ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Υλοποίηση Συστήματος Σύνθεσης Μουσικής Μέσω Κίνησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Χρήστου Α. Γαρούφη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής ΕΜΠ

Αθήνα, Μάρτιος 2018

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Υλοποίηση Συστήματος Σύνθεσης Μουσικής Μέσω Κίνησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Χρήστου Α. Γαρούφη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12/03/2018 .

.....	.....	.....
Πέτρος Μαραγκός, Καθηγητής ΕΜΠ	Κωνσταντίνος Τζαφέστας, Επικ. Καθηγητής ΕΜΠ	Γεράσιμος Ποταμιάνος, Αν. Καθηγητής Παν. Θεσσαλίας

Αθήνα, Μάρτιος 2018.

.....
Χρήστος Α. Γαρούφης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών ΕΜΠ

Περίληψη

Το θέμα της παρούσας διπλωματικής εργασίας είναι η δημιουργία ενός συστήματος σύνθεσης μουσικής, και της πλατφόρμας επικοινωνίας του με το χρήστη, το οποίο θα κάνει, με χρήση του αισθητήρα κίνησης του Kinect της Microsoft, tracking σε 2 άτομα και, ανάλογα με την κίνησή τους, θα συνθέτει μουσική σε πραγματικό χρόνο.

Ο εντοπισμός της στάσης των χεριών των δύο χρηστών γίνεται μέσω της ενσωματωμένης λειτουργίας skeleton tracking του Kinect. Συγκεκριμένα, με χρήση του ενσωματωμένου συστήματος παρακολούθησης μπορούν να προσδιοριστούν (με αρκετή ακρίβεια) οι θέσεις 25 σημείων του σώματος. Σε ότι αφορά τη στάση των χεριών, έχουμε ορίσει 12 πρότυπες στάσεις χεριών, οι οποίες αντιστοιχούν σε συγκεκριμένες νότες. Για τον προσδιορισμό της στάσης των χεριών, χρησιμοποιείται ένας απλός αλγόριθμος ταιριάσματος (matching), αφού πρώτα εξάγουμε σαν χαρακτηριστικά τα κανονικοποιημένα διανύσματα κατεύθυνσης των χεριών – τα οποία είναι ανεξάρτητα από τα χαρακτηριστικά του παίχτη – χρήστη.

Επιπλέον, για την επικοινωνία του παιχνιδιού με τους χρήστες, έχει εκπαιδευτεί ένα σύστημα αναγνώρισης δυναμικών χειρονομιών. Οι χειρονομίες είναι κωδικοποιημένες ως θέσεις των συνδέσμων που αναγνωρίζει η κάμερα, και για την αναγνώρισή τους, αφού εξήχθσαν και εδώ τα κανονικοποιημένα διανύσματα κατεύθυνσης ως γεωμετρικά χαρακτηριστικά, χρησιμοποιήθηκαν αλγόριθμοι που χρησιμοποιούνται για την αναγνώριση και ταξινόμηση χρονοσειρών. Πραγματοποιήθηκε πειραματισμός τόσο με κρυφά μοντέλα Markov, είτε μέσω της διακριτοποίησης των δεδομένων μέσω μίας παρόμοιας με την BoW προσέγγιση, είτε μέσω της εκμάθησης των παρατηρήσεων ως μίγμα Γκαουσιανών, όσο και με διάφορες παραλλαγές του αλγορίθμου του κοντινότερου γείτονα, είτε σε ότι αφορά τη μείωση της διαστατικότητας, είτε την καλύτερη προσαρμογή του στο πρόβλημα της σύγκρισης χρονοσειρών. Τον ταξινομητή συμπληρώνουν τόσο ένας ανιχνευτής κίνησης, όσο και ένας μηχανισμός απόρριψης χειρονομιών που δεν βρίσκονται στο σύνολο εκπαίδευσης.

Η μουσική νότα που αντιστοιχεί σε κάθε διαφορετική στάση χεριών καθορίζεται με την αντιστοίχιση διαφορετικής συχνότητας σε καθεμιά από αυτές, με την απαραίτητη προσοχή ώστε παραπλήσιες στάσεις να αντιστοιχούν σε κοντινές ηχητικά νότες. Για την παραγωγή των ημιτονικών κυμάτων, χρησιμοποιήθηκε η βιβλιοθήκη ανοιχτού κώδικα portaudio.

Λέξεις κλειδιά: Διαδραστική εφαρμογή, σύνθεση μουσικής από κίνηση, Kinect, αναγνώριση χειρονομιών, σκελετικά δεδομένα, αλγόριθμος ταιριάσματος

Abstract

The scope of the following thesis is the programming of a virtual music box, and its user interface, which will, using the motion sensor of a Microsoft Kinect, perform skeleton tracking on 2 people, and, dependent on their movements, will compose music.

The detection of the hand pose of the two users is achieved using the skeleton tracking feature of the Kinect. More specifically, using the built-in tracking system, the positions of 25 joints of the human body can be defined with satisfying accuracy. Regarding the static hand pose, we have defined 12 template hand stances, which correspond to specific musical notes. For hand pose classification, we are using a simple template matching algorithm, using as features the normalized direction vectors of the hands, which are largely independent from a player's person - specific characteristics.

Furthermore, to model a natural interface between the application and the users, a system for detecting and classifying dynamic hand gestures has been built. Again, using the initial encoding of the joint positions in (x, y, z) coordinates, the normalized direction vectors were extracted as geometric characteristics, and algorithms that are normally used for identifying and classifying time series were used on them. We experimented with a) Hidden Markov Models, either via discretizing the input data using a simplified Bag-of-Words approach, or via learning the observations as a mixture of Gaussians and b) some variations of the vanilla nearest-neighbour algorithm, by either reducing the dimensionality of the search space and the feature space, or adapting the algorithm to the problem of comparing time-series. The bare classifier is supported by an activity detector, as well as a control mechanism that rejects gestures that were not learned.

The musical note corresponding to each different hand pose is defined by matching a unique frequency to each possible hand pose, with care being taken to match similar poses to close musically notes. The open-source portaudio library was used to produce the sine waves.

Keywords: Interactive app, motion-based music synthesis, Kinect, gesture recognition, skeletal data, template matching.

Ευχαριστίες

Η παρούσα διπλωματική εργασία σηματοδοτεί τη λήξη των προπτυχιακών σπουδών μου στο Εθνικό Μετσόβιο Πολυτεχνείο. Με αφορμή, λοιπόν, την περάτωσή της, νιώθω την υποχρέωση να ευχαριστήσω:

Τον καθηγητή κ. Πέτρο Μαραγκό, για την ευκαιρία που μου έδωσε να εξερευνήσω τα ερευνητικά μου ενδιαφέροντα με την εκπόνηση της παρούσας διπλωματικής εργασίας, καθώς και τη μεταδιδακτορικό ερευνητή του εργαστηρίου CVSP, Δρ. Νάνσυ Ζλατίντση, και το διδακτορικό φοιτητή του εργαστηρίου, Πέτρο Κούτρα, για την καίρια συμβολή τους στο να πάρει η παρούσα εργασία την τελική της μορφή.

Τους καθηγητές του Πανεπιστημίου Ιωαννίνων, κ. Αριστείδη Λύκα και κ. Χριστόφορο Νίκου, για το ενδιαφέρον που έδειξαν και για το δανεισμό του Kinect κατά το διάστημα στο οποίο βρισκόμουν εκτός Αθηνών, καθώς και όλους όσους, ανεβάζοντας τη δουλειά τους στο διαδίκτυο, προωθούν την ελεύθερη διακίνηση της γνώσης.

Το φιλικό μου περιβάλλον, και ιδίως τους Κώστα, Άλκη, Γιώργο και Άρτεμη, και όλα τα άτομα με τα οποία είχα την ευκαιρία να συναναστραφώ στα πλαίσια της φοίτησής μου στο Εθνικό Μετσόβιο Πολυτεχνείο, είτε στα πλαίσια της συνεργασίας μεταξύ φοιτητών και φίλων - σημείο στο οποίο θέλω να σταθώ ειδικά στα παιδιά με τα οποία εκπονούσαμε παράλληλα τις διπλωματικές μας -, είτε για την αφορμή που μου έδωσαν να διευρύνω τους ορίζοντές μου, είτε, απλά, για τις όμορφες στιγμές τις οποίες μοιραστήκαμε κατά τη διάρκεια των φοιτητικών μας χρόνων (είτε για όλα τα παραπάνω).

Τέλος, θεωρώ απαραίτητο να ευχαριστήσω την οικογένειά μου για την ψυχολογική στήριξη που μου έδωσε - και ιδιαίτερα, τον πατέρα μου, Αχιλλέα, για τη συνεχή στήριξη και καθοδήγηση που μου παρείχε αδιάλειπτα, και την αδερφή μου, Νεφέλη, για την αστείρευτη υπομονή και ανοχή της.

Χρήστος Γαρούφης,

4/1/2018

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Περιεχόμενα

1 Εισαγωγή

1.1 Όραση Υπολογιστών και Μηχανική Μάθηση

Η όραση υπολογιστών είναι το επιστημονικό πεδίο το οποίο ασχολείται με την ανάλυση και εξαγωγή πληροφορίας υψηλού επιπέδου από ψηφιακές εικόνες ή εικονοσειρές (video) από πλευράς υπολογιστικών συστημάτων. Πρακτικά, δηλαδή, ασχολείται με την αυτοματοποίηση διαδικασιών και εργασιών του ανθρώπινου οπτικού συστήματος. Κατά κανόνα, αυτό επιτυγχάνεται με την ανάλυση και επεξεργασία οπτικού υλικού, με ενδιάμεσο στόχο την εξαγωγή χαρακτηριστικών από αυτό, και τελικό στόχο τη μετατροπή του σε συμβολική πληροφορία. Για τη μετατροπή των οπτικών δεδομένων – όπως αυτά λαμβάνονται από κατάλληλους αισθητήρες – σε μορφή που αναπαριστά συμβολική πληροφορία, χρησιμοποιούνται μοντέλα από τη γεωμετρία, τη στατιστική, τη φυσική και τη θεωρία βελτιστοποίησης, ενώ επιμέρους υποπροβλήματα της όρασης υπολογιστών αποτελούν η αναγνώριση αντικειμένων, η εκτίμηση κίνησης, η εκτίμηση πόζας αντικειμένων, ο εντοπισμός γεγονότων, η ανακατασκευή τρισδιάστατης εικόνας, και η αποθορυβοποίηση/ αποκατάσταση εικόνας.

Υπό αυτό το πρίσμα, είναι εμφανής η σύνδεση της όρασης υπολογιστών με τη μηχανική μάθηση και την αναγνώριση προτύπων. Και οι δύο ερευνητικές περιοχές πηγάζουν από την τεχνητή νοημοσύνη, και εστιάζουν στη δυνατότητα ταξινόμησης ομοειδών αντικειμένων σε κατηγορίες (κλάσεις). Η βασική διαφορά των δύο πεδίων είναι ότι ενώ κατά την αναγνώριση προτύπων δίνεται έμφαση στη στατιστική περιγραφή πληροφορίας, και στη μοντελοποίηση των κλάσεων ταξινόμησης μέσω κάποιων χαρακτηριστικών (γεννητικά μοντέλα), οι τεχνικές μηχανικές μάθησης συνήθως αποτελούν προβλήματα βελτιστοποίησης, χρησιμοποιούν δεδομένα και δίνεται έμφαση στην εύρεση του συνόρου των περιοχών του χώρου που αντιστοιχούν σε κάθε κλάση ταξινόμησης (διακριτικά μοντέλα). Τα προβλήματα των περιοχών αυτά χωρίζονται, κατά κανόνα, σε επιβλεπόμενα ή μη επιβλεπόμενα, αναλόγως με το αν γνωρίζουμε ή όχι εκ των προτέρων την κλάση που ανήκουν τα προς ταξινόμηση αντικείμενα. Η αναγνώριση αντικειμένων αποτελεί παράδειγμα προβλήματος επιβλεπόμενης μάθησης στον τομέα της όρασης υπολογιστών, ενώ η κατάτμηση εικόνας, κατά κανόνα, παράδειγμα προβλήματος μη επιβλεπόμενης μάθησης.

1.2 Αλληλεπίδραση Ανθρώπου - Μηχανής και Εφαρμογές

Ως αλληλεπίδραση ανθρώπου – μηχανής ορίζουμε το επιστημονικό πεδίο το οποίο έχει ως αντικείμενο τη σχεδίαση και τη χρήση υπολογιστικής τεχνολογίας, με έμφαση στις διεπαφές μεταξύ ανθρώπων και υπολογιστών. Οι ερευνητές στο πεδίο αυτό ασχολούνται τόσο με τους τρόπους με τους οποίους οι άνθρωποι αλληλεπιδρούν με τους υπολογιστές, όσο και με την ανάπτυξη τεχνολογιών που επιτρέπουν την επικοινωνία μεταξύ ανθρώπων και υπολογιστών με καινοτόμες μεθόδους.

Εμφανής είναι, επίσης, η σύνδεση μεταξύ της αλληλεπίδρασης μεταξύ ανθρώπου – υπολογιστή και της πρόσφατης επιστημονικής προόδου στα πεδία της όρασης υπολογιστών, της μηχανικής μάθησης και της επεξεργασίας φυσικής γλώσσας (natural language processing). Η πρόοδος αυτή είναι ο παράγοντας ο οποίος έχει καταστήσει δυνατή – ή μελλοντικά εφικτή – την αντικατάσταση των παρουσών διεπαφών με άλλες, πιο χρηστικές από τον μέσο άνθρωπο και πιο «φυσικές» (για παράδειγμα, αναφέρεται η παράλληλη χρήση πληκτρολογίων και συστημάτων type-to-write).

Σε ό,τι αφορά τη σύμμειξη της αλληλεπίδρασης ανθρώπου – υπολογιστή και όρασης υπολογιστών, αξίζει να επισημάνουμε το ότι καθώς, πλέον, είναι δυνατή η αναγνώριση παρουσίας ανθρώπων, και η εξαγωγή χαρακτηριστικών σε σχέση με τις χειρονομίες του, την κατεύθυνση του βλέμματός του, κτλ, παρέχονται πολλές δυνατότητες επέκτασης του πεδίου αυτού. Ενδεικτικά αναφέρεται, ως εργασία – ορόσημο, η εργασία των Viola, Jones [17], η οποία εξήγαγε εύρωστα χαρακτηριστικά μέσω πολλών απλών συνδυασμένων ταξινομητών (Haar features) με στόχο τον εντοπισμό και παρακολούθηση προσώπων.

1.3 Αναγνώριση Χειρονομιών

Ένα ακόμα επιστημονικό πεδίο, το οποίο, έχοντας αφετηρία την όραση υπολογιστών και τη μηχανική μάθηση, εφαρμόζει τα παραπάνω με τελικό στόχο τη βελτίωση της αλληλεπίδρασης μεταξύ ανθρώπων και υπολογιστών, είναι αυτό της αναγνώρισης χειρονομιών (gesture recognition), το οποίο ορίζεται ως η απόπειρα να ερμηνευτούν οι ανθρώπινες χειρονομίες με χρήση μαθηματικών (και άρα προγραμματίσιμων σε υπολογιστές) αλγορίθμων. Επί του παρόντος, έχει δοθεί έμφαση ερευνητικά στην αναγνώριση συναισθήματος μέσω των εκφράσεων του προσώπου [18] και σε αναγνώριση χειρονομιών των χεριών, με ενδεικτικό πεδίο εφαρμογής την αναγνώριση του λεξιλογίου της νοηματικής γλώσσας [19], ωστόσο σε αυτό μπορούν να υπαχθούν και άλλες περιοχές έρευνας, όπως ο προσδιορισμός της στάσης σώματος [20], του τύπου βαδίσματος [21], ή και ανθρωπίνων συμπεριφορών, βάσει της στάσης σώματος [22].

Οι αλγόριθμοι που χρησιμοποιούνται για τους σκοπούς αυτούς μπορούν να διακριθούν στις εξής κατηγορίες, αναλόγως με τη μεθοδολογία που ακολουθείται:

- Αλγόριθμοι που χρησιμοποιούν τρισδιάστατο μοντέλο [23]: Κατά κανόνα, ένα μέλος του σώματος μεταφράζεται ως ένα σύνολο γραμμών και κόμβων στο χώρο, και η όποια χειρονομία πραγματοποιείται μπορεί να εξαχθεί από τη σχετική θέση και τις αλληλεπιδράσεις μεταξύ των επιμέρους στοιχείων. Παρότι χρησιμοποιείται, ως μεθοδολογία, σε εφαρμογές όρασης που δεν απαιτούν απόκριση σε πραγματικό χρόνο ή σε γραφική υπολογιστών, δεν είναι πάντοτε ελκυστική λόγω της υπολογιστικής της πολυπλοκότητας.

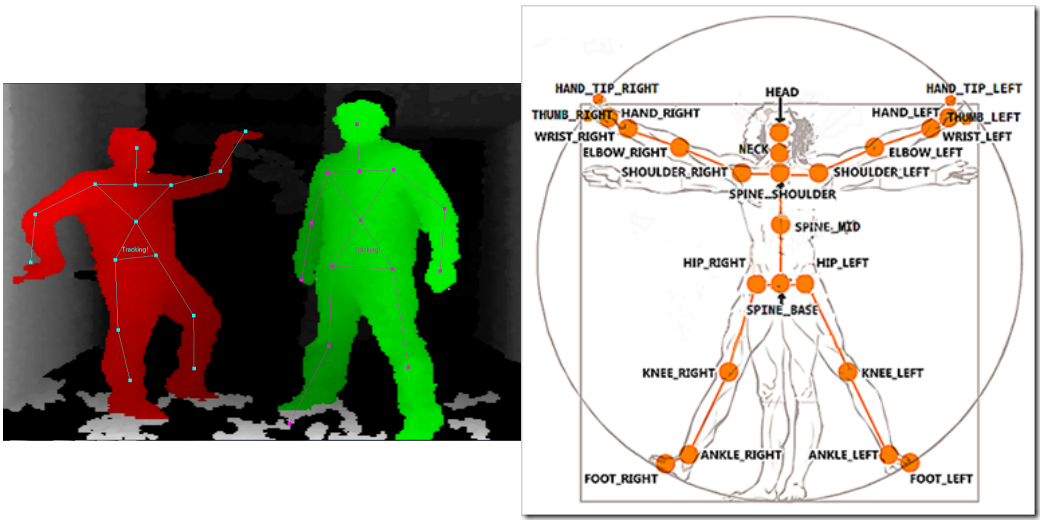
- Αλγόριθμοι που χρησιμοποιούν απλουστευμένο σκελετικό μοντέλο [24]: Αντί της αναπαράστασης του ανθρώπινου σώματος (ή τμημάτων αυτού) ως ένα σύνθετο πλέγμα τρισδιάστατων επιφανειών, εξάγεται μία σκελετική αναπαράσταση χρησιμοποιώντας τα μήκη των επιμέρους τμημάτων και τις διευθύνσεις τους. Το υπολογιστικό πλεονέκτημα αυτής της μεθόδου είναι ότι επιτυγχάνεται η αναπαράσταση του σώματος, του χεριού κτλ με χρήση κάποιων σημείων/κλειδιών (keypoints).

- Αλγόριθμοι που χρησιμοποιούν μοντέλα βασισμένα στην εμφάνιση (appearance-based) [25]: Και σε αυτή την περίπτωση, η αναπαράσταση των μελών του σώματος ανάγεται σε ορισμένα σημεία – κλειδιά, όμως η επεξεργασία γίνεται απευθείας από διδιάστατη εικόνα, και όχι με χρήση τρισδιάστατου μοντέλου. Σε αυτή την περίπτωση, συνήθη είσοδο του αλγορίθμου επεξεργασίας της χειρονομίας αποτελεί ή το περίγραμμα του μέλους που εκτελεί τη χειρονομία, ή η επιφάνεια που αντιστοιχεί σε αυτό, και γίνεται σύγκριση με μία βάση χειρονομιών.

Αξίζει να επισημάνουμε, τέλος, τη διάκριση μεταξύ στατικών και δυναμικών χειρονομιών. Μία στατική χειρονομία εξελίσσεται μόνο στο χωρικό επίπεδο. Αντιθέτως, σε περιπτώσεις εκτέλεσης δυναμικών χειρονομιών, μας αφορά η εξέλιξη του φαινομένου στο χρόνο, και συνεπώς απαιτούν διαφορετικό χειρισμό κατά την αναγνώριση και ταξινόμησή τους.

1.4 Το Kinect

Η συσκευή Kinect της Microsoft πρώτης γενιάς [26] (στην οποία, στη συνέχεια του κειμένου της διπλωματικής, θα αναφερόμαστε ως το Kinect) κυκλοφόρησε στην αγορά το 2010, και το 2013, κυκλοφόρησε η συσκευή 2ης γενιάς, Kinect v2. Από άποψη υλικού, αποτελείται από μία κάμερα χρώματος (RGB), μία κάμερα time-of-flight που μεταδίδει δεδομένα βάθους, και μία υπέρυθρη κάμερα (IR) οι οποίες μεταδίδουν δεδομένα σε ρυθμό 30 fps. Οι προδιαγραφές αυτές επιτρέπουν την τρισδιάστατη παρακολούθηση αντικειμένων, και την εφαρμογή διαφόρων αλγορίθμων όρασης υπολογιστών ή μηχανικής μάθησης, αναλόγως με τον τελικό σκοπό της εφαρμογής. Επιπλέον, είναι εξοπλισμένο και με ορισμένες built-in διεπαφές (είτε με χρήση ορισμένων χειρονομιών, είτε με συγκεκριμένες λεκτικές εντολές), ενώ παρέχει και τη δυνατότητα οπτικής παρακολούθησης σε 1 ή 2 άτομα, βάσει εκτιμήσεων των θέσεων των σκελετικών αρθρώσεων. Με χρήση της, έχει αναπτυχθεί πληθώρα εφαρμογών, με ενδεικτικές περιοχές εφαρμογής τη ρομποτική, ενσωματώνοντας το σε διάφορους βραχίονες, την ιατρική, είτε επιτρέποντας, μέσω εικόνων βάθους, το χειρισμό συστημάτων τηλεχειρουργικής, είτε μέσω του εντοπισμού διαταραχών στην κίνηση ατόμων με διάφορες παθήσεις (και έτσι, χρησιμοποιώντας τις πρώτες ως πιθανή ένδειξη κάποιας πάθησης), και άλλα.



Σχήμα 1.1: α) Εικόνα tracking από το Kinect. Φαίνονται οι σκελετοί των δύο ανθρώπων ως εκτιμήσεις θέσεων των αρθρώσεων [1]. β) Η αναπαράσταση του σκελετού του ανθρώπου, όπως αυτή επιστρέφεται από το Skeleton Tracking του Kinect. Μας ενδιαφέρουν κυρίως οι θέσεις των αρθρώσεων που επιστρέφονται στα χέρια και στον κορμό [2].

1.5 Ανάλυση Χρονοσειρών

Αν ορίσουμε ως χρονοσειρά οποιαδήποτε αλληλουχία δεδομένων έχει χρονική συσχέτιση, τότε είναι εμφανές από τα παραπάνω ότι, σε περιπτώσεις αναπαράστασης μίας χειρονομίας μέσω keypoints, η ανάλυση και επεξεργασία των δυναμικών χειρονομιών εμπίπτει στο πεδίο της ανάλυσης χρονοσειρών (time-series analysis), η οποία αφορά στην επεξεργασία δεδομένων με χρονική συνάφεια.

Στα πλαίσια της μηχανικής μάθησης, η ανάλυση χρονοσειρών χρησιμοποιείται για την εξαγωγή συγκεκριμένων – στατιστικών, για παράδειγμα – χαρακτηριστικών, ή το μετασχηματισμό, των όποιων χρονοσειρών εισόδου, με τελικό σκοπό την ευκολότερη διακρισιμότητα μεταξύ χρονοσειρών που ανήκουν σε διαφορετικές κατηγορίες, είτε για την ακριβέστερη ταξινόμησή τους, είτε για την ομαδοποίησή τους.

1.6 Σκοπός, Συνεισφορές και Διάρθρωση της Διπλωματικής Εργασίας

Σκοπός της παρούσας διπλωματικής εργασίας είναι η υλοποίηση μίας εφαρμογής Kinect, μέσω της οποίας δύο παίκτες θα μπορούν, μέσω των κινήσεων των χεριών τους, να συνθέτουν μουσική σε πραγματικό χρόνο. Ως δεδομένα εισόδου της εφαρμογής, θα χρησιμοποιούμε μόνο τα παρεχόμενα από το Kinect σκελετικά δεδομένα, και όχι δεδομένα χρώματος ή βάθους. Ως προς τη λειτουργία της εφαρμογής, ο ένας

παίκτης, αναλόγως με την κίνησή των χεριών του, θα παράγει συγκεκριμένες μουσικές νότες, ενώ ο άλλος, πραγματοποιώντας χειρονομίες, θα ελέγχει την έναρξη και τον τερματισμό της εφαρμογής, καθώς και το ρυθμό με τον οποίον παράγονται οι διαδοχικές νότες. Συνεπώς, απαραίτητη είναι η χρήση αλγορίθμων, οι οποίοι θα μπορούν να αναγνωρίζουν συγκεκριμένες στάσεις των χεριών από σκελετικές αναπαραστάσεις του σώματος. Υπό το πλαίσιο αυτό, συγκαταλέγουμε στις συνεισφορές της παρούσας διπλωματικής:

- Τη χρήση, ως χαρακτηριστικών προς αναγνώριση τόσο στατικών, όσο και δυναμικών χειρονομιών, των κανονικοποιημένων - κατά διεύθυνση - μηκών των τμημάτων των χεριών μέχρι και από τον αγκώνα (κεφάλαιο 4.2), τα οποία και λαμβάνουμε από σκελετικά δεδομένα.

- Την υλοποίηση ενός pipeline τριών σταδίων για τον εντοπισμό και την ταξινόμηση δυναμικών χειρονομιών, αποτελούμενο από έναν ανιχνευτή κίνησης, έναν ταξινομητή, και ενός συστήματος το οποίο θα απορρίπτει χειρονομίες που εντοπίζονται αλλά δεν υπερβαίνουν ένα κατώφλι ομοιότητας.

- Τον πειραματισμό με εξελικτικούς αλγορίθμους για την αντιστοίχιση παραπλήσιων στάσεων των χεριών σε κοντινούς συχνοτικά ήχους.

- Το συνδυασμό των παραπάνω για την υλοποίηση μίας εφαρμογής, μέσω της οποίας θα συντίθεται μουσική μέσω κίνησης σε πραγματικό χρόνο.

Σε ό,τι αφορά τη διάρθρωση της εργασίας:

- Στο κεφάλαιο 2 γίνεται μία βιβλιογραφική ανασκόπηση πάνω σε αλγορίθμους που χρησιμοποιούνται τόσο για την εκτίμηση στάσης σώματος σε εικόνες και εικονοσειρές, όσο και για την ταξινόμηση χειρονομιών από σκελετικά δεδομένα, και γίνεται αναφορά σε παρόμοια συστήματα τα οποία συνθέτουν μουσική μέσω κίνησης.

- Στο κεφάλαιο 3 αναλύονται θεωρητικά τα εργαλεία από το χώρο της αναγνώρισης προτύπων που χρησιμοποιήσαμε για την υλοποίηση των επιμέρους υποσυστημάτων, και περιγράφεται ο αλγόριθμος που χρησιμοποιεί για το Kinect για τον προσδιορισμό της στάσης σώματος.

- Στο κεφάλαιο 4, περιγράφεται αναλυτικά η λειτουργία της εφαρμογής Kinect, και, όπου κρίνεται απαραίτητο, αιτιολογούνται οι σχεδιαστικές επιλογές που πραγματοποιήθηκαν για κάθε υποσύστημα.

- Στο κεφάλαιο 5, επιχειρείται η ποσοτική αξιολόγηση της εφαρμογής που αναπτύξαμε, τόσο σε επίπεδο επιμέρους υποσυστημάτων, όσο και σε πραγματικές συνθήκες λειτουργίας.

- Τέλος, στο κεφάλαιο 6, αναφέρονται κάποιες πιθανές επεκτάσεις της διπλωματικής, είτε σε επίπεδο αλγορίθμων που μπορούν να χρησιμοποιηθούν, είτε σε επίπεδο επιπλέον δυνατοτήτων και λειτουργικότητας του προγράμματος.

- Ορισμένα μαθηματικά θέματα – ή ζητήματα που δεν άπτονται άμεσα επί της εργασίας – αναλύονται περαιτέρω στο συμπληρωματικό υλικό.

2 Σχετική Βιβλιογραφία

2.1 Εκτίμηση Στάσης Σώματος

2.1.1 Γενικά

Όπως έχει αναφερθεί, η τελική εφαρμογή που αναπτύσσεται χρησιμοποιεί ως δεδομένα τις θέσεις των αρθρώσεων των χεριών των χρηστών της εφαρμογής. Τα δεδομένα αυτά λαμβάνονται από το ενσωματωμένο σύστημα skeleton tracking του Kinect. Γι' αυτό το λόγο, προτού αναλυθούν οι βασικές αρχές του αλγορίθμου που χρησιμοποιούνται για το σκοπό αυτό από το Kinect, κρίνεται απαραίτητο να γίνει μία ανασκόπηση των κυριότερων μεθόδων που έχουν αναπτυχθεί με σκοπό την εκτίμηση ανθρώπινης πόζας – καθώς η εκτίμηση της στάσης των χεριών μπορεί να θεωρηθεί τμήμα του προβλήματος αυτού.

Δεδομένου ότι το πρόβλημα της εκτίμησης πόζας, συχνά, θεωρεί ως προαπαιτούμενο την επίλυση του προβλήματος της αναγνώρισης ανθρώπων, θα αναφερθούμε αρχικά στις βασικότερες μεθόδους που αναπτύχθηκαν για την επίλυση αυτού του προβλήματος. Καθώς, με τη σειρά της, η αναγνώριση ανθρώπων μπορεί να θεωρηθεί υποσύνολο του γενικότερου προβλήματος του εντοπισμού αντικειμένων (object detection), αρκετές ερευνητικές προσεγγίσεις έχουν δώσει έμφαση στην εύρεση κατάλληλων περιγραφητών, οι οποίοι – ιδανικά – αποτελούν μία αναπαράσταση αντικειμένων, κατάλληλη για χρήση σε εφαρμογές ταξινόμησης. Ο γνωστότερος από αυτούς είναι τα Ιστογράμματα Προσανατολισμένων Παραγώγων (Histograms Of Oriented Gradients, HoGs), τα οποία εισήχθησαν στη βιβλιογραφία από τους Dalal et al. [27], και πρακτικά αποτελούν μοντελοποίηση της κατεύθυνσης των ακμών ενός αντικειμένου. Ένας ακόμα περιγραφητής για αντικείμενα είναι ο Μετασχηματισμός Χαρακτηριστικών Αμετάβλητος ως προς την Κλίμακα (Scale-Invariant Feature Transform, SIFT), που εισήχθη στη βιβλιογραφία από τον Lowe [28], περιγράφει αντικείμενα βάσει των σημείων ενδιαφέροντος (keypoints) τους, και βελτιώσεις του χρησιμοποιούνται ακόμα και σήμερα σε εφαρμογές.

Οι εργασίες των Felzenswalb et al. [29] [30] έκαναν δημοφιλή τα Παραμορφώσιμα Μοντέλα Τμημάτων (Deformable Part Models, DPMs) και τις φωτογραφικές δομές (pictorial structures), χρησιμοποιώντας τα HOGs όχι ως ολικό περιγραφητή, αλλά κατασκευάζοντας πολλούς ταξινομητές για διάφορα τμήματα ενός αντικειμένου, και εισάγοντας σχέσεις εξάρτησης μεταξύ τους. Αυτά είχαν ως αποτέλεσμα την επιτάχυνση της διαδικασίας αναγνώρισης και τη βελτίωση της ακρίβειας, μέσω της εξαγωγής δεδομένων και για τις σχετικές θέσεις των τμημάτων ενός αντικειμένου.

Οι παραπάνω εργασίες, ωστόσο, στόχευαν κατά κύριο λόγο στον εντοπισμό αντικειμένων και όχι στην εξαγωγή πόζας (καθώς τα αποτελέσματα των ανιχνευτών τμημάτων χρησιμοποιούνταν κυρίως ως στοιχείο ύπαρξης ή όχι αντικειμένου). Από τις πρώτες εργασίες που ενοποίησαν το παραπάνω πλαίσιο με τους υπάρχοντες

αλγορίθμους εκτίμησης πόζας ήταν αυτές των Andriluka et al.[31] [20]. Στην περίπτωση αυτή, το προτεινόμενο πλαίσιο για εντοπισμό και εκτίμηση πόζας αποτελείται από ισχυρούς αυτοτελείς ανιχνευτές για τα διάφορα μέλη του σώματος στο κατώτερο επίπεδο, και ένα πιθανοτικό μοντέλο που περιγράφει τις συνδέσεις μεταξύ των μελών στο υψηλότερο επίπεδο. Μία επιπλέον επέκταση στο παραπάνω προσέφερε η εργασία των Yang et al. (2011) [32], καθώς η αναζήτηση για τα διάφορα μέλη του σώματος σε διάφορους προσανατολισμούς (που απαιτούσε πολλά ταιριάσματα μεταξύ του εκπαιδευμένου detector και της εικόνας) αντικαταστάθηκε από τη χρήση άνω του ενός (αλλά λιγότερων) ανιχνευτών ανά μέλος, ο καθένας «εξειδικευμένος» σε διαφορετική διεύθυνση.

Σε διαφορετική κατεύθυνση από τις παραπάνω εργασίες κινείται η εργασία των Ivekovic et al. [33], οι οποίοι χρησιμοποιούν τον αλγόριθμο Βελτιστοποίησης Σμήνους Σωματιδίων (Particle Swarm Optimization, PSO), με στόχο την εκτίμηση στάσης του άνω σώματος (upper body pose estimation) σε βίντεο που λαμβάνεται από πολλαπλές όψεις. Για το σκοπό αυτό, θεωρούν δεδομένο ένα αρχικοποιημένο μοντέλο της σκελετικής – κινηματικής αλυσίδας, και μέσω του PSO βελτιστοποιούν την εκτίμησή τους σε κάθε frame του βίντεο, χρησιμοποιώντας ως παράμετρο το μοντέλο δέρματος ώστε να ευθυγραμμίσουν με αυτό την κινηματική αλυσίδα.

Αρκετές από τις πρόσφατες ερευνητικές εργασίες στον τομέα αυτό έχουν εκμεταλλευθεί την ανάπτυξη πιο σύνθετων αρχιτεκτονικών νευρωνικών δικτύων, η οποία έγινε δυνατή χάρη στην αύξηση των υπολογιστικών πόρων και των διαθέσιμων δεδομένων. Ενδεικτικά αναφέρουμε την εργασία των Toshev et al. [34], οι οποίοι χρησιμοποιούν μία αρχιτεκτονική Συνελικτικού Νευρωνικού Δικτύου (Convolutional Neural Network, CNN) παρόμοια με αυτή που αναπτύχθηκε το 2012 από τους Krizhevsky et al. [35], αυτή των Tompson et al. [36], οι οποίοι χρησιμοποιούν ως τελικό ταξινομητή τη σύνθεση ενός CNN και ενός Τυχαίου Πεδίου Markov (MRF) για εκτίμηση πόζας σε εικόνες, καθώς και αυτήν των Pfister et al. [37], οι οποίοι χρησιμοποιούν μία αρχιτεκτονική καθαρά βασισμένη σε CNNs για την επίλυση του ίδιου προβλήματος σε βίντεο.

Τέλος, αναφέρουμε και την εργασία των Kostrikov et al. [38], οι οποίοι χρησιμοποιούν τυχαία δάση απόφασης (regression random forests) για την εύρεση της τρισδιάστατης πόζας από RGB (διδιάστατες, δηλαδή) εικόνες, σε συνδυασμό με τις προαναφερθείσες pictorial structures.

2.1.2 Χρήση Δεδομένων Βάθους

Παράλληλα, τμήμα της έρευνας επικεντρώθηκε στη χρήση δεδομένων βάθους (που μπορούν να μετασχηματιστούν σε απόσταση), τα οποία λαμβάνονται από κατάλληλους αισθητήρες. Τα βασικά πλεονεκτήματα που προκύπτουν από τη χρήση δεδομένων βάθους αφορούν στο ότι αυξάνουν τη διακρισιμότητα μεταξύ μη γειτονικών αντικειμένων με παραπλήσιο χρώμα, καθώς και στο ότι προσθέτουν μία τρίτη διάσταση,

αναπαριστώντας το χώρο με τη μορφή νεφών σημείων (point clouds). Αυτό διευκολύνει αρκετά το πρόβλημα του εντοπισμού πόζας, καθώς μπορούμε τόσο να εξάγουμε ευκολότερα τη σιλουέτα του ανθρώπου – είτε με κατάλληλη προεπεξεργασία των δεδομένων, είτε μέσω ταιριάσματος τρισδιάστατων μοντέλων για τα διάφορα μέλη του σώματος, όπως στην εργασία των Xia et al. [39].

