Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Ευφυής Καταγραφή 3Δ Ανθρώπινης Κίνησης μέσω Δεικτών με χρήση Βαθιάς Μάθησης και Χαρτών Βάθους Πολλαπλών Προβολών

Διδακτορικη Διατριβη

του

ΑΝΑΡΓΥΡΟΥ Ε. ΧΑΤΖΗΤΟΦΗ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Αθήνα, Αύγουστος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Ευφυής Καταγραφή 3Δ Ανθρώπινης Κίνησης μέσω Δεικτών με χρήση Βαθιάς Μάθησης και Χαρτών Βάθους Πολλαπλών Προβολών

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΑΝΑΡΓΥΡΟΥ Ε. ΧΑΤΖΗΤΟΦΗ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 26η Αυγούστου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.	Πέτρος Δάρας Ερευνητής Α΄ Ε.Κ.Ε.Τ.Α.	Κωνσταντίνος Καρπούζης Επικ. Καθηγητής Παντείου Παν.
(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*

.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γιώργος Στάμου Καθηγητής Ε.Μ.Π. *(Υπογραφή)*	Αθανάσιος Βουλοδήμος Επίκ. Καθηγητής Ε.Μ.Π.

.....
Γεώργιος Καρυδάκης
Αναπλ. Καθηγητής Παν. Αιγαίου

Αθήνα, Αύγουστος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Διδακτορικής Διατριβής, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Διδακτορική μου Διατριβή και κατά συνέπεια αποτυχία απόκτησης του Διδακτορικού Τίτλου, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Διδακτορική Διατριβή προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Ανάργυρος Χατζητοφής
Διδάκτωρ ΕΜΠ

23η Αυγούστου 2022

Περίληψη

Η παρούσα διδακτορική διατριβή διερευνά την υψηλή ακρίβεια που επιτυγχάνουν τα επαγγελματικά συστήματα καταφραφής ανθρώπινης κίνησης με τη χρήση δεικτών, την αξιοσημείωτη ικανότητα των μοντέλων όρασης υπολογιστών βαθιάς μάθησης να επιλύουν προβλήματα που σχετίζονται με την εκτίμηση της ανθρώπινης πόζας, καθώς και τις πρόσφατες τεχνολογικές εξελίξεις στους αισθητήρες ανίχνευσης βάθους καταναλωτικής ποιότητας και χαμηλού κόστους για την ανάπτυξη καινοτόμων 'υβριδικών' μεθόδων καταγραφής κίνησης.

Κατά τη μελέτη και έρευνα στο συγκεκριμένο πεδίο για την επίτευξη του ερευνητικού στόχου της διατριβής πραγματοποιήθηκαν μια σειρά από ερευνητικές συνεισφορές όπως, μεταξύ άλλων, η μοντελοποίηση του προβλήματος καταγραφής κίνησης με δείκτες από δεδομένα βάθους πολλαπλών χρονικά και χωρικά συσχετισμένων προβολών, η εισαγωγή κατάλληλων αναπαραστάσεων ανεξάρτητων από τον τρόπο καταγραφής των αρχικών δεδομένων ή τη θέση και λειτουργία των αισθητήρων, καθιστώντας τα μοντέλα ικανά να αντεπεξέλθουν αποτελεσματικά ακόμα και με σχετικά περιορισμένα δεδομένα εκπαίδευσης, η ανάπτυξη νέων τεχνικών αποκωδικοποίησης 3Δ συντεταγμένων από 2Δ και 2.5Δ δεδομένα εισόδου με υψηλή ακρίβεια για την αξιοποίηση των μοντέλων σε πραγματικού χρόνου τεχνολογίες καταγραφής κίνησης. Επιπρόσθετα, για την πραγματοποίηση των παραπάνω κρίθηκε απαραίτητη η δημιουργία δύο (2) μεγάλων συνόλων δεδομένων, των DMC και HUMAN4D. Η δημιουργία τους ήταν απαραίτητη λόγω της περιορισμένης δημόσιας διάθεσης δεδομένων βάθους πολλαπλών προβολών που καταγράφουν κινήσεις υποκειμένων με τοποθετημένους δείκτες στο σώμα τους.

Αρχικά, με τη δημιουργία του συνόλου δεδομένων DMC επιβεβαιώσαμε τη βασική ιδέα καταγραφής κίνησης με χρήση βαθιάς μάθησης και πολλαπλών χαρτών βάθους με οπισθοανακλαστικά υλικά χωρίς υψηλές απαιτήσεις σε ακρίβεια συγχρονισμού και ground-truth δεδομένων πόζας. Κατόπιν, δημιουργήσαμε το σύνολο δεδομένων HUMAN4D το οποίο παρέχει δεδομένα βάθους από πολλαπλές κάμερες ταυτόχρονης λήψης με συγχρονισμό υλισμικού και εξαιρετικά ακριβείς 3Δ πόζες για εποπτεία και αξιολόγηση, το οποίο κατά σειρά, μας επέτρεψε την πραγματοποίηση περαιτέρω έρευνας γύρω από σύγχρονες μεθόδους εκτίμησης πόζας/κίνησης και τεχνικών αναπαράστασης 3Δ συντεταγμένων. Τέλος, ως ακολουθία των παραπάνω πραγματοποιήθηκε η ανάπτυξη μιας νέας, state-of-the-art μεθόδου εκτίμησης 3Δ πόζας/κίνησης από δεδομένα βάθους με τη χρήση σφαιρικών οπισθοανακλαστικων δεικτών.

Λέξεις Κλειδιά

καταγραφή κίνησης, βαθιά μάθηση, ταυτοποίηση δεικτών, χάρτες βάθους, πολλαπλές προβολές, 3Δ αποκωδικοποίηση συντεταγμένων, 3Δ απόδοση πολλαπλών προβολών, 3Δ επίβλεψη πολλαπλών προβολών

Abstract

The present PhD thesis explores the high accuracy achieved by professional marker-based motion capture systems, the remarkable ability of deep learning models to solve human pose estimation and tracking problems, and recent technological developments in consumer-grade and low-cost depth sensing sensors for the development of novel "hybrid" methods for motion capture.

To achieve the research objectives of this thesis, a number of research contributions took place, such as the modelling of marker-based motion capture from spatio-temporally aligned multi-view depth maps, the introduction of appropriate scale and translation invariant representations independent of the way the original data is captured or the pose and operation of the sensors, making the models able to cope effectively even with relatively limited training data, the development of new techniques for decoding 3D coordinates from 2D and 2.5D input data with high accuracy for real-time operation. In addition, to achieve the above, it was necessary to create 2 large datasets, DMC and HUMAN4D. Their creation was necessary due to the limited availability of public multi-view depth data with body-attached markers on the subjects.

Initially, by creating the DMC dataset, we validated the basic concept of motion capture using the deep learning paradigm and multiple depth maps with retro-reflective materials without high requirements on synchronization accuracy and ground-truth poses. Then, we created the HUMAN4D dataset which provides depth data from multiple cameras with hardware synchronization and highly accurate 3D pose data for supervision and evaluation, which in turn allowed us to conduct further research around state-of-the-art pose/motion estimation methods and 3D coordinate representation techniques. Finally, as a result of the above, we carried out the conceptualization and development of a novel, state-of-the-art method for estimating 3D pose/motion from depth data using spherical retro-reflective markers.

Keywords

motion capture, deep learning, marker labeling, depth maps, multi-projection, 3D coordinate decoding, 3D multi-view rendering, multi-view supervision

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Πρόλογος

Με την παρούσα διατριβή φτάνω στον τελικό προορισμό της ακαδημαϊκής μου πορείας, εκτιμώντας φυσικά τόσο τον προορισμό όσο και το μακρύ αυτό ταξίδι. Είμαι σε θέση να πω με βεβαιότητα πως ο δρόμος δεν ήταν πάντοτε εύκολος, αλλά πάντοτε ήταν επιμορφωτικός. Σε αυτό το ταξίδι λοιπόν θα ήθελα να απονείμω την αναγνώριση και τις ευχαριστίες μου σε συγκεκριμένους ανθρώπους, ο ρόλος των οποίων ήταν καθοριστικός για την ολοκλήρωση αυτής της διατριβής.

Θα ήθελα καταρχήν να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Στέφανο Κόλλια ο οποίος μου προσέφερε τη δυνατότητα να πραγματοποιήσω την παρούσα διδακτορική διατριβή υπό την επίβλεψη και καθοδήγησή του στο Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ), στη σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών (ΗΜΜΥ), σε συνεργασία με το Εθνικό Κέντρο Έρευνας και Τεχνολογικής Ανάπτυξης (ΕΚΕΤΑ), στο Ινστιτούτο Πληροφορικής και Τηλεματικής (ΙΠΤΗΛ).

Μεγίστης σημασίας θεωρώ τη συνεισφορά του αναπληρωτή καθηγητή κ. Κωνσταντίνου Καρπουζη και μέλους της τριμελούς επιτροπής που αποτέλεσε τον συνδετικό κρίκο και ανιδιοτελώς συνέβαλε τα μέγιστα για αυτή την σύμπραξη. Θα ήθελα να τονίσω επίσης το σημαντικό ρόλο που θεωρώ ότι διαδραματίζουν άνθρωποι σαν αυτόν στη στελέχωση ακαδημαϊκών ιδρυμάτων και στην ενθάρρυνση νέων σπουδαστών να βρουν κίνητρα και να μετουσιώσουν τις δυνατότητες τους, ορμώμενος από το γεγονός ότι είχε επικοινωνήσει μαζί μου στη φάση των προπτυχιακών μου σπουδών όταν ενημερώθηκε για ένα έξω-πανεπιστημιακό μου εγχείρημα και προσφέρθηκε να με στηρίξει με κάθε δυνατό τρόπο όπως και έκανε.

Επίσης, θερμά θα ήθελα να ευχαριστήσω τον ερευνητή Α' Βαθμίδας κ. Πέτρο Δάρα και μέλος της τριμελούς επιτροπής που από την πρώτη στιγμή της συνεργασίας μας το 2012 στο ΕΚΕΤΑ-ΙΠΤΗΛ, είχαμε ιδανική συνεργασία και με αντιμετώπισε με τον καλύτερο δυνατό τρόπο σε επαγγελματικό και προσωπικό επίπεδο, καθώς και για το σπουδαιότατο ρόλο και την καθοδήγηση στην πορεία αυτής της διατριβής. Επίσης, εισάκουσε την πρόθεση μου για την πραγματοποίηση διδακτορικών σπουδών παράλληλα με το ερευνητικό έργο που πραγματοποιούσε το εργαστήριό μας, κάνοντας οποιαδήποτε δυνατή ενέργεια για την επίτευξη συμφωνίας για τις σπουδές αυτές με ανώτατους ακαδημαϊκούς φορείς.

Φυσικά δε θα μπορούσα να μην αναγνωρίσω και ευχαριστήσω τους συνεργάτες και μέλη του εργαστηρίου Visual Computing Lab - VCL του ΕΚΕΤΑ-ΙΠΤΗΛ που κατά τα χρόνια που πραγματοποίησα το διδακτορικό με ενέπνευσαν με ιδέες αλλά και με βοήθησαν με πράξεις για να πάρει η έρευνα την τελική της μορφή.

Ευχαριστώ επίσης τους καθηγητές κ. Σταφυλοπάτη, κ. Στάμου, κ. Βουλοδήμο και κ. Καρυδάκη, οι οποίοι δέχτηκαν να είναι μέλη της επταμελούς επιτροπής.

Δεδομένων των δυσκολιών και της σκληρής δουλειάς που απαιτεί η εκπόνηση μιας διδα-

κτορικής διατριβής, και ειδικότερα λόγω της παράλληλης εργασίας μου στο ερευνητικό κέντρο, δε θα μπορούσα να μην ευχαριστήσω τη σύζυγο μου Δήμητρα Ροπούλου για τη στήριξη, τα άγχη, τις αγωνίες και τις πιέσεις που περάσαμε μαζί για την αποπεράτωση του ερευνητικού αυτού έργου. Τέλος, θα ήθελα να ευχαριστήσω από καρδιάς τους γονείς μου που έκαναν τα πάντα για να με στηρίξουν σε οτιδήποτε, με καθοδήγησαν και μου έδωσαν τη δυνατότητα να πραγματοποιήσω τις προπτυχιακές σπουδές μου που ήταν αρχή όλων, μου προσέφεραν τη δυνατότητα να έχω τις επιλογές που επιθυμούσα και με τον τρόπο τους με έκαναν να διαμορφώσω την προσωπικότητα και τον τρόπο σκέψης μου που θεωρώ πως με έχει βοηθήσει σημαντικά μέχρι και σήμερα.

Κεφάλαιο 1

Εισαγωγή

Κατοικούμε σε ένα κόσμο δυναμικό, γεμάτο από κινούμενα όντα, υλικά και αντικείμενα διαφόρων σχημάτων, χρωμάτων και υφών, δημιουργώντας στον άνθρωπο την ανάγκη να καταγράφει, αναλύει και μελετά εκτενώς τα χωροχρονικά δεδομένα γύρω του. Τα τελευταία χρόνια, την ανάγκη αυτή ενθαρρύνει η μεγάλη ανάπτυξη της τεχνολογίας και ειδικότερα η υψηλή επεξεργαστική ισχύς και παραγωγή αισθητήρων χαμηλού κόστους που επιτρέπουν την καταγραφή / ανίχνευση / μέτρηση 3Δ χωρικών δεδομένων στο χρόνο, όπως για παράδειγμα οι αισθητήρες ανίχνευσης βάθους. Τη σύγχρονη ψηφιακή εποχή που διανύουμε, ο άνθρωπος βρίσκεται και θα βρίσκεται στο επίκεντρο της τρέχουσας τεχνολογίας αλλά και των επερχόμενων, μελλοντικών τεχνολογιών, σε μια προσπάθεια ψηφιοποίησης όλων των ανθρωπίνων πτυχών, από τον τρόπο κοινωνικοποίησής του μέχρι την ψηφιοποίηση της κίνησης του και της ίδιας του της παρουσίας σε ψηφιακά ή μεικτά περιβάλλοντα [13, 14, 15, 16].

Σήμερα, οι ψηφιακά σχεδιασμένοι ή σκαναρισμένοι 3Δ χαρακτήρες κινούνται με τη χρήση τεχνολογιών καταγραφής κίνησης και χρησιμοποιούνται σε διάφορους τεχνολογικούς και βιομηχανικούς τομείς. Οι τεχνολογίες αυτές δίνουν τη δυνατότητα παραγωγής πολυμέσων για χρήση σε εμπειρίες που παρέχουν απομακρυσμένη εικονική παρουσία και συν-παρουσία (π.χ. 3Δ τηλε-διασκέψεις [17], μουσεία μεικτής πραγματικότητας [18], κ.λπ.). Στην παρούσα διδακτορική διατριβή, πραγματοποιείται έρευνα στα πεδία της Υπολογιστικής Όρασης και Βαθιάς Μάθησης για την ανάπτυξη μεθόδων και μοντέλων καταγραφής και ψηφιοποίησης 3Δ κίνησης με χρήση πολλαπλών χαρτών βάθους από χαμηλού κόστους, συμβατικούς αισθητήρες και οπισθοανακλαστικούς δείκτες.

Κίνητρο αποτελεί ο περιορισμός χρήσης τέτοιων τεχνολογιών λόγω του απαγορευτικού κόστους και τις πολυπλοκότητας εγκατάστασης των υπαρχόντων συστημάτων, ενώ αντίθετα, στόχος είναι η εύκολη και ανοικτή πρόσβαση στις τεχνολογίες αυτές για τις γενιές του αύριο που θα αλληλεπιδρούν καθημερινά με μέσα ανθρώπινης ψηφιοποίησης. Ενδεικτικά σε επίπεδο υλισμικού, το κόστος μεθόδων σαν τις προτεινόμενες μπορεί να αγγίζει το υπο-εικοσαπλάσιο σε σχέση με τις υπάρχουσες λύσεις του εμπορίου.

1.1 Υπολογιστική Όραση και Βαθιά Μάθηση

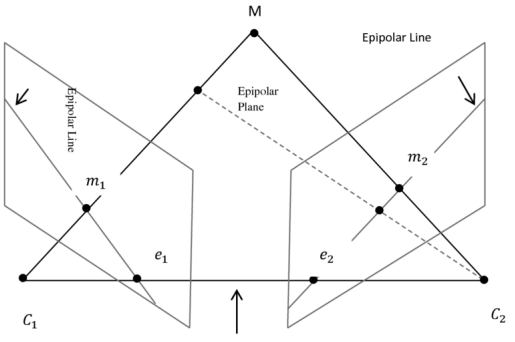
Η εξέλιξη τόσο σε επίπεδο λογισμικού όσο και υλισμικού είναι ραγδαία, προσφέροντας τη δυνατότητα επεξεργασίας και ανάλυσης πάσης φύσεως δεδομένων με ταχύτητα και ακρίβεια. Εξέχουσα θέση καταλαμβάνει η επεξεργασία σημάτων συναφών με εκείνα τα οποία λαμβάνει

και επεξεργάζεται ο ανθρώπινος εγκέφαλος μέσω των αισθήσεων (όσφρηση, γεύση, αφή, ακοή, όραση). Συγκεκριμένα, με τον όρο **Υπολογιστική Όραση** αναφερόμαστε στον επιστημονικό κλάδο που αφορά στη μελέτη μεθόδων και αλγορίθμων που επιτρέπουν σε μηχανές ή απλές υπολογιστικές μονάδες την αντίληψη και ερμηνεία 'οπτικών' σημάτων/δεδομένων.

Παράλληλα, ταχύτατα εξελισσόμενη πρόοδο παρουσιάζει η ανάπτυξη και εφαρμογή μοντέλων τεχνητής νοημοσύνης. Η τεχνητή νοημοσύνη αποτελεί την επιστήμη που έχει ως επίκεντρο τη μελέτη και ανάπτυξη υπολογιστικών μονάδων ή συστημάτων που προσομοιώνουν την επεξεργασία και αντίληψη δεδομένων με τροπισμό όμοιο με του ανθρωπίνου εγκεφάλου. Σημαντικό κλάδο της τεχνητής νοημοσύνης αποτελεί η Μηχανική Μάθηση που επιτρέπει στις 'μηχανές' να μαθαίνουν από τα δεδομένα που έχουν στη διάθεσή τους προς ανάλυση, χωρίς να χρειάζεται η επίβλεψη και ο αυστηρός έλεγχος από τον άνθρωπο. **Βαθιά Μάθηση** είναι ένα υποσύνολο της μηχανικής μάθησης όπου νευρωνικά δίκτυα με τρία ή περισσότερα επίπεδα εκπαιδεύονται να προσομοιώνουν τη συμπεριφορά του ανθρώπινου εγκεφάλου μέσω της παρατήρησης και επίβλεψης τους σε, συνήθως αλλά όχι πάντα, μεγάλα σύνολα δεδομένων. Η όραση υπολογιστών και η βαθιά μάθηση είναι άρρηκτα συνδεδεμένες προσφέροντας πληθώρα δυνατοτήτων, όπως αντίστοιχα συνδέονται στον ανθρώπινο εγκέφαλο οι ικανότητες όρασης-αντίληψης-απόκρισης.

Την τελευταία δεκαετία, οι κλάδοι αυτοί έχουν παρουσιάσει ιδιαίτερη άνθηση με αποτέλεσμα να διευρύνεται η έρευνα και η εφαρμογή τους σε συνεχώς περισσότερους τομείς για την επίλυση όλο και πιο σύνθετων προβλημάτων. Στόχος, μεταξύ άλλων, είναι η χωροχρονική κατανόηση και ανάλυση σκηνών και δράσεων μέσω του εντοπισμού και αναγνώρισης ανθρώπινων κινήσεων και συμπεριφορών, καθώς και αντικειμένων και των χρήσεων τους από τον άνθρωπο. Η δυνατότητα καταγραφής 3Δ δεδομένων μέσω αισθητήρων-καμερών είναι ένα πολύτιμο εργαλείο το οποίο επιτρέπει τη χωροχρονική αντίληψη διαδραστικών ή στατικών σκηνών μέσω της **Υπολογιστικής Όρασης Βάθους**. Η υπολογιστική όραση βάθους είναι ένας πολύ ενεργός τομέας της υπολογιστικής όρασης με καινοτομίες που κυμαίνονται από εφαρμογές όπως η ανακατασκευή 3Δ μοντέλων και η χρήση επαυξημένης πραγματικότητας, έως θεμελιώδεις καινοτομίες όπως η βαθμονόμηση φυσικών σκηνών και η ανίχνευση ημι-διαφανών/διαφανών αντικειμένων. Η ανίχνευση βάθους αδιαμφισβήτητα συμβάλλει σημαντικά στην διεύρυνση των δυνατοτήτων που προσφέρουν η όραση υπολογιστών και η μηχανική μάθηση, ιδιαίτερα σε συστήματα που προορίζονται να δρουν και να αλληλεπιδρούν στο χώρο, π.χ. σε τηλε-εμβάθυνσης, ρομποτικές ή/και ιατρικές εφαρμογές. Σήμερα, η χαμηλού κόστους υπολογιστική όραση βάθους σε συνδυασμό με την ανάπτυξη μοντέλων βαθιάς μάθησης έχει πληθώρα εφαρμογών σε όλα τα είδη των ταχέως κινούμενων καταστάσεων, από τον αυτοματισμό εργοστασίων έως τον εξοπλισμό αυτοκινήτων

Σχήμα 1.1: *Η τριγωνοποίηση είναι μια μέθοδος που χρησιμοποιείται για τον προσδιορισμό της θέσης ενός 3Δ σημείου με βάση τους νόμους της τριγωνομετρίας που δηλώνουν ότι αν είναι γνωστές η μία πλευρά και οι δύο γωνίες ενός τριγώνου, μπορούν να υπολογιστούν οι άλλες δύο πλευρές και η τρίτη γωνία του.*



και μη επανδρωμένων αεροσκαφών, συμβάλλοντας στην πρόοδο της μηχανικής όρασης και όχι μόνο.

Πέραν από την υπολογιστική όραση βάθους που επιτυγχάνει μερική αναγνώριση μιας 3Δ σκηνής από μια προβολή, υπάρχουν περίπλοκες γεωμετρικές σχέσεις που καταγράφονται ακόμη πιο εύστοχα και λεπτομερώς με τη χρήση πολλαπλών προβολών μέσω της **Υπολογιστικής Όρασης Πολλαπλών Προβολών**. Οι σχέσεις αυτές εξαρτώνται από τη θέση, την περιστροφή και τις εσωτερικές παραμέτρους προβολής, καθώς και από τη δομή της ίδιας της σκηνής. Η υπολογιστική όραση πολλαπλών προβολών έχει προφανές πλεονέκτημα έναντι της μονής προβολής λόγω του μεγαλύτερου και πιο πλήρους εύρους κάλυψης. Όταν μια 3Δ σκηνή προβάλλεται σε κάποια σημεία θέασης, υπάρχουν περιοχές οι οποίες αποκρύπτονται σε ορισμένες από τις όψεις, αλλά είναι ορατές σε άλλες. Ένα σύστημα πολλαπλών καμερών είναι σε θέση να καταγράφει αντικείμενα που είναι μερικώς ή ακόμη και πλήρως καλυμμένα σε κάποιες από τις προβολές. Αυτό επιτυγχάνεται αξιοποιώντας ταυτόχρονη πληροφορία από πολλές διαφορετικές γωνίες θέασης, προσφέροντας τη δυνατότητα υπολογισμού και σύνθεσης 3Δ πληροφορίας με ακρίβεια. Για παράδειγμα, μία ευρέως γνωστή και διαδεδομένη μέθοδος είναι η τριγωνοποίηση (Σχήμα 1.1) η οποία χρησιμοποιείται για τον προσδιορισμό της θέσης ενός 3Δ σημείου με βάση τους νόμους της τριγωνομετρίας, ανάμεσα σε άλλες όπως η στερεοσκοπική εκτίμηση βάθους, η φωτογραμμετρία, ο ταυτόχρονος εντοπισμός και χαρτογράφηση, κ.ο.κ.

1.2 Καταγραφή 3Δ Ανθρώπινης Κίνησης

Με τον όρο **Καταγραφή 3Δ Ανθρώπινης Κίνησης** αναφερόμαστε στην ψηφιοποίηση της κίνησης του σώματος που εκτελούν ένα η περισσότερα άτομα. Η ψηφιοποιημένη κίνηση αναπαριστάται μέσω των 3Δ θέσεων και περιστροφών των αρθρώσεων του σώματος ανά στιγμιότυπο, θέτοντας μία ιεραρχική σχέση μεταξύ τους σε μορφή δέντρου (γράφου). Εκτεταμένες ερευνητικές προσπάθειες έχουν αφιερωθεί στην ανάπτυξη μεθόδων καταγραφής κίνησης. Σήμερα, οι οπτικές λύσεις, ιδίως αυτές που βασίζονται σε δείκτες, θεωρούνται απαραίτητες σε διάφορους βιομηχανικούς τομείς, όπως η παραγωγή ταινιών και VFX, ο αθλητισμός, η υγεία, τα βιντεοπαιχνίδια και οι εφαρμογές μεικτής πραγματικότητας (XR). Οι λύσεις αυτές επιτρέπουν την καταγραφή και ψηφιοποίηση υψηλής πιστότητας των κινήσεων του σώματος, των χεριών και του προσώπου σε πραγματικό ή δεύτερο χρόνο που πραγματοποιούνται από ανθρώπους και όχι μόνο στο φυσικό περιβάλλον. Αυτό επιτρέπει τη χρήση τους σε διάφορες εφαρμογές, όπως οι 3Δ κινούμενοι ψηφιακοί χαρακτήρες βιντεοπαιχνιδιών ή κινουμένων σχεδίων, η αλληλεπίδραση ανθρώπου-μηχανής, ο έλεγχος ρομποτικών μονάδων, η παρακολούθηση σωματικής άσκησης και πολλά άλλα.

Η καταγραφή μπορεί να επιτευχθεί με διαφόρων ειδών μέσα, εξέχουσα θέση μεταξύ των οποίων καταλαμβάνουν οι μέθοδοι που βασίζονται σε υπολογιστική όραση. Ειδικότερα, για καταγραφή υψηλής ακρίβειας συνηθίζεται η χρήση υπολογιστικής όρασης βάθους ή/και πολλαπλών προβολών. Ακόμη πιο ειδικά, η καταγραφή κίνησης με δείκτες (marker-based) και τεχνικές όρασης υπολογιστών αποτελεί τη λύση εκλογής όπου πολλαπλές συγχρονισμένες υπέρυθρες κάμερες υψηλής ανάλυσης και συχνότητας τοποθετημένες γύρω από μια ορισμένη περιοχή λήψης, καταγράφουν κινήσεις ατόμων με προσαρτημένους ειδικούς δείκτες στο σώμα

τους. Τα τελευταία χρόνια ωστόσο, όπως θα αναλυθεί και στη συνέχεια στο Κεφ. 2, έντονη ερευνητική δραστηριότητα παρατηρείται στην προσέγγιση και ανάπτυξη μεθόδων καταγραφής 3Δ κίνησης με χρήση μηχανικής, και ειδικότερα, βαθιάς μάθησης. Στον τομέα αυτό επικεντρώνεται και η παρούσα έρευνα, όπως αυτό παρουσιάζεται στη συνέχεια της διδακτορικής διατριβής.

1.3 Στόχος και Ερευνητικές Συνεισφορές Διδακτορικής Διατριβής

Εδώ και δεκαετίες, η καταγραφή κίνησης με χρήση δεικτών αποτελεί το βασικό πρότυπο για την καταγραφή και παρακολούθηση κίνησης υψηλής πιστότητας. Ωστόσο, παρά την ακρίβεια χιλιοστού που προσφέρουν τα συστήματα αυτά, η χρήση τους είναι περιορισμένη, λόγω υψηλού κόστους εξοπλισμού, αδειών λογισμικού, συντήρησης και άλλων. Με άλλα λόγια, οι υπάρχουσες λύσεις καταγραφής κίνησης που βασίζονται σε δείκτες δεν ικανοποιούν την ανάγκη για ευέλικτες και χαμηλού κόστους επιλογές. Από την άλλη, τα πρόσφατα μοντέλα βαθιάς μάθησης καταγραφής κίνησης που δε χρησιμοποιούν δείκτες ή άλλες προσαρτήσεις στο σώμα, αν και αποτελεσματικά και πιο ευέλικτα, δεν μπορούν να επιτύχουν τα ίδια επίπεδα ποιότητας (σταθερότητας, συνέπειας και ακρίβειας) ελλείψει ισχυρών προκαθορισμών. Εστιάζοντας σε αυτό το ερευνητικό και τεχνολογικό κενό, η προτεινόμενη διδακτορική διατριβή καθοδηγείται από 3 κύρια, αδιαμφισβήτητα, σύγχρονα δεδομένα:

- την υψηλή ακρίβεια που επιτυγχάνουν τα επαγγελματικά συστήματα στην καταγραφή των αρθρωτών κινήσεων του ανθρωπίνου σώματος (και όχι μόνο) με τη χρήση δεικτών προσαρτημένων στο σώμα,
- την αξιοσημείωτη ικανότητα των μοντέλων όρασης υπολογιστών βαθιάς μάθησης να επιλύουν προβλήματα που σχετίζονται με την παρακολούθηση και εκτίμηση της ανθρώπινης πόζας και
- τις πρόσφατες τεχνολογικές εξελίξεις στις κάμερες ανίχνευσης βάθους καταναλωτικής ποιότητας και χαμηλού κόστους.

Συνδυάζοντας τα παραπάνω στοιχεία με το βέλτιστο δυνατό και αποδοτικό τρόπο, στόχος της διδακτορικής διατριβής είναι η ανάπτυξη και ενδελεχής διερεύνηση καινοτόμων 'υβριδικών' μεθόδων καταγραφής κίνησης. Συγκεκριμένα, στόχος είναι η ανάπτυξη των πρώτων μοντέλων βαθιάς μάθησης που θα επιτρέπουν τη χρήση αισθητήρων χαμηλού κόστους και οπισθοανακλαστικών υλικών ή επαγγελματικών δεικτών του εμπορίου, επιτυγχάνοντας ισάξια ποιοτικά καταγραφή κίνησης με τα επαγγελματικά συστήματα που χρησιμοποιούν οπισθοανακλαστικούς δείκτες.

Κατά τη διάρκεια της μελέτης και έρευνας στο συγκεκριμένο τομέα, για την επίτευξη του παραπάνω ερευνητικού στόχου ήταν απαραίτητη η αντιμετώπιση των παρακάτω προκλήσεων:

- Η μοντελοποίηση του προβλήματος καταγραφής κίνησης με δείκτες από δεδομένα βάθους πολλαπλών χρονικά και χωρικά συσχετισμένων προβολών. Αυτό επετεύχθη με τη σχεδίαση και ανάπτυξη μοντέλων βαθιών νευρωνικών δικτύων τα οποία δίνουν τη δυνατότητα

προβλέψεων υψηλής ποιότητας ως προς την ακρίβεια και τη σταθερότητα καταγραφής των κινήσεων, ξεπερνώντας ταυτόχρονα τους ποιοτικούς περιορισμούς που θέτουν στα δεδομένα εισόδου αισθητήρες ή κάμερες χαμηλού κόστους, και κατ΄ επέκταση χαμηλών χαρακτηριστικών και δυνατοτήτων.

- Η έλλειψη ground-truth δεδομένων για την εκπαίδευση τέτοιων μοντέλων. Όπως αναφέρεται στο Κεφ. 7, ιδιαίτερη σημασία δόθηκε στην απλοποίηση του προβλήματος με την εισαγωγή και χρήση κατάλληλων αναπαραστάσεων ανεξάρτητων από τον τρόπο καταγραφής των αρχικών δεδομένων, ή τη θέση και λειτουργία των ίδιων των αισθητήρων, για παράδειγμα, μεταθέτοντας τα 3Δ δεδομένα σε έναν πιο ομογενοποιημένο και κανονικοποιημένο χώρο. Εκπαιδεύοντας με τέτοιου είδους αναπαραστάσεις, τα μοντέλα είναι ικανά να αντεπεξέλθουν αποτελεσματικά, ακόμα και με σχετικά περιορισμένα δεδομένα εκπαίδευσης.

- Η εξαγωγή 3Δ συντεταγμένων από 2Δ και 2.5Δ δεδομένα εισόδου. Έρευνα διεξήχθη στη μελέτη και ανάπτυξη τεχνικών που επιτρέπουν την εκτίμηση 3Δ συντεταγμένων με υψηλή ακρίβεια και απόδοση.

Στην πρόσφατη βιβλιογραφία, μια πληθώρα αλγορίθμων και μοντέλων βαθιάς μάθησης εστιάζουν στην εκτίμηση της 3Δ πόζας, ωστόσο, μόνο λίγες μέθοδοι προσεγγίζουν το πρόβλημα με τη χρήση δεδομένων βάθους πολλαπλών προβολών. Αυτό οφείλεται στην πολυπλοκότητα χρήσης και ρύθμισης τέτοιων συστημάτων, καθώς και στην έλλειψη χωροχρονικά συσχετισμένων χαρτών βάθους πολλαπλών προβολών με ground-truth δεδομένα πόζας. Για το σκοπό αυτό, αφού επιβεβαιώσαμε τη βασική ιδέα καταγραφής κίνησης με χρήση βαθιάς μάθησης και πολλαπλών χαρτών βάθους με οπισθοανακλαστικά υλικά χωρίς υψηλές απαιτήσεις σε ακρίβεια συγχρονισμού και ground-truth δεδομένων πόζας (Κεφ. 3), δημιουργήσαμε το σύνολο δεδομένων HUMAN4D (Κεφ. 4) το οποίο παρέχει δεδομένα βάθους από πολλαπλές κάμερες ταυτόχρονης λήψης (το πρώτο σύνολο με συγχρονισμό υλισμικού σε δεδομένα βάθους πολλαπλών προβολών) και εξαιρετικά ακριβείς 3Δ πόζες (με χρήση επαγγελματικού συστήματος καταγραφής κίνησης) για εποπτεία και αξιολόγηση. Το HUMAN4D μας επέτρεψε την πραγματοποίηση έρευνας και ανάπτυξης σύγχρονων μεθόδων εκτίμησης πόζας και καταγραφής κίνησης (Κεφ. 5) και τεχνικών αναπαράστασης 3Δ συντεταγμένων (Κεφ. 6) για να οδηγηθεί η διατριβή προς την τελική της κατεύθυνση με την ανάπτυξη και πειραματισμό νέων προσεγγίσεων εκτίμησης 3Δ πόζας από δεδομένα βάθους με τη χρήση σφαιρικών οπισθοανακλαστικων δεικτών (Κεφ. 7 - DeMoCap). Ως εκ τούτου, οι τεχνικές και μέθοδοι που μελετήθηκαν και αναπτύχθηκαν είναι εφαρμόσιμες στο σύνολο δεδομένων HUMAN4D και την εκδοχή του με τα δεδομένα υπερύθρων (Κεφ. 7.1), διερευνώντας εκτενώς το συνδυασμό βαθιάς μάθησης και υπολογιστικής όρασης βάθους για την αντικατάσταση των παραδοσιακών οπτικών συστημάτων που βασίζονται σε τριγωνοποίηση ή άλλες παρεμφερείς τεχνικές.

Η χρήση δεικτών και η καταγραφή κίνησης μέσω της παρακολούθησής τους στο χρόνο είναι δισήμαντη για τα μοντέλα βαθιάς μάθησης που αναπτύχθηκαν και παρουσιάζονται στην παρούσα διατριβή. Από τη μία, η προσάρτηση δεικτών προσφέρει υψηλή ακρίβεια στην καταγραφή λόγω της ακριβούς αλλά και σχετικά αβίαστης ανίχνευσης τους στα καταγραφόμενα οπτικά δεδομένα. Μοντέλα που δέχονται ως δεδομένα εισόδου τους δείκτες που είναι προσαρτημένοι

στο σώμα, επικεντρώνονται αποκλειστικά στις πληροφορίες που καλούνται να επιλύσουν και σχετίζονται με την πόζα του σώματος ανά στιγμιότυπο, χωρίς περιττή πληροφορία από τα δεδομένα της σκηνής, όπως συμβαίνει με τα δεδομένα χρώματος ή τα πυκνά δεδομένα βάθους. Από την άλλη, η συνολική πληροφορία εισόδου στα μοντέλα είναι ένα αραιό νέφος 3Δ σημείων, δηλ. ένας σχετικά μικρός αριθμός (π.χ. < 100) από 3Δ σημεία. Αυτό δίνει τη δυνατότητα ανάπτυξης ευέλικτων αρχιτεκτονικών και εκπαίδευσης μοντέλων που απαιτούν σχετικά χαμηλή υπολογιστική ισχύ και, άρα, είναι εφαρμόσιμα σε τεχνολογίες καταγραφής κίνησης πραγματικού χρόνου, σε αντίθεση με τα πυκνά δεδομένα εικόνων χρώματος, χαρτών βάθους ή πυκνών 3Δ δεδομένων.

1.4 Δομή της Διατριβής

Το περιεχόμενο της διατριβής είναι δομημένο ως εξής:

- Στο Κεφ. 2 επικεντρωνόμαστε στην επισκόπηση του ερευνητικού πεδίου παρουσιάζοντας μεθόδους οπτικών δεδομένων, πρόσφατα ερευνητικά μοντέλα και τεχνικές εκτίμησης ανθρώπινης πόζας και κίνησης, τεχνικών αναπαράστασης συντεταγμένων καθώς και σχετικά σύνολα δεδομένων.

- Στο Κεφ. 3, προτείνουμε μια νέα μέθοδο καταγραφής κίνησης (DeepMoCap), που βασίζεται στην σήμανση ανακλαστικού υλικού (ιμάντες και επιθέματα τοποθετημένα στο σώμα του ατόμου προς καταγραφή) με τη χρήση βαθιάς μάθησης και πολλαπλών συγχρονισμένων και χωρικά συσχετισμένων αισθητήρων υπέρυθρης ακτινοβολίας και βάθους.

- Στο Κεφ. 4, προτείνουμε το HUMAN4D, το πρώτο σύνολο δεδομένων που παρέχει συγχρονισμένα με χρήση συγχρονισμού υλισμικού καρέ χρώματος-βάθους μαζί με δεδομένα κίνησης και ήχου, με τη χρήση χαμηλού κόστους καμερών βάθους και συστημάτων καταγραφής κίνησης τελευταίας τεχνολογίας.

- Στο Κεφ. 5, διερευνούμε την αποτελεσματικότητα υπαρχόντων μοντέλων βαθιάς μάθησης εκτίμησης 2Δ και 3Δ πόζας.

- Στο Κεφ. 6, διερευνούμε και συγκρίνουμε μεθόδους και τεχνικές κωδικοποίησης του ground truth χάρτη θερμότητας (heatmaps) και αποκωδικοποίησης του αντίστοιχου προβλεπόμενου χάρτη σε 2Δ συντεταγμένες.

- Στο Κεφ. 7, παρουσιάζουμε το DeMoCap, ένα καινοτόμο μοντέλο βαθιάς μάθησης με στόχο την καταγραφή κίνησης συνδυάζοντας τη χρήση σφαιρικών οπισθοανακλαστικών δεικτών με χάρτες βάθους πολλαπλών προβολών από στερεοσκοπικούς αισθητήρες βάθους χαμηλού κόστους.

- Στο Κεφ. 8, συνοψίζουμε τα συμπεράσματα της μελέτης και τις μελλοντικές ερευνητικές κατευθύνσεις.

Κεφάλαιο 2

Υπόβαθρο και Επισκόπηση Ερευνητικού Πεδίου

Ο ερευνητικός τομέας καταγραφής ανθρώπινης κίνησης μπορεί να οριστεί ως ένα υπερσύνολο των μεθόδων υπολογισμού στατικής ανθρώπινης πόζας και ανθρώπινης πόζας στο χρόνο (ανθρώπινης κίνησης), περιλαμβάνοντας μεγάλη ποικιλία ερευνητικών προσεγγίσεων. Η κατηγοριοποίηση των προσεγγίσεων αυτών μπορεί να βασίζεται:

- στη χρήση δεικτών[1] ή όχι,
- στη χρήση οπτικών ή αδρανειακών, μεταξύ άλλων, αισθητήρων,
- σε δεδομένα από εικόνες χρώματος, χάρτες βάθους ή εικόνες υπερύθρων ή συνδυασμούς,
- στον υπολογισμό ανθρώπινης πόζας στο 2Δ ή 3Δ χώρο,
- στην καταγραφή ενός ή πολλών ατόμων ταυτόχρονα,
- σε πραγματικό, κοντά σε πραγματικό ή σε δεύτερο χρόνο.

Στη συνέχεια, επικεντρωνόμαστε σε μεθόδους οπτικών δεδομένων και παρουσιάζουμε πρόσφατα ερευνητικά μοντέλα και τεχνικές εκτίμησης ανθρώπινης πόζας και κίνησης. Στο κλείσιμο του κεφαλαίου, παραθέτουμε συνοπτικά τα πιο διαδεδομένα ανοικτά σύνολα δεδομένων πολλαπλών προβολών που περιλαμβάνουν δεδομένα 3Δ ανθρώπινης πόζας και κίνησης που αξιοποιούνται σήμερα για την εκπαίδευση και επαλήθευση μοντέλων βαθιάς μάθησης.

2.1 Μέθοδοι Εκτίμησης Ανθρώπινης Πόζας

Κατά την τελευταία δεκαετία, πληθώρα ερευνητών εργάζονται εντατικά στην ανάπτυξη απλών και ευέλικτων προσεγγίσεων χρησιμοποιώντας πόρους χαμηλού κόστους [19]. Οι περισσότερες πρόσφατες μέθοδοι επικεντρώνονται στη όραση μονής προβολής, κυρίως χρησιμοποιώντας έγχρωμες εικόνες [20, 21, 22, 23, 24, 25], καθώς και, λιγότερο συχνά, χρησιμοποιώντας χάρτες βάθους [26, 27, 28, 29]. Ακόμα λιγότερες προσεγγίζουν την εκτίμηση της 3Δ πόζας από χρωματικές ροές πολλαπλών προβολών [6, 30, 31, 32, 33, 34], ενώ η εκτίμηση από χάρτες βάθους πολλαπλών προβολών είναι σχετικά ανεξερεύνητη [35].

