จะใช้วิธีการติดเครื่องหมายตามข้อต่อต่างๆ บนร่างกาย หรือติดเครื่องหมายลงบนชุดที่สวมใส่ แล้วใช้กล้องตรวจจับเครื่องหมายที่ติดไว้ จากนั้นจะใช้ซอฟต์แวร์ทำการวิเคราะห์ตำแหน่งของเครื่องหมายเพื่อสร้างตำแหน่งของข้อต่อของร่างกายในรูปแบบสามมิติ การทำเครื่องหมายตามตำแหน่งต่างๆ บนร่างกายอีกวิธีการหนึ่ง จะใช้วิธีการติดอุปกรณ์บอกตำแหน่งไว้ที่ร่างกายโดยตรง วิธีการนี้จึงไม่ต้องใช้กล้องในการตรวจจับ และทำให้ได้ตำแหน่งของข้อต่อในสามมิติโดยตรง

การตรวจจับการเคลื่อนไหวโดยปราศจากการทำเครื่องหมายตามตำแหน่งต่างๆ บนร่างกาย จะใช้กล้องตรวจจับภาพการเคลื่อนไหว แล้วใช้ซอฟต์แวร์ทำการวิเคราะห์ภาพที่กล้องตรวจจับได้ เพื่อแยกภาพร่างกายออกจากภาพพื้นหลัง จากนั้นจะทำการวิเคราะห์ภาพร่างกาย เพื่อสร้างตำแหน่งของข้อต่อของร่างกายในรูปแบบสามมิติ

2.6.1 Evaluating a Dancer's Performance using Kinect-based Skeleton Tracking

งานวิจัยของ D. Alexiadis, P. Kelly, P. Daras, N. E. O'Connor, T. Boubekeur,และ M. B. Moussa (2011) ได้ทำการศึกษาเกี่ยวกับการวัดประสิทธิภาพของนักเต้นโดยใช้กล้อง Kinect ในการตรวจจับการเคลื่อนไหว ในการให้ข้อมูลการตรวจจับงานวิจัยนี้ใช้การเปรียบเทียบตำแหน่งของข้อต่อของนักเต้นสองคนที่ได้จากการตรวจจับตามเวลาจริงมาเป็นส่วนหนึ่งในการให้คะแนนการเต้น ผลการศึกษาของงานวิจัยนี้แสดงให้เห็นว่าสามารถนำข้อมูลตำแหน่งของข้อต่อและมือ ไปใช้งานในด้านการตรวจจับการเคลื่อนไหวได้ภาพที่ 2.13

ภาพที่ 2.13 การตรวจจับการเคลื่อนไหวของคนและโมเดล
ที่มา: (http://doras.dcu.ie/16574/)