Οι ερευνητικές προσπάθειες διακρίνονται σε δύο βασικές κατηγορίες: αυτές τις οποίες – κατόπιν αρχικοποίησης – αντιμετωπίζουν το πρόβλημα της εκτίμησης της στάσης του σώματος ως πρόβλημα τοπικής βελτιστοποίησης, και συνεπώς σε κάθε επανάληψη εκτέλεσης του αλγορίθμου βελτιστοποιούν την εκτίμησή τους, και σε αυτές που προσεγγίζουν το πρόβλημα ως πρόβλημα εκμάθησης δεδομένων. Η πρώτη κατηγορία, εφόσον επιλύει ένα πρόβλημα τοπικής βελτιστοποίησης, εξαρτάται από την καλή αρχικοποίησή της, κάτι που την κάνει αρκετά βολική για χρήση σε βίντεο (όπου μπορούμε να λαμβάνουμε ως αρχική εκτίμηση στάσης το προηγούμενο frame, χωρίς μεγάλη απόκλιση), υπό τη συνθήκη ότι δε θα έχουμε απότομες μεταβολές, ενώ η δεύτερη μπορεί να επιλύσει το ίδιο πρόβλημα και σε ακίνητες εικόνες (still images).

Σε ό,τι αφορά την πρώτη προσέγγιση, αρκετές από τις πρώιμες προσπάθειες στον τομέα αυτό (Bray et al. [40], Demirdjian et al. [41], Grest et al. [42]) αντιμετώπισαν το πρόβλημα της εξαγωγής στάσης σώματος από δεδομένα βάθους ως πρόβλημα βελτιστοποίησης υπό περιορισμούς, χρησιμοποιώντας έναν επαναληπτικό αλγόριθμο Πλησιέστερου Σημείου (Iterative Closest Point, ICP). Ο αλγόριθμος αυτός στοχεύει στην εύρεση του κατάλληλου μετασχηματισμού (υπέρθεση περιστροφής και μεταφοράς) μεταξύ της στάσης του σώματος και της αρχικής μας υπόθεσης για αυτήν.

Ενδιαφέρον παρουσιάζει η εργασία των Baak et al. [43], η οποία χρησιμοποιεί μία συγχώνευση των δύο προσεγγίσεων που αναφέρθηκαν στην εισαγωγή του κεφαλαίου. Συγκεκριμένα, από κάθε frame ενός video, εξάγεται καταρχήν η ανθρώπινη σιλουέτα – ως σύνολο σημείων στο χώρο (x, y, z), και με εφαρμογή μίας παραλλαγής του αλγορίθμου του Dijkstra [44], βρίσκονται τα γεωδαιτικά ακρότατα της σιλουέτας, τα οποία συνήθως αντιστοιχούν στις θέσεις του κεφαλιού, των χεριών και των ποδιών, και με χρήση μιας βάσης δεδομένων αντίστοιχων τοπικών ακροτάτων, εξάγεται μία υποψήφια πόζα – ως η πλησιέστερη. Παράλληλα με το παραπάνω, εξάγεται και μία δεύτερη υποψήφια πόζα με τεχνικές τοπικής βελτιστοποίησης επί της πόζας που βρέθηκε στο προηγούμενο frame, και τελικά επιλέγεται, με χρήση μετρικών απόστασης, η βέλτιστη από τις δύο.

Παρόμοιο συνδυασμό των δύο προσεγγίσεων χρησιμοποιούν και οι Ye et al. [45], οι οποίοι αφού πρώτα προεπεξεργαστούν τα δεδομένα βάθους, εξάγουν από αυτά τη σιλουέτα, δημιουργώντας περιστροφές του για να πετύχουν μη-μεταβλητότητα στην οπτική γωνία, και εν συνεχεία συγκρίνουν τα προκύπτοντα point clouds με μία βάση δεδομένων αντιστοίχων, σε έναν υποχώρο μικρότερης διάστασης που προέκυψε με εφαρμογή PCA. Η τελική πόζα βελτιστοποιείται μέσω ταιριάσματος του βέλτιστου

point cloud στην αρχική σιλουέτα. Αυτό γίνεται με χρήση του αλγορίθμου Συναφών Ολισθαινόντων Σημείων (Coherent Drifting Point algorithm, CDP), ο οποίος αποτελεί μία βελτίωση επί του ICP και εισήχθη στη βιβλιογραφία από τους Myronenko et al. [46]. Παρά το γεγονός ότι παρουσιάζουν καλύτερα αποτελέσματα σε σχέση με τον αλγόριθμο που χρησιμοποιείται στο Kinect, τονίζεται ότι ο αλγόριθμος δεν είναι κατάλληλος για χρήση σε εφαρμογές πραγματικού χρόνου.

Τέλος, κατόπιν της κυκλοφορίας του Kinect, αρκετές ερευνητικές προσπάθειες στράφηκαν από την ανάπτυξη αλγορίθμων για εκτίμηση της ολικής στάσης σώματος – τομέας στον οποίο το ενσωματωμένο σύστημα του Kinect θεωρείται ακόμα state of the art – στην επίλυση του ίδιου προβλήματος για μεμονωμένα μέλη του σώματος, για παράδειγμα της εκτίμησης της στάσης των χεριών με εμφανή εφαρμογή σε διαδραστικά συστήματα αναγνώρισης χειρονομιών. Εδώ, η πλειονότητα των εργασιών χρησιμοποιεί κάποιες βελτιωμένες εκδοχές του αλγορίθμου τυχαίων δασών για τη λήψη απόφασης, όπως στις περίπτωσεις της εργασίας των Xu et al. [47] και αυτής των Tang et al. [48].

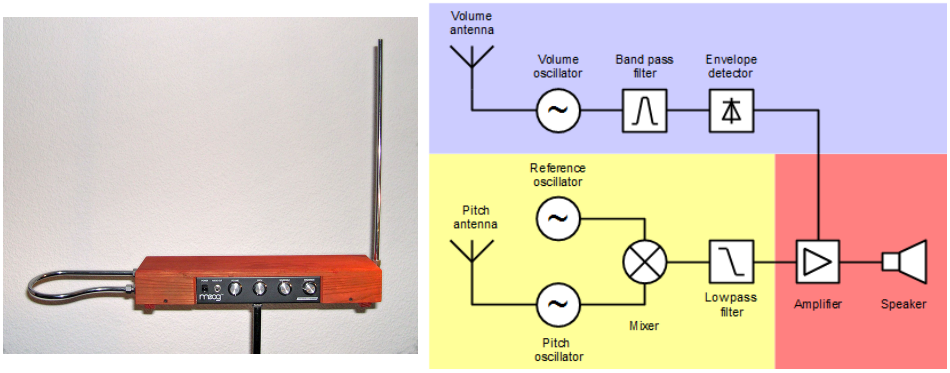
2.2 Σύνθεση Μουσικής από Κίνηση

2.2.1 Ιστορική Αναδρομή

Μία από τις αναρίθμητες δυνατότητες που έχει προσφέρει η σύγχρονη τεχνολογική πρόοδος αφορά στην αποσύνδεση της σύνθεσης μουσικής από τη χρήση μουσικών - με την παραδοσιακή έννοια - οργάνων. Αυτό μπορεί να γίνει είτε με την ευθεία χρήση του ηλεκτρονικού υπολογιστή ως εργαλείου σύνθεσης - μεγάλο μέρος της σύγχρονης ηλεκτρονικής μουσικής παράγεται από μονομελή συγκροτήματα, που επεξεργάζονται υπάρχοντα ηχητικά samples, τα οποία αναπαριστούν ήχους μουσικών οργάνων ή οτιδήποτε άλλο, είτε με τη χρήση πιο απτών μέσων ως εργαλείων σύνθεσης. Παραδείγματα της δεύτερης περίπτωσης αποτελούν η χρήση εικόνας ή κειμένου, συχνά μέσω της χαρτογράφησης κάποιου ενδιάμεσου χαρακτηριστικού (όπως το συναίσθημα) τόσο σε μουσικά κομμάτια, όσο και χαρακτηριστικών της εικόνας ή του κειμένου, καθώς και η χρήση της κίνησης για τη σύνθεση μουσικής [49].

Ωστόσο, η σύνθεση μουσικής μέσω κίνησης ήταν εφικτή πολύ πριν την εφεύρεση του σύγχρονου ηλεκτρονικού υπολογιστή. Συγκεκριμένα, κατά τη δεκαετία του 1920, ο Σοβιετικός φυσικός Lev Termen (ευρύτερα γνωστός ως Leon Theremin) υλοποίησε για πρώτη φορά το ομώνυμο μουσικό όργανο, η ιστορική σημασία του οποίου έγκειται στο ότι ήταν το πρώτο μουσικό όργανο το οποίο παρήγαγε ήχο χωρίς φυσική επαφή.

Η αρχή λειτουργίας του είναι σχετική απλή: Κάθε theremin είναι εξοπλισμένο με δύο κεραίες, μία σε κάθε άκρο του. Ο παίκτης του theremin μπορεί να ελέγχει τόσο τη συχνότητα του παραγόμενου σήματος εξόδου - το οποίο αποτελεί ημιτονοειδές μεταβαλλόμενης συχνότητας - όσο και την έντασή του, αναλόγως με την απόσταση



Σχήμα 2.1: α) Ένα theremin. Εμφανείς οι δύο κεραίες, οι οποίες δημιουργούν δύο χωρητικότητες με τα χέρια του παίκτη, επιτρέποντας την παραγωγή ηχητικών σημάτων. β) Μπλοκ διάγραμμα του εσωτερικού ενός theremin. [3]

των χεριών του από κάθε κεραία. Συγκεκριμένα, δεχόμενοι ότι μεταξύ των κεραιών και των χεριών του παίκτη δημιουργούνται δύο πυκνωτές, έχουμε τη λειτουργία δύο LC ταλαντωτών στο κύκλωμα του theremin. Σε ότι αφορά τη συχνότητα, το σήμα που παράγεται από τον αντίστοιχο LC ταλαντωτή πολλαπλασιάζεται με ημιτονοειδές δεδομένης συχνότητας αναφοράς, και στη συνέχεια φιλτράρεται προκειμένου να διατηρηθεί η βαθυπερατή συνιστώσα του προκύπτοντος σήματος. Σχετικά με την ένταση, το σήμα που παράγεται από τον αντίστοιχο LC ταλαντωτή φιλτράρεται ζωνοπερατά αυτή τη φορά, και με χρήση ενός κυκλώματος που δίνει ως έξοδο την περιβάλλουσα του κυκλώματος εισόδου (φωρατής περιβάλλουσας) ρυθμίζεται η ένταση.

Βέβαια, από τότε, έχουν αναπτυχθεί ποικίλες άλλες μέθοδοι, προκειμένου η κίνηση να μεταφράζεται μουσικά σε ήχο. Παραδείγματα αποτελούν η χρήση ηλεκτρονικών αισθητήρων, οι οποίοι επιτρέπουν τη χρήση του χώρου ως μουσικό όργανο, και η εξαγωγή χαρακτηριστικών κίνησης, είτε με χρήση γυρόμετρων και επιταχυνσιόμετρων [50], είτε με χρήση απομακρυσμένου αισθητήρα κίνησης [51]. Διαφοροποιήσεις, επίσης, υφίστανται στο αν η χρήση της κίνησης λειτουργεί πρωτογενώς ως συνθετικό μέσο, ή συμπληρωματικά, επηρεάζοντας λόγου χάρη την ταχύτητα της μουσικής αναλόγως με το πόσο έντονη είναι.

Πριν προχωρήσουμε στην παρουσίαση σχετικών εφαρμογών, ας κάνουμε μία αναφορά στα formats με τα οποία μπορεί να αποθηκευτεί και να αναπαραχθεί, ηλεκτρονικά, ήχος. Πέραν της απευθείας αποθήκευσης του παραγόμενου ήχου, σε αρχεία τύπου .wav, συχνά χρησιμοποιούμε αρχεία που ακολουθούν το πρωτόκολλο MIDI. Αρχεία που ακολουθούν το πρωτόκολλο αυτό δεν αποθηκεύουν τον ήχο αυτό καθ' αυτό, αλλά μία συμβολική αναπαράσταση μουσικών γεγονότων, όπως η ένταση, η συχνότητα και η διάρκεια κάθε νότας, καθώς και το αντίστοιχο μουσικό όργανο. Το καθένα από τα παραπάνω γεγονότα κωδικοποιείται με τη χρήση ενός 7-bit αριθμού, από το 0 ως το 127. Αρκετά βολική, επίσης, είναι η αναπαράσταση μίας μουσικής κυματομορφής

ως άθροισμα ημιτονοειδών (μέσω ανάλυσης Fourier), καθώς μειώνει σημαντικά τον αριθμό των προς αποθήκευση παραμέτρων.

2.2.2 Σχετικές Εφαρμογές

Τα τελευταία χρόνια, έχει αναπτυχθεί πληθώρα εφαρμογών που, χρησιμοποιώντας ως είσοδο δεδομένα κίνησης, παράγουν κάποιο είδος μουσικής εξόδου, τόσο στον ακαδημαϊκό χώρο όσο και ως εμπορικές εφαρμογές. Σε ό,τι αφορά τις εμπορικές εφαρμογές, συνήθως αποτελούν ανεξάρτητα projects, των οποίων η χρηματοδότηση επιχειρήθηκε μέσω crowdfunding (και, αρκετά συχνά, δε μετουσιώθηκαν ποτέ σε τελικό προιόν λόγω της μη εκπλήρωσης του στόχου). Ενδεικτικά αναφέρουμε τα παρακάτω:

- Point Motion [52]: Αποτελεί προιόν της ομώνυμης εταιρίας στη Βοστώνη. Κατά το αρχικό concept, προσέφερε δύο δυνατές λειτουργίες: τόσο την απευθείας σύνθεση μουσικής, μέσω σύνδεσης κινήσεων σε συγκεκριμένα μουσικά συμβάντα, όσο και τη δυνατότητα προσθήκης special effects σε υπάρχοντα μουσικά κομμάτια, μέσω συγκεκριμένων κινήσεων. Σε επίπεδο software, η εφαρμογή χρησιμοποιεί upper body skeleton tracking για την εκτίμηση της στάσης του σώματος, και κατά συνέπεια μπορεί να λειτουργήσει με οποιαδήποτε ενσωματωμένη σε υπολογιστή κάμερα. Το τελικό προιόν χρησιμοποιείται σε εφαρμογές ψηφιακής υγείας.

- Motus [50]: Προέρχεται από το T2M Creative Lab, με έδρα τη Λιθουανία. Προσφέρει τη δυνατότητα χρήσης 8 εικονικών οργάνων, αντιστοιχίζοντας κινήσεις σε νότες. Σε αντίθεση με την παραπάνω εφαρμογή, δε χρησιμοποιεί δεδομένα κάμερας για την αναγνώριση των διαφόρων κινήσεων, αλλά απευθείας αισθητήρες κίνησης - επιταχυνσιόμετρα, γυρόμετρα, μαγνητόμετρα και υψόμετρα.

- Interactive Music Battle [51]: Αναπτύχθηκε από την Phonotonic. Λειτουργεί με 2 παίκτες, και χρησιμοποιώντας, και σε αυτήν την περίπτωση, άμεσους αισθητήρες κίνησης, μετατρέπει τις κινήσεις σε ήχο. Με στόχο την εξοικείωση των παικτών με τις κινήσεις, η εφαρμογή περιλαμβάνει και ένα εκπαιδευτικό learning mode. Από τις εφαρμογές που αναφέρουμε εδώ, είναι και η μόνη που κατάφερε να μετατραπεί σε κάποιο τελικό εμπορικό προιόν, διατηρώντας τον αρχικό της προσανατολισμό ως ψυχαγωγική εφαρμογή.

Σε ό,τι αφορά σχετικές εφαρμογές που έχουν αναπτυχθεί σε ακαδημαϊκό περιβάλλον, ένα μεγάλο τμήμα τους [53] [54] χρησιμοποιεί το Kinect, με στόχο την εξαγωγή μετρικών σχετικών με τη θέση ή την ταχύτητα των χεριών των παικτών και την αντιστοίχισή τους σε διάφορα MIDI events. Παρόμοια προσέγγιση ακολουθείται στο

[55], στο οποίο δεδομένα σχετικά με τη θέση των χεριών του παίκτη μπορούν να αντιστοιχιστούν σε συγκεκριμένα μουσικά όργανα (synthesizer ή drums), με επιπλέον δυνατότητες ρύθμισης της έντασης της μουσικής και της συχνότητας του synthesizer, θυμίζοντας έτσι τη διαδικασία σύνθεσης μέσω theremin.

Ωστόσο, οι προαναφερθείσες εργασίες δε λαμβάνουν υπόψη την εγγενή καθυστέρηση που προκαλεί η μεταφορά δεδομένων από το Kinect, στον υπολογιστή και από εκεί στο MIDI controller. Αυτό επιχειρεί να επιλύσει το [56], στο οποίο με χρήση πρώτων και δεύτερων ροπών προβλέπονται οι μελλοντικές θέσεις των χεριών, με στόχο την εύρυθμη λειτουργία ενός εικονικού προσομοιωτή drum kit.

Σε διαφορετικό πλαίσιο κινείται τόσο το [57], στο οποίο, μέσω Kinect, παρακολουθούνται οι κινήσεις των παικτών της εφαρμογής οι οποίες και προσομοιάζουν - όπως και στο [55] - παίξιμο εικονικών μουσικών οργάνων, αλλά σε μεγαλύτερη ποικιλία (καλύπτεται τόσο κιθάρα, όσο και drums), όσο και το [58]. Η εργασία αυτή παρουσιάζει ενδιαφέρον, από την άποψη ότι επιχειρείται η αντιστοίχιση συγκεκριμένων δυναμικών χειρονομιών σε συγκεκριμένα ηχητικά γεγονότα, με κριτήριο την χωρική και ηχητική τους συνάφεια αντίστοιχα. Για το σκοπό αυτό, χρησιμοποιούνται δύο όμοια SOM με διδιάστατο grid, για τη χαρτογράφηση των χειρονομιών και των ηχητικών συμβάντων αντίστοιχα, και αναλόγως του αποτελέσματος της χαρτογράφησης γίνεται η αντιστοίχιση (η κάθε χειρονομία αντιστοιχίζεται στο ηχητικό συμβάν που έχει χαρτογραφηθεί στη θέση του grid με τις ίδιες συντεταγμένες).

Αξίζει να επισημάνουμε, τέλος, πέραν της εμφανούς ψυχαγωγικής χρήσης των παραπάνω εφαρμογών, ότι η σύνδεση της μουσικής με πιο απτές έννοιες, όπως η κίνηση, μπορεί να προσφέρει και εκπαιδευτικά, μέσω της εκμάθησης αφηρημένων εννοιών που σχετίζονται με τη μουσική ως νοητές επεκτάσεις των απτών εννοιών. Μεταξύ άλλων, με το ζήτημα αυτό ασχολείται η εργασία των Antle et al. [59], στην οποία μελετάται το κατά πόσο η σύνδεση μουσικών εννοιών με ορισμένες κινήσεις, μέσω συγκεκριμένης χαρτογράφησης, βοηθάει στην κατανόηση των πρώτων. Σημαντικό αποτέλεσμα της έρευνας αυτής είναι ότι η κατανόηση ήταν ευκολότερη, στις περιπτώσεις όπου η χαρτογράφηση σχετιζόταν με χωρικές συνδέσεις, παρά με κινήσεις του σώματος.

2.3 Ταξινόμηση Χειρονομιών από Σκελετικά Δεδομένα

Όπως αναφέρθηκε στην ενότητα 2.1.2, από την κυκλοφορία του Kinect v1 (2010) και έπειτα, τμήμα της επιστημονικής κοινότητας έστρεψε την προσοχή του στην προσπάθεια αξιοποίησης των σκελετικών δεδομένων που παρεχόταν, προκειμένου να εξαχθεί κάποιο ανώτερο επίπεδο πληροφορίας, σχετικά με την ανθρώπινη κίνηση, την αναγνώριση συγκεκριμένων στάσεων σώματος και χειρονομιών, κτλ.

Αρκετά δημοφιλής επιλογή αλγορίθμου, μεταξύ των ερευνητικών ομάδων, για το πρόβλημα της ταξινόμησης χειρονομιών, είναι η, κατόπιν μετατροπής των αρχικών σκελετικών δεδομένων σε κάποια ενδιάμεση αναπαράσταση, σύγκρισή τους με κάποιες πρότυπες χειρονομίες με χρήση κάποιας μετρικής απόστασης – συνήθως ευκλίδειας απόστασης κατόπιν εύρεσης του βέλτιστου μονοπατιού χρονικής στρέβλωσης. Η προσέγγιση αυτή αναφέρεται στις εργασίες των Raptis et al. [60], Reyes et al. [61], Celebi et al. [62] και Zhang et al. [63]. Η διαφορά αυτών των εργασιών έγκειται στην επιλογή της ενδιάμεσης αναπαράστασης που επιλέγεται. Συγκεκριμένα, στο [60] εξάγονται χαρακτηριστικά βασισμένα στις χρονικές συσχετίσεις μεταξύ των γωνιών των επιμέρους αρθρώσεων, και χρησιμοποιείται μία απόσταση τύπου dynamic time warping προκειμένου να προσδώσουν ευρωστία στο σύστημα, καθώς και να εξάγουν μία μετρική που δείχνει πόσο καλά εκτελέστηκε η χειρονομία (σε σύγκριση με μία πρότυπη χειρονομία, η οποία έχει προκύψει από τα δεδομένα εκπαίδευσης του μοντέλου). Η ενδιάμεση αναπαράσταση στο [63] αφορά στην εξαγωγή κάποιων επιμέρους κινήσεων μέσω των αποστάσεων μεταξύ των συνδέσμων, και την περιγραφή των – συνθετότερων – χειρονομιών ως χρονικές συνενώσεις των απλούστερων κινήσεων. Αντίθετα, τόσο στο [61] όσο και στο [62] δε χρησιμοποιείται κάποιου είδους ενδιάμεση αναπαράσταση, αλλά οι (x, y, z) συντεταγμένες διάφορων αρθρώσεων όπως παρέχονται από το Kinect. Για την ταξινόμηση, ωστόσο, χρησιμοποιούνται παραλλαγές της πολυδιάστατης εκδοχής του αλγορίθμου κοντινότερου γείτονα με χρήση δυναμικής χρονοστρέβλωσης [64], κατά την οποία το κάθε χαρακτηριστικό δε συμμετέχει εξίσου στην εξαγωγή του βέλτιστου μονοπατιού, αλλά με ένα βάρος που προκύπτει κατόπιν εκπαίδευσης (και, στην περίπτωση του [62], είναι διαφορετικό για κάθε χειρονομία).

Διαφορετικές προσεγγίσεις είναι αυτές που παρουσιάζονται στις εργασίες των Papadopoulos et al. [65], Negin et al. [66] και Hussein et al. [67]. Στην πρώτη περίπτωση εξάγονται ως χαρακτηριστικά οι γωνίες των σκελετικών αρθρώσεων και τις γωνιακές ταχύτητές τους, και χρησιμοποιούνται πλήρως συνδεδεμένα μοντέλα Markov, με τις μεταβλητές να μοντελοποιούνται ως μοντέλα μίξης Γκαουσιανών. Στις υπόλοιπες, χρησιμοποιούνται γραμμικές SVMs για την ταξινόμηση των επιμέρους χειρονομιών, αφού πρώτα έχουν εξαχθεί ενδιάμεσες αναπαραστάσεις. Οι αναπαραστάσεις αυτές στο [66] αποτελούνται από τα βέλτιστα γεωμετρικά χαρακτηριστικά, όπως έχουν βρεθεί με χρήση τυχαίων δασών, ενώ στο [67], από τους πίνακες συνδιασποράς μεταξύ των σκελετικών δεδομένων σε διαφορετικά frames.

3 Θεωρητικά Εργαλεία

Όπως αναφέραμε και στην εισαγωγή, σε αυτή την ενότητα καλύπτεται το απαιτούμενο υπόβαθρο από τους χώρους της Αναγνώρισης Προτύπων και της Θεωρίας Βελτιστοποίησης. Ο ενδιαφερόμενος αναγνώστης μπορεί να βρει επιπλέον πληροφορίες για τα ζητήματα που αναλύονται στα [68], [69], [70], [71].

3.1 Αλγόριθμοι Ομαδοποίησης

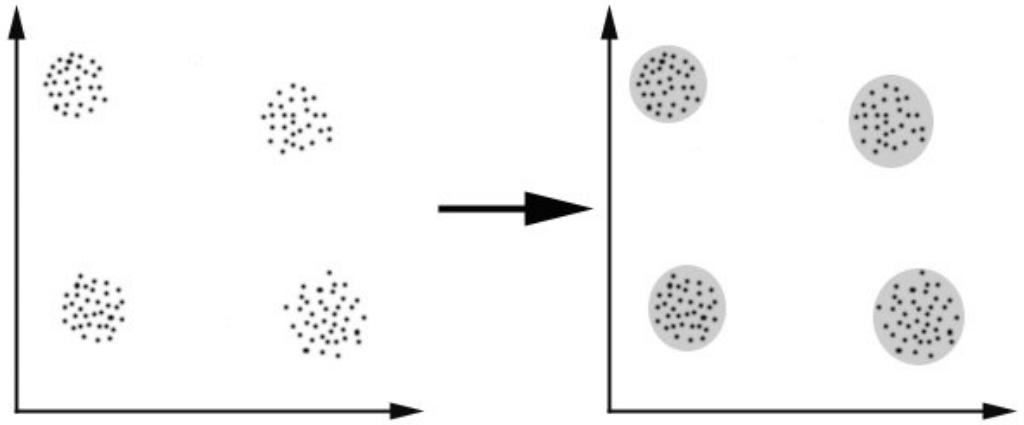
Όπως έχει αναφερθεί, τα προβλήματα της μηχανικής μάθησης διακρίνονται σε επιβλεπόμενα και μη επιβλεπόμενα. Στα προβλήματα επιβλεπόμενης μάθησης (supervised learning), συνήθως ο αλγόριθμος καλείται να αναπαράγει ένα μοντέλο – συνήθως κατηγοριοποίησης, ή παλινδρόμησης – το οποίο παρέχεται ρητά μέσω των δεδομένων εκπαίδευσης, τα οποία έχουν ταμπέλες (labels) που αναφέρουν την κατηγορία στην οποία ανήκουν. Αντίθετα, στα προβλήματα μη επιβλεπόμενης (unsupervised) μάθησης, το μοντέλο το οποίο καλείται να αναπαράγει ο αλγόριθμος δεν είναι ξεκάθαρο, καθώς τα δεδομένα εκπαίδευσης δεν έχουν labels.

Ένα από τα γνωστότερα προβλήματα μη επιβλεπόμενης μάθησης είναι το πρόβλημα της ομαδοποίησης ή συσταδοποίησης (clustering) δεδομένων. Στόχος, στην περίπτωση αυτή, είναι ο διαχωρισμός ενός συνόλου ετερογενών δεδομένων – τα οποία παρίστανται ως διανύσματα σε έναν ν-διάστατο χώρο χαρακτηριστικών – σε ομάδες/κατηγορίες, τα στοιχεία των οποίων είναι επιθυμητό να έχουν όσο γίνεται κοινά χαρακτηριστικά μεταξύ τους, και όσο γίνεται διαφορετικά χαρακτηριστικά από τα στοιχεία των υπολοίπων κατηγοριών. Τυπικά παραδείγματα είναι ο αλγόριθμος των κ-μέσων (k-means) και η εκπαίδευση ενός συνόλου k Γκαουσιανών, κατάλληλα αρχικοποιημένων.

Ο αλγόριθμος των κ-μέσων μπορεί να συνοψιστεί ως εξής: Έστω ένα σύνολο σημείων $X \in R^n$, τα οποία θέλουμε να ομαδοποιήσουμε σε κ κατηγορίες. Αρχικά, αρχικοποιούμε τυχαία κ σημεία στο χώρο R^n, τα οποία και είναι τα κέντρα (centroids) των κ clusters στο χώρο. Στη συνέχεια, επαναλαμβάνουμε τα εξής τρία βήματα μέχρι να υπάρξει σύγκλιση:

- Υπολογισμός της απόστασης κάθε σημείου $x_i \in X$ από όλα τα centroids.
- Ανάθεση (assignment) του κάθε σημείου $x_i \in X$ στο cluster, με κριτήριο την ελάχιστη απόσταση από το κέντρο του cluster.
- Επανυπολογισμός των κέντρων του κάθε cluster, ως μέσος όρος των σημείων του χώρου που ανήκουν σε αυτόν.

Συνήθως, ως κριτήριο σύγκλισης παίρνουμε την πραγματοποίηση Τ επαναλήψεων, τη σταθεροποίηση των centroids ανάμεσα σε διαδοχικές επαναλήψεις (καθώς αυτή θα συνεπάγεται σταθεροποίηση των clusters, δείγμα καλής ομαδοποίησης αν δεν έχουμε πέσει σε κάποιο τοπικό ελάχιστο), ή τη μείωση της διασποράς σε κάθε cluster κάτω



Σχήμα 3.1: Εφαρμογή αλγορίθμου ομαδοποίησης σε ένα διδιάστατο σύνολο δεδομένων. Όπως είναι εμφανές και από το αριστερά σχήμα, τα δεδομένα χωρίζονται σε 4 κατηγορίες. Το (αναμενόμενο) αποτέλεσμα εφαρμογής του αλγορίθμου φαίνεται στο δεξιά σχήμα. [4]

από μία προκαθορισμένη τιμή, καθώς αυτό συνεπάγεται ικανοποιητική ομαδοποίηση του χώρου, και κάθε cluster θα αποτελείται από παραπλήσια στο χώρο σημεία.

Στην περίπτωση του αλγορίθμου μίγματος Γκαουσιανών για ομαδοποίηση, μπορούμε και εδώ να δούμε δύο διακριτά βήματα: το βήμα προσδοκίας (Expectation step ή E-step) και το βήμα μεγιστοποίησης (Maximization step ή M-step). Συγκεκριμένα, η δομή του αλγορίθμου είναι η εξής:

- Αρχικοποίηση των a priori πιθανοτήτων για κάθε cluster (a_k), των μέσων τιμών (μ_k) και των πινάκων συνδιασποράς (Σ_k) για κάθε cluster (ή για κάθε Γκαουσιανή στην περίπτωση που μοντελοποιήσουμε κάθε cluster με άνω της μίας Γκαουσιανής). Στη συνέχεια, μέχρι τη σύγκλιση (η οποία και ορίζεται με παρόμοια κριτήρια με προηγουμένως), επαναλαμβάνουμε τα:

- E-step: Για όλα τα σημεία $x_i \in X$ υπολογίζουμε την πυκνότητα πιθανότητας για κάθε cluster $C_k \in C$, βάσει του τύπου της πολυμεταβλητής Gaussian κατανομής για κάθε cluster, λαμβάνοντας υπόψιν και την a priori πιθανότητα. Έπειτα, υπολογίζουμε τα βάρη w_{ik} , τα οποία δίνουν την πιθανότητα το σημείο $x_i \in C_k$. Αυτά εκφράζουν τις κανονικοποιημένες πυκνότητες πιθανότητας για κάθε cluster, και αν έχουμε N clusters, υπολογίζονται βάσει του τύπου:

$$w_{ik} = \frac{(P(x_i \in C_k))}{(\sum_{j=1}^{N}(x_i \in C_j))} \tag{3.1}$$

- M-step: Έστω M ο συνολικός αριθμός των σημείων προς ομαδοποίηση. Τότε ορίζουμε, ως το 'soft' άθροισμα των σημείων που ανήκουν σε ένα cluster:

$$M_k = \sum_{i=1}^{M}(w_{ik}) \tag{3.2}$$

Τότε, ανανεώνουμε τις a priori πιθανότητες κάθε cluster, τις μέσες τιμές, και τους πίνακες συνδιασποράς, βάσει των εξής τύπων:

$$a_{knew} = \frac{M_k}{M} \tag{3.3}$$

$$\mu_{knew} = \frac{1}{M_k}\sum_{i=1}^{M}(w_{ik}x_i) \tag{3.4}$$

$$\Sigma_{knew} = \frac{1}{M_k}\sum_{i=1}^{M}(w_{ik}d_{ik}d_{ik}^T), d_{ik} = x_i - \mu_{knew}, \in R^n \tag{3.5}$$

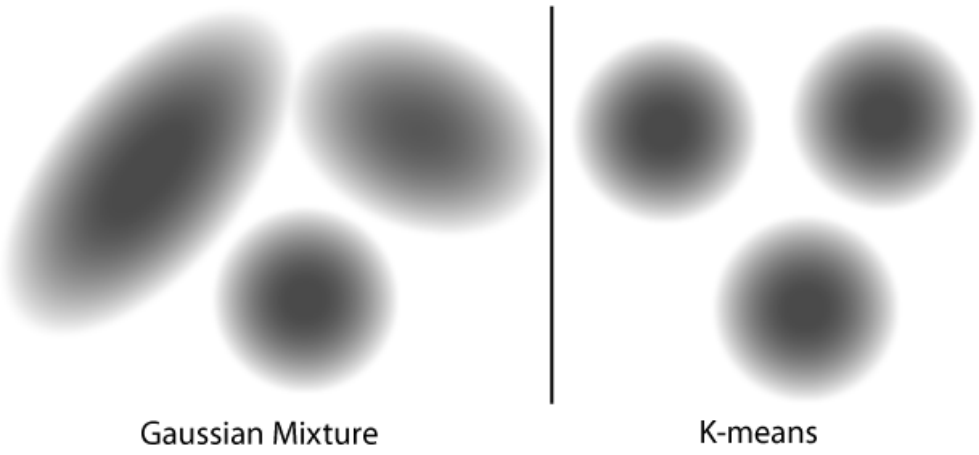
Το βασικό πλεονέκτημα αυτής της μεθόδου έναντι του αλγορίθμου των κ-μέσων είναι ότι ενώ τα clusters που προκύπτουν από τον k-means έχουν σφαιρικό σχήμα (καθώς κριτήριο ένταξης σε ένα cluster είναι η ευκλίδεια απόσταση), τα clusters που προκύπτουν από τον αλγόριθμο μίγματος Γκαουσιανών μπορούν να έχουν ελλιπτικό σχήμα, με διευθύνσεις των αξόνων παράλληλες στους άξονες (αν ο πίνακας συνδιασποράς είναι διαγώνιος) ή όχι (αν ο πίνακας συνδιασποράς έχει μη μηδενικά στοιχεία). Επιπλέον, ενώ ο αλγόριθμος των κ-μέσων οδηγεί σε σκληρή ομαδοποίηση, καθώς ένα σημείο μπορεί είτε να ανήκει είτε όχι σε ένα cluster, η χρήση μίγματος Γκαουσιανών για περιγραφή ενός συνόλου clusters μπορεί να οδηγήσει σε χαλαρή ομαδοποίηση, καθώς σημεία που βρίσκονται κοντά στα όρια δύο clusters κατανέμονται, εν μέρει, και στα δύο. Ωστόσο, είναι εξαρτώμενος από την καλή αρχικοποίηση των παραμέτρων του, και γι΄ αυτό συχνά τρέχουμε κάποιες επαναλήψεις του αλγορίθμου k-means για να παραχθούν οι αρχικές εκτιμήσεις των παραμέτρων.

3.2 Μετρικές Απόστασης

Έστω δύο σημεία x, y, στο χώρο R^n. Ως απόσταση των δύο σημείων ορίζουμε μία μετρική $d(x, y)$, η οποία υποδηλώνει πόσο απέχουν τα δύο σημεία μεταξύ τους – εναλλακτικά, σε τι βαθμό οι συντεταγμένες τους είναι όμοιες. Ιστορικά, έχουν αναπτυχθεί διάφορες μετρικές απόστασης μεταξύ δύο σημείων. Η πιο ευρέως διαδεδομένη είναι η – γνωστή σε όλους μας – Ευκλίδεια απόσταση, που ορίζεται ως:

$$d(x, y) = \sqrt{(x - y)^T(x - y)}. \tag{3.6}$$

Μία άλλη μετρική απόστασης, η οποία χρησιμοποιείται σε στατιστικές εφαρμογές – συγκεκριμένα, όταν ο στόχος είναι η εξαγωγή της απόστασης ενός σημείου από



Σχήμα 3.2: Σύγκριση του σχήματος των συστάδων που προκύπτουν από χρήση των αλγορίθμων μίγματος Γκαουσιανών (αριστερά) και κ-μέσων (δεξιά) σε διδιάστατα δεδομένα. Εμφανής η ευελιξία του σχήματος των συστάδων στην πρώτη περίπτωση. [5]

μία κατανομή με δεδομένα χαρακτηριστικά μ, Σ, είναι η απόσταση Mahalanobis. Σε αυτήν την περίπτωση, η απόσταση μεταξύ των σημείων ορίζεται ως:

$$d(x, \mu, \Sigma) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}. \tag{3.7}$$

Και για την οποία έχουμε να παρατηρήσουμε τα ακόλουθα:

- Η απόσταση Mahalanobis εκφυλίζεται στην Ευκλίδεια, στην περίπτωση που έχουμε μοναδιαίους πίνακες συνδιασποράς.
- Όσο μεγαλύτερη είναι η διασπορά μίας κατανομής, τόσο μικρότερη θα είναι η απόσταση μεταξύ της κατανομής και του υποψήφιου σημείου.