[1] παθητικών (οπισθοανακλαστικού υλικού) ή ενεργητικών (τύπου LED)

2.1.1 Μέθοδοι Εκτίμησης Ανθρώπινης Πόζας Μονής Προβολής

Έντονη ερευνητική προσπάθεια έχει αφιερωθεί στην εργασία εκτίμησης 2Δ/3Δ πόζας, αναπτύσσοντας μεθόδους αποτελεσματικές σε απαιτητικές "in-the-wild" βάσεις δεδομένων. Αρχικά, άξιος αναφοράς είναι ο αισθητήρας Microsoft Kinect (Kinect) [36], ο πρώτος χαμηλού κόστους αισθητήρας χρώματος-βάθους (RGB-D) για την εκτίμηση βάθους και 3Δ ανθρώπινης πόζας, επιταχύνοντας την ανάπτυξη μεθόδων υπολογιστικής όρασης που χρησιμοποίησαν τα δεδομένα του ή συγκρίθηκαν με τη μέθοδό του [37, 38, 39, 40]. Γενικότερα, το Kinect επέτρεψε τη μαζική παραγωγή και ευρεία κυκλοφορία RGB-D καμερών χαμηλού κόστους, επιτρέποντας σε μια ευρεία κοινότητα ερευνητών να εργαστούν με τα δεδομένα αυτά και να αναπτύξουν πληθώρα αποτελεσματικών μεθόδων [26, 41, 42, 43, 44].

Πιο πρόσφατα, με την εδραίωση των μοντέλων βαθιάς μάθησης, παρουσιάστηκε μεταξύ άλλων η αρχιτεκτονική εκτίμησης 2Δ πόζας πολλαπλών σταδίων, ονομαζόμενη Pose Machines (PM) [45]. Τα μοντέλα που βασίστηκαν σε αυτή την αρχιτεκτονική επέδειξαν μεγάλη αποτελεσματικότητα αξιοποιώντας την έμμεση εκμάθηση χαρακτηριστικών και συσχετίσεων μεταξύ εικόνας και ανθρώπινης σωματοδομής (κινηματικής ιεραρχίας των διαφόρων μερών του σώματος). Αργότερα, η PM επεκτάθηκε με την ανάπτυξη της Convolutional Pose Machines (CPM) [1, 46, 47] αρχιτεκτονικής, επιτρέποντας την εκμάθηση χαρακτηριστικών τόσο της εικόνας όσο και του ευρύτερου χωρικού πλαισίου απευθείας από τα δεδομένα εκπαίδευσης. Ειδικότερα στο OpenPose [1], σε κάθε στάδιο, οι χάρτες χαρακτηριστικών της εικόνας εισόδου και το αποτέλεσμα που εκτιμάται από το προηγούμενο στάδιο, δηλαδή οι χάρτες εμπιστοσύνης (ή θερμότητας) και τα 2Δ διανυσματικά πεδία, συγχωνεύονται και χρησιμοποιούνται ως είσοδος στο επόμενο, βελτιώνοντας διαδοχικά τις προβλέψεις με χρήση επίβλεψης μεταξύ των σταδίων αυτών.

Τελευταία, ακόμη νεότερες τεχνικές δικτύων βαθιάς μάθησης, προερχόμενες αρχικά από άλλα πεδία εφαρμογής, έχουν εφαρμοστεί με στόχο την εκτίμηση της ανθρώπινης πόζας και κίνησης, όπως για παράδειγμα οι Transformers [48, 49, 50, 51], τα Neural Radiance Fields - NeRF [52, 53, 54], και άλλα.

2.1.2 Μέθοδοι Εκτίμησης Ανθρώπινης Πόζας Πολλαπλών Προβολών

Στο [55], παρουσιάζεται μια μέθοδος πολλαπλών προβολών πραγματικού χρόνου για ταυτόχρονη καταγραφή κίνησης πολλών ατόμων. Πολλαπλές χωρικά συσχετισμένες RGB-D κάμερες τοποθετούνται γύρω από το χώρο καταγραφής. Οι τελικές 3Δ εκτιμήσεις επιτυγχάνονται με τη συγχώνευση εκτιμήσεων 2Δ πόζας μιας όψης από CPM [1, 46], αντιπροβάλοντάς τις στον 3Δ χώρο μέσω της πληροφορίας βάθους. Οι Shafaei et al. [56] χρησιμοποιούν πολλαπλές RGB-D κάμερες σταδιοποιώντας την εκτίμηση πόζας πολλαπλών προβολών σε 3 βήματα: (i) κατάτμηση των εικονοστοιχείων βάθους, (ii) συγκέντρωση εκτιμήσεων πολλαπλών προβολών και (iii) συνεκτίμηση / συγχώνευση 3Δ πόζας. Εφαρμόζοντας πρόσφατες τεχνικές κατάτμησης δεδομένων βάθους, ένα Convolutional Neural Network (CNN) εκπαιδεύεται σε αμιγώς συνθετικά δεδομένα. Οι αρθρώσεις εντοπίζονται με ακρίβεια χωρίς να απαιτείται οποιαδήποτε πρότερη γνώση. Στη συνέχεια, οι θέσεις των αρθρώσεων του σώματος ανακτώνται συνδυάζοντας εκτιμήσεις από πολλαπλές προβολές σε πραγματικό χρόνο, αντιμετωπίζοντας το

πρόβλημα της εκτίμησης πόζας ως ένα πρόβλημα γραμμικής παλινδρόμησης. Οι Shuai et al. [57] προτείνουν μια μέθοδο εκτίμησης 3Δ πόζας με την τεχνική προσαρμογής ενός αρθρωτού μοντέλου ανθρωπίνου σώματος σε ένα πυκνό νέφος σημείων καταγεγραμμένο μέσω πολλαπλών αισθητήρων βάθους. Το μοντέλο αποτελείται από αρθρωτά ελλειψοειδή βασισμένα σε σφαιρικές αρμονικές συναρτήσεις μετατόπισης για την εκτίμηση της 3Δ πόζας του υποκειμένου που καταγράφεται.

Πιο πρόσφατα, στο [6] οι συγγραφείς παρουσιάζουν δύο παραλλαγές μιας τεχνικής βαθιάς μάθησης βασισμένης στην τριγωνοποίηση, μια αλγεβρική και μια ογκομετρική, με στόχο την εκτίμηση της 3Δ πόζας από πολλαπλές 2Δ έγχρωμες προβολές. Η αλγεβρική βασίζεται στην εφαρμογή σταθμισμένης τριγωνοποίησης με βάρη βασισμένα στον υπολογισμό εμπιστοσύνης ανά εκτίμηση προβολής του μοντέλου, ενώ η ογκομετρική βασίζεται στην πυκνή γεωμετρική συνάθροιση εκτιμώμενων 2Δ θερμικών χαρτών από πολλαπλές προβολές σε ένα κοινό, δομημένο 3Δ χώρο. Ο συναθροισμένος βαθμονομημένος όγκος βελτιώνεται στη συνέχεια μέσω 3Δ συνελίξεων για την τελική εκτίμηση 3Δ θερμικών χαρτών που επιτρέπουν τη μοντελοποίηση της ανθρώπινης πόζας.

Οι συγγραφείς του [34] προτείνουν το VoxelPose, μια προσέγγιση ταυτόχρονης εκτίμησης 3Δ πόζας πολλαπλών ατόμων με χρήση δεδομένων πολλαπλών προβολών. Σε αντίθεση με τις προαναφερθείσες μεθόδους πολλαπλών όψεων, των οποίων η αντιστοιχία μεταξύ των όψεων βασίζεται σε εκτιμήσεις 2Δ πόζας, το VoxelPose λειτουργεί εξ' ολοκλήρου στον 3Δ χώρο. Τα χαρακτηριστικά γνωρίσματα από τις προβολές των καμερών συγκεντρώνονται στον 3Δ χώρο και τροφοδοτούνται σε ένα 3Δ δίκτυο για τον εντοπισμό πολλαπλών υποκειμένων στον χώρο, προβλέποντας έναν αριθμό προτάσεων 3Δ κελιών από τον όγκο των 3Δ χαρακτηριστικών. Στη συνέχεια, δημιουργείται ένας νέος, πιο λεπτομερής όγκος χαρακτηριστικών γνωρισμάτων γύρω από την κάθε πρόβλεψη, και τροφοδοτείται σε ένα δίκτυο παλινδρόμησης 3Δ πόζας. Παρά τις συχνές αποκρύψεις μεταξύ των ατόμων στην ίδια σκηνή, η προσέγγιση παραμένει ακριβής.

Άλλες προσεγγίσεις βασίζονται σε παραμετρικά μοντέλα του ανθρωπίνου σώματος [58, 59, 60, 61] εμπερικλείοντας παραμέτρους σχετικά με την δομή και την αρθρωτή ιεραρχία του σκελετού που επιτρέπουν την έκφραση της στάσης και της κίνησης του σώματος. Στο [43], οι συγγραφείς επιστρατεύουν τέτοια μοντέλα για την προσαρμογή τους σε δεδομένα από πολλαπλές οπτικές γωνίες. Αξιοποιώντας τις εντοπίσεις των χαρακτηριστικών του προσώπου, του σώματος και των χεριών με τη χρήση 2Δ ανιχνευτών από πολλαπλές προβολές, εξάγονται 3Δ συντεταγμένες που με τη χρήση παραμετρικών μοντέλων επιτρέπουν την αναγνώριση πρόσθετων, εξωτερικών παραλλαγών του σώματος, όπως μαλλιών και ρούχων.

2.2 Μέθοδοι Καταγραφής Ανθρώπινης Κίνησης

Πέρα από τη στιγμιαία εκτίμηση 2Δ και 3Δ πόζας, έχουν προταθεί μέθοδοι που επιβάλλουν τη διαδοχική γεωμετρική συνοχή και χρονική συσχέτιση μεταξύ των καρέ είτε για την αποφυγή λαθών, όπως για παράδειγμα λόγω στιγμιαίας υποβαθμισμένης ποιότητας των δεδομένων εισόδου (π.χ., θόλωση των εικόνων κατά την πραγματοποίηση γρήγορων κινήσεων ή στιγμιαίες αποκρύψεις μερών του σώματος), είτε για την από κοινού αξιοποίηση των χωροχρονικών χαρακτηριστικών και συσχετίσεων μιας ακολουθίας.

2.2.1 Μέθοδοι Καταγραφής Ανθρώπινης Κίνησης χωρίς τη χρήση δεικτών

Στο [62], οι συγγραφείς επεκτείνουν τη CPM αρχιτεκτονική ενσωματώνοντας ένα μοντέλο χωροχρονικής σχέσης. Δεδομένης της ταυτόχρονης εκμάθησης των 2Δ θέσεων των αρθρώσεων και των χωροχρονικών τους συσχετίσεων σε ένα ενοποιημένο πλαίσιο και της χωροχρονικής κανονικοποίησης κατά τη διαδικασία μάθησης, το μοντέλο επιτυγχάνει υψηλή ακρίβεια και δυνατότητες γενίκευσης. Η οπτική συσχέτιση μεταξύ διαδοχικών καρέ λαμβάνεται υπόψη με την εισαγωγή ενός πρόσθετου επιπέδου πληροφορίας στην αρχιτεκτονική που συσχετίζει χρονικά τους χάρτες χωρικής πιθανότητας των αρθρώσεων. Ομοίως, οι Luo κ.ά. [63] επεκτείνουν τη CPM αρχιτεκτονική με στόχο να συμπεριλάβουν τη χωροχρονική συσχέτιση μεταξύ των καρέ. Ανασχεδιάζοντας τη CPM ως ένα επαναλαμβανόμενο νευρωνικό δίκτυο (Recurrent Neural Network - RNN) και εισάγοντας μονάδες μακράς-βραχυπρόθεσμης μνήμης (Long-Short Term Memory - LSTM) μεταξύ διαδοχικών καρέ, στοχεύουν στην αποτελεσματική εκμάθηση των χρονικών εξαρτήσεων. Αυτή η αρχιτεκτονική, ονομαζόμενη LSTM Pose Machines, αποτυπώνει τις γεωμετρικές σχέσεις των αρθρώσεων στο χρόνο, αυξάνοντας τη σταθερότητα καταγραφής της κίνησης.

Μια άλλη προσέγγιση υπό παρόμοιο πρίσμα με κάποιες από τις προαναφερθείσες μεθόδους, παρουσιάζεται στο [33]. Αρχικά, ένα συνελικτικό βαθύ νευρωνικό δίκτυο εκτιμά από κοινού τις 2Δ πόζες μέσω μιας τεχνικής συγχώνευσης χαρακτηριστικών πολλαπλών όψεων, το οποίο επιτρέπει την ακριβή εκτίμηση. Στη συνέχεια, εφαρμόζοντας ένα αναδρομικό μοντέλο εικονογραφικής δομής (recursive pictorial structure model - RPSM), οι 3Δ συντεταγμένες ανακτώνται από τις 2Δ εκτιμήσεις πολλαπλών όψεων και βελτιώνεται σταδιακά, καθώς το RPSM διακριτοποιεί αναδρομικά τον όγκο γύρω από κάθε άρθρωση μέσω της εκτίμησης σε προηγούμενο χρόνο.

Στο [11], προτείνεται μια νέα προσέγγιση καταγραφής κίνησης πολλαπλών προβολών πολλαπλών ατόμων. Με βάση τους εκτιμώμενους χάρτες θερμότητας και τα πεδία συσχέτισης μερών του σώματος (Part Affinity Fields - PAFs) του OpenPose [1], η προτεινόμενη μέθοδος συσχετίζει τις εκτιμήσεις των αρθρώσεων ανά όψη, τις εκτιμήσεις πολλαπλών όψεων μεταξύ τους και τη χρονική ακολουθία συνεχόμενων καρέ με την εισαγωγή ενός γραφήματος 4Δ συσχέτισης. Ο 4Δ γράφος συσχέτισης επιλύεται ενιαία εκτιμώντας με ακρίβεια τις θέσεις των αρθρώσεων εφαρμόζοντας ευρετική αναζήτηση και τον αλγόριθμο bundle Kruskal, όπως ορίζεται από τους συγγραφείς.

2.2.2 Μέθοδοι Καταγραφής Ανθρώπινης Κίνησης με χρήση Δεικτών

Όπως αναφέρθηκε στο Κεφ. 1.2, τα κλασικά οπτικά συστήματα καταγραφής κίνησης που προϋποθέτουν χρήση δεικτών αποτελούν αναμφίβολα τις κορυφαίες λύσεις εδώ και δεκαετίες. Παρ' όλα αυτά, η επίλυση γνωστών τους μειονεκτημάτων όπως το υψηλό τους κόστος, η ανάγκη επεξεργασίας σε δεύτερο χρόνο για τη διόρθωση αστοχιών, και η πολυπλοκότητα εγκατάστασής τους, θεωρείται πρόκληση και προσελκύει το ενδιαφέρον της ερευνητικής κοινότητας. Παρακάτω, παρουσιάζονται πρόσφατες μέθοδοι και μοντέλα μηχανικής βαθιάς μάθησης εφαρμοσμένα σε τέτοιου τύπου δεδομένα για την αυτόματη διόρθωση της καταγε-

γραμμένης πορείας των δεικτών κατά την κίνηση, καθώς επίσης και την εκτίμηση των θέσεων των αρθρώσεων, που σχεδόν σε όλα τα κλασσικά συστήματα μέχρι σήμερα πραγματοποιείται με μεθόδους πρόσθιας ή ανάστροφης κινηματικής σε συνδυασμό με την παρακολούθηση των δεικτών.

Οι Bascones κ.α. [64] αντιμετωπίζουν την ταυτοποίηση των δεικτών σε πραγματικό χρόνο ως ένα πρόβλημα ταξινόμησης, εκπαιδεύοντας 'αδύναμους' ταξινομητές σε ένα σύνολο συνεργαζόμενων μερικών επιλυτών. Οι συγγραφείς του [65] παρουσιάζουν μια αποτελεσματική μέθοδο για την ταυτοποίηση και την παρακολούθηση των δεικτών, ιδιαίτερα αποδοτική για μεγάλους χώρους καταγραφής και αραιά σύνολα δεικτών. Ο Holden προτείνει μια γρήγορη μέθοδο [66] για την ταυτόχρονη επίλυση των 3Δ θέσεων των δεικτών και των αρθρώσεων με χρήση τεχνικών αποθορυβοποίησης οπτικών δεδομένων δεικτών με χρήση μηχανικής μάθησης. Αυτή η προσέγγιση, η οποία είναι σταθερά αποτελεσματική στην καταγραφή λανθασμένων 3Δ θέσεων των δεικτών, αντικαθιστά την εκτίμηση της κίνησης με χρήση κινηματικής, εξαλείφοντας ταυτόχρονα την ανάγκη για χειροκίνητη επεξεργασία και αποθορυβοποίηση των δεδομένων. Πρόσφατα, οι Perepichka κ.α. [67] εισήγαγαν μια μέθοδο που ανιχνεύει και επιδιορθώνει με υψηλή αξιοπιστία την καταγεγραμμένη πορεία των δεικτών, αντικαθιστώντας τα λανθασμένα τμήματα με συνθετικά. Χρησιμοποιώντας το μοντέλο επίλυσης κίνησης από δείκτες που περιγράφηκε παραπάνω [66], συντίθεται μια αρχική 3Δ κίνηση. Χρησιμοποιώντας την ως αναφορά, ανιχνεύονται μέσω σύγκρισης τα λανθασμένα τμήματα της αρχικής κίνησης και αντικαθίστανται από τα αντίστοιχα της κίνησης αναφοράς, προσαρμόζοντας τα φυσικά ώστε να διατηρούν την σωματοδομή του υποκειμένου.

Οι Han κ.α. [10] παρουσιάζουν ένα μοντέλο ταυτοποίησης δεικτών για την καταγραφή της κίνησης των αρθρώσεων των χεριών-δακτύλων, προσεγγίζοντάς το πρόβλημα από τη σκοπιά της εκτίμησης 2Δ συντεταγμένων και εφαρμόζοντάς το πειραματικά σε ενθόρυβες ακολουθίες περιλαμβάνοντας μη εμφανείς (καλυπτόμενους από μέρη του σώματος - occluded) και ψευδείς (ghost) δείκτες. Το μοντέλο έχει ταχεία απόκριση και είναι ακριβές δεδομένου ότι εκπαιδεύτηκε σε μεγάλο όγκο συνθετικών δεδομένων, αν και δεν αξιολογήθηκε σε ιδιαίτερα θορυβώδη και απαιτητικά σύνολα δεικτών. Το μοντέλο εκτιμά τις 3Δ συντεταγμένες των δεικτών με τη χρήση πλήρως συνδεδεμένων επιπέδων (fully connected layers) στο δίκτυο, τα οποία είναι επιρρεπή σε πολώσεις όταν η διαθεσιμότητα δεδομένων είναι περιορισμένη, δυσχεραίνοντας την ικανότητα γενίκευσης του μοντέλου.

Παρόλο που οι κλασικές προσεγγίσεις δεικτών αποτελούν τη βασική επιλογή για επαγγελματική καταγραφή 3Δ κίνησης, οι παραπάνω ερευνητικές εργασίες αυτής της ενότητας προσπαθούν να ξεπεράσουν γνωστά τους προβλήματα όπως η αυτοματοποιημένη ταυτοποίηση δεικτών, η αποθορυβοποίηση ψευδών δεικτών, η ανάκτηση μη καταγεγραμμένων δεικτών, η λανθασμένη εναλλαγή δεικτών κατά την καταγραφή, καθώς και η επίλυση της ανθρώπινης κίνησης από δείκτες.

2.3 Αναπαράσταση Συντεταγμένων

Στην παρούσα διατριβή, προσεγγίζουμε την καταγραφή κίνησης κυρίως ως ένα συνεργατικό έργο ταυτοποίησης και παλινδρόμησης 3Δ συντεταγμένων και, ως εκ τούτου, παραθέτουμε μια επισκόπηση των σύγχρονων προσεγγίσεων κωδικοποίησης και αποκωδικοποίησης

συντεταγμένων. Πολλά προβλήματα υπολογιστικής όρασης τίθενται ως εργασίες εκτίμησης συντεταγμένων, όπου μοντέλα βαθιάς μάθησης έχουν αποδειχθεί πολύ αποτελεσματικά. Στην πρόσφατη βιβλιογραφία, οι σύγχρονες μέθοδοι για τον εντοπισμό συντεταγμένων χωρίζονται σε τρεις κύριες κατηγορίες: την άμεση παλινδρόμηση, την πρόβλεψη μέσω χαρτών θερμότητας και τη χωρική παλινδρόμηση.

Οι μέθοδοι άμεσης παλινδρόμησης, που έχουν χρησιμοποιηθεί σε σημαντικές ερευνητικές μεθόδους όπως η εκτίμηση πόζας στο DeepPose [68] και η επισήμανση δεικτών καταγραφής κίνησης [10], αγνοούν σε μεγάλο βαθμό τη χωρική φύση των εικόνων στο 2Δ επίπεδο λόγω της μονοδιάστατης αναπαράστασής τους. Έτσι, αυτές οι μέθοδοι καταλήγουν να προσεγγίζουν $2\Delta/3\Delta$ συντεταγμένες στηριζόμενες στην εκφραστική δύναμη των μοντέλων βαθιάς μάθησης και όχι σε χωρικές συσχετίσεις των δεδομένων εισόδου και των εξαγόμενων χαρακτηριστικών. Ωστόσο, αξίζει να σημειωθεί ότι η άμεση παλινδρόμηση επιτρέπει άμεση εποπτεία κατά την εκπαίδευση με συναρτήσεις απώλειας ευθέως συσχετισμένες με την απόσταση μεταξύ των συντεταγμένων, τον άμεσο στόχο δηλαδή του βασικού προβλήματος.

Οι μέθοδοι εκτίμησης συντεταγμένων που βασίζονται σε πυκνούς χάρτες θερμότητας χρησιμοποιούν πλήρως συνελικτικά νευρωνικά δίκτυα για να προβλέψουν τη χωρική πιθανότητα για κάθε εικονοστοιχείο του χάρτη. Κατά την εκπαίδευση, κατασκευάζονται χάρτες θερμότητας για την επίβλεψη των προβλέψεων όπου, ως επί το πλείστον, χρησιμοποιούνται 2Δ Γκαουσιανές κατανομές με σταθερή διακύμανση. Τέτοιες τεχνικές έχουν χρησιμοποιηθεί σε ευρέως διαδεδομένες ερευνητικές εργασίες εκτίμησης πόζας [1, 46], παρουσιάζοντας γενικότερα υψηλότερη απόδοση και καλύτερη γενίκευση από την άμεση παλινδρόμηση λόγω της χωρικής τους φύσης, επιτρέποντας στις προβλέψεις να μην έχουν χωρική πόλωση. Σε αυτές τις μεθόδους, οι εκτιμώμενες συντεταγμένες προσδιορίζονται με αποκωδικοποίηση *ArgMax* ή εναλλακτικών ευριστικών προσεγγίσεων [69], όπως αναλύσουμε στο Κεφ. 6, στον εκτιμώμενο θερμικό χάρτη. Το κύριο μειονέκτημα αυτών των μεθόδων είναι η χρήση ενδιάμεσων συναρτήσεων απωλειών που εκπαιδεύουν το δίκτυο να προβλέπει θερμικούς χάρτες προσεγγίζοντας τους ανακατασκευασμένους, όχι τις ground truth συντεταγμένες. Επομένως, η εποπτεία κατά την εκπαίδευση είναι έμμεση, σε αντίθεση με την άμεση παλινδρόμηση όπου είναι άμεση, όπως προαναφέρθηκε.

Οι μέθοδοι χωρικής παλινδρόμησης έχουν αποδειχθεί οι πλέον αποτελεσματικές συνδυάζοντας τα πλεονεκτήματα της άμεσης παλινδρόμησης και της πρόβλεψης μέσω πυκνών θερμικών χαρτών. Ειδικότερα, τα μοντέλα που εκπαιδεύονται με χωρική παλινδρόμηση καταφέρνουν να αντεπεξέλθουν καλύτερα και να ξεπεράσουν την χωρική πόλωση με τη βοήθεια των χωρικών χαρτών θερμότητας, επιτρέποντας ταυτόχρονα την εποπτεία με χρήση συναρτήσεων απώλειας βασισμένες στην απόσταση μεταξύ των προβλεπόμενων και των ground truth συντεταγμένων. Ενώ οι μέθοδοι πρόβλεψης πυκνών θερμικών χαρτών εκπαιδεύουν το δίκτυο να προβλέπει θερμικούς χάρτες που ταιριάζουν με τους προκαθορισμένους, τα δίκτυα χωρικής παλινδρόμησης μαθαίνουν τον βέλτιστο θερμικό χάρτη που κωδικοποιεί την ακριβέστερη θέση των συντεταγμένων ενός σημείου σε επίπεδο υπο-εικονοστοιχείου. Η χωρική παλινδρόμηση, η αλλιώς το κέντρο μάζας ενός θερμικού χάρτη πιθανοτήτων [70], εισήχθη ως έννοια σχεδόν ταυτόχρονα ως *integral regression* [71], ως *differentiable spatial-to-numerical transform* (DSNT) [72], και ως *Soft-argmax* [73].

2.4 Σύνολα Δεδομένων Πολλαπλών Προβολών

Τις τελευταίες δεκαετίες, η ερευνητική κοινότητα έχει δείξει αυξημένο ενδιαφέρον για την μελέτη τεχνολογιών σχετικών με την προσομοίωση, αναπαραγωγή ή/και ψηφιοποίηση της ανθρώπινης φύσης (π.χ., κίνησης, συμπεριφοράς, ομιλίας, κτλ.) και ειδικότερα τεχνολογίες που σχετίζονται με την ψηφιακή ανθρώπινη αναπαράσταση (εικονικούς/ψηφιακούς ανθρώπους). Μια ποικιλία μεθόδων υπολογιστικής όρασης στοχεύουν σε ανοικτά ερευνητικά προβλήματα που χρησιμοποιούν δεδομένα κίνησης, 3Δ αναπαράστασης, εικόνας, ήχου, κ.α. Σε αυτή την ενότητα, παραθέτουμε σύνολα δεδομένων [2, 74, 75, 76, 77, 78] ανθρώπινης αναπαράστασης, πόζας και κίνησης σχετικά με το αντικείμενο της παρούσας διδακτορικής διατριβής, σημειώνοντας τις δυνατότητες που προσφέρει το καθένα στην ερευνητική κοινότητα. Ακολουθεί μια σύντομη επισκόπηση αυτών των συνόλων δεδομένων, ενώ ο πίνακας 2.1 συνοψίζει τα χαρακτηριστικά και τις ιδιότητές τους.

MHAD [77]: Ένα από τα πρώτα δημόσια διαθέσιμα σύνολα δεδομένων που προσφέρουν από κοινού RGBD δεδομένα και δεδομένα κίνησης είναι το (Berkeley) MHAD. Το σύνολο δεδομένων MHAD περιέχει χωροχρονικά συσχετισμένα δεδομένα που έχουν καταγραφεί με ένα επαγγελματικό σύστημα καταγραφής κίνησης με ενεργούς δείκτες [79] μαζί με 12 RGB και 2 MS Kinect v2 (RGBD) κάμερες, 6 φορητούς αδρανειακούς αισθητήρες (μόνο επιταχυνσιόμετρα) και 4 μικρόφωνα, καταγράφοντας τα ηχητικά σήματα κατά την εκτέλεση των δραστηριοτήτων. Το σύνολο δεδομένων αποτελείται από 659 συνολικά ακολουθίες δεδομένων, περιλαμβάνοντας 11 διαφορετικές δραστηριότητες, εκτελεσμένες από 12 διαφορετικά άτομα, επανεκτελώντας την κάθε δραστηριότητα 5 φορές. Αν και το MHAD επιτρέπει την έρευνα για την εκτίμηση πόζας πολλαπλών όψεων και όχι μόνο, οι συσκευές MS Kinect v2 είναι μόνο 2 και δεν υποστηρίζουν ακριβή συγχρονισμό, με αποτέλεσμα την ύπαρξη χωροχρονικών διαφορών μεταξύ των απεικονιζόμενων χαρτών βάθους και των 3Δ δεδομένων κίνησης, περιορίζοντας με αυτόν τον τρόπο την κοινή τους αξιοποίηση.

Human3.6M [74]: Το Human3.6M (H36M) περιέχει 3,6 εκατομμύρια 3Δ ανθρώπινες πόζες 5 γυναικών και 6 ανδρών. Τα υποκείμενα εκτελούν ένα σύνολο κινήσεων (που έχουν καταγραφεί με 10 κάμερες καταγραφής κίνησης) από καθημερινές ανθρώπινες δραστηριότητες (λήψη φωτογραφιών, ομιλία στο τηλέφωνο, φαγητό, κάθισμα, κ.λπ.), μαζί με συγχρονισμένες εικόνες από 4 συγχρονισμένες RGB κάμερες, χάρτες βάθους από 1 αισθητήρα βάθους Time-of-Flight (ToF) και ακριβείς 3Δ σαρώσεις σώματος των υποκειμένων. Το H36M αποτελεί ένα από τα πιο ευρέως χρησιμοποιούμενα σύνολα δεδομένων για ερευνητικές εργασίες όρασης υπολογιστών συσχετισμένες με ανθρώπινη παρακολούθηση-καταγραφή. Ωστόσο, μόνο οι RGB κάμερες υποστηρίζουν ακριβή συγχρονισμό, υπάρχει μόνο ένας αισθητήρας βάθους με χαμηλή ανάλυση βάθους, ενώ το σύνολο των καμερών καταγραφής κίνησης είναι σχετικά περιορισμένο (10). Τέλος, οι πρόσφατες ερευνητικές μελέτες επικεντρώνονται σε λήψεις πολλαπλών υποκειμένων (π.χ., κοινωνικών δραστηριοτήτων και άλλων αλληλεπιδράσεων), όπως σε μεταγενέστερα σύνολα δεδομένων [2, 76], σε αντίθεση με το H36M που περιέχει μόνο ακολουθίες ενός προσώπου.

CMUPanoptic [2]: Το CMUPanoptic (CMU) είναι ένα από τα μεγαλύτερα δημόσια σύνολα δεδομένων όσον αφορά τον αριθμό των προβολών κάμερας (521), το οποίο αποτυπώνει αλληλεπιδράσεις έως και οκτώ (8) ατόμων που εκτελούν κοινωνικές δραστηριότητες με μη

στρατευμένη συμπεριφορά και προκαθορισμένη εμφάνιση. Το σύνολο δεδομένων έχει καταγραφεί με τη χρήση του Panoptic Studio [2], ενός συστήματος μαζικής καταγραφής πολλαπλών προβολών που αποτελείται από 480 VGA, 31 HD και 10 RGBD (Kinect v2) κάμερες, κατανεμημένες στην επιφάνεια ενός σφαιρικού θόλου. Πέρα από τις ανθρώπινες πόζες, το σύνολο CMU περιλαμβάνει 3Δ δεδομένα των χαρακτηριστικών του προσώπου και 2Δ/3Δ δεδομένα για τις πόζες των χεριών. Το CMU αποτελεί επί του παρόντος ένα από τα πλουσιότερα δημόσια διαθέσιμα σύνολα δεδομένων στον τομέα, ωστόσο, παρά τη χωροχρονική του ρύθμιση, δεν παρέχει απόλυτα ακριβή συγχρονισμό, δεδομένου ότι επιτυγχάνεται μέσω τροποποίησης των συσκευών Kinect v2 μέσω της συστοιχίας των μικροφώνων τους, ανίκανη να παρέχει ακρίβεια συγχρονισμού συγκρίσιμη με την εργοστασιακή υποστήριξη συγχρονισμού υλισμικού. Τέλος, οι κινήσεις δεν έχουν καταγραφεί με τη χρήση επαγγελματικού συστήματος καταγραφής κίνησης με δείκτες που προσφέρει απόλυτη ακρίβεια στις καταγραφές, αλλά μια προσέγγιση πολλαπλών προβολών χωρίς δείκτες, αρκετά ακριβής ωστόσο.

HUMBI [76]: Ένα άλλο μεγάλο και δημόσια διαθέσιμο σύνολο δεδομένων πολλαπλών προβολών είναι το HUMBI, το οποίο επικεντρώνεται στις εκφράσεις του ανθρώπινου σώματος με έμφαση στην ενδυμασία, με στόχο να διευκολύνει τη μοντελοποίηση της εμφάνισης, της γεωμετρίας και της αναπαράστασης του βλέμματος, του προσώπου, των χεριών, του σώματος και του ενδύματος από πολλά και διαφορετικά άτομα. Το HUMBI ξεπερνά κατά πολύ τα λοιπά διαθέσιμα δημόσια σύνολα δεδομένων όσον αφορά τον αριθμό των προβολών (107 συγχρονισμένες κάμερες HD) και των υποκειμένων (772 διαφορετικά υποκείμενα με διαφοροποιήσεις ως προς το φύλο, την εθνικότητα, την ηλικία και τη φυσική κατάσταση). Το σύνολο δεδομένων περιλαμβάνει 5 βασικές εκφράσεις του σώματος: το βλέμμα, το πρόσωπο, τα χέρια, το σώμα και το ένδυμα. Με τη χρήση του παραμετρικού μοντέλου SMPL [58], το HUMBI παρέχει πληροφορίες σχετικά με την 3Δ γεωμετρία των υποκειμένων μαζί με τους αντίστοιχους άτλαντες πληροφορίας χρώματος.

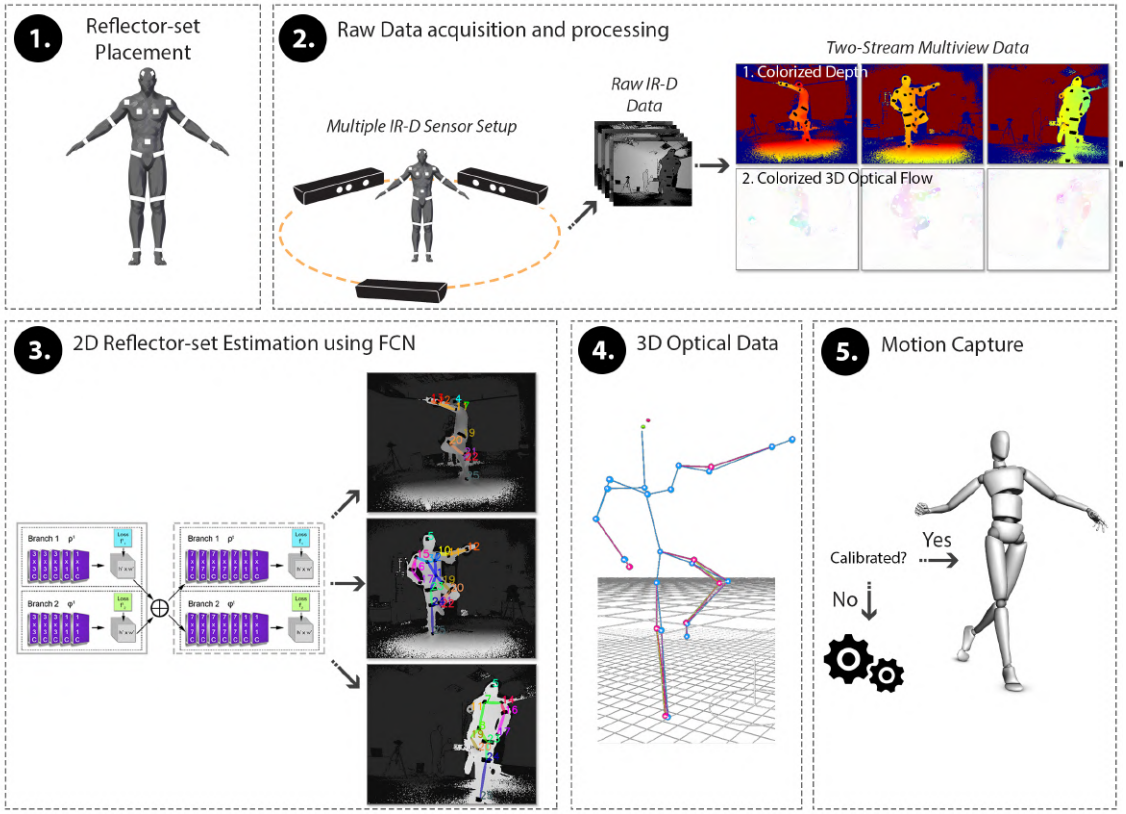
	MHAD$_{(2013)}$ [77]	**Human3.6M**$_{(2014)}$ [74]	**CMUPanoptic**$_{(2015)}$ [2]	**HUMBI**$_{(2018)}$ [76]
Body Pose	✓	✓	✓	✓
Marker-based MoCap	✓	✓	✗	✗
Body Part Segments	✗	✗	✗	✓
Multi-view RGB	✓	✓	✓	✓
Multi-view Depth	✓	✗	✓	✗
3D Meshes	✗	✗	✗	✓
Point-clouds	✗	✗	✓	✗
Audio Cues	✓	✗	✗	✗
Gaze Features	✗	✗	✗	✓
Hand Features	✗	✗	✓	✓
Facial Features	✗	✗	✓	✓
Rigged Characters	✗	✓	✗	✗
Multi-person	✗	✗	✓	✗

Πίνακας 2.1: *Σύνοψη των σύγχρονων συνόλων δεδομένων ως προς τα διαθέσιμα χαρακτηριστικά και τις λειτουργίες τους.*

Κεφάλαιο 3

DeepMoCap: Καταγραφή ανθρώπινης κίνησης με χρήση βαθιάς μάθησης, αισθητήρων βάθους και οπισθο-ανακλαστικής ταινίας

Στο παρόν κεφάλαιο, προτείνεται μια νέα μέθοδος καταγραφής κίνησης ενός ατόμου (DeepMoCap), με τη χρήση πολλαπλών συγχρονισμένων και χωρικά συσχετισμένων αισθητήρων υπέρυθρης ακτινοβολίας και βάθους (IR-D) και ανακλαστικούς ιμάντες και επιθέματα (μη τυποποιημένους δείκτες) τοποθετημένα στο σώμα του ατόμου προς καταγραφή. Συγκεκριμένα, η μέθοδος αναπτύχθηκε αξιοποιώντας ένα πρωτότυπο σύνολο δεδομένων το DMC (DMC2.5D και DMC3D), που παρουσιάζεται στη συνέχεια του κεφαλαίου. Η μέθοδος αυτή διερευνά την καταγραφή κίνησης που βασίζεται στον 2Δ εντοπισμό των δεικτών σε εικόνες βάθους και οπτικής ροής και, στη συνέχεια, στον 3Δ χώρο. Εισάγοντας μια μη παραμετρική αναπαράσταση για την κωδικοποίηση της χρονικής συσχέτισης μεταξύ ζευγών χρωματισμένων χαρτών βάθους και 3Δ οπτικής ροής, προτείνεται μια αρχιτεκτονική πολλαπλών σταδίων *Πλήρως Συνελικτικού Δικτύου* (FCN) για την από κοινού εκμάθηση των θέσεων των δεικτών και της χρονικής τους εξάρτησης μεταξύ διαδοχικών στιγμιοτύπων. Οι εξαγόμενες 2Δ θέσεις αντιπροβάλονται στον 3Δ χώρο με την αξιοποίηση της πληροφορίας βάθους, με αποτέλεσμα την εκτίμηση 3Δ δεδομένων. Η 3Δ ανθρώπινη πόζα προσεγγίζεται με τεχνικές βελτιστοποίησης της εφαρμογής ενός αρθρωτού μοντέλου στις εκτιμώμενες 3Δ θέσεις των δεικτών. Το μοντέλο που προτείνεται υπερτερεί στο σύνολο δεδομένων DMC2.5D κάνοντας χρήση της μετρικής Percentage of Correct Keypoints (PCK), ενώ τα αποτέλεσματα της τελικής καταγραφής 3Δ κίνησης αξιολογούνται έναντι προσεγγίσεων συνδυασμού δεδομένων υπολογισμού πόζας από RGB-D δεδομένα και αδρανειακών δεδομένων στο DMC3D, υπερτερώντας έναντι της αμέσως καλύτερης μεθόδου κατά 4,5% στη συνολική ακρίβεια 3D PCK.



Σχήμα 3.1: *Επισκόπηση της μεθόδου DeepMoCap. Μετά την τοποθέτηση των ανακλαστικών δεικτών στο σώμα (1), λαμβάνονται και υποβάλλονται σε επεξεργασία χωροχρονικά συσχετισμένα IR-D δεδομένα (2) για να τροφοδοτηθεί το FCN δίκτυο με χρωματισμένους χάρτες βάθους και 3Δ οπτικής ροής (3). Η απόκριση του FCN, δηλαδή οι εκτιμήσεις του συνόλου δεικτών πολλαπλών όψεων, συγχωνεύεται για την εξαγωγή των 3Δ δεδομένων (4) και, τέλος, για την εκτίμηση της 3Δ πόζας και κίνησης του υποκειμένου (5).*

Το DeepMoCap αποτελεί μια γρήγορη, σχεδόν πραγματικού χρόνου, προσέγγιση βασισμένη σε δείκτες που καταναλώνει δεδομένα IR-D πολλαπλών προβολών και επιτρέπει την 3Δ καταγραφή κίνησης ενός ατόμου. Τα βήματα της προτεινόμενης μεθόδου, που απεικονίζονται στο Σχήμα 3.1, συνοψίζεται ως εξής:

- Στο σώμα του υποκειμένου τοποθετούνται οι μη τυποποιημένοι δείκτες, δηλ. ένα σύνολο από οπισθοανακλαστικούς ιμάντες και επιθέματα.
- Τοποθετώντας πολλαπλούς χωροχρονικά συσχετισμένους αισθητήρες γύρω από το υποκείμενο για την πλήρη καταγραφή των κινήσεων, καταγράφονται και επεξεργάζονται IR-D δεδομένα, δίνοντας ζεύγη πολλαπλών προβολών χρωματισμένου βάθους και 3Δ οπτικής ροής.
- Κάθε τέτοιο ζεύγος τροφοδοτείται στο προτεινόμενο μοντέλο με αποτέλεσμα την εκτίμηση του συνόλου των 2Δ θέσεων των δεικτών ανά προβολή.
- Οι εκτιμήσεις του συνόλου των δεικτών αντιπροβάλλονται στον 3Δ χώρο αξιοποιώντας τα δεδομένα βάθους και τις εσωτερικές και εξωτερικές παραμέτρους του κάθε αισθητήρα. Τα προκύπτοντα 3Δ σύνολα σημείων συγχωνεύονται, με αποτέλεσμα την εκτίμηση των τελικών 3Δ δεδομένων των δεικτών και των αρθρώσεων.