2.6.2 Free Viewpoint Virtual Try-On With Commodity Depth Cameras

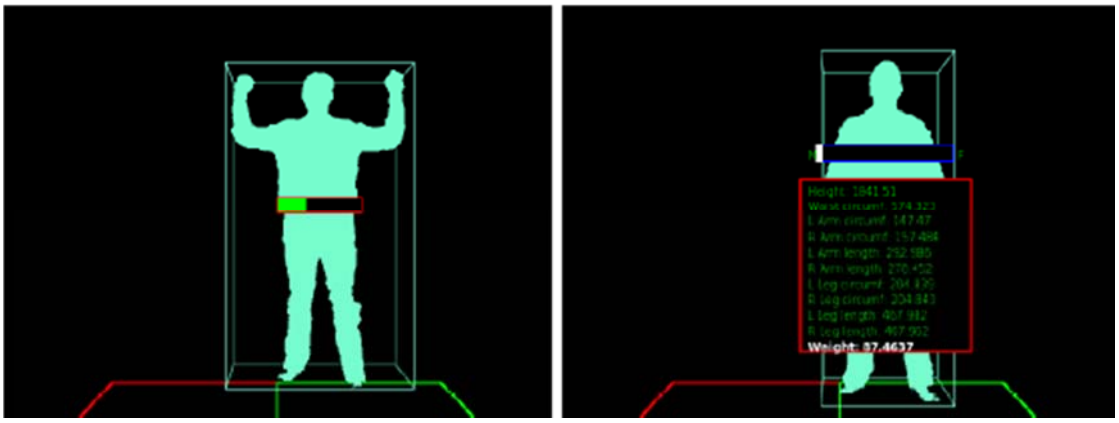
งานวิจัยของ S. Hauswiesner, M. Straka และ G. Reitmayr (2011) ได้เสนอระบบการลองเสื้อผ้าจาก ที่บ้านผ่านระบบเสมือนจริง โดยใช้กล้อง Kinect ตรวจจับผู้ใช้ และตรวจจับเสื้อผ้าที่ต้องการ จากนั้นระบบจะทำการสร้างแบบจำลองสามมิติของผู้ใช้และเสื้อผ้า แล้วนำมาแสดงประกอบกันเป็นภาพผู้ใช้สวมใส่เสื้อผ้าตามที่ได้เลือกไว้ โดยแสดงท่าทางเดียวกันกับผู้ใช้ตามเวลาจริง ตามที่แสดงในภาพที่ 2.14 งานวิจัยนี้ทดสอบระบบที่เสนอโดยการใช้แบบสอบถามสอบถามผู้ทดลอง ซึ่งปรากฏผลตอบรับที่ดีจากผู้ทดลอง

ภาพที่ 2.14 ระบบการลองเสื้อผ้าเสมือนจริง

ที่มา: (http://dl.acm.org/citation.cfm?id=2087759)

2.6.3 Real Time Extraction of Body Soft Biometric from 3D Videos

งานวิจัยของ C. Velardo และ J. Dugelay (2011) ได้เสนอระบบการบ่งบอกคุณลักษณะภายนอกของบุคคล (Body Soft Biometric) โดยใช้กล้อง Kinect ตรวจจับร่างกายของผู้ใช้ จากนั้นจะนำข้อมูลการตรวจจับที่ได้ไปคำนวณหาส่วนสูง น้ำหนัก และเพศของผู้ใช้คนนั้นโดยประมาณ ตามที่แสดงใน ภาพที่ 2.15 งานวิจัยนี้แสดงให้เห็นถึงแนวทางในการระบุตัวตนของบุคคลโดยใช้กล้อง Kinect



ภาพที่ 2.15 ระบบการบ่งบอกคุณลักษณะภายนอกของบุคคลโดยใช้กล้องKinect

ที่มา (http://dl.acm.org/citation.cfm?id=2072298.2072454)

2.6.4 Workshop on Digital Media and Digital Content Management

งานวิจัยของ M. K. M. bin Sidik, M. S. bin Sunar, I. bin Ismail, M. K. bin Mokhtar และ N. binti M.Jusoh (2011) ได้ทำการศึกษาเกี่ยวกับการปฏิสัมพันธ์ระหว่างผู้ใช้กับคอมพิวเตอร์โดยการตรวจจับการเคลื่อนไหวด้วยการใช้ภาพความลึก (Depth Image) การศึกษานี้ได้แสดงให้เห็นถึงโครงสร้างของระบบการวิเคราะห์การเคลื่อนไหวของร่างกายมนุษย์ ซึ่งประกอบด้วย 4 ขั้นตอนดังนี้