Ως μετρική απόστασης μπορεί να χρησιμοποιηθεί και η απόσταση συνημιτόνου, η οποία βρίσκει κυρίως εφαρμογές σε εφαρμογές εξόρυξης κειμένου (text mining). Προσόν της αποτελεί η υπολογιστική απλότητα, ειδικά στην περίπτωση όπου τα διανύσματα είναι κανονικοποιημένα, αφού μπορεί να υπολογιστεί ως:

$$d_{cos}(x, y) = 1 - cos(x, y) = 1 - \frac{(x^T y)}{(\|x\|\|y\|)}. \tag{3.8}$$

Ενδιαφέρον παρουσιάζει η σχέση μεταξύ της ευκλίδειας απόστασης δύο διανυσμάτων και της απόστασης συνημιτόνου μεταξύ τους, κατά την περίπτωση όπου τα προς σύγκριση διανύσματα έχουν το ίδιο μέτρο. Συγκεκριμένα, έστω $x, y \in R^n$, με $\|x\| = \|y\| = a$. Τότε:

$$d^2_{eucl}(x, y) = (x-y)^T(x-y) = x^T x + y^T y - 2\langle x, y\rangle = \|x\|^2 + \|y\|^2 - 2\|x\|\|y\| cos(x, y)$$

$$= 2a^2(1 - cos(x, y)) = 2a^2 d_{cos}(x, y)$$

Από όπου συμπεραίνουμε ότι, σε αυτή την περίπτωση, η απόσταση συνημιτόνου είναι ανάλογη του τετραγώνου της ευκλίδειας απόστασης, και άρα θα εξάγουν τα ίδια διαγράμματα Voronoi – και κατά συνέπεια, και τα ίδια με την ευκλίδεια απόσταση.

3.3 Αλγόριθμος Κοντινότερου Γείτονα

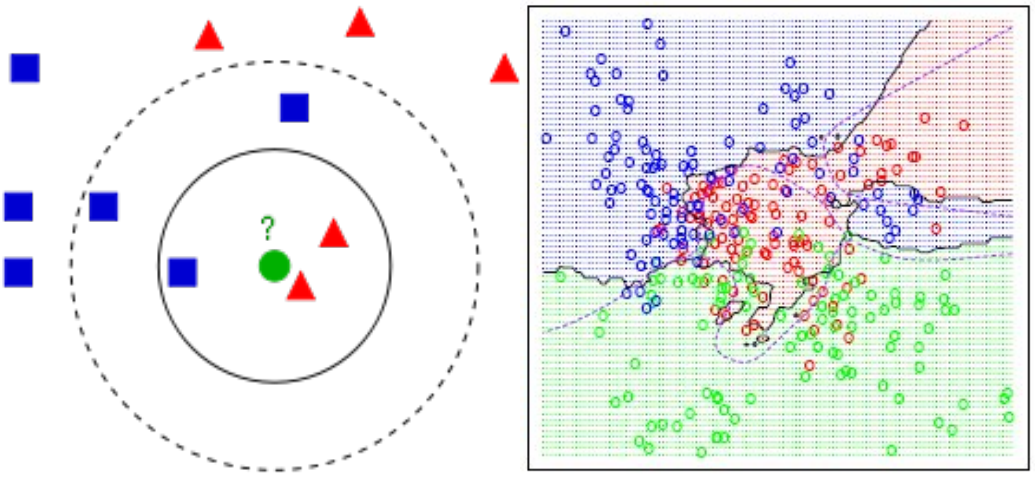
Ο αλγόριθμος του κοντινότερου γείτονα (nearest – neigbour algorithm) αποτελεί έναν αλγόριθμο μηχανικής μάθησης ο οποίος χρησιμοποιείται για ταξινόμηση. Διαισθητικά, μπορεί να εξηγηθεί εύκολα, καθώς αντιστοιχίζουμε κάθε προς ταξινόμηση αντικείμενο στο πλησιέστερό του. Φορμαλιστικά:

Έστω ένα σημείο στο χώρο R^n, έστω y_i, αγνώστου κατηγορίας, και ένα σύνολο κ σημείων X, έστω $x_i, i = 1 \dots k$ στο χώρο R^n, τα οποία γνωρίζουμε σε ποια κατηγορία ανήκουν. Για να ταξινομήσουμε σωστά το σημείο y_i, υπολογίζουμε την απόσταση του σημείου από όλα τα σημεία x_i, χρησιμοποιώντας κάποια από τις μετρικές απόστασης που αναφέρθηκαν στην προηγούμενη ενότητα.

Μία πιο σύνθετη, αλλά πιο εύρωστη σε θόρυβο στα αρχικά δεδομένα, παραλλαγή του αλγορίθμου αυτού είναι ο αλγόριθμος των κ κοντινότερων γειτόνων (k-nearest neighbor algorithm, ή knn). Συνοπτικά, αντί να ταξινομήσουμε το σημείο στην κατηγορία στην οποία ανήκει το πλησιέστερό του, λαμβάνουμε υπόψη την κατηγορία στην οποία ανήκουν τα κ πλησιέστερα σημεία. Η τελική ταξινόμηση γίνεται κατόπιν ψηφοφορίας των κ πλησιέστερων σημείων, όπου όλα τα σημεία έχουν ίσα βάρη, ή βάρη αναλόγως με την απόσταση.

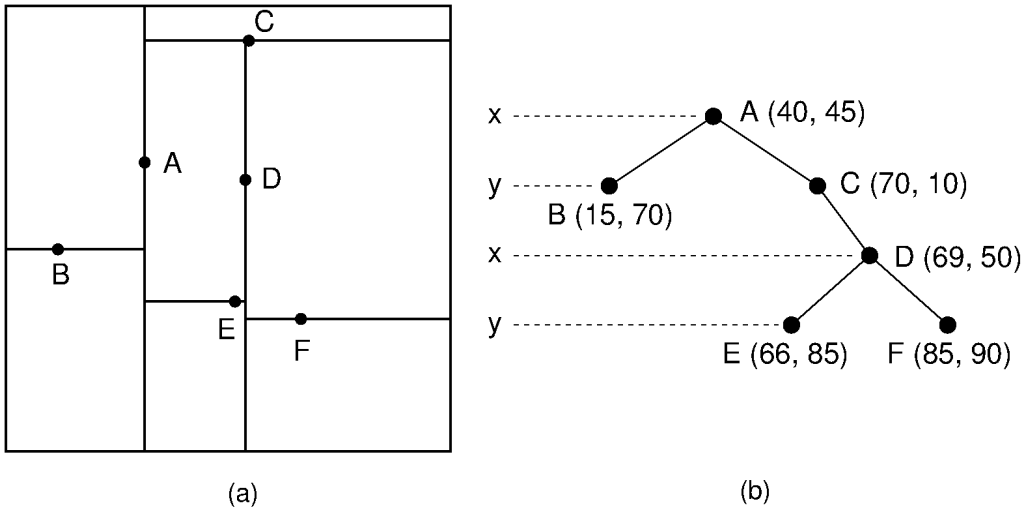
Μία ιδιαίτερη – και απλοποιημένη – υποπερίπτωση του αλγορίθμου αυτού είναι ο αλγόριθμος ταιριάσματος (template matching). Αυτός προκύπτει, αν απαιτήσουμε, από κάθε κατηγορία, να έχουμε μοναδικό δείγμα – έτσι, σε αυτήν την περίπτωση, ο χώρος δειγμάτων μας έχει διάσταση ίση με τον αριθμό των κατηγοριών στις οποίες ένα πρότυπο μπορεί να ταξινομηθεί.

Ο αλγόριθμος πλησιέστερου γείτονα μπορεί να πετύχει πολύ καλές επιδόσεις σε προβλήματα ταξινόμησης, στα οποία υπάρχουν αρκετά δεδομένα προκειμένου να περιγράψουν ικανοποιητικά τη δομή του χώρου. Αυτό είναι το μεγάλο του πλεονέκτημα αλλά και το μεγάλο του μειονέκτημα, καθώς – σε αντίθεση με άλλους αλγορίθμους εκπαίδευσης, των οποίων το μοντέλο είναι συγκεκριμένο και ο όγκος των δεδομένων επηρεάζει μόνο την ταχύτητα εκπαίδευσης – η ταχύτητα εκτέλεσης του αλγορίθμου εξαρτάται γραμμικά από τον αριθμό των δεδομένων εκπαίδευσης. Κατά συνέπεια, προκειμένου να επιτευχθεί επιταχυμένη εκτέλεση του αλγορίθμου, είναι απαραίτητο να χρησιμοποιηθεί κάποια τεχνική συμπύκνωσης (condensing) του αρχικού συνόλου σημείων, ώστε να διατηρηθεί κατά το μέγιστο η περιγραφή του χώρου R^n.



Σχήμα 3.3: α) Αριστερά: οπτικοποίηση του κανόνα απόφασης Κοντινότερων Γειτόνων για δυαδική ταξινόμηση. Αν ο αριθμός των γειτόνων που λαμβάνουμε υπόψη μας κ = 3, το σημείο μας ταξινομείται ως τρίγωνο, αν κ = 5, ως τετράγωνο. [6] β) Δεξιά, οι επιφάνειες απόφασης που προκύπτουν σε ενδεικτικό πρόβλημα ταξινόμησης σε 3 κατηγορίες. [7]

Για τη συμπύκνωση μπορούν να χρησιμοποιηθούν διάφορες μεθοδολογίες. Η μία από αυτές αφορά στην κατασκευή ενός κ-διάστατου δυαδικού δέντρου (kd-tree) από το αρχικό σύνολο δεδομένων. Τα kd-trees αποτελούν μία δομή δεδομένων αντίστοι-



Σχήμα 3.4: Αριστερά, ένας διδιάστατος χώρος σημείων, και δεξιά, ένα αντίστοιχο kd-tree. Η επιλογή των σημείων – κόμβων έγινε τυχαία, κάτι το οποίο οδηγεί σε μη βέλτιστο κατασκευαστικά δέντρο. [8]

χη ενός δυαδικού δέντρου, όπου το αρχικό επίπεδο αντιστοιχεί στο αρχικό σύνολο δεδομένων, και κάθε κόμβος διαχωρίζει το αρχικό σύνολο δεδομένων σε δύο υποσύνολα, κατά τέτοιο τρόπο ώστε να διατηρείται η χωρική συνεκτικότητα των δύο υποσυνόλων (σε κάθε επίπεδο, η γραμμή διαχωρισμού είναι αντίστοιχη με τον άξονα που αντιστοιχεί σε μία διάσταση). Έτσι, αν υποθέσουμε ότι, στο τελικό επίπεδο ενός kd-tree, κάθε κόμβος θα αντιστοιχεί σε μοναδικό σημείο του αρχικού συνόλου – το οποίο είναι μεγέθους N -, για την εκτέλεση του αλγορίθμου του κοντινότερου γείτονα, απαιτούνται – κατά μέσο όρο, καθώς δεν υπάρχει καμία εγγύηση ότι ο αλγόριθμος κατασκευής του δυαδικού δέντρου είναι βέλτιστος – συγκρίσεις με $log(N)$ σημεία, αντί για N σημεία, οδηγώντας σε σημαντικά βελτιωμένη πολυπλοκότητα ($O(logN)$). Το μειονέκτημα της μεθόδου αυτής είναι ότι δεν είναι εφαρμόσιμη σε σύνολα δεδομένων μεγάλης διάστασης: συγκεκριμένα, όπως αναφέρεται στην εργασία του Indyk [72], αν η διαστατικότητα του χώρου μας είναι d, και το σύνολο των σημείων N, πρέπει να ισχύει η συνθήκη $N \gg 2^d$. Σε τέτοιες περιπτώσεις, μπορούμε να πετύχουμε συμπύκνωση του δειγματικού χώρου με χρήση κάποιου αλγορίθμου ομαδοποίησης, όπως για παράδειγμα τον αλγόριθμο των κ-μέσων.

Εναλλακτικά, αντί να μειώσουμε το μέγεθος του δειγματικού χώρου, μπορούμε να μειώσουμε τη διάστασή του. Αυτό μπορεί να επιτευχθεί με χρήση διαφόρων μεθόδων μείωσης της διαστατικότητας, η γνωστότερη από τις οποίες είναι η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis, PCA). Η λογική πίσω από την εφαρμογή της μεθόδου PCA είναι ότι, αν έχουμε ένα ν-διάστατο σύνολο δεδομένων, και προσπαθήσουμε να ταιριάξουμε ένα ελλειψοειδές σε αυτά, τότε η κύρια πληροφορία που περιγράφει το σύνολο δεδομένων βρίσκεται στις διαστάσεις (άξονες) με μεγάλο μήκος. Αυτό, βέβαια, απαιτεί το σύνολο δεδομένων μας να έχει κανονικοποιηθεί.

Μαθηματικώς ορισμένα, έχουμε τα εξής: Έστω ένας πίνακας δεδομένων X, nxm , ο οποίος, δηλαδή, αποτελείται από n, διάστασης m, δείγματα. Αρχικά, μετασχηματίζουμε τα δεδομένα, ώστε κάθε διάσταση – χαρακτηριστικό να έχει μηδενική μέση τιμή και μοναδιαία διασπορά, προκειμένου να εξισώσουμε το σχετικό βάρος τους. Στη συνέχεια, υπολογίζουμε τον mxm πίνακα συνδιασποράς των δεδομένων, Σ, και υπολογίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματά του. Τότε, θα ισχύουν τα ακόλουθα:

Εφόσον ο πίνακας Σ (ως πίνακας συνδιασποράς) είναι συμμετρικός, και θετικά ορισμένος, οι ιδιοτιμές του θα είναι θετικές και τα ιδιοδιανύσματά του θα δημιουργούν έναν ορθογώνιο χώρο [1]. Εάν ορίσουμε P τον πίνακα των κανονικοποιημένων – ώστε να έχουν μέτρο ίσο με τη μονάδα – ιδιοδιανυσμάτων του Σ, τότε ο πίνακας $T = XP$ αποτελεί την προβολή των δεδομένων X στον χώρο με κύριους άξονες αυτούς που ορίζει ο P, οι οποίοι και είναι ορθογώνιοι μεταξύ τους. Ο καθένας από αυτούς τους άξονες θα συνεισφέρει στη συνολική διασπορά του συνόλου δεδομένων, αν ορίσουμε

[1] Περισσότερα σχετικά με αυτό, στο παράρτημα.

λ_i την αντίστοιχη ιδιοτιμή του ιδιοδιανύσματος – άξονα, κατά ένα βαθμό c_i:

$$c_i = \frac{\lambda_i}{\sum_{i=1}^{m}(\lambda_i)}. \tag{3.9}$$

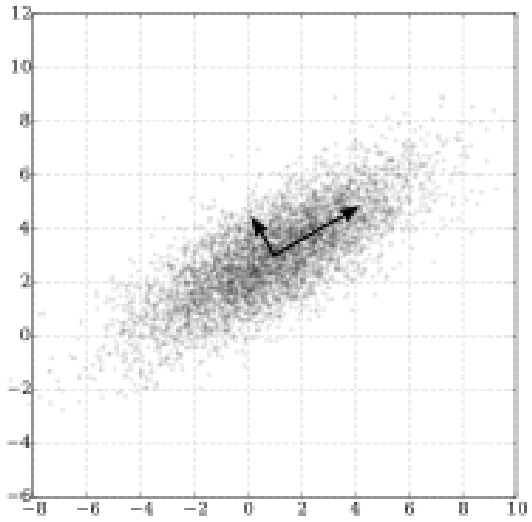
Συνεπώς, μπορούμε να διατηρήσουμε το μεγαλύτερο μέρος της διασποράς μεταξύ του συνόλου δεδομένων, διατηρώντας ορισμένες μόνο διευθύνσεις, οι οποίες και αντιστοιχούν σε ορισμένες στήλες του P. Αν P_{new} ο μετασχηματισμένος πίνακας P ώστε να έχει ως στήλες μόνο τα k ιδιοδιανύσματα τα οποία συνεισφέρουν στη διασπορά του συνόλου δεδομένων πάνω από ένα ορισμένο ποσοστό, τότε ο τελικός πίνακας $T = XP_{new}$, διαστάσεων nxk, αποτελεί την αναπαράσταση του αρχικού συνόλου δεδομένων σε έναν χώρο χαμηλότερης διαστατικότητας, έχοντας διατηρήσει το μεγαλύτερο μέρος της αρχικής πληροφορίας.

Ενδιαφέρον παρουσιάζει και η συμπεριφορά των μετρικών απόστασης μεταξύ των σημείων στο χώρο, κατόπιν της εφαρμογής του μετασχηματισμού PCA. Από τη στιγμή που οι νέοι πρωταρχικοί άξονες του χώρου, τον οποίον καθορίζει ο μετασχηματισμός, έχουν προκύψει από απλή περιστροφή του αρχικού χώρου R^n, χωρίς επιπλέον κλιμάκωση, η ευκλίδεια απόσταση μεταξύ των σημείων θα παραμένει ίδια. Επιπλέον, καθώς το μήκος ενός άξονα, όπως είδαμε, εξαρτάται από το μέτρο της ιδιοτιμής που αντιστοιχεί σε αυτόν, είναι εύλογο να συμπεράνουμε ότι ακόμα και διατήρηση ελάχιστων πρωτευόντων αξόνων θα διατηρήσει, σε μεγάλο βαθμό, ανεπηρέαστη την ευκλίδεια μετρική απόστασης.

3.4 Δυναμική Χρονοστρέβλωση

Σε προβλήματα ανάλυσης και σύγκρισης χρονοσειρών, ή εν γένει δεδομένων τα οποία έχουν χρονική συνάφεια και συσχέτιση, η χρήση του απλού αλγορίθμου κοντινότερου γείτονα εμφανίζει προβλήματα. Συγκεκριμένα, αν επιθυμούμε να συγκρίνουμε μία – έστω μονοδιάστατη – χρονοσειρά μήκους n, έστω $x \in R^n$, με ένα σύνολο m πρότυπων χρονοσειρών $S \in R^{mxn}$, ο αλγόριθμος κοντινότερου γείτονα συγκρίνει μεταξύ τους τα σημεία που αντιστοιχούν στις ίδιες χρονικές στιγμές, αμελώντας παραμέτρους όπως η γενικότερη μορφή της χρονοσειράς, η περίοδός της σε περίπτωση που είναι περιοδική, κτλ. Τα παραπάνω προβλήματα έρχεται να επιλύσει σε μεγάλο βαθμό ο αλγόριθμος της δυναμικής χρονοστρέβλωσης, ο οποίος αποτελεί έναν αλγόριθμο δυναμικού προγραμματισμού, που εφαρμοζόμενος σε δύο χρονοσειρές x_1, x_2 προσπαθεί να βρει το βέλτιστο μονοπάτι χρονικών στιγμών (optimal warping path), ώστε οι χρονοσειρές, κατόπιν της ευθυγράμμισης μέσω του μονοπατιού αυτού, να συμπίπτουν κατά το μέγιστο δυνατό βαθμό.

Φορμαλιστικά, έστω οι χρονοσειρές (μίας μεταβλητής) $x_1(t), x_2(t)$, μήκους ίσου με n. Θέλουμε α) να βρούμε μία μετρική απόστασης μεταξύ των χρονοσειρών που να λαμβάνει υπόψη της την πιθανή χρονική ευθυγράμμιση, και β) να «ευθυγραμμίσουμε» στο χρόνο τις δύο σειρές μεταβάλλοντας τη μία. Και για τα δύο προβλήματα, είναι



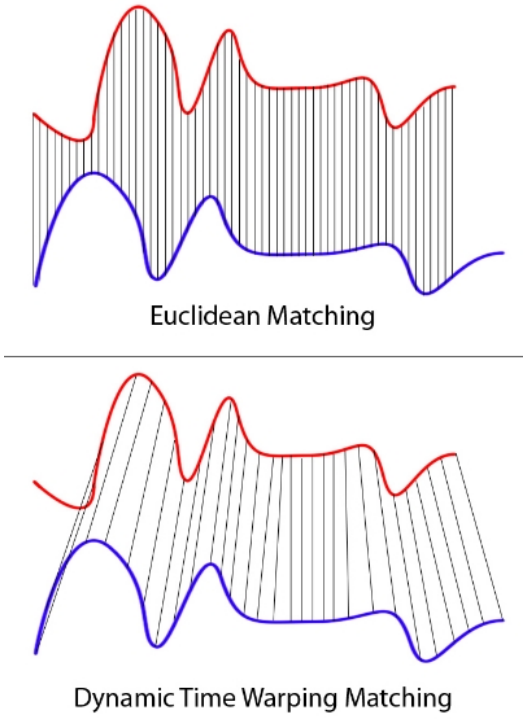
Σχήμα 3.5: Ενδεικτική απεικόνιση δισδιάστατου συνόλου δεδομένων. Φαίνονται οι διευθύνσεις των αξόνων που θα προκύψουν, κατόπιν εφαρμογής PCA, το μήκος των οποίων αντιστοιχεί στη συνεισφορά τους στη συνολική διασπορά των δεδομένων. [9]

απαραίτητη η εύρεση της μήτρας αποστάσεων, D, η οποία είναι τετραγωνική διάστασης nxn, με το σημείο $D(i,j)$ να δηλώνει τη συνολική, έχοντας λάβει υπόψη και τις προηγούμενες αποστάσεις, απόσταση μεταξύ των σημείων $x_1(i), x_2(j)$. Για τον υπολογισμό της, ακολουθούνται τα εξής βήματα.

- Αρχικοποιούνται όλες οι αποστάσεις στο άπειρο (πρακτικά σε μία πολύ μεγάλη τιμή), με εξαίρεση το σημείο $D(1,1) = 0$.
- Για κάθε ζεύγος τιμών (i,j), υπολογίζεται η «κανονική» απόσταση μεταξύ των σημείων $x_1(i), x_2(j), d(x_1(i), x_2(j))$, βάσει κάποιας μετρικής απόστασης.
- Στη συνέχεια, υπολογίζεται το συνολικό κόστος μεταξύ των σημείων $x_1(i)$, $x_2(j)$, ως:

$$D(i,j) = d(x_1(i), x_2(j)) + min(D(i,j-1), D(i-1,j-1), D(i-1,j)) \quad (3.10)$$

Η χρήση των σημείων αυτών του πίνακα για τον υπολογισμό της απόστασης $D(i,j)$ υπονοεί την τήρηση συνθηκών συνέχειας και μονοτονίας (προς τα εμπρός για τουλάχιστον μία από τις δύο σειρές) του μονοπατιού στρέβλωσης. Η τελική μετρική απόστασης μεταξύ των δύο χρονοσειρών προκύπτει ως $D(n,n)$.

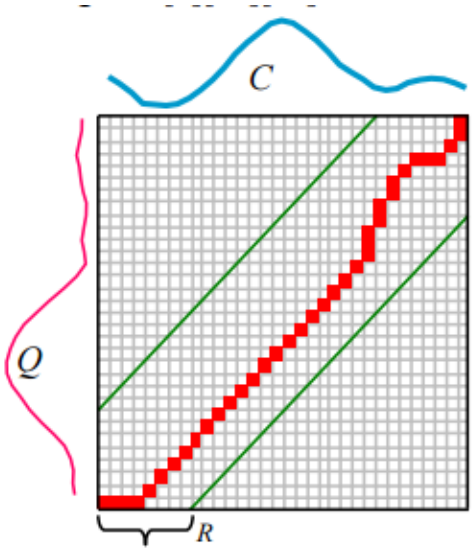


Σχήμα 3.6: Απόπειρα ταιριάσματος δύο χρονοσειρών χρησιμοποιώντας ως μετρική απόστασης α) την Ευκλίδεια απόσταση μεταξύ χρονικά ίδιων instances β) την Ευκλίδεια απόσταση μεταξύ των instances που έχουν προκύψει ως ταυτόχρονα με εφαρμογή του αλγορίθμου dtw. Εμφανής η ανωτερότητα της δεύτερης προσέγγισης σε ότι αφορά την ικανότητα σύλληψης της ομοιότητας μεταξύ των χρονοσειρών. [10]

Από τον πίνακα D, εύκολα μπορούμε να ανακτήσουμε το βέλτιστο μονοπάτι στρέβλωσης. Για να το κάνουμε αυτό, ξεκινάμε από το τελευταίο στοιχείο του πίνακα, και σε κάθε βήμα, αν βρισκόμαστε στο σημείο (i, j) του πίνακα, συγκρίνουμε με τα σημεία $(i, j-1), (i-1, j)$, και $(i-1, j-1)$, και εισάγουμε το σημείο στο οποίο ο πίνακας D έχει την ελάχιστη τιμή στο βέλτιστο μονοπάτι στρέβλωσης. Αφού φτάσουμε στο σημείο $(1, 1)$, αντιστρέφουμε το μονοπάτι προκειμένου να ξεκινάει από τη θέση $(1, 1)$ και να καταλήγει στη θέση (n, n).

Συχνά, επιθυμούμε να επιβάλλουμε κάποιους περιορισμούς τοπικότητας στην έρευνά μας. Για το σκοπό αυτό, συχνά δεν αφήνουμε τις μεταβλητές i, j να λαμβάνουν και οι δύο όλες τις τιμές στο $[1, n]$, αλλά περιορίζουμε την j στο $[i - w, i + w]$, όπου w το εύρος του παραθύρου έρευνας. Η ζώνη έρευνας που προκύπτει με αυτόν τον τρόπο ονομάζεται ζώνη Sakoe – Chiba.

Επί του αλγορίθμου αυτού, έχουν προταθεί διάφορες παραλλαγές, βελτιώσεις και επεκτάσεις. Άξιες αναφοράς είναι η χρήση των πρώτων διαφορών των χρονοσειρών



Σχήμα 3.7: Οπτικοποίηση παραδείγματος εφαρμογής του αλγορίθμου δυναμικής χρονοστρέβλωσης με τοπικούς περιορισμούς (η περιοχή έρευνας είναι ανάμεσα στις πράσινες γραμμές). Με κόκκινο, το τελικό βέλτιστο μονοπάτι στρέβλωσης. [11]

(ως προσέγγιση παραγώγων) ως χαρακτηριστικά, αντί για τις τιμές τους. που προτάθηκε από τους Keogh et al. [73], και η επέκταση του αλγορίθμου σε χρονοσειρές άνω των μία διαστάσεων, κατά την οποία πρακτικά εφαρμόζουμε σε όλες τις διαστάσεις της χρονοσειράς ξεχωριστά τον αλγόριθμο και αθροίζουμε τους επιμέρους πίνακες στρέβλωσης, D, προκειμένου να βρούμε τόσο την τελική μετρική απόστασης, όσο και το βέλτιστο μονοπάτι στρέβλωσης, που χρησιμοποιήθηκε από τους ten Holt et al. [64] για το σκοπό αναγνώρισης χειρονομιών.

3.5 Μοντέλα Markov

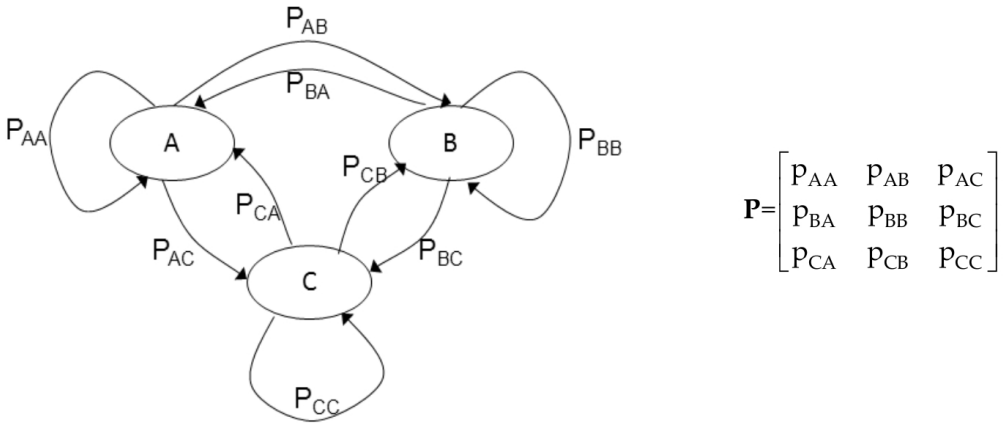
3.5.1 Γενικά

Τα μοντέλα Markov (Markov models) είναι ένα στοχαστικό μοντέλο, που χρησιμοποιείται για τη μοντελοποίηση συστημάτων, για τα οποία υποθέτουμε ότι:

- Ο χώρος καταστάσεων X μπορεί να χωριστεί σε κάποιες διακριτές καταστάσεις.
- Η τρέχουσα κατάσταση εξαρτάται μόνο από ένα πεπερασμένο ιστορικό προηγούμενων καταστάσεων, πληρείται, δηλαδή, η υπόθεση Markov (η οποία είναι γνωστή και ως ιδιότητα απώλειας μνήμης).

Ο αριθμός των παρελθόντων καταστάσεων από τις οποίες εξαρτάται η τρέχουσα κατάσταση ονομάζεται τάξη της ακολουθίας Markov. Η απλούστερη περίπτωση ακολουθίας Markov είναι η ακολουθία Markov πρώτης τάξης, κατά την οποία η

κάθε κατάσταση (που αντιστοιχεί στη χρονική στιγμή t) εξαρτάται μόνον από την προηγούμενή της (που αντιστοιχεί στη χρονική στιγμή $t-1$). Συμβολίζουμε, δηλαδή, την πιθανότητα να βρισκόμαστε σε κάποια κατάσταση τη χρονική στιγμή t ως $P(X_t) = P(X_t|X_{t-1})$, για να δηλώσουμε την εξάρτηση από την κατάσταση X_{t-1}. Αυξάνοντας την τάξη του μοντέλου, αυξάνουμε και τη συνθετότητά του, για παράδειγμα σε ένα μοντέλο Markov τρίτης τάξης έχουμε εξάρτηση της παρούσας κατάστασης από τις τρεις προηγούμενές της. Πάντα μπορούμε να υποβιβάσουμε μία αλυσίδα Markov υψηλότερης τάξης σε μία αλυσίδα Markov πρώτης τάξης, αυξάνοντας τις μεταβλητές του χώρου καταστάσεων.



Σχήμα 3.8: Απεικόνιση μίας αλυσίδας Markov (πρώτης τάξης), 3 καταστάσεων. Βλέπουμε επίσης τον πίνακα μεταβάσεων της αλυσίδας. [12]

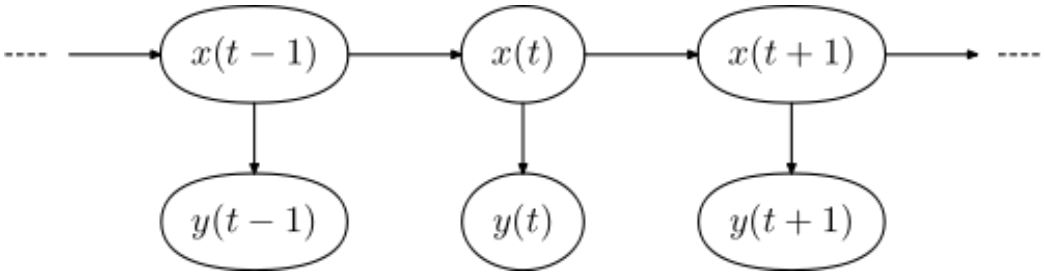
Όπως είπαμε, στις αλυσίδες Markov, γνωρίζουμε πάντα σε ποια κατάσταση βρισκόμαστε, το σύστημα είναι δηλαδή πλήρως παρατηρήσιμο. Αντίθετα, στα κρυφά μοντέλα Markov (hidden Markov models, HMMs), το σύστημα είναι μερικώς μόνο παρατηρήσιμο. Δεν μπορούμε να ξέρουμε σε ποια κατάσταση του συστήματος βρισκόμαστε σε κάθε χρονική στιγμή. Μας είναι, αντίθετα, φανερή, κάποια έξοδος του συστήματος (η οποία ανήκει σε ένα σύνολο παρατηρήσεων, Y), η οποία και εξαρτάται επίσης πιθανοτικά από την κατάσταση στην οποία είμαστε (και μόνο από αυτήν), δηλαδή ισχύει – με την υπόθεση ότι το μοντέλο παραμένει πρώτης τάξης:

$$P(X_t) = P(X_t|X_{t-1}) \tag{3.11}$$

$$P(Y_t) = P(Y_t|X_t) \tag{3.12}$$

Ας υποθέσουμε, τώρα, ότι έχουμε X: χώρο καταστάσεων, συνολικά με m καταστάσεις, και Y: χώρο παρατηρήσεων, συνολικά με n παρατηρήσεις. Ορίζουμε τους εξής πίνακες:

- Πίνακας μετάβασης $T : mxm$ με $T(i,j) = P(X_t = j|X_{t-1} = i)$



Σχήμα 3.9: Απεικόνιση της χρονικής εξέλιξης ενός κρυφού μοντέλου Markov. Παρατηρούμε ότι, πέραν των μεταβάσεων/αλλαγών κατάστασης στο πεδίο του χρόνου, κάθε χρονική στιγμή έχουμε και μία παρατήρηση/έξοδο του συστήματος η οποία είναι σε εμάς ορατή. [13]

- Πίνακας παρατηρήσεων $E : mxn$ με $E(i, j) = P(Y_t = j|X_t = i)$

Οι ορισμοί αυτοί είναι πολύ βολικοί, καθώς διευκολύνουν την εύρεση της πιθανότητας να βρισκόμαστε τη χρονική στιγμή t σε οποιαδήποτε κατάστασηi του συνόλου καταστάσεων X. Συγκεκριμένα, θα ισχύει για τον πίνακα καταστάσεων:

$$X_t = T^T X_{t-1} \tag{3.13}$$

Ενώ για τις παρατηρήσεις:

$$Y_t = E^T X_t \tag{3.14}$$

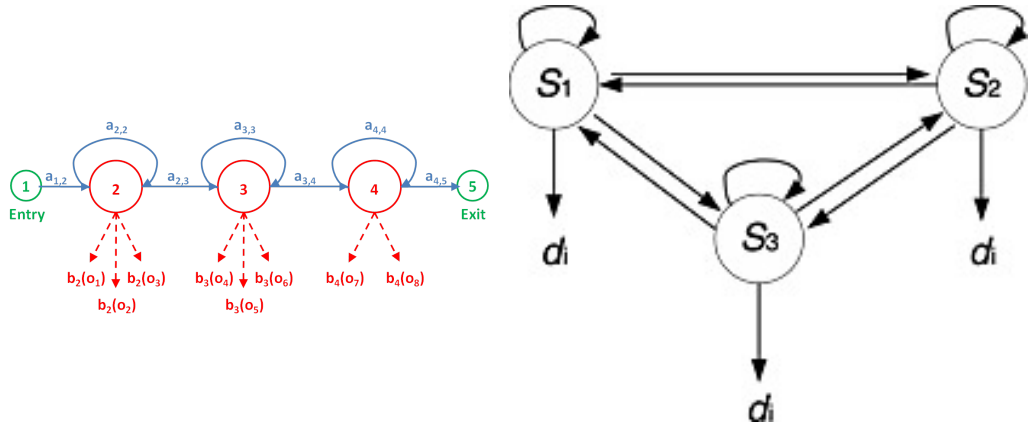
.

Αυτό δίνει μία γρήγορη λύση στο πρόβλημα της εκτίμησης της πιθανότητας εμφάνισης μίας συγκεκριμένης ακολουθίας παρατηρήσεων $Y_1, Y_2, \ldots, Y_n$, επαναλαμβάνοντας τους ανωτέρω πολλαπλασιασμούς πινάκων.

Αυτό επίσης διευκολύνει και την εύρεση της πιθανότερης κατάστασης δεδομένου ενός συνόλου παρατηρήσεων. Δεδομένης μίας ακολουθίας $Y_1, Y_2, \ldots, Y_n$ γνωστών καταστάσεων και γνωστών των a priori πιθανοτήτων να είμαστε σε κάθε κατάσταση του μοντέλου Markov, θέλουμε να βρούμε την πιθανότερη ακολουθία καταστάσεων $X_1, X_2, \ldots, X_n$. Λόγω του ότι πληρείται η συνθήκη Markov, η αποτίμηση αυτή μπορεί να γίνει ανά βήμα, δηλαδή να βρούμε την πιθανότερη κατάσταση για τη χρονική στιγμή t=1, να βρούμε με βάση αυτή την πιθανότερη για τη χρονική στιγμή t=2, κοκ. Ο αλγόριθμος αυτός είναι γνωστός ως αλγόριθμος Viterbi, και αποτελείται από επαναλαμβανόμενες εκτελέσεις του αλγορίθμου forward, με τη διαφορά ότι μας ενδιαφέρει η μεγιστοποίηση της ποσότητας:

$$\Pi(T^T X_{t-1})$$

Τα κρυφά μοντέλα Markov πρώτης τάξης χωρίζονται σε δύο βασικές κατηγορίες. Τα πλήρως συνδεδεμένα (ή fully connected) μοντέλα, στα οποία κάθε μετάβαση είναι



Σχήμα 3.10: Συγκριτική απεικόνιση ενός forward-only HMM (αριστερά) [14], και ενός πλήρως συνδεδεμένου HMM (δεξιά). [15]

πιθανή (ο πίνακας μεταβάσεων, T, μπορεί να έχει όλα τα στοιχεία του μη μηδενικά), και τα forward-only, στα οποία μόνο δύο ειδών μεταβάσεις είναι δυνατές, η παραμονή στην ίδια κατάσταση και η μετάβαση στην επόμενη. Στην περίπτωση αυτή, τα μόνα μη μηδενικά στοιχεία του πίνακα είναι τα στοιχεία της διαγωνίου και τα αμέσως επόμενα τους, δηλαδή ισχύει:

$$T(i,j) = 0 \forall \{i,j\} \neq \{\{k,k\},\{k,k+1\}\}, k \in X - \{X_m\}$$

3.5.2 Εκπαίδευση Κρυφών Μοντέλων Markov - ο Αλγόριθμος Baum – Welch.