3.1 Σύνολο Δεδομένων DMC

Για την ανάπτυξη της μεθόδου που αποτελεί την πρώτη ερευνητική εργασία που αξιοποιεί αισθητήρες βάθους και οπισθοανακλαστικό υλικό και χρήση βαθιάς μάθησης για την καταγραφή ανθρώπινης κίνησης, η δημιουργία ενός ειδικά σχεδιασμένου συνόλου δεδομένων, του ***DMC***, κρίθηκε απαραίτητη λόγω της περιορισμένης δημόσιας διάθεσης δεδομένων βάθους πολλαπλών προβολών που καταγράφουν κινήσεις υποκειμένων με τοποθετημένους ειδικά κατασκευασμένους οπισθοανακλαστικούς δείκτες στο σώμα τους. Το DMC σύνολο περιλαμβάνει δύο δημόσια υποσύνολα δεδομένων, τα ***DMC2.5D*** και ***DMC3D***[1], που περιλαμβάνουν ακολουθίες υποκειμένων που πραγματοποιούν φυσικές ασκήσεις με μη τυποποιημένους δείκτες τοποθετημένους στο σώμα τους. Η χρήση μη τυποποιημένων δεικτών από οπισθοανακλαστική ταινία προτιμήθηκε αντί των εμπορικά τυποποιημένων σφαιρικών δεικτών μικρής διαμέτρου για λόγους υψηλότερης ευκρίνειας από τους αισθητήρες.

3.1.1 Στόχος

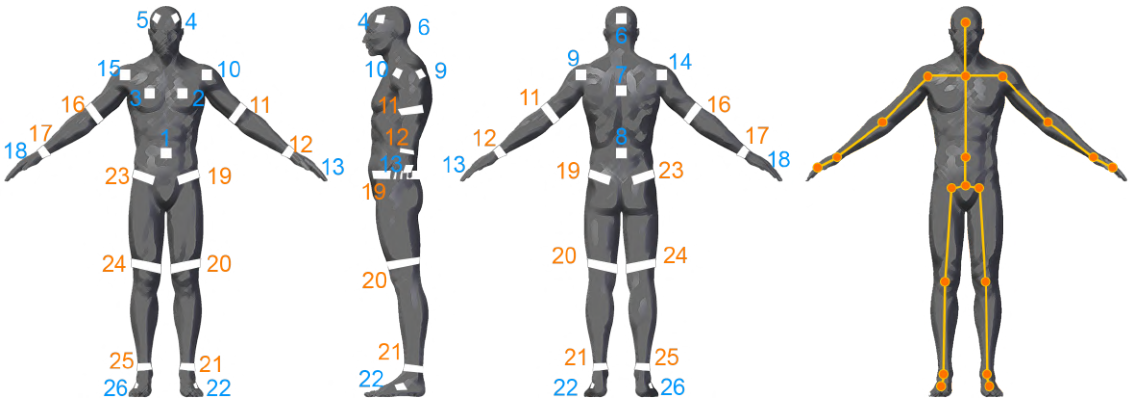
Τα DMC υποσύνολα δεδομένων δημιουργήθηκαν με στόχο την αξιοποίησή τους για i) την εκπαίδευση των μοντέλων εντοπισμού και ταυτοποίησης δεικτών και καταγραφής κίνησης, ii) την αξιολόγησή τους σε σύγκριση με πρόσφατες μεθόδους καταγραφής κίνησης και ground truth δεδομένα και, τέλος, iii) την επαλήθευση της βασικής κεντρικής ιδέας ότι δείκτες από ανακλαστικό υλικό μπορούν να εντοπισθούν με χρήση δεδομένων βάθους και να ταυτοποιηθούν με χρήση βαθιάς μάθησης, στοχεύοντας στην εκτίμηση της ανθρώπινης πόζας του υποκειμένου με αποδοτικό και αποτελεσματικό τρόπο.

3.1.2 Μεθοδολογία Δημιουργίας

Τοποθέτηση δεικτών. Η τοποθέτηση των μη τυποποιημένων δεικτών σχεδιάστηκε έτσι ώστε να παρέχει αξιόπιστα και ακριβή δεδομένα καταγραφής κίνησης, επιτρέποντας τον υπολογισμό υψηλού αριθμού βαθμών ελευθερίας (Degrees-of-Freedom - DoF) των αρθρώσεων του σώματος. Η τοποθέτηση που επιλέχθηκε απεικονίζεται στο Σχήμα 3.2, αποτελούμενη από ένα σύνολο 26 δεικτών $R_i \in \{R_1, \ldots, R_{26}\}$, 16 τετραγωνικών και 10 σε μορφή ιμάντα, επιτρέποντας την καταγραφή κίνησης αρθρωτής δομής 40 DoFs. Η ταυτόχρονη χρήση ιμάντων όσο και τετραγωνικών δεικτών επιλέχθηκε αφενός λόγω του ότι οι ιμάντες είναι πλήρως ορατοί (360°) στα κυλινδρικά μέρη του σώματος (π.χ. χέρια - πόδια), και αφετέρου οι τετραγωνικοί χρησιμοποιήθηκαν στα μέρη του σώματος όπου η τοποθέτηση ιμάντων δεν ήταν εφικτή, δηλαδή στον κορμό, το κεφάλι, τις πατούσες και τις παλάμες. Με στόχο τη διάκριση μεταξύ του θώρακα και της πλάτης (πρόσθιας και οπίσθιας όψης), οι αντίστοιχοι τετραγωνικοί δείκτες δεν τοποθετήθηκαν συμμετρικά. Στην πρόσθια όψη, δύο δείκτες τοποθετήθηκαν στο κεφάλι, δύο στο στήθος και ένας στην κοιλιακή χώρα, ενώ στην οπίσθια όψη, ένας τοποθετήθηκε στο κεφάλι, ένας στην πλάτη και ένας στη σπονδυλική στήλη, στο ύψος της οσφυϊκής χώρας. Το ανακλαστικό υλικό που χρησιμοποιήθηκε για τη δημιουργία του συνόλου των δεικτών είναι μια ευρέως διαθέσιμη ανακλαστική ταινία πλάτους 5 cm. Ακολουθώντας τις οδηγίες που απεικονίζονται στο Σχήμα 3.2, η τοποθέτηση των δεικτών είναι μια γρήγορη διαδικασία (διαρκεί

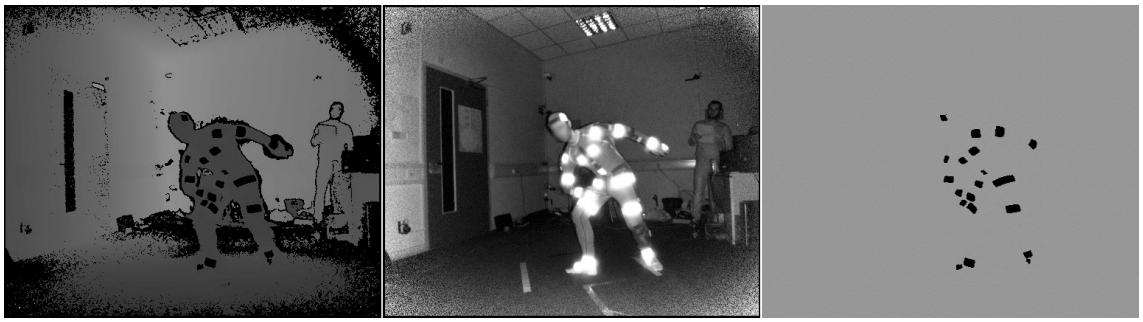
[1] https://vcl.iti.gr/deepmocap/dataset

λιγότερο από 2 λεπτά), δεδομένου ότι δεν απαιτεί υψηλή ακρίβεια (αρκεί να τοποθετηθούν κατά προσέγγιση στις αντίστοιχες θέσεις των μερών του σώματος).



Σχήμα 3.2: *Προτεινόμενη τοποθέτηση δεικτών. Τοποθέτηση ανακλαστικών ιμάντων (πορτοκαλί) και τετραγωνικών δεικτών (μπλε) στο σώμα του υποκειμένου.*

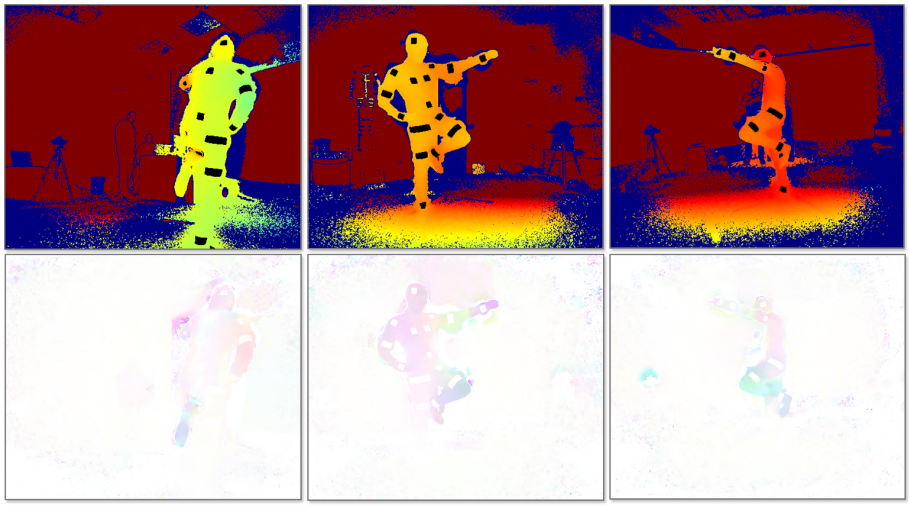
Λήψη και επεξεργασία δεδομένων. Μετά την τοποθέτηση των δεικτών, το υποκείμενο είναι έτοιμο για καταγραφή. Θεωρούμε τη χρήση N αισθητήρων IR-D, και επομένως N προβολών, $v \in \{1, \ldots, N\}$. Με τη χρήση της διάταξης πολλαπλών Kinect for Xbox One που προτείνεται στο [80], καταγράφηκαν χωροχρονικά συσχετισμένα IR-D δεδομένα πολλαπλών όψεων. Οι περιοχές των ανακλαστικών δεικτών έχουν διακριτές τιμές εικονοστοιχείων στις εικόνες IR $\mathcal{I}^v_{IR}$ (Σχήμα 3.3 - μέση), επομένως, εφαρμόζοντας δυαδικό σκληρό κατωφλίωμα, εξάγεται εύκολα η δυαδική μάσκα $\mathcal{I}^v_{IR_m}$ (Σχήμα 3.3 - δεξιά). Οι αντίστοιχες περιοχές στις ακατέργαστες εικόνες βάθους $\mathcal{I}^v_D$ (Σχήμα 3.3 - αριστερά) έχουν μηδενικές τιμές λόγω των οπισθοανακλάσεων που 'τυφλώνουν' τις μετρήσεις του αισθητήρα.



Σχήμα 3.3: *Ακατέργαστα δεδομένα βάθους και υπέρυθρης ακτινοβολίας.*

Κατόπιν, τα IR-D δεδομένα πολλαπλών όψεων υποβλήθηκαν σε επεξεργασία για την καλύτερη αξιοποίησή τους στην εκπαίδευση μοντέλων μηχανικής μάθησης. Αφενός, πραγματοποιήθηκε από κοινού επεξεργασία στα καρέ βάθους και υπερύθρων για τον υπολογισμό 3Δ οπτικής ροής $\mathcal{I}^v_F$ των κινήσεων του σώματος [81]. Συγκεκριμένα, υπολογίστηκαν τα διανύσματα 3Δ κίνησης μεταξύ δύο ζευγών εικόνων IR-D. Στη συνέχεια, η 3Δ ροή απεικονίστηκε κανονικοποιώντας τις τιμές κάθε άξονα και μετατρέποντας τα διανύσματα 3Δ κίνησης σε εικόνα 3 καναλιών. Παράλληλα, οι χάρτες βάθους $\mathcal{I}^v_D$ χρωματίστηκαν εφαρμόζοντας την παλέττα χρωματικού χάρτη JET. Τέλος, η μάσκα $\mathcal{I}^v_{IR_m}$ αφαιρέθηκε από τις χρωματισμένες εικόνες βάθους θέτοντας μηδενικές τιμές χρώματος (μαύρο) στις περιοχές των δεικτών στο

χρωματισμένο βάθος, $\mathcal{I}^v_{CD}$, διευκολύνοντας την εντόπιση των δεικτών. Εφαρμόσαμε χρωματισμό των δεδομένων προκειμένου να επιτραπεί η χρήση του προτεινόμενου μοντέλου, τα πρώτα 10 επίπεδα του οποίου είναι κοινά με το VGG-19 [82]. Ένα παράδειγμα επεξεργασμένων δεδομένων πολλαπλών προβολών από το ίδιο στιγμιότυπο παρουσιάζεται στο Σχήμα 3.4.



Σχήμα 3.4: *Είσοδος δύο ροών πολλαπλών προβολών.* ***(Πάνω)*** *Χρωματισμένοι χάρτες βάθους, με αφαίρεση μάσκας.* ***(Κάτω)*** *Χρωματισμένη 3Δ οπτική ροή.*

3.1.3 Υποσύνολα και Στατιστικά

Υποσύνολο Δεδομένων DMC3D

Το σύνολο δεδομένων DMC3D αποτελείται από δεδομένα βάθους-υπερύθρων και πόζες πολλαπλών προβολών, καθώς και από δεδομένα αισθητήρων αδράνειας (Inertial Measurement Units - IMUs) και ground truth δεδομένων κίνησης. Συγκεκριμένα, χρησιμοποιήθηκαν 3 αισθητήρες Kinect for Xbox One για τη λήψη των δεδομένων βάθους-υπερύθρων και πόζας, και 9 IMUs XSens MT [83] για την καταγραφή αδρανειακών δεδομένων από 9 μέρη του σώματος (κεφάλι, κορμό, βραχίονες, πήχεις, μηρούς, κνήμες). Για τη λήψη των ground truth δεδομένων, χρησιμοποιήθηκε το σύστημα PhaseSpace Impulse X2 [84]. Το PhaseSpace Impulse X2, που αποτελεί και αυτό ένα σύστημα καταγραφής κίνησης βασισμένο σε δείκτες, θεωρήθηκε κατάλληλο για τη δημιουργία του συνόλου δεδομένων λόγω της χρήσης ενεργών (LED) αντί παθητικών δεικτών που θα ήταν το ίδιο ορατοί με τους μη τυποποιημένους και θα υπήρχε σύγχυση στο διαχωρισμό τους. Φωτογραφία από καταγραφές με PhaseSpace Impulse X2 απεικονίζεται στο Σχήμα 3.5.

Η προετοιμασία του συνόλου δεδομένων DMC3D απαιτούσε τη χωροχρονική συσχέτιση των συστημάτων καταγραφής (Kinect, PhaseSpace, XSens MTs). Η λύση που προτείνεται από τους Alexiadis κ.α. [80] και χρησιμοποιήθηκε για τις καταγραφές διασφάλισε την καταγραφή χωροχρονικά συσχετισμένων καρέ βάθους-υπερύθρων και πόζας Kinect . Ο συγχρονισμός μεταξύ Kinect και IMU επετεύχθη με τη χρήση κοινού χρόνου καταγραφής των δεδομένων. Τέλος, ορίζοντας το χρονικό βήμα της υψίσυχνης συχνότητας καταγραφής του PhaseSpace Impulse X2 ως χρόνο αναφοράς, ο συγχρονισμός των ground truth δεδομένων πραγματοποιήθηκε χειροκίνητα με την ανεύρεση της απόλυτης αρχικής χρονικής διαφοράς εκκίνησης

μεταξύ των ακολουθιών.

Σχήμα 3.5: *Φωτογραφία από καταγραφές 2 ατόμων με σύστημα PhaseSpace Impulse X2.*

Στατιστικά. Συνολικά καταγράφηκαν 10 υποκείμενα, 2 γυναίκες και 8 άνδρες, έχοντας τοποθετημένους 26 μη τυποποιημένους παθητικούς και 38 τυποποιημένους ενεργητικούς δείκτες και 9 αισθητήρες αδράνειας στο σώμα, καταγράφηκαν να εκτελούν 15 φυσικές ασκήσεις, οι οποίες παρουσιάζονται στον Πίνακα 3.1. Το πλήρες σύνολο δεδομένων περιέχει περισσότερα από 80,000 καρέ 3 προβολών βάθους-υπερύθρων και πόζας, τις εσωτερικές παραμέτρους των καμερών καθώς και τις παραμέτρους χωρικής συσχέτισης και τα αντίστοιχα αδρανειακά και ground truth δεδομένα.

Πίνακας 3.1: *Δεδομένα ανά υποκείμενο στο σύνολο δεδομένων DMC3D.*

Physical Exercise	**# of Repetitions**	**# of Frames**	**Type**
Walking on the spot	10–20	200–300	Free
Single arm raise	10–20	300–500	Bilateral
Elbow flexion	10–20	300–500	Bilateral
Knee flexion	10–20	300–500	Bilateral
Closing arms above head	6–12	200–300	Free
Side steps	6–12	300–500	Bilateral
Jumping jack	6–12	200–300	Free
Butt kicks left-right	6–12	300–500	Bilateral
Forward lunge left-right	4–10	300–500	Bilateral
Classic squat	6–12	200–300	Free
Side step + knee-elbow	6–12	300–500	Bilateral
Side reaches	6–12	300–500	Bilateral
Side jumps	6–12	300–500	Bilateral
Alternate side reaches	6–12	300–500	Bilateral
Kick-box kicking	2–6	200–300	Free

Υποσύνολο Δεδομένων DMC2.5D

Ένα δεύτερο σύνολο 2.5Δ δεδομένων (DMC2.5D) συλλέχθηκε με στόχο την εκπαίδευση, επαλήθευση και αξιολόγηση του μοντέλου που περιγράφεται στο Κεφ. 3.4.1. Εφαρμόζοντας την επεξεργασία των καταγεγραμμένων δεδομένων IR-D όπως περιγράφεται στο Κεφ. 3.1.2, δημιουργήθηκαν χρωματισμένα ζεύγη δεδομένων βάθους και 3Δ οπτικής ροής ανά προβολή.
Στατιστικά. Τα δείγματα επιλέχθηκαν τυχαία από 8 από τα 10 υποκείμενα, εξαιρώντας 2 από αυτά για την αξιολόγηση των μεθόδων. Πάνω από 300,000 2Δ συντεταγμένες δεικτών (2Δ προβολές των 3Δ θέσεων) επισημάνθηκαν σε πάνω από 25,000 τυχαία ζεύγη διαδοχικών δειγμάτων (του τρέχοντος και του προηγούμενου καρέ στην εικόνα), περιλαμβάνοντας ποικιλία στάσεων και κινήσεων. 20,000, 3,000 και 2,000 δείγματα χρησιμοποιήθηκαν για την εκπαίδευση, επαλήθευση και αξιολόγηση του μοντέλου, αντίστοιχα.

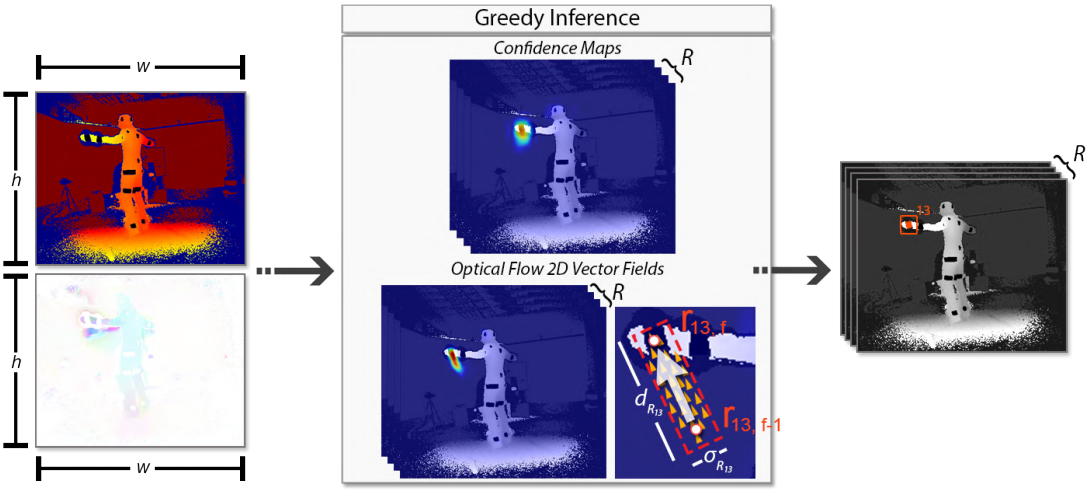
Σχήμα 3.6: *Επισήμανση δεικτών με χρήση εντοπισμού κηλίδων σε δυαδική μάσκα IR, οπτικοποιημένη όμως σε δεδομένα IR, $\mathcal{I}^v_{IR}$, για λόγους καλύτερης κατανόησης. Οι λανθασμένες εκτιμήσεις εμφανίζονταν όταν υπήρχε επικάλυψη μεταξύ των δεικτών όπως στις περιπτώσεις που εμφανίζονται στην εικόνα.*

Η επισήμανση πραγματοποιήθηκε με την ανάπτυξη και χρήση τεχνικών υπολογιστικής όρασης. Συγκεκριμένα, με εφαρμογή ανίχνευσης κηλίδων στη δυαδική εικόνα IR $\mathcal{I}^v_{IR_m}$ εξάγαμε τις 2Δ θέσεις των δεικτών, και στη συνέχεια, τις αντίστοιχες 3Δ θέσεις με αντιπροβολή της πληροφορίας βάθους. Τέλος, οι δείκτες επισημάνθηκαν συγκρίνοντας την ευκλείδεια 3Δ απόσταση ανά καρέ μεταξύ των εξαγόμενων 3Δ θέσεων και των 3Δ θέσεων των δεικτών των ground truth δεδομένων. Ωστόσο, η αυτοματοποιημένη επισήμανση ήταν λανθασμένη στις περιπτώσεις που οι περιοχές των μη τυποιημένων δεικτών είχαν επικάλυψη (Σχήμα 3.6) ή οι πόζες ήταν πολύπλοκες. Τα προβλήματα στις σύνθετες πόζες (μεγάλες κάμψης και αναδιπλώσεις σώματος) προέκυψαν λόγω της διαφοράς τοποθέτησης μεταξύ των διαφορετικού τύπου δεικτών, προκαλώντας σύγχυση στη διαδικασία επισήμανσης. Έτσι, απαιτήθηκε περαιτέρω χειροκίνητη επεξεργασία προκειμένου να διορθωθεί το σύνολο δεδομένων.

3.2 Εκτίμηση 2Δ θέσεων δεικτών

Η κύρια πρόκληση της προτεινόμενης μεθόδου είναι ο αποτελεσματικός εντοπισμός και αναγνώριση των δεικτών που τοποθετούνται στο σώμα του υποκειμένου. Μελετώντας την πρόσφατη βιβλιογραφία στον 2Δ εντοπισμό σε εικόνες χρώματος, η αποτελεσματικότητα των βαθιών νευρωνικών δικτύων σε σύνθετες εργασίες όπως η εκτίμηση της αρθρωτής 2Δ πόζας είναι αξιοσημείωτη και, ως εκ τούτου, θεωρήθηκε κατάλληλη προσέγγιση για την παρούσα

ερευνητική μέθοδο. Συγκεκριμένα, αναπτύχθηκε μια προσέγγιση βαθιάς μάθησης που επεκτείνει την αρχιτεκτονική CPM πολλαπλών σταδίων, προκειμένου να εντοπιστούν και να ταυτοποιηθούν οι δείκτες στο σώμα. Ένα πλήρως συνελικτικό δίκτυο πολλαπλών σταδίων εκπαιδεύτηκε να λειτουργεί με ενδιάμεσους χάρτες εμπιστοσύνης και διανυσματικά πεδία 2Δ οπτικής ροής, αντί για τα διανυσματικά πεδία μεταξύ 2Δ σημείων που ενώνουν δύο συνεχής αρθρώσεις στο σώμα Part Affinity Fields (PAF), όπως αρχικά παρουσιάστηκαν στο [1].

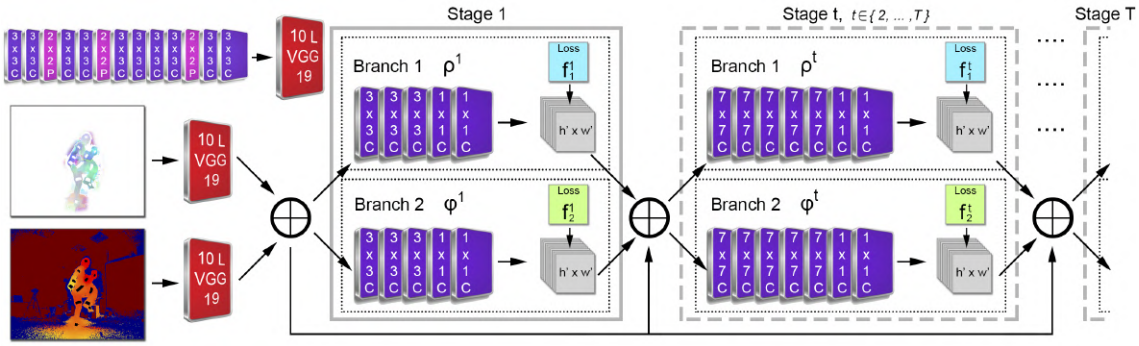


Σχήμα 3.7: *Εκτίμηση συνόλου 2Δ συντεταγμένων δεικτών από χάρτες εμπιστοσύνης και διανυσματικά πεδία 3Δ οπτικής ροής. Η 2Δ θέση για τον ανακλαστήρα R_{13} εκτιμάται λαμβάνοντας υπόψη τον προβλεπόμενο χάρτη θερμότητας και την οπτική ροή μεταξύ των προβλέψεων $\boldsymbol{r}_{13,f-1}$ και $\boldsymbol{r}_{13,f}$.*

Παρά τις ομοιότητες μεταξύ εκτίμησης 2Δ πόζας και του συνόλου των δεικτών, καθώς και τα δύο αποτελούν λύσεις προβλημάτων εντοπισμού 2Δ συντεταγμένων σε έγχρωμες εικόνες, υπάρχουν αξιοσημείωτες διαφορές. Οι Cao κ.ά. [1] αντιμετωπίζουν αποτελεσματικά το πρόβλημα της εκτίμησης 2Δ πόζας σε μεγάλα, “in-the-wild" σύνολα RGB δεδομένων [85], με αποτέλεσμα ακριβείς εκτιμήσεις σε μια τεράστια ποικιλία δεδομένων που παρουσιάζουν διάφορες πόζες σε διαφορετικά περιβάλλοντα, συνθήκες φωτισμού, ρουχισμό, κτλ. Αντίθετα, η εκτίμηση των δεικτών εφαρμόζεται σε πιο ‘ελεγχόμενες’ συνθήκες: i) τα δεδομένα βάθους βρίσκονται εντός ενός συγκεκριμένου και σχετικά σταθερού εύρους τιμών, ii) οι περιοχές των δεικτών έχουν σαφώς διακριτές τιμές στις εικόνες IR, iii) υπάρχει μόνο ένα υποκείμενο προς εντοπισμό και, στις περισσότερες περιπτώσεις, iv) το υποκείμενο ενεργεί στο κέντρο της σκηνής. Από την άλλη πλευρά, η εργασία εκτίμησης των δεικτών είναι πιο περίπλοκη όσον αφορά i) την εκτίμηση μεγαλύτερου αριθμού δεικτών σε σύγκριση με τις αρθρώσεις που ανιχνεύονται από τις CPM μεθόδους και ii) το γεγονός ότι επιβάλλουμε στο δίκτυο να εντοπίζει τους δείκτες μόνο όταν είναι ορατοί. Για παράδειγμα, οι δείκτες R_{15} και R_{14} είναι και οι δύο τοποθετημένοι στον δεξί ώμο, αλλά στην μπροστινή και την πίσω πλευρά του σώματος αντίστοιχα, ενώ η άρθρωση του δεξιού ώμου στις 2Δ πόζες για τις CPM μεθόδους είναι μοναδική και κοινή για όλες τις όψεις.

Η μέθοδος απεικονίζεται στο Σχήμα 3.7. Ένα ζεύγος εικόνων, το χρωματισμένο βάθος $\mathcal{I}^v_{CD}$ και ο αντίστοιχος χάρτης 3Δ οπτικής ροής $\mathcal{I}^v_F$, δίνονται ως είσοδος. Το μοντέλο εκτιμά ταυτόχρονα ένα σύνολο 2Δ χαρτών εμπιστοσύνης S των θέσεων των δεικτών και ένα σύνο-

λο 2Δ διανυσματικών πεδίων L - το τελευταίο αντιστοιχεί στην πεδία οπτικής ροής από το προηγούμενο πλαίσιο στο επόμενο, κωδικοποιώντας τη χρονική συσχέτιση μεταξύ διαδοχικών καρέ. Και τα δύο σύνολα περιέχουν $R = 26$ στοιχεία, ένα ανά δείκτη $R_i \in R_1, \ldots, R_{26}$, το σύνολο $S = (S_{R_1}, S_{R_2}, \ldots, S_{R_{26}})$, $S_{R_i} \in R^{w \times h}$, όπου w και h είναι το πλάτος και το ύψος των θερμικών χαρτών αντίστοιχα, και το σύνολο $L = (L_{R_1}, L_{R_2}, \ldots, L_{R_{26}})$, όπου $L_{R_i} \in R^{2 \times w \times h}$. Ένα τελικό βήμα εφαρμόζεται στους εξαγόμενους χάρτες θερμότητας και τα πεδία οπτικής ροής, με αποτέλεσμα την εκτίμηση των 2Δ θέσεων του συνόλου των δεικτών.



Σχήμα 3.8: *FCN Αρχιτεκτονική δύο ροών - δύο διακλαδώσεων πολλαπλών σταδίων. Το αποτέλεσμα κάθε σταδίου $t \in \{2, \ldots, T-1\}$ και το σύνολο χαρακτηριστικών $\boldsymbol{F}$ συγχωνεύονται και δίνονται ως είσοδος στο επόμενο στάδιο.*

3.2.1 Αρχιτεκτονική

Η αρχιτεκτονική που παρουσιάζεται στο Σχήμα 3.8, εισάγει ένα νέο σχήμα δύο ροών - δύο κλάδων πολλαπλών σταδίων βασισμένο στο CPM - η οποία δέχεται ως είσοδο 'χρωματισμένα' δεδομένα βάθους, $\mathcal{I}^v_{CD}$, και 3Δ δεδομένα οπτικής ροής, $\mathcal{I}^v_F$. Και οι δύο είσοδοι-ροές επεξεργάζονται χωριστά από ένα δίκτυο 10 επιπέδων (τα πρώτα 10 επίπεδα του VGG-19 [82]), δημιουργώντας δύο σύνολα χαρτών χαρακτηριστικών $\mathbf{F}_D$ και $\mathbf{F}_{OF}$, αντίστοιχα. Διαδοχικά, λαμβάνει χώρα ένα πρώιμο στάδιο συγχώνευσης, με τη συνένωση των χαρτών χαρακτηριστικών των δύο ροών, $\mathbf{F} = \mathbf{F}_{OF} \oplus \mathbf{F}_D$.

Στο πρώτο στάδιο $t = 1$, το σύνολο των συγχωνευμένων χαρτών χαρακτηριστικών $\mathbf{F}$ δίνεται και στους δύο κλάδους που παράγουν χάρτες εμπιστοσύνης, $\mathbf{S}^t = \rho^t(\mathbf{F})$, και 2Δ διανυσματικά πεδία, $\mathbf{L}^t = \varphi^t(\mathbf{F})$, όπου ρ^t και φ^t αποτελούν την έξοδο του κάθε FCN κλάδου, αντίστοιχα. Για όλα τα επόμενα $T-1$ στάδια, όπου T δηλώνει το συνολικό αριθμό σταδίων, οι προβλέψεις και από τους δύο κλάδους στο προηγούμενο στάδιο, μαζί με το σύνολο χαρακτηριστικών $\mathbf{F}$, συγχωνεύονται και χρησιμοποιούνται για την παραγωγή βελτιωμένων προβλέψεων με βάση τις εξισώσεις παρακάτω:

$$\begin{aligned} \mathbf{S}^t &= \rho^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}), t \in \{2, \ldots, T\} \\ \mathbf{L}^t &= \varphi^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}), t \in \{2, \ldots, T\} \end{aligned} \quad , \tag{3.1}$$

όπου ο αριθμός των σταδίων T είναι ίσος με 6, που καθορίστηκε πειραματικά με την αξιολόγηση των αποτελεσμάτων στο σύνολο δεδομένων επαλήθευσης για $T = 3$ και $T = 6$ στάδια, όπως προτείνεται στις [1, 46], αντίστοιχα. Στο τέλος κάθε σταδίου, εφαρμόζονται δύο συναρτήσεις απωλειών $\mathcal{L}^t_{\mathbf{S}}$ και $\mathcal{L}^t_{\mathbf{L}}$, για την επίβλεψη και καθοδήγηση του δικτύου. Στο στάδιο

t, για το σημείο $\mathbf{p} = (x, y)$, $\mathbf{p} \in \mathbb{R}^2$, οι συναρτήσεις απωλειών δίνονται ως εξής:

$$\begin{aligned} \mathcal{L}_{\mathbf{S}}^t &= \sum_{r=1}^{R} \sum_{\mathbf{p}} ||\mathbf{S}_r^t(\mathbf{p}) - \mathbf{S}_r^*(\mathbf{p})||_2^2 \\ \mathcal{L}_{\mathbf{L}}^t &= \sum_{r=1}^{R} \sum_{\mathbf{p}} ||\mathbf{L}_r^t(\mathbf{p}) - \mathbf{L}_r^*(\mathbf{p})||_2^2 \end{aligned} \tag{3.2}$$

όπου $\mathbf{S}_{R_i}^*$ και $\mathbf{L}_{R_i}^*$ αντιπροσωπεύουν τους ground truth χάρτες θερμότητας και τα διανυσματικά πεδία, αντίστοιχα. Με την ενδιάμεση εποπτεία σε κάθε στάδιο t, αποφεύγουμε την πιθανότητα εξάλειψης του ρυθμού ανατροφοδότησης λόγω του βάθους του ενιαίου δικτύου κατά την εκπαίδευση. Η συνολική συνάρτηση απωλειών $\mathcal{L}$ του δικτύου δίνεται από:

$$\mathcal{L} = \sum_{t=1}^{T} (\mathcal{L}_{\mathbf{S}}^t + \mathcal{L}_{\mathbf{L}}^t) \tag{3.3}$$

3.2.2 Χάρτες θερμότητας και πεδία οπτικής ροής

Χάρτες θερμότητας: Κάθε χάρτης θερμότητας είναι μια 2Δ αναπαράσταση της πιθανότητας ανά εικονοστοιχείο της εύρεσης της προβολής ενός δείκτη πάνω σε ένα χάρτη. Η προτεινόμενη μέθοδος ερευνά την καταγραφή κίνησης ενός ατόμου, επομένως, σε κάθε χάρτη θα πρέπει να υπάρχει μια μοναδική μέγιστη τιμή ανά δείκτη. Έστω $x_{R_i,f} \in \mathbb{R}^2$ η ground truth 2Δ θέση της προβολής του δείκτη R_i στην εικόνα, στο καρέ f. Για κάθε 2Δ θέση $\mathbf{p} \in \mathbb{R}^2$, ο ground truth χάρτης θερμότητας $\mathbf{S}_{R_i,f}^*$ δίνεται από:

$$\mathbf{S}_{R_i,f}^*(\mathbf{p}) = \exp(-\frac{||\mathbf{p} - x_{R_i,f}||_2^2}{\sigma^2}) \qquad , \tag{3.4}$$

όπου το σ ορίζει την κατανομή της 2Δ πιθανότητας. Για τον υπολογισμό των 2Δ συντεταγμένων των δεικτών, εφαρμόζεται εντοπισμός των συντεταγμένων μέγιστης τιμής στους εκτιμώμενους χάρτες θερμότητας, θέτοντας παράλληλα τη μέγιστη τιμή σε τιμή εμπιστοσύνης της πρόβλεψης, $E_{R_i,f}^{\mathbf{S}}$.

Πεδία οπτικής ροής: Στην παρόν κεφάλαιο, η αναπαράσταση χαρακτηριστικών 2Δ διανυσματικών πεδίων που προτείνεται στην [1], χρησιμοποιείται με στόχο τη χρονική συσχέτιση. Διατηρώντας τόσο την πληροφορία της 2Δ θέσης όσο και την κατεύθυνση σε μια περιοχή της εικόνας, ορίζεται ένα 2Δ διανυσματικό πεδίο για κάθε δείκτη συνδέοντας τις θέσεις του μεταξύ των καρέ $f-1$ και f. Έστω $x_{R_i,f-1}, x_{R_i,f} \in \mathbb{R}^2$ είναι οι πραγματικές 2Δ θέσεις του δείκτη R_i στα καρέ $f-1$ και f, αντίστοιχα. Η δημιουργία του χάρτη θερμότητας για κάθε 2Δ θέση $\mathbf{p} \in \mathbb{R}^2$ του $\mathbf{L}_{R_i,f}^*$ δίνεται από:

$$\begin{aligned} \mathbf{L}_{R_i,f}^*(\mathbf{p}) &= \begin{cases} \mathbf{v}, & if\ \mathbf{p}\ on\ optical\ flow\ field \\ 0, & otherwise \end{cases} \\ \mathbf{v} &= \frac{x_{R_i,f} - x_{R_i,f-1}}{||x_{R_i,f} - x_{R_i,f-1}||_2} \end{aligned} \tag{3.5}$$

Το σύνολο των σημείων που ανήκουν στο πεδίο οπτικής ροής περιλαμβάνει τα σημεία εντός ενός κατωφλίου ευκλείδειας απόστασης από το ευθύγραμμο τμήμα μεταξύ των 2Δ θέσεων του

δείκτη, το οποίο δίνεται από:

$$\begin{aligned} 0 \leq \mathbf{v} \cdot (\mathbf{p} - x_{R_i,f-1}) \leq d_{R_i} \\ |\mathbf{v}_\perp \cdot (\mathbf{p} - x_{R_i,f-1})| \leq \sigma_{R_i} \end{aligned} \quad , \tag{3.6}$$

όπου σ_{R_i} είναι το πλάτος του πεδίου σε εικονοστοιχεία, d_{R_i} είναι η ευκλείδεια απόσταση των 2Δ θέσεων του δείκτη R_i μεταξύ διαδοχικών καρέ στην εικόνα, $d_{R_i} = ||x_{R_i,t} - x_{R_i,t-1}||_2$, και $\mathbf{v}_\perp$ είναι ένα διάνυσμα κάθετο στο $\mathbf{v}$. Ως παράδειγμα, το πεδίο οπτικής ροής 2Δ για τον δείκτη R_{13} απεικονίζεται στο Σχήμα 3.7.

Κατά τη διάρκεια της εκτίμησης των δεικτών, η οπτική ροή, που κωδικοποιεί τις χρονικές συσχετίσεις, μετράται με τον υπολογισμό του γραμμικού ολοκληρώματος επί του αντίστοιχου πεδίου οπτικής ροής κατά μήκος του 2Δ διανύσματος που συνδέει τις θέσεις των πιθανών δεικτών μεταξύ δύο διαδοχικών καρέ. Έστω $\mathbf{r}_{R_i,f}$ και $\mathbf{r}_{R_i,f-1}$ οι εκτιμώμενες θέσεις για τον δείκτη R_i στο τρέχον καρέ f και στο προηγούμενο $f-1$, αντίστοιχα. Το προβλεπόμενο πεδίο οπτικής ροής $\mathbf{L}_{R_i,f}$ λαμβάνεται δειγματοληπτικά κατά μήκος του ευθύγραμμου τμήματος για να μετρηθεί η εμπιστοσύνη χρονικής συσχέτισης μεταξύ των προβλεπόμενων θέσεων των δεικτών στο χρόνο κατά:

$$E^{\mathbf{L}}_{R_i,f} = \int_0^1 \mathbf{L}_{R_i,f}(\mathbf{p}(u)) \cdot \frac{\mathbf{r}_{R_i,f} - \mathbf{r}_{R_i,f-1}}{||\mathbf{r}_{R_i,f} - \mathbf{r}_{R_i,f-1}||_2} du \quad , \tag{3.7}$$

όπου $\mathbf{p}(u)$ παρεμβάλλει τις θέσεις των ανακλαστήρων $\mathbf{r}_{R_i,f}$ και $\mathbf{r}_{R_i,f-1}$ μεταξύ διαδοχικών καρέ, όπως δίνεται από:

$$\mathbf{p}(u) = (1-u) \cdot \mathbf{r}_{R_i,f-1} + u \cdot \mathbf{r}_{R_i,f} \tag{3.8}$$

Με άλλα λόγια, το ολοκλήρωμα προσεγγίζεται με τη δειγματοληψία και το άθροισμα ομοιόμορφα κατανεμημένων τιμών του u.

