Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Πανεπιστήμιο Θεσσαλίας

Διπλωματική Εργασία

Αλληλεπίδραση ανθρώπου – ρομπότ με χρήση του αισθητήρα Kinect

Human-Robot interaction using the Kinect sensor

Συγγραφέας:
Νικολάου Δημήτριος

Επιβλέποντες:
Ποταμιάνος Γεράσιμος
Μπέλλας Νικόλαος

Πανεπιστήμιο Θεσσαλίας

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Φεβρουάριος 2017

Περίληψη

Η δυνατότητα αναγνώρισης φωνής για αλληλεπίδραση με υπολογιστές έχει βρει εφαρμογές εδώ και πολλά χρόνια. Πλέον έχει αρχίσει και η εξάπλωση συστημάτων όρασης υπολογιστών με τον ίδιο σκοπό. Σε αυτήν την εργασία μελετάμε τον συνδυασμό των δύο παραπάνω μεθόδων ώστε να αναπτύξουμε ένα πολυτροπικό (multimodal) σύστημα χειρισμού ενός ρομποτικού οχήματος.

Το κομμάτι της φωνητικής αναγνώρισης υλοποιήθηκε στην πλατφόρμα IrisTK, η οποία αποτελεί ένα εξειδικευμένο πλαίσιο ανάπτυξης συστημάτων αλληλεπίδρασης με υπολογιστές. Για την αναγνώριση χειρονομιών υλοποιήθηκε πρόγραμμα σε γλώσσα C++, το οποίο δέχεται είσοδο από τον αισθητήρα Kinect v2.

Φωνητικές εντολές και χειρονομίες χρησιμοποιούνται ως μέθοδοι απομακρυσμένου χειρισμού ενός ρομποτικού οχήματος, ελεγχόμενο από μητρική Raspberry Pi 3 Model B, ενώ η επικοινωνία μεταξύ του ρομπότ και των εφαρμογών ελέγχου γίνεται μέσω ασυρμάτου δικτύου με network sockets. Το ρομπότ επιστρέφει στον χρήστη ζωντανή ροή βίντεο από μια κάμερα που βρίσκεται τοποθετημένη πάνω του. Δίνεται ιδιαίτερη έμφαση στην εκπαίδευση νέων χειρονομιών και στην εισαγωγή τους στο σύστημα.

Abstract

Voice recognition in human – computer interaction has been in use for many years already. Recently we have also seen computer vision systems in support of the same goal. In the present thesis, we focus on the combination of techniques from these two fields in order to implement a multimodal control system for a robot car.

The voice recognition part was developed on IrisTK, a specialized framework for creating multimodal interaction systems. For the gesture recognition part, a C++ program was developed to analyze input from the Kinect V2 sensor.

Voice commands and hand gestures are used as remote control methods for a robot car, which is controlled by a Raspberry Pi 3 Model B motherboard. The wireless connection between the robot and the control applications consists of two network sockets. The robot captures and sends a live video feed using a mounted camera. Emphasis is given upon the training and integration of new gestures.

Περιεχόμενα

Κατάλογος Σχημάτων

Σχήμα 5.8: Ταυτόχρονη αναγνώριση εντολής «Στοπ» και «Τερματισμός».
Σχήμα 5.9: Σύγχυση μεταξύ εντολής «Μπροστά» και «Στοπ».
Σχήμα 5.10: Ο αριθμός των frames μέχρι την αναγνώριση. 9 frames αντιστοιχούν σε 297ms.
Σχήμα 5.11: Η μέση καθυστέρησης εκτέλεσης εντολής.

Κατάλογος Πινάκων

1 Εισαγωγή

Οι υπολογιστές είναι πλέον ένα αναπόσπαστο κομμάτι της ζωής του σύγχρονου ανθρώπου. Βελτιώνουν και απλουστεύουν την καθημερινότητά μας ενώ ταυτόχρονα αποτελούν απαραίτητα εργαλεία και για πιο εξειδικευμένες εργασίες. Τα τελευταία χρόνια, χάρις στην ταχύτατη τεχνολογική πρόοδο, δίνεται η δυνατότητα εξερεύνησης νέων τρόπων αλληλεπίδρασης με τους υπολογιστές. Ο τομέας της Επικοινωνίας Ανθρώπου – Μηχανής μελετά την ανάπτυξη, αξιολόγηση και εφαρμογή μεθόδων αλληλεπίδρασης μεταξύ ανθρώπων και υπολογιστικών συστημάτων [1]. Για τον άνθρωπο, η ομιλία και η κίνηση συνιστούν τους πιο φυσικούς και διαισθητικούς τρόπους επικοινωνίας [2].

1.1 Αναγνώριση ομιλίας από υπολογιστές

Η αναγνώριση ομιλίας από υπολογιστές επιτρέπει την μετάφραση της ομιλούμενης γλώσσας σε κείμενο. Βασίζεται σε διάφορους επιστημονικούς τομείς, όπως η επεξεργασία σημάτων, η αναγνώριση προτύπων και η γλωσσολογία [3].

Οι πρώτες προσπάθειες αναγνώρισης ομιλίας ξεκίνησαν την δεκαετία του 1930 από την Bell Labs [4]. Δύο δεκαετίες αργότερα είχαν ήδη αναπτυχθεί συστήματα αναγνώρισης τα οποία μπορούσαν να ξεχωρίσουν ένα περιορισμένο λεξιλόγιο περίπου δέκα λέξεων [5]. Την δεκαετία του 60’ Σοβιετικοί ερευνητές εφηύραν τον αλγόριθμο “Dynamic time warping” (DTW), ο οποίος έδωσε την δυνατότητα αναγνώρισης συνεχούς ομιλίας με λεξιλόγιο 200 λέξεων [6]. Αργότερα μπήκε στο παιχνίδι η IBM, η οποία υλοποίησε μία γραφομηχανή φωνητικής λειτουργίας που μπορούσε να διαχειριστεί λεξικό 20.000 λέξεων [7]. Στα μέσα της δεκαετίας του 1990 τα συστήματα αναγνώρισης ομιλίας είχαν εξελιχθεί αρκετά ώστε να διατεθούν στην αγορά και να γνωρίσουν εμπορική επιτυχία [8].

Στη σύχρονη εποχή η αναγνώριση ομιλίας είναι πλήρως διαδεδομένη. Τα μοντέρνα λειτουργικά συστήματα υπολογιστών και κινητών παρέχουν έξυπνους εικονικούς βοηθούς (Apple’s Siri, Microsoft’s Cortana, Amazon Alexa, Google Now), οι οποίοι αναγνωρίζουν και εκτελούν ομιλούμενες εντολές. Πέρα από την καθημερινή χρήση, υπηρεσίες πληροφοριών, εφαρμόζουν τεχνικές ανάλυσης ηχογραφημένων ομιλιών για την εύρεση λέξεων-κλειδιών (keyword spotting) [9].

1.2 Ανίχνευση κίνησης και χειρονομιών από υπολογιστές

Η ανίχνευση κίνησης αφορά τον εντοπισμό αλλαγής θέσης ενός αντικειμένου σε σχέση με το περιβάλλον του. Η αναγνώριση χειρονομιών είναι η διαδικασία της μαθηματικής ερμηνείας ανθρώπινων κινήσεων από μια υπολογιστική μηχανή [10]. Οι χειρονομίες μπορούν να προέλθουν από κινήσεις ή καταστάσεις οποιουδήποτε σημείου του σώματος αλλά συνήθως αφορούν κινήσεις των χεριών.

Επιτρέπουν τον χειρισμό του υπολογιστή χρησιμοποιώντας μόνο τα χέρια, δηλαδή χωρίς τη βοήθεια συσκευών όπως το πληκτρολόγιο, το ποντίκι, οθόνες αφής και άλλα χειριστήρια. Έτσι κάνει πιο φυσική την αλληλεπίδραση του ανθρώπου με την μηχανή.

Για την ανίχνευση και τον προσδιορισμό χειρονομιών είναι απαραίτητη η χρήση κάποιου εργαλείου εισόδου. Οι κυριότερες μέθοδοι που έχουν αναπτυχθεί μέχρι σήμερα χρησιμοποιούν εργαλεία όπως:

- Γάντια (Wired gloves): με ενσωματωμένα επιταχυνσιόμετρα ή συσκευές μαγνητικού προσανατολισμού. Μπορούν να ανιχνεύσουν την κίνηση και την περιστροφή του χεριού, την κάμψη των δακτύλων με αρκετά μεγάλη ακρίβεια [11].
- Κάμερες βάθους: εξειδικευμένες κάμερες που χρησιμοποιούν τεχνικές δομημένου φωτός (structured light) ή time-of-flight (χρόνος κάλυψης απόστασης από το φως). Μπορούν να εξάγουν χάρτη βάθους (depth map) και να δημιουργήσουν μια προσεγγιστική 3-d αναπαράσταση του σκηνικού. Είναι αρκετά αποτελεσματικές στην αναγνώριση χειρονομιών [12].

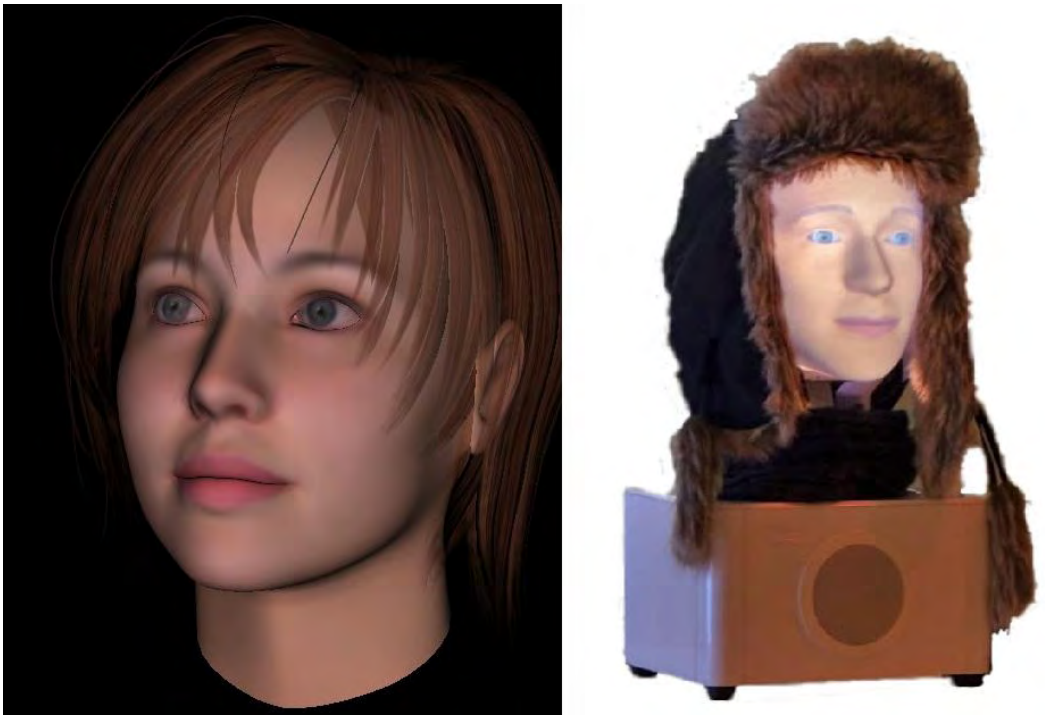
Σήμερα, η αναγνώριση και καταγραφή κινήσεων χρησιμοποιείται ευρέως στον τομέα της ψυχαγωγίας. Οι σύγχρονες παιχνιδομηχανές προσφέρουν ειδικά συστήματα ελέγχου κίνησης (Microsoft Kinect, Playstation Move, Nintendo Wii) [13]. Επίσης οι μεγάλες κινηματογραφικές παραγωγές χρησιμοποιούν συστήματα καταγραφής κινήσεων (motion-capture) για να μεταφέρουν τις φυσικές κινήσεις των ηθοποιών σε ψηφιακά κατασκευασμένους χαρακτήρες της οθόνης [14].

Πέρα απ’ την ψυχαγωγία, γίνονται προσπάθειες εφαρμογής της αναγνώρισης χειρονομιών και σε άλλους τομείς, όπως για παράδειγμα η ιατρική. Ερευνητές

του πανεπιστήμιου Ben Gurion του Ισραήλ ανέπτυξαν ένα σύστημα που δίνει στους γιατρούς τη δυνατότητα εν ώρα εγχείρισης να μελετάνε εικόνες ή ακτινογραφίες του ασθενή με χειρονομίες, χωρίς να αγγίζουν την οθόνη ή το πληκτρολόγιο [15].

1.3 Πολυτροπική (multimodal) αλληλεπίδραση

Τα πολυτροπικά συστήματα αλληλεπίδρασης δέχονται είσοδο και παράγουν έξοδο με παραπάνω από έναν τρόπους. Ένα τέτοιο σύστημα, το οποίο χρησιμοποιήθηκε και στην παρούσα εργασία, αποτελεί το IrisTK. Ουσιαστικά αποτελεί έναν Java-based σκελετό (framework) για την ανάπτυξη πολυτροπικών συστημάτων και προσφέρει δυνατότητες εισόδου με αναγνώριση ομιλίας, θέσης και προσώπου μέσω Kinect, και εξόδου μέσω ενός 3d ψηφιακού ή φυσικού προσώπου (Σχήμα 1.1) [16].



Σχήμα 1.1: Αριστερά: Το ψηφιακό πρόσωπο για αλληλεπίδραση με τον χρήστη. Δεξιά: Το ρομποτικό πρόσωπο (από [16]).

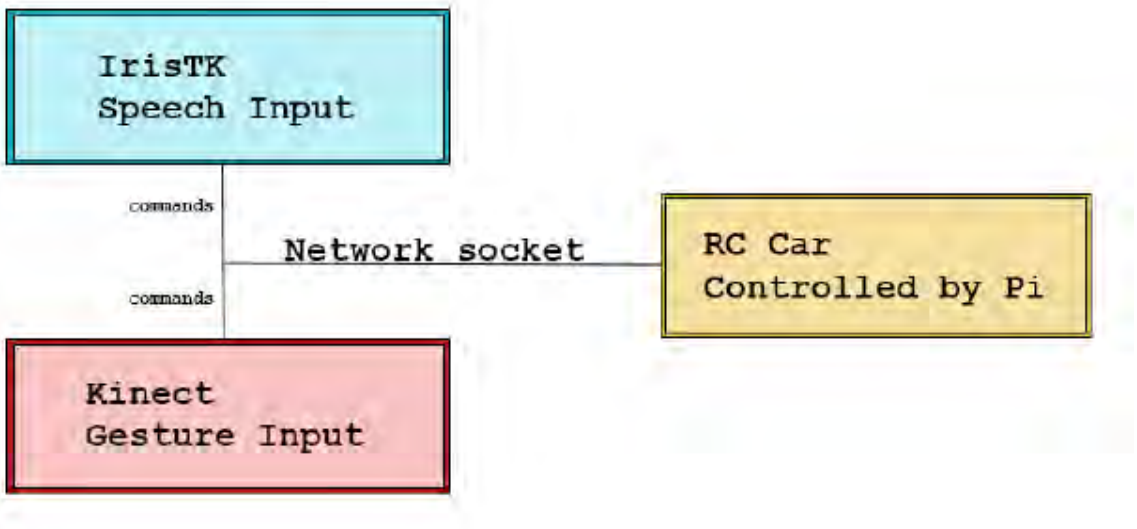
1.4 Η παρούσα υλοποίηση

Η παρούσα διπλωματική εργασία αποτελεί την υλοποίηση ενός συστήματος απομακρυσμένου χειρισμού οχήματος-ρομπότ μέσω φωνητικών εντολών και χειρονομιών.