Σε ό,τι αφορά την εκπαίδευση αυτή καθ'αυτή ενός κρυφού μοντέλου Markov – κατά την οποία πρακτικά προσπαθούμε, ξεκινώντας από κάποιες αρχικές εκτιμήσεις, να βελτιστοποιήσουμε τους πίνακες μετάβασης, T, παρατήρησης, E, και εκ των προτέρων πιθανοτήτων, π, ώστε να μπορέσουμε να μοντελοποιήσουμε μία συγκεκριμένη κατηγορία παρατηρήσεων , οι συχνότερα χρησιμοποιούμενοι αλγόριθμοι είναι αυτοί των Baum – Welch και Viterbi. Μία συγκριτική αποτίμηση της αποτελεσματικότητας των δύο αλγορίθμων αποτυπώνεται στην εργασία των Rodriguez et al. [74].

Ο αλγόριθμος Baum – Welch είναι, στην πραγματικότητα, μία ακόμη εφαρμογή του αλγορίθμου Προσδοκίας – Μεγιστοποίησης (EM). Συγκεκριμένα, έστω το κρυφό μοντέλο Markov, για το οποίο έχουμε κάποιες αρχικές εκτιμήσεις των πινάκων μετάβασης T, παρατηρήσεων E, και των εκ των προτέρων πιθανοτήτων, π (συμβολίζουμε τα παραπάνω – τις γενικές παραμέτρους του μοντέλου Markov - ως θ). Ο αλγόριθμος Baum – Welch αποτελείται από τα εξής τρία βήματα, τα οποία και επαναλαμβάνονται μέχρι να υπάρξει σύγκλιση (η κάθε επανάληψη του αλγορίθμου αποτελείται από τα παρακάτω βήματα, ξεχωριστά για κάθε δείγμα παρατηρήσεων που έχουμε διαθέσιμο):

- Πέρασμα προς τα μπροστά (forward pass) του συνόλου παρατηρήσεων: Έστω

$\alpha_j(t) = P(Y_1, \ldots, Y_t, X_t = j|\theta)$ η πιθανότητα να βρισκόμαστε στην κατάσταση j, και να έχουμε τη σειρά παρατηρήσεων $Y_1, \ldots, Y_t$. Τότε, θα έχουμε, για κάθε κατάσταση:

- Ορίζουμε ως $\alpha_j(1)$ το γινόμενο της a priori πιθανότητας να είμαστε στην κατάσταση j με την πιθανότητα παρατήρησης της Y_1 γι αυτήν την κατάσταση, δηλαδή:
$$\alpha_j(1) = \pi_j e_j(Y_1) \tag{3.15}$$
- Για κάθε επόμενη χρονική στιγμή, υπολογίζουμε το $\alpha_j(t+1)$ ως το γινόμενο της συνολικής πιθανότητας να είμαστε στην κατάσταση j – η οποία υπολογίζεται αναδρομικά, αθροίζοντας, για όλες τις καταστάσεις i, τα γινόμενα των προηγούμενων τιμών των πιθανοτήτων (για τη χρονική στιγμή t) με τις αντίστοιχες πιθανότητες μετάβασης στην κατάσταση j – επί την αντίστοιχη πιθανότητα παρατήρησης.
$$\alpha_j(t+1) = e_j(Y_{t+1}) \sum_{i=1}^{N} (a_i(t) t_{ij}) \tag{3.16}$$

- Πέρασμα προς τα πίσω (backward pass) του συνόλου παρατηρήσεων: Έστω $\beta_j(t) = P(Y_{t+1}, \ldots, Y_T|X_t = j, \theta)$ η πιθανότητα εμφάνισης της σειράς των μελλοντικών παρατηρήσεων αν γνωρίζουμε τις παραμέτρους του μοντέλου και την τωρινή κατάσταση. Αρχικοποιούμε $\beta_j(T) = 1$, και απο εκεί υπολογίζουμε αναδρομικά και προς τα πίσω τα $\beta_j(t)$, αθροίζοντας στο χώρο καταστάσεων το γινόμενο των προηγουμένων (επομένων χρονικά) β_i, της πιθανότητας εμφάνισης της παρατήρησης Y_{t+1} (η οποία και δεν συμπεριλήφθηκε στον προηγούμενο υπολογισμό) και της πιθανότητας μετάβασης από την κατάσταση j στην κατάσταση i).
$$\beta_j(t) = \sum_{i=1}^{N} (b_i(t+1) t_{ji} e_i(Y_{t+1})) \tag{3.17}$$

- Βήμα ανανέωσης (update step):
  - Καταρχήν, υπολογίζουμε τις ενδιάμεσες πιθανότητες $\gamma_i(t) = P(X_t = i|Y, \theta)$ (που δίνουν την πιθανότητα για συγκεκριμένο μοντέλο Markov, δεδομένης της ακολουθίας καταστάσεων που έχουμε, να βρισκόμαστε στην κατάσταση i) και $\xi_{ij}(t) = P(X_t = i, X_{t+1} = j|Y, \theta)$ (που δίνουν την πιθανότητα να έχουμε τη μετάβαση από την κατάσταση i στην κατάσταση j για συγκεκριμένο μοντέλο Markov), βάσει των τύπων (για κάθε χρονική στιγμή, τα στοιχεία των γ, ξ έχουν άθροισμα 1):
$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{\sum_j (\alpha_j(t)\beta_j(t))} \tag{3.18}$$

$$\xi_{ij}(t) = \frac{\alpha_i(t)t_{ij}\beta_j(t+1)e_j(Y_{t+1})}{\sum_i \sum_j (\alpha_i(t)t_{ij}\beta_j(t+1)e_j(Y_{t+1}))} \tag{3.19}$$

– Τέλος, αφού επαναλάβουμε τους ανωτέρω υπολογισμούς για το σύνολο των δειγμάτων εκπαίδευσης (έστω E), ανανεώνουμε τις παραμέτρους του κρυφού μοντέλου Markov, ως:

$$\pi_i = \frac{(\sum_{samples} \gamma_i(1))}{E} \tag{3.20}$$

$$t_{ij} = \frac{\sum_{samples} \sum_t \xi_{ij}(t)}{\sum_{samples} \sum_t (\gamma_i(t))} \tag{3.21}$$

$$e_i(v_k) = \frac{\sum_{samples} \sum_t (\|_{U_t=v_k})\gamma_i(t)}{\sum_{samples} \sum_t (\gamma_i(t))} \tag{3.22}$$

Από τα παραπάνω, ο πρώτος τύπος αντιστοιχεί στην εκτίμηση της πιθανότητας να βρισκόμαστε στην κατάσταση i για τη χρονική στιγμή $t = 1$), στον δεύτερο, έχουμε τον υπολογισμό των στοιχείων του πίνακα μεταβάσεων (η πιθανότητα μετάβασης από την κατάσταση i στην κατάσταση j υπολογίζεται ως ο λόγος της πιθανότητας εμφάνισης της ακολουθίας καταστάσεων i, j διά την πιθανότητα εμφάνισης της κατάστασης i), ενώ στον τρίτο, τον υπολογισμό των στοιχείων του πίνακα παρατηρήσεων (όπου ο όρος $\|_{Y_t=v_k}$ αποτελεί δείκτρια συνάρτηση).

3.5.3 Μοντέλα Markov με Συνεχείς Μεταβλητές Παρατήρησης

Τα παραπάνω ισχύουν στις περιπτώσεις όπου η μεταβλητή παρατήρησης είναι διακριτή. Στις περιπτώσεις όπου η μεταβλητή παρατήρησης είναι συνεχής, οι συνήθως χρησιμοποιούμενες λύσεις είναι δύο:

- Η διακριτοποίηση του χώρου παρατηρήσεων, μέσω ενός codebook.
- Η μοντελοποίηση των παρατηρήσεων σε κάθε κατάσταση του κρυφού μοντέλου Markov ως μίγμα Γκαουσιανών (mixture of Gaussians, moG).

Στην πρώτη από τις δύο περιπτώσεις, απλώς παρακάμπτουμε το πρόβλημα της ύπαρξης συνεχών μεταβλητών παρατήρησης, μετατρέποντας τις σε διακριτές. Για να το πετύχουμε αυτό, είναι απαραίτητο να χωρίσουμε το πεδίο ορισμού των μεταβλητών παρατήρησης, R^n σε κ συστάδες σημείων, εφαρμόζοντας κάποιον από τους αλγορίθμους ομαδοποίησης που αναφέρθηκαν παραπάνω. Το σύνολο των διακριτών labels, που προκύπτουν για τα διάφορα σημεία του χώρου, με αυτή τη μέθοδο, λέγεται λεξικό (dictionary) ή κωδικό βιβλίο (codebook), και το κάθε label, κωδική λέξη (codeword). Από το σημείο αυτό και μετά, για την εκπαίδευση του μοντέλου ακολουθείται ακριβώς η διαδικασία που αντιστοιχεί στα διακριτά μοντέλα Markov, όπως και για την αποτίμηση, αφού τα συνεχή σημεία εισόδου διακριτοποιηθούν βάσει της αντιστοίχισης

που είχε παραχθεί αρχικά.

Στη δεύτερη περίπτωση, περιγράφουμε τις συνεχείς μεταβλητές παρατήρησης, που αντιστοιχούν σε κάθε κατάσταση, με βάση τις στατιστικές ιδιότητές τους: τη μέση τιμή τους και τους πίνακες συνδιασποράς τους. Συνεπώς, αλλάζει η μοντελοποίηση των κρυφών μοντέλων Markov: Διατηρείται ο πίνακας μετάβασης καταστάσεων, ο οποίος δηλώνει τις πιθανότητες μετάβασης από κατάσταση σε κατάσταση, όμως η πιθανότητα εμφάνισης κάποιας παρατήρησης δε μοντελοποιείται πλέον με τη μορφή πίνακα, όπου το στοιχείο $E(i, j)$ δηλώνει την πιθανότητα εμφάνισης της παρατήρησης j στην κατάσταση i, αλλά ως πυκνότητα πιθανότητας που ακολουθεί την κανονική κατανομή, δηλαδή $X \sim N(\mu, \Sigma)$, όπου X η εκάστοτε παρατήρηση. Συνεπώς, οι αλγόριθμοι εκπαίδευσης σε αυτήν την περίπτωση δεν πρέπει να εκπαιδευτούν προκειμένου να βελτιστοποιήσουν – όπως στην προηγούμενη περίπτωση – τη λογαριθμική πιθανοφάνεια και τους πίνακες παρατηρήσεων και μεταβάσεων, αλλά τη λογαριθμική πιθανοφάνεια, τους πίνακες μεταβάσεων και τις παραμέτρους των πιθανοτικών κατανομών με τις οποίες μοντελοποιούμε τις παρατηρήσεις που αντιστοιχούν – δηλαδή, τη μέση τιμή μ και τον πίνακα συνδιασποράς, Σ των γκαουσιανών κατανομών. Αξίζει να σημειωθεί ότι οι μεταβλητές που αντιστοιχούν σε κάθε κατάσταση δε μοντελοποιούνται απαραίτητα με μία Γκαουσιανή κατανομή, αλλά συχνά με υπέρθεση (μίξη) Γκαουσιανών.

Για την περίπτωση αυτή, η εφαρμογή του αλγορίθμου forward θα ακολουθεί το γενικό τύπο:

$$P(X_t = i, O_t = a) = \sum_j (P(X_{t-1} = j)T(j, i))P(O_t = \alpha | X_t = i) \tag{3.23}$$

όπου η $P(O_t = a|X_t = i)$ ακολουθεί την κανονική κατανομή (ή αποτελείται από άθροισμα κανονικών κατανομών).

Αντίστοιχα, για την εκπαίδευση των κρυφών μοντέλων Markov, ορίζοντας $e_{jl}(Y_t)$ την τιμή της πυκνότητας πιθανότητας της l-οστής Γκαουσιανής κατανομής που αντιστοιχεί στην κατάσταση j, και υποθέτοντας τη χρήση L γκαουσιανών για τη μοντελοποίηση μιας κατάστασης, ο αλγόριθμος Baum – Welch τροποποιείται ως εξής:

- Στα πρώτα δύο βήματα εκτέλεσης – την εκτέλεση του αλγορίθμου forward και την εκτέλεση του αλγορίθμου backward – πρακτικά ακολουθούμε τα ίδια βήματα με προηγουμένως, με μοναδική διαφορά ότι πλέον, για να εκφράσουμε την πιθανότητα εμφάνισης μίας παρατήρησης σε κάποια κατάσταση, αν ορίσουμε το βάρος της l-οστής Γκαουσιανής ως c_l (προφανώς ισχύει $\sum_{l=1}^{L} c_l = 1$), δε χρησιμοποιούμε πλέον το στοιχείο $E(i, j)$ (καθώς αυτό πλέον δεν υφίσταται) αλλά τον όρο

$$\sum_{l=1}^{L} (c_{il} P(Y_t \ N(\mu_{il}, \Sigma_{il}))) = \sum_{l=1}^{L} (b_{il}(Y_t)). \tag{3.24}$$

- Κατά το τρίτο βήμα, υπολογίζουμε ομοίως τις ενδιάμεσες παραστάσεις $\gamma_i(t) = P(X_t = i|Y, \theta)$ και $\xi_{ij}(t) = P(X_t = i, X_{t+1} = j|Y, \theta)$ για κάθε κατάσταση – ο προηγούμενος τύπος ισχύει αυτούσιος για την $\gamma_i(t)$ και με την αντικατάσταση του όρου πιθανοφάνειας για την $\xi_{ij}(t)$ που έγινε και προηγουμένως.

- Στη συνέχεια, υπολογίζουμε για κάθε συστατική Γκαουσιανή την ποσότητα $\gamma_{il}(t)$ που δηλώνει την πιθανότητα να βρισκόμαστε στην i-οστή κατάσταση και στην l-οστή Γκαουσιανή από όσες την περιγράφουν, μέσω του τύπου:

$$\gamma_{il} = \gamma_i c_{il} \frac{b_{il}(Y_t)}{\sum_{l=1}^{L}(b_{il}(Y_t))} \tag{3.25}$$

- Τέλος, αφού επαναλάβουμε τα ανωτέρω για όλα τα δείγματα, ανανεώνουμε τον πίνακα μεταβάσεων του κρυφού μοντέλου Markov καθώς και τις παραμέτρους των Γκαουσιανών που μοντελοποιούν τις πιθανότητες εμφάνισης των παρατηρήσεων σε κάθε κατάσταση, ως:

$$\pi_i = \frac{(\sum_{samples} \gamma_i(1))}{E} \tag{3.26}$$

$$t_{ij} = \frac{\sum_{samples} \sum_t \xi_{ij}(t)}{\sum_{samples} \sum_t (\gamma_i(t))} \tag{3.27}$$

Οι οποίοι πρακτικά είναι οι ίδιοι τύποι με την προηγούμενη περίπτωση για τις εκ των προτέρων πιθανότητες και τον πίνακα μετάβασης. Για τις μέσες τιμές και τους πίνακες συνδιασποράς, καθώς και τα ποσοστά συμμετοχής της κάθε Γκαουσιανής στη μοντελοποίηση των παρατηρήσεων ισχύουν οι παρακάτω κανόνες ανανέωσης:

$$c_{il} = \frac{\sum_{samples} \sum_t \gamma_{il}(t)}{\sum_{samples} \sum_t \gamma_i(t)} \tag{3.28}$$

$$\mu_{il} = \frac{\sum_{samples} \sum_t (\gamma_{il}(t) O_t)}{\sum_{samples} \sum_t (\gamma_{il}(t))} \tag{3.29}$$

$$\Sigma_{il} = \frac{\sum_{samples} \sum_t (\gamma_{il}(t) d_{til} d_{til}^T)}{\sum_{samples} \sum_t (\gamma_{il}(t))} \tag{3.30}$$

Που είναι ακριβώς οι τύποι για υπολογισμό μέσων τιμών και συνδιασπορών σε μοντέλα μίξης γκαουσιανών, με τα $\sum_t(\gamma_{il}(t))$ ως βάρη, και θέτοντας $d_{til} = O_t - \mu_{il}$.

3.6 Τυχαία Δάση Απόφασης

3.6.1 Δέντρα Απόφασης

Σε γενικές γραμμές, ένα δέντρο απόφασης (decision forest) αποτελεί ένα εργαλείο που μοντελοποιεί τις παραμέτρους ενός προβλήματος απόφασης με τη χρήση μίας

δενδρικής δομής. Είναι χρήσιμα τόσο στον κλάδο της ανάλυσης αποφάσεων (decision analysis), όσο και σε προβλήματα Μηχανικής Μάθησης και Εξόρυξης Δεδομένων.

Γενικά, η δομή ενός δένδρου απόφασης είναι η εξής: Ξεκινώντας από έναν κεντρικό κόμβο - ρίζα (root node), στοχεύουμε, προχωρώντας προς τα κάτω, να μοντελοποιήσουμε πιθανοτικά κάθε πιθανό ενδεχόμενο. Για το σκοπό αυτό, εναλλάσσουμε κόμβους - φύλλα (στους οποίους γίνεται αποτίμηση της παρούσας κατάστασης) με κόμβους απόφασης (στους οποίους μελετώνται διαφορετικές δυνατές αποφάσεις) και κόμβους πιθανοτήτων (όπου μελετώνται οι συνέπειες τυχαίων συμβάντων, ανεξαρτήτων από τις ληφθείσες αποφάσεις). Όταν έχουμε πετύχει επαρκή μοντελοποίηση του προβλήματος, μετατρέπουμε τους κατώτερους κόμβους - φύλλα σε τερματικούς κόμβους.

Γενικά, τα δέντρα απόφασης έχουν το πλεονέκτημα ότι, λόγω της δομής τους, είναι εύκολα κατανοητά και προσφέρουν δυνατότητες ερμηνείας των αποτελεσμάτων.

3.6.2 Εκπαίδευση Δέντρων Απόφασης

Ωστόσο, κατά τη μεταφορά του παραπάνω μοντέλου σε προβλήματα Μηχανικής Μάθησης, το μόνο το οποίο πρακτικά διατηρείται είναι η δομή του δέντρου. Πιο συγκεκριμένα, χρησιμοποιούμε τα δέντρα απόφασης ως προβλεπτικούς μηχανισμούς σε προβλήματα ταξινόμησης ή παλινδρόμησης, διατυπώνοντας, στους κόμβους απόφασης, ερωτήσεις σχετικά με κάποια χαρακτηριστικά των στοιχείων εισόδου, προκειμένου να καταλήξουμε είτε σε κάποια αποτίμηση για την κλάση στην οποία το στοιχείο ανήκει (είτε αυστηρά είτε με τη χρήση πιθανοτικής κατανομής), είτε για την τιμή του αναλόγως τον τύπο του προβλήματος.

Φυσικά, είναι δυνατό τα δέντρα απόφασης σε αυτές τις περιπτώσεις να κατασκευαστούν με το χέρι. Μολαταύτα, ένας από τους λόγους για τους οποίους αποτελούν ισχυρούς ταξινομητές είναι η δυνατότητα αυτοματοποιημένης εκπαίδευσης τους, βάσει της χρήσης αλγορίθμων οι οποίοι, σε κάθε κόμβο του δέντρου, επιλέγουν το ιδανικότερο κριτήριο προκειμένου να βελτιστοποιήσουμε τα αποτελέσματα της ταξινόμησης. Σε γενικές γραμμές, συνήθης διαδικασία για την εκπαίδευση ενός δέντρου απόφασης για ταξινόμηση (ή παλινδρόμηση), με χρήση ενός συνόλου δεδομένων X με χαρακτηριστικά (είτε διακριτά, είτε συνεχή) x_i είναι η ακόλουθη:

- Ξεκινώντας από το αρχικό σύνολο δεδομένων, εύρεση του κριτηρίου διαχωρισμού που μεγιστοποιεί κάποια μετρική σχετιζόμενη είτε με την εντροπία είτε με την απώλεια πληροφορίας, και δημιουργία δύο κόμβων - παιδιών του αρχικού.
- Αναδρομική εφαρμογή του παραπάνω στους νέους κόμβους, μέχρι να μην μπορεί να προκύψει κάποιο επιπλέον όφελος.

Για την περίπτωση δέντρων ταξινόμησης, οι πιο συχνά χρησιμοποιούμενες μετρικές είναι δύο: η καθαρότητα κατά Gini (Gini impurity) και το κέρδος πληροφορίας.

- Για την πρώτη περίπτωση, έχουμε να ορίσουμε την ομοιογένεια ενός συνόλου, το οποίο αποτελείται από στοιχεία τα οποία ανήκουν σε κάποια κλάση $j, j = 1, ..., J$ ως το άθροισμα, για κάθε κλάση, του γινομένου της πιθανότητας ένα τυχαίο στοιχείο να ανήκει σε αυτή την κλάση επί την πιθανότητα να μην ανήκει. Είναι εμφανές από τον παρακάτω τύπο ότι έχουμε ελαχιστοποίηση αν $p_i = 0$ ή $p_i = 1$, οπότε και ένα σύνολο είναι πλήρως ομοιογενές ως προς μία κλάση.

$$I(p) = \sum_{i=1}^{J} p_i(1 - p_i) \tag{3.31}$$

  Συνεπώς, επιλέγουμε το κριτήριο που ελαχιστοποιεί το άθροισμα των παραπάνω ποσοτήτων σε όλα τα προκύπτοντα splits.

- Για τη δεύτερη περίπτωση, είναι απαραίτητο να ορίσουμε την έννοια της εντροπίας. Η εντροπία ενός συνόλου S, το οποίο αποτελείται από στοιχεία τα οποία ανήκουν σε κάποια κλάση $j, j = 1, ..., J$ ορίζεται, από τη θεωρία πληροφορίας, ως:

$$E(S) = -\sum_{i=1}^{J} p_i log_2(p_i) \tag{3.32}$$

  Όπως μπορούμε να δούμε, συμπεριφέρεται αντίστοιχα με την καθαρότητα κατά Gini σε ό,τι αφορά τα σημεία μηδενισμού της. Και πάλι αντίστοιχα με πριν, επιλέγουμε το κριτήριο διαχωρισμού που ελαχιστοποιεί τη συνολική εντροπία σε όλα τα splits.

Τέλος, στην περίπτωση όπου έχουμε κάποιο πρόβλημα παλινδρόμησης, λόγω του ότι η μεταβλητή που πρέπει να προβλέψουμε δεν είναι πλέον διακριτή αλλά συνεχής, επιθυμούμε το κριτήριο διαχωρισμού να μεγιστοποιεί τη διασπορά στα σύνολα των προκυπτώντων splits. Για ένα σύνολο S, με στοιχεία συνεχείς μεταβλητές, ορίζουμε τη διασπορά ως:

$$Var(S) = \frac{1}{2|S|} \sum_{i \in S} \sum_{j \in S} (x_i - x_j)^2 \tag{3.33}$$

3.6.3 Συνδυασμένοι Ταξινομητές με Χρήση Δέντρων Απόφασης

Στα βασικά πλεονεκτήματα ενός δέντρου απόφασης συγκαταλέγονται η απλότητα στην κατανόησή τους, η ικανότητα χειρισμού τόσο αριθμητικών όσο και κατηγορικών δεδομένων, η υψηλή απόδοση σε μεγάλα σύνολα δεδομένων και το γεγονός ότι προσομοιάζουν, στη λογική τους, την ανθρώπινη πορεία σκέψης και απόφασης. Ωστόσο, είναι αρκετά ευάλωτα σε προβλήματα υπερφόρτωσης (overfitting). Οι κύριες λύσεις στο πρόβλημα αυτό είναι τόσο το κλάδεμα (pruning) των προκυπτόντων δέντρων, αφαιρώντας κόμβους που δε συνεισφέρουν ιδιαίτερα, είτε η παράλληλη χρήση πολλών

δέντρων ως συνδυασμένο ταξινομητή.

Μία πρώτη προσέγγιση πάνω στο θέμα αυτό κατασκευάζει διαδοχικά ένα σύνολο από δέντρα απόφασης, με το κάθε επόμενο δέντρο να δίνει έμφαση στα παραδείγματα του συνόλου εκπαίδευσης τα οποία έχουν ταξινομηθεί, από τα προηγούμενα, εσφαλμένα. Με αυτόν τον τρόπο, μπορεί σταδιακά να χτιστεί ένας ισχυρός ταξινομητής.

Ωστόσο, πλέον η ευρύτερα χρησιμοποιούμενη προσέγγιση στο πρόβλημα αυτό αφορά στην κατασκευή τυχαίων δασών απόφασης (Random Decision Forests, RDFs). Αυτό που επιδιώκεται, με τη χρήση της μεθόδου αυτής, είναι η κατασκευή ενός πιο εύρωστου ταξινομητή, χρησιμοποιώντας πολλά τυχαία δέντρα. Για το σκοπό αυτό, θεωρείται απαραίτητο τα δέντρα να μην είναι συσχετισμένα μεταξύ τους. Αυτό επιτυγχάνεται χρησιμοποιώντας τους εξής δύο μηχανισμούς:

- Το καθένα από τα προς εκπαίδευση δέντρα δε χρησιμοποιεί το σύνολο των δεδομένων εκπαίδευσης, αλλά ένα τυχαίο υποσύνολο. Αυτή η τεχνική ονομάζεται bootstrap aggregating (bagging), και η πατρότητά της στο συγκεκριμένο πεδίο αποδίδεται στον Breiman [75].

- Επιπλέον, για τα δεδομένα εκπαίδευσης δε χρησιμοποιούμε, σε κάθε δέντρο, όλα τα χαρακτηριστικά, αλλά ένα υποσύνολό τους, τυχαία επιλεγμένο. Η μέθοδος αυτή ονομάζεται Μέθοδος Τυχαίου Υπόχωρου (Random Subspace Method), και σε αυτή την περίπτωση αυτός που ευθύνεται για την εισαγωγή της στη βιβλιογραφία είναι ο Ho [76]. Αν έχουμε ένα σύνολο p δεδομένων, χρησιμοποιούμε $\sqrt{p}$ ανά δέντρο, σύμφωνα με τους συγγραφείς.

3.6.4 Ο Αλγόριθμος που Χρησιμοποιείται από το Kinect

Στο σημείο αυτό, θα αναλύσουμε σε επίπεδο υλοποίησης το σύστημα που χρησιμοποιεί το Kinect προκειμένου να εντοπίζει τη σκελετική πόζα ενός ανθρώπου, όπως περιγράφεται στη σχετική εργασία των Shotton et al. [24].

Αρκετές από τις προηγούμενες εργασίες που αναφέρθηκαν στο κεφάλαιο 2 αντιμετωπίζουν το πρόβλημα της εκτίμησης της στάσης σώματος ως πρόβλημα τοπικής βελτιστοποίησης, χρησιμοποιώντας ως αρχική εκτίμηση τη στάση που βρέθηκε ως βέλτιστη σε προηγούμενο frame. Αντίθετα, στον αλγόριθμο που χρησιμοποιείται στο Kinect δε λαμβάνεται υπόψη πρότερη πληροφορία, αλλά μόνο τα δεδομένα του παρόντος frame – δηλαδή θα μπορούσε να εφαρμοστεί, με κάποιες τεχνικές τροποποιήσεις, και σε ακίνητες εικόνες οι οποίες περιέχουν δεδομένα βάθους.

Ο αλγόριθμος εντοπισμού ανθρώπων του Kinect δουλεύει κατά μέλη, δηλαδή εντοπίζει πρώτα διάφορα μέλη του σώματος – κάτι χρήσιμο και για την κατασκευή της σκελετικής πόζας – και από αυτά, εν συνεχεία, αποφαίνεται για την ύπαρξη ή όχι ανθρώπου σε μία δεδομένη περιοχή της εικόνας. Για το σκοπό αυτό, κατασκευάστηκε

μία βάση δεδομένων από καθαρά συνθετικά δεδομένα για την εκπαίδευση (τα οποία προέκυψαν από πολλών ειδών τροποποιήσεις σε πραγματικά δεδομένα, προκειμένου η ταξινόμηση να μην επηρεάζεται από παράγοντες κλίμακας, οπτικής γωνίας, θορύβου της κάμερας, κτλ), και ένας συνδυασμός πραγματικών και συνθετικών δεδομένων για την αποτίμηση.

Αφού, λοιπόν, συλλέχθηκαν και παρήχθηκαν τα απαιτούμενα δεδομένα, και πραγματοποιήθηκε η απόδοση labels σε αυτά (από κάθε πόζα, σε κάθε εικονοστοιχείο δόθηκε ως label το μέλος του σώματος στο οποίο άνηκε), έπρεπε να δημιουργηθεί ένα σύνολο χαρακτηριστικών, τα οποία στη συνέχεια θα μπορούν να χρησιμοποιηθούν από οποιονδήποτε ταξινομητή της επιλογής μας. Ως χαρακτηριστικό, επιλέχθηκε η απόσταση βάθους μεταξύ δύο εικονοστοιχείων:

$$f(x) = d(x + \frac{u}{d(x)}) - d(x + \frac{v}{d(x)}) \tag{3.34}$$

όπου d(x) η τιμή του βάθους της εικόνας στο σημείο x και τα u, v διανύσματα θέσης (τα οποία κανονικοποιούνται με την παράμετρο d(x) προκειμένου τα τελικά χαρακτηριστικά να μένουν ανεπηρέαστα από το βάθος της εικόνας, καθώς το τελικό διάνυσμα u ή v θα έχει, τελικά, δεδομένη τιμή στο σύστημα συντεταγμένων του κόσμου). Αναλόγως με τις τιμές και τις διευθύνσεις των u, v, προκύπτουν ως χαρακτηριστικό αποστάσεις βάθους σε οποιαδήποτε xz διεύθυνση.

Έχοντας τα χαρακτηριστικά έτοιμα, το επόμενο βήμα είναι να χτιστεί ένας συγκεκριμένος ταξινομητής πάνω σε αυτά. Συγκεκριμένα, ο ταξινομητής που επιλέχθηκε για το σκοπό αυτό ήταν ένα τυχαίο δάσος, αποτελούμενο από 3 δυαδικά δέντρα, το καθένα εκ των οποίων είχε βάθος ίσο με 20 κόμβους. Για την εκπαίδευση των δέντρων, χρησιμοποιήθηκε ένας αλγόριθμος ελαχιστοποίησης εντροπίας, καθώς παράγονταν, ανά κόμβο, 2000 συνδυασμοί διανυσμάτων κατεύθυνσης u, v και 50 πιθανά κατώφλια απόφασης T, και επιλέγονταν σε κάθε βήμα το βέλτιστο σύμφωνα με το προαναφερθέν κριτήριο.

Ωστόσο, τα παραπάνω δέντρα παρέχουν πληροφορία για το μέλος του σώματος σε επίπεδο εικονοστοιχείων, όχι σε επίπεδο εικόνας. Για το σκοπό αυτό, για όλα τα εικονοστοιχεία που έχουν ταξινομηθεί σε συγκεκριμένο μέλος από μία εικόνα, χρησιμοποιείται ένας Γκαουσιανός πυρήνας, με κατάλληλο βάρος ώστε να λαμβάνεται υπόψη τόσο η πιθανότητα ταξινόμησης όσο και το βάθος του στοιχείου. Τα κέντρα των εν λόγω πυρήνων χρησιμοποιούνται ως η προτεινόμενη θέση για τους συνδέσμους, έναντι του ολικού μέσου στις τρεις διαστάσεις των σημείων που έχουν ταξινομηθεί, καθώς η χρησιμοποιηθείσα προσέγγιση είναι λιγότερο ευάλωτη σε μακρινά σημεία (outliers).

3.7 Βελτιστοποίηση

3.7.1 Γενικά

Γενικά μιλώντας, η έννοια της βελτιστοποίησης αφορά στη λήψη της καλύτερης δυνατής απόφασης σε μία οποιαδήποτε δεδομένη κατάσταση. Το ποια είναι, σε κάθε περίπτωση, η καλύτερη απόφαση δεν είναι μονοσήμαντο, αλλά εξαρτάται από το στόχο μας, καθώς και τυχόν περιορισμούς. Δηλαδή, ένα πρόβλημα βελτιστοποίησης καθορίζεται από:

- Τη δεδομένη κατάσταση, επί της οποίας πρέπει να ληφθεί μία απόφαση.
- Το στόχο που έχουμε θέσει.
- Τυχόν περιορισμούς, οι οποίοι περιορίζουν τις υπό εξέταση αποφάσεις.

Μετατρέποντας τα ανωτέρω σε μαθηματικά φορμαλιστική μορφή, έχουμε τα εξής: Έστω μία συνάρτηση $f(x)$, όπου $x \in X \subseteq R^n$. Ονομάζουμε την f συνάρτηση κόστους του προβλήματος, και προφανώς απαιτούμε να έχει ως σύνολο τιμών το R. Ονομάζουμε το X επιτρεπτό σύνολο του προβλήματος βελτιστοποίησης, καθώς αποτελείται από τον υποχώρο στον οποίον επιτρέπουμε να ανήκει το σημείο x. Αυτό που επιδιώκουμε είναι η εύρεση του σημείου $x_0 \in X$, το οποίο ελαχιστοποιεί τη συνάρτηση κόστους. Σε ότι αφορά τη δομή του συνόλου X, διακρίνουμε τις εξής περιπτώσεις:

- Αν δεν υπάρχουν περιορισμοί (είτε ισοτικοί είτε ανισοτικοί) επί των στοιχείων του x, τότε έχουμε ένα μη - περιορισμένο πρόβλημα βελτιστοποίησης και το επιτρεπτό σύνολο του προβλήματος είναι το R.
- Στην περίπτωση ωστόσο που υπάρχουν περιορισμοί, τότε το επιτρεπτό σύνολο του προβλήματος X αποτελείται από τα σημεία x^*, τα οποία ικανοποιούν όλους τους δοθέντες περιορισμούς. Το X μπορεί να είναι και το κενό σύνολο.

Στόχος του προβλήματος βελτιστοποίησης είναι η εύρεση του $x_0 \in X : argmin(f(x))$. Ένα πρόβλημα βελτιστοποίησης μπορεί να οριστεί ως κυρτό ή μη - κυρτό, αναλόγως με το παρακάτω:

- Αν οποιοδήποτε τοπικό ελάχιστο της $f(x)$ είναι και ολικό, το πρόβλημα βελτιστοποίησης είναι κυρτό.
- Αντίθετα, σε μη - κυρτά προβλήματα βελτιστοποίησης (που είναι και η πιο συνηθισμένη και ρεαλιστική περίπτωση), δεν υπάρχει καμία εγγύηση ότι κάποιο τοπικό ελάχιστο της $f(x)$ αποτελεί και ολικό της ελάχιστο.

Αξίζει να αναφέρουμε ότι, πρακτικά, μεγάλο μέρος των αλγορίθμων μηχανικής μάθησης επιλύουν, πρακτικά, προβλήματα βελτιστοποίησης. Συνήθως, το ζητούμενο είναι οι παράμετροι του μοντέλου που υλοποιεί την επιθυμητή πρόβλεψη, ενώ ως συνάρτηση κόστους χρησιμοποιούμε κάποια μετρική σφάλματος μεταξύ της επιθυμητής

εξόδου και της εξόδου του συστήματος.

Τέλος, ορίζουμε τα προβλήματα γραμμικού προγραμματισμού (linear programming) και τετραγωνικού προγραμματισμού (quadratic programming), ως εξής:

- Γραμμικός προγραμματισμός: minimize $f(x) = a^T x + b, a, x \in R^n, b \in R$
- Τετραγωνικός προγραμματισμός: minimize $f(x) = x^T Ax + c^T x + b, A \in R^{nxn}, c, x \in R^n, b \in R$

3.7.2 Κατηγοριοποίηση Τεχνικών Βελτιστοποίησης

Αναλόγως με το κατά πόσο ένας αλγόριθμος βελτιστοποίησης τερματίζει σε πεπερασμένο χρόνο, αποδεικνύεται μαθηματικά ότι συγκλίνει σε πεπερασμένο χρόνο στη λύση, ή απλώς στοχεύει στην έυρεση μίας ικανοποιητικά καλής λύσης αντί της βέλτιστης, διακρίνουμε τις ακόλουθες υποκατηγορίες τους:

- Αλγόριθμοι με εγγυημένο τερματισμό σε πεπερασμένο αριθμό βημάτων. Ωστόσο, πρακτικά τέτοιοι αλγόριθμοι χρησιμοποιούνται είτε για συγκεκριμένου τύπου προβλήματα (γραμμικού ή τετραγωνικού προγραμματισμού), είτε για προβλήματα διακριτής βελτιστοποίησης.
- Αναλυτικές επαναληπτικές μέθοδοι. Αυτές χρησιμοποιούνται για την επίλυση μη - γραμμικών ή τετραγωνικών προβλημάτων ελαχιστοποίησης, αρχικοποιώντας τυχαία την υποψήφια λύση του προβλήματος και εξετάζοντας κατά κανόνα το διάνυσμα πρώτων παραγώγων, ή τον Εσσιανό πίνακα δευτέρων παραγώγων, της συνάρτησης, με στόχο τη βελτίωση της εκτιμώμενης λύσης ανά βήμα. Ενδεικτικά αναφέρουμε μία δημοφιλή μέθοδο ανά κατηγορία:
  - Η μέθοδος της κατάβασης κλίσης (gradient descent), η οποία σε κάθε βήμα υπολογίζει την κατεύθυνση όπου η κλίση της συνάρτησης είναι η ελάχιστη, και μετακινεί το ελάχιστο σε εκείνη τη διεύθυνση. Βελτιώσεις επί της μεθόδου αυτής, συχνά, βελτιστοποιούν προς το μήκος του βήματος, ώστε στο πέρας του κάθε βήματος εκτέλεσης του αλγορίθμου να είμαστε στο τοπικό ελάχιστο της συνάρτησης σε αυτή την διεύθυνση.
  - Η μέθοδος Newton - Raphson, η οποία, υποθέτοντας σε κάθε βήμα ότι η συνάρτηση προς βελτιστοποίηση είναι τετραγωνική, υπολογίζει ως επόμενη εκτίμηση του ελαχίστου της συνάρτησης το υποθετικό ολικό ελάχιστο της συνάρτησης - υπάρχει κλειστός τύπος για την περίπτωση τετραγωνικής συνάρτησης.