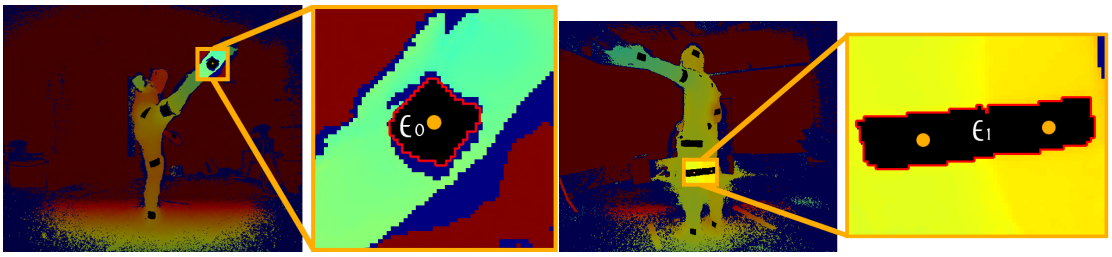
Αυθαίρετο συμπέρασμα: Τέλος, εισάγεται ένα αυθαίρετο βήμα εξαγωγής των τελικών προβλέψεων, λαμβάνοντας υπόψη τις χρονικές συσχετίσεις μεταξύ διαδοχικών εκτιμήσεων των 2Δ θέσεων των δεικτών. Οι τιμές εμπιστοσύνης $E^{\mathbf{S}}_{R_i,f}$ και $E^{\mathbf{L}}_{R_i,f}$ που δίνονται από τους χάρτες εμπιστοσύνης και τα 2Δ διανύσματα οπτικής ροής αντίστοιχα συνυπολογίζονται ώστε να προκύψει η ολική, τελική τιμή εμπιστοσύνης $E_{R_i,f}$ για κάθε εκτίμηση συντεταγμένων ανά δείκτη. Αναλυτικότερα, η κύρια συνιστώσα της $E_{R_i,f}$ είναι η $E^{\mathbf{S}}_{R_i,f}$, ωστόσο, σταθμίζουμε την εμπιστοσύνη $E^{\mathbf{L}}_{R_i,f}$ με βάση έναν παράγοντα $w^{\mathbf{L}}_{R_i,f} = (1 - E^{\mathbf{S}}_{R_i,f})$ που αυξάνεται όταν το $E^{\mathbf{S}}_{R_i,f}$ μειώνεται ως εξής:

$$E_{R_i,f} = E^{\mathbf{S}}_{R_i,f} + w^{\mathbf{L}}_{R_i,f} \cdot E^{\mathbf{L}}_{R_i,f} \tag{3.9}$$

Με αυτόν τον τρόπο, όταν η πρόβλεψη ενός χάρτη εμπιστοσύνης οδηγεί σε εκτιμήσεις χαμηλής εμπιστοσύνης $E^{\mathbf{S}}_{R_i,f}$, η συνολική εμπιστοσύνη $E_{R_i,f}$ επηρεάζεται πιο έντονα από την εμπιστοσύνη της οπτικής ροής. Το τελικό αποτέλεσμα της εκτίμησης του συνόλου των δεικτών δίνεται με την εφαρμογή ευριστικού εντοπισμού μέγιστης παραμέτρου (standard coordinate decoding, βλ. Κεφ. 6) στις προβλέψεις των δεικτών με βάση τη συνολική εμπιστοσύνη $E_{R_i,f}$.

3.2.3 2Δ σε 3Δ αντιπροβολή δεδομένων

Με δεδομένες τις 2Δ θέσεις των δεικτών στους χάρτες βάθους $\mathcal{I}_D^v$, εφαρμόζεται μια τεχνική 3Δ εκτίμησης και αντιπροβολής για την ακριβή εξαγωγή των αντίστοιχων 3Δ θέσεων. Λαμβάνοντας υπόψη ότι οι θέσεις των δεικτών δίνονται αποκλειστικά όταν είναι πλήρως ορατοί, η εκτίμηση ενός δείκτη θεωρείται έγκυρη μόνο εάν ανήκει σε μια περιοχή με περισσότερα από έναν αριθμό b_{min} μαύρων εικονοστοιχείων στο $\mathcal{I}_{CD}^v$, διαφορετικά απορρίπτεται. Το ελάχιστο αποδεκτό μέγεθος εικονοστοιχείων για μια περιοχή ορίστηκε $b_{min} = 5$, καθώς το ίδιο μέγεθος περιοχής χρησιμοποιήθηκε για τη σήμανση του συνόλου δεδομένων DMC.

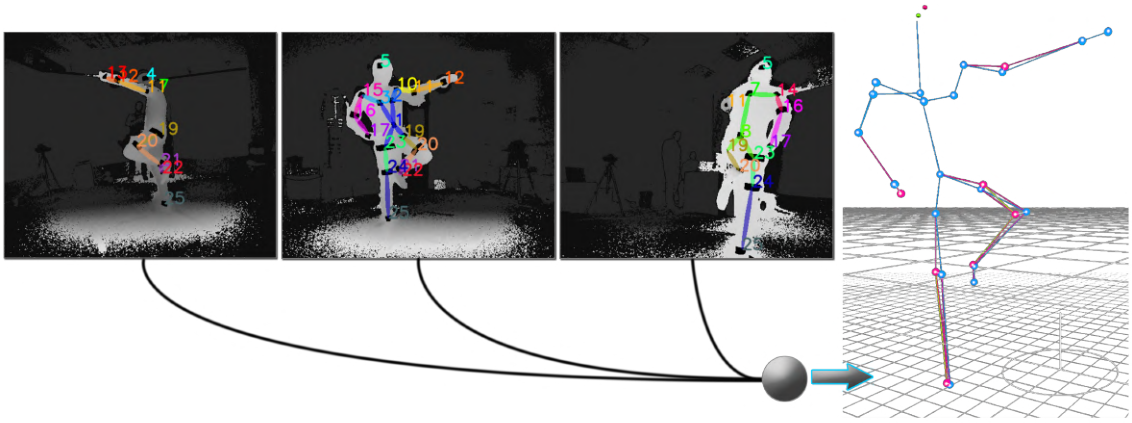


Σχήμα 3.9: *Εντοπισμός περιγράμματος (κόκκινα εικονοστοιχεία) των δεικτών για την εκτίμηση του βάθους και της 3Δ θέσης.* ***(α)*** *Περίγραμμα περιοχής ενός δείκτη,* $\mathcal{E}_0$. ***(β)*** *Περίγραμμα συγχωνευμένης περιοχής πολλαπλών δεικτών,* $\mathcal{E}_1$.

Στο Σχήμα 3.9, παρουσιάζονται δύο από τις πιθανές περιπτώσεις σε σχέση με την αντιπροβολή των δεικτών. Στην πρώτη περίπτωση (Σχήμα 3.9α), την απλή και πιο συχνή, $\mathcal{E}_0 \in \mathcal{I}_{CD}^v$ είναι η ανιχνευόμενη περιοχή για έναν δείκτη $R_i \in \{R_1, \ldots, R_{26}\}$. Ανακτώντας ένα σύνολο εικονοστοιχείων $\mathcal{P}_{R_i}^v$ του περιγράμματος της περιοχής $\mathcal{E}_0$ στην εικόνα $\mathcal{I}_{CD}^v$ (κόκκινα εικονοστοιχεία στο Σχήμα 3.9α) και αντιστοιχίζοντας τα εικονοστοιχεία αυτά στο $\mathcal{I}_D^v$, λαμβάνονται οι αντίστοιχες τιμές βάθους του $\mathcal{P}_{R_i}^v$. Χρησιμοποιώντας μόνο τις μη μηδενικές τιμές βάθους του $\mathcal{P}_{R_i}^v$, η διάμεση τιμή d_{R_i} θεωρείται η τιμή βάθους του δείκτη R_i από τον αισθητήρα v, ενώ η 2Δ συντεταγμένες (x, y) θεωρούμε το κέντρο του περιγράμματος .

Η δεύτερη περίπτωση, αν και όχι τόσο συχνή, απεικονίζεται στο σχήμα 3.9β, όπου δύο ή περισσότεροι δείκτες ανήκουν στην ίδια περιοχή $\mathcal{E}_1$. Εξετάζοντας την επικάλυψη μεταξύ των περιοχών των ανακλαστήρων, δηλαδή όταν $n > 1$ δείκτες μοιράζονται την ίδια 'μαύρη' περιοχή, τα εικονοστοιχεία του περιγράμματος ομαδοποιούνται σε n ομάδες, με βάση τις 2Δ εντοπίσεις συντεταγμένες των εικονοστοιχείων και τις τιμές βάθους, και στη συνέχεια αντιστοιχίζονται στους αντίστοιχους δείκτες. Στη συνέχεια, προσδιορίζεται το d_{R_i} για κάθε δείκτη R_i με βάση το αντίστοιχο σύνολο των ομαδοποιημένων εικονοστοιχείων. Μετά την αντιστοίχιση μεταξύ δεικτών και περιοχών στην εικόνα, το 2Δ κέντρο βάρους $\mathbf{p}_{R_i}$ κάθε περιοχής $\mathcal{E}_{R_i}$ αντιστοιχίζεται χωρικά σε 3Δ συντεταγμένες χρησιμοποιώντας την τιμή βάθους d_{R_i} και τις εγγενείς παραμέτρους του αντίστοιχου αισθητήρα IR-D, δίνοντας την 3Δ θέση $\mathbf{P}_{R_i}^v$ του δείκτη R_i από το σημείο θέασης v.

3.2.4 Συγχώνευση 3Δ θέσεων δεικτών



Σχήμα 3.10: *Πολλαπλές 2Δ εκτιμήσεις συνόλων ανακλαστήρων (**αριστερά**) απεικονίζονται χωρικά σε ένα συνολικό 3Δ καρτεσιανό σύστημα συντεταγμένων, με αποτέλεσμα την εξαγωγή 3Δ οπτικών δεδομένων (**δεξιά**).*

Χρησιμοποιώντας τις παραμέτρους χωρικής συσχέτισης κάθε αισθητήρα, οι εξαγόμενες 3Δ θέσεις μεταφέρονται σε ένα ενιαίο καρτεσιανό σύστημα, όπως φαίνεται στο Σχήμα 3.10. Για τους τετραγωνικούς δείκτες, οι οποίοι είναι ορατοί από μία μόνο πλευρά, η ίδια οπισθοανακλαστική περιοχή καταγράφεται από όλους τους αισθητήρες και, ως εκ τούτου, η 3Δ αντιπροβολή αποδίδει μόνο ελαφρώς διαφορετικές εκτιμήσεις. Έτσι, τα 3Δ σημεία $\mathbf{P}^{v}_{R_i}$ για έναν τετραγωνικό δείκτη R_i από όλες τις προβολές v συγχωνεύονται σε ένα ενιαίο 3Δ σημείο $\mathbf{P}_{R_i}$, λαμβάνοντας υπόψη την τιμή εμπιστοσύνης $E^{v}_{R_i}$ της εκτίμησης του δείκτη R_i ανά προβολή v. Το 3Δ σημείο $\mathbf{P}_{R_i}$ υπολογίζεται ως το σταθμισμένο κέντρο βάρους των 3Δ δεικτών από όλες τις προβολές:

$$w^{v}_{R_i} = E^{v}_{R_i} / \sum_{v=1}^{N} E^{v}_{R_i} \qquad \mathbf{P}_{R_i} = \frac{1}{N} \sum_{r=1}^{N} w_{R_i} \mathbf{P}^{v}_{R_i} \tag{3.10}$$

Για τους δείκτες-ιμάντες, δεδομένου ότι διαφορετικά τμήματα των δεικτών είναι ορατά ανά αισθητήρα, τα 3Δ σημεία εκτιμώνται σε διαφορετικές 3Δ θέσεις γύρω από το τμήμα του άκρου στο οποίο τοποθετείται ο ιμάντας. Σε αυτή την περίπτωση, το επιθυμητό 3Δ σημείο βρίσκεται κατά προσέγγιση στο κέντρο του 'δακτυλίου' όπου βρίσκονται αυτά τα 3Δ σημεία, προσαρτώνται και εξάγονται ως εξής: έστω $\mathbf{n}^{v_i}_{R_i}$ και $\mathbf{n}^{v_j}_{R_i}$ τα 3Δ μοναδιαία διανύσματα των περιοχών του δείκτη για τις όψεις v_i και v_j, $v_i, v_j \in \{v_1, \ldots v_N\}$, αντίστοιχα. Αυτά τα διανύσματα ορίζονται ως τα μοναδιαία διανύσματα των 3Δ σημείων που ορίζονται από την αντιπροβολή των αντίστοιχων συνόλων εικονοστοιχείων $\mathcal{P}^{v_i}_{R_i}$ και $\mathcal{P}^{v_j}_{R_i}$ σε 3Δ καρτεσιανές συντεταγμένες. Τα πλησιέστερα 3Δ σημεία $\mathbf{P}^{v_{i,j}}_{R_i}$ και $\mathbf{P}^{v_{j,i}}_{R_i}$ μεταξύ των διανυσματικών ευθυών των $\mathbf{n}^{v_i}_{R_i}$ και $\mathbf{n}^{v_j}_{R_i}$, για ένα ζεύγος όψεων $v_i - v_j$ δίνονται με την εφαρμογή των εξισώσεων:

$$\begin{aligned} b &= \mathbf{n}_{R_i}^{v_i} \bullet \mathbf{n}_{R_i}^{v_j} & d &= 1 - b^2 \\ c &= \mathbf{n}_{R_i}^{v_i} \bullet (\mathbf{P}_{R_i}^{v_i} - \mathbf{P}_{R_i}^{v_j}) & f &= \mathbf{n}_{R_i}^{v_j} \bullet (\mathbf{P}_{R_i}^{v_i} - \mathbf{P}_{R_i}^{v_j}) \\ s &= (b \cdot f - c)/d & t &= (f - c \cdot b)/d \\ \mathbf{P}_{R_i}^{v_{i,j}} &= \mathbf{P}_{R_i}^{v_i} + s \cdot \mathbf{n}_{R_i}^{v_i} & \mathbf{P}_{R_i}^{v_{j,i}} &= \mathbf{P}_{R_i}^{v_j} + t \cdot \mathbf{n}_{R_i}^{v_j} \end{aligned} \tag{3.11}$$

Κατόπιν, εφαρμόζοντας την εξίσωση (3.10) για όλα τα $N_{R_i} = 2 \cdot m_{R_i}$ εξαγόμενα 3Δ σημεία, όπου m είναι ο συνολικός αριθμός των ζευγών, εκτιμάται η 3Δ θέση του ανακλαστικού ιμάντα R_i στο κέντρο του μέρους του σώματος. Ακόμη και όταν μόνο ένας αισθητήρας καταγράφει έναν δείκτη-ιμάντα, εάν η δομή του σώματος και η τοποθέτηση των δεικτών έχει ήδη υπολογιστεί, η 3Δ θέση μπορεί να εκτιμηθεί χρησιμοποιώντας το μοναδιαίο διάνυσμα και την ακτίνα του μέρους του σώματος (άκρου).

Το πλήρες σύνολο των 3Δ θέσεων των δεικτών ανά καρέ πολλαπλών προβολών f, δηλαδή τα εξαγόμενα 3Δ οπτικά δεδομένα των δεικτών, συμβολίζεται ως $\mathbf{P}_f$, ενώ η συνολική τιμή εμπιστοσύνης $C_f^{R_i}$ για κάθε δείκτη R_i θεωρείται ως η μέση τιμή της εμπιστοσύνης εκτίμησης $E_{R_i}^v$ για κάθε προβολή $v-$, $v \in \{1, \ldots .N\}$, $C_f^{R_i} = \frac{1}{N} \sum_{v=1}^{N} E_{R_i}^v$. Για να βελτιωθεί η ποιότητα και η σταθερότητα των εκτιμήσεων των 3Δ σημείων στο χρόνο, εφαρμόζεται φίλτρο Kalman [86].

3.3 Καταγραφή Κίνησης

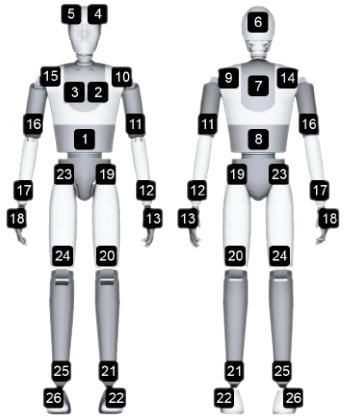
Το τελικό στάδιο της προτεινόμενης μεθόδου στοχεύει στην 3Δ καταγραφή ανθρώπινης κίνησης με την αξιοποίηση 3Δ οπτικών δεδομένων δεικτών. Σε αυτό το σημείο, τα εξαγόμενα 3Δ δεδομένα $\mathbf{P}_f$ αντιστοιχούν σε μια αναπαράσταση πόζας που αποτελείται από κοινές 3Δ θέσεις και περιστροφές. Ομοίως με τις ερευνητικές δουλειές [87, 88], προτείνεται μια τεχνική προσαρμογής σωματοδομής η οποία προσαρμόζει ένα αρθρωτό πρότυπο μοντέλο στην πραγματική δομή του σώματος του υποκειμένου. Στη συνέχεια, το αρθρωτό μοντέλο κινείται με την εφαρμογή πρόσθιας κινηματικής.

Η προτεινόμενη δομή αρθρωτού σώματος αποτελείται από 20 αρθρώσεις, $j \in J$ $D_{DoF} = 40$ βαθμών ελευθερίας. Αποτελείται από $L_i \in L, i \in \{0, \ldots, 6\}$ ιεραρχικά επίπεδα αρθρώσεων όπου οι δείκτες αντιστοιχούν σε συγκεκριμένα μέρη του σώματος. Έτσι, ένα υποσύνολο δεικτών $\mathcal{R}_{S_j} \subset \{R_1, \ldots, R_{26}\}, j \in J$, μετακινεί την άρθρωση του σώματος $j \in J$ σύμφωνα με την ιεραρχία των αρθρώσεων του σώματος. Η αντιστοιχία μεταξύ των αρθρώσεων και των δεικτών απεικονίζεται στον Πίνακα 3.2, ενώ η αντιστοίχιση δεικτών-μερών του σώματος απεικονίζεται στο Σχήμα 3.11.

Αρχικά, το πρότυπο μοντέλο προσαρμόζεται χωρίς υψηλή ακρίβεια με βάση την πρώτη ακολουθία 3Δ οπτικών δεδομένων δεικτών $\mathbf{P}_f$. Στη συνέχεια, δεδομένων των 3Δ θέσεων $\mathbf{P}_j$, $j \in J$ ανά καρέ, εκτελείται μια διαδικασία βελτιστοποίησης ανά οστό για την ακριβή προσαρμογή των τμημάτων του προτύπου στα πραγματικά μήκη του σώματος του υποκειμένου. Το βήμα αυτό εκτελείται διαδοχικά στα οστά που ακολουθούν τα επίπεδα ιεραρχίας των αρθρώσεων, από L_0 έως L_6. Πιο συγκεκριμένα, μετά την αρχική προσαρμογή του προτύπου στα 3Δ οπτικά δεδομένα, δίνεται η 3Δ θέση του γοφού $\mathbf{P}_{HIPS}$ με ιεραρχικό επίπεδο L_0, επιτρέποντας τη διαδοχική εκτίμηση των υπόλοιπων μηκών των οστών. Με βάση την υπόθεση ότι τα μήκη

Πίνακας 3.2: *Αρθρώσεις πρότυπου μοντέλου, ιεραρχικό επίπεδο, βαθμοί ελευθερίας και αντιστοιχία με τα υποσύνολα δεικτών.*

Body Joint, $j \in J$	Level L_i	DoFs	Subset $\mathcal{R}_{S_j}$	S_j (#)	Retro-reflectors
Hips	L_0	6	$\mathcal{R}_{S_0}$	4	$\{R_1, R_8, R_{19}, R_{23}\}$
Spinebase	L_1	3	$\mathcal{R}_{S_1}$	2	$\{R_1, R_8\}$
Neck	L_2	3	$\mathcal{R}_{S_2}$	3	$\{R_2, R_3, R_7\}$
Head	L_3	-	$\mathcal{R}_{S_3}$	3	$\{R_4, R_5, R_6\}$
Left Shoulder	L_3	3	$\mathcal{R}_{S_4}$	2	$\{R_9, R_{10}\}$
Left Elbow	L_4	1	$\mathcal{R}_{S_5}$	1	$\{R_{11}\}$
Left Wrist	L_5	3	$\mathcal{R}_{S_6}$	1	$\{R_{12}\}$
Left Hand	L_6	-	$\mathcal{R}_{S_7}$	1	$\{R_{13}\}$
Right Shoulder	L_3	3	$\mathcal{R}_{S_8}$	2	$\{R_{14}, R_{15}\}$
Right Elbow	L_4	1	$\mathcal{R}_{S_9}$	1	$\{R_{16}\}$
Right Wrist	L_5	3	$\mathcal{R}_{S_{10}}$	1	$\{R_{17}\}$
Right Hand	L_6	-	$\mathcal{R}_{S_{11}}$	1	$\{R_{18}\}$
Left Hip	L_1	3	$\mathcal{R}_{S_{12}}$	1	$\{R_{19}\}$
Left Knee	L_2	1	$\mathcal{R}_{S_{13}}$	1	$\{R_{20}\}$
Left Ankle	L_3	3	$\mathcal{R}_{S_{14}}$	1	$\{R_{21}\}$
Left Foot	L_4	-	$\mathcal{R}_{S_{15}}$	1	$\{R_{22}\}$
Right Hip	L_1	3	$\mathcal{R}_{S_{16}}$	1	$\{R_{23}\}$
Right Knee	L_2	1	$\mathcal{R}_{S_{17}}$	1	$\{R_{24}\}$
Right Ankle	L_3	3	$\mathcal{R}_{S_{18}}$	1	$\{R_{25}\}$
Right Foot	L_4	-	$\mathcal{R}_{S_{19}}$	1	$\{R_{26}\}$



Σχήμα 3.11: *Αντιστοιχία μεταξύ δεικτών και μερών του σώματος.*

των οστών είναι σταθερά (άκαμπτα οστά) και χρησιμοποιώντας αποκλειστικά οπτικά δεδομένα $\mathbf{P}_f$ υψηλής αξιοπιστίας ($C_f^{R_i} > 0.6$, πειραματικά καθορισμένα), εφαρμόζεται ένα βήμα δημιουργίας τυχαίων σωματιδίων ανά επίπεδο σε $L-1$ φάσεις. Πιο συγκεκριμένα, ένα σύνολο $\mathcal{G}_j$ σωματιδίων με $G = 500$ (πειραματικά καθορισμένο) δημιουργείται γύρω από τη θέση της j-άρθρωσης $\mathbf{P}_j$ που δίνεται από τη χωρικά συσχετισμένη τοποθέτηση του προτύπου με χρήση του $\mathbf{P}_f$. Μετά τη δημιουργία των σωματιδίων, τα σωματίδια $\mathcal{G}_j$ ακολουθούν τα 3Δ δεδομένα με χρήση πρόσθιας κινηματικής. Το σωματίδιο $g_j \in \mathcal{G}_j$ που κινείται πιο άκαμπτα μεταξύ $\mathbf{P}_{j_l}$ και $\mathbf{P}_{j_{l+1}}$, θεωρείται ως το πλησιέστερο στην πραγματική σχετική θέση της άρθρωσης j_{l+1}. Η αντικειμενική συνάρτηση που εκτιμά αυτό το σωματίδιο δίνεται από:

$$\begin{aligned} D_0 &= ||\mathbf{P}_{j_l,0} - \mathbf{P}_{g_{j_{l+1}},0}||_2 + ||\mathbf{P}_{j_{l+1,0}} - \mathbf{P}_{g_{j_{l+1}},0}||_2 \\ \arg\min_{g_{j_{l+1}}} D(g_{j_{l+1}}) &= \frac{1}{F}\sum_{f=1}^{F}(D_0 - ||\mathbf{P}_{j_l,f} - \mathbf{P}_{g_{j_{l+1}},f}||_2 + ||\mathbf{P}_{j_{l+1},f} - \mathbf{P}_{g_{j_{l+1}},f}||_2) \end{aligned}, \quad (3.12)$$

όπου το D_0 δηλώνει το άθροισμα των 3Δ ευκλείδειων αποστάσεων στο αρχικό καρέ του $(l) - (l+1)$ επιπέδου προσαρμογής μεταξύ της 3Δ θέσης του σωματιδίου $g_{j_{l+1}}$ και των αρθρώσεων j_l και j_{l+1}, όπως προκύπτουν από την τελευταία προσαρμογή του προτύπου, και το $\mathbf{P}_{g_{j_{l+1}},f}$ δηλώνει την 3Δ θέση του σωματιδίου $g_{j_{l+1}}$ στο καρέ $f \in \{1,\ldots,F\}$, όπου F είναι ο συνολικός αριθμός των αρχικών καρέ που χρησιμοποιούνται για την προσαρμογή ενός μέρους του σώματος επιπέδου $l+1$. Με παρόμοιο τρόπο προσαρμόζονται οι αρθρώσεις του επόμενου επιπέδου και τα αντίστοιχα οστά. Οι γωνιακές κινήσεις των μερών του σώματος, ιδίως οι κάμψεις του αγκώνα και του γόνατος, επιτρέπουν ταχύτερη και αποτελεσματικότερη σύγκλιση της διαδικασίας προσαρμογής ανά οστό.

3.4 Πειράματα

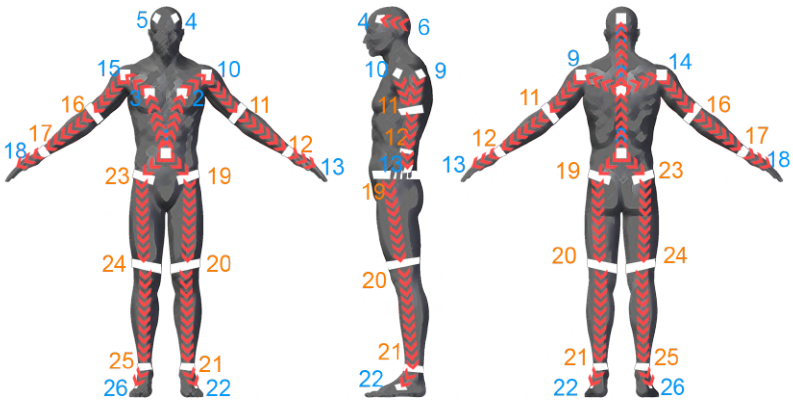
Για την αξιολόγηση της προτεινόμενης μεθόδου, διεξήχθησαν δύο τύποι πειραμάτων, τα οποία παρουσιάζονται και συζητούνται στην παρούσα ενότητα. Ο πρώτος τύπος αφορά την αξιολόγηση της εκτίμησης των 2Δ συντεταγμένων των δεικτών στο σύνολο δεδομένων DMC2.5D, ενώ, στο δεύτερο, τα αποτελέσματα της καταγραφής κίνησης συγκρίνονται με εναλλακτικές μεθόδους καταγραφής σε σχέση με τα ground truth δεδομένα του συνόλου DMC3D. Αρχικά, αξιολογούνται διαφορετικές αρχιτεκτονικές, αναδεικνύοντας την υπεροχή του προτεινόμενου μοντέλου. Η ακριβής εκτίμηση του συνόλου των 2Δ συντεταγμένων των δεικτών εξαλείφει τα σφάλματα στην εξαγωγή 3Δ δεδομένων και, κατά συνέπεια, στο τελικό αποτέλεσμα καταγραφής κίνησης. Έτσι, στη συνέχεια, εφαρμόζοντας το προτεινόμενο μοντέλο στο DMC3D σύνολο δεδομένων αξιολόγησης, εξάγονται 3Δ δεδομένα από ακολουθίες πολλαπλών προβολών και χρησιμοποιούνται αντίστοιχα για την καταγραφή κίνησης.

3.4.1 Πειραματικά Αποτελέσματα 2Δ Μεθόδου

Πειραματικό Πλαίσιο

Το μοντέλο που παρουσιάστηκε προσεγγίζει την εκτίμηση των 2Δ δεικτών, ένα παρόμοιο, αλλά ταυτόχρονα διαφορετικό πρόβλημα όρασης υπολογιστών σε σύγκριση με την εκτίμηση της 2Δ πόζας. Με στόχο την αξιολόγηση της παρούσας προσέγγισης και της εισαγόμενης επέκτασης σε σχέση με τις χρονικές συσχετίσεις μεταξύ των 2Δ σημείων των δεικτών από καρέ σε καρέ, οι υφιστάμενες μέθοδοι εκτίμησης 2Δ πόζας προσαρμόζονται κατάλληλα και εκπαιδεύονται για την επίλυση της παρούσας πρόκλησης.

Αναλυτικότερα, οι μέθοδοι των Wei *et al.* [46] και Cao *et al.* [1], που περιλαμβάνονται στη βιβλιοθήκη OpenPose[2], προσαρμόστηκαν και εκπαιδεύτηκαν με το σύνολο δεδομένων DMC2.5D. Οι αρθρώσεις του σώματος έχουν αντικατασταθεί από τις θέσεις τοποθέτησης των ανακλαστήρων, ενώ αξίζει να σημειωθεί ότι τα πεδία συσχέτισης των αρθρώσεων (PAFs) έχουν τροποποιηθεί λόγω της διαφορετικής τοποθέτησης των υποσυνόλων των δεικτών στην πρόσθια και οπίσθια πλευρά του σώματος. Η προσαρμοσμένη συσχέτιση των δεικτών απεικονίζεται στο Σχήμα 3.12. Επιπλέον, δεδομένου ότι οι μέθοδοι αξιολογούνται στην εκτίμηση δεικτών ενός υποκειμένου, παρόλο που ο κλάδος PAFs συμβάλλει στην εκμάθηση πρόβλεψης των χαρτών εμπιστοσύνης, η έξοδός του δεν λαμβάνεται υπόψη για την τελική εκτίμηση του συνόλου των δεικτών. Τέλος, αξιολογείται μια προσέγγιση δύο κλάδων χρωματισμένου βάθους και 3Δ οπτικής ροής δύο κλάδων παρόμοια με την OpenPose [1] (OpenPose [1] + 3D OF), η οποία παρουσιάζει αξιοσημείωτα αποτελέσματα.



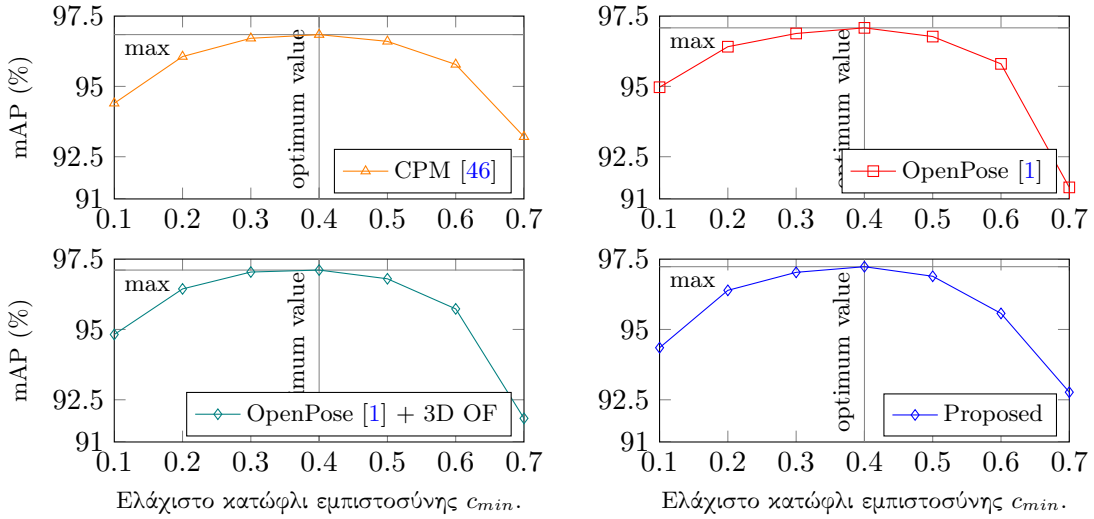
Σχήμα 3.12: *Τα κόκκινα βέλη απεικονίζουν τις κατευθυντικές συσχετίσεις μεταξύ των δεικτών για την προσαρμογή των πεδίων συχέτισης των μερών του σώματος, όπως προτείνεται στην [1]. Οι πορτοκαλί και μπλε αριθμήσεις υποδεικνύουν τους δείκτες-ιμάντες και τους τετραγωνικούς δείκτες, αντίστοιχα.*

Όσον αφορά τις μετρικές αξιολόγησης, το προτεινόμενο μοντέλο αξιολογείται στο σύνολο δεδομένων DMC2.5D μετρώντας τη Μέση Ακρίβεια (AP) ανά δείκτη και τη συνολική Μέση Ακρίβεια (mAP) για ολόκληρο το σύνολο. Αυτό επιτυγχάνεται με χρήση του κατωφλίου Percentage of Correct Keypoints (PCK) [89], δηλαδή, μια πρόβλεψη θεωρείται αληθής εάν εμπίπτει σε μια περιοχή εικονοστοιχείων γύρω από το ground truth 2Δ σημείο. Η περιοχή αυτή ορίζεται πολλαπλασιάζοντας το πλάτος και το ύψος του οριοθετημένου πλαισίου που

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose

περιέχει το υποκείμενο στην εικόνα με έναν παράγοντα α που ελέγχει το σχετικό κατώφλι.

Θέτοντας $\alpha = 0.05$, το σύνολο επαλήθευσης του συνόλου δεδομένων DMC2.5D χρησιμοποιήθηκε για να υποδείξει το βέλτιστο ελάχιστο όριο εμπιστοσύνης c_{min} για το υψηλότερο mAP ανά μέθοδο, με στόχο τη δίκαιη σύγκριση μεταξύ τους. Το c_{min} αντιστοιχεί στο ελάχιστο όριο εμπιστοσύνης για να θεωρηθεί έγκυρη μια πρόβλεψη δείκτη, δηλαδή $E_{R_i,f} > c_{min}$. Τα αποτελέσματα παρουσιάζονται στο Σχήμα 3.13, όπου απεικονίζεται η μετρική mAP ανά μέθοδο σε σχέση με το κατώφλι εμπιστοσύνης. Η μέγιστη τιμή mAP στο σύνολο επικύρωσης επιτεύχθηκε για $c_{min} = 0.4$ για όλες τις μεθόδους, επομένως, θεωρείται βέλτιστο για τα πειράματα στο σύνολο δοκιμών DMC2.5D.



Σχήμα 3.13: *Μέση Ακρίβεια με βάση τα PCK κατώφλια ορθών εκτιμήσεων ($a = 0.05$) σε σχέση με το κατώφλι εμπιστοσύνης, $mAP(c_{min})$.*

Αποτελέσματα και συζήτηση

Μελετώντας τα αποτελέσματα AP ανά δείκτη που παρουσιάζονται στον πίνακα 3.3, γίνεται αντιληπτή η αποτελεσματικότητα των μεθόδων στην εκτίμηση του συνόλου των δεικτών. Η προτεινόμενη μέθοδος υπερτερεί έναντι των υπόλοιπων μεθόδων για το 80,7% των δεικτών (δηλ. 21 από τους 26). Ειδικότερα, η μέση ακρίβεια AP ανά δείκτη βελτιώνεται για τους δείκτες που τοποθετούνται στα χέρια και τα πόδια ($R_{13}, R_{18}, R_{22}, R_{26}$), και τους πλησιέστερους σε αυτούς, δηλαδή τους καρπούς και τους αστραγάλους ($R_{12}, R_{17}, R_{21}, R_{25}$), οι οποίοι τοποθετούνται στα μέρη του σώματος με τη μεγαλύτερη κινητικότητα και, επομένως, τις πιο γρήγορες και μη περιορισμένες κινήσεις. Ως εκ τούτου συμπεραίνουμε ότι η χρονική πληροφορία που δίνεται ως είσοδος αλλά και κωδικοποιείται στο προτεινόμενο μοντέλο βελτιώνει την πρόβλεψη των δεικτών αυτών, ενώ για τους δείκτες που η AP είναι ελαφρώς χαμηλότερη ($R_8, R_{10}, R_{13}, R_{14}, R_{20}$), υποθέτουμε ότι η εκτιμώμενη οπτική ροή δεν ήταν αρκετά ακριβής ή πλούσια σε πληροφορία ώστε να ενισχύσει την εμπιστοσύνη της πρόβλεψης και, επομένως, την ακρίβεια των εκτιμήσεων.

Πίνακας 3.3: *AP για PCK με $\alpha = 0.05$, για κάθε έναν από τους 26 δείκτες.*

%	**R01**	**R02**	**R03**	**R04**	**R05**	**R06**	**R07**	**R08**	**R09**	**R10**	**R11**	**R12**	**R13**
[46]	96.81	96.11	99.22	95.06	90.98	85.26	98.78	**99.76**	95.25	94.70	96.99	93.33	85.92
[1]	96.95	95.36	98.77	96.69	91.08	85.26	98.78	99.51	96.00	95.89	97.15	93.79	87.64
[1] + 3D OF	96.81	96.61	99.45	94.85	89.98	85.53	98.78	99.45	96.00	**96.27**	97.25	93.49	**87.66**
Proposed	**98.10**	**97.31**	**99.48**	**97.35**	**91.36**	**86.20**	**99.00**	98.27	**96.64**	95.32	**97.81**	**95.19**	87.13
%	**R14**	**R15**	**R16**	**R17**	**R18**	**R19**	**R20**	**R21**	**R22**	**R23**	**R24**	**R25**	**R26**
[46]	83.42	93.21	98.24	96.25	75.23	98.39	98.38	94.52	74.28	94.88	99.30	97.18	64.73
[1]	84.58	92.88	98.66	95.95	76.19	98.39	**98.44**	94.88	74.06	96.99	99.52	97.87	71.18
[1] + 3D OF	**86.33**	94.11	98.44	95.92	72.29	98.40	98.42	95.14	78.68	96.53	99.21	97.76	70.56
Proposed	85.61	**94.79**	**98.81**	**97.50**	**77.17**	**99.30**	98.18	**96.61**	**79.23**	**97.93**	**100.0**	**98.73**	**73.96**

Ωστόσο, για τις λοιπές μεθόδους, η πρόβλεψη των ακραίων δεικτών είναι αδύναμη σε σύγκριση με των υπολοίπων. Αυτό πιθανώς οφείλεται στο γεγονός ότι οι δείκτες αυτοί δεν είναι συχνά ορατοί και είναι τοποθετημένοι σε μικρές αποστάσεις από τους γειτονικούς τους δείκτες. Για να τονιστεί αυτή η διαφορά, τα αποτελέσματα mAP παρουσιάζονται στον Πίνακα 3.4 με και χωρίς τη συμμετοχή των ακραίων δεικτών.

Πίνακας 3.4: *mAP για PCK με* $\alpha = 0.05$*, με και χωρίς τους δείκτες τοποθετημένους στα άκρα.*

Method	**Total**	**Total (without End-Reflectors)**
CPM [46]	92.16%	95.27%
OpenPose [1]	92.79%	95.61%
OpenPose [1] + 3D OF	92.84%	95.67%
Proposed	**93.73%**	**96.77%**

Η προτεινόμενη προσέγγιση υπερτερεί των συγκρινόμενων μεθόδων, παρουσιάζοντας απόλυτο ποσοστό βελτίωσης του mAP ίση με 1.47%, 0.94% και 0.89% με συμμετοχή των ακραίων δεικτών και 1.5%, 1.16% και 0.9% χωρίς αυτούς, σε σύγκριση με τις μεθόδους [46], [1] και [1] + 3D OF, αντίστοιχα. Αξίζει να σημειωθεί ότι η προσέγγιση δύο κλάδων που λαμβάνει υπόψη την 3Δ οπτική ροή ([1] + 3D OF) επιτυγχάνει υψηλότερο mAP από την [1, 46], πράγμα που σημαίνει ότι η χρονική πληροφορία που παρέχεται από τον κλάδο 3Δ οπτικής ροής κωδικοποιείται στο χώρο χαρακτηριστικών του μοντέλου, με αποτέλεσμα υψηλότερη ακρίβεια εντοπισμού.

Τέλος, προτού τροφοδοτηθεί η μέθοδος καταγραφής κίνησης με τις εκτιμήσεις του συνόλου των 2Δ δεικτών, εφαρμόζεται μια διαδικασία φιλτραρίσματος που βασίζεται σε δύο θεμελιώδεις θεωρήσεις. Αρχικά, οι εκτιμήσεις των δεικτών θεωρούνται αποδεκτές μόνο όταν οι δείκτες είναι ορατοί: i) εάν η περιοχή όπου εκτιμάται η θέση του δείκτη δεν ανήκει σε μια συγκεκριμένη χρωματική (μαύρη) περιοχή μεγέθους μεγαλύτερου ή ίσου με $b_{min} = 5$ εικονοστοιχεία, η εκτίμηση αυτή απορρίπτεται, και ii) όταν ανιχνεύονται περισσότεροι από ένας δείκτες στην ίδια θέση (απόλυτη απόσταση μικρότερη από 3 εικονοστοιχεία, πειραματικά καθορισμένη), οι δείκτες με χαμηλότερη τιμή εμπιστοσύνης απορρίπτονται. Δεύτερον, οι δείκτες είναι μοναδικοί σε μια εικόνα, καθώς μέσω του DeepMoCap προσεγγίζουμε το πρόβλημα καταγραφής κίνησης ενός υποκειμένου μόνο - εάν ανιχνευθούν περισσότερες από μία εκτιμήσεις του ίδιου δείκτη, αυτή με την υψηλότερη εμπιστοσύνη θεωρείται έγκυρη.

Τα αποτελέσματα AP μετά το φιλτράρισμα παρουσιάζονται στον Πίνακα 3.5. Όπως φαίνεται, τα αποτελέσματα για όλους τους δείκτες και για όλες τις μεθόδους είναι ίσα ή μεγαλύτερα από τα αντίστοιχα αποτελέσματα πριν από το φιλτράρισμα. Σε αυτό το πείραμα, το προτεινόμενο μοντέλο ξεπερνά τις υπόλοιπες μεθόδους σε 14 από τους 26 (53.84%) δείκτες.