Χρησιμοποιήθηκε το IrisTK framework για τον προγραμματισμό της λειτουργίας φωνητικών εντολών και ο αισθητήρας Kinect v2 για τον προγραμματισμό κίνησης μέσω χειρονομιών.

Οι ενέργειες που υποστηρίζονται από τη λειτουργία φωνητικών εντολών είναι οι εντολές κίνησης (εμπρόσθια κίνηση, οπισθοδρόμιση, αριστερή και δεξιά στροφή, σταμάτημα), εντολή φωτογραφίας η οποία αποτυπώνει μια φωτογραφία από την κάμερα που βρίσκεται πάνω στο όχημα, και τέλος εντολή για δημιουργία κάτοψης που παρουσιάζει όλα τα εμπόδια που εμφανίστηκαν στις φωτογραφίες που τραβήχθηκαν απ' το ρομπότ. Οι ενέργειες που υποστηρίζονται από τη λειτουργία χειρονομιών είναι μόνο οι εντολές κίνησης. Τέλος, το ρομπότ επιστρέφει ροή βίντεο στον χρήστη μέσω της κάμερας του.

Το όχημα ελέγχεται από ένα Raspberry Pi 3 το οποίο κινεί το όχημα ανάλογα με την εντολή που λαμβάνει κάθε φορά. Οι εντολές αποστέλονται μέσω network sockets. Ένα συνοπτικό διάγραμμα του συστήματος παρουσιάζεται στο σχήμα 1.2.



Σχήμα 1.2: Διάγραμμα της υλοποίησης.

1.5 Περιεχόμενο και δομή της εργασίας

1.5.1 Συνεισφορά της διπλωματικής

Η παρούσα διπλωματική ασχολείται με τον προγραμματισμό ενός real-time συστήματος με είσοδο φωνητικές εντολές, χειρονομίες ή τον συνδυασμό τους και έξοδο την κίνηση ενός ρομπότ-οχήματος σε γνωστό ή άγνωστο χώρο. Το ρομπότ δεν είναι απαραίτητο να βρίσκεται στο οπτικό πεδίο του χρήστη καθώς παρέχει ζωντανή ροή βίντεο.

Ερευνάται ακόμα η χρονική απόκριση του κάθε τύπου εισόδου, το ποσοστό επιτυχίας αναγνώρισης τόσο των φωνητικών εντολών όσο και των χειρονομιών.

Ένα άλλο αντικείμενο μελέτης της διπλωματικής είναι η δημιουργία και εισαγωγή νέων χειρονομιών στο σύνολο αναγνώρισης του Kinect, το οποίο πραγματοποιήθηκε σε περιβάλλον προγραμματισμού C++.

Τέλος, σημειώνεται ότι στην παρούσα εργασία δίνεται μεγαλύτερη έμφαση στην υλοποίηση αναγνώρισης χειρονομιών παρά σε αυτήν των φωνητικών εντολών, καθότι η τελευταία αποτελεί αντικείμενο μελέτης συναδέλφου [17].

1.5.2 Οργάνωση περιεχομένου της διπλωματικής

Το περιεχόμενο της διπλωματικής είναι οργανωμένο σε πέντε συνολικά κεφάλαια. Πέραν της εισαγωγής, τα υπόλοιπα τέσσερα είναι τα παρακάτω:

- **Κεφάλαιο 2:** Στο κεφάλαιο αυτό αναλύεται η αρχιτεκτονική του συστήματος, τα μηχανικά μέρη του ρομπότ και η συνδεσμολογία των διάφορων εξαρτημάτων.

- **Κεφάλαιο 3:** Σε αυτό το κεφάλαιο αναλύεται η συμπεριφορά του συστήματος με είσοδο χειρονομίες. Παρουσιάζεται ο αισθητήρας Kinect v2 καθώς και το σύνολο των αποδεκτών χειρονομιών.

- **Κεφάλαιο 4:** Στο κεφάλαιο αυτό γίνεται περιγραφή της λειτουργίας του ρομπότ με είσοδο αποκλειστικά φωνητικές εντολές. Παρουσιάζεται ο τρόπος λειτουργίας του IrisTK, το λεξικό εντολών και οι γραμματικές που χρησιμοποιήθηκαν.

- **Κεφάλαιο 5:** Σε αυτό το κεφάλαιο γίνεται εκτενής ανάλυση της διαδικασίας εκπαίδευσης νέων χειρονομιών και της εισαγωγής τους στο σύστημα, των προβλημάτων που αντιμετωπίστηκαν και των πειραμάτων για την ακρίβεια αναγνώρισης και καθυστέρηση εκτέλεσης των εντολών.

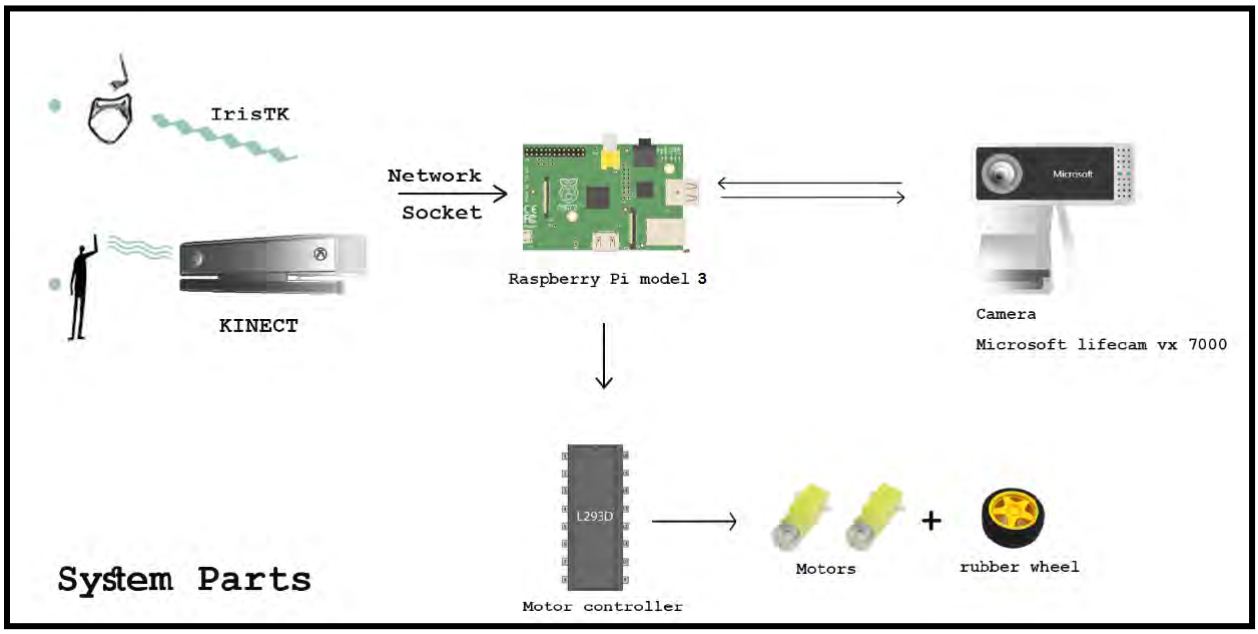
- **Κεφάλαιο 6:** Στο τελευταίο αυτό κεφάλαιο παρουσιάζονται συμπεράσματα σχετικά με την εργασία, τα αποτελέσματα της, η συνεισφορά της και πιθανή μελλοντική δουλειά με βάση αυτήν.

2 Αρχιτεκτονική Συστήματος

2.1 Όχημα ελεγχόμενο από ενσωματωμένο Raspberry Pi

Για την υλοποίηση του συστήματός μας κατασκευάσαμε το παρακάτω ρομπότ-όχημα. Το όχημα κινείται με τη βοήθεια δύο ροδών οι οποίες ελέγχονται από έναν motor controller.

Ο motor controller τροφοδοτείται από τέσσερις μπαταρίες τύπου AA και ελέγχεται μέσω των pins του Raspberry Pi. Τέλος, το raspberry pi ελέγχεται μέσω δικτύου χρησιμοποιώντας network socket, μέσω του οποίου αποστέλονται οι εντολές ελέγχου του ρομπότ είτε μέσω του IrisTK για φωνητική είσοδο, είτε μέσω προγράμματος C++ για είσοδο από το Kinect. Η αρχιτεκτονική του συστήματος φάινεται στο σχήμα 2.1.



Σχήμα 2.1: Σχεδιάγραμμα αρχιτεκτονικής συστήματος.

2.1.1 Μηχανικά μέρη του οχήματος

Το ρομπότ αποτελείται από τα εξής μέρη:

- Shadow chassis: εξωτερικό προστατευτικό περίβλημα
- 2 Rubber wheels: ροδάκια που δένουν πάνω στο εξωτερικό περίβλημα
- Breadboard Mini: μέσω του οποίου γίνεται η διασύνδεση
- Jumper Wires: για την καλωδίωση
- 2 DC Gear Motors με χαρακτηριστικά
    - Περιστροφές ανά λεπτο (RPM): 125r/min
    - Ένταση (current): 80 – 100 mA
    - Λόγος μετάδοσης: 48:1
    - Ροπή (output torque): 0.8kg/cm
    - Διαστάσεις: 70x22x18mm
    - Τάση (voltage): 3V
- L293D Motor Controller: ελεγκτής κίνησης των motors
- 4 μπαταρίες για την τροφοδοσία του Motor Controller

Για τη λήψη φωτογραφιών και βίντεο από το ρομπότ χρησιμοποιήθηκε μια web camera Microsoft VX7000. Στον πίνακα 2.1 φαίνονται τα κυριότερα χαρακτηριστικά της.

Microsoft LifeCam VX-7000	
Webcam Length	121 mm
Webcam Width	69 mm
Webcam Depth/Height	26 mm
Webcam Weight	132 grams
Webcam Cable Length	1.8 m
Imaging Features	
Sensor	CMOS 2.0 MP sensor technology
Video Resolution	2 megapixel (1600 x 1190 pixels)
Still Image Resolution	7.6 megapixel (3200 x 2380 pixels) interpolated
Field of View	71^{o} diagonal field of view
Imaging Features	Digital pan, digital tilt, 5x digital zoom Fixed Focus, Automatic image adjustment

Πίνακας 2.1: Κυριότερα τεχνικά χαρακτηριστικά της LifeCam VX-7000.

2.1.2 Raspberry Pi 3 Model B

Το Raspberry Pi 3 αποτελεί τον «εγκέφαλο» του ρομπότ. Είναι υπεύθυνο για την επικοινωνία μέσω δικτύου με το IrisTK και την εφαρμογή αναγνώρισης χειρονομιών, καθώς και για τον συντονισμό της κίνησης των τροχών.

Σχήμα 2.2: Raspberry Pi 3 Model B.

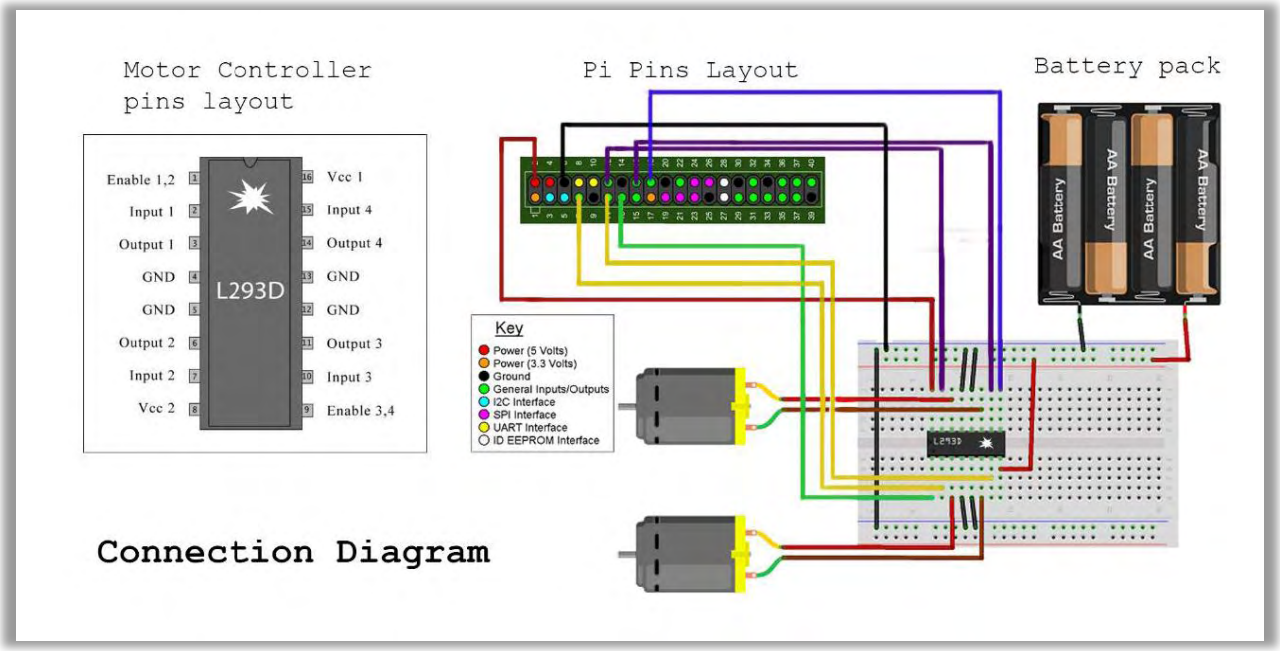
Τα τεχνικά χαρακτηριστικά του είναι:

- Broadcom BCM2837 64bit ARMv7 Quad Core Processor @1.2GHz
- 1GB RAM
- BCM43134 Wi-Fi on board
- Bluetooth Low Energy (BLE) on board
- 40pin extended GPIO
- 4 x USB 2 ports
- 4 pole Stereo output and Composite Video Port
- Full size HDMI
- CSI camera port for connecting the Raspberry Pi camera
- Micro SD port for loading your operating system and storing data
- Micro USB power source (supports up to 2.4 Amps)

Το Raspberry Pi γενικά δεν ανήκει στην κατηγορία μικροελεγκτών κλασσικού τύπου (single board microcontrollers) που χρησιμοποιούνται σε παρόμοια projects, (όπως το Arduino) λόγω της μεγάλης σχετικά κατανάλωσης που έχει, η οποία δεν το κάνει τόσο αυτόνομο όσο τα άλλα. Η υπολογιστική του δύναμη όμως είναι πολλές φορές μεγαλύτερη και έχει την δυνατότητα να υποστηρίξει πλήρη λειτουργικά όπως τα Windows (Internet of Things Edition) ή κάποιες διανομές των Linux (Ubuntu Raspbian Jessie - βασισμένο στο Debian, Ubuntu Mate, Ubuntu Snappy) [18]. Προτιμήθηκε η διανομή Raspbian η οποία περιέχει προεγκατεστημένα αρκετά εργαλεία απαραίτητα για το project.