1. Initialization เป็นขั้นตอนในการเตรียมความพร้อมก่อนที่จะเริ่มทำการตรวจจับการเคลื่อนไหว

2. Tracking เป็นขั้นตอนในการแบ่งแยกระหว่างผู้ถูกตรวจจับกับฉากพื้นหลัง

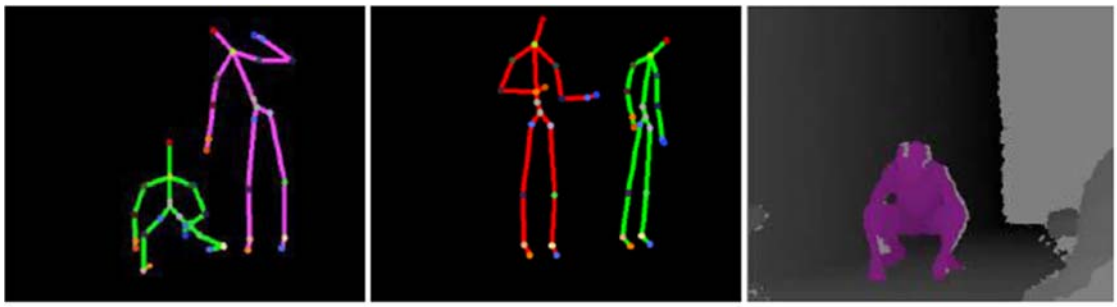
3. Pose Estimation เป็นขั้นตอนในการประมาณการแสดงท่าทางของผู้ถูกตรวจจับ

4. Recognition เป็นขั้นตอนในการวิเคราะห์และจำแนกการกระทำของผู้ถูกตรวจจับ

ผลการศึกษาของงานวิจัยนี้แสดงให้เห็นถึงประโยชน์ของการใช้ภาพความลึกในการตรวจจับการเคลื่อนไหวของร่างกายมนุษย์ ซึ่งการใช้ภาพความลึกจะช่วยลดขั้นตอนและเวลาในการแบ่งแยกระหว่างผู้ถูกตรวจจับกับฉากพื้นหลัง

2.6.5 Third Chinese Conference on Intelligent Visual Surveillance

งานวิจัยของ X. Yu, L. Wu, Q. Liu และ H. Zhou (2011)ได้เสนอวิธีการประเมินการแสดงอารมณ์ด้านลบของเด็กด้วยข้อมูลที่ได้รับจากการตรวจจับการเคลื่อนไหวโดยใช้กล้อง Kinect ตามที่แสดงในภาพที่ 2.16 งานวิจัยนี้นำข้อมูลการตรวจจับการเคลื่อนไหวที่ได้ไปวิเคราะห์ด้วยอัลกอริทึมในการวิเคราะห์พฤติกรรมแบบ Stochastic Grammar งานวิจัยนี้ทำการทดสอบระบบต้นแบบเปรียบเทียบกับการบันทึกภาพเคลื่อนไหวและการจดบันทึกพฤติกรรมที่เด็กแสดงออก ผลของงานวิจัยนี้แสดงให้เห็นถึงระบบต้นแบบที่ช่วยในการวิเคราะห์พฤติกรรมและการแสดงอารมณ์ด้านลบของเด็ก และช่วยในการแก้ไขพฤติกรรมของเด็ก



ภาพที่ 2.16 ตัวอย่างพฤติกรรมของเด็กที่บันทึกโดยใช้กล้อง Kinect

ที่มา: (http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6153167)

2.6.6 Development and Evaluation of Low Cost Game-Based Balance Rehabilitation Tool Using the Microsoft Kinect Sensor

งานวิจัยของ B. Lange, C. Chang, E. Suma, B. Newman, A. S. Rizzo และ M. Bolas (2011) ได้เสนอเกมสำหรับฝึกการทรงตัวของผู้ป่วยทางระบบประสาท โดยใช้กล้อง Kinect ในการตรวจจับการเคลื่อนไหวของผู้ป่วย ในการให้ข้อมูลตำแหน่งของข้อต่อ แล้วนำข้อมูลการตรวจจับการเคลื่อนไหวที่ได้ไปควบคุมตัวการ์ตูนในเกม งานวิจัยนี้ทำการทดสอบเกมที่นำเสนอกับผู้ป่วยทางระบบประสาท ผลของงานวิจัยนี้แสดงให้เห็นถึงเกมที่ช่วยในการพัฒนาการทรงตัวของผู้ป่วยที่มีความเพลิดเพลิน และผู้ป่วยมีความพึงพอใจมาก