Πίνακας 3.5: *AP για PCK με $\alpha = 0.05$, για κάθε έναν από τους 26 δείκτες μετά το φιλτράρισμα.*

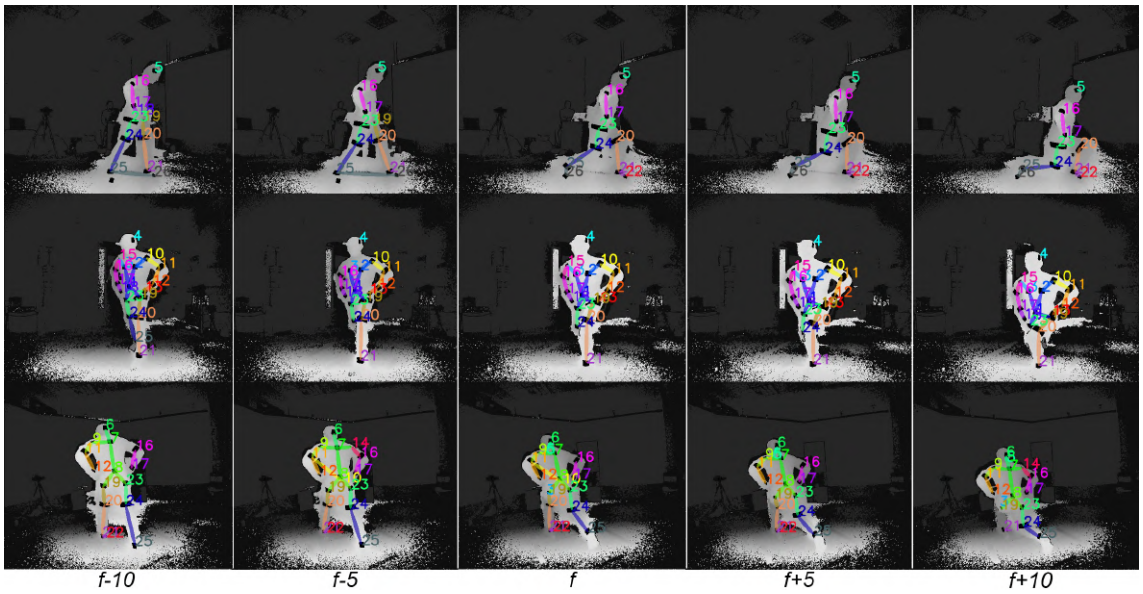
%	**R01**	**R02**	**R03**	**R04**	**R05**	**R06**	**R07**	**R08**	**R09**	**R10**	**R11**	**R12**	**R13**
[46]	96.95	96.39	99.45	97.30	91.74	86.81	100.0	**100.0**	96.62	**97.46**	98.15	94.16	90.50
[1]	96.95	96.19	98.77	98.31	**93.98**	86.81	100.0	99.76	96.87	**97.46**	98.55	94.64	90.79
[1] + 3D OF	96.81	96.88	99.45	97.30	91.72	86.81	100.0	99.70	96.87	**97.46**	98.55	94.05	**91.67**
Proposed	**98.52**	**98.61**	**99.81**	**98.65**	92.58	**87.35**	100.0	**100.0**	**98.90**	96.59	**100.0**	**95.50**	89.60
%	**R14**	**R15**	**R16**	**R17**	**R18**	**R19**	**R20**	**R21**	**R22**	**R23**	**R24**	**R25**	**R26**
[46]	83.78	95.93	99.03	97.19	78.81	98.60	99.41	96.59	74.08	97.59	99.40	98.36	68.64
[1]	**90.19**	96.40	**99.32**	97.17	**82.82**	98.60	99.37	**97.25**	75.66	**99.09**	99.61	98.60	69.52
[1] + 3D OF	**90.19**	**96.75**	99.14	98.30	78.87	98.50	**99.50**	96.83	**81.69**	98.25	99.30	98.52	72.68
Proposed	88.90	95.10	99.13	**97.82**	78.20	**99.63**	98.50	96.93	79.49	98.25	**100.0**	**99.05**	**78.21**

Τα αποτελέσματα mAP για το σύνολο των δεικτών μετά το φιλτράρισμα, με και χωρίς τη συμμετοχή των ακραίων δεικτών, παρουσιάζονται στον Πίνακα 3.6, επίσης παρουσιάζοντας μεγαλύτερη ακρίβεια από τις αντίστοιχες τιμές πριν το φιλτράρισμα. Η προτεινόμενη μέθοδος υπερτερεί των CPM [46], OpenPose [1] και OpenPose [1] + 3D OF παρουσιάζοντας απόλυτη αύξηση ίση με 1.25%, 0.49% και 0.37% με τον υπολογισμό των ακραίων δεικτών και 1.06%, 0.47% και 0.61% χωρίς αυτούς, αντίστοιχα.

Πίνακας 3.6: *mAP για PCK με $\alpha = 0.05$, με και χωρίς τους δείκτες τοποθετημένους στα άκρα.*

Method	Total	Total (without End-Reflectors)
CPM [46]	93.57%	96.41%
OpenPose [1]	94.33%	97.00%
OpenPose [1] + 3D OF	94.45%	96.86%
Proposed	**94.82%**	**97.47%**

Ποιοτικά αποτελέσματα του προτεινόμενου μοντέλου σε διαδοχικά καρέ απεικονίζονται στο Σχήμα 3.14, ενώ περισσότερα ποιοτικά αποτελέσματα έχουν δημοσιοποιηθεί (https://vcl.iti.gr/deepmocap).



Σχήμα 3.14: *Οπτικοποίηση των αποτελεσμάτων του προτεινόμενου μοντέλου σε διαδοχικά καρέ βάθους πολλαπλών όψεων. Παρουσιάζονται 5 διαδοχικά καρέ πολλαπλών όψεων οριζοντίως, από το καρέ $f - 10$ έως το $f + 10$ με βήμα πλαισίου ίσο με 5.*

3.4.2 Πειραματικά Αποτελέσματα 3Δ Μεθόδου

Πειραματικό Πλαίσιο

Με την εισαγωγή ενός αποτελεσματικού μοντέλου για την 2Δ εκτίμηση του συνόλου των δεικτών, αξιολογείται το τελικό αποτέλεσμα καταγραφής κίνησης. Εφαρμόζοντας το προτεινόμενο μοντέλο σε ακολουθίες πολλαπλών όψεων, τα 3Δ οπτικά δεδομένα εξάγονται και

τροφοδοτούνται στην προτεινόμενη μέθοδο καταγραφής κίνησης. Για τα πειράματα αυτά, επιλέχθηκαν ακολουθίες αποτελλούμενες από περίπου 6×10^3 καρέ συνολικά από 2 διαφορετικά υποκείμενα που είχαν εξαιρεθεί από το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου. Λαμβάνοντας υπόψη τα ground truth δεδομένα του συνόλου DCM3D, δηλαδή, τα δεδομένα κίνησης που καταγράφηκαν με το PhaseSpace Impulse X2 [84], το αποτέλεσμα της καταγραφής κίνησης συγκρίνεται με τα αποτελέσματα του αισθητήρα Kinect for Xbox One με την υψηλότερη ποιότητα ανά καρέ, μια μέθοδο που συγχωνεύει δεδομένα κίνησης Kinect και αδρανειακά δεδομένα από 9 IMU (Fusion) [90], και μια δεύτερη ακόμη πιο ισχυρή μέθοδο με αδρανειακούς αισθητήρες παρόμοια με τη [90] (Fusion++) που συγχωνεύει δεδομένα ground truth για την αρχικοποίηση και τον εντοπισμό του κέντρου βάρους του σώματος αντί για τη χρήση δεδομένων Kinect. Αξίζει να σημειωθεί ότι οι πολύ λανθασμένες εκτιμήσεις του Kinect που προκαλούν εξαιρετικά εσφαλμένες εκτιμήσεις της 3Δ θέσης του κέντρου βάρους έχουν εξαιρεθεί από τις ακολουθίες αξιολόγησης, διατηρώντας μόνο τις εκτιμήσεις που έχουν νόημα για σύγκριση.

Οι 'αδρανειακές' μέθοδοι θεωρήθηκαν κατάλληλες για σύγκριση λόγω της σταθερότητας καταγραφής τους, χωρίς να επηρρεάζονται από αποκρύψεις μερών του σώματος όπως συνηθίζεται με τις οπτικές μεθόδους. Άλλες προσεγγίσεις βασιζόμενες σε RGB-D ή 3Δ δεδομένα δεν ελήφθησαν υπόψη λόγω των ελλειπόντων τμημάτων βάθους και, ως εκ τούτου, των ελλειπόντων 3Δ δεδομένων για τα μέρη του σώματος του υποκειμένου στα οποία τοποθετήθηκαν οι ρετροαναλαστικοί δείκτες, με αποτέλεσμα να επηρρεάζονται τα αποτελέσματα και παρατηρείται μη δίκαιη σύγκριση. Τέλος, μέθοδοι 3Δ καταγραφής κίνησης από απλές RGB κάμερες θεωρήθηκαν εκτός πεδίου εφαρμογής λόγω της εισόδου 2Δ δεδομένων.

Όσον αφορά τις μετρικές αξιολόγησης, το DeepMoCap αξιολογείται στο σύνολο δεδομένων DMC3D χρησιμοποιώντας τις μετρικές Mean Average Error (MAE), Root Mean Squared Error (RMSE) και 3D PCK @ $a_{3D} = 20$ cm για την 3Δ ευκλείδεια απόσταση μεταξύ του αποτελέσματος των μεθόδων και των ground truth δεδομένων σε 12 αρθρώσεις που περιλαμβάνουν τους ώμους, τους αγκώνες, τους καρπούς, τους γοφούς, τα γόνατα και τους αστραγάλους. Στο 3D PCK, μια εκτίμηση θεωρείται σωστή όταν η 3Δ ευκλείδεια απόσταση της από την αντίστοιχη ground truth θέση είναι μικρότερη από το κατώφλι a_{3D}.

Αποτελέσματα και συζήτηση

Τα συνολικά αποτελέσματα της σύγκρισης μεταξύ των μεθόδων παρουσιάζονται στον πίνακα 3.7, όπου φαίνεται η υπεροχή της προτεινόμενης μεθόδου σε σύγκριση με τις υπόλοιπες. Τα συνολικά MAE, RMSE και 3D PCK για όλες τις ακολουθίες είναι $9,02$ cm, $10,06$ cm και $92,25\%$, αντίστοιχα, παρουσιάζοντας τα καλύτερα αποτελέσματα μεταξύ των πειραματικών μεθόδων. Οι Fusion++ [90], Fusion [90] και Best Kinect [36] ακολουθούν την προτεινόμενη μέθοδο παρουσιάζοντας 88.75%, 85.93% και 83.37% στην ακρίβεια 3D PCK, αντίστοιχα. Επιπλέον, αξίζει να αναφερθεί ότι η προτεινόμενη μέθοδος παρουσιάζει συνολικό μέσο σφάλμα κάτω από 10 cm (MAE).

Πίνακας 3.7: *Συγκριτική αξιολόγηση των αποτελεσμάτων καταγραφής κίνησης των συγκρινόμενων μεθόδων, παρουσιάζοντας τις συνολικές μετρικές MAE, RMSE και 3D PCK ($\alpha_{3D} = 20$ cm).*

Method	**MAE (cm)**	**RMSE (cm)**	**3D PCK ($a = 20$ cm) [89]**
Best Kinect [36]	15.35	16.06	82.03%
Fusion [90]	12.31	12.91	85.93%
Fusion++ [90]	10.66	11.30	88.75%
Proposed	**9.02**	**10.06**	**92.25%**

Στον Πίνακα 3.8, παρουσιάζονται αποτελέσματα 3D PCK ανά άσκηση, δίνοντας στοιχεία σχετικά με τα δυνατότητες των μεθόδων στις διαφορετικές κινήσεις του σώματος. Το Deep-

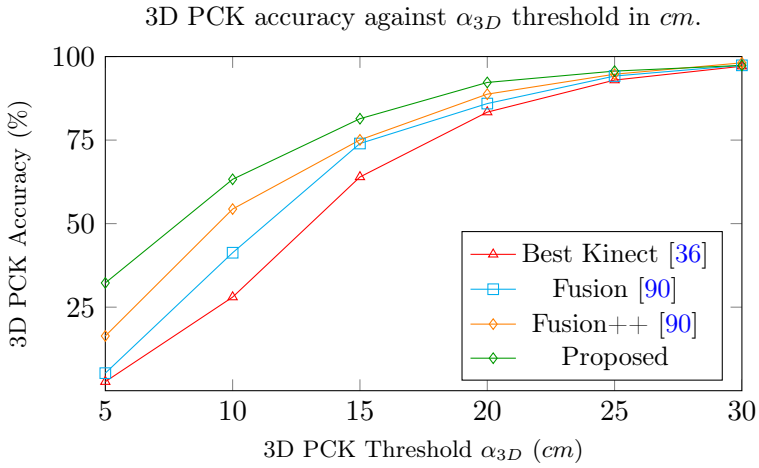
Πίνακας 3.8: *Αξιολόγηση ανά άσκηση με χρήση 3Δ PCK μετρική με $\alpha_{3D} = 20$ cm.*

Exercise	**Best Kinect [36]**	**Fusion [90]**	**Fusion++ [90]**	**Proposed**
Walking on the spot	96.60%	**100.00%**	97.54%	**100.00%**
Single arm raise	93.57%	96.19%	97.38%	**100.00%**
Elbow flexion	91.12%	**100.00%**	**100.00%**	97.40%
Knee flexion	88.36%	94.11%	**100.00%**	98.80%
Closing arms above head	82.48%	80.08%	83.33%	**88.62%**
Side steps	85.00%	88.33%	**93.33%**	87.50%
Jumping jack	95.48%	84.18%	87.57%	**96.05%**
Butt kicks left-right	81.87%	80.99%	86.26%	**90.94%**
Forward lunge left-right	57.31%	87.93%	86.05%	**92.01%**
Classic squat	59.60%	78.67%	83.05%	**90.40%**
Side step + knee-elbow	77.78%	80.25%	81.94%	**89.81%**
Side reaches	89.24%	84.55%	87.88%	**91.52%**
Side jumps	90.00%	89.31%	92.78%	**93.47%**
Alternate side reaches	68.01%	74.19%	77.69%	**82.53%**
Kick-box kicking	74.07%	70.14%	76.39%	**84.72%**

MoCap επιτυγχάνει υψηλότερη ακρίβεια από τις υπόλοιπες μεθόδους σε 12 από τις 15 ασκήσεις συνολικά (80%), καταγράφοντας αποτελεσματικά τις περισσότερες από αυτές. Αναλυτικότερα, το *Walking on the spot*, μια απλή και αργή κίνηση, καταγράφεται αποτελεσματικά από όλες τις μεθόδους. Οι ασκήσεις *Elbow flexion* και *Knee flexion*, οι οποίες είναι απλές περιστροφικές κινήσεις των αρθρώσεων, αποτυπώνονται με μεγάλη ακρίβεια από όλες τις μεθόδους, ειδικά από τις 'αδρανειακές'. Αξίζει να σημειωθεί ότι η DeepMoCap παρουσιάζει χαμηλότερη ακρίβεια για την άσκηση *Side-steps* από τη Fusion++ πιθανώς λόγω της τοποθέτησης των χεριών στους δείκτες που είναι τοποθετημένοι στα ισχία με αποτέλεσμα να συγχωνεύονται ρετροανακλαστικές περιοχές, γεγονός που δυσκολεύει την ανίχνευση και την ταυτοποίηση των συγκεκριμένων δεικτών. Ωστόσο, σε πιο σύνθετες ασκήσεις όπως οι *Butt kicks left-right* και *Forward lunge left-right* όπου υπάρχουν αποκρύψεις για το Kinect και εκτάσεις του σώματος για τους αδρανειακούς αισθητήρες που είναι τοποθετημένοι στο σώμα, το DeepMoCap παρουσιάζει περίπου 5% υψηλότερη απόλυτη ακρίβεια 3D PCK από το Fusion++. Για την *Jumping jack*, η οποία είναι μια γρήγορη και σύνθετη άσκηση όπου συμμετέχουν πλήρως όλα τα μέρη του σώματος, το DeepMoCap επιτυγχάνει ακρίβεια 96.05% 3D PCK και ακολουθεί το Best Kinect [36], ενώ οι 'αδρανειακές' προσεγγίσεις αποτυγχάνουν να συλλάβουν σωστά τις θέσεις των ώμων λόγω ίσως της άκαμπτης κίνησης του σώματος του κορμού, παρουσιάζοντας χαμη-

λότερη ακρίβεια. Για τις ασκήσεις *Alternate side reaches* και *Kick-box kicking*, οι οποίες είναι οι πιο απαιτητικές, η ακρίβεια 3D PCK της προτεινόμενης μεθόδου είναι κατά 4.84% και 8.33% υψηλότερη σε σύγκριση με την αμέσως επόμενη καλύτερη μέθοδο (Fusion++), αντίστοιχα. Επιπλέον, θα πρέπει να σημειωθεί ότι όλες οι ασκήσεις αποτυπώνονται από το DeepMoCap με ακρίβεια 3D PCK υψηλότερη από 82.53%, γεγονός που δείχνει χαμηλή διακύμανση μεταξύ διαφορετικών τύπων κινήσεων του σώματος.

Στο διάγραμμα που παρουσιάζεται στο Σχήμα 3.15, δίνεται η συνολική 3Δ ακρίβεια PCK σε σχέση με τις τιμές κατωφλίου a_{3D}. Το DeepMoCap παρουσιάζει υψηλότερη απόδοση για όλα τα a_{3D}, παρουσιάζοντας ικανοποιητικά αποτελέσματα ήδη από χαμηλές τιμές κατωφλίου (π.χ. 32.25% και 63.27% για $\alpha_{3D} = 5$ cm και $\alpha_{3D} = 10$ cm, αντίστοιχα), σε σύγκριση με την αμέσως επόμενη καλύτερη μέθοδο που παρουσιάζει 16.38% και 54.36%, αντίστοιχα. Λαμβάνοντας υπόψη το γεγονός ότι η εκτίμηση των αρθρώσεων ποικίλλει μεταξύ διαφορετικών προσεγγίσεων καταγραφής κίνησης με αποτέλεσμα την ύπαρξη μιας σταθερής μετατόπισης μεταξύ των εκτιμώμενων 3Δ θέσεων, καταλήγουμε στο συμπέρασμα ότι η DeepMoCap παρουσιάζει υψηλή απόδοση παρουσιάζοντας 32.25% των εκτιμήσεων κατά μέσο όρο να είναι πιο κοντά από 5 cm από τα ground truth δεδομένα.



Σχήμα 3.15: *Συγκριτική αξιολόγηση των μεθόδων καταγραφής κίνησης χρησιμοποιώντας τα συνολικά 3D PCK αποτελέσματα για διάφορες τιμές κατωφλίου α_{3D} σε cm.*

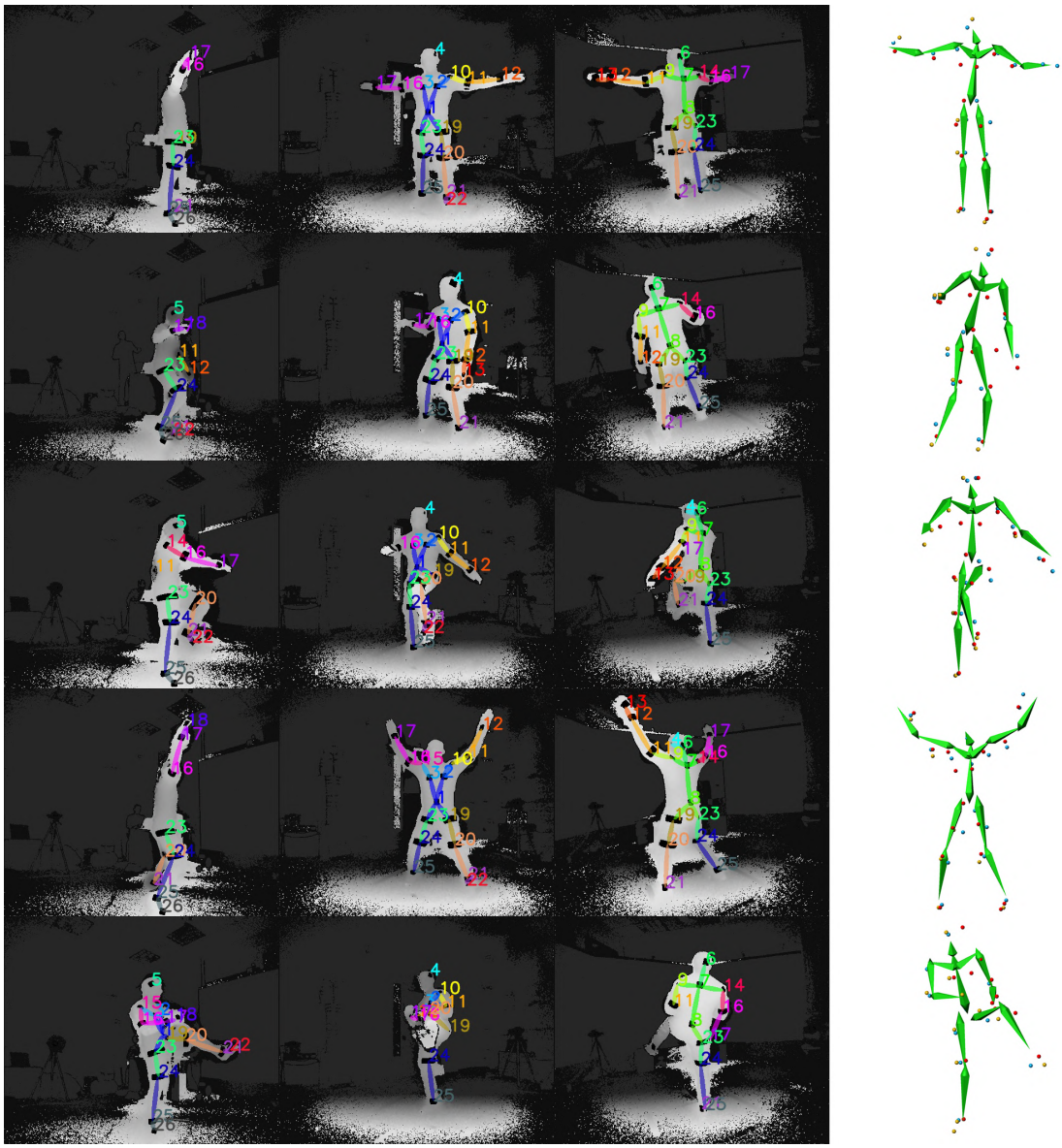
Στον πίνακα 3.9, παρουσιάζονται οι μετρικές MAE και RMSE ανά άρθρωση. Παρατηρείται ότι η προτεινόμενη μέθοδος έχει τα χαμηλότερα σφάλματα για 9 από τις 12 (75%) αρθρώσεις για τις MAE και τις RMSE μετρικές. Για τις αρθρώσεις *Shoulders* και *Right Elbow*, η Fusion++ [90] παρουσιάζει ελαφρώς καλύτερα αποτελέσματα από την DeepMoCap πιθανώς λόγω της καλύτερης τοποθέτησης της δομής του σκελετού. Οι αρθρώσεις του κάτω μέρους του σώματος (γοφοί, γόνατα και πόδια) παρουσιάζουν 6, 05 cm και 7, 08 cm ολικό μέσο όρο και μέσο όρο τετραγωνικών σφαλμάτων, αντίστοιχα.

Ποιοτικά αποτελέσματα που απεικονίζουν το 3Δ αποτέλεσμα της προτεινόμενης προσέγγισης παρουσιάζονται στο Σχήμα 3.16. Συγκεκριμένα, απεικονίζεται η είσοδος πολλαπλών προβολών με τις 2Δ εκτιμήσεις του συνόλου των δεικτών και τα αντίστοιχα αποτελέσματα της 3Δ καταγραφής κίνησης μαζί με τα οπτικά δεδομένα. Όπως φαίνεται, το DeepMoCap προσεγγίζει τη καταγραφή κίνησης παρόμοια με τον τρόπο που λειτουργούν οι παραδοσιακές οπτικές

Πίνακας 3.9: *Πειραματικά αποτελέσματα των μεθόδων καταγραφής κίνησης χρησιμοποιώντας τις μετρικές MAE και RMSE ανά άρθρωση (σε cm).*

Joint	**Best Kinect [36]**		**Fusion [90]**		**Fusion++ [90]**		**Proposed**	
	MAE	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Left Shoulder	12.83	13.25	8.16	**8.37**	**7.89**	8.53	11.41	12.63
Right Shoulder	15.59	15.90	9.71	10.28	**8.62**	**9.19**	11.09	11.76
Left Elbow	16.04	17.45	16.46	17.17	15.90	16.60	**13.25**	**14.84**
Right Elbow	19.37	19.67	11.61	12.61	**10.88**	**11.68**	12.25	13.36
Left Hand	16.01	17.95	14.52	15.87	13.53	14.44	**11.94**	**12.58**
Right Hand	21.24	21.55	13.10	14.24	12.42	13.29	**12.04**	**13.05**
Left Hip	8.63	8.82	9.99	10.20	6.33	6.45	**4.18**	**4.69**
Right Hip	10.89	11.16	10.59	10.81	5.94	6.11	**4.53**	**4.99**
Left Knee	10.79	11.73	12.55	13.10	9.97	10.45	**5.12**	**5.82**
Right Knee	15.13	15.99	12.17	12.64	8.92	9.45	**7.24**	**8.16**
Left Foot	17.74	18.34	16.35	17.11	15.57	16.40	**7.00**	**8.82**
Right Foot	19.91	20.86	12.48	12.53	11.93	12.97	**8.24**	**10.00**

λύσεις καταγραφής κίνησης με δείκτες, ωστόσο με πιο ευέλικτο και χαμηλού κόστους τρόπο. Πιο πολλά ποιοτικά αποτελέσματα είναι διαθέσιμα στο (https://vcl.iti.gr/deepmocap).



Σχήμα 3.16: *Δείγματα των αποτελεσμάτων της μεθόδου απεικονίζονται ανά γραμμή. Στην αριστερή πλευρά παρουσιάζονται η είσοδος πολλαπλών προβολών μαζί με τις εκτιμήσεις του συνόλου των 2Δ δεικτών από το μοντέλο, ενώ, στη δεξιά πλευρά, παρουσιάζονται τα αντίστοιχα αποτελέσματα 3Δ δεικτών και πόζας από την καταγραφή της κίνησης.*

3.4.3 Ανάλυση επιδόσεων

Η ανάλυση επιδόσεων εκτέλεσης πραγματοποιήθηκε με τη μέτρηση του συνολικού χρόνου που απαιτήθηκε για τον υπολογισμό των δεδομένων κίνησης από τις ακολουθίες αξιολόγησης. Για περίπου 6×10^3 ζεύγη 3 προβολών χρωματισμένων καρέ βάθους και 3Δ οπτικής ροής, άρα 18×10^3 ζεύγη μίας όψης, η αρχική επεξεργασία των δεδομένων διήρκεσε 1796 s (∼100 ms ανά δείγμα), ενώ η πρόβλεψη του μοντέλου διήρκησε 3136 s (∼174 ms ανά δείγμα). Έτσι, η προτεινόμενη μέθοδος επιτυγχάνει την εκτίμηση του συνόλου των 2Δ δεικτών σε περίπου 6 καρέ ανά δευτερόλεπτο, ενώ η καταγραφή της κίνησης από 3Δ οπτικά δεδομένα πραγματοποιείται σε πραγματικό χρόνο απαιτώντας λιγότερο από 10 ms. Σε σχέση με την είσοδο, το αρχικό μέγεθος του πλαισίου είναι 424×512, το οποίο κατά τη διάρκεια των δοκιμών διαμορφώθηκε

σε 368×444. Έτσι, το DeepMoCap εκτελεί καταγραφή κίνησης σε περίπου 2 fps για είσοδο 3 όψεων των 368×444, ενώ η πολυπλοκότητα της απόδοσης σε σχέση με τον αριθμό των όψεων, δηλαδή τα ζεύγη εικόνων εισόδου, είναι $\mathcal{O}(n)$. Η ανάλυση κατά το χρόνο εκτέλεσης πραγματοποιήθηκε σε ένα μηχάνημα εξοπλισμένο με μία GPU NVIDIA GeForce Titan X. Ο κώδικας (https://github.com/tofis/deepmocap) και τα εργαλεία συνόλου δεδομένων της προτεινόμενης μεθόδου είναι δημόσια διαθέσιμα για να ενθαρρύνουν την περαιτέρω έρευνα στην ερευνητική αυτή περιοχή.

3.5 Συμπεράσματα

Στην παρόν κεφάλαιο, παρουσιάζεται μια μέθοδος οπτικής καταγραφής κίνησης βαθιάς μάθησης, χρησιμοποιώντας πολλαπλούς αισθητήρες υπερύθρων-βάθους και οπισθοανακλαστικούς, μη τυποποιημένους δείκτες. Αποτελεί μια γρήγορη και ευέλικτη προσέγγιση που εξάγει αυτόματα ταυτοποιημένα 3Δ οπτικά δεδομένα και εκτελεί άμεση καταγραφή κίνησης χωρίς την ανάγκη μετα-επεξεργασίας. Αυτό επιτυγχάνεται με την εισαγωγή ενός δικτύου δύο ροών, πολλαπλών σταδίων το οποίο εισάγει μια μη παραμετρική αναπαράσταση για την κωδικοποίηση της χρονικής συσχέτισης μεταξύ ζευγών χρωματισμένων χαρτών βάθους και καρέ 3Δ οπτικής ροής, με αποτέλεσμα τον 2Δ εντοπισμό και ταυτοποίηση των δεικτών. Αυτό το βήμα επιτρέπει την εξαγωγή 3Δ οπτικών δεδομένων από πολλαπλούς χωροχρονικά συσχετισμένους αισθητήρες με τελικό στόχο την καταγραφή της ανθρώπινης κίνησης. Για σκοπούς έρευνας και αξιολόγησης, δημιουργήθηκαν δύο νέα δημόσια σύνολα δεδομένων. Η προτεινόμενη μέθοδος αξιολογήθηκε όσον αφορά την εκτίμηση του συνόλου των 2Δ δεικτών και την ακρίβεια της σύλληψης κίνησης σε αυτά τα σύνολα δεδομένων, ξεπερνώντας τις πρόσφατες και εύρωστες μεθόδους και στις δύο εργασίες.

Όσον αφορά τους περιορισμούς, η καταγραφή από πλευρικές όψεις και οι εξαιρετικά πολύπλοκες στάσεις του σώματος που αποκρύπτουν ή συγχέουν τους δείκτες στις υπέρυθρες εικόνες αποτελούν κύριο πρόβλημα προς επίλυση. Η συγχώνευση των δεικτών κατά την προβολή τους εντείνεται λόγω του μεγέθους των δεικτών, καθώς και της τοποθέτησής τους. Φυσικά, αυτοί οι περιορισμοί μπορούν να μετριαστούν με την αύξηση του αριθμού των αισθητήρων γύρω από το υποκείμενο, αυξάνοντας ωστόσο το κόστος και την πολυπλοκότητα της μεθόδου. Επίσης, αξίζει να τονιστεί ότι η ακρίβεια συγχρονισμού των καμερών δεν είναι υψηλή δεδομένου ότι δεν υπάρχει συγχρονισμός υλισμικού, μονάχα λογισμικού, με αποτέλεσμα διαφορές μεταξύ των στιγμιοτύπων των διαφόρων προβολών, ιδιαίτερα στις πολύ γρήγορες κινήσεις.

Αναλύοντας τα παραπάνω, οδηγούμαστε στην ανάγκη δημιουργίας ενός νέου συνόλου δεδομένου, με χρήση τυποποιημένων σφαιρικών δεικτών μικρής διαμέτρου και αισθητήρες βάθους που επιτρέπουν το συγχρονισμό υλισμικού μεταξύ των δεδομένων των αισθητήρων. Η δημιουργία του συνόλου αυτού περιγράφεται λεπτομερώς στο επόμενο κεφάλαιο.

Κεφάλαιο 4

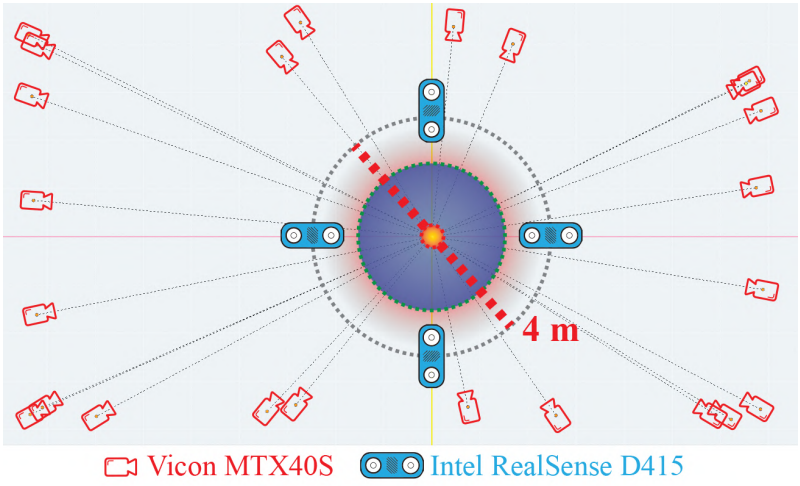
Σύνολο Δεδομένων HUMAN4D

Το αυξημένο ερευνητικό ενδιαφέρον που έχει επέλθει τα τελευταία χρόνια για την ανάπτυξη καινοτόμων μεθόδων εκτίμησης ανθρώπινης πόζας και κίνησης με χρήση υπολογιστικής όρασης και βαθιάς μάθησης, καθώς και το ενδιαφέρον της βιομηχανίας για την καταγραφή/παραγωγή 4Δ πολυμέσων, αυξάνουν την ανάγκη για δημιουργία νέων, υψηλής ποιότητας συνόλων δεδομένων. Εντούτοις προς το παρόν, λίγα σύνολα δεδομένων επιτυγχάνουν να καλύψουν (και μάλιστα μερικώς) αυτές τις ανάγκες, όπως συζητήθηκε στο Κεφ. 2.4.

Πολλές μέθοδοι προσεγγίζουν τις ερευνητικές αυτές προκλήσεις με δεδομένα από κάμερες χρώματος μονής προβολής ή πολλαπλές συγχρονισμένες κάμερες, χρώματος μόνο. Ωστόσο, εξ΄ ορισμού, τα 2Δ δεδομένα δεν μπορούν να ανταπεξέλθουν στις αυξημένες απαιτήσεις υπολογισμού 3Δ σχημάτων, μορφών και αρθρωτών κινήσεων, τουλάχιστον στην ακρίβεια που μπορούν τα αντίστοιχα 3Δ. Η επιλογή αυτή, πέραν από το καθαρά ερευνητικό ενδιαφέρον για τα διαφορετικά αυτά πεδία, ενισχύεται πιθανώς από την έλλειψη συγχρονισμένων μέσω υλισμικού δεδομένων βάθους (3Δ) πολλαπλών προβολών σε δημόσια σύνολα.

Για το σκοπό αυτό, και κυρίως για να ικανοποιήσουμε την ανάγκη της παρούσας διατριβής στη διερεύνηση καινοτόμων μεθόδων εκτίμησης πόζας και καταγραφής κίνησης, δημιουργήσαμε το σύνολο δεδομένων HUMAN4D. Το σύνολο αυτό καλύπτει σε μεγάλο βαθμό σημαντικές ανάγκες στο πεδίο παρέχοντας δεδομένα κίνησης υψηλής ακρίβειας μαζί με δεδομένα χρώματος-βάθους πολλαπλών προβολών με συγχρονισμό υλισμικού που παρέχει μεγάλη ακρίβεια. Συγκεκριμένα, η κύρια συνεισφορά του HUMAN4D είναι η δημοσίευση ενός συνόλου επισημασμένων, χωροχρονικά συσχετισμένων δεδομένων χρώματος-βάθους πολλαπλών προβολών και δεδομένων κίνησης, προκειμένου να καταστεί δυνατή η εκτεταμένη έρευνα νέων μεθόδων όρασης υπολογιστών, καθώς και ψηφιακών γραφικών. Εξ όσων γνωρίζουμε, το HUMAN4D είναι το πρώτο σύνολο δεδομένων που παρέχει συγχρονισμένα με χρήση συγχρονισμού υλισμικού καρέ χρώματος-βάθους μαζί με δεδομένα κίνησης και ήχου, με τη χρήση χαμηλού κόστους καμερών βάθους και συστημάτων καταγραφής κίνησης τελευταίας τεχνολογίας, καθώς και μικροφώνων στα σώματα των ηθοποιών, αντίστοιχα.

Σχήμα 4.1: Φωτογραφίες από την προετοιμασία και λήψη του συνόλου δεδομένων HUMAN4D. Η αίθουσα είναι εξοπλισμένη με 24 κάμερες Vicon MX T40S που είναι σταθερά τοποθετημένες στους τοίχους, ένα φορητό σύστημα καταγραφής πολλαπλών προβολών με 4 αισθητήρες βάθους Intel RealSense D415 που εγκαταστάθηκαν προσωρινά για τη λήψη των RGB-D δεδομένων, καθώς και φορητά μικρόφωνα για τους ηθοποιούς.



Σχήμα 4.2: *Κάτοψη χώρου με τις θέσεις των 24 καμερών Vicon MXT40S και των 4 αισθητήρων Intel RealSense D415.*

4.1 Μεθοδολογία Δημιουργίας

4.1.1 Σύστημα 4Δ Καταγραφής

Η καταγραφή του συνόλου δεδομένων πραγματοποιήθηκε σε επαγγελματικό στούντιο (Artanim Foundation[1]), όπου, πέρα από το σύστημα σύλληψης κίνησης, εγκαταστάθηκε προσωρινά ειδικός φορητός RGB-D εξοπλισμός πολλαπλών καμερών, όπως απεικονίζεται στο Σχήμα 4.1. Χρησιμοποιήθηκαν 24 κάμερες καταγραφής κίνησης μαζί με 4 στερεοσκοπικούς αισθητήρες βάθους και μικρόφωνα με χρήση συγχρονισμού υλισμικού και λογισμικού (βλ. Κεφ. 4.1.4 για λεπτομέρειες). Οι 24 κάμερες ήταν τοποθετημένες στους τοίχους, κοντά στην οροφή, όπως συνηθίζεται για αυτά τα συστήματα, για μεγιστοποίηση του ωφέλιμου όγκου καταγραφής. Ο μεγάλος αριθμός καμερών αυξάνει την ακρίβεια της καταγραφής λόγω της εξάλειψης πιθανών αποκρύψεων από άλλα μέρη του σώματος, παρέχοντας ground-truth πόζες υψηλής ακρίβειας για το σύνολο δεδομένων. Ως τελικός κοινός χώρος καταγραφής για όλα τα συστήματα ορίστηκε μια περιοχή περίπου $4m \times 4m$, έτσι ώστε τα σώματα των ηθοποιών να βρίσκονται τουλάχιστον εν μέρει στο οπτικό πεδίο των RGB-D αισθητήρων κατά τη διάρκεια των δραστηριοτήτων, παραμένοντας ταυτόχρονα στο ωφέλιμο εύρος καταγραφής βάθους. Οι κάμερες αυτές τοποθετήθηκαν στις 4 γωνίες της σκηνής σε σχήμα σταυρού. Η κάτοψη της διάταξης απεικονίζεται στο Σχήμα 4.2.

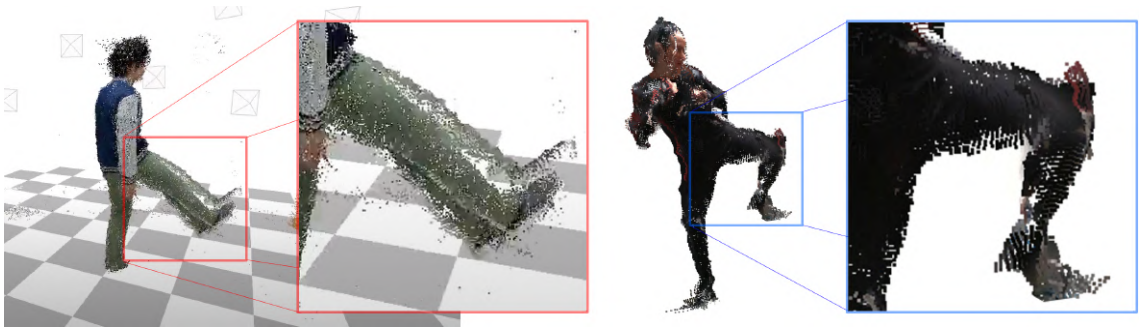
Στη συνέχεια, περιγράφονται λεπτομερώς οι μέθοδοι καταγραφής και οι τεχνικές που εφαρμόστηκαν για τη λήψη του συνόλου δεδομένων.