Τέλος, το Pi τροφοδοτείται από ένα Power Bank ZALMAN ZM-PB84IW (8400MAH).

Στο σχήμα 2.3 φαίνονται τα pin layouts του Pi και του motor controller, καθώς και η συγκεκριμένη συνδεσμολογία που χρησιμοποιήθηκε.



Σχήμα 2.3: Συνδεσμολογία Pi – motors.

Παρατειρείται ότι κάθε motor απαιτεί τρία σήματα. Το ένα είναι σήμα ελέγχου, αν είναι Low το motor δεν λειτουργεί, ενώ αν είναι High, το motor περιστρέφεται προς τη φορά που του προσδίδουν τα άλλα δύο σήματα, το ένα για την φορά του ρολογιού και το άλλο για την αντίστροφη φορά. Στον πίνακα 2.2 περιέχει τις εισόδους και τις εξόδους των motors.

Motor1E	Motor2E	Motor1A	Motor2B	Motor2A	Motor2B	Κίνηση
0	0	X	X	X	X	Stop
0	1	X	X	X	X	DF
1	0	X	X	X	X	DF
1	1	1	0	1	0	Forward
		0	1	0	1	Backward
		1	0	0	1	Right
		0	1	1	0	Left
		0	0	0	0	DF
		0	0	0	1	
		0	0	1	0	
		0	0	1	1	
		0	1	0	0	
		0	1	1	1	
		1	0	0	0	
		1	0	1	1	
		1	1	0	0	
		1	1	0	1	
		1	1	1	0	
		1	1	1	1	

Πίνακας 2.2: Οι συνδυασμοί των σημάτων.

Ο παραπάνω πίνακας δίχνει όλους τους δυνατούς συνδυασμούς των σημάτων των pins του motor controller και σε ποια κίνηση του ρομπότ μεταφράζονται. DF σημαίνει dysfunction, αντίστοιχη κίνηση δεν υπάρχει και δεν προβλέπεται να έρθουν σε αυτήν την κατάσταση τα motors. Εάν ωστόσο γίνει κάτι τέτοιο, η κίνηση θα είναι ελλατωματική και απρόβλεπτη.

3 Λειτουργία με Είσοδο Χειρονομιών μέσω Kinect v2

3.1 Kinect v2

Το Kinect είναι μία συσκευή ανίχνευσης κίνησης και ήχου από την Microsoft. Η πρώτη έκδοση κυκλοφόρησε τον Νοέμβριο του 2010 για την κονσόλα παιχνιδιών Xbox 360, ενώ η τρέχουσα αναβαθμισμένη έκδοση v2 κυκλοφόρησε τον Νοέμβριο του 2013 για την κονσόλα Xbox One. Ο σκοπός του είναι να επιτρέψει στον χρήστη τον έλεγχο και την αλληλεπίδραση με την κονσόλα μέσω κινήσεων, χειρονομιών και φωνητικών εντολών, χωρίς την χρήση συμβατικού χειριστηρίου .

3.2 Προκλήσεις κατά την αναγνώριση ατόμων

Για να αποτελέσει το Kinect μια επιτυχημένη εναλλακτική λύση, αντί του χειριστηρίου, έπρεπε να ξεπεράσει τρία βασικά προβλήματα.

Το πρώτο έχει να κάνει με τον σαφή διαχωρισμό των παικτών που βρίσκονται στο οπτικό πεδίο της συσκευής. Άτομα σε διαφορετικές αποστάσεις από την κάμερα έχουν φαινομενικά διαφορετικό μέγεθος, γεγονός που μπορεί να προκαλέσει πρόβλημα στην αλληλεπίδραση τους. Το φαινομενικό μέγεθος είναι αντιστρόφως ανάλογο της απόστασης, συνεπώς μία καλή λύση θα ήταν η μέτρηση της απόστασης κάθε ατόμου από την συσκευή [19].

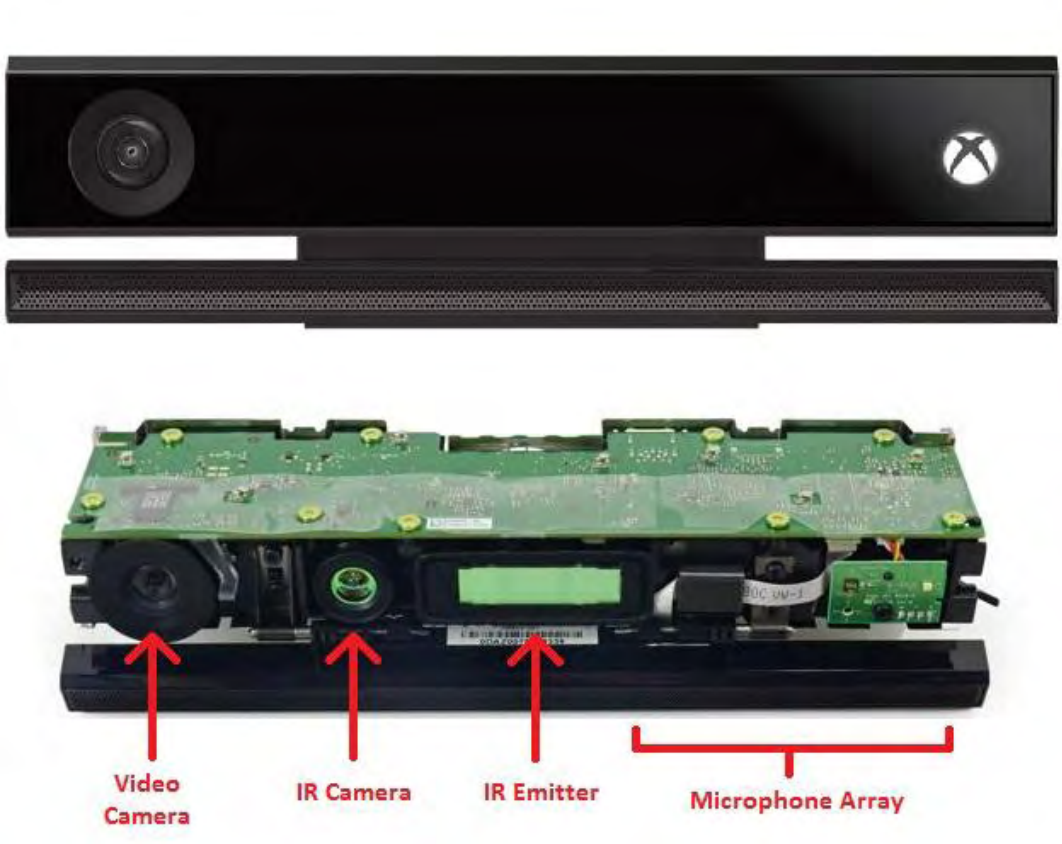
Το δεύτερο πρόβλημα σχετίζεται με την ταχύτητα ανταπόκρισης και την σταθερή λειτουργία του Kinect. Η εξαγωγή δεδομένων βάθους από ζωντανή ροή βίντεο είναι μια διαδικασία που απαιτεί αρκετούς μαθηματικούς υπολογισμούς. Η λύση για την άμεση και αδιάλειπτη λειτουργία ήταν η ενσωμάτωση ολοκληρωμένων κυκλωμάτων πάνω στη συσκευή, ώστε να εκτελούνται εκεί οι υπολογισμοί. Έτσι το Kinect καταφέρνει να παρέχει βίντεο βάθους με εντυπωσιακά μικρή καθυστέρηση, κάτω από 14ms [20].

Η τρίτη πρόκληση που έπρεπε να αντιμετωπιστεί αφορά τον φωτισμό του χώρου. Το Kinect θα έπρεπε να λειτουργεί χωρίς πρόβλημα σε οποιεσδήποτε συνθήκες φωτισμού, ακόμα και σε ακραίες περιπτώσεις όπως σε ένα σκοτεινό δωμάτιο ή σε ανομοιόμορφα φωτισμένους χώρους. Αυτό επετεύχθη με την χρήση υπέρυθρων ακτίνων και ενός ειδικού αισθητήρα υπερύθρων [20].

Έτσι η Microsoft κατέληξε σε μία συσκευή με δυνατότητα καταγραφής βίντεο βάθους χώρου, νυχτερινής λήψης αλλά και συνηθισμένου έγχρωμου βίντεο.

3.3 Τα μέρη του Kinect v2

Το Kinect v2 αποτελείται από δύο τμήματα. Την βάση, η οποία περιέχει μια συστοιχία τεσσάρων μικροφώνων και το άνω τμήμα που περιέχει δύο κάμερες.



Σχήμα 3.1: Πάνω: Η μπροστινή όψη του Kinect v2.
Κάτω: Οι κάμερες, τα μικρόφωνα και ο εκπομπός υπερύθρων ακτίνων.

Η πρώτη κάμερα είναι μία κάμερα έγχρωμου βίντεο ευρείας γωνίας με ανάλυση 1080p (1920 x 1080 pixels) και συχνότητα καταγραφής 30 fps (frames per second). Η δεύτερη κάμερα αποτελεί έναν αισθητήρα υπέρυθρων ακτίνων που αποδίδει βίντεο νυχτερινής όρασης (night vision) ή βίντεο βάθους χώρου (depth map) με ανάλυση 512 x 424 pixels και συχνότητα 30 fps [21].

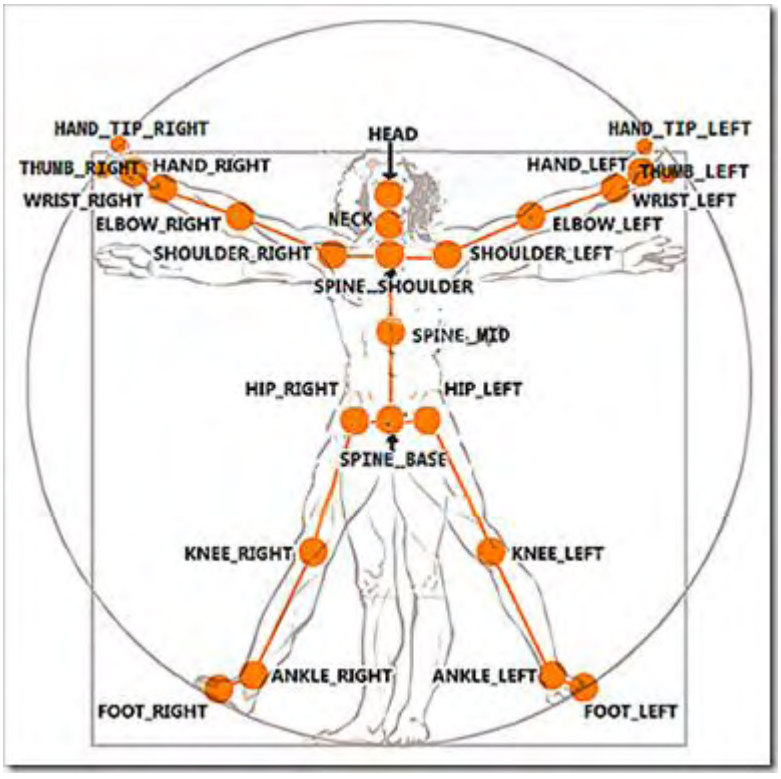
Το Kinect περιέχει επίσης κάποια ολοκληρωμένα κυκλώματα για την επί τόπου εξαγωγή δεδομένων βάθους απ' το βίντεο. Έτσι το depth map δημιουργείται πάνω στο Kinect, αντί να στέλνεται ένας μεγάλος όγκος ανεπεξέργαστων

δεδομένων σε κάποιον υπολογιστή, επιβαρύνοντας τον με επιπλέον υπολογισμούς [22].