Ωστόσο, οι παραπάνω μέθοδοι είναι ευάλωτες τόσο στον εγκλωβισμό σε τοπικά ελάχιστα της συνάρτησης, όσο και - ειδικά στην περίπτωση όπου ο χώρος λύσεων του προβλήματος είναι πολυδιάστατος - σε μακροσκελείς υπολογισμούς.

- Ευριστικές μέθοδοι (heuristics). Οι ευριστικές μέθοδοι χρησιμοποιούνται σε περιπτώσεις όπου οι συμβατικοί αλγόριθμοι βελτιστοποίησης είτε αποτυγχάνουν να βρουν λύση, είτε είναι πολύ αργοί για οποιαδήποτε πρακτική χρησιμότητα. Ωστόσο, αυτό επιτυγχάνεται με τη θυσία είτε πλήρους θεμελίωσης, είτε ακρίβειας, για επίτευξη υψηλότερης ταχύτητας. Συχνά, οι ευριστικές δε στοχεύουν - πόσο μάλλον καταλήγουν - στην εύρεση της βέλτιστης λύσης, αλλά μίας η οποία είναι αρκετά καλή δεδομένων των συνθηκών του προβλήματος. Για την καθοδήγηση της ευριστικής διαδικασίας έρευνας στο χώρο καταστάσεων, χρησιμοποιείται συχνά συγκεκριμένο πλαίσιο, το οποίο ορίζεται από την εκάστοτε μετα-ευριστική (meta-heuristic).

Κλείνοντας αυτή την υποενότητα, αξίζει να αναφέρουμε ότι οι ευριστικές μέθοδοι επίλυσης είναι κατάλληλες - με τη χρήση μεταευριστικών για κατάλληλη καθοδήγηση, και κωδικοποίηση των καταστάσεων - για την επίλυση προβλημάτων που ορίζονται σε διακριτό χώρο καταστάσεων. Αντιθέτως, αυτο δεν ισχύει για τις αναλυτικές μεθόδους.

3.7.3 Εξελικτικοί Αλγόριθμοι

Δεχόμενοι τη θεωρία εξέλιξης του Δαρβίνου για την προέλευση των ειδών, μπορούμε να ισχυριστούμε ότι, για την συνεχή πρόοδο και προσαρμογή των έμβιων οργανισμών στο διαρκώς μεταβαλλόμενο γήινο περιβάλλον, θετικούς παράγοντες αποτελούν το μέγεθος του πληθυσμού των εν λόγω οργανισμών, οι διαδικασίες της αναπαραγωγής και της μετάλλαξης, καθώς μόνο μέσω αυτών μπορεί να προκύψουν πληθυσμιακά μέλη με διαφορετικά χαρακτηριστικά, καθώς και η διαδικασία της φυσικής επιλογής μέσω της οποίας, τελικά, επιβιώνουν οι καλύτερα προσαρμοσμένοι οργανισμοί στις παρούσες συνθήκες.

Αυτή είναι και η λογική με την οποία λειτουργεί η οικογένεια μετα-ευριστικών γνωστή και ως εξελικτικοί αλγόριθμοι. Σε γενικές γραμμές, όλες οι πιθανές περιπτώσεις χρήσης τους για την επίλυση προβλημάτων βελτιστοποίησης αποτελούνται από συγκεκριμένα βήματα:

- Αρχικοποίηση πληθυσμού: Αντί, ως αρχική υπόθεση για τη λύση του προβλήματος, να πάρουμε μία τιμή, λαμβάνουμε ως αρχικές εκτιμήσεις ένα σύνολο τιμών, το οποίο έχει προκύψει τυχαία. Το σύνολο αυτό το ονομάζουμε πρώτη γενιά του πληθυσμού.

- Αρχική αποτίμηση πληθυσμού: Αντικαθιστούμε στη συνάρτηση κόστους όλα τα μέλη του αρχικού πληθυσμού, προκειμένου να υπολογίσουμε τις τιμές της συνάρτησης κόστους για αυτά. Στη συνέχεια, και για πεπερασμένο αριθμό επαναλήψεων:

  - Επιλέγουμε τα καλύτερα, βάσει της απόδοσή τους βάσει της συνάρτησης κόστους, μέλη της παρούσας γενιάς του πληθυσμού.

  - Παράγουμε την επόμενη γενιά του πληθυσμού, βάσει των μηχανισμών της μετάλλαξης (όπου προκαλούμε μικρές μεταβολές στα μέλη του πληθυσμού), και της αναπαραγωγής (όπου συνδυάζουμε μεταξύ τους μέλη του πληθυσμού με ικανοποιητικό κόστος).
  - Επανεκτιμούμε την απόδοση της επόμενης γενιάς του πληθυσμού βάσει της συνάρτησης κόστους.
  - Αφαιρούμε από τον πληθυσμό τα λιγότερο αποδοτικά στελέχη του (μέσω ενός μηχανισμού φυσικής επιλογής), για να διατηρήσουμε το μέγεθος κάθε γενιάς πληθυσμού.

- Τέλος, επιλέγουμε, ως λύση του προβλήματος βελτιστοποίησης, το μέλος του πληθυσμού το οποίο έχει πετύχει την καλύτερη απόδοση κατά την τελευταία επανάληψη εκτέλεσης του αλγορίθμου.

4 Περιγραφή Εφαρμογής Σύνθεσης Μουσικής

4.1 Γενική Αρχιτεκτονική της Εφαρμογής

Όπως αναφέρθηκε και στην εισαγωγή, σκοπός της παρούσας διπλωματικής εργασίας ήταν η υλοποίηση μίας εφαρμογής, η οποία με χρήση των σκελετικών δεδομένων που παρέχει το σύστημα παρακολούθησης του Kinect, θα επιτρέπει σε δύο άτομα/παίκτες να συνθέτουν μουσική. Τελικώς, καταλήξαμε σε μία αρχιτεκτονική που αποτελείται από τα εξής υποσυστήματα:

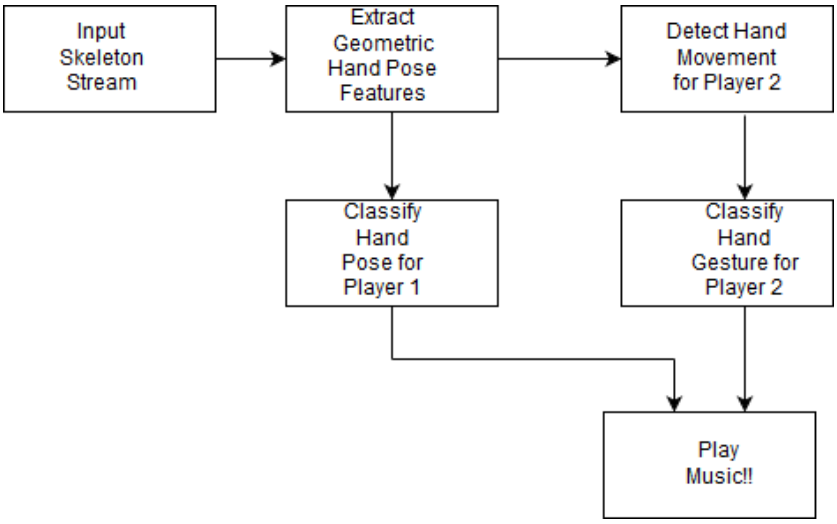
- Καταρχήν, είναι απαραίτητη η ύπαρξη ενός συστήματος το οποίο θα λαμβάνει τα σκελετικά δεδομένα, όπως αυτά παρέχονται από το Kinect, και θα τα επεξεργάζεται ώστε να εξάγει, και για τους δύο παίκτες, κατάλληλα χαρακτηριστικά. Ο λόγος για αυτό είναι ότι τα χαρακτηριστικά που παρέχονται απευθείας από το σύστημα παρακολούθησης του Kinect αφορούν απλά στις (x,y,z) θέσεις των αρθρώσεων, οι οποίες από μόνες τους αποτελούν ασθενές χαρακτηριστικό, υπό την έννοια ότι δε γενικεύουν.

- Σε ό,τι αφορά τον πρώτο παίκτη, ο ρόλος του έγκειται στην παραγωγή μουσικών νοτών. Οι μουσικές νότες που παράγονται εξάγονται μονοσήμαντα από συγκεκριμένες στάσεις (στατικές χειρονομίες) των χεριών του παίκτη. Συνεπώς, απαραίτητη είναι η ανάπτυξη τόσο ενός συστήματος που θα ταξινομεί τις στατικές πόζες σε συγκεκριμένες κατηγορίες, όσο και μίας αντιστοίχισης συγκεκριμένων στατικών ποζών σε συγκεκριμένες νότες.

- Τέλος, καθώς επιθυμούμε την ενσωμάτωση κάποιας περαιτέρω λειτουργικότητας στο σύστημά μας, κρίθηκε απαραίτητη η ανάπτυξη ενός συστήματος φυσικής διεπαφής, υπεύθυνος για τη λειτουργία του οποίου είναι ο δεύτερος παίκτης. Το σύστημα αυτό λειτουργεί βάσει της αντιστοίχισης συγκεκριμένων χειρονομιών σε συγκεκριμένες ενέργειες, όπως η έναρξη λειτουργίας του συστήματος, η αυξομείωση του ρυθμού, και άλλες, όπως περιγράφουμε αναλυτικότερα στην ενότητα 4.4.1. . Η διεπαφή αυτή μπορεί να θεωρηθεί ότι αποτελείται από τα εξής διακριτά υποσυστήματα: έναν ανιχνευτή κίνησης (activity detector) ο οπίος θα ενεργοποιείται όποτε εντοπίζεται κάποια κίνηση που θα μπορούσε δυνητικά να αποτελεί χειρονομία, και έναν ταξινομητή ο οποίος θα ταξινομεί τη χειρονομία που εκτελείται σε κάποια από τις κατηγορίες (ή καμία), και θα επιστρέφει το αποτέλεσμα της ταξινόμησης στο κυρίως πρόγραμμα, προκειμένου να πραγματοποιηθεί η αντίστοιχη ενέργεια.

Σημειώνεται, επιπλέον, ότι ο ρόλος του κάθε παίκτη καθορίζεται από τη θέση του σχετικά με την κάμερα - ο παίκτης αριστερά της κάμερας είναι αυτός που κινεί τα χέρια του παράγοντας νότες, ενώ ο παίκτης δεξιά της, ο υπεύθυνος για τον έλεγχο του προγράμματος και του ρυθμού.

Στις επόμενες υποενότητες, θα δώσουμε περισσότερες πλήροφορίες για τα παραπάνω υποσυστήματα, και τις σχεδιαστικές επιλογές που πραγματοποιήσαμε σε αυτά. Προς το παρόν, παρουσιάζουμε σε επίπεδο block diagram τα διάφορα υποσυστήματα στο Σχήμα 4.1.

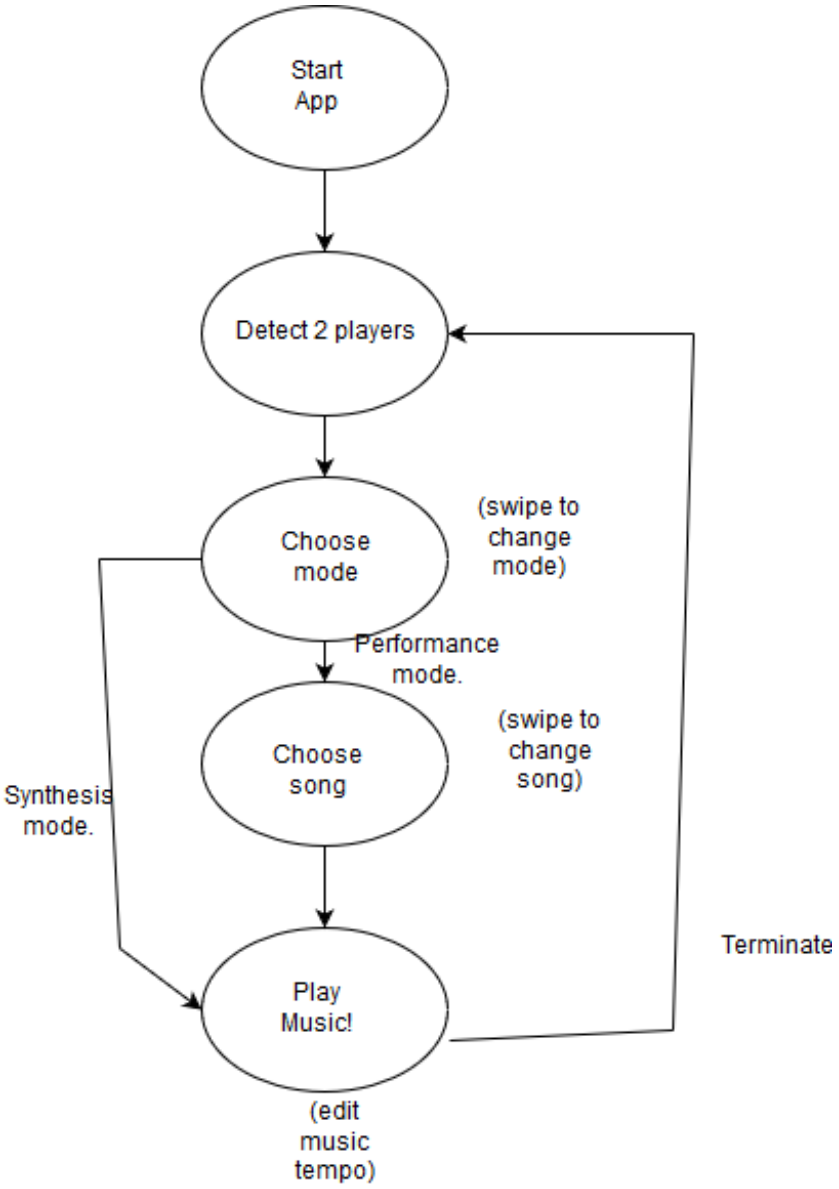


Σχήμα 4.1: Βασικό διάγραμμα του συστήματός μας, σε επίπεδο υποσυστημάτων.

Σε ό,τι αφορά τις δυνατότητες που θα παρέχει η εφαρμογή στους χρήστες, έχουν προγραμματιστεί δύο διακριτά modes λειτουργίας: Το performance mode, στο οποίο οι δύο παίκτες θα συνθέτουν μουσική της επιλογής τους, και το synthesis mode, κατά το οποίο οι δύο παίκτες θα πρέπει να προσεγγίσουν κατά το δυνατό περισσότερο κάποιο δοθέν κομμάτι. Ως προς τις δυνατότητες των παικτών, το εύρος νοτών που θα μπορούν να καλύψουν ισοδυναμεί με μία οκτάβα (7 νότες και 5 ημιτόνια), ενώ η διεπαφή θα πρέπει να παρέχει τις δυνατότητες εκκίνησης/επιβεβαίωσης/τερματισμού, μετακίνησης στα διάφορα μενού επιλογών, καθώς και αυξομείωσης του ρυθμού. Στο Σχήμα 4.2, βλέπουμε μία γενική μορφή του διαγράμματος ροής της εφαρμογής.

Σε επίπεδο διεργασιών που αποτελούν τη συνολική εφαρμογή, διακρίνουμε τις ακόλουθες:

- Την κεντρική διεργασία. Αυτή είναι υπεύθυνη τόσο για τη λήψη, όσο και για την επεξεργασία των δεδομένων από το Kinect, ώστε να εξαχθούν τα κατάλληλα χαρακτηριστικά. Επιπλέον, έχει τον έλεγχο της πορείας του προγράμματος, μέσω του feedback που λαμβάνει από τον ταξινομητή χειρονομιών.
- Τον ανιχνευτή δραστηριότητας. Αυτός λαμβάνει τα δεδομένα κίνησης του δεύτερου παίκτη από την κεντρική διεργασία, και σε περίπτωση που θεωρηθεί ότι έχει ξεκινήσει να πραγματοποιεί κάποια χειρονομία, ξεκινάει να μεταφέρει τα δεδομένα για το επόμενο 1 δευτερόλεπτο στον ταξινομητή χειρονομιών. Επίσης,



Σχήμα 4.2: Αφαιρετικό διάγραμμα ροής της εφαρμογής μας.

έχει την ευθύνη δημιουργίας ενδιάμεσων δεδομένων ανάμεσα σε διαδοχικές επικοινωνίες με την κεντρική διεργασία.

- Τον ταξινομητή δυναμικών χειρονομιών, ο οποίος ξεκινάει τη διαδικασία ταξινόμησης με το που ολοκληρωθεί η λήψη δεδομένων σχετικών με την παρούσα χειρονομία από τον ανιχνευτή δραστηριότητας, και την ταξινομεί σε κάποια κατηγορία ή την απορρίπτει. Αναλόγως του αποτελέσματος της ταξινόμησης, επιστρέφει το αποτέλεσμα στην κεντρική διεργασία.

- Τον ταξινομητή στατικών χειρονομιών, ο οποίος ανά τακτά χρονικά διαστήματα (η συχνότητα καθορίζεται από τον δεύτερο παίκτη) ταξινομεί σε κάποια κατηγορία τη στάση των χεριών του πρώτου παίκτη, βάσει των δεδομένων που λαμβάνει από την κεντρική διεργασία.

- Τέλος, το music player, που δίνει ως έξοδο τη νότα που αντιστοιχεί στη στάση των χεριών του πρώτου παίκτη, όπως έχει αποσταλεί από τον ταξινομητή στατικών χειρονομιών.

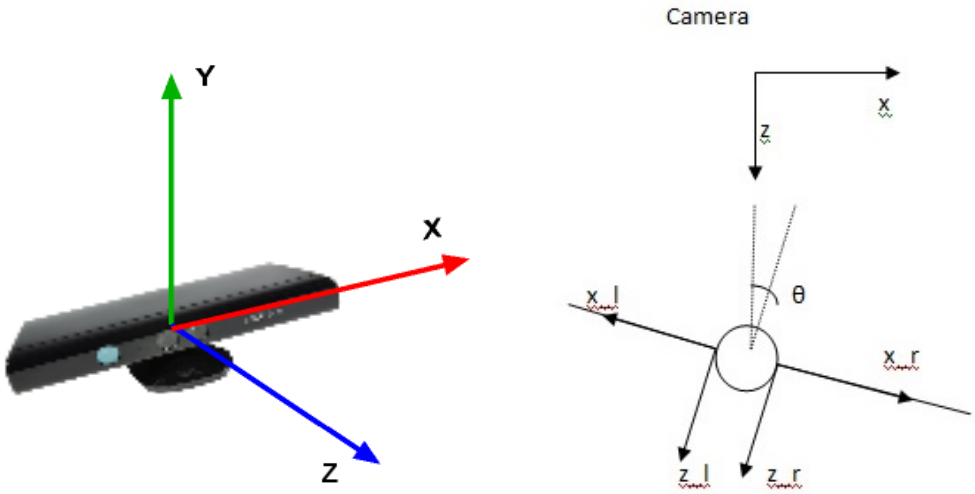
4.2 Εξαγωγή Γεωμετρικών Χαρακτηριστικών

Και για τους δύο παίκτες, είτε σε χρονικά στατικό είτε σε δυναμικό πλαίσιο, χρειαζόμαστε μία αναπαράσταση των σκελετικών δεδομένων τέτοια ώστε να διευκολύνει το χαρακτηρισμό συγκεκριμένων στάσεων των χεριών. Έτσι, για το κάθε τμήμα του κάθε χεριού (από τον αγκώνα και πάνω, και από τον αγκώνα και κάτω) εξάγεται ένα τρισδιάστατο διάνυσμα προσανατολισμού (x', y, z'). Η επιλογή των διανυσμάτων κατεύθυνσης ως χαρακτηριστικό έγινε με κριτήριο την καθολικότητά του – δεν επηρεάζεται σε μεγάλο βαθμό από χαρακτηριστικά συγκεκριμένα ανά άτομο, αλλά γενικεύει ικανοποιητικά. Για να εξαχθεί το διάνυσμα αυτό ωστόσο, πρέπει να λάβουμε υπόψη α) αν επεξεργαζόμαστε το δεξί ή το αριστερό χέρι (καθώς και στις δύο περιπτώσεις αντιστοιχούν τα ίδια πρότυπα διανύσματα, αλλά οι άξονες x στο σύστημα αναφοράς κάθε χεριού έχουν αντίθετη φορά), β) τη γωνία στρέψης του παίκτη (κατά την οποία πρέπει να στραφεί κάθε φορά το διάνυσμα που εξάγεται – που αντιστοιχεί στο πλαίσιο αναφοράς του Kinect – να ευθυγραμμιστεί με το πλαίσιο αναφοράς του χρήστη).

Αρχικά, λοιπόν, εξάγουμε τη γωνία στροφής του παίκτη/χρήστη (έστω θ). Για να την εκτιμήσουμε, θεωρούμε την κατεύθυνση του χρήστη ως αυτήν, που είναι κάθετη στον κορμό του, επί του xz επιπέδου (καθώς η y συνιστώσα αντιστοιχεί σε ύψος). Από τα παρεχόμενα από το Kinect δεδομένα, μπορούμε να ορίσουμε την κάθετη στον κορμό του χρήστη διεύθυνση ως αυτή που είναι κατά το δυνατόν καθετότερη τόσο στο ευθύγραμμο τμήμα που συνδέει τις αρθρώσεις που αναπαριστούν τους δύο ώμους, όσο και σε αυτό που συνδέει αυτές που αναπαριστούν τα άνω τμήματα των ποδιών. Έτσι, υπολογίζουμε τις γωνίες θ_1, θ_2 οι οποίες αντιστοιχούν σε διεύθυνση κάθετη στις προαναφερθείσες, και εξάγουμε την τελική γωνία στρέψης ως το ημιάθροισμα των δύο ανωτέρω γωνιών. Για να μεταβούμε από το επίπεδο της κάμερας στο επίπεδο του χρήστη, συνεπώς, απαιτείται η περιστροφή κατά τη γωνία θ, η οποία και συμβολίζεται με τον παρακάτω πίνακα περιστροφής:

$$R = \begin{bmatrix} cos(\theta) & 0 & -sin(\theta) \\ 0 & 1 & 0 \\ sin(\theta) & 0 & cos(\theta) \end{bmatrix}$$

Στη συνέχεια, εξάγουμε τα διανύσματα προσανατολισμού και για τα δύο χέρια στο επίπεδο της κάμερας. Ξεκινώντας από την αναπαράσταση (x, y, z) που παρέχεται



Σχήμα 4.3: α) Το default (x, y, z) σύστημα συντεταγμένων του Kinect. [16] β) Τα συστήματα αναφοράς για τα δύο χέρια του χρήστη (οι άξονες x, z, εφόσον μας αφήνει αδιάφορους ο y άξονας κατά τη στροφή) σε συνάρτηση με το σύστημα αναφοράς της κάμερας του Kinect, και η γωνία στροφής θ.

από το Kinect για τις θέσεις των διάφορων αρθρώσεων, παίρνουμε τα μήκη των δύο χεριών, τόσο κατά κατεύθυνση όσο και κατά απόλυτη τιμή, βάσει των τύπων:

$$x\_upper = x\_shoulder - x\_elbow \tag{4.1}$$
$$x\_lower = x\_elbow - x\_wrist \tag{4.2}$$
$$y\_upper = y\_shoulder - y\_elbow \tag{4.3}$$
$$y\_lower = y\_elbow - y\_wrist \tag{4.4}$$
$$z\_upper = z\_shoulder - z\_elbow \tag{4.5}$$
$$z\_lower = z\_elbow - z\_wrist \tag{4.6}$$
$$len\_upper = \sqrt{x\_upper^2 + y\_upper^2 + z\_upper^2} \tag{4.7}$$
$$len\_lower = \sqrt{x\_lower^2 + y\_lower^2 + z\_lower^2} \tag{4.8}$$

Για να πάρουμε το κανονικοποιημένο διάνυσμα κατεύθυνσης για κάθε τμήμα του χεριού, διαιρούμε το μήκος του τμήματος στον αντίστοιχο άξονα με το συνολικό μήκος. Έστω $n = [x\_norm, y\_norm, z\_norm]$ το κανονικοποιημένο διάνυσμα για το κάθε τμήμα του χεριού. Τότε, παρατηρούμε ότι, για τα τμήματα που ανήκουν στο δεξί χέρι, η περιστροφή από το επίπεδο της κάμερας στο επίπεδο του χρήστη μπορεί να

πραγματοποιηθεί με τους τύπους:

$$x\_user = cos(\theta)x\_cam - sin(\theta)z\_cam \tag{4.9}$$
$$y\_user = y\_cam \tag{4.10}$$
$$z\_user = sin(\theta)x\_cam + cos(\theta)z\_cam \tag{4.11}$$
$$\tag{4.12}$$

Ή με πολλαπλασιασμό πινάκων:

$$n\_user = Rn\_cam \tag{4.13}$$

Αντίστοιχα, για τα τμήματα που ανήκουν στο αριστερό χέρι, θα ισχύουν οι τύποι:

$$x\_user = -cos(\theta)x\_cam - sin(\theta)z\_cam \tag{4.14}$$
$$y\_user = y\_cam \tag{4.15}$$
$$z\_user = -sin(\theta)x\_cam + cos(\theta)z\_cam \tag{4.16}$$
$$\tag{4.17}$$

Ή, με πολλαπλασιασμό πινάκων:

$$n\_user = Rn'\_cam, n'\_cam = \begin{bmatrix} -x\_cam \\ y\_cam \\ z\_cam \end{bmatrix} \tag{4.18}$$

Τέλος, συνενώνουμε τα παραχθέντα τρισδιάστατα διανύσματα για το κάθε τμήμα των χεριών, παράγοντας το τελικό 12-διάστατο διάνυσμα χαρακτηριστικών:

$$n\_final = \begin{bmatrix} n\_user\_upper\_left \\ n\_user\_lower\_left \\ n\_user\_upper\_right \\ n\_user\_lower\_right \end{bmatrix} \tag{4.19}$$

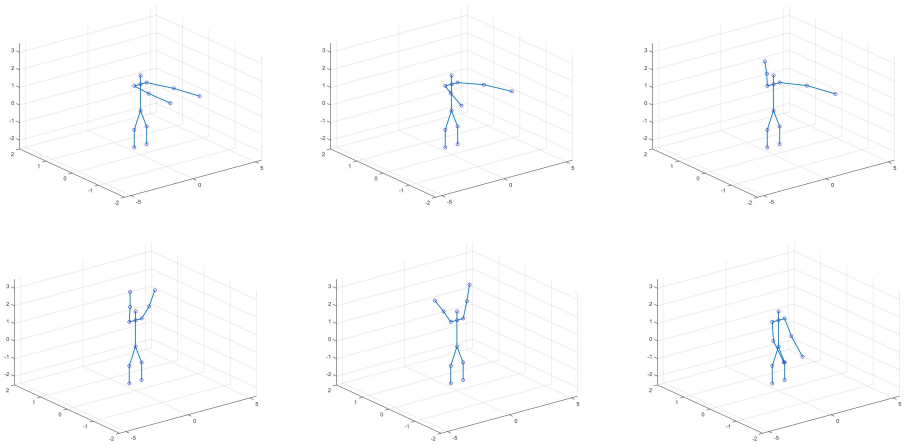
4.3 Υποσύστημα Ταξινόμησης Στατικών Ποζών και Παραγωγής Μουσικής

Σε αυτή την περίπτωση, στόχος είναι, δοθείσης της σκελετικής αναπαράστασης της στάσης των χεριών του παίκτη, η ταξινόμησή της σε κάποια κατηγορία, η οποία και αντιστοιχεί μονοσήμαντα σε κάποια μουσική νότα. Συνεπώς, είναι καταρχήν αναγκαίο να ορίσουμε έναν χώρο από υποψήφιες στάσεις χεριών, να καθορίσουμε τους όρους υπό τους οποίους θα εκτελείται η ταξινόμηση, και να αντιστοιχίσουμε μονοσήμαντα σε καθεμιά από αυτές κάποια μουσική νότα. Έχουμε θέσει ως προδιαγραφή του συστήματος την ικανότητα αναπαραγωγής μίας πλήρους οκτάβας (η οποία αποτελείται από 12 νότες). Συνεπώς, αναγκαίος ήταν ο καθορισμός των 12 στάσεων χεριών που θα αντιστοιχίσουμε στις νότες της οκτάβας. Οι στατικές στάσεις που χρησιμοποιούμε είναι οι παρακάτω (σε σήμανση {στάση αριστερού χεριού, στάση δεξιού χεριού}).

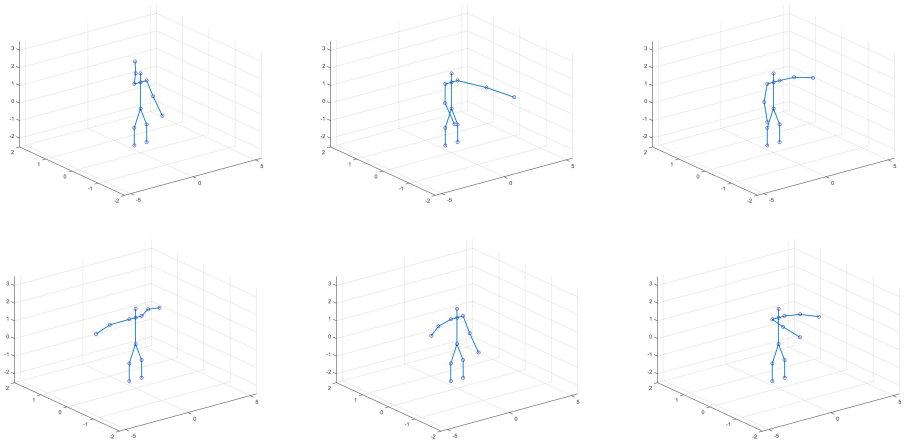
- {Up, Up}
- {Up, Down}
- {Up, Front}
- {Down, Down}
- {Down, Front}
- {Down, Stretched}
- {Stretched, Stretched}
- {Stretched, Down}
- {Front, Front}
- {Front, Stretched}
- {Diagonally Front, Diagonally Front}
- {Diagonally Up, Diagonally Up}

Για την παραγωγή των προτύπων διανυσμάτων της καθεμίας, πραγματοποιήθηκαν 5 εκφάνσεις της κάθε στάσης μπροστά από ένα Kinect, και στη συνέχεια, αφού μετατρέψαμε τα προερχόμενα από το Kinect χαρακτηριστικά στα αντίστοιχα διανύσματα κατεύθυνσης, τα συμπιέσαμε, με χρήση του αλγορίθμου των κ-μέσων, ώστε σε κάθε κατηγορία να αντιστοιχιστεί μοναδική πρότυπη στάση. Στην πράξη, η στάση των χεριών του πρώτου παίκτη ταξινομείται, ανά τακτά χρονικά διαστήματα, σε μία από τις 12 ανωτέρω στάσεις με χρήση απλού template matching. Αν θέλουμε να βρούμε το σύνολο των διανυσμάτων (x, y, z) που ταξινομούνται, τελικά, σε κάθε στάση χεριού, αρκεί η χάραξη του αντίστοιχου τρισδιάστατου διαγράμματος Voronoi με κέντρα, τα κέντρα των προτύπων. Στα Σχήματα 4.3-4.4 βλέπουμε οπτικοποιήσεις των αναπαραστάσεων των χειρονομιών αυτών σε σκελετικό επίπεδο, όπως προέκυψαν από τη χρήση του αλγορίθμου κ-μέσων.

Σε ό,τι αφορά το τελικό κομμάτι του συστήματος, το οποίο παράγει την αντίστοιχη μουσική νότα, αυτό υλοποιήθηκε με χρήση της βιβλιοθήκης ανοιχτού λογισμικού portaudio [77]. Οι παραγόμενοι ήχοι αντιστοιχούν σε έναν ημιτονικό παλμό συγκεκριμένης συχνότητας. Η αντιστοίχιση της συχνότητας που αντιστοιχεί σε κάθε ακέραιο δείκτη, ο οποίος αντιπροσωπεύει συγκεκριμένη στάση των χεριών (όπως έχει προκύψει κατά το προηγούμενο τμήμα του συστήματος) γίνεται με τη χρήση ενός πίνακα, ο οποίος περιέχει τις συχνότητες των 7 νοτών (και 5 ημι-νοτών) μίας οκτάβας. Αναφέρεται ότι γειτονικές νότες ηχητικά είναι παραπλήσιες σε ότι αφορά τις συχνότητές τους, και ότι υψηλή συχνότητα αντιστοιχεί σε πιο οξείς ήχους.



Σχήμα 4.4: Απεικονίσεις των στάσεων χεριών (από πάνω προς τα κάτω και από αριστερά προς τα δεξιά: (Front, Front), (Diagonal Stretched, Diagonal Stretched), (Up, Front), (Up, Up), (Diagonal Up, Diagonal Up), (Down, Down)



Σχήμα 4.5: Απεικονίσεις των στάσεων χεριών (από πάνω προς τα κάτω και από αριστερά προς τα δεξιά: (Up, Down), (Down, Front), (Down, Stretched), (Stretched, Stretched), (Stretched, Down), (Front, Stretched)

Ενδιαφέρον παρουσιάζει ο τρόπος με τον οποίον έγινε η αντιστοίχιση μεταξύ των προτύπων στατικών χειρονομιών και των νοτών. Όπως μπορεί να γίνει εύκολα αντιληπτό, μία επιθυμητή προδιαγραφή του συστήματος αφορά στην τοπικότητα μεταξύ παραπλησίων ηχητικά νοτών (και άρα συχνοτήτων). Στο [58], η αντιστοίχιση έγινε με την παράλληλη εκπαίδευση 2 SOMs. Εδώ, προσεγγίζουμε το πρόβλημα αυτό ως πρόβλημα διακριτής βελτιστοποίησης. Συγκεκριμένα, έχοντας ορίσει:

- Μία μετρική απόστασης μεταξύ των διάφορων στάσεων των χεριών, $d(i, j)$

Note Name	Corresponding Frequency (Hz)
D4#	311.12
E4	329.63
F4	349.23
F4#	369.99
G4	392
G4#	415.31
A4	440
A4#	466.16
B4	493.88
C5	523.35
C5#	554.37
D5	587.33

Table 4.1: Οι μουσικές συχνότητες που αντιστοιχούν σε κάθε ακέραιο δείκτη, οι οποίες και καλύπτουν μία οκτάβα.

- Την αντιστοίχιση, στη στάση k, του δείκτη $idx(k)$ (ο οποίος δηλώνει τη μουσική νότα που αντιστοιχεί σε κάθε στάση και όσο μεγαλύτερος είναι, τόσο πιο υψίσυχνη είναι η παραγόμενη νότα)

Στοχεύουμε να ελαχιστοποιήσουμε το παρακάτω άθροισμα, ως προς τη διάταξη των i, j :

$$\sum_{i=1}^{C} \sum_{j=i+1}^{C} (\frac{idx(i) - idx(j)}{d(i,j)^2}) \tag{4.20}$$

Η αιτιολογία πίσω από την επιλογή της συνάρτησης αυτής προς ελαχιστοποίηση είναι ότι θέλουμε να ποινικοποιήσουμε τις μεγάλες διαφορές στους δείκτες, για τα ζεύγη διανυσμάτων κατεύθυνσης που αντιστοιχούν σε παραπλήσιες στάσεις (και άρα με μικρή μετρική απόστασης μεταξύ τους), επιβραβεύοντας τις μικρές διαφορές για τα αντίστοιχα διανύσματα.

Το πρόβλημα αποτελεί, βάσει του φορμαλισμού του κεφαλαίου 3.7, ένα πρόβλημα διακριτής βελτιστοποίησης. Για την επίλυσή του, χρησιμοποιούμε έναν εξελικτικό αλγόριθμο, του οποίου το κάθε βήμα ισοδυναμεί με τα ακόλουθα:

- Αρχικά, αρχικοποιούμε 50 τυχαίους συνδυασμούς δεικτών. Στη συνέχεια, και για 50 επαναλήψεις:

  - Υπολογίζουμε τη συνάρτηση κόστους (εξίσωση 4.20) για κάθε συνδυασμό δεικτών.
  - Επιλέγουμε τους 4 καλύτερους από αυτούς.
  - Ανανεώνουμε τον πίνακα υποψήφιων συνδυασμών δεικτών, χρησιμοποιώντας
    * Τους 4 καλύτερους αυτούσιους
    * 9 τυχαίες μεταλλάξεις για τον καθένα από αυτούς (κάθε μετάλλαξη προκύπτει από 2 swaps μεταξύ δεικτών και αντίστοιχων θέσεων),
    * 10 καινούριους τυχαίους συνδυασμούς.