4.1.2 Καταγραφή χρώματος-βάθους πολλαπλών προβολών

Το HUMAN4D είναι το πρώτο σύνολο δεδομένων που προσέφερε συγχρονισμένα RGB-D δεδομένα πολλαπλών προβολών με χρήση συγχρονισμού υλισμικού. Τα περισσότερα από τα υπάρχοντα σύνολα δεδομένων χρησιμοποιούν RGB κάμερες [74] ή παλαιότερες εκδόσεις

[1] http://artanim.ch/

του αισθητήρα Microsoft Kinect για την καταγραφή RGB-D [2], οι οποίες δεν υποστηρίζουν συγχρονισμό υλισμικού, οπότε καταγράφουν σε διαφορετικές χρονικές στιγμές, απαιτώντας λύσεις συγχρονισμού βασισμένες σε αλγορίθμους για την ομαδοποίηση των καρέ με βάση τη χρονοσήμανση τους, μη προσφέροντας ωστόσο την ίδια ακρίβεια. Στο HUMAN4D, χρησιμοποιούμε αντ΄ αυτών τον αισθητήρα Intel RealSense D415, ο οποίος προσφέρει αυτή τη λειτουργικότητα [91]. Οι αισθητήρες D415, όταν συνδεθούν κατάλληλα μεταξύ τους, μπορούν να ρυθμιστούν σε λειτουργία συγχρονισμού "master", ώστε να αποστέλλουν το σήμα καταγραφής, είτε σε λειτουργία "subordinate" για τη λήψη του. Για την καταγραφή του HUMAN4D, μία συσκευή ορίστηκε ως master, παρέχοντας το σήμα συγχρονισμού, και οι υπόλοιπες 3 ως subordinate, σηματοδοτώντας την κοινή στιγμή καταγραφής και εξαλείφοντας την ανάγκη εξωτερικής σηματοδότησης. Η συνεισφορά και σημασία της καταγραφής με



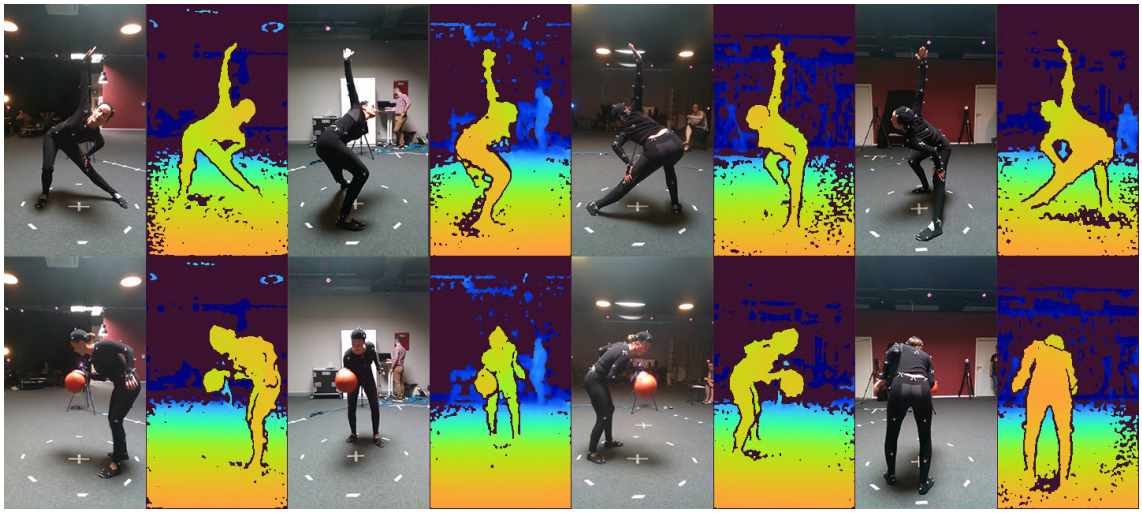
Σχήμα 4.3: *Έγχρωμα σημειακά νέφη από το CMU σύνολο [2] (Αριστερά) και το HUMAN4D (Δεξιά) παρουσιάζουν τα πλεονεκτήματα του συγχρονισμού υλισμικού. Στο CMU, όπου οι συσκευές Kinect for Xbox One έχουν τροποποιηθεί για σκοπούς συγχρονισμού, το πόδι του υποκειμένου αλλοιώνεται σε μια σχετικά αργή κίνηση (π.χ. απλή ανύψωση του ποδιού) λόγω της ύπαρξης χρονικών διαφορών μεταξύ των στιγμών καταγραφής των συσκευών. Στο HUMAN4D, το πόδι αποτυπώνεται κατάλληλα σε μια γρήγορη κίνηση (δηλ. γροθιές και κλωτσιές).*

συγχρονισμό ταυτόχρονης σηματοδότησης υλισμικού για την έρευνα και ανάπτυξη μεθόδων βασισμένων σε δεδομένα πολλαπλών προβολών, όπως για παράδειγμα η δημιουργία 3Δ δεδομένων επιφανείας ή υπολογισμού πόζας και κίνησης απεικονίζεται στο Σχήμα 4.3. Εκεί, συγκρίνονται 3Δ δεδομένα (σημειακά νέφη) που συντίθενται με αξιοποίηση των συσχετισμένων RGB-D δεδομένων από τα HUMAN4D και CMU Datasets [2], επιδεικνύοντας την μακράν πιο ακριβή χωρική συσχέτιση στο χρόνο του HUMAN4D έναντι του CMU. Αξίζει να σημειωθεί ότι το CMU σύνολο αποτελούσε τη στιγμή της δημιουργίας του HUMAN4D το μοναδικό σύνολο δεδομένων που παρείχε συγχρονισμένους χάρτες βάθους τροποποιώντας τις συσκευές Kinect for Xbox One.

Όσον αφορά τη λήψη δεδομένων βάθους, οι αισθητήρες χρησιμοποιήθηκαν στη λειτουργία ‘υψηλής ακρίβειας¨, προσφέροντας μόνο τις εκτιμήσεις βάθους υψηλής εμπιστοσύνης, παράγοντας επομένως ακριβή αλλά σχετικά αραιά δεδομένα βάθους. Αξίζει να σημειωθεί ότι ρυθμίσαμε τους αισθητήρες εκμεταλλευόμενοι τις δυνατότητες χωρικού φιλτραρίσματος για την καλύτερη δυνατή ποιότητα καταγραφής βάθους. Τα δεδομένα καταγράφηκαν με τη χρήση του συστήματος καταγραφής[2] που προτάθηκε από τους Sterzentsenko κ.α. [92], ενώ η χωρική

[2] https://github.com/VCL3D/VolumetricCapture

συσχέτιση μεταξύ των αισθητήρων επετεύχθη με χρήση της μεθόδου χωρικής συσχέτισης πολλαπλών προβολών βάθους που προτάθηκε από τους Papachristou κ.α. [93]. Δείγματα RGB-D δεδομένων του συνόλου απεικονίζονται στο Σχήμα 4.4.



Σχήμα 4.4: *Δείγματα συγχρονισμένων RGB-D δεδομένων πολλαπλών προβολών (4 καρέ το καθένα) από τις δραστηριότητες "stretching_n_talking" (πάνω) και "basketball_dribbling" (κάτω). Οι χάρτες βάθους έχουν χρωματιστεί με τη χρήση του χρωματικού χάρτη TURBO [3].*

4.1.3 Καταγραφή κίνησης

Για τη λήψη των ground truth δεδομένων κίνησης, χρησιμοποιήθηκε ένα επαγγελματικό σύστημα καταγραφής. Το σύστημα αποτελούνταν από 24 κάμερες Vicon MXT40S (Vicon, Oxford Metrics, UK) με συχνότητα καταγραφής στα 120Hz. Κάθε ηθοποιός φορούσε μια ειδική στολή καταγραφής κίνησης με 53 προσαρτημένους ανακλαστικούς δείκτες, οι οποίοι οδηγούν στην εκτίμηση πόζας 33 αρθρώσεων. Το πυκνό σχετικά σύνολο δεικτών σε συνδυασμό με τον μεγάλο αριθμό καμερών κίνησης (24) μας επέτρεψε να καταγράψουμε δεδομένα υψηλής ακρίβειας για την εκπαίδευση, επαλήθευση και αξιολόγηση στατιστικών μοντέλων.

Για τους σκοπούς ορθής ρύθμισης και επαλήθευσης του συστήματος πριν τις καταγραφές, ζητήθηκε από κάθε ηθοποιό να εκτελέσει ένα πλήρες εύρος κινήσεων όλων των αρθρώσεων. Η διαδικασία αυτή εξασφάλισε ότι οι θέσεις των αρθρώσεων αντιστοιχούσαν σωστά στο σύνολο των εντοπισμένων δεικτών. Πριν από κάθε δραστηριότητα, ζητήθηκε από τους ηθοποιούς να ξεκινούν σε 'Στάση Τ' και στη συνέχεια να πραγματοποιούν τη δραστηριότητα που τους είχε ανατεθεί.

4.1.4 Συγχρονισμός και χωρική συσχέτιση

Ο συγχρονισμός μεταξύ των συστημάτων καταγραφής αποτελεί απαραίτητη προϋπόθεση για τη δημιουργία συνόλων δεδομένων πολλαπλών, διαφορετικού τύπου αισθητήρων. Οι κάμερες καταγραφής κίνησης λειτουργούν εξ' ορισμού με συγχρονισμό υλισμικού. Όσον αφορά τη ρύθμιση λήψης των δεδομένων χρώματος-βάθους, όπως ήδη αναφέραμε, οι αισθητήρες Intel

RealSense D415 προσφέρουν επίσης συγχρονισμό υλισμικού. Σχετικά με τον συγχρονισμό των διαφορετικών συστημάτων, επετεύχθη αλγοριθμικά ορίζοντας ως χρόνο αναφοράς το χρόνο καταγραφής του συστήματος κίνησης και λαμβάνοντας υπόψιν τη χρονοσήμανση των δεδομένων χρώματος-βάθους και ήχου. Συγκεκριμένα, δεδομένης της συχνότητας καταγραφής κίνησης ίσης με 120 Hz, το χρονικά πλησιέστερο δείγμα κίνησης σε κάθε χρονοσήμανση ομαδοποιημένων RGB-D καρέ θεωρήθηκε η αντίστοιχη πόζα, δίνοντας μια χαμηλή χρονική διαφορά t_d, όπου $t_d \leq \frac{1}{120}/2$ ms, $t_d \leq 4.16$ ms. Η αρχική χρονική διαφορά μεταξύ των συστημάτων εντοπίστηκε στην αρχή κάθε ακολουθίας με τη βοήθεια μιας κλακέτας εξοπλισμένης με 2 δείκτες, επιτρέποντας σε όλα τα συστήματα να καταγράψουν τη χρονική στιγμή του κλεισίματος της κατά την εκκίνηση της καταγραφής. Αναλυτικότερα, για τις ακολουθίες δεδομένων καταγραφής κίνησης, αναλύθηκαν τα σήματα 3Δ θέσης των δεικτών της κλακέτας για την ανίχνευση του κλεισίματός της με τον εντοπισμό της χρονικής στιγμής κατά την οποία η ευκλείδεια απόσταση μεταξύ των δεικτών ήταν η ελάχιστη. Για τα ηχητικά δεδομένα, το κρότος της κλακέτας προκάλεσε μια εύκολα ανιχνεύσιμη κορυφή στο ηχητικού σήματος, ενώ για τα δεδομένα RGB-D, ανιχνεύθηκε χειροκίνητα.

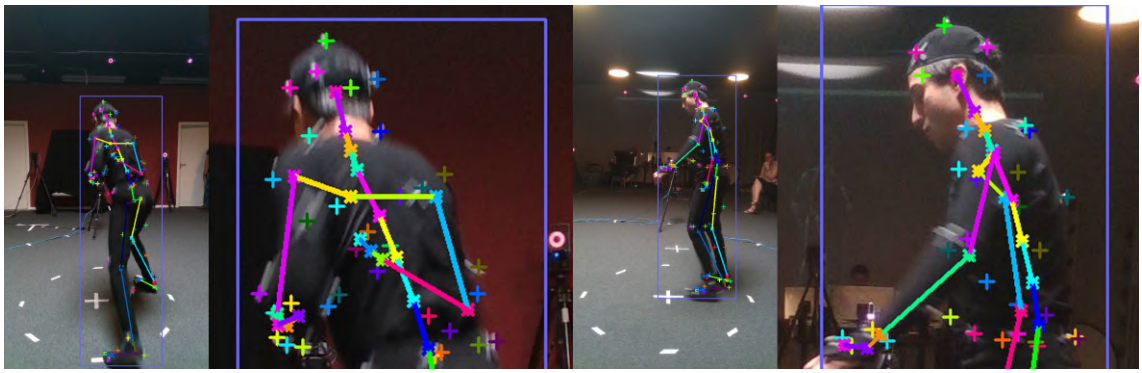
Για τη χωρική συσχέτιση των συστημάτων, το σύστημα κίνησης ρυθμίστηκε μία μονάχα φορά πριν από τις λήψεις, ενώ το σύστημα πολλαπλών RGB-D αισθητήρων επαναρυθμιζόταν ανά υποκείμενο (κάθε υποκείμενο εκτέλεσε όλες τις δραστηριότητες συνεχόμενα). Η χωρική ευθυγράμμιση μεταξύ των δύο αυτών συστημάτων επετεύχθη με την εφαρμογή μιας ημι-αυτόματης τεχνικής, καταγράφοντας σύντομες ακολουθίες δεικτών πριν από τη λήψη κάθε υποκειμένου. Για αυτές τις ακολουθίες, ενεργοποιήθηκε η υπέρυθρη ροή δεδομένων των αισθητήρων αντί της έγχρωμης. Λεπτομέρειες της χωρικής συσχέτισης μεταξύ των συστημάτων αναφέρονται στο Κεφ. 7.1.3.

4.1.5 Καταγραφή Ήχου

Η αξιοποίηση ήχου σε συνδυασμό με οπτικά δεδομένα έχει παρουσιάσει σημαντικά αποτελέσματα σε διάφορους ερευνητικούς τομείς, όπως η αναγνώριση ανθρώπινων συναισθημάτων [94], η ανάλυση σκηνών [95], η αναγνώριση ανθρώπινης δραστηριότητας [96] και άλλα. Για το σκοπό αυτό, επιθυμώντας την καταγραφή αλληλεπιδράσεων μεταξύ ατόμων, που σπανίως βρίσκεις σε δημόσια σύνολα δεδομένων, καταγράφηκαν δεδομένα ήχου κατά τη διάρκεια εκτέλεσης ορισμένων δραστηριοτήτων. Συγκεκριμένα, 6 δραστηρίοτητες (βλ. Πίνακα 4.1) περιλαμβάνουν ήχο, είτε ως μονόλογο (ενός ατόμου), είτε ως συνομιλία μεταξύ δύο ηθοποιών με βάση το σχετικό σενάριο. Όπως προαναφέρθηκε, για την καταγραφή χρησιμοποιήθηκαν ασύρματα μικρόφωνα που τοποθετήθηκαν στο σώμα για την καταγραφή των ηχητικών ακολουθιών. Η ηχογράφηση πραγματοποιήθηκε σε συχνότητα 48 kHz.

4.2 Επεξεργασία και επισήμανση συνόλου δεδομένων

Η χωροχρονική συσχέτιση μεταξύ των συστημάτων και η υψίσυχνη και ακριβής 3Δ καταγραφή κίνησης, επιτρέπουν την εξαγωγή ground truth δεδομένων που προβάλλονται με ακρίβεια στα RGB-D δεδομένα. Με ένα σύνολο $J = 33$ j-αρθρώσεων , μια 3Δ πόζα ανά καρέ f απεικονίζεται σε κάθε μεμονωμένο καρέ του αισθητήρα s. Στη συνέχεια, εφαρμόζο-



Σχήμα 4.5: *Δισδιάστατες προβολές υψηλής ακρίβειας των 3Δ θέσεων αρθρώσεων και δεικτών. Η ακρίβεια προβολής των δεικτών είναι σαφώς ορατή αναδεικνύοντας τη σημασία της χωροχρονικής συσχέτισης μεταξύ των 3Δ θέσεων των δεδομένων κίνησης και των RGB-D δεδομένων.*

ντας ανάστροφο μετασχηματισμό ανά θέση και περιστροφή κάμερας και προβάλλοντας τις 3Δ θέσεις των αρθρώσεων στα RGB-D δεδομένα, οι 2Δ συντεταγμένες $\mathcal{K}$ υπολογίζονται από την εξίσωση:

$$\mathcal{K}(f,s,j) = \pi(\mathbf{T}_{g\to l}(x_{f,s,j}), \mathbf{K}_s), \tag{4.1}$$

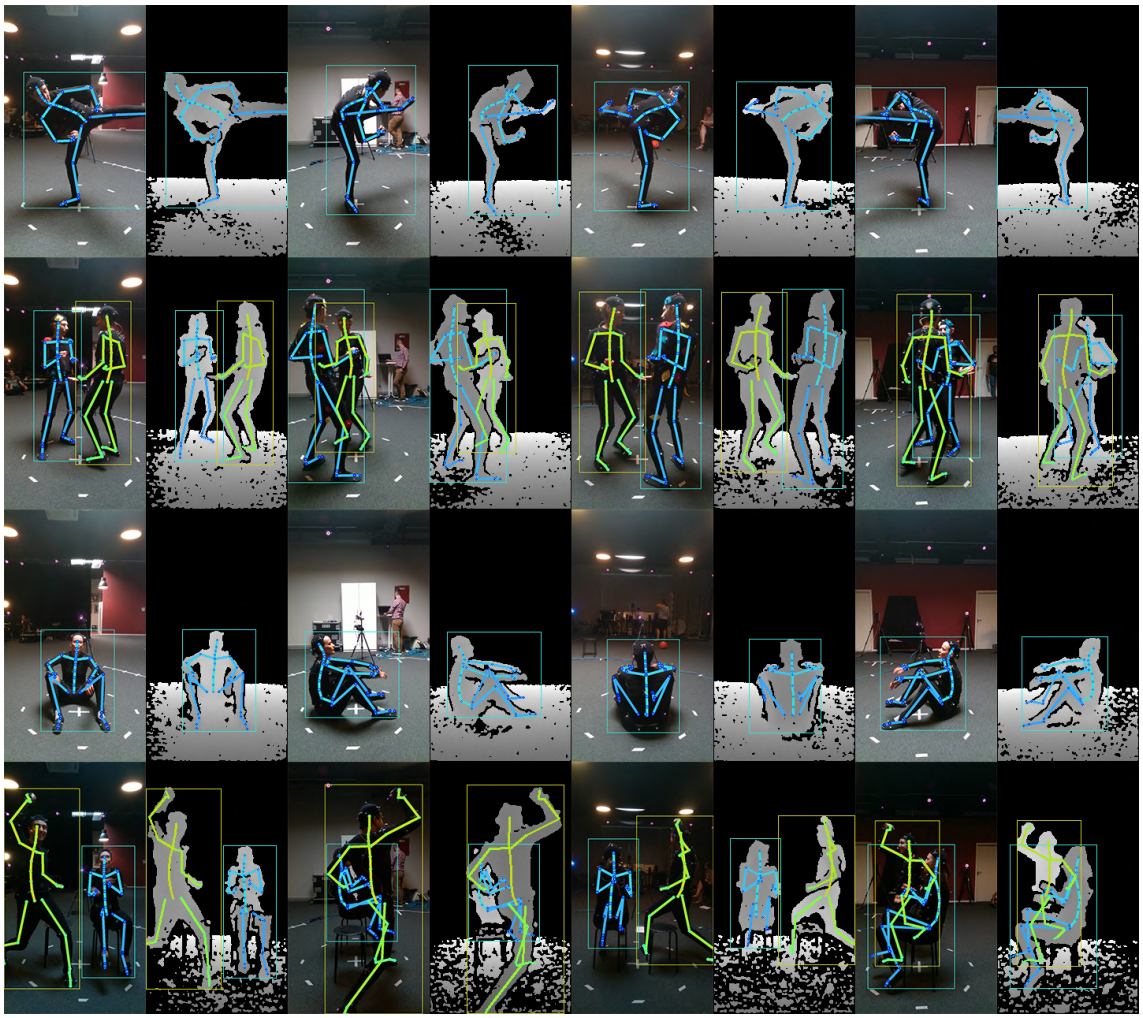
όπου $x_{f,s,j} \in \mathbb{R}^3$ είναι η 3Δ θέση της άρθρωσης j, $\mathbf{T}_{g\to l}$ είναι ο μετασχηματισμός από το καθολικό (g) σύστημα συντεταγμένων στο τοπικό (l) του αισθητήρα s με το βέλος να δείχνει την κατεύθυνση του μετασχηματισμού. Το π δηλώνει τη συνάρτηση προβολής που μετασχηματίζει τις 3Δ συντεταγμένες σε 2Δ σημεία πάνω στην εικόνα χρησιμοποιώντας τον πίνακα εσωτερικών παραμέτρων του αισθητήρα, $\mathbf{K_s}$. Τα αποτελέσματα αυτής της επεξεργασίας απεικονίζονται στα Σχήματα 4.5 και 4.6 .

Τέλος, λαμβάνοντας υπόψη τις 3Δ θέσεις των δεικτών και τις αντίστοιχες προβολές τους στο πεδίο θέασης του αισθητήρα (χρησιμοποιώντας την Εξ. (4.1)), εξάγουμε τα 3Δ και 2Δ πλαίσια που οριοθετούν τα άτομα ανά καρέ, προσαρμόζοντας ένα ορθογώνιο ελαφρώς μεγεθυμένο (2% του μεγέθους της διάστασης ανά πλευρά) πλαίσιο γύρω από τις 3Δ θέσεις και τις 2Δ προβολές, αντίστοιχα.

4.3 Στατιστικά

Για τη δημιουργία του συνόλου δεδομένων προσλήφθηκαν 4 επαγγελματίες ηθοποιοί, 2 γυναίκες και 2 άνδρες, επιδιώκοντας την υψηλότερη δυνατή ποιότητα των καταγραφών όσον αφορά την αυθεντικότητα των αναπαραστάσεων. Στο πλαίσιο του HUMAN4D, έγιναν διαθέσιμα τα εξής:

- Πολυτροπικά δεδομένα 5 δραστηριοτήτων δύο ατόμων και 14 ενός ατόμου (19 συνολικά), συμπεριλαμβανομένων φυσικών ασκήσεων, καθημερινών και κοινωνικών δραστηριοτήτων, περιλαμβάνοντας συνολικά 10 ακολουθίες δύο ατόμων και 56 ενός. Στον Πίνακα 4.1, παρουσιάζονται λεπτομέρειες σχετικά με τις δραστηριότητες του συνόλου.



Σχήμα 4.6: *Σημάνσεις πόζας και τετραγωνικού περιγράμματος σε κάποια δείγματα δεδομένων χρώματος και βάθους. Σε κάθε σειρά απεικονίζονται οι 4 διαφορετικές προβολές των καρέ από τυχαία δείγματα του συνόλου από διαφορετικές δραστηριότητες ενός και δύο ατόμων.*

- Παράμετροι προβολής και χωρικής συσχέτισης των καμερών που επιτρέπουν τη 2Δ προβολή των 3Δ δεδομένων στις διάφορες κάμερες και αντίστροφα.
- 12 ηχητικές καταγραφές για ορισμένες από τις δραστηριότητες όπου οι ηθοποιοί έπρεπε να μιλούν και να δρουν βάσει σεναρίων (βλ. Πίνακα 4.1).
- Υψηλής ακρίβειας συγχρονισμός μεταξύ των RGB-D αισθητήρων καθώς και συγχρονισμός μεταξύ των διαφορετικών συστημάτων καταγραφής δεδομένων με χρονοσήμανση.

4.4 Σκοπός δημιουργίας του συνόλου δεδομένων HUMAN4D

Στον κατακλυσμό της πρόσφατης βιβλιογραφίας, μια πληθώρα αλγορίθμων και μοντέλων βαθιάς μάθησης εστιάζουν στην εκτίμηση 3Δ πόζας, ωστόσο, μόνο λίγες μέθοδοι προσεγγίζουν το πρόβλημα με τη χρήση δεδομένων βάθους πολλαπλών όψεων. Αυτό πιθανώς ο-

Πίνακας 4.1: *Λεπτομέρειες σχετικά με τις φυσικές, καθημερινές και κοινωνικές δραστηριότητες του HUMAN4D.*

	activity	**# frames**	**audio**	**type**
Single-person	*running*	2,050	✗	physical
	jumping_jack	1,974	✗	physical
	bending	2,156	✗	physical
	punching_n_kicking	2,079	✗	physical
	basketball_dribbling	2,124	✗	physical
	laying_down	4,082	✗	physical
	sitting_down	3,288	✗	daily
	sitting_on_a_chair	2,797	✗	daily
	talking	2,377	✗	daily
	object_dropping_n_picking	1,768	✗	daily
	stretching_n_talking	2,787	✗	physical
	talking_n_walking	2,889	✗	daily
	watching_scary_movie	2,194	✗	daily
	in-flight_safety_announcement	6,192	✓	daily
Multi-person	*watching_football_together*	1,760	✓	social
	dancing_together	1,356	✓	social
	physical_examination	2,328	✓	social
	whispering	3,045	✓	social
	card_trick	3,060	✓	social
		50,306		

φείλεται στην πολυπλοκότητα ανάπτυξης και ρύθμισης των συστημάτων λήψης δεδομένων από πολλαπλούς αισθητήρες-κάμερες, και κατ΄ επέκταση, στην έλλειψη χωροχρονικά συσχετισμένων χαρτών βάθους με δεδομένα ground-truth. Για το σκοπό αυτό, μέσω του HUMAN4D, στοχεύουμε να καταστήσουμε δυνατή την έρευνα προς αυτή την κατεύθυνση ενθαρρύνοντας την ερευνητική κοινότητα να αναπτύξει και να πειραματιστεί με νέες προσεγγίσεις εκτίμησης 3Δ πόζας παρέχοντας υψηλής ακρίβειας συγχρονισμένα δεδομένα βάθους και χρώματος, μαζί με εξαιρετικά ακριβείς 3Δ ground-truth πόζες για ορθή εποπτεία και αξιολόγηση.

Πέραν από την προσφορά στην ερευνητική κοινότητα, το σύνολο HUMAN4D αποτέλεσε ακρογωνιαίο λίθο στην πραγματοποίηση της παρούσας διατριβής. Αρχικά, όπως θα περιγραφεί αναλυτικά και στη συνέχεια, το σύνολο αξιοποιήθηκε για την διερεύνηση σύγχρονων μοντέλων/μεθόδων εκτίμησης 2Δ και 3Δ πόζας (Κεφ. 5) με στόχο την αξιολόγηση και βαθύτερη κατανόηση της κάθε προσέγγισης σχετικά με την εκτίμηση συντεταγμένων στο χώρο, καθώς και με υπάρχουσες τεχνικές αποκωδικοποίηση συντεταγμένων (Κεφ. 6) που λειτουργούν επιτελικά πάνω στις εκτιμήσεις των μοντέλων βαθιάς μάθησης. Τέλος, το σύνολο δεδομένων πρακτικά καταγράφηκε εις διπλούν ως προς τις ακολουθίες που εκτέλεσαν τα υποκείμενα, μία αρχικά καταγράφοντας συγχρονισμένα και χωρικά συσχετισμένα δεδομένα χρώματος-βάθους και μια δεύτερη με δεδομένα έγχρωμης υπέρυθρης πληροφορίας και βάθους τα οποία μελετήθηκαν και αξιοποιήθηκαν για την ανάπτυξη μιας νέας μεθόδου καταγραφής κίνησης με δείκτες. Συγκεκριμένα, με τη χρήση των δεδομένων αυτών, και τη μελέτη πάνω σε υπάρχουσες μεθόδους εκτίμησης πόζας και αποκωδικοποίησης συντεταγμένων, πραγματοποιήθηκε η ανάπτυξη του πρώτου μοντέλου βαθιάς μάθησης που επιτρέπει τη χρήση αισθητήρων πολύ χαμηλού κόστους

και επαγγελματικών οπισθοανακλαστικών δεικτών, επιτυγχάνοντας ισάξια ποιοτικά καταγραφή κίνησης με τα επαγγελματικά συστήματα που χρησιμοποιούν τέτοιους δείκτες (Κεφ. 7).

Κεφάλαιο 5

Διερεύνηση και αξιολόγηση σύγχρονων μεθόδων εκτίμησης πόζας και καταγραφής κίνησης

Το σύνολο δεδομενων HUMAN4D το οποίο παρουσιάστηκε στο προηγούμενο κεφάλαιο (Κεφ. 4), μας επιτρέπει την έρευνα σε εργασίες υπολογιστικής όρασης που σχετίζονται με την ανθρώπινη ψηφιοποίηση (αναπαράσταση και κίνηση), παρέχοντας χωροχρονικά συσχετισμένα δεδομένα χαρτών χρώματος και βάθους από πολλαπλές προβολές, μαζί με ακριβείς 2Δ και 3Δ πόζες. Πολλές πρόσφατες έρευνες επικεντρώνονται σε προσεγγίσεις εκτίμησης πόζας ενός και πολλών ατόμων ταυτόχρονα σε "in-the-wild" δεδομένα χρώματος μονής προβολής [24, 97, 98, 99], βάθους [100, 101], χρώματος μόνο πολλαπλών προβολών [6, 102] και χρώματος-βάθους πολλαπλών προβολών [103, 104], μεταξύ άλλων. Στο κεφάλαιο αυτό διερευνούμε την αποτελεσματικότητα υπαρχόντων μοντέλων και τεχνικών σε εκτίμηση 2Δ και 3Δ δεδομένων πόζας με στόχο την καλύτερη κατανόηση των δυνατοτήτων αλλά και των περιορισμών τους. Τα βασικά κριτήρια επιλογής των μεθόδων που διερευνούμε παρακάτω είναι να είναι ανοικτού κώδικα και να είναι εφαρμόσιμες στο HUMAN4D.

5.1 Υποσύνολα Δεδομένων Αξιολόγησης του HUMAN4D

Για τη συγκριτική αξιολόγηση στο HUMAN4D, χωρίζουμε το σύνολο δεδομένων σε δύο υποσύνολα, ενός ατόμου (H4D1) και δύο ατόμων (H4D2), προκειμένου να μειώσουμε τον όγκο της επεξεργασίας των δεδομένων, ενώ ταυτόχρονα να αξιολογήσουμε τις μεθόδους με δείγματα που καλύπτουν ένα μεγάλο εύρος διαφορετικών ανθρώπινες στάσεων. Για το σκοπό αυτό, υποδειγματολειπτούμε τυχαία 100 καρέ από κάθε ακολουθία, παραλείποντας τα καρέ στην αρχή όπου τα υποκείμενα στέκονται σε T-Pose[1], καταλήγοντας σε συνολικά 5600 και 1000 καρέ ενός και πολλών ατόμων, αντίστοιχα.

5.2 Εκτίμηση 2Δ Πόζας Μονής Προβολής

Λαμβάνοντας υπόψη τις 2Δ πόζες ανά προβολή, αξιολογούμε σύγχρονες μεθόδους εκτίμησης πόζας από έγχρωμες εικόνες. Εφαρμόζουμε τις μεθόδους στις έγχρωμες προβολές και των 4 καμερών, εξάγοντας τις συνολικές μετρικές σφάλματος ανά καρέ ως το μέσο όρο των

[1]Το T-Pose αναφέρεται σε μια προεπιλεγμένη πόζα όπου ένας χαρακτήρας στέκεται με τα χέρια του οριζόντια τεντωμένα.

σφαλμάτων ανά προβολή.

5.2.1 Ερευνητικές Μέθοδοι

Επιλέγουμε δύο ευρέως γνωστές μεθόδους εκτίμησης 2Δ πόζας, μια bottom-up και μια top-down [105], για να αξιολογήσουμε την αποτελεσματικότητά τους στις έγχρωμες εικόνες του HUMAN4D. Αρχικά, επιλέγουμε την OpenPose των *Cao et al.* [4], μια bottom-up μέθοδο βαθιάς μάθησης που συνδυάζει χάρτες θερμότητας με διανυσματικά πεδία ιεραρχικής συσχέτισης μεταξύ των αρθρώσεων για την εκτίμηση 2Δ πόζας πολλών ατόμων σε πραγματικό χρόνο, όπως αναφέρθηκε ξανά στο Κεφ. 3. Στην παρούσα αξιολόγηση χρησιμοποιήσαμε την τελευταία κατά την φάση των πειραμάτων επίσημη υλοποίηση της μεθόδου[2]. Εν συνεχεία, αξιολογήσαμε τη μέθοδο AlphaPose, μια επίσης μέθοδο βαθιάς μάθησης που προτάθηκε από τους Fang *et al.* [5]. Η μέθοδος AlphaPose αποτελεί μια top-down μέθοδο εκτίμησης 2Δ πόζας πραγματικού χρόνου, η οποία συντηρείται και εξελίσσεται από τους συγγραφείς της συνεχώς τα τελευταία χρόνια. Κατά τη φάση των πειραμάτων, χρησιμοποιήσαμε την τελευταία έκδοση της μεθόδου, όπως αυτή είχε επίσημα δημοσιευτεί από τους ίδιους τους συγγραφείς[3].

Επιπρόσθετα, πειραματιστήκαμε με την επίσημη υλοποίηση του VNect[4], από τους Mehta *et al.* [20], ένα από τα πρώτα μοντέλα που προσέγγισε την εκτίμηση της 3Δ πόζας εξ' ολοκλήρου από εικόνες χρώματος, και τη μέθοδο A2j[5], από τους Xiong *et al.* [100], για την εκτίμηση 3Δ πόζας αποκλειστικά από χάρτες βάθους μονής προβολής. Ωστόσο, τα μοντέλα αυτά δεν είχαν ικανοποιητική εφαρμογή στο HUMAN4D, πιθανώς λόγω των διαφορών του με τα σύνολα που χρησιμοποιήθηκαν για την εκπαίδευση τους ως προς τα χαρακτηριστικά τους. Για το A2j για παράδειγμα, τα δεδομένα βάθους που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου έχουν ληφθεί με το αισθητήρα βάθους Asus Xtion PRO ο οποίος βασίζεται στην εκπομπή και αναγνώριση δομημένης υπέρυθρης ακτινοβολίας και καταγράφει χάρτες βάθους διαφορετικής ανάλυσης και θορύβου σε σύγκριση με το στερεοσκοπικό αισθητήρα βάθους της Intel, τον Intel RealSense D415. Ωστόσο, και αυτή η μελέτη αποδείχθηκε χρήσιμη καθώς συνέβαλε στην επιλογή των μοντέλων και των τεχνικών που χρησιμοποιήθηκαν στις μεθόδους που αναπτύχθηκαν στη συνέχεια της διατριβής.

5.2.2 Μετρικές

Παρόμοια με τις μετρικές που χρησιμοποιήσαμε στο DeepMoCap (βλ. Κεφ. 3.4), για τη μέτρηση της ακρίβειας εκτίμησης των συντεταγμένων των αρθρώσεων του σώματος, μετράμε τη μέση ακρίβεια (mAP) για τις κοινές αρθρώσεις μεταξύ των 2 μεθόδων[6] σε σύγκριση με τα ground truth με χρήση της μετρικής Percentage of Correct Keypoints-head (*PCKh*) [106]. Η μετρική *PCKh* αποτελεί μια παραλλαγή του Percentage of Correct Keypoints (*PCK*) [107], ορίζοντας ένα κατώφλι α ως το ποσοστό του μήκους του τμήματος του κεφαλιού (από το λαιμό έως την κορυφή του κεφαλιού), αντί της μακράς πλευράς του πλαισίου που οριοθετεί το

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose/tree/b5bffe18a8021f5f3ed98f19441b658647d9a8c3

[3] https://github.com/MVIG-SJTU/AlphaPose/tree/a22d3d6047b05be6ed94567c520d2a20d28d0407

[4] http://gvv.mpi-inf.mpg.de/projects/VNect

[5] https://github.com/zhangboshen/A2J/tree/60b45312c5009b2053d014510c08806c2c91e950

[6] Τις αρθρώσεις που εντοπίζονται και από τις δύο μεθόδους και δεν υπάρχει απόκλιση μεταξύ τους ως προς την ακριβή θέση εντοπισμού.

υποκείμενο στην εικόνα, με στόχο να καταστήσει τη μετρική ανεξάρτητη από τη συγκεκριμένη πόζα του σώματος, την εκάστοτε άρθρωση και την προβολή στο φακό της κάμερας. Για το σκοπό αυτό, μια πρόβλεψη για ένα καρέ f και μια πόζα s θεωρείται σωστή εάν το ευκλείδειο σφάλμα της 2Δ απόστασης $\epsilon_{f,p}$ εμπίπτει σε μια κυκλική περιοχή εικονοστοιχείων γύρω από τις ground truth συντεταγμένες με ακτίνα $r = \alpha L_{head}$, δηλ:

$$PCKh(f,s,j) = \begin{cases} 1, & \epsilon_{f,s}(j) \leq \alpha L_{head} \\ 0, & \epsilon_{f,s}(j) > \alpha L_{head} \end{cases} \tag{5.1}$$

$$AP_{PCKh}(f,\mathbf{s}) = \frac{1}{\mathcal{J}_s} \sum_{j=1}^{\mathcal{J}_s} PCKh(f,s,j) \tag{5.2}$$

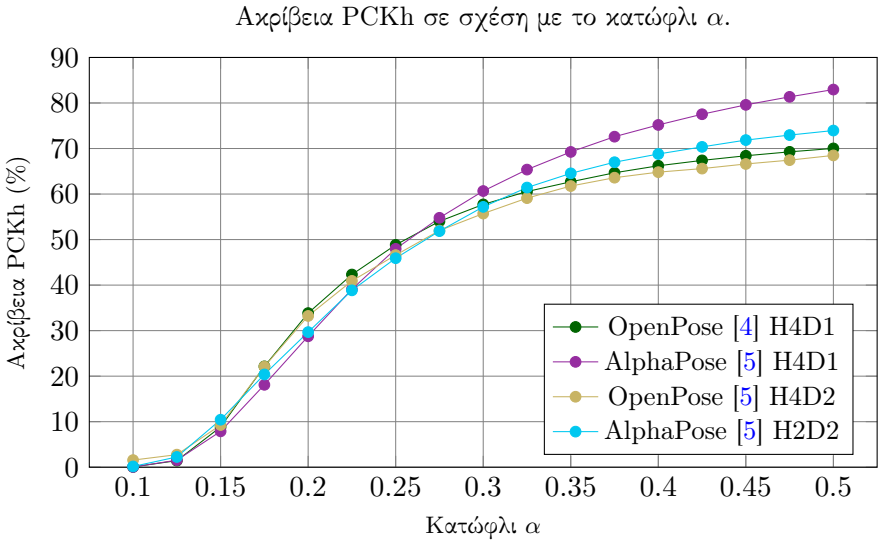
όπου, όπως προείπαμε, L_{head} είναι το μήκος του τμήματος της κεφαλής και α είναι το ποσοστιαίο κατώφλι που ελέγχει την ορθότητα της πρόβλεψης.

5.2.3 Αποτελέσματα

Παρουσιάζουμε ξεχωριστά τα αποτελέσματα των μεθόδων στα σύνολα H4D1 και H4D2 για να διακρίνουμε καλύτερα την αποτελεσματικότητά τους σε δεδομένα ενός και πολλών ατόμων, αντίστοιχα. Αρχικά, παρόμοια με τα αποτελέσματα σε άλλα δημόσια σύνολα δεδομένων, η AlphaPose υπερτερεί της OpenPose παρουσιάζοντας υψηλότερη ακρίβεια τόσο στα σύνολα αναφοράς ενός όσο και πολλών ατόμων του HUMAND. Ωστόσο, παρόλο που και οι δύο μέθοδοι παρουσιάζουν χαμηλότερη ακρίβεια στα δεδομένα πολλαπλών υποκειμένων του H4D2, τα οποία είναι πολύ πιο απαιτητικά λόγω των αποκρύψεων μεταξύ των σωμάτων των υποκειμένων, αξίζει να σημειωθεί ότι η διαφορά μεταξύ των αποτελεσμάτων ενός και πολλαπλών ατόμων της OpenPose μεθόδου είναι πιο περιορισμένη ($\sim 1.5\%$), ενώ η AlphaPose παρουσιάζει αισθητή πτώση, περίπου 9%. Λαμβάνοντας υπόψη ότι η απόσταση μεταξύ των υποκειμένων και των αισθητήρων είναι μικρή, από 1 έως 2 μέτρα, και στα περισσότερα δείγματα δύο ατόμων υπάρχουν έντονες αποκρύψεις για ορισμένους από τους αισθητήρες, μπορούμε πιθανώς να υποθέσουμε ότι το OpenPose, ως "bottom-up" προσέγγιση, συμπεριφέρεται πιο 'σταθερά' στις αποκρύψεις, ωστόσο η AlphaPose, ως "top-bottom", είναι γενικά πιο ακριβής αλλά επηρεάζεται πιο έντονα από τις τελευταίες. Για να παρέχουμε επιπλέον πληροφορίες στον αναγνώστη, μαζί με τα αποτελέσματα στο HUMAN4D, αναφέρουμε επίσης τα σχετικά αποτελέσματα των μεθόδων σε άλλα σύνολα δεδομένων, τα MPII [78] και COCO [85] χρησιμοποιώντας τη μετρική *PCKh* με $\alpha = 0.5$, όπως παρουσιάζονται στον πίνακα 5.1. Τέλος, η γραφική παράσταση στο Σχήμα 5.1 απεικονίζει τη συσχέτιση μεταξύ του *PCKh* mAP σε σχέση με το κατώφλι α και για τις δύο μεθόδους και για τα δύο υποσύνολα.

Πίνακας 5.1: *Αποτελέσματα εκτίμησης 2Δ πόζας των μεθόδων OpenPose [4] και AlphaPose [5] με χρήση της μετρικής* $AP_{PCKh\text{-}0.5}$.

mAP (%)	**MPII [78]**	**COCO [85]**	**H4D1**	**H4D2**
Cao *et al.* OpenPose [4]	72.50	64.20	**70.02**	**68.48**
Fang *et al.* AlphaPose [5]	82.10	71.00	**82.95**	**73.94**



Σχήμα 5.1: *Οι μέθοδοι OpenPose [4] και AlphaPose [5] εφαρμόζονται στις 4 προβολές των καμερών στα υποσύνολα H4D1 και H4D2, εξάγοντας τις συνολικές μετρικές σφάλματος ανά καρέ πολλαπλών προβολών με μέσο όρο σφαλμάτων ανά άρθρωση.*

5.3 Εκτίμηση 3Δ Πόζας Πολλαπλών Προβολών

Στη συνέχεια, αξιολογούμε την εκτίμηση 3Δ πόζας πολλαπλών προβολών στο HUMAN4D, αξιοποιώντας τις έγχρωμες εικόνες και τις αντίστοιχες εσωτερικές και εξωτερικές παραμέτρους του φακού των καμερών και συγκρίνοντας με τις αντίστοιχες ground truth 3Δ πόζες του.