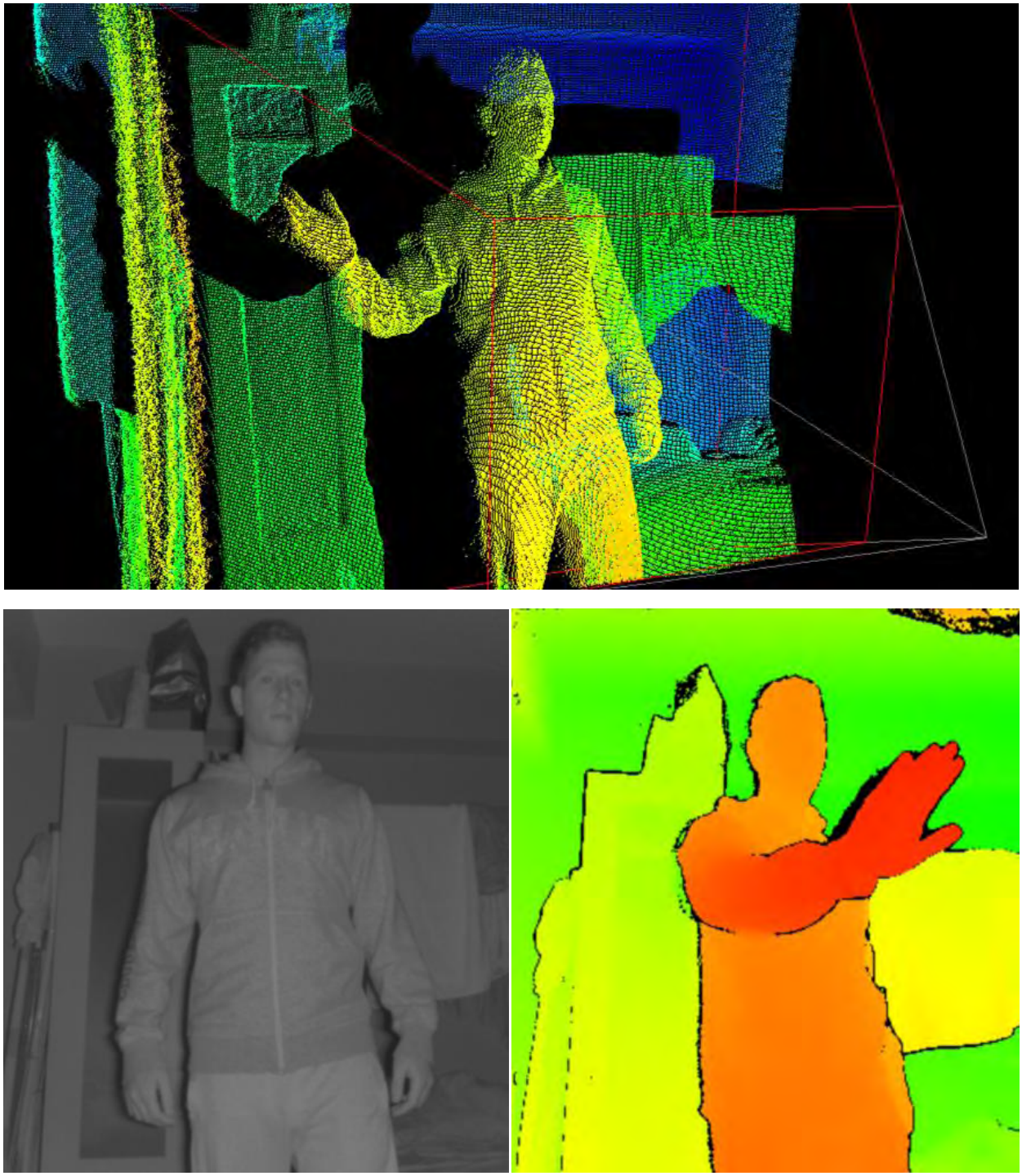
3.4 Δυνατότητες του Kinect v2

Το Kinect αποτελεί ένα ισχυρό εργαλείο όρασης υπολογιστών και αλληλεπίδρασης ανθρώπου-μηχανής. Οι δυνατότητες που παρέχει είναι:

- Ανίχνευση έως και 6 ατόμων ταυτόχρονα
- Κατασκευή σκελετού 25 αρθρώσεων για κάθε άτομο
- Κατασκευή χάρτη βάθους χώρου (depth map)
- Αναγνώριση προσώπου αλλά και εκφράσεων του προσώπου
- Αναγνώριση πολύπλοκων χειρονομιών
- Καταγραφή βίντεο νυχτερινής όρασης (night vision)



Σχήμα 3.2: Οι αρθρώσεις που αναγνωρίζονται και η ανακατασκευή του σκελετού (από [23]).

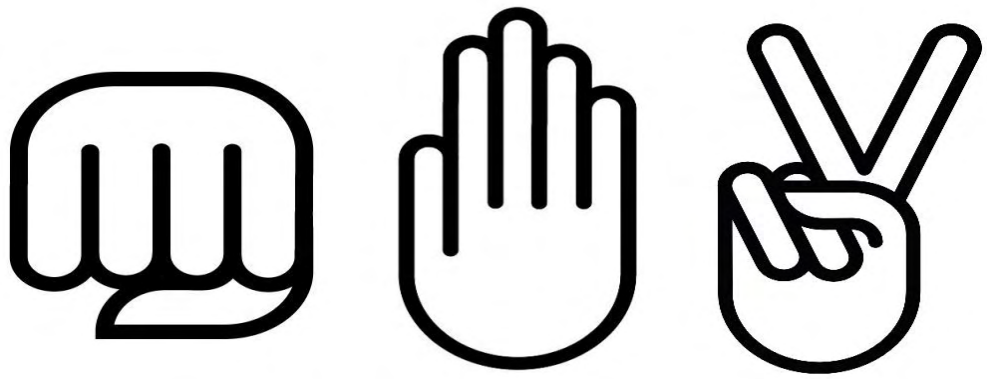


Σχήμα 3.3: Πάνω: Χάρτης βάθους χώρου (depth map). Κάτω αριστερά: Εικόνα νυχτερινής όρασης. Κάτω δεξιά: Εικόνα βάθους. Όσο πιο θερμά τα χρώματα, τόσο πιο μικρή η απόσταση απ' την κάμερα.

3.5 Αναγνώριση χειρονομιών

Στην παρούσα εργασία, ο χρήστης καθοδηγεί το ρομπότ μέσω χειρονομιών. Το Kinect μπορεί πολύ εύκολα να αναγνωρίσει τρεις προκαθορισμένες καταστάσεις χεριού:

- Κλειστή γροθιά
- Ανοιχτή παλάμη
- Lasso (κλειστή γροθιά με προέκταση του δείκτη και του μέσου)



Σχήμα 3.4: Οι τρεις καταστάσεις χεριού που αναγνωρίζονται απ' το Kinect. Από αριστερά προς τα δεξιά: Κλειστή, Ανοιχτή, Lasso (από [24]).

Οι τρεις παραπάνω καταστάσεις δεν αρκούν για την ανάπτυξη ενός συστήματος καθοδήγησης του ρομπότ. Κρίνεται απαραίτητο να εμπλουτίσουμε το σύστημα με κάποιες επιπλέον χειρονομίες, για να καλύψουμε περισσότερες από τρεις εντολές. Ευτυχώς το Kinect SDK (Software Development Kit) που παρέχει η Microsoft δίνει την δυνατότητα στους προγραμματιστές να δημιουργήσουν και να εισάγουν καινούργιες χειρονομίες στο πρόγραμμά τους.

3.5.1 Ευρετική (heuristic) αναγνώριση

Στην ευρετική αναγνώριση, ο προγραμματιστής γράφει απευθείας κώδικα για να περιγράψει μια νέα χειρονομία. Εάν για παράδειγμα θέλει να εισάγει μια κίνηση κατά την οποία ο χρήστης φέρνει την κλειστή γροθιά του πάνω απ' το κεφάλι του, τότε μπορεί πολύ απλά να ελεγχθεί αν **α)** το χέρι βρίσκεται ψηλότερα απ' το κεφάλι και **β)** αν το χέρι είναι στην κλειστή κατάσταση.

Η ευρετική προσέγγυση αποτελεί μία εύκολη και γρήγορη λύση για απλές περιπτώσεις σαν την παραπάνω. Όμως η ακριβής περιγραφή μιας

πολυπλοκότερης χειρονομίας είναι πολύ δυσκολότερη και πιθανόν να οδηγήσει σε αστοχίες αναγνώρισης.

3.5.2 Αναγνώριση με μηχανική μάθηση

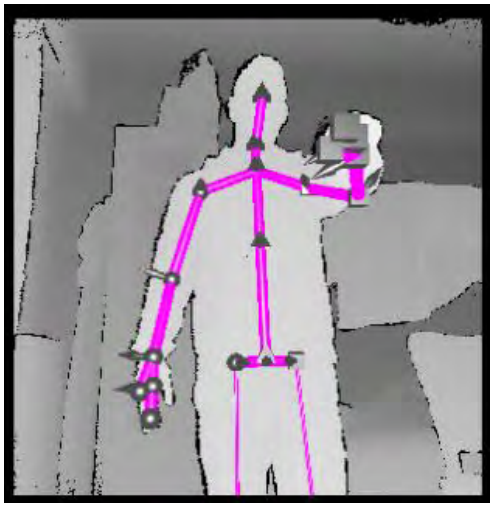
Σε πιο πολύπλοκες περιπτώσεις είναι προτιμότερο να εκπαιδεύσουμε το πρόγραμμα να αναγνωρίζει μία χειρονομία, παρά να την περιγράψουμε με κώδικα. Η Microsoft παρέχει το Gesture Builder, ένα ισχυρό εργαλείο για αυτόν ακριβώς τον σκοπό.

Η εκπαίδευση γίνεται τροφοδοτώντας το Gesture Builder με πολλά διαφορετικά clips της χειρονομίας με σημειωμένη την αρχή και το τέλος της. Εάν η εκπαίδευση δεν παράγει την επιθυμητή ακρίβεια αναγνώρισης, μπορούμε να προσθέσουμε και άλλα clips [25].

3.6 Σύνολο αποδεκτών χειρονομιών

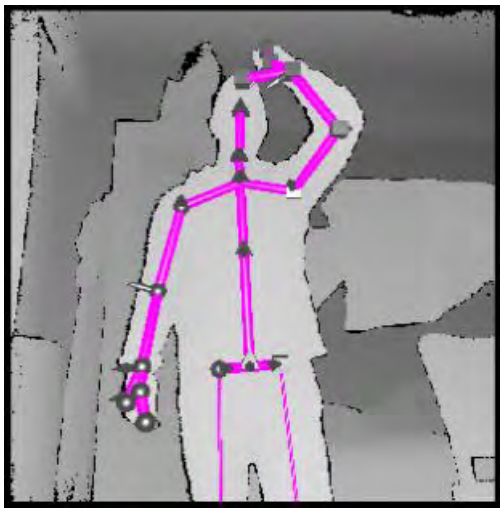
Με τη διαδικασία της μηχανικής μάθησης καταλήγουμε στην προσθήκη 6 χειρονομιών για την αλληλεπίδραση με το ρομπότ. Οι χειρονομίες και οι εντολές που αντιστοιχούν στην κάθε μία φαίνονται παρακάτω.

1. Εντολή «Μπροστά»: Ο χρήστης σηκώνει το δεξί του χέρι στο ύψος του ώμου, τεντωμένο προς τα εμπρός, με κλειστή γροθιά. Το ρομπότ μετακινείται προς τα εμπρός για όσο το χέρι βρίσκεται σε αυτή την στάση.



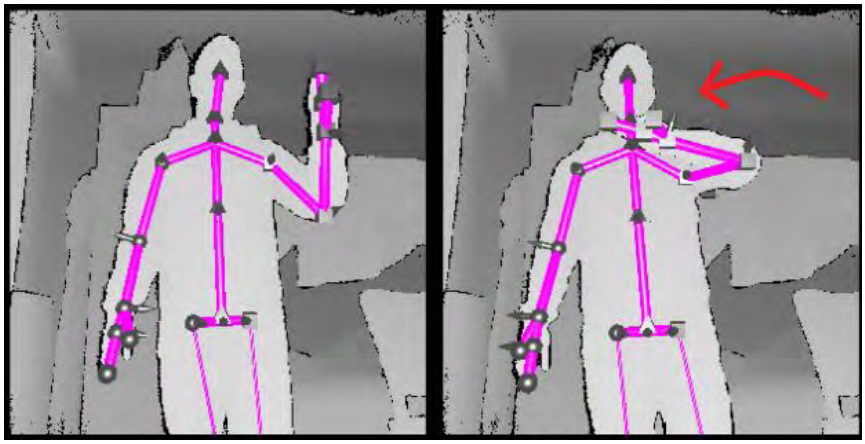
Σχήμα 3.5: Η χειρονομία που αντιστοιχεί στην εντολή «Μπροστά».

2. Εντολή «Πίσω»: Ο χρήστης σηκώνει το δεξί του χέρι ψηλά και προς τα πίσω, έτσι ώστε η κλειστή γροθιά του να βρίσκεται λίγο πάνω απ' το κεφάλι του. Το ρομπότ μετακινείται προς τα πίσω για όσο το χέρι βρίσκεται σε αυτή την στάση.



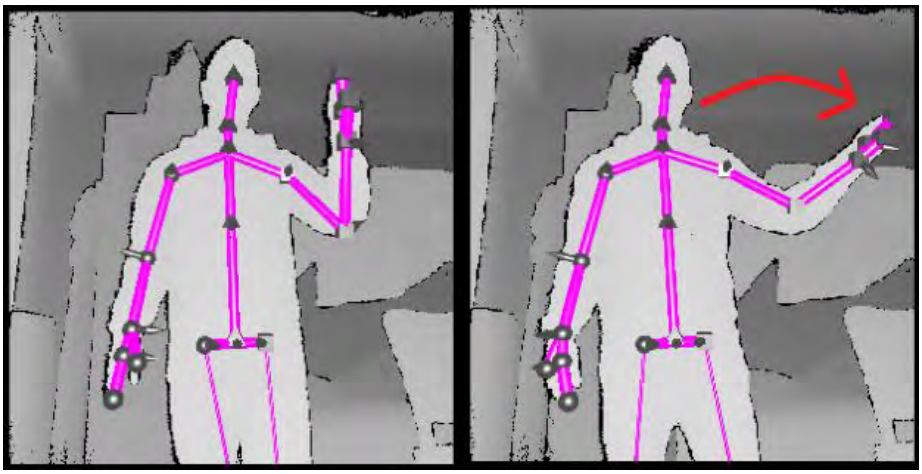
Σχήμα 3.6: Η χειρονομία που αντιστοιχεί στην εντολή «Πίσω».

3. Εντολή «Αριστερά»: Ο χρήστης σηκώνει το δεξί χέρι του στο ύψος του προσώπου (λυγισμένο) σχηματίζοντας την προκαθορισμένη χειρονομία lasso. Έπειτα γέρνει το χέρι του προς τα αριστερά. Το ρομπότ κάνει αριστερή περιστροφή γύρω απ' τον άξονά του για όσο το χέρι βρίσκεται λυγισμένο προς τα αριστερά.



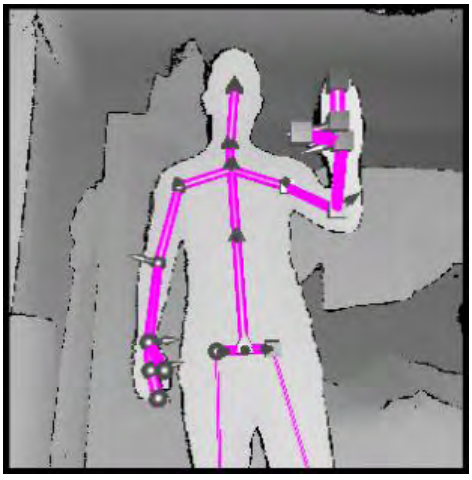
Σχήμα 3.7: Η χειρονομία που αντιστοιχεί στην εντολή «Αριστερά».

4. Εντολή «Δεξιά»: Ο χρήστης φέρνει το δεξί του χέρι στην ίδια αρχική θέση, όπως και στην εντολή 3. Όμως σε αυτή την περίπτωση γέρνει το χέρι του προς τα δεξιά. Το ρομπότ κάνει δεξιά περιστροφή γύρω απ' τον άξονά του για όσο το χέρι βρίσκεται λυγισμένο προς τα δεξιά.