- Τελικά, χρησιμοποιούμε τη διάταξη η οποία έχει δώσει το ελάχιστο κόστος.

Static Gesture Code	Pose Description (Left Hand, Right Hand)	Note
S0	(Down, Down)	D4#
S1	(Down, Front)	E4
S2	(Down, Stretched)	F4
S3	(Stretched, Stretched)	F4#
S4	(Stretched, Down)	G4
S5	(Front, Stretched)	G4#
S6	(Diagonal Stretch, Diagonal Stretch)	A4
S7	(Front, Front)	A4#
S8	(Up, Front)	B4
S9	(Diagonal Up, Diagonal Up)	C5
S10	(Up, Up)	C5#
S11	(Up, Down)	D5

Table 4.2: Αριστερά, οι κωδικές ονομασίες που έχουν δοθεί στις στατικές στάσεις σώματος/χεριών – κεντρικά, οι περιγραφές των στάσεων βάσει της θέσης του αριστερού και του δεξιού χεριού του χρήστη, και δεξιά η αντίστοιχη νότα.

Δεχόμενοι ότι στην στάση με τη κωδική ονομασία Si δίνεται ο δείκτης i, στον Πίνακα 4.2. βλέπουμε την τελική αντιστοίχιση στάσεων – δεικτών.

4.4 Υποσύστημα Εντοπισμού και Ταξινόμησης Χειρονομιών

4.4.1 Γενικά

Και εδώ, η προτεινόμενη λύση αποτελείται από δύο διακριτά τμήματα: ένα υποσύστημα το οποίο, εξάγοντας στατιστικά από τα χαρακτηριστικά που έχουν εξαχθεί,

θα αναμένει την έναρξη κάποιας κίνησης η οποία, δυνητικά, θα μπορούσε να αποτελεί κάποια από τις χειρονομίες που χρησιμοποιούμε (activity detector), και ένα υποσύστημα - ταξινομητή το οποίο, ενεργοποιούμενο κατάλληλα από το προηγούμενο, θα μπορεί να ταξινομεί την επακόλουθη κίνηση σε κάποια από τις αποδεκτές κατηγορίες, ή σε καμία.

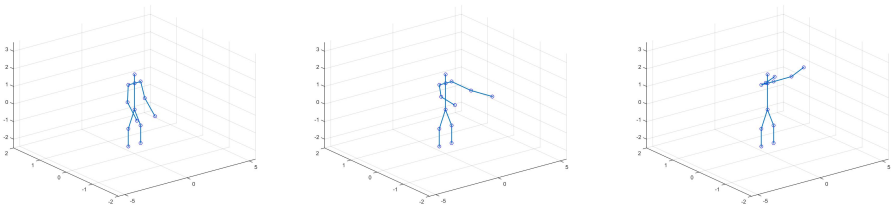
4.4.2 Η Βάση Δεδομένων Microsoft Research Data Center Dataset

Για το σκοπό της υλοποίησης των συστημάτων, χρειαστήκαμε ένα σύνολο δεδομένων, το οποίο να περιέχει αναπαραστάσεις χειρονομιών ως σκελετικά δεδομένα. Αυτό το οποίο, τελικά, χρησιμοποιήθηκε ήταν το Microsoft Research Data Center Dataset. Το dataset αυτό συλλέχθηκε το 2012, στα πλαίσια της εργασίας των Fothergill et al. [78], με χρήση ενός Kinect, και αποτελείται από συνολικά περίπου 7000 εκτελέσεις, 12 συνολικά χειρονομιών/κινήσεων σώματος, οι οποίες και είναι κωδικοποιημένες ως G1 ... G12. Σε όλα τα instances της βάσης δεδομένων, η θέση των διάφορων μελών του ανθρώπου μοντελοποιείται όπως εξάγεται από το middleware του Kinect, δηλαδή με τη μορφή 25 συνδέσμων σε σημεία – κλειδιά (keypoints) του σώματος, στις συντεταγμένες (x,y,z) των οποίων σε σχέση με το επίπεδο της κάμερας έχουμε πρόσβαση. Αυτό κάνει την εν λόγω βάση δεδομένων κατάλληλη για χρήση από την εφαρμογή και το χτίσιμο ενός ταξινομητή (classifier) για τη διακρισιμότητά τους, καθώς α) ούτως ή άλλως χρησιμοποιούμε το skeleton tracking του Kinect για τον εντοπισμό της στάσης σώματος του χρήστη, οπότε έχουμε πρόσβαση στα δεδομένα αυτά και β) η αναπαράσταση του σώματος με χρήση κάποιων keypoints είναι υπολογιστικά βολική.

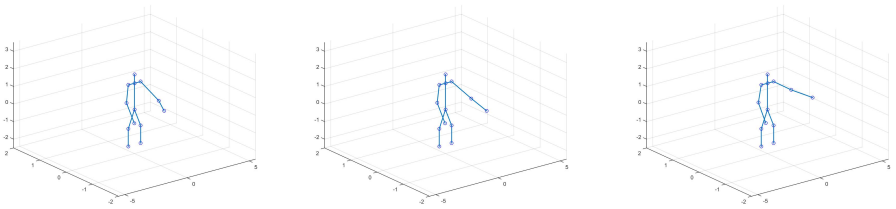
Από τη βάση αυτή δεδομένων, επιλέχθηκε ένα υποσύνολο 5 χειρονομιών που περιέχονται σε αυτήν – με κριτήρια τη διακρισιμότητα και κατά το δυνατό χρήση μόνο των χεριών κατά την εκτέλεση τους. Οι χειρονομίες αυτές είναι οι G1, G3, G5, G9 και G11 , οι οποίες και παρουσιάζονται στα Σχήματα 4.6 - 4.10, ενώ στον Πίνακα 4.3, δίνουμε τη σημασία της καθεμίας από τις χειρονομίες αυτές σε επίπεδο προγράμματος.

Gesture Name (via MSDRC)	Assigned Meaning
G1	Confirm
G3	Next Menu Option
G5	Decrease Music Tempo
G9	Stop Current Session
G11	Increase Music Tempo

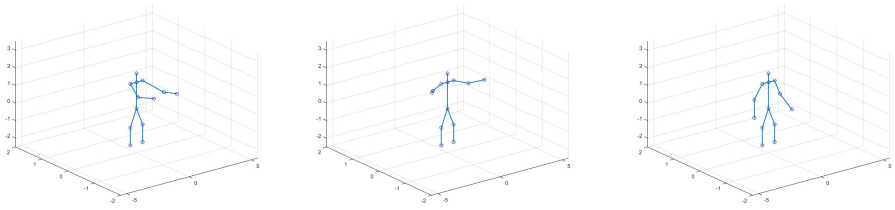
Table 4.3: Αριστερά, οι κωδικές ονομασίες των διάφορων δυναμικών χειρονομιών – δεξιά, το νόημα των χειρονομιών για τη διεπαφή της εφαρμογής μας.



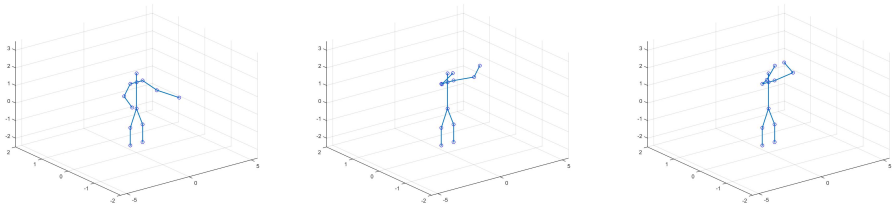
Σχήμα 4.6: Τρία χρονικά στιγμιότυπα εκτέλεσης της χειρονομίας G1



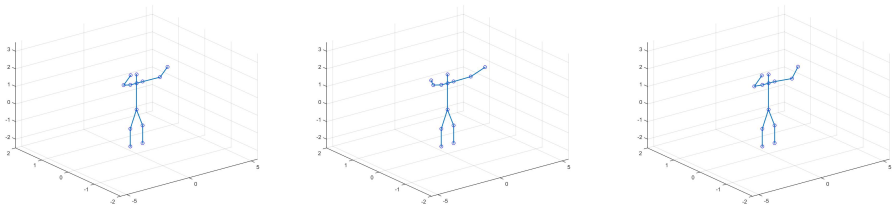
Σχήμα 4.7: Τρία χρονικά στιγμιότυπα εκτέλεσης της χειρονομίας G3



Σχήμα 4.8: Τρία χρονικά στιγμιότυπα εκτέλεσης της χειρονομίας G5



Σχήμα 4.9: Τρία χρονικά στιγμιότυπα εκτέλεσης της χειρονομίας G9



Σχήμα 4.10: Τρία χρονικά στιγμιότυπα εκτέλεσης της χειρονομίας G11

4.4.3 Ανιχνευτής Δραστηριότητας

Γενικώς, μπορούμε να προχωρήσουμε στην παραδοχή ότι, σε επίπεδο εικονοσειρών, για να έχουμε εκτέλεση κάποια δράσης/χειρονομίας, είναι απαραίτητη κάποια μεταβολή μεταξύ διαδοχικών εικόνων. Σε περιπτώσεις όπου έχουμε ως είσοδο συνεχή ροή εικόνων, χρησιμοποιούμε την οπτική ροή μεταξύ των εικόνων ως μέτρο αυτής της μεταβολής.

Εδώ, θέλουμε μία εκτίμηση της ταχύτητας των επιμέρους αρθρώσεων, θεωρώντας όμως την κίνηση του κάθε τμήματος του χεριού ανεξάρτητη των υπολοίπων, κάτι που μπορούμε να πετύχουμε λαμβάνοντας ως ακίνητη την άρθρωση - βάση του κάθε τμήματος και εξετάζοντας την μετατόπιση της ακραίας ως προς αυτήν. Αυτό γίνεται τόσο με χρήση των αρχικών συντεταγμένων για κάθε σημείο, είτε με χρήση του 12-διάστατου

χαρακτηριστικού που έχουμε εξάγει. Στη δεύτερη περίπτωση, έχουμε το επιπρόσθετο πλεονέκτημα ότι το μέγεθός μας είναι ανεξάρτητο του μήκους κάθε τμήματος. Πιο συγκεκριμένα:

Στο κάθε τμήμα χεριού, αντιστοιχεί μία τρισδιάστατη συνιστώσα [x_norm, y_norm, z_norm]. Αφαιρώντας τα μεγέθη αυτά για δύο διαδοχικά frames, έχουμε:

$$\begin{bmatrix} x\_norm(t+1) \\ y\_norm(t+1) \\ z\_norm(t+1) \end{bmatrix} - \begin{bmatrix} x\_norm(t) \\ y\_norm(t) \\ z\_norm(t) \end{bmatrix} = \begin{bmatrix} x\_norm(t+1) - x\_norm(t) \\ y\_norm(t+1) - y\_norm(t) \\ z\_norm(t+1) - z\_norm(t) \end{bmatrix}$$

$$= \frac{1}{l} \begin{bmatrix} l_x(t+1) - l_x(t) \\ l_y(t+1) - l_y(t) \\ l_z(t+1) - l_z(t) \end{bmatrix} = \frac{1}{l} \begin{bmatrix} v_x(t) \\ v_y(t) \\ v_z(t) \end{bmatrix} \tag{4.21}$$

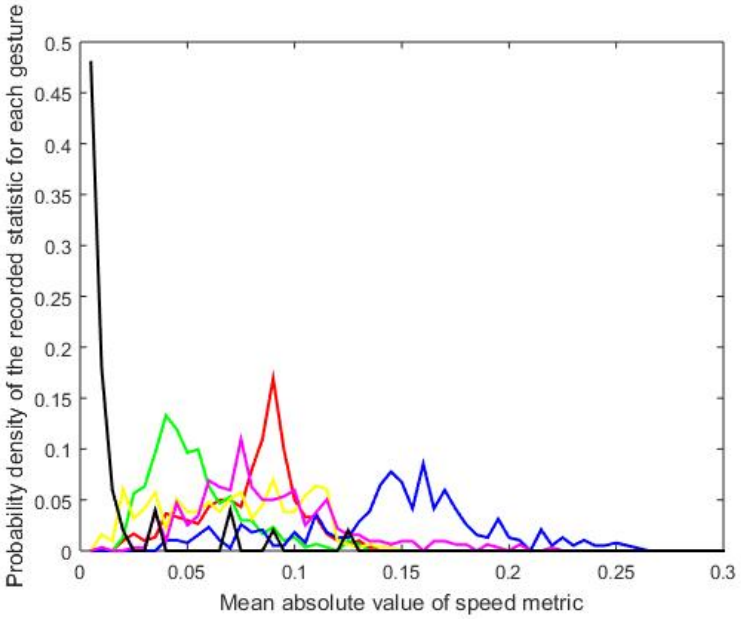
Όπου στο τελευταίο βήμα οι ταχύτητες αντιστοιχούν στα κάτω άκρα των τμημάτων των χεριών, αν θεωρήσουμε τα πάνω ακίνητα.

Εύκολα μπορούμε, συνεπώς, να εξάγουμε μία μετρική ταχύτητας για τα 4 άκρα (αγκώνας και καρπός για κάθε χέρι), υπολογίζοντας το μέτρο του παραπάνω διανύσματος. Ωστόσο, το μέγεθος αυτό είναι ασταθές σε επίπεδο frame, καθώς επηρεάζεται σημαντικά από την ύπαρξη θορύβου. Για το σκοπό αυτό, κρίνεται απαραίτητο να το φιλτράρουμε. Έτσι, τελικά παίρνουμε το μέσο όρο των κατά διεύθυνση ταχυτήτων, σε βάθος 30 frames.

Θεωρούμε, λοιπόν, ότι έχουμε κάποιου είδους δραστηριότητα όταν η μέση ταχύτητα αυτή υπερβαίνει κάποια τιμή κατωφλίου (threshold value). Για την εύρεση μίας καλής τιμής κατωφλίου, εξάγουμε το στατιστικό αυτό α) από τις εκτελέσεις των χειρονομιών όπως μας δίνονται στη βάση δεδομένων και β) από instances στα οποία δεν έχουμε εκτέλεση χειρονομίας. Στο Σχήμα 4.11, φαίνονται οι δειγματικές κατανομές πυκνότητας πιθανότητας του μεγέθους αυτού τόσο για τις διάφορες χειρονομίες (έγχρωμες γραμμές), όσο και για τα instances μη – δραστηριότητας (μαύρη γραμμή). Καθώς επιθυμούμε α) στις περισσότερες εκτελέσεις των χειρονομιών, το κατώφλι αυτό να υπερβαίνεται και β) σε συνθήκες ηρεμίας, ακόμα και αν λάβουμε υπόψη μας τον εγγενή θόρυβο του αισθητήρα, η μέση ταχύτητα των αρθρώσεων να είναι κάτω από αυτό, επιλέγουμε ως τιμή κατωφλίου $T = 0.02$.

4.4.4 Ταξινομητής Χειρονομιών I: Αρχική Επιλογή Ταξινομητή

Από τη στιγμή, λοιπόν, που ο ανιχνευτής δραστηριότητας έχει δώσει θετικό σήμα, το οποίο και σηματοδοτεί την έναρξη κάποιας χειρονομίας, για τα επόμενα frames, τα εξαχθέντα χαρακτηριστικά μεταδίδονται στον ταξινομητή χειρονομιών, προκειμένου να αποφανθεί για τον τύπο της χειρονομίας που εκτελέστηκε. Με επισκόπηση επί



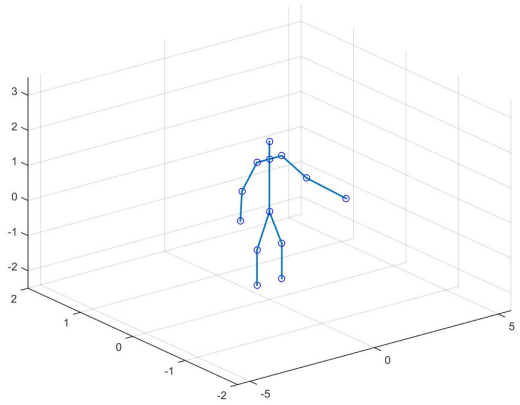
Σχήμα 4.11: Στατιστική απεικόνιση των μετρικών ταχύτητας κατά την εκτέλεση των διάφορων χειρονομιών (έγχρωμο) και υπό συνθήκες ηρεμίας (μαύρο). Πέραν της διακρισιμότητας μεταξύ χειρονομιών και ηρεμίας (για την οποία και χρησιμοποιείται), παρατηρούμε ότι η μετρική αυτή είναι ακατάλληλη για την διάκριση μεταξύ των χειρονομιών.

των δειγμάτων της βάσης δεδομένων, ορίστηκε ως αναμενόμενη/επιθυμητή διάρκεια κάθε χειρονομίας το 1 δευτερόλεπτο, που ισοδυναμεί, για τις συνθήκες μας, με 31 frames.

Για τους σκοπούς της ταξινόμησης, πραγματοποιήθηκε, σε πρώτη φάση, πειραματισμός με τους εξής τύπους ταξινομητών:

- Διακριτοποιημένα κρυφά μοντέλα Markov
- Συνεχή κρυφά μοντέλα Markov
- Αλγόριθμος Κοντινότερου Γείτονα.

Σχετικά με το πειραματικό πρωτόκολλο που ακολουθήσαμε, σε όλες τις περιπτώσεις είχαμε διαχωρισμό των δεδομένων σε δεδομένα εκπαίδευσης και επαλήθευσης, με αναλογία 4:1, και πραγματοποίησαμε 5-fold cross validation. Η εκπαίδευση των μοντέλων πραγματοποιήθηκε, αφού εξήγαμε ως δεδομένα τα χαρακτηριστικά διανύσματα κατεύθυνσης για κάθε instance, το οποίο αντιστοιχεί σε ένα frame. Ακολουθούν λεπτομέρειες για τη διαδικασία που ακολουθήσαμε σε κάθε περίπτωση.



Σχήμα 4.12: Ενδεικτική οπτικοποίηση της κωδικής λέξης 44, από κάποιο codebook (μεγέθους 80) που παρήχθη για τη διακριτοποίηση των συνεχών δεδομένων (διανυσμάτων κατεύθυνσης)

- Διακριτά HMMs: Ξεκινώντας από το σύνολο των διανυσμάτων κατεύθυνσης που έχουν εξαχθεί για κάθε frame, τα συμπυκνώνουμε σε κάποιο σύνολο οπτικών λέξεων (codebook) προκειμένου να το φέρουμε σε διακριτή μορφή - κατάλληλη για την εκπαίδευση διακριτών HMMs. Για τη δημιουργία του codebook, χρησιμοποίησαμε τον αλγόριθμο μίγματος Γκαουσιανών, προκειμένου οι διάφορες συστάδες – λέξεις του codebook να έχουν μεγαλύτερη ευελιξία ως προς το σχήμα τους. Για λόγους ευστάθειας, όπως προαναφέρθηκε, χρησιμοποιήθηκε ο αλγόριθμος των κ-μέσων για την αρχικοποίηση των παραμέτρων εισόδου του αλγορίθμου μίγματος Γκαουσιανών. Εν συνεχεία, εκπαιδεύτηκαν 5 HMMs, το καθένα μονοσήμαντα αντίστοιχο σε μία χειρονομία. Τα HMMs είναι πρώτης τάξης και forward-only (επιτρέπουν μεταβάσεις μόνο προς τις «επόμενες» καταστάσεις), και ως παρατηρήσεις θεωρούμε τις λέξεις του codebook. Η εκπαίδευση πραγματοποιήθηκε μέσω του αλγορίθμου Baum - Welch. Πρακτικά, θεωρούμε ότι κάθε χειρονομία αποτελείται από καταστάσεις, και η κάθε κατάσταση αντιστοιχεί σε συγκεκριμένη στάση χεριών/οπτική λέξη.

- Συνεχή HMMs[2] : Εδώ, η μετατροπή των δεδομένων σε μορφή κατάλληλη για την εκπαίδευση των μοντέλων δεν αποτέλεσε ξεχωριστό βήμα. Αντίθετα, καθώς η εκπαίδευση των μοντέλων έγινε μέσω της παραλλαγής του αλγορίθμου Baum-Welch, η έυρεση των κατάλληλων Γκαουσιανών, ώστε να μοντελοποιηθούν οι παρατηρήσεις που αντιστοιχούν σε κάθε κατάσταση, και η προσαρμογή των παραμέτρων του κάθε HMM συμβαίνουν παράλληλα. Όπως και προηγουμένως, εκπαιδεύσαμε 5 HMMs, ένα για κάθε χειρονομία. Η αρχικοποίηση των

[2]Χρησιμοποιήθηκε το πακέτο συναρτήσεων ανοιχτού λογισμικού netlab [79]

στατιστικών παραμέτρων των μοντέλων έγινε με βάση μία εκτίμηση του πίνακα μεταβάσεων, βασισμένη στον αριθμό των καταστάσεων και στο γεγονός ότι οι ακολουθίες μας είχαν μήκος 31 frames.

- Αλγόριθμος Κοντινότερου Γείτονα: Σε πρώτη φάση, χρησιμοποιήσαμε τον απλό αλγόριθμο του κοντινότερου γείτονα. Η μοναδική τροποποίηση που πραγματοποιήσαμε στα δεδομένα εκπαίδευσης αφορούσε την αποσύνδεση του χρονικού παράγοντα από τη δομή τους, μετατρέποντας τον $31x12$ πίνακα σε μονοδιάστατο διάνυσμα (μήκους 372).

Στον Πίνακα 4.3, βλέπουμε μία συγκριτική απεικόνιση της ακρίβειας ταξινόμησης που αντιστοιχεί σε καθένα από τα 3 βασικά set-ups αλγορίθμων που χρησιμοποιήθηκαν. Στο ενδιάμεσο κελί, επισημαίνονται οι βέλτιστες υπερπαράμετροι που βρέθηκαν για κάθε περίπτωση. Ο ενδιαφερόμενος αναγνώστης μπορεί να βρει περαιτέρω πληροφορίες για την ταξινόμηση τόσο για κάθε πιθανό συνδυασμό υπερπαραμέτρων, όσο και για τις λανθασμένες ταξινομήσεις ανά κλάση, με χρήση της αντίστοιχης Μήτρας Σύγχυσης (Confusion Matrix), στο παράρτημα.

Algorithm	Optimal Setup	Accuracy
Discrete HMMs	7 states, 80 codewords	93.11%
Continuous HMMs	5 states, 5 mixture comps	97.11%
Nearest Neighbor	1 voting neighbor	**98.57%**

Table 4.4: Συγκριτική απεικόνιση της απόδοσης, επί της ταξινόμησης του συνόλου δεδομένων μας, των 3 αλγορίθμων που μελετήσαμε

Σχετικά με τα αποτελέσματα, αυτό που δεν προκαλεί εντύπωση είναι η σημαντική βελτίωση που προσφέρει η μοντελοποίηση των παρατηρήσεων με χρήση συστατικών Γκαουσιανών, έναντι της διακριτοποίησής τους, καθώς είναι εμφανές ότι η πρώτη περιγραφή είναι στατιστικά πληρέστερη. Αυτό, πάλι, που προκαλεί εντύπωση είναι το γεγονός ότι ένας φαινομενικά απλοϊκός, και χωρίς να προσφέρει χρονική συσχέτιση μεταξύ των δεδομένων, αλγόριθμος, συγκεκριμένα αυτός του κοντινότερου γείτονα, είναι ικανός να ξεπεράσει σε ικανότητα ταξινόμησης ένα μοντέλο με ικανοποιητική μαθηματική θεμελίωση, το οποίο λαμβάνει υπόψη του τη χρονική εξέλιξη του προς ταξινόμηση φαινομένου. Μία μερική εξήγηση του παραπάνω μπορεί να προσφέρει ο μεγάλος αριθμός δειγμάτων εκπαίδευσης ανά κατηγορία.

4.4.5 Ταξινομητής Χειρονομιών II: Βελτίωση του Αρχικού Μοντέλου

Παρ΄όλα αυτά, όπως αναφέραμε και στην ενότητα 3.3, ο χρόνος εκτέλεσης ενός ταξινομητή βασισμένου στον αλγόριθμο του κοντινότερου γείτονα είναι γραμμικά ε-

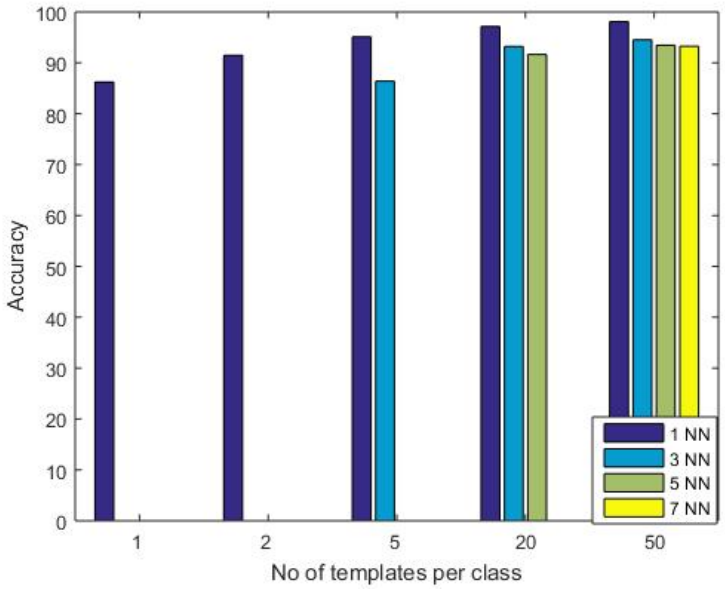
ξαρτώμενος από το μέγεθος των δεδομένων τα οποία χρησιμοποιεί κατά τη διαδικασία της αποτίμησης. Αυτό, σε συνδυασμό με την παράλειψη του χρονικού παράγοντα, μας οδήγησε σε επιπλέον πειράματα σχετικά τόσο με τη συμπίεση, ως προς κάποια διάσταση, του αρχικού συνόλου δεδομένων, όσο και με την εφαρμογή Δυναμικής Χρονοστρέβλωσης προκειμένου να εισάγουμε τη χρονική συσχέτιση ως παράγοντα σύγκρισης. Τέλος, επιχειρείται, ως βελτιωτικός παράγοντας της ταξινόμησης, η απόδοση βαρών στα χαρακτηριστικά.

Μείωση Μεγέθους του Δειγματικού Χώρου: Ως πρώτο πείραμα, επιχειρήθηκε η μείωση του μεγέθους του χώρου δειγμάτων, με ελεύθερη παράμετρο τον αριθμό των templates ανά χειρονομία. Η δημιουργία των templates επετεύχθη εφαρμόζοντας συσταδοποίηση με χρήση του αλγορίθμου των κ – μέσων. Έχουμε να παρατηρήσουμε ότι, όπως φαίνεται τόσο στο Σχήμα 4.7, όσο και στον Πίνακα 4.5, ακόμα και για χρήση μόλις μίας template τα αποτελέσματα είναι – σχετικά – ικανοποιητικά, ενώ με χρήση 20 η ακρίβεια ταξινόμησης πλησιάζει τα ποσοστά επιτυχών ταξινομήσεων της αρχικής υλοποίησης, χρησιμοποιώντας μόλις το 5% της αρχικής βάσης δεδομένων. Επιπλέον, η χρήση περισσότερων «ψηφοφόρων» γειτόνων για ταξινόμηση επηρεάζει, πλέον, εμφανώς αρνητικά την αποτελεσματικότητά της. Αυτό αποδίδεται στη χαμηλή πυκνότητα του δειγματικού χώρου, η οποία εισάγει υψηλότερη τυχαιότητα στις ταξινομήσεις και παράγει πιο «απλοϊκές» επιφάνειες απόφασης.

	1 Neighbor	3 Neighbors	5 Neighbors	7 Neighbors
1 Template	86.22%	-	-	-
2 Templates	91.47%	-	-	-
5 Templates	95.12%	86.41%	-	-
20 Templates	**97.17%**	93.23%	91.66%	-
50 Templates	98.10%	94.52%	93.48%	93.26%

Table 4.5: Ποσοτικοποίηση του Σχήματος 4.13.

Μείωση Διάστασης του Δειγματικού Χώρου: Ως δεύτερο πείραμα, επιχειρείται η μείωση της διάστασης των διανυσμάτων που περιγράφουν τις χειρονομίες, με χρήση του μετασχηματισμού PCA. Έχοντας επιλέξει, βάσει των προηγούμενων αποτελεσμάτων, να εκφράσουμε κάθε κλάση χειρονομίας με χρήση 20 προτύπων (καθώς πετυχαίνουμε ικανοποιητικά ποσοστά ταξινόμησης έχοντας μειώσει σημαντικά τον αριθμό προτύπων με τα οποία συγκρίνουμε την κάθε είσοδο), θέλουμε να δούμε κατά πόσο η επιλογή/κατασκευή κάποιων πρωτευόντων χαρακτηριστικών μπορεί να αναπαράξει τα ίδια αποτελέσματα με χαμηλότερο χρόνο επεξεργασίας. Ο μετασχηματισμός υπολογίζεται μέσω των $20x5 = 100$ προτύπων που έχουμε διατηρήσει για



Σχήμα 4.13: Το ποσοστό επιτυχών αναγνωρίσεων ως συνάρτηση του αριθμού των προτύπων που χρησιμοποιούμε σε κάθε κλάση (από αριστερά προς δεξιά, για κάθε περίπτωση, αυξάνεται ο αριθμός των ψηφοφόρων γειτόνων). Παρατηρούμε τόσο την, αρχικά απότομη, στη συνέχεια ομαλή αύξηση του ποσοστού επιτυχών αναγνωρίσεων, όσο και την αρνητική επίδραση της αύξησης του αριθμού των ψηφοφόρων γειτόνων.

εκπαίδευση. Τα αποτελέσματα είναι ενθαρρυντικά, και απεικονίζονται στον Πίνακα 4.6:

Dimensionality	Accuracy
1	51.34%
2	78.62%
5	94.34%
10	96.80%
20	**97.31%**
50	97.17 %
Baseline	97.17%

Table 4.6: Απεικόνιση του ποσοστού επιτυχών αναγνωρίσεων επί του χρησιμοποιούμενου υποσυνόλου της βάσης δεδομένων MSDRC, ως συνάρτηση του αριθμού διαστάσεων που διατηρούμε στα πρότυπα διανύσματα. Παρατηρούμε ότι ακόμα και για μείωση των 372 σε 10 διαστάσεις, επιτυγχάνεται παραπλήσια επιτυχία στην ταξινόμηση.

Δυναμική Χρονοστρέβλωση: Επιπλέον, επιχειρήθηκε η εφαρμογή του αλγορίθμου της δυναμικής χρονοστρέβλωσης πάνω στα προηγούμενα. Συγκεκριμένα, εφαρμόζεται η παραλλαγή του αρχικού αλγορίθμου για πολυδιάστατες χρονοσειρές [64]. Καθώς η πολυπλοκότητα του αλγορίθμου είναι της τάξης $O(n^2)$ για σύγκριση χρονοσειρών μεγέθους n, ή $O(w^2)$ αν επιτρέψουμε παράθυρο έρευνας εύρους w, χρησιμοποιούμε μειωμένο αριθμό προτύπων ανά χειρονομία προς ταξινόμηση, και καθώς θέλουμε να διατηρήσουμε τη χρονική δομή της χειρονομίας ως χρονοσειρά, δεν εφαρμόζουμε Ανάλυση Κύριων Συνιστωσών (καθώς οι μειωμένες διαστάσεις δε θα έχουν κάποια χρονική συσχέτιση). Παρακάτω παρατίθενται τα αποτελέσματα που πήραμε για διάφορες τιμές του παραθύρου έρευνας (βάσει της ζώνης Sakoe – Chiba) του αλγορίθμου (μηδενικό – baseline, 5 frames, 10 frames και ελεύθερο – σε όλη τη διάρκεια του instance), ενώ ο αριθμός των προτύπων ανά χειρονομία μεταβάλλεται από 1 μέχρι και 20. Βάσει των αποτελεσμάτων των προηγουμένων πειραμάτων, χρησιμοποιούμε αναζήτηση ενός κοντινότερου γείτονα. Διαισθητικά, αναμένουμε ορισμένη βελτίωση σε σύγκριση με το baseline, ειδικά στις περιπτώσεις όπου χρησιμοποιούμε λιγότερα πρότυπα, με πιθανότητα χειρότερων ποσοστών – συγκριτικά πάντα με το baseline – όπου χρησιμοποιούμε α) περισσότερα πρότυπα και β) κανέναν χρονικό περιορισμό στο παράθυρο έρευνας. Τα αποτελέσματα παρατίθενται στον Πίνακα 4.7, και σε μεγάλο βαθμό επιβεβαιώνουν τις εμπειρικές εκτιμήσεις μας, καθώς ναι μεν για χαμηλό αριθμό προτύπων η χρήση δυναμικής χρονοστρέβλωσης βελτιώνει τα ποσοστά επιτυχίας, για υψηλότερο ωστόσο επιδρά λιγότερο, ενώ – αμελώντας τις αυξομειώσεις που οφείλονται σε τυχαιοκρατικούς παράγοντες – το αποτέλεσμα της ταξινόμησης σπάνια διαφέρει ανάμεσα στην περίπτωση μήκους ζώνης 10 και σε αυτήν χωρίς περιορισμένη ζώνη έρευνας.

-	Baseline	5-width search	10-width search	Unconstrained
1 Templates	86.22%	88.91%	89.31%	89.35 %
2 Templates	91.47%	92.93%	93.07%	93.04 %
5 Templates	95.12%	95.68%	96.46%	96.02 %
20 Templates	97.17%	**97.84%**	97.43%	97.73 %

Table 4.7: Ποσοστά επιτυχών ταξινομήσεων επί του χρησιμοποιούμενου υποσυνόλου της βάσης δεδομένων MSDRC, για διάφορες τιμές των προτύπων ανά χειρονομία και μεταβάλλοντας ως παράμετρο το παράθυρο έρευνας του αλγορίθμου δυναμικής χρονοστρέβλωσης (ως baseline θεωρείται η μη – εφαρμογή του).

Απόδοση Βαρών Στα Χαρακτηριστικά: Τέλος, εξετάστηκε η απόδοση βαρών στα διάφορα χαρακτηριστικά, ανάλογα με τη σημαντικότητά τους. Η εξέταση γίνεται σε δύο επίπεδα: Κατά πόσο τα βάρη είναι χρονοανεξάρτητα για το κάθε χαρακτηριστικό ή αλλάζουν ανάλογα με το χρονικό σημείο στο οποίο βρισκόμαστε, και κατά πόσο η επιλογή διαφορετικών βαρών ανά κλάση οδηγεί σε καλύτερα αποτελέσμα-

τα ταξινόμησης σε σχέση με την επιλογή ενιαίων βαρών για όλες τις κλάσεις. Για τον υπολογισμό των βαρών, χρησιμοποιήθηκε ο εξής τύπος, ο οποίος δίνει υψηλά βάρη σε χαρακτηριστικά με υψηλή διασπορά μεταξύ των κλάσεων (inter-class variance), αλλά χαμηλή μεταξύ των στοιχείων που ανήκουν σε μία κλάση (intra-class variance):

$$w(i) = \frac{interclass\_variance}{\sum_{classes} intraclass\_variance} \tag{4.22}$$

Στην περίπτωση χρήσης τοπικών βαρών, αντί για το άθροισμα των ενδο-κλασικών διασπορών, χρησιμοποιούμε απλά την διασπορά των στοιχείων της εκάστοτε κλάσης. Επιπλέον, για να μην προκληθεί προκατάληψη υπέρ κάποιας κλάσης, κανονικοποιούμε τα βάρη ώστε να αθροίζουν στη μονάδα. Για την εμφάνιση κάποιου ουσιώδους αποτελέσματος, χρησιμοποιήσαμε ως πειραματικό set-up τη χρήση δυναμικής χρονοστρέβλωσης με παράθυρο Sakoe-Chiba μήκους 5, και δύο προτύπων ανά κλάση. Τα αποτελέσματα του πειραματισμού αυτού φαίνονται στον Πίνακα 4.8.

Συμπεράσματα Βάσει Των Ανωτέρω Πειραμάτων: Βάσει των αποτελεσμάτων των ανωτέρω πειραμάτων, και με δεδομένη την απαίτηση για λειτουργία της εφαρμογής σε πραγματικό χρόνο, αποφασίστηκαν τα ακόλουθα:

- Συμπίεση του αρχικού συνόλου δεδομένων, με χρήση 20 δειγμάτων - εκπροσώπων ανά κατηγορία.
- Μη - χρήση του μετασχηματισμού Ανάλυσης σε Κύριες Συνιστώσες.
- Απόδοση χρονοανεξάρτητων βαρών στα χαρακτηριστικά.
- Μη χρήση του αλγορίθμου δυναμικής χρονοστρέβλωσης, για χρονική ευθυγράμμιση - καθώς χρονικά, η ισοδύναμη διάταξη του συστήματος απαιτεί μοναδικό δείγμα ανά κατηγορία, στην οποία περίπτωση - όπως είδαμε - το ποσοστό επιτυχών ταξινομήσεων ήταν χαμηλότερο.

Weight Globality	Time-dependence	Accuracy
-	-	92.93%
Global	No	**93.55%**
Local	No	89.69%
Global	Yes	92.26%
Local	Yes	92.07%

Table 4.8: Αποτελέσματα του πειραματισμού με την απόδοση βαρών στα διάφορα χαρακτηριστικά. Παρατηρούμε ότι η μόνη περίπτωση στην οποία έχουμε βελτιωμένη συμπεριφορά είναι αυτή της χρήσης καθολικών και χρονοανεξάρτητων βαρών.