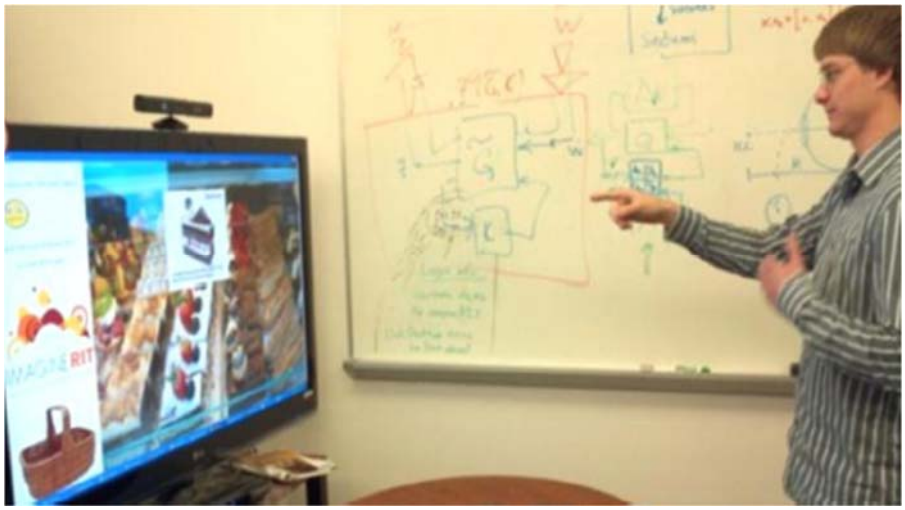
2.6.7 Controller-free exploration of medical image data: experiencing the Kinect

งานวิจัยของ L. Gallo, A. P. Placitelli และ M. Ciampi (2011) ได้เสนอระบบการควบคุมการแสดงรูปภาพทางการแพทย์ด้วยท่าทางของผู้ใช้ ตามที่แสดงในภาพที่ 2.17งานวิจัยนี้ใช้กล้อง Kinect ในการตรวจจับท่าทางของผู้ใช้ ในการให้ข้อมูลตำแหน่งของข้อต่อ แล้วนำข้อมูลตำแหน่งของข้อต่อที่มือและแขนไปวิเคราะห์เป็นคำสั่งในการควบคุมการแสดงรูปภาพ ผลการศึกษาของงานวิจัยนี้แสดงให้เห็นถึงระบบการควบคุมการแสดงรูปภาพทางการแพทย์ที่ปราศจากการใช้อุปกรณ์ควบคุม มีค่าใช้จ่ายน้อย และสามารถหาซื้ออุปกรณ์ได้ตามร้านค้าทั่วไป

ภาพที่ 2.17 ระบบการควบคุมการแสดงรูปภาพทางการแพทย์ด้วยท่าทางของผู้ใช้
ที่มา: (www.researchgate.net/.../224255259_Controller-free_exploration_of_medical_image_data_Experiencing_the_Kinect)