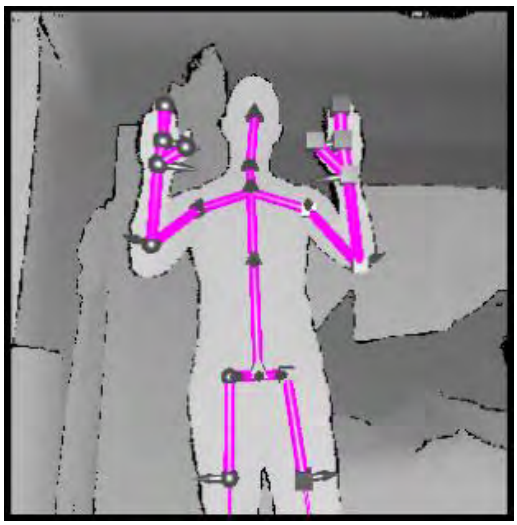
Σχήμα 3.8: Η χειρονομία που αντιστοιχεί στην εντολή «Δεξιά».

5. Εντολή «Στοπ»: Ο χρήστης σηκώνει το δεξί του χέρι έχοντας ανοιχτή την παλάμη του. Το ρομπότ μένει σταθερό. Εάν βρισκόταν σε κίνηση τότε σταματάει αμέσως.



Σχήμα 3.9: Η χειρονομία που αντιστοιχεί στην εντολή «Στοπ».

6. Εντολή «Τερματισμός»: Ο χρήστης σηκώνει και τα δύο χέρια στο ύψος του προσώπου έχοντας ανοιχτές παλάμες. Η χειρονομία αυτή σημαίνει το τέλος της αλληλεπίδρασης και τον τερματισμό του προγράμματος.



Σχήμα 3.10: Η χειρονομία που αντιστοιχεί στην εντολή «Τερματισμός».

4 Λειτουργία με Είσοδο Φωνητικές Εντολές

4.1 IrisTK Framework

Το IrisTK δημιουργήθηκε από τους Gabriel Skantze και Samer Al Moubayed το 2013. Σκοπός του ήταν να παρέχει ένα εργαλείο προγραμματισμού real-time συστημάτων που περιέχουν πρόσωπο με πρόσωπο αλληλεπίδραση ενός ή περισσότερων χρηστών με το σύστημα ή μεταξύ τους. Πέραν της κύριας χρήσης του, αποτελεί ένα πολύ γρήγορο εργαλείο προγραμματισμού συστημάτων που βασίζονται σε αναγνώριση φωνής, εξαγωγή χαρακτηριστικών προσώπου, χειρονομιών ή κίνησης σε συνδυασμό με επικοινωνία ανθρώπου μηχανής.

Το IrisTK Framework είναι προγραμματισμένο σε γλώσσα Java και χρησιμοποιεί XML αρχεία για τη γρήγορη ανάπτυξη συστημάτων αλληλεπίδρασης βασισμένα σε γεγονότα, μέσω της μετατροπής του XML αρχείου σε κώδικα Java. Δίνει τη δυνατότητα στον χρήστη να χρησιμοποιεί διάφορα πρωτόκολλα γραμματικών για να δημιουργήσει τις γλώσσες που θα αναγνωρίζονται [26].

4.1.2 Γραμματική SRGS

Το IrisTK χρησιμοποιεί μόνο γραμματικές χωρίς συμφραζόμενα για να δημιουργήσει τη γλώσσα που μπορεί να αναγνωρίσει από φωνητική είσοδο. Γραμματικές δηλαδή για τις οποίες ισχύει: G = (V, Σ, R, S) όπου:

- V είναι το σύνολο μη τερματικών συμβόλων
- Σ είναι το σύνολο των τερματικών συμβόλων
- R είναι το σύνολο κανόνων
- S είναι η αρχική κατάσταση

Η προεπιλεγμένη μορφή του IrisTK χρησιμοποιεί το πρωτόκολλο SRGS του οργανισμού W3C (world wide web consortium) το οποίο είναι σε μορφή XML και προτιμήθηκε στην παρούσα υλοποίηση. Εκτός του SRGS, υποστηρίζονται και άλλες μορφές όπως το ABNF format που δεν αποτυπώνεται σε XML, και άρα είναι πιο ευανάγνωστο. Για να συνθέσει ομιλία, το IrisTK χρησιμοποιεί τον κατάλογο UPS (universal phone set). Ο UPS περιέχει σύμβολα κατανοητά από υπολογιστή, τα οποία αντιστοιχίζονται σε φωνήματα βασισμένα στην προφορά κάποιων καθορισμένων λέξεων της γλώσσας.

UPS	SAMPA	IPA	English	UPS	SAMPA	IPA	English	UPS	SAMPA	IPA	English
P	p	p	put	G	g	g	gut	OI	OI	ɔ.ɪ	toy
B	b	b	big	NG	N	ŋ	sing	AI	aI	ɑ.ɪ	bite
M	m	m	mat	H	h	h	help	IYX	I@	i.ə	fear
F	f	f	fork	CH	tS	t.ʃ / tʃ	chin	EHX	e@	ɛ.ə	stairs
V	v	v	vat	JH	dZ	d.ʒ / dʒ	joy	UWX	U@	u.ə	lure
TH	T	θ	thin	I	i:	i	feel	OWX	@U	o.ə	boa
DH	D	ð	then	U	u:	u	too	AOX	U@	ɔ.ə	four
T	t	t	talk	IH	I	ɪ	fill	_s			SILENCE
D	d	d	dig	UH	U	ʊ	book				
N	n	n	no	O	@U	o	go				
DX	t	ɾ	butter (US)	AX	@	ə	ago				
S	s	s	sit	EH	e	ɛ	pet				
Z	z	z	zap	ER	3:	ɜ	bird (UK)				
L	l	l	lid	AH	{	ʌ	cut				
SH	S	ʃ	she	AO	Q	ɔ	dog				
ZH	Z	ʒ	pleasure	AE	{	æ	cat				
R	r	ɻ	red	AA	A:	ɑ	father				
J	j	j	yard	Q	Q	ɒ	hot				
W	w	w	with	EI	eI	e.ɪ	late				
K	k	k	cut	AU	aU	ɑ.ʊ	foul				

Πίνακας 4.1: UPS phone set (από [27]).

4.2 Περιγραφή λειτουργίας του συστήματος

Αρχικά το σύστημα περιμένει κάποια εντολή προς το ρομπότ. Αν η είσοδος που λάβει ανήκει στην αναγνωρίσιμη γλώσσα γίνεται αποδεκτή και θέτει την προβλεπόμενη τιμή σε μια ειδική μεταβλητή με αναγνωριστικό action. Διαφορετικά, ζητάει εκ νέου είσοδο. Το σχήμα 4.1 δείχνει τον κώδικα σε μορφή SRGS XML και στον πίνακα 4.2 φαίνονται όλες οι αποδεκτές προτάσεις.

```
<?xml version="1.0" encoding="utf-8"?>
<grammar xml:lang="en-US" version="1.0" root="root"
      xmlns="http://www.w3.org/2001/06/grammar"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.w3.org/2001/06/grammar
http://www.iristk.net/xml/srgs.xsd" tag-format="semantics/1.0">

  <rule id="root" scope="public">
      <one-of>
            <item>forward<tag>out.action=1</tag></item>
            <item>go forward<tag>out.action=1</tag></item>
            <item>go straight<tag>out.action=1</tag></item>
            <item>move forward<tag>out.action=1</tag></item>
            <item>move straight<tag>out.action=1</tag></item>

            <item>move backwards<tag>out.action=2</tag></item>
            <item>move back<tag>out.action=2</tag></item>
            <item>go backwards<tag>out.action=2</tag></item>
            <item>go back<tag>out.action=2</tag></item>
            <item>backwards<tag>out.action=2</tag></item>

            <item>left<tag>out.action=3</tag></item>
            <item>turn left<tag>out.action=3</tag></item>
            <item>go left<tag>out.action=3</tag></item>

            <item>stop<tag>out.action=5</tag></item>

            <item>right<tag>out.action=4</tag></item>
            <item>turn right<tag>out.action=4</tag></item>
            <item>go right<tag>out.action=4</tag></item>

            <item>exit<tag>out.action=0</tag></item>

            <item>take a photo<tag>out.action=10</tag></item>
            <item>take a picture<tag>out.action=10</tag></item>
            <item>take a photograph<tag>out.action=10</tag></item>
            <item>photo<tag>out.action=10</tag></item>
            <item>picture<tag>out.action=10</tag></item>
            <item>photograph<tag>out.action=10</tag></item>

            <item>yes<tag>out.yes=1</tag></item>
            <item>no<tag>out.no=1</tag></item>


      </one-of>
  </rule>

</grammar>
```

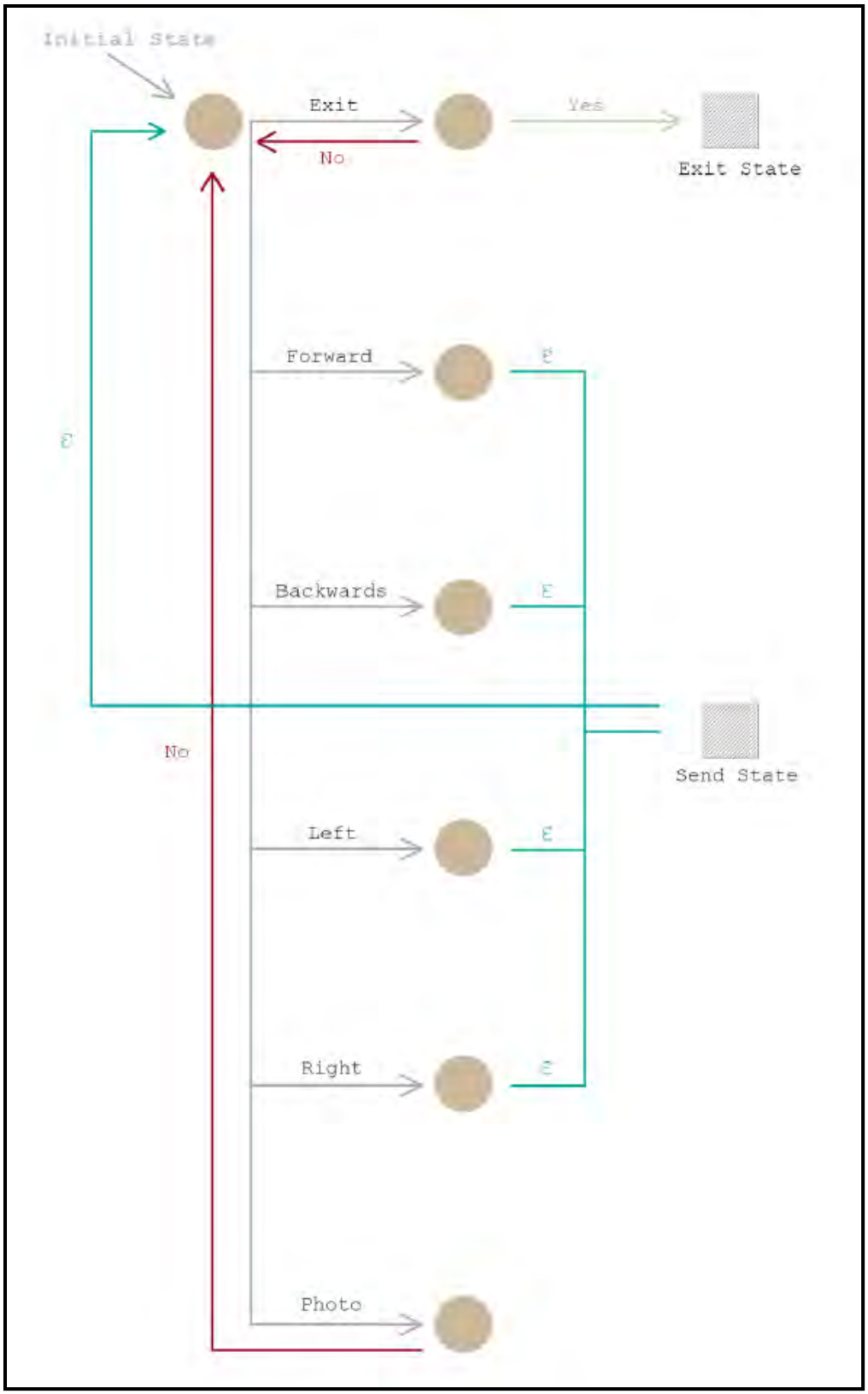
Σχήμα 4.1: Γλώσσα της υλοποίησης σε μορφή XML (από [17]).

	Προτάσεις που ανήκουν στη γλώσσα					
Μπροστά	Forward	Go forward	Move forward	Go straight	Move straight	-
Πίσω	Backwards	Go back	Go backwards	Move back	Move backwards	-
Αριστερά	Left	Turn left	Go left	-	-	-
Δεξιά	Right	Turn right	Go right	-	-	-
Στοπ	Stop	-	-	-	-	-
Φωτογραφία	Photograph	Picture	Photo	Take a photo	Take a photograph	Take a picture
Όχι	No	-	-	-	-	-
Ναι	Yes	-	-	-	-	-
Τερματισμός	Exit	-	-	-	-	-

Πίνακας 4.2: Όλες οι αποδεκτές προτάσεις που ανήκουν στη γλώσσα.

Εάν ο χρήστης δώσει ως είσοδο μια λέξη ή φράση που ανήκει στη γλώσσα και άρα αναγνωρίζεται, η μεταβλητή action θα πάρει την κατάλληλη τιμή και θα την επιστρέψει στη ροή του προγράμματος, ώστε να αποστείλει στο ρομπότ την αντίστοιχη εντολή. Εάν η είσοδος δεν αναγνωριστεί, δεν γίνει κατανοητή λόγω θορύβου ή αν δεν ειπωθεί τίποτα απολύτως, τότε το σύστημα διαλόγου του IrisTK θα ζητήσει εκ νέου είσοδο. Μόλις αναγνωριστεί η εντολή, αποστέλεται στο ρομπότ μέσω network socket.

Στην επόμενη σελίδα φαίνεται το FSM με όλες τις πιθανές μεταβάσεις καταστάσεων. Το Sent State αντιπροσωπεύει την αποστολή εντολής στο ρομπότ, ενώ το Exit State σημαίνει αποστολή ειδοποίησης τερματισμού στο ρομπότ και κλείσιμο του προγράμματος.



Σχήμα 4.2: FSM της λειτουργίας του συστήματος.