4.4.6 Ταξινομητής Χειρονομιών ΙΙΙ: Επανακατωφλίωση

Είθισται οι αλγόριθμοι ταξινόμησης να δίνουν στην έξοδό τους, πέρα από την απόφανσή τους σχετικά με την κατηγορία όπου ανήκει το προς ταξινόμηση δείγμα, και κάποιο σκορ ή βαθμό εμπιστοσύνης (confidence score), ο οποίος δηλώνει τη βεβαιότητα με την οποία πραγματοποιήθηκε η ταξινόμηση. Στην περίπτωση χρήσης ταξινομητών βασισμένων στον αλγόριθμο του κοντινότερου γείτονα, μπορούμε να θεωρήσουμε ως σκορ του αλγορίθμου την απόσταση του δείγματος από το πλησιέστερό του. Κατά συνέπεια, είναι βάσιμη η υπόθεση ότι, αν η απόσταση αυτή για κάποιο δείγμα είναι αρκετά υψηλή, είναι πιθανό το δείγμα προς ταξινόμηση να μην ανήκει σε κάποια κατηγορία. Για το σκοπό αυτό, και καθώς - προφανώς - σε ένα υποσύστημα αναγνώρισης και ταξινόμησης χειρονομιών σε πραγματικό χρόνο, υπάρχει πάντα η πιθανότητα να εκτελεστεί κάποια άγνωστη - για το σύστημα - χειρονομία, κρίθηκε απαραίτητο να ενσωματωθεί στο σύστημα και ένας μηχανισμός απόρριψης όσων χειρονομιών ταξινομούνται σε κάποια κατηγορία, με απόσταση πάνω από ένα συγκεκριμένο κατώφλι T. Για την εύρεση του κατωφλίου, πραγματοποιήσαμε τα κάτωθι:

Χρησιμοποιώντας διαχωρισμό των δεδομένων σε δεδομένα εκπαίδευσης, επαλήθευσης και αποτίμησης με αναλογία 3:1:1, και εφαρμόζοντας και εδώ 5-fold cross validation, εναλλάσσοντας κυκλικά τα 5 folds, αρχικά κατασκευάσαμε μία συμπιεσμένη αναπαράσταση του συνόλου χειρονομιών μας, από τα δεδομένα εκπαίδευσης, χρησιμοποιώντας 20 πρότυπα ανά κλάση. Στη συνέχεια, από τα δεδομένα επαλήθευσης, βρέθηκε η βέλτιστη τιμή για το κατώφλι απόστασης. Για το σκοπό αυτό, υπολογίστηκε, για την κάθε έκφανση από αυτά, η ελάχιστη απόσταση από α) το πρότυπο της κλάσης στην οποία ταξινομήθηκε η έκφανση, β) οποιοδήποτε πρότυπο ανήκει σε κάποια κλάση στην οποία η έκφανση δεν ταξινομήθηκε. Παράγοντας, εν συνεχεία, δειγματικές κατανομές πυκνότητας πιθανότητας από τα παραπάνω, επιλέξαμε το κατώφλι ώστε α) να μη χάνονται ορθές εκφάνσεις της χειρονομίας (θεωρώντας ορθή την έξοδο του ταξινομητή) και β) να μη γίνονται δεκτές εκφάνσεις που αντιστοιχούν σε άλλες κινήσεις - τις οποίες και μοντελοποιούμε με χρήση των υπόλοιπων χειρονομιών. Δηλαδή, αν T το επιλεχθέν κατώφλι, ορίζοντας τις πιθανοτικές κατανομές:

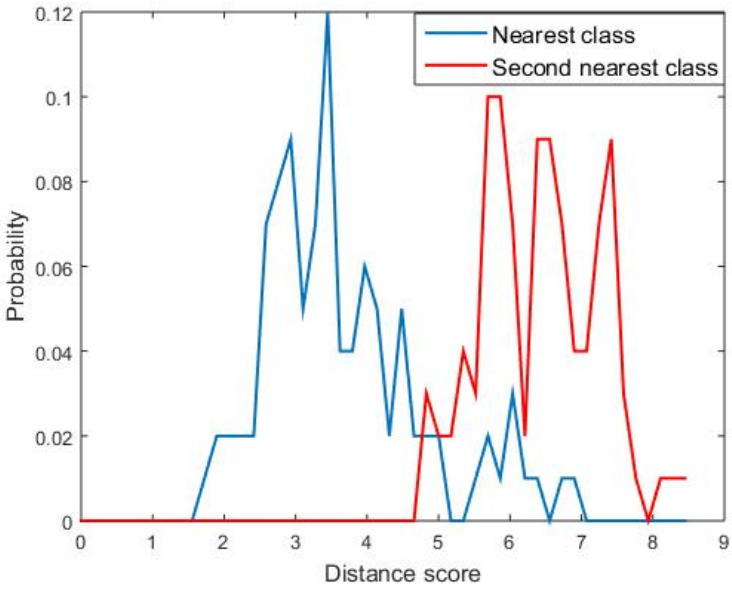
$$p1 = P(min(d(instance, corr\_templ)) > T) \tag{4.23}$$

$$p2 = P(min(d(instance, incorr\_templ)) < T) \tag{4.24}$$

θέλουμε να ελαχιστοποίησουμε την ποσότητα:

$$P(T) = p1 * p2 \tag{4.25}$$

Διερευνήθηκαν οι εξής δύο περιπτώσεις: είτε να παραχθεί ένα γενικό (global) κατώφλι για όλες τις χειρονομίες, είτε να παραχθεί ξεχωριστό κατώφλι για κάθε χειρονομία. Η μεθοδολογία που ακολουθήσαμε, και στις δύο περιπτώσεις, ήταν γενικά η ίδια, με τη διαφορά ότι στην πρώτη περίπτωση το κατώφλι εξήχθη από τις συνολικές δειγματικές συναρτήσεις πυκνότητας πιθανότητας των αποστάσεων, ενώ στη δεύτερη από τις επιμέρους. Και για τις δύο περιπτώσεις, για την αποτίμηση, υπολογίσαμε



Σχήμα 4.14: Ενδεικτική οπτικοποίηση των δύο πιθανοτικών κατανομών του σκορ ταξινόμησης για την α) πλησιέστερη, β) δεύτερη πλησιέστερη κατηγορία. Στην περίπτωση αυτή, μία καλή τιμή κατωφλίου θα ήταν κοντά στο 5.

τις αποστάσεις των δεδομένων αποτίμησης από το πλησιέστερο, σε αυτά, πρότυπο ανά κλάση. Θεωρητικά, το κατώφλι απόστασης θα πρέπει να υπερβαίνεται για κάθε κλάση, εκτός από αυτήν στην οποία ταξινομήθηκε το εκάστοτε instance. Ωστόσο, στην πράξη, έχουμε τα εξής τρία διακριτά δυνατά ενδεχόμενα:

- Υπέρβαση του κατωφλίου για όλες τις κατηγορίες, που ισοδυναμεί με απόρριψη της χειρονομίας από το σύστημά μας.
- Υπέρβαση του κατωφλίου για όλες τις κατηγορίες πλην μίας, που αποτελεί και την επιθυμητή συμπεριφορά.
- Υπέρβαση του κατωφλίου για όλες τις κατηγορίες πλην άνω της μίας, που ισοδυναμεί με την εν δυνάμει λανθασμένη αποδοχή μίας χειρονομίας.

Οπτικοποιούμε τα αποτελέσματα αυτού του πειραματισμού, και για τις δύο περιπτώσεις μας, στον Πίνακα 4.9. Ως ποσοστό επιτυχίας θεωρούμε το ποσοστό των περιπτώσεων όπου είχαμε τη δεύτερη συμπεριφορά. Καθώς - όπως αναμέναμε - η χρήση διαφορετικού κατωφλίου για κάθε χειρονομία αποδίδει καλύτερα, προχωρήσαμε στη χρήση τοπικών κατωφλίων για κάθε τύπο χειρονομίας. Δηλαδή, αναλόγως με το αποτέλεσμα της ταξινόμησης, η αποδοχή ή όχι της χειρονομίας γίνεται με τη χρήση διαφορετικού κατωφλίου.

Threshold Type	Non-classified	Uni-classified	Multi-classified
Global	5.7%	85%	9.3%
Local	4.6%	**89.8%**	5.6%

Table 4.9: Τα ποσοστά κατά τα οποία η ύπαρξη κατωφλίου στη μετρική απόστασης κρατάει καμία, μία ή περισσότερες χειρονομίες ως πιθανόν αποδεκτές ανά εκτέλεση. Όπως αναμένουμε, η χρήση τοπικών κατωφλίων υπερτερεί της χρήσης ενός ολικού κατωφλίου.

5 Πειραματική Αποτίμηση της Εφαρμογής

5.1 Γενικά

Όπως είδαμε και στην προηγούμενη ενότητα, μπορεί να θεωρηθεί ότι το σύστημά μας αποτελείται από τρία διακριτά υποσυστήματα: Το σύστημα ταξινόμησης της πόζας των χεριών, το σύστημα αναγνώρισης και ταξινόμησης χειρονομιών και το μετατροπέα συγκεκριμένης πόζας σε νότα. Για το λόγο αυτό, πριν προχωρήσουμε στην αξιολόγηση της εφαρμογής ως ενιαίας ολότητας, κρίνεται απαραίτητη η αξιολόγηση των επιμέρους υποσυστημάτων της.

Με δεδομένο ότι, στον υπολογιστή στον οποίο γράφτηκε ο κώδικας της εφαρμογής και συνδέθηκε το Kinect προκειμένου να πραγματοποιηθούν τα παρακάτω πειράματα, το Kinect έστελνε τα σκελετικά δεδομένα με ρυθμό μικρότερο των 30 fps (με τον οποίον είχαν ληφθεί τα δεδομένα εκπαίδευσης), έπρεπε να βρεθεί κάποια λύση προκειμένου να ελαχιστοποιήσουμε την απώλεια πληροφορίας. Κρίθηκε ότι η παρεμβολή ενδιάμεσων τιμών, αναλόγως του διαστήματος που μεσολαβούσε μεταξύ δύο διαδοχικών αποστολών, στα σταλθέντα δεδομένα ήταν προτιμότερη της επανεκπαίδευσης του συστήματος αναγνώρισης χειρονομιών. Η παρεμβολή πραγματοποιείται γραμμικά, μεταξύ των αρχικών και των τελικών δεδομένων.

5.2 Πειραματική Αξιολόγηση των Επιμέρους Τμημάτων του Συστήματος

Για την αξιολόγηση των επιμέρους τμημάτων του συστήματος, πραγματοποιήθηκαν τα παρακάτω πειράματα.

5.2.1 Αποτίμηση Ανιχνευτή και Ταξινομητή Δυναμικών Χειρονομιών

Καταρχήν, σε μία προσπάθεια κοινής αξιολόγησης τόσο του ανιχνευτή δραστηριότητας, όσο και του συστήματος ορθής ταξινόμησης των χειρονομιών, πραγματοποιήθηκαν, σε πραγματικές συνθήκες, 8 εκφάνσεις της κάθε χειρονομίας. Ως προς τον ανιχνευτή δραστηριότητας, θεωρούμε ότι έχει πετύχει εφόσον:

- Δεν εντοπίσει εσφαλμένη έναρξη χειρονομίας.
- Δεν έχει «αφήσει» κάποια χειρονομία απαρατήρητη.

Ως προς τον ταξινομητή, θεωρούμε ότι έχει πετύχει εφόσον έχει αναγνωρίσει ορθά τον τύπο της χειρονομίας. Τα αποτελέσματα του δεύτερου σκέλους του πειραματι-

σμού παρατίθενται στον Πίνακα 5.1. – σε ότι αφορά το πρώτο πείραμα, έχουμε να παρατηρήσουμε ότι:

- Δεν είχαμε false positives, δηλαδή περιπτώσεις όπου ο ανιχνευτής κίνησης εντόπισε κίνηση ενώ δεν υπήρχε.
- Δεν είχαμε ενεργοποίηση του ανιχνευτή χειρονομιών σε δύο εκφάνσεις της χειρονομίας G3, κάτι που ωστόσο πιθανόν οφείλεται σε προσωρινή απώλεια του συστήματος παρακολούθησης.
- Από τις εναπομείνασες 38 εκφάνσεις, είχαμε ορθή ταξινόμηση της χειρονομίας στις 35. Το ποσοστό επιτυχίας του ταξινομητή ισούται με 92.10%. Όπως βλέπουμε και από την αντίστοιχη μήτρα σύγχυσης, η συνηθέστερη αποτυχημένη ταξινόμηση αφορά σε εκφάνσεις της χειρονομίας G11 που ταξινομούνται ως G5.

	Pred. G1	Pred. G3	Pred. G5	Pred. G9	Pred. G11
True G1	8	-	-	-	-
True G3	-	6	-	-	-
True G5	1	-	7	-	-
True G9	-	-	-	8	-
True G11	-	-	2	-	6

Table 5.1: Πίνακας Σύγχυσης, ο οποίος προέκυψε κατόπιν της εκτέλεσης 8 εκφάνσεων κάθε χειρονομίας σε πραγματικό χρόνο. Αν εξαιρέσουμε τις 2 εκφάνσεις κατά τις οποίες δεν είχαμε ενεργοποίηση κανενός ταξινομητή, το συνολικό ποσοστό επιτυχίας ισούται με $35/38 = 92.10\%$.

5.2.2 Αποτίμηση Ταξινομητή Στατικών Χειρονομιών

Για την αποτίμηση του στατικού ταξινομητή χειρονομιών, το πειραματικό πρωτόκολλο που ακολουθήθηκε ήταν διαφορετικό, από τη στιγμή που θέλουμε αναγνώριση κάποιας στάσης σε συνεχή ροή, αντί μεμονωμένων instances. Έτσι, σε ενιαίες λήψεις, πραγματοποίηθηκαν εκφάνσεις όλων των στάσεων σώματος που αντιστοιχούν σε συγκεκριμένες νότες, για διάστημα προσεγγιστικά ίσο με 5 δευτερόλεπτα για κάθε στάση. Η ανωτέρω διαδικασία επαναλήφθηκε 5 φορές, και στον Πίνακα 5.2. παραθέτουμε τα αποτελέσματα, με τη μορφή πίνακα σύγχυσης (όπου αποτιμούμε την ορθότητα της ταξινόμησης ανά δευτερόλεπτο – με αποτέλεσμα να έχουμε συνολικά 25 instances ανά στάση). Όπως παρατηρούμε, στη συντριπτική πλειοψηφία των περιπτώσεων, έχουμε ορθή ταξινόμηση της στατικής χειρονομίας, με το ποσοστό ορθών ταξινομήσεων να ισούται με 99.33%.

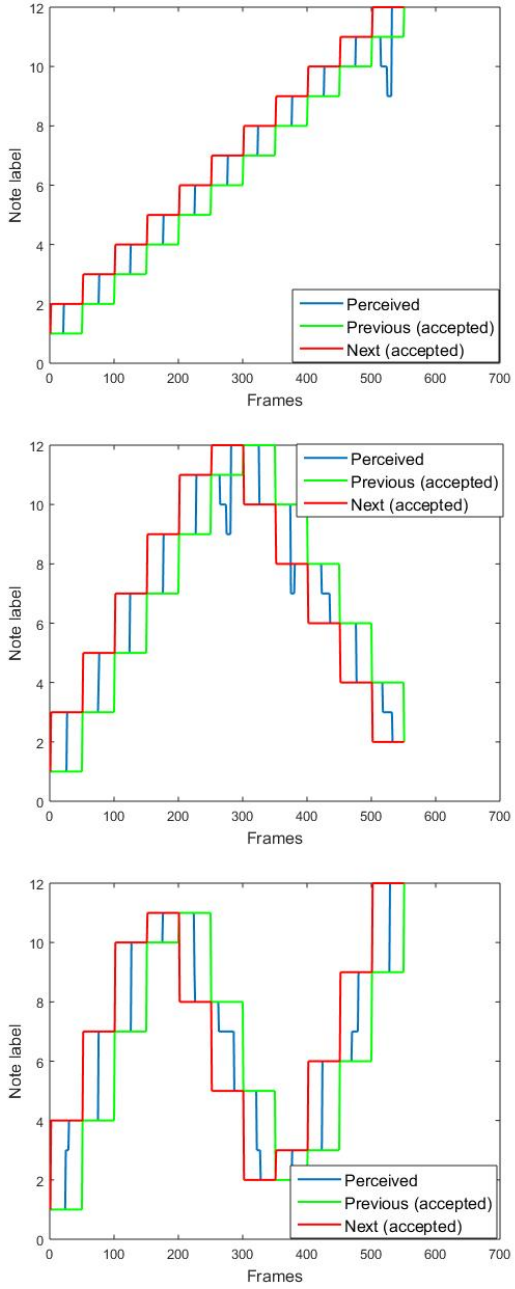
	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
S0	25											
S1		25										
S2			25									
S3				25								
S4					25							
S5						25						
S6						2	23					
S7								25				
S8									25			
S9										25		
S10											25	
S11												25

Table 5.2: Πίνακας Σύγχυσης, ο οποίος προέκυψε κατόπιν πραγματοποίησης των στατικών στάσεων χεριών. Η δομή είναι ίδια με τον προηγούμενο πίνακα, με τις στήλες να αντιστοιχούν στην έξοδο του ταξινομητή και τις γραμμές στην ground truth. Το συνολικό ποσοστό ακρίβειας ισούται με $298/300 = 99.33\%$.

5.2.3 Αποτίμηση Συστήματος Απόδοσης Νότας ανά Στατική Χειρονομία

Τέλος, κρίθηκε απαραίτητη η αξιολόγηση του μηχανισμού απόδοσης ξεχωριστής μουσικής νότας σε κάθε στάση σώματος, ανεξάρτητα από το σύστημα εντοπισμού της εκάστοτε στάσης. Για το σκοπό αυτό, προσομοιώθηκαν στο matlab όλες οι πιθανές μεταβάσεις μεταξύ στάσεων σώματος με απόσταση 1, 2, και 3 νότες αντίστοιχα (συνολικά 3 αλληλουχίες 11 μεταβάσεων, εφόσον προσθέσαμε και επιπλέον μεταβάσεις στις δύο τελευταίες περιπτώσεις). Θεωρούμε ότι έχουμε επιτυχή μετάβαση μεταξύ των δύο θέσεων, όταν καμία ενδιάμεση θέση μεταξύ της αρχικής και της τελικής (θεωρώντας ταχύτητα ίση με $30fps$, παρεμβάλλουμε 49 ενδιάμεσες στάσεις ανάμεσα σε δύο συγκεκριμένες) δεν αποδοθεί σε κάποια τρίτη θέση (που ισοδυναμεί με αντιστοιχία σε τρίτη νότα). Παραθέτουμε α) στα διαγράμματα 5.1α - 5.1γ οπτικοποιήσεις των μεταβάσεων, β) και μία αριθμητική αποτίμηση του παραπάνω στον Πίνακα 5.3. Αξίζει να σημειωθεί ότι τα σφάλματα αυτά δεν θεωρούνται σφάλματα του ταξινομητή, εφόσον η ενδιάμεση στάση εισόδου μπορεί να απέχει αρκετά από όλες τις πρότυπες.

Ως σχόλια, αναφέρουμε ότι τα αποτελέσματα που λάβαμε είναι ικανοποιητικά. Στις περισσότερες περιπτώσεις, οι στατικές χειρονομίες που έχουν αντιστοιχιστεί σε γειτονικές ηχητικά νότες είναι, πράγματι, γειτονικές χωρικά, και όλες οι ενδιάμεσες θέσεις αντιστοιχίζονται σε μία από τις δύο νότες. Επίσης, όπως αναμέναμε, όσο αυξάνεται η δυσκολία της μετάβασης - η ηχητική απόσταση, δηλαδή, δύο στάσεων, τόσο πιο



Σχήμα 5.1: Σε καθένα από τα παραπάνω διαγράμματα, με πράσινο και κόκκινο έχουμε, για κάθε χρονική στιγμή, τις αποδεκτές τιμές εξόδου, οι οποίες αντιστοιχούν στην προηγούμενη και στην επόμενη πρότυπη στάση. Η έξοδος του ταξινομητή, στα ίδια διαγράμματα, εικονίζεται με μπλε.

Distance	# of Transition Frames	# of "Mistakes"	Accuracy
1	551	17	96.91 %
2	501	35	93.01 %
3	451	46	89.80 %
Total	1503	98	**93.48** %

Table 5.3: Ποσοτικοποίηση των παραπάνω διαγραμμάτων. Όπως αναμένουμε, η αύξηση της απόστασης μεταξύ δύο στάσεων χεριών προκαλεί τη μείωση της δυνατότητας άμεσης μετάβασης από τη μία στην άλλη.

πιθανό είναι να μην μπορεί να πραγματοποιηθεί άμεση μετάβαση από τη μία στάση την άλλη. Αυτό δεν αποτελεί πρόβλημα σε κάθε περίπτωση, αφού η παραγωγή ήχου δε γίνεται συνεχόμενα αλλά κατά διαστήματα, αλλά πιθανόν να αποτελέσει, σε περίπτωση καθυστέρησης της μετάδοσης δεδομένων από τον streamer στον αντίστοιχο ταξινομητή, ή από αυτόν στο υποσύστημα αναπαραγωγής μουσικής.

5.3 Πειραματική Αξιολόγηση του Online Συστήματος

Σε αυτό το σημείο, έχοντας ελέγξει και επιβεβαιώσει τη λειτουργικότητα των επιμέρους υποσυστημάτων του συστήματος, είμαστε έτοιμοι να θέσουμε το συνολικό σύστημα σε λειτουργία.[3] Για την αξιολόγησή του, θα χρησιμοποιήσουμε το performance mode, στο οποίο και θα εκτελεστούν συνολικά από 3 εκφάνσεις καθεμιάς από τις 3 test cases (λόγω προβληματικής καταγραφής, τελικά χρησιμοποίησαμε 2 εκφάνσεις της πρώτης test case, και 3 στις υπόλοιπες δύο περιπτώσεις), οι οποίες φαίνονται συμβολικά στον Πίνακα 5.4 και διαγραμματικά (σε επίπεδο χρόνου/συχνότητας) στο Σχήμα 5.2. Οι testcases εκτελέστηκαν πλήρως, από 2 άτομα ανά περίπτωση, εκ των οποίων ο ένας κινούσε τα χέρια του παράγοντας τις αντίστοιχες νότες, και ο δεύτερος είχε τον έλεγχο της λειτουργίας του συστήματος (οι οδηγίες που δόθηκαν αφορούσαν τα χρονικά σημεία κατά τα οποία έπρεπε να πραγματοποιηθεί κίνηση, ενώ σε κάθε περίπτωση προηγήθηκε εξοικείωση των συμμετεχόντων με το λεξιλόγιο).

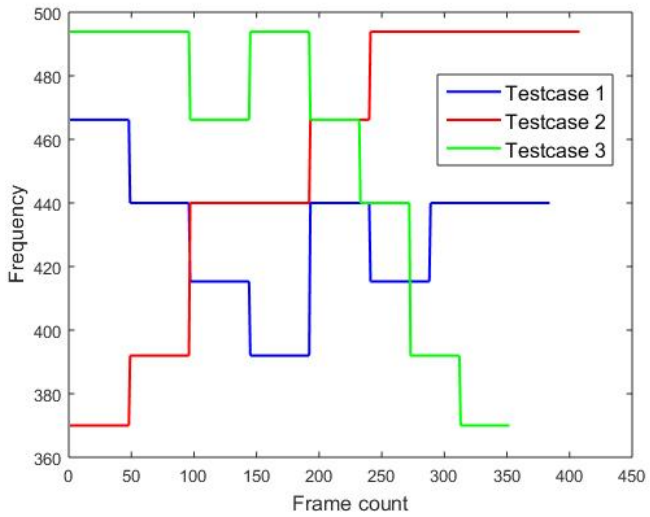
Η αξιολόγηση έγινε σε τρία επίπεδα: Πρώτον, κατά πόσο οι διάφορες χειρονομίες έναρξης και τερματισμού, καθώς και αλλαγής ρυθμού, αποτιμούνται σωστά. Δεύτερον, σε ποιο βαθμό οι νότες που εκλαμβάνει το σύστημα αντιστοιχούν στην εκφραζόμενη, και τρίτον, σε τι βαθμό, τελικά, μπορεί να αναπαραχθεί από το σύστημα ένα οποιοδήποτε μουσικό κομμάτι.

Σε ότι αφορά το πρώτο, η διαδικασία αξιολόγησης είναι παρόμοια με αυτή της

[3]Ευχαριστώ, για την ενεργή συμμετοχή τους σε αυτό το σκέλος του πειραματικού, την αδερφή μου, τον πατέρα μου και τον Κώστα

Testcase	Note Notation
1	A#, A, G#, G, A, G#, A, A
2	F#, G, A, A, A# (down), B, B, B
3	B, B, A#, B (up), A#, A, G, F#

Table 5.4: Τα 3 testcases που χρησιμοποιήθηκαν για την αξιολόγηση του συστήματος. Σε παρενθέσεις, τα σημεία στα οποία πραγματοποιείται χειρονομία αλλαγής ρυθμού.



Σχήμα 5.2: Απεικόνιση των τριών δοκιμαστικών κομματιών - testcases, ως διάγραμμα χρόνου - συχνότητας.

ενότητας 5.2.1. Βρέθηκαν, κατά τη διάρκεια των εκτελέσεων, οι χειρονομίες που αντιστοιχούν σε έναρξη, τερματισμό και αλλαγή ρυθμού, και συγκρίθηκαν με την επισημειωμένη ground truth σε κάθε περίπτωση. Παρατίθενται, με τη μορφή μήτρας σύγχυσης, τα αποτελέσματα του ανωτέρω στον Πίνακα 5.5:

Έχουμε να παρατηρήσουμε ότι τα αποτελέσματα είναι υποδεέστερα από τα αντίστοιχα στην ενότητα 5.2.1, όπου απλά είχαμε πραγματοποίηση χειρονομιών μπροστά από ένα Kinect, καθώς εδώ το συνολικό ποσοστό επιτυχων ταξινομήσεων ισούται με 79.17%. Αυτό μπορεί να οφείλεται τόσο στο μειωμένο frame rate μετάδοσης δεδομένων, όσο και σε πιθανές αλληλοεπικαλύψεις μεταξύ των δύο παικτών.

Σε ό,τι αφορά το δεύτερο, και έχοντας προχωρήσει στην υπόθεση ότι όλες οι στάσεις σώματος του παίκτη αντιστοιχούν στην ορθή για κάθε χρονική στιγμή, κατατάσσουμε κάθε έκφανση νότας σε κάποια από τις 3 εξής κατηγορίες:

- Ορθά ταξινομημένη.

	Pred. G1	Pred. G5	Pred. G9	Pred. G11
True G1	8	2		
True G5	1	2		
True G9		1	7	
True G11			1	2

Table 5.5: Μήτρα Σύγχυσης, στην οποία παρατίθεται η ακρίβεια ταξινόμησης, σε πραγματικό χρόνο, χειρονομιών που πραγματοποίηθηκαν κατά την εκτέλεση της τελικής εφαρμογής. Το συνολικό ποσοστό ακρίβειας, σε αυτή την περίπτωση, ισούται με $19/24 = 79.17\%$.

- Εσφαλμένα ταξινομημένη είτε κατά έναν τόνο, είτε κατά μία χρονική στιγμή.
- Εσφαλμένα ταξινομημένη κατά άνω του ενός τόνου ή άνω της μίας χρονικής στιγμής.

Η εισαγωγή της δεύτερης κατηγορίας καλύπτει τόσο ελαφρά σφάλματα εκ μέρους του παίκτη, όσο και αποτυχή συγχρονισμό του συστήματος αποστολής δεδομένων με το σύστημα παραγωγής μουσικής νότας. Τα αποτελέσματα φαίνονται οπτικοποιημένα στο Σχήμα 5.3:

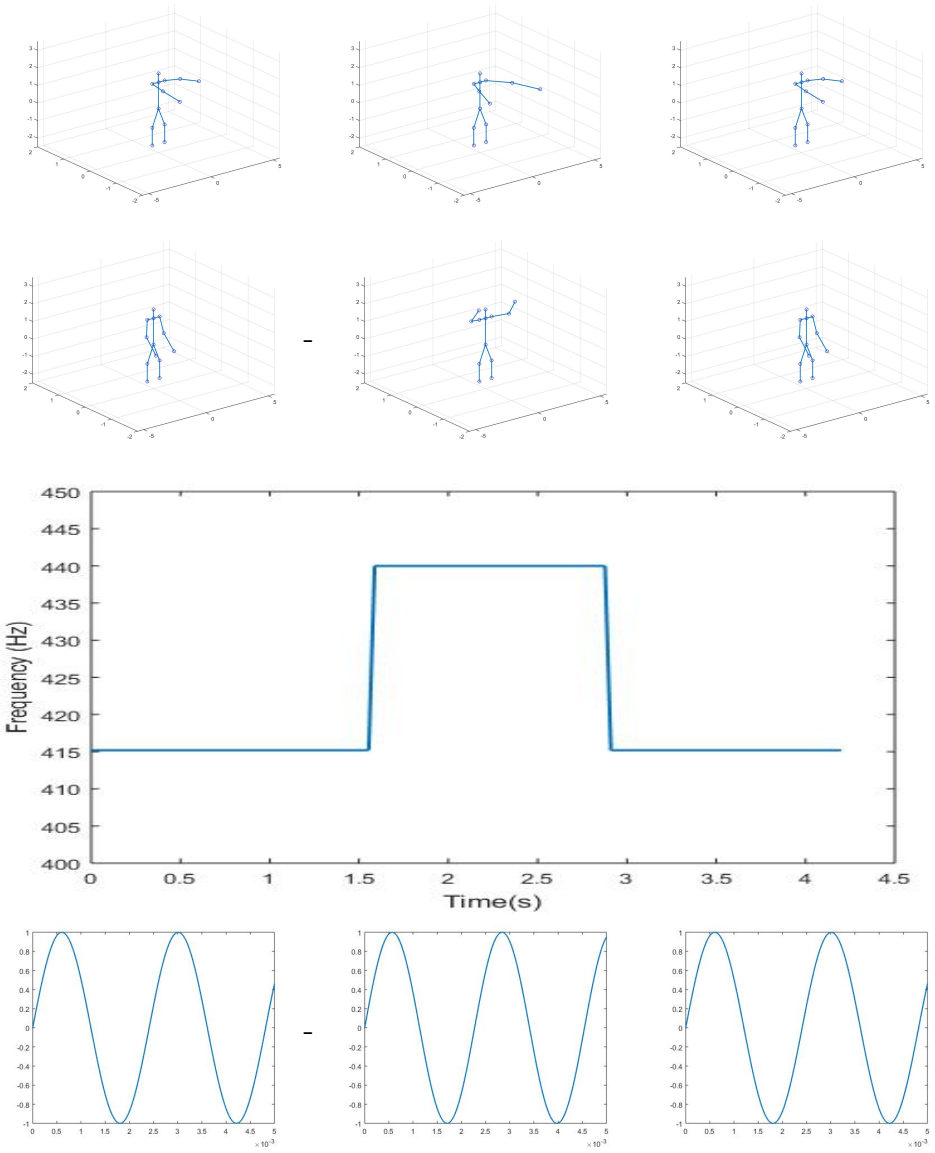
Τέλος, σε ότι αφορά το τρίτο ζητούμενο, για την ποσοτικοποίηση της ικανότητας της αναπαραγωγής ενός κομματιού, κρίθηκε απαραίτητη η ανάπτυξη μίας μετρικής. Δεδομένου ότι η ομοιότητα δύο τραγουδιών, μεταξύ άλλων, μπορεί να αναλυθεί σε ομοιότητα στην τονικότητα και ομοιότητα στο ρυθμό, για κάθε τόνο, ορίζουμε τις μετρικές:

$$note\_sim = (1 - \frac{|index\ of\ played\ note - index\ of\ real\ note|}{11}) \tag{5.1}$$

$$tempo\_sim = (1 - \frac{|tempo - real\ tempo|}{50}) \tag{5.2}$$

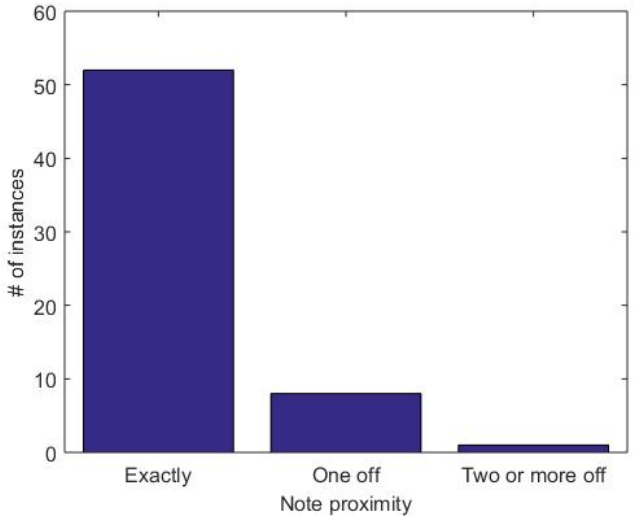
Όπου, για κάθε νότα, υπολογίζουμε την απόλυτη διαφορά της «πραγματικής» με την αντιλαμβανόμενη νότα (διά το 11 που είναι η μέγιστη απόσταση), και την απόλυτη διαφορά του αντιληπτού από τον κανονικό ρυθμό – που ορίζεται ως ο αριθμός των frames που μεσολαβούν μεταξύ δύο διαδοχικών νοτών (διά το 50). Για να ενοποιήσουμε, σε επίπεδο νότας, τις μετρικές αυτές, χρησιμοποιούμε ως συνολική μετρική το γινόμενό τους:

$$c = (note\_sim)\ x\ (tempo\_sim) \tag{5.3}$$



Σχήμα 5.3: Στην πρώτη σειρά, τρία στιγμιότυπα από εκτελέσεις νοτών του πρώτου παίκτη, ενώ στη δεύτερη, από τις κινήσεις του δεύτερου παίκτη (στο δεύτερο στιγμιότυπο έχουμε εκτέλεση της χειρονομίας G11 που αντιστοιχεί σε αύξηση ρυθμού). Η προκύπτουσα έξοδος για αυτή την ακολουθία κινήσεων φαίνεται στις δύο κάτω σειρές, στο πεδίο της συχνότητας (ο y άξονας αντιστοιχεί σε Hz) και - αποσπασματικά - στο πεδίο του χρόνου.

Για ένα σύνολο κομματιού που αποτελείται από N νότες, υπολογίζουμε το μέσο όρο των μετρικών για κάθε νότα. Στον Πίνακα 5.6. βλέπουμε τα αποτελέσματα που λάβαμε για τις εκφάνσεις αυτές:



Σχήμα 5.4: Οπτικοποίηση του αριθμού των εκφάνσεων κάθε νότας (στατικής στάσης) που αντιστοιχούν σε ορθή, ελαφρώς εσφαλμένη ή εσφαλμένη ταξινόμηση.

Instance	Accuracy Metric	Observations
1	1	
1	0.9318	1 note quite off
2	0.988	
2	0.825	Tempo down interpreted as up
2	0.965	
3	0.977	
3	1	
3	0.605	Tempo up interpreted as stop

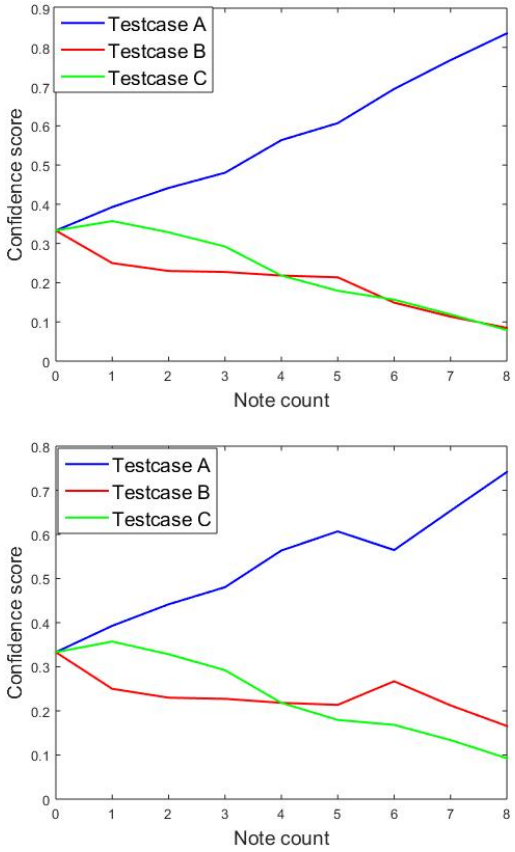
Table 5.6: Τελική ποσοτικοποίηση της μετρικής που μας δίνει την ικανότητα αναπαραγωγής κομματιών από την εφαρμογή. Στις παρατηρήσεις, τα αίτια για τις περιπτώσεις όπου η μετρική ήταν σχετικά χαμηλή.