5.3.1 Ερευνητικές Μέθοδοι

Επιλέγουμε μια πρόσφατη μέθοδο που προτάθηκε από τους Iskakov *et al.* [6]. Η μέθοδος βασίζεται σε μια τεχνική διαφορίσιμης τριγωνοποίησης (Learnable Triangulation), συνδυάζοντας 3Δ πληροφορία από πολλαπλές χωροχρονικά συσχετισμένες 2Δ έγχρωμες προβολές. Ειδικότερα, η $\text{LT}_{(\text{alg.})}$ [6] είναι μια ενιαία top-bottom μέθοδος εκτίμησης 3Δ διαφορίσιμης τριγωνοποίησης με την προσθήκη βαρών εμπιστοσύνης ανά εκτίμηση συντεταγμένων άρθρωσης. Πραγματοποιήσαμε τα πειράματα μόνο στο υποσύνολο αναφοράς HD41, καθώς η μέθοδος εκτιμά 3Δ πόζες ενός ατόμου, χρησιμοποιώντας την τελευταία επίσημη δημοσιευμένη υλοποίηση από τους συγγραφείς κατά τον καιρό των πειραμάτων[7].

5.3.2 Μετρικές

Όσον αφορά τις μετρικές, χρησιμοποιούμε τις μετρικές σφάλματος Mean Per Joint Position (MPJP) [20] και Root Mean Squared Per Joint Position (RMSPJP), οι οποίες αν και οι δύο επηρεάζονται από μεγάλες ακραίες τιμές, η τελευταία ενσωματώνει καλύτερα τη διακύμανση των εκτιμήσεων και την πόλωσή τους. Για ένα καρέ f και έναν σκελετό s, οι MPJP και RMSPJP υπολογίζονται ως εξής:

$$\epsilon_{f,s}(j) = ||\hat{x}_{f,s}(j) - x_{f,s}(j))||_2 \tag{5.3}$$

[7] https://github.com/karfly/learnable-triangulation-pytorch

$$\mathcal{E}_{MPJP}(f,s) = \frac{1}{\mathcal{J}_s}\sum_{j=1}^{\mathcal{J}_s}\epsilon_{f,s}(j) \tag{5.4}$$

$$\mathcal{E}_{RMSPJP}(f,s) = \sqrt{\frac{1}{\mathcal{J}_s}\sum_{j=1}^{\mathcal{J}_s}\epsilon_{f,s}^2(j)} \tag{5.5}$$

όπου $\mathcal{J}_s$ είναι ο συνολικός αριθμός των αρθρώσεων του σκελετού s. Τέλος, χρησιμοποιούμε επίσης τη μέση mAP με τη μετρική 3D PCK [108] ανά άρθρωση, όπου μια εκτίμηση θεωρείται σωστή όταν το σφάλμα της 3Δ ευκλείδειας απόστασης, δηλαδή του $\epsilon_{f,s}(j)$, είναι μικρότερο από ένα κατώφλι απόστασης α_{3D}, ως εξής:

$$PCK_{3D}(f,s,j) = \begin{cases} 1, & \epsilon_{f,s}(j) \leq \alpha_{3D} \\ 0, & \epsilon_{f,s}(j) > \alpha_{3D} \end{cases} \tag{5.6}$$

$$AP_{PCK_{3D}}(f,s) = \frac{1}{\mathcal{J}_s}\sum_{j=1}^{\mathcal{J}_s} PCK_{3D}(f,s,j) \tag{5.7}$$

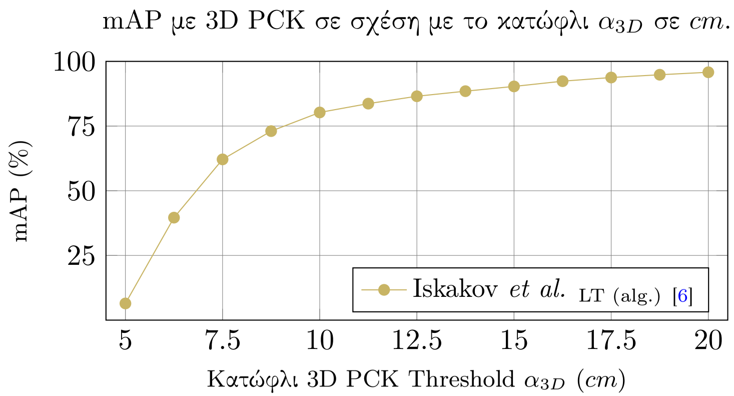
για ένα καρέ f και μια πόζα s, αντίστοιχα.

5.3.3 Αποτελέσματα

Οι κλασικοί αλγόριθμοι τριγωνοποίησης υποθέτουν ότι οι 2Δ συντεταγμένες σημείων από κάθε προβολή συμβάλλουν εξίσου στην εκτίμηση των 3Δ συντεταγμένων σημείων τριγωνοποίησης. Το σημαντικό πλεονέκτημα της μεθόδου $\text{LT}_{\text{(alg.)}}$ είναι ότι η συμβολή των 2Δ εκτιμήσεων των αρθρώσεων που δεν μπορούν να εκτιμηθούν αξιόπιστα (π.χ. λόγω απόκρυψης των αρθρώσεων ή άλλου σφάλματος) στο τελικό αποτέλεσμα της τριγωνοποίησης, ελέγχεται από ένα βαθύ νευρωνικό δίκτυο. Συγκεκριμένα, στους συντελεστές του πίνακα μετασχηματισμού της κάθε προβολής εφαρμόζονται μαθησιακά βάρη. Ένας αναμενόμενος περιορισμός της μεθόδου είναι ότι αποτυγχάνει όταν ορισμένα μέρη του σώματος βρίσκονται εκτός του οπτικού πεδίου των καμερών, οδηγώντας σε εσφαλμένες εκτιμήσεις. Τα ποσοτικά αποτελέσματα της μεθόδου στο H4D1, αλλά και επιπλέον αποτελέσματα στο σύνολο δεδομένων CMU [2], παρουσιάζονται στον Πίνακα 5.2. Το Σχήμα 5.2 απεικονίζει τη συσχέτιση μεταξύ του mAP έναντι του κατωφλίου α_{3D}. Ποιοτικά αποτελέσματα σχετικά με τις εκτιμώμενες 3Δ πόζες έναντι των ground truth του HUMAN4D απεικονίζονται στο Σχ. 5.3, όπου η μέθοδος $\text{LT}_{\text{(alg.)}}$ παρουσιάζεται ακριβείς σε 'ανοικτές' πόζες όπου οι επαφές και κατ' επέκταση οι αποκρύψεις μεταξύ μερών του σώματος είναι περιορισμένες (επιτυχείς περιπτώσεις στις επάνω σειρές), ενώ αντίθετα η ακρίβεια περιορίζεται παρουσία αποκρύψεων (αποτυχημένες περιπτώσεις στις κάτω σειρές).

Πίνακας 5.2: *Αποτελέσματα εκτίμησης ανθρώπινης πόζας ενός ατόμου στα σύνολα H4D1 και CMU[2].*

Datasets	**CMU**	**HUMAN4D (H4D1)**			
Metrics	**MPJP** (cm)	**MPJP** (cm)	**RMSPJP** (cm)	**mAP** (**PCK**$_{\alpha_{3D}=10cm}$)	**mAP** (**PCK**$_{\alpha_{3D}=12.5cm}$)
Iskakov *et. al* LT (alg.) [6]	2.13	8.42	9.56	80.26%	86.52%



Σχήμα 5.2: *Συγκριτική αξιολόγηση του Algebraic Learnable Triangulation [6] στο H4D1 με χρήση mAP 3D PCK έναντι του κατωφλίου α_{3D} σε cm.*

Σχήμα 5.3: Ποιοτικά αποτελέσματα της διαφορίσιμης τριγωνοποίησης (alg.) που προτάθηκε από τους Iskakov et al. [6]. Η επάνω και η κάτω σειρά απεικονίζουν επιτυχείς και εσφαλμένες προβλέψεις, αντίστοιχα. Οι ***μπλε*** *και* ***κόκκινες*** *αναπαραστάσεις αντιστοιχούν σε* ***ground truth*** *και* ***εκτιμώμενες*** *πόζες.*

5.4 Συμπεράσματα

Από την ανάλυση που πραγματοποιήθηκε στο παρόν κεφάλαιο, καταλήγουμε αρχικά στο ότι οι FCN 2Δ αρχιτεκτονικές σε συνδυασμό με κατάλληλες μεθόδους αποκωδικοποίησης συντεταγμένων από χάρτες θερμότητας είναι αξιοσημείωτα αποτελεσματικές για εκτίμηση 2Δ ανθρώπινης πόζας. Επίσης, με την κατάλληλη διαχείριση ανάλογα με το πρόβλημα προς επίλυση, οι χάρτες θερμότητας μπορούν να θεωρηθούν εκφραστικοί και εφαρμόσιμοι και για αποκωδικοποίηση 3Δ συντεταγμένων. Τέλος, με βάση τη αποτελεσματικότερη συμπεριφορά των "top-bottom" μεθόδων, μπορούμε να θεωρήσουμε ότι οι εκτιμήσεις είναι ακριβέστερες όταν οι είσοδος του μοντέλου είναι επικεντρωμένη στην περιοχή που βρίσκεται το υποκείμενο προς εντοπισμό. Εκμεταλλευόμαστε όλες τις παραπάνω παρατηρήσεις και τις αξιοποιούμε κατάλληλα για την ανάπτυξη μιας αποτελεσματικής μεθόδου καταγραφής κίνησης στο Κεφ. 7.

Κεφάλαιο 6

Διερεύνηση αναπαράστασης συντεταγμένων

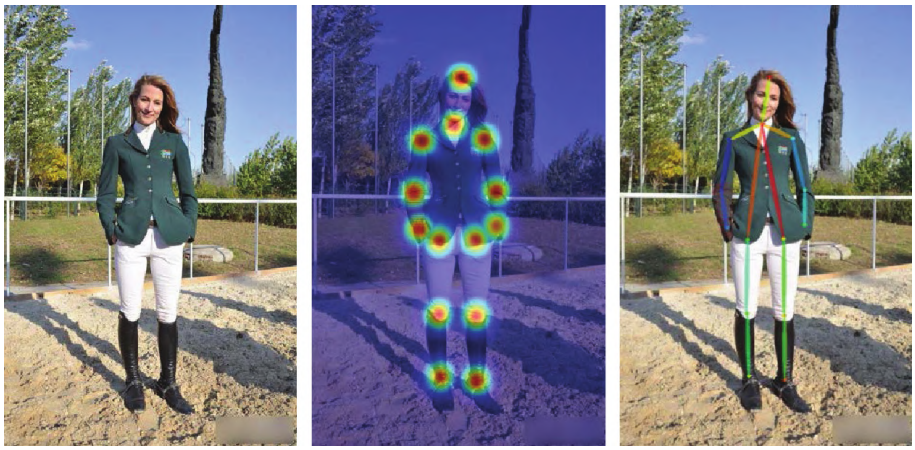
Όπως έχουμε ήδη αναφέρει και χρησιμοποιήσει στην παρούσα διδακτορική διατριβή σε προηγούμενα κεφάλαια, μια σειρά από 2Δ και 3Δ ερευνητικές προσεγγίσεις εκτίμησης συντεταγμένων βασίζονται στην αναπαράστασή τους ως ενός χάρτη πιθανότητας (ή αλλιώς χάρτη θερμότητας) που επιτρέπει την εκμάθηση και τη χωρική κωδικοποίηση και αποκωδικοποίηση των συντεταγμένων σε 2Δ πλέγματα/χάρτες (Εικόνα 6.1). Δεδομένου ότι ανάλογα με τη μέθοδο κωδικοποίησης-αποκωδικοποίησης, διαφέρει και η ακρίβεια εκτίμησης, στο παρόν κεφάλαιο, στοχεύουμε να διερευνήσουμε την αναπαράσταση 2Δ χαρτών πιθανοτήτων αναδεικνύοντας τη σημασία της κωδικοποίησης του ground truth χάρτη πιθανότητας και της αποκωδικοποίησης του προβλεπόμενου χάρτη πιθανότητας σε 2Δ συντεταγμένες.

Η εκτίμηση σημειακών συντεταγμένων σε N-διάστατα πλέγματα είναι ένα ευρέως διαδεδομένο ζητούμενο σε διάφορους τομείς της υπολογιστικής όρασης όπως η εκτίμηση ανθρώπινης πόζας και πόζας αντικειμένου, η παρακολούθηση των χαρακτηριστικών και των κινήσεων του προσώπου, η ανάλυση σκηνών πλήθους και πολλοί άλλοι. Για τις σύγχρονες μεθόδους μηχανικής μάθησης, η αναπαράσταση θερμικού χάρτη, δηλαδή μια N-διάστατη κατανομή πιθανοτήτων (Γκαουσιανή στις περισσότερες περιπτώσεις) με κέντρο τις συντεταγμένες κάθε N-διάστατου σημείου, θεωρείται η de facto αναπαράσταση των συντεταγμένων του. Οι αναπαραστάσεις θερμικών χαρτών επιτρέπουν τη χρήση αποδοτικών N-διάστατων πλήρως συνελικτικών δικτύων, μαθαίνοντας πυκνά χωρικά χαρακτηριστικά γύρω από τη ground truth σημειακή θέση και προσφέροντας έναν πιο φυσικό χωρικά τρόπο εκτίμησης των συντεταγμένων.

Ωστόσο, παρά τη δημοτικότητα της αναπαράστασης συντεταγμένων με θερμικούς χάρτες, ελάχιστες ερευνητικές δουλειές [109] διερευνούν σε βάθος την αναπαράσταση θερμικών χαρτών. Παρακάτω, αξιολογούμε διάφορες τεχνικές *κωδικοποίησης συντεταγμένων σε θερμικό χάρτη* και *αποκωδικοποίησης συντεταγμένων από θερμικό χάρτη* που υπάρχουν στη βιβλιογραφία, και αξιολογούμε την κάθε προσέγγιση ώστε να οδηγηθούμε σε σαφή συμπεράσματα όσων αφορά τη χρήση θερμικών χαρτών στη συνέχεια της διατριβής.

6.1 Αρχιτεκτονικές Νευρωνικών Δικτύων

Αρχικά, επιλέγουμε και χρησιμοποιούμε ευρέως γνωστές αρχιτεκτονικές βαθιάς μάθησης για την πρόβλεψη θερμικών χαρτών προς εκπαίδευση ώστε να αναλύσουμε τα αποτελέσματα ανεξαρτήτως δικτύου. Συγκεκριμένα χρησιμοποιήσαμε για τα μοντέλα μας τις παρακάτω FCN αρχιτεκτονικές:



Σχήμα 6.1: *Η χρήση χαρτών θερμότητας για την αναπαράσταση (κωδικοποίηση και αποκωδικοποίηση) 2Δ συντεταγμένων πάνω σε εικόνες έχει αποδειχθεί ιδιαίτερα αποτελεσματική σε προβλήματα εκτίμησης 2Δ ανθρώπινης πόζας (η εικόνα χρησιμοποιήθηκε από την ερευνητική δουλειά [7] και δεν αποτελεί προϊόν της διατριβής).*

- ***Stacked Hourglass (SH) [97]***, με χάρτες χαρακτηριστικών ακμής 128 στοιχείων, 4 σταδίων και 18 επιπέδων εξόδου.
- ***HRNet [110]***, με χάρτες χαρακτηριστικών ακμής 32 στοιχείων, 4 σταδίων και 18 επιπέδων εξόδου.

6.2 Αναπαράσταση Συντεταγμένων

6.2.1 Αποκωδικοποίηση Συντεταγμένων

Θεωρούμε έναν θερμικό χάρτη πιθανότητας $\mathbf{H}^j$ που κωδικοποιεί τις 2Δ συντεταγμένες κάθε j άρθρωσης, όπως εκτιμάται από τα εκπαιδευμένα μοντέλα μας. Στα πειράματά μας, αποκωδικοποιούμε τις συντεταγμένες από το $\mathbf{H}^j$ με 4 διαφορετικές μεθόδους αποκωδικοποίησης. Στην περιγραφή που ακολουθεί, N_u και N_v είναι το πλάτος και το ύψος του χάρτη θερμότητας και $\mathbf{p} \in \Omega$ αντιπροσωπεύει μια 2Δ συντεταγμένη εικονοστοιχείου στο πεδίο του χάρτη θερμότητας Ω.

- ***ArgMax***: χρησιμοποιώντας απευθείας τις συντεταγμένες της θέσης m με τη μέγιστη τιμή ενεργοποίησης ως εξής:

$$m = (u, v)_j = \arg\max_{\mathbf{p} \in \Omega}(\mathbf{H}^j(\mathbf{p})) \tag{6.1}$$

- ***Standard***: μετατοπίζοντας ελαφρώς (sub-pixel shifting) τις εκτιμώμενες συντεταγμένες μεταξύ της μέγιστης ενεργοποίησης m και της δεύτερης μέγιστης ενεργοποίησης s κατά:

$$p = (u, v)_j = m + 0.25 \cdot \frac{s - m}{||s - m||_2} \tag{6.2}$$

Η μέθοδος αυτή αποτελεί μία από τις συνήθεις τεχνικές αποκωδικοποίησης συντεταγμένων που προβλέπει τη μέγιστη ενεργοποίηση με εμπειρική μετατόπιση υπο-εικονοστοιχείου

(0.25 pixel) προς τη δεύτερη μέγιστη ενεργοποίηση στο χώρο του χάρτη θερμότητας Ω.

- ***Taylor-expansion***: διερεύνηση της δομής κατανομής του προβλεπόμενου χάρτη θερμότητας για την εξαγωγή συμπερασμάτων σχετικά με την υποκείμενη μέγιστη ενεργοποίηση, υποθέτοντας ότι το προβλεπόμενο σήμα του χάρτη ακολουθεί Γκαουσιανή 2Δ κατανομή παρόμοια με αυτή του ground truth χάρτη κατά τη διάρκεια της επίβλεψης από:

$$\mu = (u, v)_j = m - (\mathcal{D}''(m))^{-1}\mathcal{D}'(m) \tag{6.3}$$

όπου $\mathcal{D}'(m)$ και $\mathcal{D}''(m)$ δηλώνουν την πρώτη και τη δεύτερη παράγωγο (Hessian) του προβλεπόμενου χάρτη πιθανότητας μετά την εφαρμογή λογαριθμικού μετασχηματισμού.

- ***CoM***: θεωρώντας τον χάρτη πιθανότητας $\mathbf{H}^j$ ως ένα 2Δ πλέγμα από μάζες σημείων που καθορίζονται από τις αντίστοιχες τιμές των εικονοστοιχείων του χάρτη, υπολογίζουμε ένα μοναδικό 2Δ σημείο ως κέντρο μάζας του πλέγματος κατά:

$$c = (u, v)_j = (\frac{1}{N_u}, \frac{1}{N_v}) \circ \sum_{\mathbf{p}\in\Omega} \mathbf{H}^j(\mathbf{p}) \cdot \mathbf{p} \tag{6.4}$$

όπου $\circ$ δηλώνει τον πολλαπλασιασμό κατά στοιχείο. Το CoM αποτελεί από τη φύση του μια τεχνική αποκωδικοποίησης συντεταγμένων με ακρίβεια υπο-εικονοστοιχείου.

6.2.2 Κωδικοποίηση Συντεταγμένων

Το δεύτερο σκέλος αφορά την κωδικοποίηση συντεταγμένων, όπου οι περισσότερες σύγχρονες μέθοδοι υπο-δειγματοληπτούν τις αρχικές εικόνες στην ανάλυση εισόδου του μοντέλου, κβαντίζοντας αντίστοιχα τις συντεταγμένες των 2Δ σημείων (***w/ quant***). Έτσι, οι ground truth θερμικοί χάρτες που ανακατασκευάζονται με βάση τις κβαντισμένες συντεταγμένες είναι μερικώς λανθασμένοι, οδηγώντας σε μη βέλτιστη εποπτεία και υποβαθμισμένη απόδοση του μοντέλου.

Αντ΄ αυτού, μελετούμε κατα πόσον η δημιουργία θερμικού χάρτη τοποθετώντας το κέντρο στην πραγματική, μη κβαντισμένη συντεταγμένη του 2Δ σημείου (***w/o quant***), ανεξάρτητα από την αρχιτεκτονική του δικτύου και τη μέθοδο αποκωδικοποίησης των συντεταγμένων, αποδίδει σταθερά καλύτερα.

6.2.3 Προετοιμασία Δεδομένων

Με τη χρήση των 3Δ δεδομένων πόζας/κίνησης του HUMAN4D, προβάλλουμε τις 3Δ θέσεις των αρθρώσεων στα πλέγματα του χάρτη βάθους για να λάβουμε τις 2Δ μη κβαντισμένες πόζες. Χρησιμοποιώντας το τμήμα του συνόλου δεδομένων HUMAN4D για ένα άτομο και παρόμοια τεχνική με αυτή που περιγράψαμε στο Κεφ. 5.1, δηλαδή αποκόπτοντας τα πρώτα 100 καρέ από κάθε ακολουθία για να εξαιρέσουμε τα καρέ όπου οι ηθοποιοί στέκονται σε T-Pose και υπο-δειγματοληπτώντας το υπόλοιπο τμήμα της ακολουθίας, δημιουργούμε ένα σύνολο 98.864 δειγμάτων μονής όψης 14 δραστηριοτήτων, συμπεριλαμβανομένων και των 4 χαρτών βάθους ανά καρέ πολλαπλών προβολών.

Χρησιμοποιούμε τα δεδομένα των 3 πρώτων ατόμων από όλες τις δραστηριότητες για την εκπαίδευση, ενώ χωρίζουμε τις δραστηριότητες που εκτελούνται από το τέταρτο άτομο (αθέατο) σε δύο μισά, για την επαλήθευση και τη δοκιμή. Αποφασίζουμε να χωρίσουμε τα δεδομένα με αυτόν τον τρόπο προκειμένου να αξιολογήσουμε τα μοντέλα σε ένα αθέατο υποκείμενο με διαφορετική δομή σώματος, διατηρώντας επίσης διαφορετικές δραστηριότητες μεταξύ των συνόλων επικύρωσης και δοκιμής για να παρέχουμε εμπεριστατωμένα συμπεράσματα όσον αφορά την απόδοση των μοντέλων. Τα σύνολα δεδομένων εκπαίδευσης, επαλήθευσης και δοκιμής αποτελούνται από 73.248, 12.236 και 13.380 δείγματα, αντίστοιχα.

Τέλος, πέραν της κεντρικής περικοπής στους χάρτες βάθους από 320×180 σε 320×160 ($10px$ από πάνω και κάτω) για την καλύτερη προσαρμογή στα μοντέλα πολλαπλών σταδίων, δεν εφαρμόστηκαν περαιτέρω βήματα προ-επεξεργασίας.

6.2.4 Υπερπαράμετροι

Όλα τα μοντέλα εκπαιδεύονται για 30 εποχές, επιλέγοντας για δοκιμή τα μοντέλα με τις καλύτερες επιδόσεις με βάση τις μετρικές στο δεδομένα επαλήθευσης σε όλες τις εποχές. Δεδομένου ότι στόχος μας είναι να αξιολογήσουμε την αναπαράσταση συντεταγμένων του χάρτη θερμότητας και όχι καθαρά την αποτελεσματικότητα των μοντέλων στην εκτίμηση της πόζας, χρησιμοποιούμε σταθερές και κοινές υπερπαραμέτρους εκπαίδευσης για τα διάφορα μοντέλα, χωρίς περαιτέρω διερεύνηση για βελτίωση. Για το σκοπό αυτό, τόσο για το *SH* όσο και για το *HRNet*, χρησιμοποιήσαμε τον βελτιστοποιητή **Adam** με ρυθμό μάθησης $2e-3$ και τιμές beta 0.9 και 0.999. Το seed παρέμεινε σταθερό κατά τη διάρκεια της εκπαίδευσης όλων των μοντέλων, ίσο με 1314, εξασφαλίζοντας την αναπαραγωγιμότητα. Τέλος, σχετικά με την συνάρτηση απώλειας, χρησιμοποιήσαμε τη συνάρτηση απώλειας MSE η οποία οδήγησε εύκολα όλα τα μοντέλα σε σύγκλιση.

6.2.5 Πειραματικό Πλαίσιο

Υλοποιήσαμε τα μοντέλα μας καθώς και τις λειτουργίες κωδικοποίησης και αποκωδικοποίησης των συντεταγμένων του χάρτη θερμότητας με τη χρήση ενός αρχείου ρυθμίσεων που επιτρέπει τον ορισμό και τη διαμόρφωση των διαφόρων στοιχείων και υπερπαραμέτρων[1]. Αναφέρουμε τις ακόλουθες τυπικές μετρικές εκτίμησης πόζας που χρησιμοποιήσαμε στα πειράματά μας:

RMSE: αντιπροσωπεύει την τετραγωνική ρίζα του μέσου των τετραγωνικών διαφορών μεταξύ των εκτιμώμενων συντεταγμένων και των ground truth συντεταγμένων σε εικονοστοιχεία.

PCKh: θεωρεί μια εκτιμώμενη σημειακή θέση σωστή μόνο εάν το σφάλμα ευκλείδειας απόστασης είναι μικρότερο από ένα ποσοστιαίο όριο α της ευκλείδειας απόστασης μεταξύ του τμήματος του σώματος του κεφαλιού, δηλαδή από τις θέσεις του λαιμού έως τις θέσεις των αρθρώσεων του κεφαλιού (Κεφ. 5.3.2). Στα πειράματά μας, χρησιμοποιούμε τα *PCKh-0.1* και *PCKh-0.5* με $\alpha = 10\%$ και $\alpha = 50\%$, αντίστοιχα.

[1] https://github.com/ai-in-motion/moai

6.3 Αποτελέσματα

Τα αποτελέσματα που παρουσιάζονται στον Πίνακα 6.1 μας επιτρέπουν να αξιολογήσουμε τα επιτεύγματα αποκωδικοποίησης και κωδικοποίησης συντεταγμένων κάθε μεθόδου.

Πίνακας 6.1: *Τα συνολικά αποτελέσματα PCKh-0.1, PCKh-0.5 και RMSE των μεθόδων κωδικοποίησης συντεταγμένων* ***SH_4S*** *και* ***HRNet_4S*** *στο HUMAN4D που εκπαιδεύτηκαν με τις μεθόδους* ***w/ quant*** *και* ***w/o quant****. Η τελική εκτίμηση των συντεταγμένων αξιολογείται με τη χρήση των μεθόδων αποκωδικοποίησης συντεταγμένων ArgMax, Standard, CoM και Taylor.*

Model + *Coordinate Decoding*	**PCKh-0.1** ↑	**PCKh-0.5** ↑	**RMSE** ↓
SH_4S (w/ quant) + *ArgMax*	35.38	68.14	4.12
SH_4S (w/o quant) + *ArgMax*	39.80	68.78	3.75
SH_4S (w/ quant) + *Standard*	36.37	68.36	4.07
SH_4S (w/o quant) + *Standard*	41.18	69.05	3.69
SH_4S (w/ quant) + *CoM*	37.77	70.35	3.87
SH_4S (w/o quant) + *CoM*	44.24	71.45	**3.33**
SH_4S (w/ quant) + *Taylor*	39.54	70.75	3.87
SH_4S (w/o quant) + *Taylor*	**45.17**	**71.65**	3.47
HRNet_4S (w/ quant) + *ArgMax*	37.29	68.10	3.93
HRNet_4S (w/o quant) + *ArgMax*	39.19	68.60	3.85
HRNet_4S (w/ quant) + *Standard*	38.48	68.35	3.87
HRNet_4S (w/o quant) + *Standard*	40.50	68.90	3.79
HRNet_4S (w/ quant) + *CoM*	33.95	68.38	3.94
HRNet_4S (w/o quant) + *CoM*	36.77	70.21	3.79
HRNet_4S (w/ quant) + *Taylor*	41.73	70.29	3.65
HRNet_4S (w/o quant) + *Taylor*	**44.83**	**71.52**	**3.55**

Όπως αναφέρθηκε παραπάνω, ο κύριος στόχος μας είναι να αναδείξουμε την αποτελεσματικότητα της αναπαράστασης θερμικού χάρτη με επίγνωση της κατανομής, αντί να αξιολογήσουμε την απόδοση και τη σύγκριση μεταξύ των μοντέλων. Παρόλα αυτά, στα πειράματά μας, ένα γενικό σχόλιο θα μπορούσε να είναι ότι το **SH_4S** αποδίδει καλύτερα από το **HRNet_4S** σχεδόν σε όλους τους συνδυασμούς.

6.3.1 Αποκωδικοποίηση συντεταγμένων με δυναμοσειρά Taylor

Αξιολογώντας τις διάφορες μεθόδους αποκωδικοποίησης συντεταγμένων που συμπεριλάβαμε στα πειράματά μας, όπως παρουσιάζονται στο Κεφ. 6.2.1, η αποκωδικοποίηση συντεταγμένων με δυναμοσειρά Taylor βελτιώνει την απόδοση εκτίμησης συντεταγμένων. Η θεώρηση του προβλεπόμενου χάρτη θερμότητας ως χάρτη πιθανότητας με γκαουσιανή κατανομή και η εφαρμογή της αποκωδικοποίησης συντεταγμένων με σειρά Taylor παρουσιάζει αξιοσημείωτα αποτελέσματα έναντι των υπόλοιπων μεθόδων. Συγκεκριμένα, όπως φαίνεται στον Πίνακα 6.1, η αποκωδικοποίηση Taylor αποδίδει καλύτερα σε όλα τα πειράματα για όλες τις μετρικές, με

μόνη εξαίρεση το σφάλμα RMSE στο πείραμα ***SH_4S (w/o quant) + CoM***. Εστιάζοντας στα αποτελέσματα του μοντέλου με την καλύτερη απόδοση, όπως φαίνεται στον Πίνακα 6.2, το CoM υπερισχύει έναντι του Taylor. Ειδικά στο RMSE, το οποίο είναι μια συνεχής μετρική, το CoM παρουσιάζει χαμηλότερα σφάλματα από το Taylor, ενώ, γενικά, η απόδοσή του είναι υψηλότερη κατά την αποκωδικοποίηση των θερμικών χαρτών που εκτιμώνται από το SH_4S. Αυτό σημαίνει ότι η Taylor παρουσιάζει μεγαλύτερη διακύμανση στις τιμές σφάλματος από το CoM, ένα εν μέρει αναμενόμενο εύρημα δεδομένης της φύσης τους. Η CoM χρησιμοποιεί ολόκληρο τον χάρτη θερμότητας για να καταλήξει στην τελική συντεταγμένη, ενώ η Taylor χρησιμοποιεί στατιστικές του χάρτη θερμότητας κυρίως τοπικά γύρω από τη μέγιστη ενεργοποίηση, όντας πιο ευαίσθητη στον θόρυβο του χάρτη θερμότητας. Κατά την εφαρμογή της Taylor, χρησιμοποιείται πρώτα ένα χαμηλοπερατό Γκαουσιανό φίλτρο, το οποίο αποτελεί συνήθη πρακτική πριν από την εφαρμογή της εξαγωγής παραγώγων, προκειμένου να αφαιρεθεί ο θόρυβος και να καταγραφεί μόνο η πληροφορία υψηλής συχνότητας του σήματος. Ωστόσο, αυτό μπορεί να μην ισχύει πάντα στις προβλέψεις δεδομένων εκτός κατανομής. Όταν οι προβλέψεις ακολουθούν τις υποκείμενες παραδοχές της Taylor, δηλαδή μιας 2Δ Γκαουσιανής κατανομής, τότε είναι πιο αποτελεσματική και πιο λεπτομερής η εκτίμηση. Στην περίπτωση θορύβου στην κατανομή, η πιο ολιστική φύση της CoM είναι πιο ισχυρή. Προφανώς, το πείραμα SH είναι μια τέτοια περίπτωση, αλλά παρ' όλα αυτά, η ακρίβεια του Taylor εξακολουθεί να προσφέρει μεγαλύτερη ακρίβεια.

Πίνακας 6.2: *Τα αποτελέσματα PCKh-0.1, PCKh-0.5 και RMSE του μοντέλου με την καλύτερη απόδοση (**SH_4S** (w/o quant)) με τη χρήση των μεθόδων αποκωδικοποίησης συντεταγμένων Taylor και CoM.*

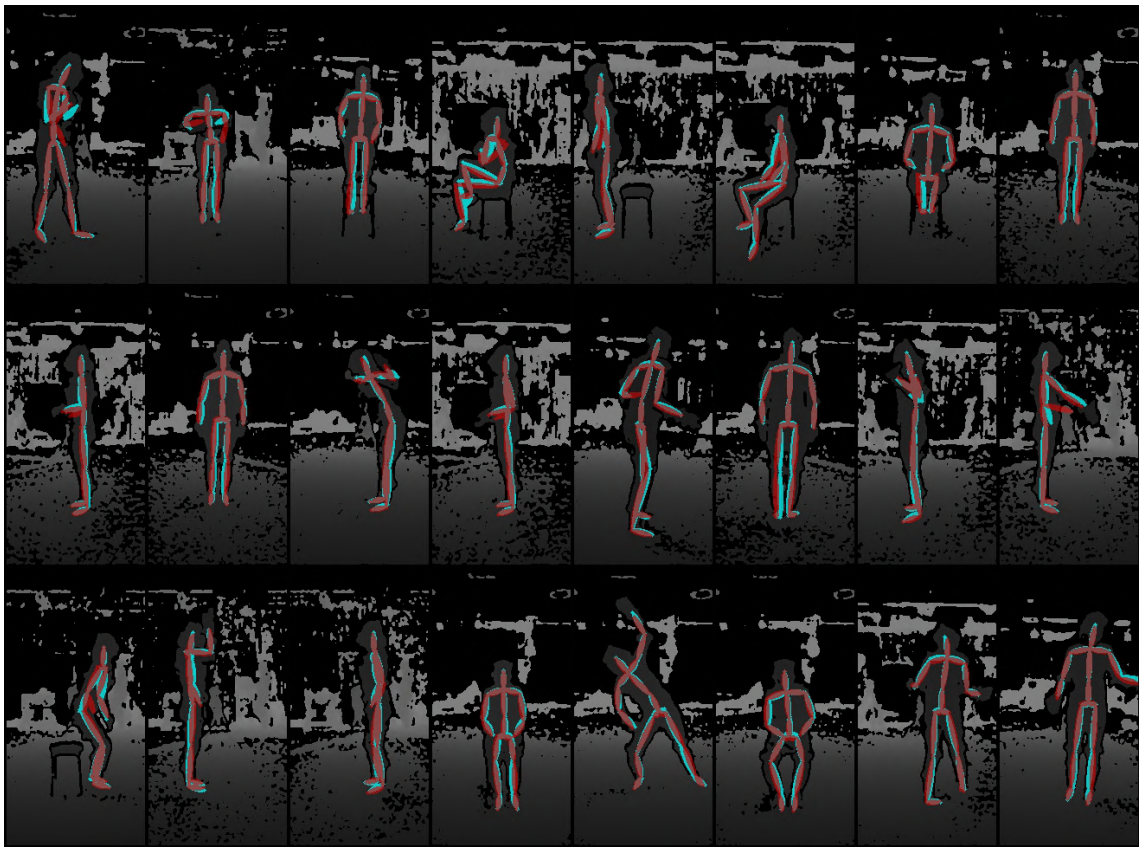
Model + *Coordinate Decoding*	**PCKh-0.1** ↑	**PCKh-0.5** ↑	**RMSE** ↓
SH_4S (w/o quant) + *CoM*	44.24	71.45	**3.33**
SH_4S (w/o quant) + *Taylor*	**45.17**	**71.65**	3.47

6.3.2 Μη κβαντισμένη κωδικοποίηση συντεταγμένων

Παρατηρώντας τώρα τα αποτελέσματα του Πίνακα 6.1 από μια άλλη οπτική γωνία, μπορούμε να συμπεράνουμε ότι η κβάντιση συντεταγμένων πριν από τη δημιουργία του χάρτη θερμότητας οδηγεί σε ανακριβή και πολωμένα ground truth δεδομένα που μειώνουν την απόδοση του μοντέλου. Πράγματι, και για τις δύο αρχιτεκτονικές, *SH* και *HRNet*, και για κάθε μέθοδο αποκωδικοποίησης συντεταγμένων, τα μοντέλα ***w/o quant*** παρουσιάζουν υψηλότερη ακρίβεια *PCKh* και χαμηλότερα σφάλματα *RMSE*, υποστηρίζοντας πλήρως αυτόν τον ισχυρισμό.

6.4 Συμπεράσματα

Συνοψίζοντας τη διερεύνηση που πραγματοποιήθηκε και παρουσιάστηκε στο παρόν κεφάλαιο, καταλήγουμε στο συμπέρασμα ότι η τεχνική Taylor εισάγει μια νέα, αποδοτική και ιδιαίτερα αποτελεσματική μέθοδο για την αποκωδικοποίηση συντεταγμένων θερμικού χάρτη που ενισχύει την ακρίβεια εκτίμησης σημείων στο N-διάστατο χώρο.



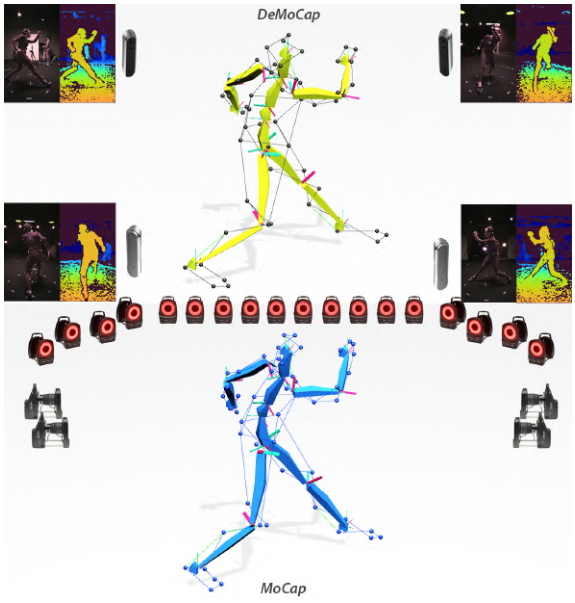
Σχήμα 6.2: *Ποιοτικά αποτελέσματα εκτίμησης θέσης με τη χρήση της αποκωδικοποίησης του χάρτη θερμότητας Taylor σε δοκιμαστικό σύνολο δεδομένων HUMAN4D (σε αθέατο υποκείμενο).* ***Κυανό*** *και* ***κόκκινο*** *χρώμα υποδεικνύουν τις πραγματικές και τις εκτιμώμενες πόζες, αντίστοιχα.*

Η Taylor υπερτερεί έναντι των τυποποιημένων ευριστικών μεθόδων αποκωδικοποίησης συντεταγμένων. Επιπλέον, παρουσιάζει υψηλότερη ακρίβεια από τις μεθόδους χωρικής παλινδρόμησης όπως η CoM (ή παραλλαγές όπως η MoM [70]), οι οποίες ωστόσο επίσης παρουσιάζουν χαμηλότερες εσφαλμένες προβλέψεις από τις τυπικές μεθόδους. Επιπλέον, η Taylor μπορεί να θεωρηθεί σημαντική στην εκτίμηση πόζας πολλαπλών ατόμων. Αυτό οφείλεται στο γεγονός ότι οι μέθοδοι χωρικής παλινδρόμησης, αν και εξαιρετικά αποτελεσματικές και πλήρως διαφοροποιήσιμες, δυσκολεύονται στην εκτίμηση περισσότερων του ενός σημείου όταν στο ίδιο θερμικό χάρτη, δηλαδή για μία συγκεκριμένη άρθρωση, υπάρχουν περισσότερα από ένα τοπικά μέγιστα στην ίδια πρόβλεψη, μια συνηθισμένη περίπτωση σε σύνολα δεδομένων πολλαπλών ατόμων.

Κεφάλαιο 7

DeMoCap: Χαμηλού κόστους καταγραφή ανθρώπινης κίνησης με χρήση βαθιάς μάθησης

Παρά την ύπαρξη αρκετών εμπορικών και ακαδημαϊκών εναλλακτικών λύσεων καταγραφής κίνησης, οι τεχνολογίες που βασίζονται σε δείκτες εξακολουθούν να παραμένουν το 'χρυσό πρότυπο' στον τομέα. Αυτό οφείλεται στην εξαιρετικά υψηλή ακρίβεια και συχνότητά, καθώς και στην εμπορική ωριμότητα για παραγωγή ψηφιακού υλικού που εξασφαλίζει αποτελέσματα υψηλής ποιότητας σε σύντομο χρονικό διάστημα.



Σχήμα 7.1: *Το DeMoCap αποτελεί την πρώτη MoCap μέθοδο βαθιάς μάθησης που βασίζεται σε δείκτες και ένα αραιό σύνολο αισθητήρων βάθους καταναλωτικής ποιότητας με πολύ χαμηλότερο κόστος και μεγαλύτερη φορητότητα και ευελιξία σε σχέση με τις εμπορικές λύσεις υψηλών προδιαγραφών.*

Παρ' όλα αυτά, η διαδικασία παραγωγής ψηφιακού υλικού των παραδοσιακών μεθόδων που χρησιμοποιούν δείκτες πάσχει από κάποια γνωστά μειονεκτήματα. Τα ακατέργαστα δεδομένα οπτικής καταγραφής κίνησης είναι συχνά λανθασμένα, λόγω απόκρυψης δεικτών ή εσφαλμένης σήμανσης από την εναλλαγή τους κατά τη διάρκεια της παρακολούθησης, με θόρυβο ή τρεμούλιασμα υψηλής συχνότητας και απαιτούν χρονοβόρα χειροκίνητη μετα-επεξεργασία. Πέραν αυτού, απαιτείται επιπλέον η προσαρμογή των αρθρωτών τμημάτων του σώματος σε υποσύνολα δεικτών και ο επαναπροσδιορισμός του σκελετού για την επίλυση του μετασχηματισμού των αρθρώσεων, καθιστώντας την αναμφίβολα μια επίπονη και χρονοβόρα διαδικασία. Συν τοις άλλοις, η πολυπλοκότητα και το κόστος των συστημάτων αυτών που απαιτούν την εγκατάσταση πολυάριθμων εξειδικευμένων υπερύθρων καμερών είναι υψηλά, καθιστώντας τα απρόσιτα στο ευρύτερο ενδιαφερόμενο κοινό.