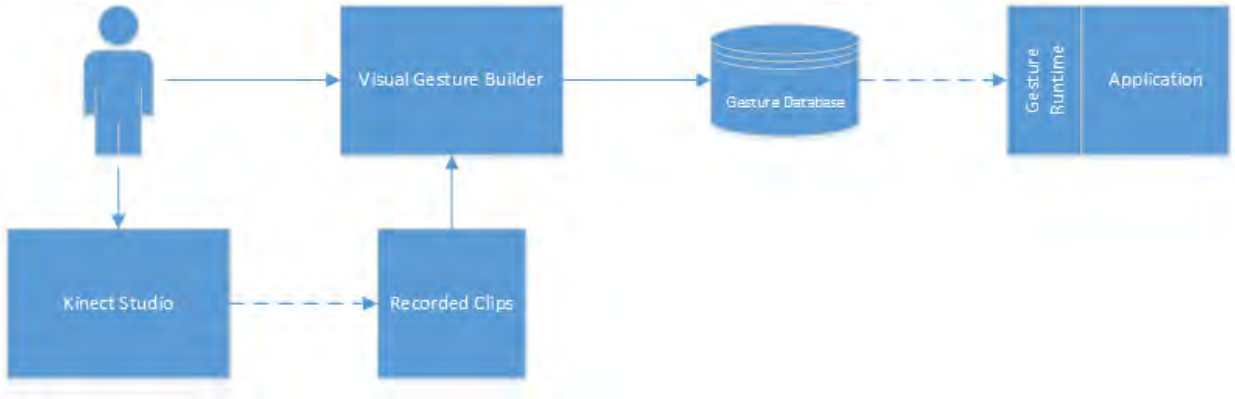
5 Εκπαίδευση Χειρονομιών και Πειράματα Ελέγχου

5.1 Visual gesture builder

Το εργαλείο που χρησιμοποιήθηκε για την δημιουργία και ενσωμάτωση των νέων χειρονομιών είναι το Visual Gesture Builder (VGB) της Microsoft. Η διαδικασία εκπαίδευσης είναι η παρακάτω:

1. Καταγραφή clip χειρονομιών (μέσω της εφαρμογής Kinect Studio)
2. Ορισμός των χρονικών σημείων του clip που εκτελείται η χειρονομία
3. Ανάλυση του clip και δημιουργία βάσης δεδομένων χειρονομίας
4. Έλεγχος του αποτελέσματος εκπαίδευσης

Η διαδικασία εκπαίδευσης τελικά δημιουργεί μία βάση δεδομένων η οποία περιέχει πληροφορίες για την αναγνώριση μιας χειρονομίας ή μίας ομάδας χειρονομιών. Έτσι ο προγραμματιστής μπορεί να εισάγει τη βάση στο πρόγραμμά του παρέχοντας την δυνατότητα αναγνώρισης των νέων χειρονομιών.



Σχήμα 5.1: Διάγραμμα δημιουργίας χειρονομιών (από [28]).

5.2 Clips εκπαίδευσης

Για την διαδικασία εκπαίδευσης χρειαζόμαστε έναν ικανοποιητικό αριθμό εκτελέσεων της χειρονομίας ώστε αυτή να γίνει κατανοητή από το VGB.

Λόγω του μεγέθους των clips, το οποίο ανέρχεται σε περίπου 25 MB/sec εξαιτίας του μεγάλου όγκου δεδομένων που περιέχονται σε αυτό (πληροφορίες βάθους, σκελετού, αρθρώσεων), ξεκινάμε την εκπαίδευση με 10 επαναλήψεις της κάθε χειρονομίας που αντιστοιχούν σε βίντεο περίπου 20 δευτερόλεπτων και μεγέθους 500 - 600 MB.

5.2.1 Clips Ελέγχου

Για να γίνει έλεγχος της αποτελεσματικότητας αναγνώρισης των χειρονομιών, παρέχουμε στο σύστημα test clips, πάνω στα οποία γίνεται δοκιμή αναγνώρισης.

Στο σχήμα 5.2 φαίνεται ένα παράδειγμα του αποτελέσματος ελέγχου για μια χειρονομία. Το μπλε σήμα δηλώνει το χρονικό διάστημα που εκτελείται η χειρονομία και είναι ορισμένο απ' τον χρήστη. Το πράσινο σήμα δηλώνει σε ποια χρονικά διαστήματα γίνεται η αυτόματη αναγνώριση της χειρονομίας με βάση την εκπαίδευση.

Σχήμα 5.2: Ο ορισμός των χειρονομιών με μπλε χρώμα και η αναγνώριση τους με πράσινο.

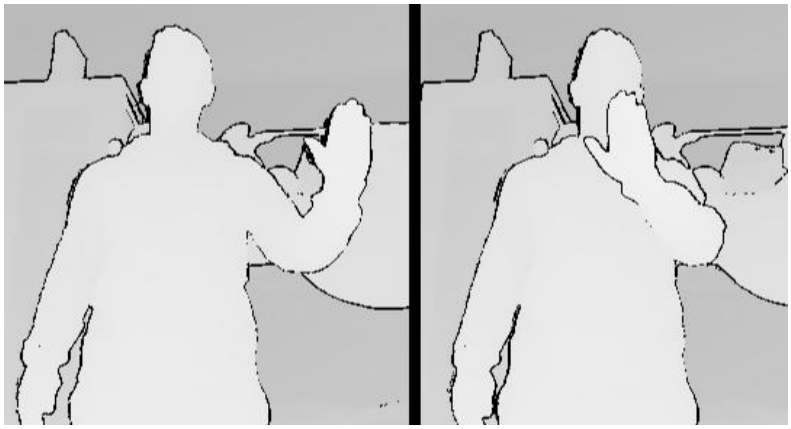
5.2.2 Confidence

Μια πολύ σημαντική πληροφορία για την διαδικασία ελέγχου αποτελεί η τιμή του confidence. Αυτή η τιμή δείχνει τον βαθμό σιγουριάς του detector για την χειρονομία που αναγνωρίζει (ή αλλιώς το πόσο ταυτίζεται η εξεταζόμενη χειρονομία με αυτές της εκπαίδευσης). Κυμαίνεται μεταξύ των ορίων [0, 1]. Μια αναγνώριση μπορεί να θεωρηθεί ικανοποιητική εάν η τιμή του confidence ξεπεράσει το 0,8.

5.3 Εκπαίδευση και έλεγχος της κάθε χειρονομίας

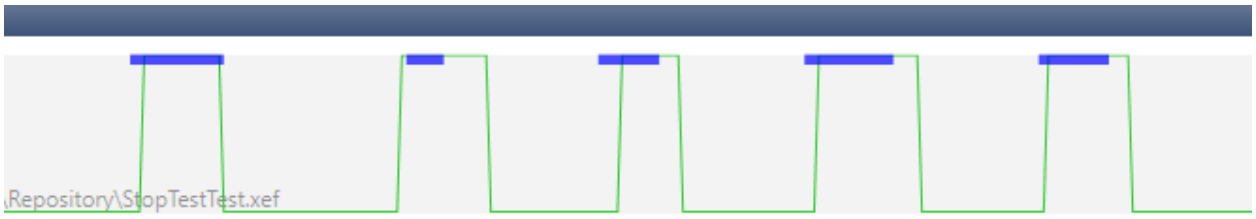
5.3.1 Εντολή «Στοπ»

Στη συγκεκριμένη εντολή ο χρήστης σηκώνει το χέρι του περίπου στο ύψος του στήθους με ανοιχτή την παλάμη. Η χειρονομία θα πρέπει να γίνεται αποδεκτή ακόμα κι αν το χέρι βρίσκεται σε διαφορετική θέση από την ιδανική. Όπως φαίνεται στο επόμενο σχήμα, εκτελούμε την χειρονομία με λίγο διαφορετικό τρόπο κάθε φορά.



Σχήμα 5.3: Διαφορετικές εκτελέσεις της χειρονομίας «Στοπ».

Μετά την εκπαίδευση, παρέχουμε ένα clip ελέγχου πέντε εκτελέσεων στο VGB και τρέχουμε την ανάλυση. Στο σχήμα 5.3 φαίνεται ότι αναγνωρίστηκαν και οι πέντε εκτελέσεις της χειρονομίας. Εξετάζοντας όμως τις τιμές του confidence για κάθε αναγνώριση, βλέπουμε ότι το αποτέλεσμα δεν είναι ικανοποιητικό. Παρατηρούμε ότι οι τιμές αυτές κυμαίνονται σε χαμηλά επίπεδα.



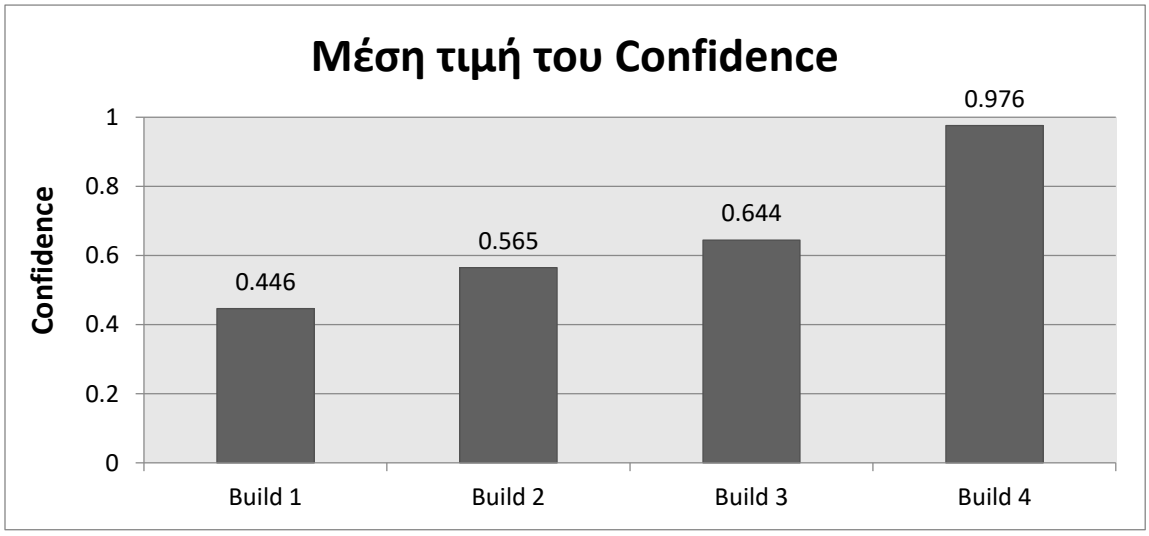
	Trial1	Trial2	Trial3	Trial4	Trial5
Confidence	0,191	0,275	0,547	0,810	0,409

Σχήμα 5.4: Αναγνώριση των χειρονομιών και το αντίστοιχο confidence.

Τα αποτελέσματα δεν είναι καθόλου ικανοποιητικά. Συνεπώς επανεκκινούμε την εκπαίδευση και αυτή την φορά συμπεριλαμβάνουμε το clip ελέγχου στο training set. Έπειτα πραγματοποιούμε έλεγχο σε νέο test clip. Στον πίνακα 5.2 φαίνονται τα αποτελέσματα των διαδοχικών εκπαιδεύσεων.

	Trial1	Trial2	Trial3	Trial4	Trial5	Average
Build 1	0,191	0,275	0,547	0,810	0,409	0,446
Build 2	0,391	0,426	0,522	1	0,487	0,565
Build 3	0,843	0,845	0,749	1	0,785	0,644
Build 4	0,883	1	1	1	1	0,976

Πίνακας 5.1: Οι τιμές του confidence για κάθε διαδοχικό έλεγχο.



Σχήμα 5.5: Η μέση τιμή του confidence των τεσσάρων διαφορετικών builds.

Χρειάστηκαν τέσσερις εκπαιδεύσεις με εικοσι-πέντε εκτελέσεις της χειρονομίας στο σύνολο για να καταλήξουμε σε ικανοποιητικές τιμές confidence.

5.3.2 Εντολή «Μπροστά»

Στην εντολή «Μπροστά» ο χρήστης τεντώνει το χέρι του μπροστά με κλειστή τη γροθιά του. Η συγκεκριμένη χειρονομία φαίνεται πως αναγνωρίζεται εύκολα απ' το VGB καθώς απ' την πρώτη κιόλας εκπαίδευση παίρνουμε υψηλές τιμές confidence. Χρειάστηκε ένα clip με δέκα εκτελέσεις.

	Trial1	Trial2	Trial3	Trial4	Trial5	Average
Build 1	1	0,998	1	1	1	0,999

Πίνακας 5.2: Οι τιμές του confidence για την εντολή «Μπροστά».

5.3.3 Εντολή «Πίσω»

Η χειρονομία «Πίσω» αναγνωρίζεται κι αυτή πολύ εύκολα μετά την πρώτη εκπαίδευση. Χρειάστηκε ένα clip με δέκα εκτελέσεις και το αποτέλεσμα που παράγεται είναι το απολύτως ιδανικό.

	Trial1	Trial2	Trial3	Trial4	Trial5	Average
Build 1	1	1	1	1	1	1

Πίνακας 5.3: Οι τιμές του confidence για την εντολή «Πίσω».

5.3.4 Εντολή «Αριστερά»

Και η χειρονομία «Αριστερά» είναι εύκολη στην αναγνώριση. Στην εκπαίδευση χρησιμοποιήθηκε clip δέκα εκτελέσεων, ενώ τα αποτελέσματα του ελέγχου φαίνονται παρακάτω.

	Trial1	Trial2	Trial3	Trial4	Trial5	Average
Build 1	0,946	1	1	0,877	1	0,964

Πίνακας 5.4: Οι τιμές του confidence για την εντολή «Αριστερά».

5.3.5 Εντολή «Δεξιά»

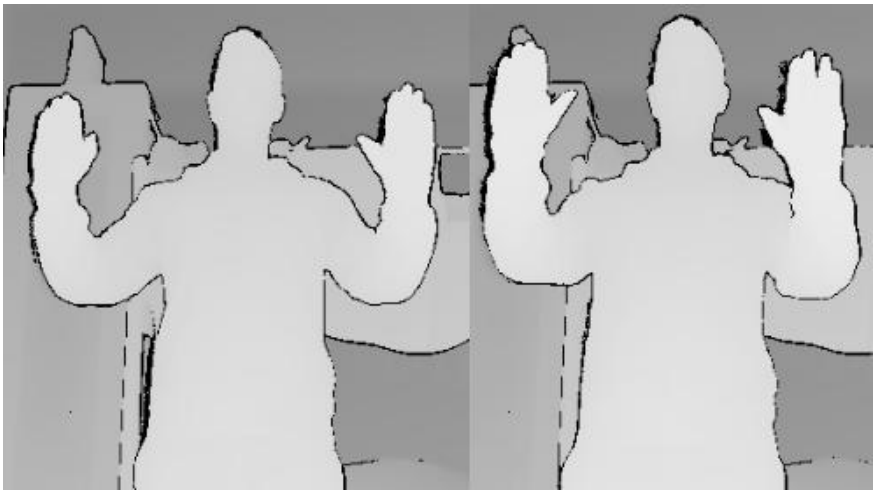
Η εκπαίδευση της χειρονομίας «Δεξιά» φαίνεται πως είναι λίγο πιο απαιτητική από τις προηγούμενες. Η πρώτη προσπάθεια δεν δίνει ικανοποιητικά αποτελέσματα. Συνολικά χρειάστηκαν δύο εκπαιδεύσεις με είκοσι εκτελέσεις.