Όπως βλέπουμε και στον Πίνακα 5.6., η μετρική ομοιότητας χαμηλώνει αισθητά σε περιπτώσεις όπου έχουμε είτε κάποια λανθασμένη αναγνώριση δυναμικής χειρονομίας, είτε κάποιο σημαντικό λάθος στην ταξινόμηση στατικών χειρονομιών. Σε περιπτώσεις όπου είχαμε - όπως φαίνεται στο Σχήμα 5.3 - λάθη μικρής έκτασης σε ότι αφορά την ταξινόμηση στατικών χειρονομιών, η μετρική ομοιότητας δεν επηρεάζεται ιδιαίτερα.

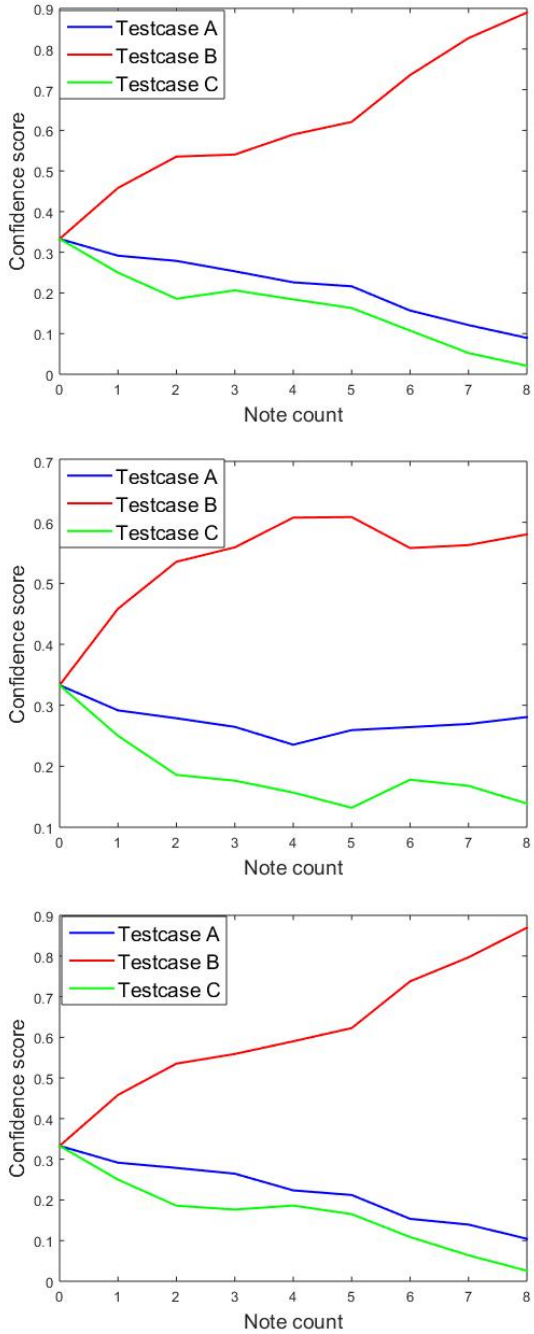
Ως ένα επιπλέον πείραμα, με αφορμή την πιθανή αυξημένη χρηστικότητα στην περίπτωση που δεν πρέπει να επιλέξουν οι χρήστες το τραγούδι που θα παιχτεί, αλλά το σύστημα αντιλαμβάνεται το ποιο τραγούδι παίζεται μέσω των παραγόμενων νοτών και ρυθμού, επιχειρήθηκε η αυτόματη ταξινόμηση των testcases που χρησιμοποίησαμε στα προηγούμενα σε κάποια από τις 3 κατηγορίες από τις οποίες προήλθαν.

Για το σκοπό αυτό, χρησιμοποιήσαμε και πάλι τη μετρική ομοιότητας (τύπος 5.3) μεταξύ κάθε παιγμένης νότας και των αντίστοιχων προτύπων για τη καθεμία από τις 3 κατηγορίες, αλλά με τη διαφορά ότι σε βάθος χρόνου πολλαπλασιάζουμε τις μετρικές αυτές, αντί να πάρουμε το μέσο όρο. Μετά το πέρας κάθε νότας, υπολογίζουμε τη - μέχρι εκείνη τη στιγμή - πιθανότητα να ανήκει το κομμάτι σε κάποια από τις 3 κατηγορίες, κανονικοποιώντας τα σκορ ώστε να αθροίζουν στη μονάδα. Στα Διαγράμματα 5.5 - 5.7, βλέπουμε την εξέλιξη των σκορ στο χρόνο για καθένα από τα 8 instances.

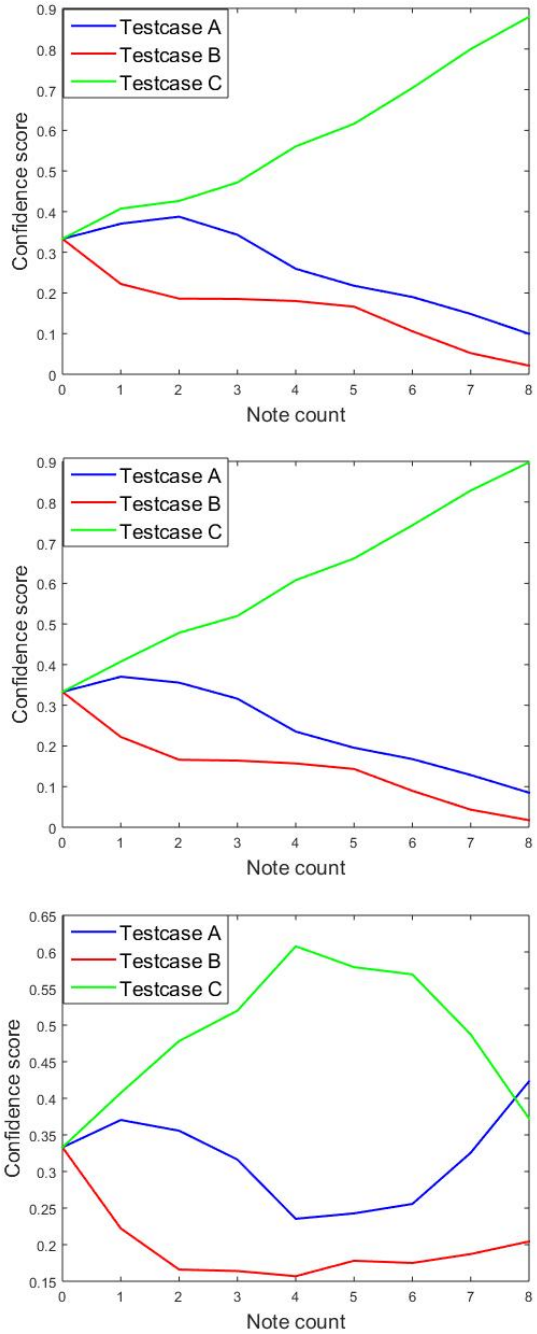
Αν έχουμε να παρατηρήσουμε κάτι επί των αποτελεσμάτων, αυτό είναι η εύστοχη (και σε περιπτώσεις ορθής εκτέλεσης, με μεγάλο περιθώριο) πρόβλεψη του εκτελούμενου κομματιού. Εξαίρεση αποτελεί το instance 3c, στο οποίο η τελική πρόβλεψη είναι εσφαλμένη. Αυτό οφείλεται στην εσφαλμένη μετάφραση της χειρονομίας αύξησης ρυθμού ως σταματήματος, που προκάλεσε μη προβλεπόμενη συμπεριφορά στον εκτιμητή (μέχρι τη στιγμή της λανθασμένης εκτίμησης, η πρόβλεψη ήταν συνεχόμενα ορθή).



Σχήμα 5.5: Εξέλιξη της προβλεπόμενης εξόδου στο χρόνο, για τα 2 instances που αντιστοιχούν στο πρώτο testcase. Με μπλε η πιθανότητα να εκτελείται το πρώτο testcase, με κόκκινο το δεύτερο, με πράσινο το τρίτο.



Σχήμα 5.6: Εξέλιξη της προβλεπόμενης εξόδου στο χρόνο, για τα 3 instances που αντιστοιχούν στο δεύτερο testcase. Με μπλε η πιθανότητα να εκτελείται το πρώτο testcase, με κόκκινο το δεύτερο, με πράσινο το τρίτο.



Σχήμα 5.7: Εξέλιξη της προβλεπόμενης εξόδου στο χρόνο, για τα 3 instances που αντιστοιχούν στο τρίτο testcase. Με μπλε η πιθανότητα να εκτελείται το πρώτο testcase, με κόκκινο το δεύτερο, με πράσινο το τρίτο.

6 Συνεισφορά, Συμπεράσματα και Μελλοντική Έρευνα

Μέχρι στιγμής, η συνεισφορά και τα συμπεράσματα της διπλωματικής μπορούν να συνοψιστούν στα ακόλουθα:

- Επιχειρήθηκε η εισαγωγή του διανύσματος κατεύθυνσης των δύο χεριών ως χαρακτηριστικό προς χρήση σε προβλήματα ταξινόμησης, όπου έχουμε διαθέσιμη την σκελετική αναπαράσταση των αρθρώσεων ενός ατόμου, όπως για παράδειγμα σε περιπτώσεις όπου έχουμε διαθέσιμη ένα Kinect.

- Πραγματοποίηθηκε εκτενής μελέτη κάποιων βασικών αλγορίθμων ταξινόμησης χρονοσειρών ως προς την παραμετροποίησή τους.

- Έγινε, επιπλέον, λεπτομερής καταγραφή διαφόρων μεθόδων συμπίεσης ενός συνόλου δεδομένων, και επαρκής πειραματισμός με αυτές, με σκοπό τη χρήση του για ταχύτερη λειτουργία ενός συστήματος ταξινόμησης.

- Από τα παραπάνω, σε ότι αφορά τη δυναμική των παραπάνω αλγορίθμων, έχουμε να παρατηρήσουμε ότι η προσέγγιση διακριτοποίησης των συνεχών μοντέλων Markov οδηγεί σε υποδεέστερα αποτελέσματα από τη μοντελοποίηση των μεταβλητών που εμπλέκουν ως συνεχείς στατιστικές κατανομές. Ο αλγόριθμος του κοντινότερου γείτονα δίνει, παραδόξως, καλύτερα αποτελέσματα ταξινόμησης, τα οποία δεν υφίστανται σημαντική μείωση αν χρησιμοποιηθεί – σε λογικά πλαίσια – τόσο συμπίεση του συνόλου με χρήση του αλγορίθμου κ – μέσων, όσο και μείωση της διαστατικότητας με χρήση PCA, ενώ η εφαρμογή δυναμικής χρονοστρέβλωσης για ευθυγράμμιση των χρονοσειρών βελτιώνει εμφανώς την αποτελεσματικότητα του αλγορίθμου, ειδικά σε περιπτώσεις όπου κάθε χειρονομία εκπροσωπείται από ελάχιστα πρότυπα.

- Τέλος, χρησιμοποιήθηκαν τα ανωτέρω για την υλοποίηση μίας – κατά τη γνώση μας – πρωτότυπης εφαρμογής σύνθεσης μουσικής από κίνηση. Η τελική απλότητα των χρησιμοποιούμενων αλγορίθμων κάνει δυνατή τη λειτουργία της σε πραγματικό ή περίπου πραγματικό χρόνο.

Παρόλα αυτά, η δουλειά που έχει πραγματοποιηθεί παραμένει ημιτελής, και επιδέχεται αρκετών επεκτάσεων, καθώς η εφαρμογή δουλεύει αποκλειστικά σε περιβάλλον Windows, με κατάλληλους drivers (το Kinect μπορεί να ενσωματωθεί σε περιβάλλον Linux και να μεταδίδει RGBD δεδομένα, αλλά χωρίς το skeleton tracking). Για να λυθούν τα παραπάνω θέματα, πιθανές επεκτάσεις του παρόντος framework στο μέλλον θα ήταν:

- Η διεπαφή προς το χρήστη περιέχει ένα περιορισμένο σύνολο χειρονομιών, οι οποίες εντάσσονται σε ενιαία τροπικότητα με τις κινήσεις των χεριών – και συνεπώς για τη διάκρισή τους, εφόσον δε συμμετέχουν 2 παίκτες αλλά ένας,

απαιτείται κάποια εκδήλωση προδιάθεσης από το χρήστη ότι θα εκτελέσει χειρονομία προς το σύστημα (συγκεκριμένη θέση στο χώρο στην οποία εκτελούνται χειρονομίες για παράδειγμα). Συνεπώς, για την ενίσχυση του συνόλου χειρονομιών, μπορούμε να χρησιμοποιήσουμε και κάποια άλλη τροπικότητα (προφορική για παράδειγμα).

- Η προσθήκη επιπλέον features στο σύστημα, με στόχο την επίτευξη περαιτέρω διαδραστικότητας με τους παίκτες. Ενδεικτικά αναφέρουμε την προσθήκη κάποιου reward system, στην περίπτωση που οι παίκτες πετύχουν καλή επίδοση.

- Για την επίτευξη λειτουργικότητας του συστήματος και σε άλλα λειτουργικά συστήματα, απαραίτητη είναι η ανάπτυξη κάποιου ανοιχτού κώδικα συστήματος εξαγωγής πόζας. Υπάρχουν κάποια διαθέσιμα (όπως το Skeltrack [80]), όμως οι προδιαγραφές τους κρίθηκαν ελλιπείς.

- Τέλος, ως σημαντική μελλοντική κατεύθυνση έρευνας εκλαμβάνεται ο πειραματισμός με επιπλέον τύπους ταξινομητών. Παρά την προσπάθεια που κατέβαλε η παρούσα διπλωματική να μην ασχοληθεί με βαθιά μάθηση, γεγονός είναι ότι η αλματώδης πρόοδος στον τομέα αυτό τα τελευταία χρόνια έχει ανοίξει νέες διόδους για τα πεδία της Όρασης Υπολογιστών και της Μηχανικής Μάθησης. Κατάλληλα για ταξινόμηση χρονοσειρών είναι τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Nets, RNNs), συνεπώς θα μπορούσαν να αποτελέσουν, επί παραδείγματι, έναν κατάλληλο ταξινομητή για το πρόβλημά μας.

7 Συμπληρωματικό Υλικό

7.1 Αναλυτικά Αποτελέσματα Ταξινόμησης των Χειρονομιών

	10 Words	20 Words	40 Words	60 Words	80 Words
2 States	86.18%	89.13%	90.32%	91.73%	92.11%
3 states	88.05%	89.80%	90.80%	92.29%	92.74%
5 states	88.83%	89.35%	90.92%	92.63%	92.48%
7 states	88.91%	89.99%	91.85%	92.67%	93.11%

Table 7.1: Αποτελέσματα της διαδικασίας ταξινόμησης του υποσυνόλου των 5 χρησιμοποιούμενων χειρονομιών, με χρήση διακριτών κρυφών μοντέλων Markov, ως προς τον αριθμό καταστάσεων και το μέγεθος του λεξικού στάσεων σώματος. Αξίζει να σημειωθεί πως η ίδια σειρά πειραμάτων πραγματοποιήθηκε και θέτοντας ως περιορισμό τη διαγωνιότητα του πίνακα συνδιασποράς, με τις διαφορές στα αποτελέσματα να βρίσκονται στα όρια του στατιστικού λάθους.

	Pred. G1	Pred. G3	Pred. G5	Pred. G9	Pred. G11
True G1	446	5	35	8	11
True G3	1	497	6	1	0
True G5	3	2	608	9	2
True G9	2	0	9	489	3
True G11	7	2	46	16	433

Table 7.2: Η μήτρα Σύγχυσης που προέκυψε από εφαρμογή της βέλτιστης περίπτωσης (80 λέξεις, 7 καταστάσεις), και για τα 5 folds. Αξίζει να σημειωθεί ότι α) είχαμε συνολικά 33 instances που δεν ταξινομήθηκαν επιτυχώς πουθενά (no gesture) και β) η μήτρα προέκυψε από επανεκτέλεση του προηγούμενου πειράματος, συνεπώς κάποιες τυχαιοκρατικές παράμετροι (το λεξιλόγιο βασικά) δεν έχουν διατηρηθεί σταθερές.

	1 Gaussian	2 Gaussians	3 Gaussians	5 Gaussians
2 states	91.66%	91.88%	92.26%	93.22%
3 states	90.10%	94.34%	95.23%	96.05%
5 states	93.48%	94.52%	95.61%	97.10%
7 states	92.18%	95.38%	96.20%	97.06%

Table 7.3: Αποτελέσματα της διαδικασίας ταξινόμησης του υποσυνόλου των 5 χρησιμοποιούμενων χειρονομιών, με χρήση συνεχών κρυφών μοντέλων Markov, ως προς τον αριθμό καταστάσεων και τον αριθμό των Γκαουσιανών με τις οποίες μοντελοποιούμε τις παρατηρήσεις σε κάθε κατάσταση. Κατά την εκπαίδευση, τέθηκε η απαίτηση να έχουμε διαγώνιους πίνακες συνδιασποράς.

	Pred. G1	Pred. G3	Pred. G5	Pred. G9	Pred. G11
True G1	480	0	8	7	10
True G3	0	514	1	3	0
True G5	13	2	621	5	2
True G9	3	0	3	497	2
True G11	7	3	11	9	482

Table 7.4: Η μήτρα Σύγχυσης που προέκυψε από εφαρμογή της βέλτιστης περίπτωσης (5 συστατικές Gaussians, 5 καταστάσεις), και για τα 5 folds. Παρότι και σε αυτήν την περίπτωση, ο πίνακας προέκυψε από επανεκτέλεση του προηγούμενου πειράματος, συνεπώς κάποιες τυχαιοκρατικές παράμετροι (η αρχικοποίηση των Gaussian clusters) δεν έχουν διατηρηθεί σταθερές, δεν έχουμε instances τα οποία έχουν μείνει χωρίς ταξινόμηση.

7.2 Αποδείξεις Ιδιοτήτων ενός Πίνακα Συνδιασποράς

Στο σημείο αυτό, παρατίθενται αποδείξεις ορισμένων ιδιοτήτων του πίνακα συνδιασποράς, Σ, οι οποίες δεν παρατέθηκαν στο κυρίως κείμενο για λόγους διατήρησης της ροής του κειμένου:

- Ο πίνακας συνδιασποράς Σ είναι συμμετρικός. Εκ κατασκευής, το στοιχείο $\Sigma(i,j)$ του πίνακα είναι ίσο με $E[(x_i - \mu_i)(x_j - \mu_j)]$ για κάθε ζεύγος των χαρακτηριστικών (x_i, x_j) στο δειγματικό σύνολο από το οποίο προέκυψε ο Σ. Το στοιχείο $\Sigma(j,i)$ του πίνακα, προφανώς, είναι ίσο με $E[(x_j - m_j)(x_i - m_i)] = E[(x_i - m_i)(x_j - m_j)]$, και εφόσον $\forall(i,j) : \Sigma(i,j) = \Sigma(j,i)$ ο πίνακας Σ είναι συμμετρικός.

- Ο πίνακας συνδιασποράς Σ είναι θετικά ημιορισμένος. Εξ ορισμού, ένας πίνακας A, nxn είναι θετικά ημιορισμένος όταν, $\forall x \in R^n : x^T Ax \geq 0$. Ο πίνακας Σ

είναι ίσος με $E[(x-\mu)(x-\mu)^T], x, \mu \in R^n$. Τότε θα έχουμε: $u^T\Sigma u = u^T E[(x-\mu)(x-\mu)^T]u = E[u^T(x-\mu)(x-\mu)^T u]$ και, θέτοντας τη βαθμωτή ποσότητα $u^T(x-\mu) = (x-\mu)^T u = a$ παίρνουμε $u^T\Sigma u = E[\alpha^2] \geq 0$, και έχουμε το ζητούμενο.

- Οι ιδιοτιμές του Σ είναι μη αρνητικές. Πράγματι, για κάθε ιδιοτιμή λ_i του Σ θα ισχύει $A\xi_i = \lambda_i\xi_i$, όπου ξ_i το αντίστοιχο ιδιοδιάνυσμα. Οπότε: $\Sigma\xi_i = \lambda_i\xi_i \iff \xi_i^T\Sigma\xi_i = \xi_i^T\lambda_i\xi_i \iff \xi_i^T\lambda_i\xi_i \geq 0$ (αφού Σ θετικά ημιορισμένος) και εφόσον ως εσωτερικό γινόμενο διανύσματος με τον εαυτό του $\xi_i^T\xi_i$ μη αρνητικό, η ιδιοτιμή λ_i είναι, επίσης, μη αρνητική.

- Τα ιδιοδιανύσματα του Σ είναι βάσεις ενός ορθογωνίου χώρου (ισοδύναμα, είναι ορθογώνια μεταξύ τους). Αρχικά, θα δείξουμε το ενδιάμεσο αποτέλεσμα ότι, για κάθε συμμετρικό πίνακα $A(A = A^T)$ και κάθε ζεύγος διανυσμάτων x, y, το $(Ax)^T y = x^T(Ay)$. Ξεκινώντας από το πρώτο μέλος, $(Ax)^T y = x^T A^T y = x^T Ay = x^T(Ay)$.
  Έχοντας αυτό υπόψιν, παίρνοντας συμμετρικό πίνακα Σ, δύο τυχαία ιδιοδιανύσματά του u_i, u_j στα οποία αντιστοιχούν ιδιοτιμές λ_i, λ_j και γνωρίζοντας ότι δύο διανύσματα u, v είναι ορθογώνια όταν το εσωτερικό τους γινόμενο μηδενίζεται: $(\Sigma u_i)^T u_j = u_i^T(\Sigma u_j) \iff (\lambda_i u_i)^T u_j = u_i^T(\lambda_j u_j) \iff \lambda_i(u_i^T)u_j = \lambda_j(u_i^T)u_j$ και, εφόσον δεν ισχύει καθολικά η απαίτηση $\lambda_i = \lambda_j$ συμπεραίνουμε ότι υποχρεωτικά $u_i^T)u_j = 0$, πληρείται δηλαδή η συνθήκη ορθογωνιότητας.

7.3 Απόδειξη ότι ο Μετασχηματισμός PCA δεν Επιδρά στην Ευκλίδεια Απόσταση

Κατά την εφαρμογή της PCA σε έναν πίνακα συνόλου δεδομένων, προκύπτει ο τελικός πίνακας $T = XP$, με P τον πίνακα του μετασχηματισμού. Οι στήλες του P αποτελούν τα κανονικοποιημένα διανύσματα του Σ (ώστε να έχουν μέτρο 1), τα οποία όπως είχαμε δει είναι ορθογώνια, συνεπώς αποτελούν μία ορθοκανονική βάση. Από εκεί, προκύπτει ότι $PP^T = 1$, συνεπώς ο πίνακας P ανήκει στην ομάδα των ορθογωνίων πινάκων (orthogonal group). Αυτοί αποτελούν υποσύνολο της Ευκλίδειας ομάδας (Euclidean group), που είναι το σύνολο των πινάκων που δρουν ως ισομετρίες στον ευκλίδειο χώρο, δεν επηρεάζουν δηλαδή την Ευκλίδεια μετρική απόστασης.

7.4 Απόδειξη ότι τα Προβλήματα Βελτιστοποίησης Τετραγωνικού Προγραμματισμού Επιδέχονται Κλειστής Λύσης

Παραδοσιακά, ο απλούστερος τρόπος επίλυσης προβλημάτων ελαχίστου συνίσταται στο μηδενισμό της παραγώγου της αντικειμενικής συνάρτησης. Οπότε, δοθείσας της

τετραγωνικής διανυσματικής συνάρτησης $P(x) = x^T Ax + b^T x + c$, έχουμε:

$$\nabla P(x) = 0 => 2Ax + b = 0 => x = -\frac{A^{-1}b}{2}$$

7.5 Το Πρόβλημα του Περιπλανώμενου Πλανόδιου

Ένα από τα γνωστότερα προβλήματα βελτιστοποίησης τα οποία επιλύονται κυρίως με χρήση ευριστικών αλγορίθμων είναι αυτό του περιπλανώμενου πλανόδιου (Travelling Salesperson Problem, TSP). Το πρόβλημα αφορά στην εύρεση του συντομότερου μονοπατιού που συνδέει ένα σύνολο πόλεων, τέτοιο ώστε να είναι κυκλικό και να επισκέπτεται κάθε πόλη μόνο μία φορά.

Σε επίπεδο μαθηματικού φορμαλισμού, μπορούμε να διατυπώσουμε το πρόβλημα του περιπλανώμενου πλανόδιου ως εξής: Έστω ένα σύνολο πόλεων, πλήθους N. Επιπλέον, ορίζουμε τα ακόλουθα μεγέθη:

$$x_{ij} = \begin{cases} 1, if\ i,j\ consecutive \\ 0, else \end{cases} \tag{7.1}$$

$$c_{ij} = d(i,j) \tag{7.2}$$

$$u_i : final\ ordering\ of\ cities \tag{7.3}$$

Τότε, πρακτικά έχουμε το ακόλουθο πρόβλημα ελαχιστοποίησης:

$$min \sum_{i=1}^{N} \sum_{j=1 \neq i}^{N} c_{ij} x_{ij},\ subject\ to \tag{7.4}$$

$$\sum_{i=1}^{N} x_{ij} = 1 \tag{7.5}$$

$$\sum_{j=1}^{N} x_{ij} = 1 \tag{7.6}$$

$$(u_i - u_j) x_{ij} > 0, \forall i > 1 \tag{7.7}$$

Όπου η πρώτη εξίσωση αποτελεί τη διατύπωση του προβλήματος ελαχιστοποίησης, οι δύο επόμενες θέτουν τις απαιτήσεις από κάθε πόλη να έχουμε μοναδική αναχώρηση και, σε κάθε πόλη, μοναδική άφιξη, και η τελευταία το τελικό μονοπάτι να είναι ένας κυκλικός γράφος (θεωρούμε ότι ξεκινάμε από την πόλη με $i = 1$).

Από τη μορφή του προβλήματος, είναι εμφανές ότι, για την έυρεση ακριβούς λύσης, καθώς για N πόλεις έχουμε N! δυνατούς συνδυασμούς, η πολυπλοκότητα αυξάνεται αναλόγως με το παραγοντικό του N (με χρήση brute force search), ή εκθετικά,

αν μοντελοποιήσουμε το πρόβλημα ως πρόβλημα γραμμικού προγραμματισμού. Συνεπώς, το πρόβλημα είναι κατάλληλο πεδίο για τη χρήση διάφορων ευριστικών και μετα-ευριστικών αλγορίθμων για την επίλυσή του.

References

[1] "Example tracking image from microsoft kinect," https://Computervisionblog.files.wordpress.com/2013/11/test2.png.

[2] "The skeleton representation used by microsoft kinect," https://www.mdpi.com/1424-8220/12/9/12126/htm.

[3] "An image and block diagram of a theremin," https://www.wikipedia.org/Theremin.

[4] "Simple clustering example," https://home.deib.poli.it/matteucc/Clustering/tutorial.html.

[5] "Comparison between gaussian mixture clustering and k-means clustering," https://shapeofdata.wordpress.com/2013/07/30/k-means.

[6] "Demonstration of-nearest neighbor algorithm," https://en.wikipedia.org/wiki/K_nearest_neighbors_algorithm.

[7] "Decision areas in k-nearest neighbor," https://onlinecourses.science.psu.edu/stat857/node/21.

[8] "Example of building a kd-tree," https://ida.liu.se/opendsa/OpenDSA/Books/Everything/html/KDtree.html.

[9] "Pca example in 2 dimensions," https://en.wikipedia.org/wiki/Principal_component_analysis.

[10] "Comparison between euclidean-based and dynamic time warping-based time-series alignment," https://sfl.scientificdata.com/data-science-blog/2016/3/dynamic-time-warping-time-series-analysis-iii.

[11] "Demonstration of a sakoe-chiba zone," https://indico.io/blog/plotlines.

[12] "A tri-state markov chain and its transition matrix," https://www.analyticsvidhya.com/blog/2014/07/markov-chain-simplified.

[13] "The state/observation dependence in a hidden markov model," https://en.wikipedia.org/wiki/Hidden_Markov_Model.

[14] "A forward-only hidden markov model," https://gekkoquant.com/2014/05/18/Hidden-markov-models-model-description-part-1-of-4.

[15] D. Song, Y. Shi, P. Zhang, Q. Huang, U. Kruschwitz, Y. Hou, and B. Wang, "Incorporating intra-query term dependencies in an aspect query language model," *Computational Intelligence*, vol. 31, no. 4, pp. 699–720, 2015.

[16] "The coordinate system used by microsoft kinect," http://gmv.cast.uark.edu/wp-content/uploads/2012/07/convert-kinect-to-standarized1.png.

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. 2001 IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR)*, Kauai, HI, USA, 2001, pp. 511–518.

[18] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces (ICMI)*, State College, PA, 2004, pp. 205–211.

[19] T. Starner and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," in *Motion-Based Recognition*. Springer, 1997, pp. 227–243.

[20] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009, pp. 1014–1021.

[21] N. F. Troje, "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns," *Journal of Vision*, vol. 2, no. 5, pp. 371–387, 2002.

[22] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani, "Probabilistic posture classification for human-behavior analysis," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 35, no. 1, pp. 42–54, 2005.

[23] J. J. Kuch and T. S. Huang, "Vision based hand modeling and tracking for virtual teleconferencing and telecollaboration," in *Proc. 1995 Int. Conf. Computer Vision (ICCV)*, Cambridge, MA, USA, 1995, pp. 666–671.

[24] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[25] J. Martin and J. L. Crowley, "An appearance-based approach to gesture-recognition," in *Proc. 9th Int. Conf. Image Analysis and Processing (ICIAP)*, Florence, Italy, 1997, pp. 340–347.

[26] "Microsoft kinect homepage," https://developer.microsoft.com/en-us/windows/kinect.

[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. 2005 IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, 2005, pp. 886–893.

[28] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 1999 IEEE Int. Conf. Computer Vision (ICCV)*, Corfu, Greece, 1999, pp. 1150–1157.

[29] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[30] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[31] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. 2008 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, USA, 2008, pp. 1–8.

[32] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. 2011 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 1385–1392.

[33] S. Ivekovic, E. Trucco, and Y. R. Petillot, "Human body pose estimation with particle swarm optimisation," *Evolutionary Computation*, vol. 16, no. 4, pp. 509–528, 2008.

[34] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, 2014, pp. 1653–1660.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[36] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, QU, Canada, 2014, pp. 1799–1807.

[37] T. Pfister, K. Simonyan, J. Charles, and A. Zisserman, "Deep convolutional neural networks for efficient pose estimation in gesture videos," in *Proc. 12th Asian Conf. Computer Vision (ACCV)*, Singapore, 2014, pp. 538–552.

[38] I. Kostrikov and J. Gall, "Depth sweep regression forests for estimating 3d human pose from images." in *Proc. British Machine Vision Conf. 2014 (BMVC)*, Nottingham, UK, 2014.

[39] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human detection using depth information by kinect," in *Proc. 2011 IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition Workshops (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 15–22.

[40] M. Bray, E. Koller-Meier, P. Muller, L. Van Gool, and N. N. Schraudolph, "3d hand tracking by rapid stochastic gradient descent using a skinning model," in *Proc. 1st European Conf. Visual Media Production (CVMP)*, London, UK, 2004.

[41] D. Demirdjian, T. Ko, and T. Darrell, "Constraining human body tracking," in *Proc. 2003 Int. Conf. Computer Vision (ICCV)*, Nice, France, 2003, pp. 1071–1078.

[42] D. Grest, J. Woetzel, and R. Koch, "Nonlinear body pose estimation from depth images," in *Deutsche Arbeitsgemeinschaft fur Mustererkennung-Symposium (DAGM)*, Vienna, Austria, 2005, pp. 285–292.

[43] A. Baak, M. Muller, G. Bharaj, H.-P. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *Consumer Depth Cameras for Computer Vision.* Springer, 2013, pp. 71–98.

[44] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[45] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys, "Accurate 3d pose estimation from a single depth image," in *Proc. 2011 IEEE Int. Conf. on Computer Vision (ICCV)*, Barcelona, Spain, 2011, pp. 731–738.

[46] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.

[47] C. Xu and L. Cheng, "Efficient hand pose estimation from a single depth image," in *Proc. 2013 IEEE Int. Conf. Computer Vision (ICCV)*, Sydney, Australia, 2013, pp. 3456–3462.

[48] D. Tang, T.-H. Yu, and T.-K. Kim, "Real-time articulated hand pose estimation using semi-supervised transductive regression forests," in *Proc. 2013 IEEE Int. Conf. Computer Vision (ICCV)*, Sydney, Australia, 2013, pp. 3224–3231.

[49] T. Winkler, "Making motion musical: Gesture mapping strategies for interactive computer music." in *Proc. 1995 Int. Computer Music Conf. (ICMC)*, Banff, AL, Canada, 1995, pp. 261–264.

[50] "Motus," http://tzm.lt/.

[51] "Phonotonic," https://www.phonotonic.net/.

[52] "Point motion," https://www.pointmotioncontrol.com.

[53] T. Berg, D. Chattopadhyay, M. Schedel, and T. Vallier, "Interactive music: Human motion initiated music generation using skeletal tracking by kinect," in *Proc. 2012 SEAMUS Conf.*, Appleton, WI, USA, 2012.

[54] M.-J. Yoo, J.-W. Beak, and I.-K. Lee, "Creating musical expression using kinect." in *Proc. Int. Conf. New Interfaces for Musical Expression*, Oslo, Norway, 2011, pp. 324–325.

[55] E. Frid, "Musik ur tomma luften: en musikleksak baserad pa kinect," *Diploma thesis, KTH Royal Institute of Technology, Stockholm*, 2012.

[56] A. Rosa-Pujazon, I. Barbancho, L. J. Tardon, and A. M. Barbancho, "A virtual reality drumkit simulator system with a kinect device," *Int. Journal of Creative Interfaces and Computer Graphics (IJCICG)*, vol. 6, no. 1, pp. 72–86, 2015.

[57] M.-H. Hsu, W. Kumara, T. K. Shih, and Z. Cheng, "Spider king: Virtual musical instruments based on microsoft kinect," in *Proc. 2013 Int. Joint Conf. Awareness Science and Technology and Ubi-Media Computing (iCAST-UMEDIA)*, Aizu-Wakamatsu, Japan, 2013, pp. 707–713.

[58] T. Grill, "Two-dimensional gesture mapping for interactive real-time electronic music instruments," *Master thesis, University of Music And Performing Arts, Vienna*, 2008.

[59] A. N. Antle, M. Droumeva, and G. Corness, "Playing with the sound maker: do embodied metaphors help children learn?" in *Proc. 7th Int. Conf. Interaction Design and Children*, Chicago, IL, USA, 2008, pp. 178–185.

[60] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," in *Proc. 2011 ACM SIGGRAPH/Eurographics Symp. Computer Animation (SCA)*, Vancouver, BC, Canada, 2011, pp. 147–156.

[61] M. Reyes, G. Dominguez, and S. Escalera, "Featureweighting in dynamic timewarping for gesture recognition in depth data," in *Proc. 2011 IEEE Int. Conf. Computer Vision Workshops (ICCV)*, Barcelona, Spain, 2011, pp. 1182–1188.

[62] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici, "Gesture recognition using skeleton data with weighted dynamic time warping," in *Proc. 8th Int. Conf. Computer Vision Theory And Applications (VISAPP)*, Barcelona, Spain, 2013, pp. 620–625.

[63] X. Zhao, X. Li, C. Pang, X. Zhu, and Q. Z. Sheng, "Online human gesture recognition from motion data streams," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 23–32.

[64] G. A. ten Holt, M. J. Reinders, and E. Hendriks, "Multi-dimensional dynamic time warping for gesture recognition," in *Proc. 13th Annu. Conf. Advanced School for Computing and Imaging*, Heijen, Netherlands, 2007.

[65] G. T. Papadopoulos, A. Axenopoulos, and P. Daras, "Real-time skeleton-tracking-based human action recognition using kinect data." in *Proc. 20th Int. Conf. Multimedia Modeling (MMM)*, Dublin, Ireland, 2014, pp. 473–483.

[66] F. Negin, F. Ozdemir, C. B. Akgul, K. A. Yuksel, and A. Ercil, "A decision forest based feature selection framework for action recognition from rgb-depth cameras," in *Proc. 10th Int. Conf. Image Analysis and Recognition (ICIAR)*, Varzim, Portugal, 2013, pp. 648–657.

[67] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations." in *Proc. 23rd Int. Joint Conf. Artificial Intelligence (IJCAI)*, Beijing, China, 2013, pp. 2466–2472.

[68] S. Haykin, *Neural Networks: A Comprehensive Foundation (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2007.

[69] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009.

[70] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[71] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*. Springer Publishing Company, Incorporated, 2015.

[72] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proc. 30th Annu. ACM Symp. Theory of Computing*, Dallas, TX, USA, 1998, pp. 604–613.

[73] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proc. 2001 SIAM Int. Conf. Data Mining*, Chicago, IL, USA, 2001, pp. 1–11.

[74] L. Rodriguez and I. Torres, "Comparative study of the baum-welch and viterbi training algorithms applied to read and spontaneous speech recognition," in *Proc. 1st Iberian Conf. Pattern Recognition and Image Analysis (IbPRIA)*, Mallorca, Spain, 2003, pp. 847–857.

[75] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[76] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[77] "The portaudio software library for c," https://www.portaudio.com.

[78] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, Austin, TX, 2012, pp. 1737–1746.

[79] "The netlab software package for matlab," http://www.ncrg.aston.ac.uk/netlab.

[80] "The skeltrack software package for c," https://github.com/joaquimrocha/Skeltrack.