2.6.8 Interactive Display Using Depth and RGB Sensors for Face and Gesture Control

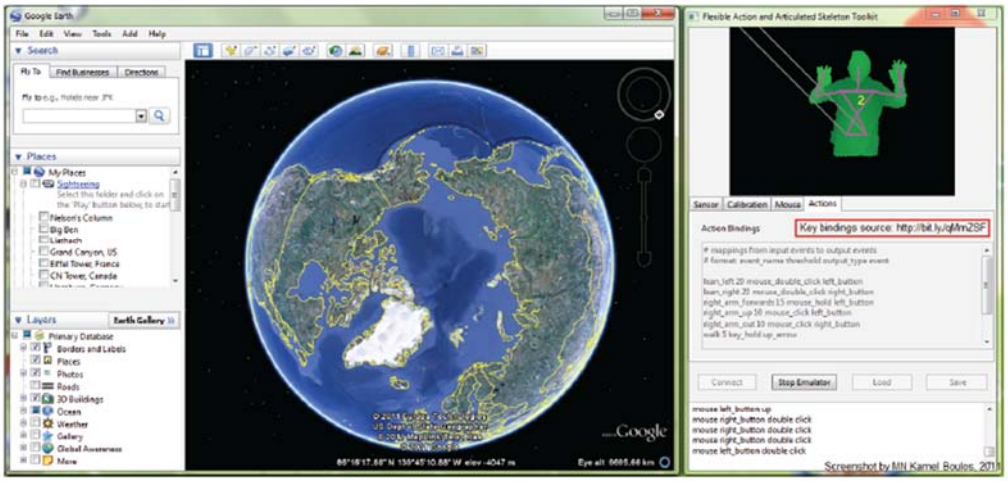
งานวิจัยของ C. Bellmore, R. Ptucha, และ A. Savakis (2011)ได้เสนอระบบแสดงผลแบบโต้ตอบ (Interactive Display System) ที่ควบคุมโดยการแสดงท่าทางของผู้ใช้ และการแสดงออกทางหน้าตาของผู้ใช้ ตามที่แสดงในภาพที่ 2.18 งานวิจัยนี้ใช้กล้อง Kinect ในการตรวจจับการเคลื่อนไหวของผู้ใช้ และใช้ภาพสีที่กล้องตรวจจับได้นำไปวิเคราะห์คำสั่งที่มาจากการแสดงออกทางหน้าตา สำหรับการควบคุมโดยใช้การแสดงท่าทาง จะนำค่าตำแหน่งของข้อต่อที่ได้รับจาก โปรแกรมที่อยู่ในรูปแบบ จุดพิกัดสามมิติไปเปลี่ยนให้เป็นจุดพิกัดสองมิติของรูปภาพก่อนนำค่าตำแหน่งของข้อต่อที่อยู่ในรูปแบบสองมิติไปวิเคราะห์คำสั่งของผู้ใช้ งานวิจัยนี้ทำการทดสอบวิธีการควบคุมระบบแสดงผล ที่นำเสนอด้วยระบบร้านค้ากับผู้ใช้จำนวนมากกว่า 100 คน ผลการศึกษาของงานวิจัยนี้แสดงให้เห็นถึงแนวทางที่เป็นไปได้ในการปฏิสัมพันธ์กับระบบด้วยร่างกายมนุษย์



ภาพที่ 2.18 ระบบแสดงผลที่ควบคุมโดยการแสดงท่าทางและการแสดงออกทางหน้าตาของผู้ใช้
ที่มา: (http://www.deepdyve.com/lp/institute-of-electrical-and-electronics-engineers/interactive-display-using-depth-and-rgb-sensors-for-face-and-gesture-7fV0UGTQ37)

2.6.9 Web GIS in practice X a Microsoft Kinect natural

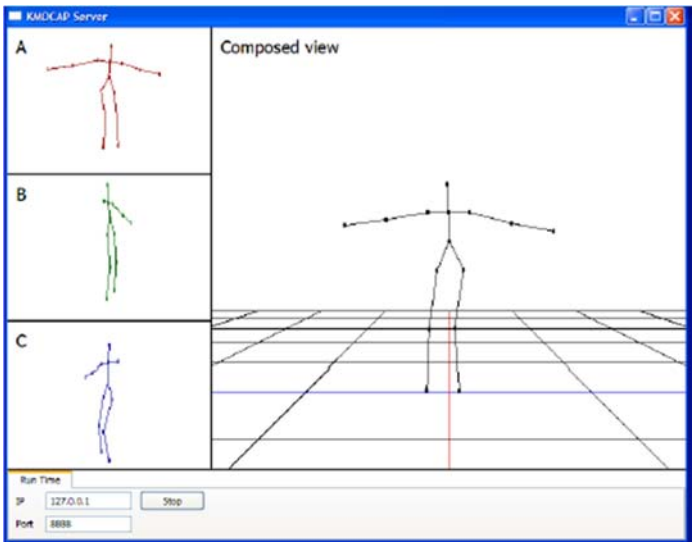
งานวิจัยของ Maged N Kamel Boulos, Bryan J Blanchard, Cory Walker, Julio Montero, Aalap Tripathy andRicardo และ Gutierrez-Osuna (2011) ได้ทำการศึกษาการใช้ Kinect ในการใช้ท่าทางและการรู้จำเสียงในการควบคุมส่วนติดต่อกับผู้ใช้บนโปรแกรมกูเกิ้ลเอิร์ธ โดยผลการศึกษาสามารถควบคุมได้ทั้งเสียงและท่าทางผลการศึกษาของงานวิจัยนี้แสดงให้เห็นว่าสามารถนำวิธีการรู้จำเสียงและท่าทางมาใช้ในการใช้เสียงตรวจจับคำพูดได้