	Trial1	Trial2	Trial3	Trial4	Trial5	Average
Build 1	0,780	0,153	0,297	0,390	0,600	0,444
Build 2	1	0,920	1	0,986	0,912	0,936

Πίνακας 5.5: Οι τιμές του confidence για την εντολή «Δεξιά».

5.3.6 Εντολή «Τερματισμός»

Η χειρονομία που αντιστοιχεί στην εντολή «Τερματισμός» έχει την ίδια ιδιαιτερότητα με την χειρονομία «Στοπ». Ο χρήστης σηκώνει τα χέρια με ανοιχτές παλάμες, αλλά το ύψος των χεριών, όπως κι η μεταξύ τους απόσταση, μπορεί να επηρεάσει την αναγνώρηση. Κατά τον έλεγχο της πρώτης εκπαίδευσης παρατηρούμε ότι οι τιμές του confidence παρουσιάζουν μεγάλες διακυμάνσεις, ανάλογα με την θέση των χεριών. Στο επόμενο σχήμα φαίνεται αυτή η διαφορά.



Σχήμα 5.6: Η εκτέλεση στα αριστερά με μεγάλο άνοιγμα χεριών δίνει confidence = 0,264 ενώ αυτή στα δεξιά με λίγο μικρότερο άνοιγμα δίνει confidence = 1.

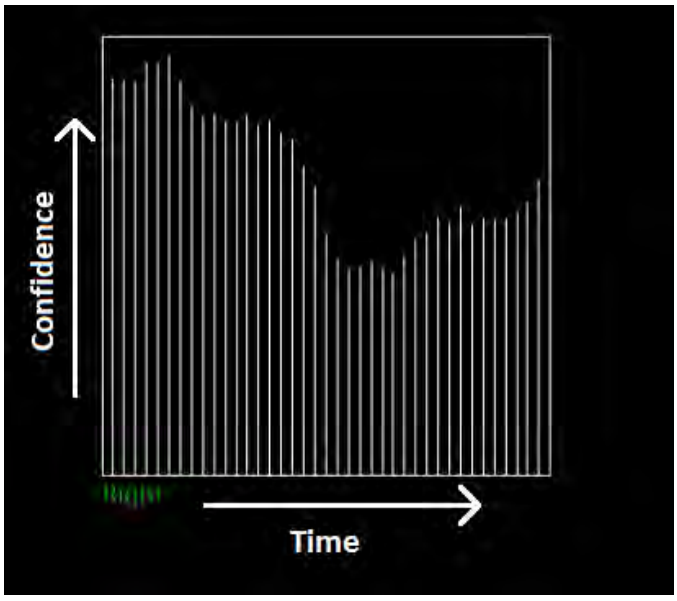
Εκκινούμε άλλη μία εκπαίδευση, προσθέτοντας clip με μεγαλύτερη ποικιλία στις θέσεις των χεριών. Έτσι καταφέρνουμε να απαλείψουμε την παραπάνω αδυναμία. Τελικά, με χρήση είκοσι εκτελέσεων της χειρονομίας καταλήγουμε στα παρακάτω αποτελέσματα.

	Trial1	Trial2	Trial3	Trial4	Trial5	Average
Build 1	0,264	1	0,456	1	0,894	0,722
Build 2	1	0,815	1	1	0,925	0,948

Πίνακας 5.6: Οι τιμές του confidence για την εντολή «Τερματισμός».

5.4 Live preview

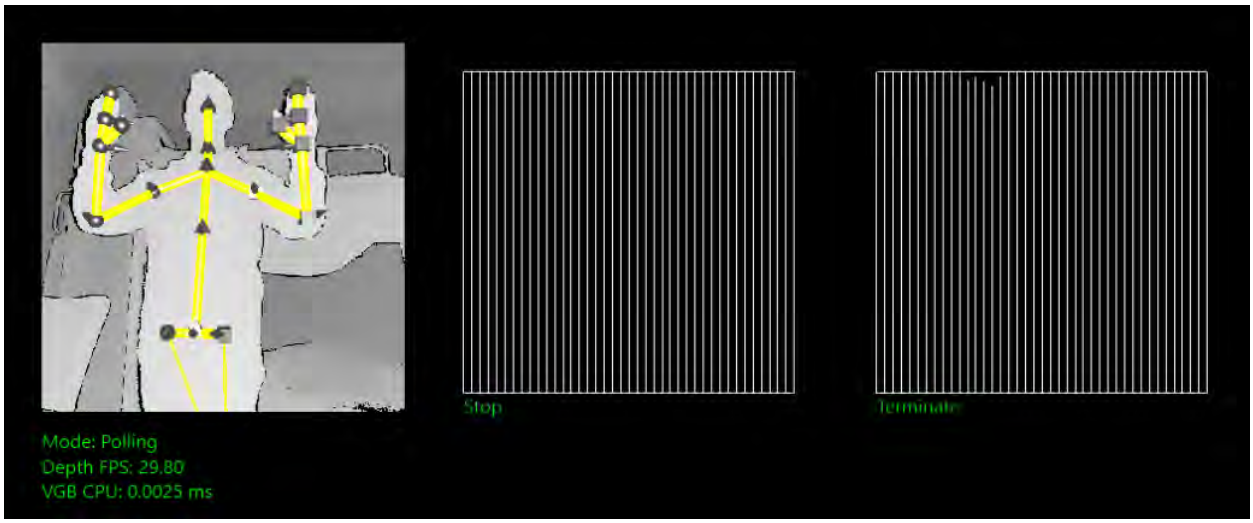
Το Gesture Builder δίνει την δυνατότητα του live ελέγχου χειρονομιών. Ο χρήστης μπορεί να κάνει δοκιμές με το Kinect βλέποντας άμεσα το αποτέλεσμα αναγνώρισης των χειρονομιών σε real time.



Σχήμα 5.7: Η μεταβολή του confidence στον χρόνο.

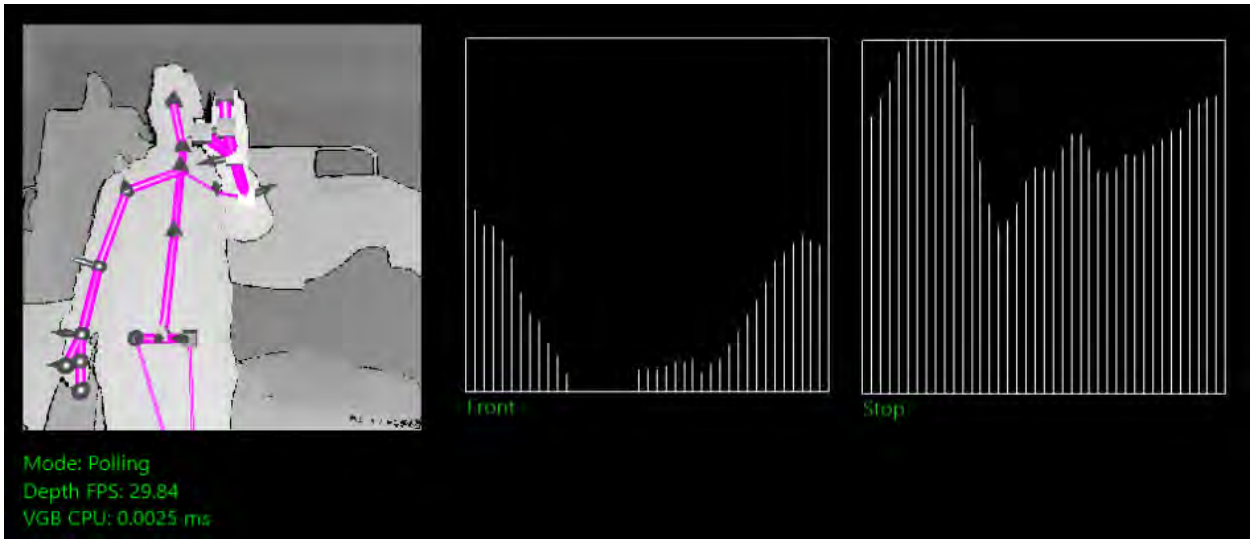
Ένα βασικό πλεονέκτημα του Live Preview είναι ότι μπορεί να ελεγχθεί ολόκληρη η βάση δεδομένων ταυτόχρονα. Με αυτόν τον τρόπο ανακαλύπτουμε ότι χειρονομίες που μοιάζουν μεταξύ τους μπορεί να αναγνωριστούν μαζί (η μία εκ των οποίων προφανώς λανθασμένα).

Πιο συγκεκριμένα, κατά την αναγνώριση της εντολής «Τερματισμός» αναγνωρίζεται και η εντολή «Στοπ», καθώς η δεύτερη περιέχεται στην πρώτη. (Στην πρώτη έχουμε σηκωμένα τα δύο χέρια με ανοιχτές παλάμες ενώ στη δεύτερη σηκωμένο μόνο το δεξί χέρι με ανοιχτή παλάμη).



Σχήμα 5.8: Ταυτόχρονη αναγνώριση εντολής «Στοπ» και «Τερματισμός».

Επίσης η χειρονομία «Μπροστά» μοιάζει λίγο με την «Στοπ», καθώς και στις δύο έχουμε προέκταση του δεξιού χεριού. Εύκολα μπορεί να προκληθεί σύγχυση και εδώ.



Σχήμα 5.9: Σύγχυση μεταξύ εντολής «Μπροστά» και «Στοπ».

Το παραπάνω πρόβλημα μπορεί να λυθεί με αρνητικά παραδείγματα κατά την εκπαίδευση. Συγκεκριμένα, όταν εκπαιδεύουμε την χειρονομία «Στοπ», παρέχουμε και ένα clip με εκτελέσεις της εντολής «Τερματισμός» ορίζοντας ρητά ότι αυτή η χειρονομία είναι διαφορετική. Το ίδιο κάνουμε και για την περίπτωση του ζευγαριού «Μπροστά» και «Στοπ». Έτσι ο detector καταφέρνει να ξεχωρίσει τις διαφορετικές περιπτώσεις και να μη συγχέει χειρονομίες που μοιάζουν μεταξύ τους.

5.5 Πειράματα ελέγχου του ρομπότ

Κατά την εκπαίδευση έγιναν οι απαραίτητοι έλεγχοι ώστε να εξακριβωθεί η σωστή αναγνώριση χειρονομιών. Τι γίνεται όμως όταν εισάγουμε την βάση δεδομένων στο πρόγραμμα χειρισμού του ρομπότ; Κρίνεται αναγκαίο να ελέγξουμε την ορθή λειτουργία του συνολικού συστήματος.

Ο στόχος του πειραματικού ελέγχου είναι να αντλήσουμε χρήσιμες πληροφορίες σχετικά με:

- Το ποσοστό επιτυχούς αναγνώρισης χειρονομιών-εντολών
- Την καθυστέρηση αναγνώρισης χειρονομίας και εκτέλεσης εντολής

5.5.1 Επιτυχία αναγνώρισης χειρονομίας

Για να εξάγουμε το ποσοστό επιτυχίας κάθε χειρονομίας εκτελούμε την καθεμία από 50 φορές και σημειώνουμε τον αριθμό των επιτυχών αναγνωρίσεων. Ο επόμενος πίνακας απεικονίζει τα ποσοστά επιτυχίας. Πρέπει να σημειωθεί ότι μια αναγνώριση θεωρείται επιτυχής εάν η τιμή του confidence είναι μεγαλύτερη ή ίση με 0,8.

	Ποσοστό επιτυχίας
Μπροστα	100%
Πίσω	96%
Αριστερά	92%
Δεξιά	100%
Στοπ	86%
Τερματισμός	96%

Πίνακας 5.7: Ποσοστά επιτυχίας αναγνώρισης χειρονομιών με confidence ≥ 0,8.

Παρατηρούμε ότι το ποσοστό επιτυχίας της εντολής «Στοπ» είναι σχετικά χαμηλό. Το γεγονός αυτό μπορεί να αποτελέσει πρόβλημα. Για παράδειγμα, εάν το ρομπότ πλησιάζει σε κάποιο εμπόδιο και η εντολή «Στοπ» αποτύχει, τότε θα έχουμε σύγκρουση.

Για να βελτιώσουμε τα ποσοστά, μπορούμε να χαμηλώσουμε το όριο του confidence, κάνοντας την αναγνώριση των εντολών πιο ελαστική. Ο παρακάτω πίνακας δείχνει τα ποσοστά με confidence μεγαλύτερο ή ίσο του 0,6.

	Ποσοστό επιτυχίας
Μπροστα	100%
Πίσω	100%
Αριστερά	100%
Δεξιά	100%
Στοπ	100%
Τερματισμός	100%

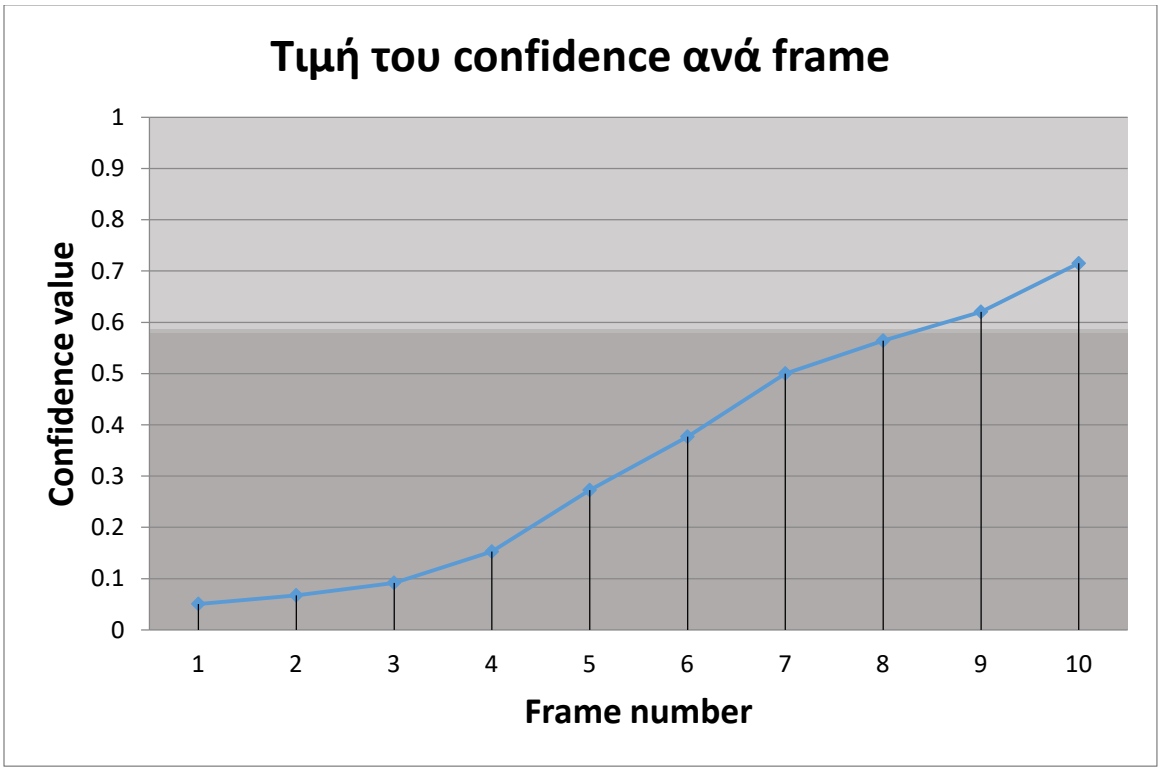
Πίνακας 5.8: Ποσοστά επιτυχίας αναγνώρισης χειρονομιών με confidence ≥ 0,6.

5.5.2 Καθυστέρηση αναγνώρισης και εκτέλεσης εντολής

Ο εξ αποστάσεως χειρισμός του ρομπότ είναι μια διαδικασία που απαιτεί αξιοπιστία και ταχύτητα. Ο χρόνος μεταξύ της πραγματοποίησης μιας χειρονομίας απ' τον χρήστη και της εκτέλεσης της εντολής απ' το ρομπότ πρέπει να είναι όσο πιο μικρός γίνεται. Στην συγκεκριμένη υλοποίηση, οι παράγοντες που επηρεάζουν αυτό το χρονικό διάστημα είναι οι παρακάτω:

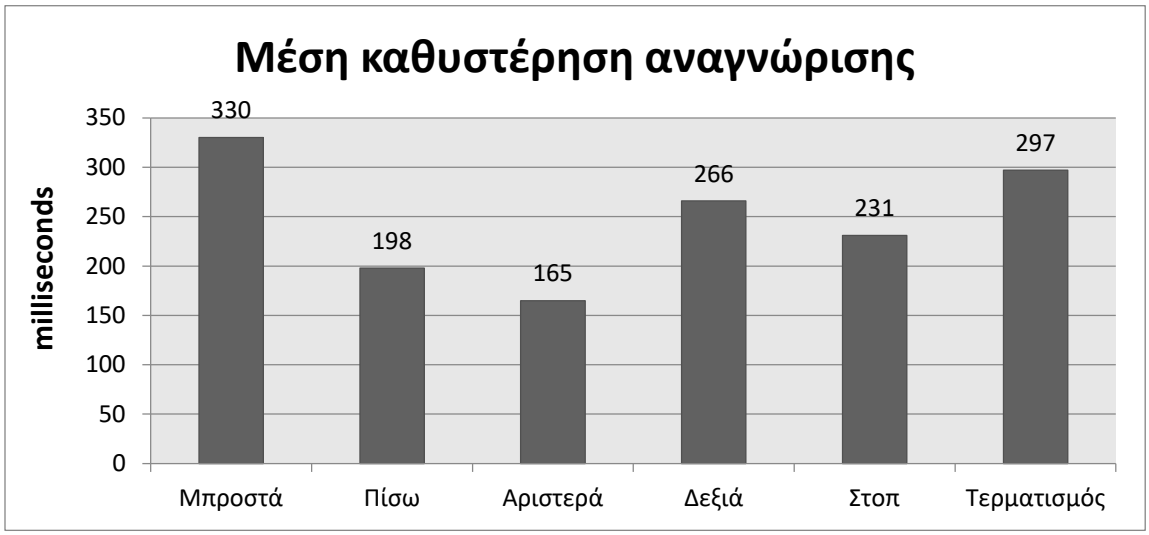
- Ταχύτητα αναγνωρισης χειρονομίας απ' το πρόγραμμα
- Καθυστέρηση επικοινωνίας μέσω του ασύρματου δικτύου
- Ταχύτητα απόκρισης και εκτέλεσης εντολής από το ρομπότ

Το πρόγραμμα αναγνώρισης δέχεται ροή βίντεο από το Kinect με ρυθμό 30 frames/sec. Αυτό σημαίνει ότι, στην ιδανική περίπτωση, δύναται να αναγνωριστεί χειρονομία κάθε 1/30 δευτερόλεπτα, ή αλλιώς 33ms. Ωστόσο η πραγματικότητα είναι διαφορετική. Τα πρώτα frames της αναγνώρισης δίνουν πολύ χαμηλές τιμές confidence και άρα η χειρονομία δεν γίνεται αμέσως αποδεκτή. Στο επόμενο διάγραμμα φαίνεται η καθυστέρηση κατά την αναγνώριση της εντολής «Μπροστά».



Σχήμα 5.10: Ο αριθμός των frames μέχρι την αναγνώριση. 9 frames αντιστοιχούν σε 297ms.

Το παρακάτω διάγραμμα δείχνει την μέση καθυστέρηση αναγνώρισης για κάθε εντολή, βάσει δείγματος 10 εκτελέσεων ανά περίπτωση.



Σχήμα 5.11: Η μέση καθυστέρηση για κάθε εντολή.

Έπειτα μετράμε το χρονικό διάστημα από τη στιγμή της αναγνώρισης έως τη στιγμή της εκτέλεσης της εντολής από το ρομπότ. Ο επόμενος πίνακας συνοψίζει τα αποτελέσματα 50 δοκιμών.

	Μέγιστη	Ελάχιστη	Μέση τιμή
Καθυστέρηση	150ms	32ms	96ms

Πίνακας 5.9: Καθυστέρηση εκτέλεσης εντολής.

Από το σχήμα 5.8 μπορούμε να εξάγουμε τον μέσο χρόνο αναγνώρισης χειρονομιών, 247ms. Σε αυτό προσθέτουμε την καθυστέρηση αποστολής και εκτέλεσης, καταλήγοντας έτσι σε μια συνολική καθυστέρηση 343ms.

6 Συμπεράσματα

6.1 Συνεισφορά της διπλωματικής εργασίας

Στην παρούσα διπλωματική εργασία παρουσιάστηκε ένα σύστημα αλληλεπίδρασης μεταξύ ενός οχήματος – ρομπότ ελεγχόμενου από Raspberry Pi 3 και ενός ή περισσότερων χρηστών. Χρησιμοποιήθηκε το IrisTK για την γρήγορη και εύκολη ανάπτυξη λογισμικού επικοινωνίας ανθρώπου – υπολογιστή καθώς και ο αισθητήρας Kinect v2 για την αναγνώριση χειρονομιών του χρήστη. Αναλύθηκε η λειτουργία των παραπάνω συστατικών μερών του συστήματος και παρουσιάστηκαν τα αποτελέσματα της προσπάθειας αυτής. Αποτελέσματα που μπορούν να θεωρηθούν ικανοποιητικά, καθώς δημιουργήθηκε ένα πλήρες σύστημα απομακρυσμένου χειρισμού του οχήματος μέσω φωνητικών εντολών και χειρονομιών σε πραγματικό χρόνο, με πολύ μικρή καθυστέρηση απόκρυσης και ικανοποιητική ακρίβεια κίνησης για τα μηχανικά μέρη που χρησιμοποιήθηκαν. Αποδείχτηκε έτσι πως ένα σύστημα επικοινωνίας ανθρώπου μηχανής είναι αρκετά εύκολο να πραγματωθεί υποστηρίζοντας συνδυαστικές εισόδους φωνής και κίνησης που μπορούν, σε διαφορετικές περιπτώσεις ανθρώπων, να αποδειχθούν πολύτιμα εργαλεία καθημερινής χρήσης και επικοινωνίας.

6.2 Μελλοντική έρευνα

Η παρούσα μελέτη μπορεί να χρησιμοποιηθεί ως παράδειγμα χρήσης της πλατφόρμας IrisTK και του αισθητήρα Kinect v2, καθώς και του συνδυασμού τους με κάποιον μικροελεγκτή, όπως το Raspberry Pi. Αν και η λειτουργικότητα του ρομπότ ήταν αρκετά περιορισμένη, η εργασία μπορεί να χρησιμοποιηθεί ως βάση πάνω στην οποία δύναται να χτιστεί κάποιο πιο πολύπλοκο σύστημα αλληλεπίδρασης ανθρώπινης φωνής ή κίνησης. Παραδείγματα μελλοντικής έρευνας αποτελούν εφαρμογές για άτομα με ειδικές ανάγκες, που μπορούν να εκμεταλλευτούν κατάλληλα τη συνδυαστική είσοδο για να διευκολυνθούν, ή να δημιουργήσουν. Όσον αφορά το εργαλείο IrisTK, μπορεί να χρησιμοποιηθεί και ως εργαλείο εκμάθησης της ανθρώπινης συμπεριφοράς και επικοινωνίας από ανθρώπους με αυτισμό ή άλλες δυσκολίες. Τέλος, στοιχεία της παρούσας εργασίας μπορούν να ληφθούν υπόψιν και για κάποια ψυχαγωγική διαδραστική εφαρμογή με χρήση του αισθητήρα Kinect v2.

Βιβλιογραφία

[1] Priyanka Shende and Prof. Archana Dehankar. “Gesture and Voice Based Real Time Control System”, In: *Global Journal of Engineering Science and Researches,* June 2014.

[2] Michael Tomasello. “Origins of Human Communication”, The MIT Press, 2008.

[3] Lawrence Rabiner and Biing-Hwang Juang. “Fundamentals of Speech Recognition”, PTR Prentice Hall, 1993, pp. 1-3.

[4] Mara Mills. “Media and Prosthesis: the Vocoder, the Artificial Larynx, and the History of Signal Processing”, Qui Parle, 2012, pp. 107-149.

[5] Lawrence Rabiner and B. H. Juang, “Automatic Speech Recognition - A Brief History of the Technology Development”, 2004.

[6] Benesty, Sondhi and Huang. “Springer Handbook of Speech Processing”, Springer, 2008, pp. 525-526.

[7] Melanie Pinola. “Speech Recognition Through the Decades: How we Ended Up with Siri”, [Online]. Available: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/. [Accessed 1-March-2017].

[8] Guzman Alvarez. “Speech Recognition Market”*,* In: *S-89.3680 Speech Recognition Seminar*, 2014.

[9] S. J. Young and L. L. Chase. “Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes*”,* In: *Computer Speech and Language,* vol. 12, no. 4, pp. 263-279, 1998.

[10] Margaret Rouse. “What is gesture recognition?”, [Online]. Available: http://whatis.techtarget.com/definition/gesture-recognition. [Accessed 5-February-2017].

[11] Thomas G. Zimmerman, Jason Lanier, Chuck Blanchard, Steve Bryson and Young Harvill. “A Hand Gesture Interface Device”, In: *SIGCHI/GI Conference*, April 1987.

[12] Zhou Ren, Jingjing Meng and Junsong Yuan. “Depth Camera Based Hand Gesture Recognition and its Applications in Human-Computer-Interaction”, 2011.

[13] Will Greenwald. “Kinect vs. PlayStation Move vs. Wii: Motion-Control Showdown”*,* [Online]. Available: http://www.pcmag.com/article2/0,2817,2372244,00.asp. [Accessed 18-February-2017].

[14] Ilana Rapp. “Motion Capture Actors: Body Movement Tells the Story”, [Online]. Available: http://www.nycastings.com/motion-capture-actors-body-movement-tells-the-story/. [Accessed 1-March-2017].

[15] Juan Wachs, Helman Stern, Yael Edan, Michael Gillam, Craig Feied, Mark Smith and Jon Handler. “Gestix: A Doctor-Computer Sterile Gesture Interface for Dynamic Environments”, In: *Soft Computing in Industrial Applications*, 2007.

[16] G. Skantze. “IrisTK: Intelligent Real-time Interactive Systems Toolkit”, [Online]. Available: http://www.iristk.net/index.html. [Accessed 20-February-2017].

[17] Dimitrios Panagiotou. “Gesture Based Interaction Using the Kinect Sensor”, Thesis, October 2016.

[18] “Supported Operating Systems for Raspberry Pi”, [Online]. Available: https://www.raspberrypi.org/downloads/. [Accessed 1-March-2017].

[19] C. Demerjian. “XBox One's Kinect sensor overcomes problems with intelligence”, SemiAccurate, [Online]. Available: http://semiaccurate.com/2013/10/16/xbox-ones-kinect-sensor-overcomes-problems-intelligence/. [Accessed 15-January-2017].

[20] Patrick O’Connor and John Sell. “The Xbox One System on a Chip and Kinect Sensor”, In: IEEE Micro 34(2):44-53, March 2014.

[21] “Kinect Hardware”, [Online]. Available: https://developer.microsoft.com/en-us/windows/kinect/hardware. [Accessed 1-March-2017].

[22] C. Demerjian. “A deep dive into Microsoft's XBox One's architecture”, SemiAccurate, [Online]. Available: http://semiaccurate.com/2013/08/29/a-deep-dive-into-microsofts-xbox-ones-architecture/. [Accessed 1-March-2017].

[23] Microsoft Corporation, “JointType Enumeration - Joint types in a skeleton”, Microsoft, [Ηλεκτρονικό]. Available: https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx. [Πρόσβαση 1 March 2017].

[24] Hand Gesture Cliparts, [Online]. Available: http://cliparts.co/hand-graphics. [Accessed 1-March-2017].

[25] Microsoft Corporation. “Visual Gesture Builder: A Data-Driven Solution to Gesture Detection”, 2013.

[26] G. Skantze. *Overview of IrisTK*, [Online]. Available: http://www.iristk.net/overview.html. [Accessed 1-March-2017].

[27] G. Skantze. *IrisTK Soeech synthesis*, [Online]. Available: http://www.iristk.net/speech_synthesis.html. [Accessed 12-February-2017].

[28] Microsoft Corporation. “Visual Gesture Builder: Overview”, [Online]. Available: https://i-msdn.sec.s-msft.com/dynimg/IC751355.png. [Accessed 1-March-2017].

[29] Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh and Mo Nours Arab, “Human-Computer Interaction: Overview on State of the Art”, In International Journal on Smart Sensing and Intelligent Systems, vol 1, March 2008.