ภาพที่ 2.19 ระบบแสดงผลที่ควบคุมโปรแกรมกูเกิ้ลเอิร์ธโดยใช้ท่าทางและเสียง
ที่มา: (http://www.ij-healthgeographics.com/content/10/1/45)

2.6.10 การประกอบรวมโครงร่างมนุษย์จากการตรวจจับการเคลื่อนไหวโดยใช้กล้อง Kinect หลายตัว

งานวิจัยของ นราวุฒิ พัฒโนทัย, พรชัย มงคลนาม และ บัณฑิต วรรธนาภา(2012)ได้ทำการศึกษาการใช้ Kinect สามตัวต่อเข้าด้วยกันในระบบ Client – Server โดยวางมุมกล้องทำมุม 120 องศาต่อกัน การศึกษาของงานวิจัยนี้แสดงให้เห็นว่าสามารถเพื่อช่วยเพิ่มโอกาสในการมองเห็นจุด Skeleton ของคนในการจับการเคลื่อนไหวได้ดีขึ้น



ภาพที่ 2.20 โครงร่างมนุษย์ที่กล้องแต่ละตัวตรวจจับได้ และโครงร่างมนุษย์ประกอบรวม
ที่มา: (www2.sit.kmutt.ac.th/.../1%20Narawut_NCIT2012%20motionMovement.pdf)

2.6.11 รีวิว Samsung Smart TV 55ES8000

ในปี 2012 บริษัท Samsung Electronics (2012)ได้นำเทคโนโลยีการรู้จำเสียงและการตรวจสอบการเคลื่อนไหวใส่กับอินเตอร์เน็ตทีวี หลักการก็คล้ายๆกับ Kinect ที่กล้องจะจับท่าทางและตัวทีวีก็คอยจะตอบสนองคำสั่งตามท่าทางของเรา ตัวอย่างเช่น "การแบมือ" โชว์หน้ากล้องก็จะเริ่มต้นคำสั่ง หลังจากนั้นตัว Cursor ลูกศรก็จะปรากฏขึ้นมาบนหน้าจอ เลื่อนมือไปในทิศทางต่างๆ เพื่อเลื่อนตัวลูกศรที่เป็น Cursor ไปในทิศทางที่เราต้องการ หากต้องการกดปุ่มใดๆบนหน้าจอ ก็กำมือ ซึ่งเปรียบเสมือนการคลิกเมาส์ การสั่งงานด้วยเสียง เพียงแค่พูดว่า "HI TV" เพื่อเปิดหรือปิดการใช้งาน Voice Control หลังจากนั้นจะมี "แถบชุดเมนูคำสั่ง" ขึ้นมาด้านล่าง

ชุดคำสั่ง Voice Control

1. "TV Power Off": ปิดทีวี
2. "Source": เปลี่ยนแหล่งสัญญาณ

3. "Channel Number”: เลือกช่องทีวีตามที่ต้องการ

4. "Channel Up หรือ Down”: เลื่อนช่องทีวีขึ้นหรือลงทีละช่อง

5. "Volume Up/Down”: เพิ่มระดับเสียงดัง/เบา

6. "Mute”: ปิดเสียง

ภาพที่ 2.21 ตัวอย่าง Smart TV รุ่น 55ES8000

ที่มา: (http://www.lcdtvthailand.com/review/detail.asp?desc=1&param_id=1277,2012)

ดังนั้นงานวิจัยนี้จึงได้เสนอและออกแบบและพัฒนาเทคนิคการผสมผสานเทคนิคระหว่างการรู้จำเสียงและเทคนิคการตรวจสอบการเคลื่อนไหวเข้าด้วยกันเพื่อใช้ในการควบคุมอินเทอร์เน็ตทีวีโดยออกแบบการควบคุมไว้ดังนี้ เปลี่ยนช่อง 3,5,7,9,11,thai PBS เพิ่มลดเสียงได้ 5 ระดับและการเปิดปิด TV