Αυτές οι δυσκολίες προσελκύουν το ενδιαφέρον της ευρύτερης ερευνητικής κοινότητας για τη διερεύνηση και την πρόταση νέων λύσεων. Από τη μία πλευρά, οι ερευνητές ψηφιακών γραφικών εντείνουν τις προσπάθειές τους στην ανάπτυξη μοντέλων βαθιάς μάθησης που

επιλύουν ή αμβλύνουν αυτά τα ζητήματα [10, 64, 66, 67], ενώ οι ερευνητές υπολογιστικής όρασης επικεντρώνονται κυρίως σε πρωτοποριακές μεθόδους χωρίς δείκτες, θέτοντας το πρόβλημα καταγραφής κίνησης κυρίως ως έργο εκτίμησης 3Δ πόζας [6, 33, 34]. Παρ' όλα αυτά, οι υπάρχουσες λύσεις που βασίζονται σε δείκτες δεν ικανοποιούν την ανάγκη για ευέλικτες και χαμηλού κόστους επιλογές, ενώ τα πρόσφατα μοντέλα βαθιάς μάθησης χωρίς δείκτες, αν και είναι αποτελεσματικά και πιο ευέλικτα, δεν μπορούν να φτάσουν τα ίδια επίπεδα ευρωστίας και ακρίβειας ελλείψει ισχυρών και ντετερμινιστικών προκαθορισμών, όπως οι δείκτες που προσαρτώνται στο σώμα.

Συνδυάζοντας τα παραπάνω, αναπτύσσουμε μια 'υβριδική' προσέγγιση καταγραφής κίνησης. Σε αυτό το κεφάλαιο παρουσιάζουμε το DeMoCap, ένα καινοτόμο μοντέλο βαθιάς μάθησης με στόχο την καταγραφή κίνησης συνδυάζοντας τη χρήση δεικτών με χάρτες βάθους πολλαπλών προβολών από στερεοσκοπικούς αισθητήρες βάθους χαμηλού κόστους. Το DeMoCap είναι το πρώτο μοντέλο μηχανικής μάθησης που επιτρέπει τη χρήση επαγγελματικών σφαιρικών οπισθοανακλαστικών δεικτών και εξοπλισμού πολύ χαμηλότερου κόστους από τις υπάρχουσες εμπορικές λύσεις, επιτυγχάνοντας υψηλής ποιότητας καταγραφή 3Δ κίνησης.

Χρησιμοποιούμε μια χωροχρονικά συσχετισμένη διάταξη πολλαπλών αλλά ολιγάριθμων καμερών ανίχνευσης βάθους καταναλωτικής ποιότητας και χαμηλού κόστους για την παρακολούθηση ενός πυκνού συνόλου σφαιρικών οπισθοανακλαστικών δεικτών προσαρτημένους πάνω στο ανθρώπινο σώμα. Για να αντιμετωπίσουμε όλες τις προκλήσεις των μεθόδων με δείκτες, όπως την αποθορυβοποίησή τους, την ταυτοποίησή τους, την παρακολούθηση και την επίλυση του μετασχηματισμού των αρθρώσεων, καθώς και τους περιορισμούς της διάταξης αισθητήρων βάθους χαμηλού κόστους, δηλαδή το χαμηλό αριθμό των σημείων θέασης, τον θόρυβο των χαρτών βάθους του αισθητήρα, την πόλωση στα δεδομένα με βάση τις ρυθμίσεις καταγραφής και το θόλωμα (blurriness) της υπέρυθρης εικόνας για την ακριβή ανίχνευση των δεικτών, προτείνουμε ένα ενιαίο, πλήρως διαφοροποιήσιμο μοντέλο βαθιάς μάθησης για την εκτίμηση της 3Δ πόζας και της κίνησης. Με αυτόν τον τρόπο, θέτουμε το πρόβλημα καταγραφής κίνησης ως ένα πρόβλημα εκτίμησης πόζας από δείκτες στο χρόνο.

Παρόλο που τα δεδομένα εισόδου είναι 3Δ, αποφεύγουμε τη χρήση 3Δ συνελίξεων, καθώς στοχεύουμε σε εφαρμογές πραγματικού και σχεδόν πραγματικού χρόνου. Εισάγουμε μια νέα τεχνική χωρικής 3Δ παλινδρόμησης από θερμικούς χάρτες που εκτιμώνται από ένα 2Δ πλήρως συνελικτικό νευρωνικό δίκτυο για την αποκωδικοποίηση των 3Δ συντεταγμένων των δεικτών και των αρθρώσεων, με πλήρως διαφοροποιήσιμο τρόπο. Υλοποιούμε διαφορετικές FCN αρχιτεκτονικές πολλαπλών σταδίων, δημιουργώντας υπερ-στάδια, δηλαδή ομαδοποιώντας τα αρχικά και τα τελικά στάδια για την αποκωδικοποίηση των 3Δ συντεταγμένων των δεικτών και των αρθρώσεων, αντίστοιχα, ορίζοντας μια ομαλή μετάβαση αναπαράστασης από τους δείκτες στην 3Δ πόζα (markers-to-pose). Με αυτή την προσέγγιση, οδηγούμε το δίκτυο να μάθει τη χωρική και ιεραρχική σχέση μεταξύ των δεικτών και της υποκείμενης 3Δ πόζας.

Για την τροφοδοσία του δικτύου μας, εφαρμόζουμε μια ογκομετρική κανονικοποίηση για την ομοιόμορφη κατανομή του νέφους των δεικτών σε έναν κυβοειδή 3Δ χώρο για υψηλότερη χωρική ανεξαρτητοποίηση και πιο αραιή κατανομή των δεδομένων. Στη συνέχεια, εισάγουμε μια τεχνική απόδοσης πολλαπλών όψεων για την προβολή των δεικτών από αντίθετα τοποθετημένες ορθογραφικές κάμερες των οποίων ο κύριος άξονας διέρχεται από το κέντρο του αραιού νέφους των δεικτών. Ειδικότερα, διατηρούμε την 3Δ πληροφορία των δεικτών διαχωρίζοντας

το σχετικό τους βάθος σε πολλαπλές όψεις (δύο αντίθετες όψεις για το τελικό μοντέλο), δημιουργώντας 'αραιούς' χάρτες βάθους των δεικτών για να τροφοδοτήσουμε το μοντέλο μας. Από αυτό το σημείο, προσεγγίζουμε την εργασία μας ως ένα πρόβλημα εκτίμησης 3Δ πόζας από χάρτες βάθους διπλής όψης. Το μοντέλο οδηγείται στην αφομοίωση της αρθρωτής σχέσης μεταξύ των δεικτών και των αρθρώσεων, εκτιμώντας διαδοχικά τις 3Δ συντεταγμένες τους σε ένα πρόσθιο πέρασμα ανά πλαίσιο πολλαπλών προβολών υπέρυθρης-βάθους, χωρίς καμία προηγούμενη γνώση της δομής του σώματος ή ρητή συσχέτιση μεταξύ των τμημάτων του σώματος και των δεικτών. Συνοψίζοντας, οι συνεισφορές μας είναι οι εξής:

- Εξ' όσων γνωρίζουμε, το DeMoCap αποτελεί το πρώτο μοντέλο βαθιάς μάθησης που χρησιμοποιεί αποδοτικά πλήρως συνελικτικά νευρωνικά δίκτυα για την ταυτόχρονη παλινδρόμηση των οπτικών δεικτών και της 3Δ πόζας από αραιά σύνολα 3Δ σημείων, τα οποία καταγράφονται με τη χρήση μιας χαμηλού κόστους διάταξης αισθητήρων βάθους πολλαπλών όψεων.

- Για την εκπαίδευση του δικτύου μας εισάγεται μια αναπαράσταση ανεξάρτητη κλίμακας μεγέθους και μετατόπισης σε έναν κανονικοποιημένο 3Δ χώρο. Εκπαιδεύοντας το μοντέλο μέσω αυτής της αναπαράστασης, το μοντέλο ξεπερνά την πόλωση στα σχετικά περιορισμένα σε αριθμό δεδομένα εκπαίδευσης και γενικεύεται η εφαρμογή και η απόδοσή του καλύτερα.

- Πραγματοποιούμε μια ομαλή μετάβαση της αναπαράστασης από τους δείκτες στην 3Δ πόζα. Το μοντέλο μας οδηγείται στην εκμάθηση της υποκείμενης δομικής σχέσης μεταξύ του ανθρώπινου σώματος και της τοποθέτησης των δεικτών, αποκωδικοποιώντας τις συσχετίσεις δεικτών και αρθρώσεων από τα αραιά και θορυβώδη νέφη σημείων 3Δ δεικτών, με αποτέλεσμα την ακριβή εκτίμηση της πόζας.

- Θέτουμε την εκτίμηση 3Δ σημείων ως κοινό στόχο 2Δ εντοπισμού και παλινδρόμησης εντός του κανονικοποιημένου 3Δ χώρου μας για την έμμεση κωδικοποίηση της 3ης διάστασης z με την εισαγωγή μιας νέας πλήρως διαφοροποιήσιμης τεχνικής για 3Δ χωρική παλινδρόμηση.

- Διαθέτουμε δημόσια το ειδικό μας σύνολο δεδομένων που περιλαμβάνει δια- και ενδοσυστημικά χωροχρονικά συσχετισμένα δεδομένα υπέρυθρης-βάθους και δεδομένα καταγραφής κίνησης. Επιπλέον, τα πρώτα έχουν καταγραφεί με συγχρονισμό υλισμικού για ακριβή χρονική συνέπεια μεταξύ των πολλαπλών προβολών.

Το υπόλοιπο κεφάλαιο έχει την ακόλουθη δομή. Στο Κεφ. 7.1 παρουσιάζεται η δημιουργία και η επεξεργασία του συνόλου δεδομένων για να εξοικειωθεί ο αναγνώστης με τη φύση της πρόκλησής μας και τον τρόπο με τον οποίο την προσεγγίζουμε. Στο Κεφ. 7.2, συζητείται η μεθοδολογική μας προσέγγιση, παρουσιάζοντας και εξηγώντας σε βάθος το σκεπτικό πίσω από τις συνεισφορές της. Στο Κεφ. 7.3, παρουσιάζουμε ποσοτικά και ποιοτικά πειραματικά αποτελέσματα της μεθόδου συνολικά και αλλά και των συνεισφορών μας ξεχωριστά. Στο Κεφ. 7.4, συζητάμε τα πλεονεκτήματα και τα μειονεκτήματα του DeMoCap σε σύγκριση με τις υπάρχουσες λύσεις καταγραφής κίνησης υψηλού κόστους με δείκτες και τις πρόσφατες προσεγγίσεις εκτίμησης της πόζας χωρίς δείκτες.

7.1 Οπτικά δεδομένα δεικτών από δεδομένα βάθους

Δημιουργήσαμε ένα ειδικό σύνολο δεδομένων από χωροχρονικά συσχετισμένα δεδομένα καταγραφής κίνησης και δεδομένα βάθους υπερύθρων πολλαπλών όψεων για να εξυπηρετήσουμε το αντικείμενό μας (βλ. Κεφ. 4.4). Το σύνολο δεδομένων μας αποτελεί την πρώτη συλλογή οπτικών δεδομένων που περιέχει χωροχρονικά συσχετισμένες έγχρωμες υπέρυθρες εικόνες και χάρτες βάθους πολλαπλών προβολών με 3Δ δεδομένα πόζας/κίνησης και δεικτών. Καταγράψαμε διάφορες δραστηριότητες που εκτελούνται από ηθοποιούς με οπισθοανακλαστικούς δείκτες που είναι προσαρτημένοι στο σώμα τους, οι οποίοι είναι ευδιάκριτοι στις εικόνες υπερύθρων. Για το σκοπό αυτό, χρησιμοποιήσαμε τους υπέρυθρους πομπούς του αισθητήρα βάθους Intel RealSense D415 για να εκπέμπουν υπέρυθρο φως στη σκηνή προκαλώντας ανταναακλάσεις στους οπισθοανακλαστικούς δείκτες (βλ. Εικ. 7.2). Οι εικόνες υπερύθρων και βάθους, οι οποίες είναι ευθυγραμμισμένες και ορίζονται στο ίδιο πεδίο 2Δ εικόνας, επιτρέπουν τον 3Δ εντοπισμό των δεικτών μίας όψης, ο οποίος εξετάζεται στο κεφάλαιο 7.1.2.

Επιτυγχάνουμε 3Δ επισήμανση δεικτών και εκτίμηση ανθρώπινης πόζας εφαρμόζοντας ακριβή συγχρονισμό και χωρική συσχέτιση μεταξύ ενός επαγγελματικού συστήματος καταγραφής κίνησης και ενός συστήματος αισθητήρων βάθους πολλαπλών όψεων, που περιγράφεται λεπτομερώς στο Κεφ. 7.1.3.

Όπως και στο HUMAN4D που καταγράφηκε ακριβώς με το ίδιο πρωτόκολλο αλλά με δεδομένα χρώματος βάθους, όχι υπέρυθρα, (Κεφ. 4.3), καταγράψαμε 4 ηθοποιούς, 2 άνδρες και 2 γυναίκες, και συμπεριλάβαμε 11 διαφορετικές δραστηριότητες διάρκειας περίπου 20 δευτερόλεπτα η κάθε μία. Συνολικά, περισσότερα από 20.000 δείγματα περιλαμβάνονται στο σύνολο δεδομένων μας, ωστόσο λεπτομέρειες σχετικά με τον τρόπο με τον οποίο το χωρίσαμε για την εκπαίδευση και την αξιολόγηση με βάση τα υποκείμενα και τις δραστηριότητες δίνονται στο Κεφ. 7.3.1.

7.1.1 Τοποθέτηση δεικτών και δομή ανθρωπίνου σώματος

Για το σύνολο δεικτών χρησιμοποιήσαμε $M = 53$ σφαιρικούς οπισθοανακλαστικούς δείκτες διαμέτρου $14mm$, οι οποίοι τοποθετήθηκαν στις στολές καταγραφής κίνησης των ηθοποιών. Θεωρούμε τα μετα-επεξεργασμένα, καθαρά δεδομένα των δεικτών $\mathbf{M}_{gt} \in \mathbb{R}^{M\times 3}$ ως ground truth. Όσον αφορά τη δομή, οι αρχικές ακολουθίες παρέχουν πόζες 33 διαφορετικών αρθρώσεων, ωστόσο απλοποιούμε τη δομή και χρησιμοποιούμε $J = 19$. Τα καθαρά δεδομένα πόζας, $\mathbf{J}_{gt} \in \mathbb{R}^{J\times 3}$, θεωρούνται ως ground truth. Δείγματα των σχολιασμένων ζευγών εικόνων βάθους-υπέρυθρης ακτινοβολίας μαζί με τη διάταξη λήψης στο επαγγελματικό στούντιο καταγραφής όπου πραγματοποιήθηκε η δημιουργία του συνόλου δεδομένων, απεικονίζονται στο Σχήμα 7.2.

Σχήμα 7.2: Η διάταξη λήψης με 24 [8] κάμερες MXT40S και ένα χαμηλού κόστους σύστημα αισθητήρων βάθους πολλαπλών προβολών εξοπλισμένο με 4 στερεοσκοπικές συσκευές ανίχνευσης βάθους Intel RealSense D415. Οι έντονες αντανακλάσεις των δεικτών προκαλούνται στις ροές υπερύθρων με την εκπομπή υπέρυθρου φωτός στους οπισθο-ανακλαστικούς δείκτες που είναι προσαρτημένοι στο σώμα των υποκειμένων. Για να περιορίσουμε το θόλωμα της εικόνας, μειώσαμε το χρόνο έκθεσης των αισθητήρων, ο οποίος, κατά συνέπεια, μείωσε τη φωτεινότητα της εικόνας, οδηγώντας σε πιο ευδιάκριτες αντανακλάσεις των δεικτών σε σύγκριση με τις προεπιλεγμένες ρυθμίσεις. Τα ζεύγη εικόνων υπέρυθρης βάθους απεικονίζονται στο κάτω μέρος του σχήματος, ενώ στην αριστερή πλευρά, οι δείκτες βασικής αλήθειας και η πόζα προβάλλονται σε μία από τις προβολές υπέρυθρης βάθους για να απεικονιστεί η χωροχρονική ευθυγράμμιση μεταξύ των συστημάτων καταγραφής κίνησης και υπέρυθρης βάθους.

7.1.2 Δεδομένα οπτικών δεικτών από πολλαπλούς αισθητήρες βάθους

Οι περισσότερες από τις πρόσφατες συσκευές ανίχνευσης βάθους χαμηλού κόστους είναι εξοπλισμένες με υπέρυθρες κάμερες και πομπούς [111, 112]. Η κάμερα βάθους Intel RealSense D415 βασίζεται στην ενεργή στερεοσκοπική όραση για τον υπολογισμό του βάθους, αποτελούμενη από δύο κάμερες και έναν πομπό υπερύθρων για τη βελτίωση της ακρίβειας βάθους σε σκηνές φτωχές σε χαρακτηριστικά υφής. Ο πομπός υπερύθρων εκπέμπει στατικό υπέρυθρο μοτίβο στη σκηνή με αποτέλεσμα οι δείκτες να ανακλούν το σήμα στις υπέρυθρες κάμερες, επιτρέποντας την ανίχνευσή τους. Θεωρούμε C χωροχρονικά ευθυγραμμισμένες κάμερες βάθους $c \in \{1, ..., C\}$, περιμετρικά τοποθετημένες γύρω από ένα χώρο σύλληψης διαμέτρου περίπου $4m$, όπως φαίνεται στο Σχήμα 7.2. Κάθε κάμερα λαμβάνει ένα ζεύγος από μία έγχρωμη υπέρυθρη εικόνα $\mathbf{I}_c(\mathbf{p}) \in \mathbb{R}^3$ και έναν χάρτη βάθους $D_c(\mathbf{p}) \in \mathbb{R}$, με $\mathbf{p} := (x, y) \in \Omega$ να είναι οι συντεταγμένες των εικονοστοιχείων στο πεδίο εικόνας Ω που ορίζεται σε ένα πλέγμα $w \times h$, με w και h να είναι το πλάτος και το ύψος του, αντίστοιχα. Οι πόζες του αισθητήρα $\mathbf{T}_c := \left[\begin{smallmatrix} \mathbf{R}_c & \mathbf{t}_c \\ \mathbf{0} & 1 \end{smallmatrix}\right]$ είναι γνωστές σε ένα κοινό σύστημα συντεταγμένων, όπου $\mathbf{R}_c$ και $\mathbf{t}_c$ δηλώνουν την περιστροφή και τη μετάθεση αντίστοιχα. Ως εκ τούτου, μπορούμε να μετατρέψουμε τις συντεταγμένες του πεδίου της εικόνας βάθους κάθε προβολής σε ένα καθολικό σύστημα συντεταγμένων ως εξής:

$$\mathcal{T}_c(\mathbf{p}) = \mathbf{T}_c \pi^{-1}(D_c(\mathbf{p}), \mathbf{K}_c, \mathbf{p}), \tag{7.1}$$

με $\mathbf{T}_c$ τη σχετική θέση από το τοπικό σύστημα συντεταγμένων του αισθητήρα c προς το καθολικό και π^{-1} τη συνάρτηση αντιπροβολής που μετασχηματίζει το εικονοστοιχείο βάθους σε 3Δ συντεταγμένες, χρησιμοποιώντας τον πίνακα εγγενών παραμέτρων του αισθητήρα $\mathbf{K}_c$. Δεδομένου ότι οι υπέρυθρες εικόνες $\mathbf{I}_c$ και οι εικόνες βάθους D_c της κάμερας c είναι ευθυγραμμισμένες και ορίζονται στο ίδιο πεδίο εικόνας Ω, εφαρμόζουμε μια γραμμική κατωφλίωση [113] για αποτελεσματική κατάτμηση των τιμών της εικόνας. Στη συνέχεια, εφαρμόζεται ένας ταχύς αλγόριθμος ανίχνευσης περιγράμματος για να προκύψουν τα κέντρα των δεικτών ανά προβολή, τα οποία στη συνέχεια αντιπροβάλλονται στον κοινό, ενιαίο 3Δ χώρο χρησιμοποιώντας την Εξ. 7.1. Κατά τη διάρκεια αυτής της διαδικασίας, τα σημεία που δεν περιέχονται σε ένα 3Δ πλαίσιο οριοθέτησης που έχει οριστεί για να περιορίσει τον 3Δ χώρο, θεωρούνται ακραία σημεία και απορρίπτονται. Έτσι, ένα αραιό νέφος δεικτών, $\mathbf{M}_r \in \mathbb{R}^{M_f \times 3}$ των M_f 3Δ σημείων εξάγεται ανά καρέ f, το οποίο περιέχει τις 3Δ συντεταγμένες των δεικτών όπως λαμβάνονται από τους πολλαπλούς αισθητήρες. Δεδομένου του θορύβου των αισθητήρων καθώς και της ανίχνευσης μονής προβολής για κάθε αισθητήρα, το M_f κυμαίνεται γύρω αλλά διαφέρει από τον πραγματικό αριθμό των δεικτών, δηλαδή $M = 53$.

Η ποιότητα του συνολικού εντοπισμού των δεικτών είναι ανάλογη με τον αριθμό των προβολών, όπως στην πλειονότητα των συστημάτων πολλαπλών προβολών, κυρίως λόγω της εξάλειψης των αποκρύψεων και, κατά συνέπεια, των ελλιπών δεικτών. Επιλέξαμε μια διάταξη 4 αισθητήρων σε σταυρωτή τοποθέτηση για τη δημιουργία του συνόλου δεδομένων μας, θεωρώντας ότι αποτελεί το συμβιβασμό μεταξύ της αποφυγής αποκρύψεων και χαμηλού κόστους. Παρ' όλα αυτά, το προτεινόμενο μοντέλο εκπαιδεύεται ούτως ή άλλως με εξαιρετικά θορυβώδη δεδομένα που προκύπτουν από την προαναφερθείσα σχετικά αδύναμη οπτική παρακολούθηση των δεικτών (αρκεί μία έγκυρη παρατήρηση τουλάχιστον σε μία από τις προβολές) με τη μορφή μιας αραιής και χωρικά κανονικοποιημένης 3Δ αναπαράστασης δεδομένων, εξαλείφοντας

την πόλωσή του στις θέσεις των καμερών, των εγγενών παραμέτρων τους ή το συστηματικό θόρυβο των αισθητήρων βάθους, όπως συζητάμε στην πειραματική μας μελέτη στο Κεφ. 7.3.5.

7.1.3 Χωροχρονική Συσχέτιση

Η υψηλή ακρίβεια συγχρονισμού μεταξύ των αισθητήρων βάθους χαμηλής συχνότητας (30 καρέ/δευτερόλεπτο στο σύνολο δεδομένων μας, βλ. Κεφ. 4.1.1 για περισσότερες λεπτομέρειες) αποτελεί βασική προϋπόθεση. Η επιλεγμένη κάμερα βάθους Intel RealSense D415 προσφέρει συγχρονισμό υλισμικού εντός και μεταξύ των αισθητήρων, επιτρέποντας συγχρονισμό υψηλής ακρίβειας 4.1.4. Όσον αφορά τον συγχρονισμό μεταξύ των συστημάτων (D415, χαμηλού κόστους - VICON, υψηλού κόστους), η συνολική χρονική μετατόπιση μεταξύ των συστημάτων ανιχνεύθηκε στην αρχή κάθε ακολουθίας με τη χρήση μίας κλακέτας εξοπλισμένη με 2 δείκτες. Στη συνέχεια, οι αλληλουχίες μεταξύ των συστημάτων με διαφορετικό ρυθμό καρέ συγχρονίστηκαν λαμβάνοντας υπόψη τα χρονικά τους βήματα και αφαιρώντας πάντα την αρχική τους χρονική διαφορά.

Η χωρική συσχέτιση μεταξύ του συστήματος VICON και του συστήματος αισθητήρων βάθους επιτυγχάνεται με μια διαδικασία δύο βημάτων. Πραγματοποιούμε μια αρχική συσχέτιση χρησιμοποιώντας τις 3Δ θέσεις των δεικτών, όπως εκτιμώνται από κάθε σύστημα, δηλαδή με βάση την τριγωνοποίηση για το VICON και με βάση το βάθος για το D415. Εκμεταλλευόμαστε την αρχική, στατική φάση στην αρχή κάθε ακολουθίας T-Pose, όπου οι δείκτες ανιχνεύονται εύκολα από το σύστημα χαμηλού κόστους, καθώς οι υπέρυθρες εικόνες είναι ευκρινείς. Εφαρμόζουμε τον αλγόριθμο επαναληπτικού πλησιέστερου σημείου (Iterative Closest Point, ICP) για να μετασχηματίσουμε αρχικά με κάποιο σφάλμα το $\mathbf{M}_r$ στο σύστημα συντεταγμένων των σημείων αναφοράς $\mathbf{M}_{gt}$, με αποτέλεσμα το $\mathbf{M}_r^{'}$. Χρησιμοποιούμε τον ICP, καθώς, όπως αναφέρεται στην ενότητα 7.1.2, ο αριθμός των ανιχνευόμενων δεικτών ποικίλλει ανά καρέ και δεν υπάρχουν άμεσες αντιστοιχίες.

Σε δεύτερη φάση, για την επίτευξη ακριβούς χωρικής καταγραφής, ακολουθούμε με ένα δεύτερο βήμα βελτίωσης της πόζας των αισθητήρων. Αρχικά, βρίσκουμε τις αντιστοιχίες μεταξύ των $\mathbf{M}_{gt}$ και $\mathbf{M}_{r,c}^{'}$, όπου $\mathbf{M}_{r,c}^{'} \subset \mathbf{M}_r^{'}$ είναι το υποσύνολο των δεικτών που ανήκουν σε κάθε αισθητήρα c. Κατόπιν, κατασκευάζουμε διμερή γραφήματα μεταξύ των $\mathbf{M}_{gt}$ και $\mathbf{M}_{r,c}^{'}$, όπου τα κόστη των ακμών αντιπροσωπεύουν τις ευκλείδειες αποστάσεις μεταξύ των 3Δ σημείων. Τέλος, εφαρμόζουμε το διμερές ταίριασμα ελάχιστου κόστους (minimum cost bipartite matching) με:

$$\mathcal{B}_{M,c}(\mathbf{M}_{r,c}^{'}, \mathbf{M}_{gt}) = \min_{M_{gt}} \sum_{i} ||x_i - y_i||_2 \tag{7.2}$$

όπου $x_i \in \mathbf{M}_{r,c}^{'}$ και $y_i \in \mathbf{M}_{gt}$. Από το $\mathcal{B}_{M,c}$, χρησιμοποιούμε μόνο τις αντιστοιχίες κάτω από ένα αυστηρό κατώφλι, εξασφαλίζοντας μια ομάδα αντιστοιχιών υψηλής ποιότητας μεταξύ του $\mathbf{M}_{gt}$ και των ανιχνευμένων δεικτών στο $\mathbf{I}_c$. Στη συνέχεια, εφαρμόζουμε συνολική προσαρμογή πόζας καμερών [114] για να βελτιώσουμε τις πόζες των αισθητήρων χρησιμοποιώντας το $\mathbf{M}_{gt}$ ως αναφορά. Αναλυτικότερα, θεωρώντας σταθερές τις παραμέτρους $\mathbf{M}_{gt}$ και τις εγγενείς παραμέτρους της κάμερας $\mathbf{K}_c$, βελτιώνουμε από κοινού και επαναληπτικά τις πόζες των καμερών για να ελαχιστοποιήσουμε τα σφάλματα επαναπροβολής. Αυτό παρέχει μια υψηλής

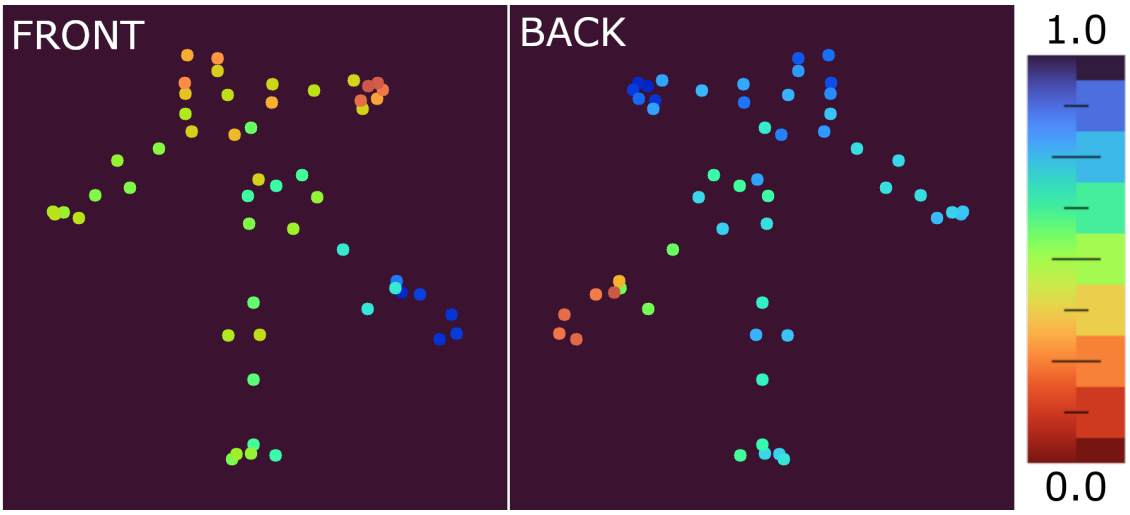
ακρίβειας χωρική συσχέτιση με βάση τα ground-truth δεδομένα οπτικών δεικτών $\mathbf{M}_{gt}$, μετασχηματίζοντας πλέον το $\mathbf{M}^{'}_{r,c}$ σε $\mathbf{M}^{''}_{r,c}$ για τους δείκτες κάθε όψης c και, κατά συνέπεια συνολικά για όλες τις όψεις σε $\mathbf{M}^{''}_{r}$, εφαρμόζοντας μια χωρική ομαδοποίηση για αποστάσεις δεικτών μικρότερες από $10mm$ ώστε να συγχωνεύονται μόνο οι δείκτες που έχουν εντοπιστεί εξαιρετικά κοντά ο ένας στον άλλο.

Αξίζει να σημειωθεί ότι αυτή η διαδικασία συσχέτισης λαμβάνεται υπόψη για τη δημιουργία του συνόλου δεδομένων και δεν απαιτείται κατά τη διάρκεια της δοκιμής του μοντέλου όπου τα ground truth δεδομένα απουσιάζουν. Τα αποτελέσματα της χωρικής και χρονικής συσχέτισης μεταξύ του VICON και του συστήματος πολλαπλών αισθητήρων παρουσιάζονται στο Σχήμα 7.2 όπου η πόζα και οι δείκτες από το VICON προβάλλονται σε ένα δείγμα υπέρυθρης εικόνας, επαληθεύοντας οπτικά την ακριβή συσχέτιση μεταξύ των 2 συστημάτων.

7.1.4 Κανονικοποιημένη Ορθογραφική Απόδοση Βάθους

Η έρευνα μας πραγματοποιείται λοιπόν με αρχική οπτική πληροφορία ένα αραιό 3Δ νέφος αντί για τις αρχικές εικόνες υπερύθρων ή βάθους, πράγμα το οποίο μας επιτρέπει να ξεπεράσουμε γνωστούς περιορισμούς των πυκνών δεδομένων. Αρκετά μοντέλα που βασίζονται σε πυκνά δεδομένα υποφέρουν από πόλωση, για παράδειγμα στις συγκεκριμένες πόζες της κάμερας καταγραφής και τις συνθήκες φωτισμού, την υπερμοντελοποίηση στα σύνολα εκπαίδευσης ή το συστηματικό θόρυβο του αισθητήρα βάθους.

Απλοποιούμε την ερευνητική μας πρόκληση θέτοντάς την ως ένα πρόβλημα χωρικής 3Δ παλινδρόμησης σε ορθογραφικά απεικονιζόμενους χάρτες βάθους υπό ένα πλαίσιο πολλαπλών όψεων, δηλαδή με τη χρήση δύο (ή περισσότερων) αντίθετων απεικονίσεων για να ξεπεραστούν οι τυχόν ασάφειες μίας προβολής. Εφαρμόζουμε έναν ογκομετρικό μετασχηματισμό κανονικοποίησης κλίμακας και μετατόπισης $\mathcal{T}_{\mathcal{N}}$ προκειμένου ο $\mathbf{M}^{''}_{r}$ να καταλαμβάνει το 80% κάθε διάστασης ενός κανονικοποιημένου κυβοειδούς 3Δ χώρου, δηλαδή να κυμαίνεται μεταξύ [0.1, 0.9] κατά μήκος των αξόνων, με αποτέλεσμα $\hat{\mathbf{M}}_r$. Αν και προφανές, αξίζει να σημειωθεί ότι ο ίδιος μετασχηματισμός κανονικοποίησης $\mathcal{T}_{\mathcal{N}}$ εφαρμόζεται στα ground truth δεδομένα $\mathbf{M}_{gt}$ και $\mathbf{J}_{gt}$, ορίζοντας έτσι τα $\hat{\mathbf{M}}_{gt}$ και $\hat{\mathbf{J}}_{gt}$ αντίστοιχα για την εποπτεία. Με αυτόν τον τρόπο, το 3Δ πλαίσιο οριοθέτησης που περιέχει κάθε δείγμα καταλαμβάνει τον ίδιο όγκο σε 3Δ επίπεδο, ενώ το περιθώριο 10% από τα όρια εξασφαλίζει την κατάλληλη συμπεριφορά των 2Δ συνελίξεων σε όλα τα επίπεδα του δικτύου. Στη συνέχεια, προβάλουμε το $\hat{\mathbf{M}}_r$ σε δύο αντίθετες όψεις αποδίδοντας δύο αραιές εικόνες βάθους σε δύο ορθογραφικές κάμερες, με το κύριο σημείο του φακού να είναι κεντραρισμένο στο βαρύκεντρο του $\hat{\mathbf{M}}_r$. Αποδίδουμε τις εικόνες βάθους σε υψηλή ανάλυση εικονοστοιχείων (800×800 συγκεκριμένα) και στη συνέχεια τις μεταβάλλουμε σε μέγεθος γραμμικά στην ανάλυση εισόδου του μοντέλου (160×160), με στόχο να εξαλείψουμε την κωδικοποίηση του σφάλματος κβαντισμού και την απώλεια πληροφορίας λόγω της κβαντισμένης απόδοσης (Κεφ. 6.3.2). Κατά την απόδοση, το εύρος [0.1, 0.9] στον άξονα "z" καθιστά τις κανονικοποιημένες θέσεις των δεικτών διακριτές σε σχέση με τις μηδενικές τιμές και στους δύο αποδιδόμενους χάρτες βάθους. Για το σκοπό αυτό, οι τιμές βάθους διατηρούν τις 3Δ θέσεις των σημείων σήμανσης, αναπαριστώντας το κανονικοποιημένο βάθος τους ως μικρές περιοχές από διασκορπισμένα εικονοστοιχεία βάθους, δημιουργώντας δύο 'αραιές' εικόνες βάθους, $\mathcal{D}_{front}$ και $\mathcal{D}_{back}$, που τροφοδοτούν το δίκτυό μας. Δείγματα από αυτούς τους



Σχήμα 7.3: ***Οπτικοποίηση δεδομένων εισόδου.*** *$\mathcal{D}_{front}$ και $\mathcal{D}_{back}$ με χρωματισμό για λόγους σαφήνειας.*

απεικονιζόμενους κανονικοποιημένους χάρτες βάθους απεικονίζονται στο Σχήμα 7.3.

Πριν από την εφαρμογή του μετασχηματισμού κανονικοποίησης $\mathcal{T}_{\mathcal{N}}$, εφαρμόζουμε μια τυχαία περιστροφική επαύξηση των δεδομένων γύρω από τους άξονες X, Y και Z κατά $[-10°, 10°]$, $[-180°, 180°]$ και $[-10°, 10°]$ αντίστοιχα, αυξάνοντας τη διακύμανση των μηκών των μελών του ανθρώπινου σώματος ανάλογα με τον προσανατολισμό τους στους άξονες του κυβοειδούς όγκου. Αξίζει να σημειωθεί ότι η 3Δ περιστροφική επαύξηση των δεδομένων, με την είσοδο να είναι ένα αραιό νέφος 3Δ σημείων, έχει τη φυσική σημασία της αλλαγής της οπτικής γωνίας απόδοσης της κάμερας. Αυτό επιτρέπει τη δημιουργία εντελώς νέων εισόδων χαρτών βάθους για το δίκτυο κατά τη διάρκεια της εκπαίδευσης, σε αντίθεση με την περιορισμένη επίδραση της ψευδο-περιστροφικής επαύξησης που εφαρμόζεται σε πυκνές αναπαραστάσεις εισόδου, όπως έγχρωμες εικόνες, πυκνούς χάρτες βάθους ή οποιαδήποτε άλλη είσοδο πυκνού 2Δ-πλέγματος.

Κατά αυτόν τον τρόπο, μας επιτρέπεται η εισαγωγή ενός αποδοτικού μοντέλου για αποτελεσματική εκτίμηση σε αραιά 3Δ δεδομένα. Αποφεύγουμε τη χρήση 3Δ συνελικτικών αρχιτεκτονικών, καθώς, παρά την αποτελεσματικότητά τους [34, 115, 116], εξακολουθούν να είναι υπολογιστικά δαπανηρές.

7.2 Μέθοδος

Το χαμηλό κόστος του DeMoCap προκύπτει κατά βάση από τη χρήση φθηνών, εμπορικών στερεοσκοπικών αισθητήρων βάθους και υπερύθρων, οι οποίοι, παρά τη θορυβώδη ανίχνευσή τους, μπορούν ικανοποιητικά να εκτιμήσουν την 3Δ θέση των οπισθοανακλαστικών δεικτών δίχως την ανάγκη συνεκτίμησης πολλαπλών προβολών, όπως συμβαίνει με τις υπέρυθρες κάμερες των επαγγελματικών συστημάτων που ο υπολογισμός γίνεται με τριγωνοποίηση μεταξύ παρατηρήσεων πολλαπλών όψεων. Το χαμηλό κόστος της στερεοσκοπικής υπέρυθρης ανίχνευσης βάθους εξακολουθεί να έχει το τίμημα της ανακρίβειας της παρατήρησης, την οποία, μαζί με τις προαναφερθείσες προκλήσεις της καταγραφής κίνησης με δείκτες, ξεπερνάμε με τη χρήση ενός μοντέλου βαθιάς μάθησης που επιτρέπει την ταυτόχρονη και αποτελεσματική

σταδιακή αλλά από κοινού εκτίμηση των 3Δ συντεταγμένων των δεικτών και των αρθρώσεων.

Με το DeMoCap, μοντελοποιούμε την καταγραφή κίνησης μέσω δεικτών ως μια σταδιακή παλινδρόμηση δεικτών σε 3Δ πόζα από θορυβώδη δεδομένα εισαγωγής, που προβάλλονται ορθογραφικά σε πολλαπλά σημεία θέασης ως χάρτες βάθους. Μέσω του μοντέλου μας επιτυγχάνουμε τα εξής:

- Ομαδοποίηση δεικτών καταγεγραμμένων ι από πολλαπλές γωνίες θέασης.
- Αποθορυβοποίηση δεικτών αγνοώντας τους δείκτες 'φαντάσματα' που προκαλούνται από λανθασμένη ανίχνευση και ανάκτηση αποκρυπτόμενων είτε μη εντοπισμένων δεικτών.
- Ανάκτηση από λανθασμένες εναλλαγές δεικτών κατά την παρακολούθηση, λόγω της διακριτής εκτίμησης ανά καρέ.
- Ταυτοποίηση/σήμανση δεικτών μέσω χωρικής παλινδρόμησης από τους θερμικούς χάρτες που αντιστοιχούν σε συγκεκριμένους δείκτες.
- Στιγμιαία εκτίμηση 3Δ πόζας από επισημασμένους 3Δ δείκτες χωρίς προηγούμενη γνώση της δομής του σώματος.

Ένα συνολικό διάγραμμα της προτεινόμενης μεθόδου απεικονίζεται στο Σχήμα 7.4. Τα καρέ πολλαπλών όψεων υπέρυθρων-βάθους που λαμβάνονται από τους αισθητήρες υποβάλλονται σε επεξεργασία για την εξαγωγή των 3Δ θέσεων των δεικτών $\mathbf{M}_r''$, οι οποίες στη συνέχεια κανονικοποιούνται και δίνουν τους κανονικοποιημένους $\hat{\mathbf{M}}_r$. Οι δείκτες $\hat{\mathbf{M}}_r$ στη συνέχεια αποδίδονται ορθογραφικά σε δύο αντικριστές εικόνες βάθους, $\mathcal{D}_{front}$ και $\mathcal{D}_{back}$. Με δεδομένο ότι οι εικόνες $\mathcal{D}_{front}$ και $\mathcal{D}_{back}$ αντιπροσωπεύουν τη χωρική κατανομή των δεικτών που είναι προσαρτημένοι στο ανθρώπινο σώμα, τις τροφοδοτούμε σε ένα πλήρως συνελικτικό μεταβατικό δίκτυο. Το μοντέλο προβλέπει διαδοχικά τους θερμικούς χάρτες των δεικτών και των αρθρώσεων στους οποίους εφαρμόζουμε μια νέα διπλής όψης και πλήρως διαφοροποιήσιμη χωρική 3Δ παλινδρόμηση για να αποκωδικοποιήσουμε με ακρίβεια τις κανονικοποιημένες 3Δ συντεταγμένες τόσο των $M = 53$ δεικτών, $\hat{\mathbf{X}}_M \in \mathbb{R}^{M\times 3}$, όσο και των $J = 19$ αρθρώσεων, $\hat{\mathbf{X}}_J \in \mathbb{R}^{J\times 3}$. Τέλος, εφαρμόζουμε τον αντίστροφο μετασχηματισμό κλίμακας και μετατόπισης $\mathcal{T}_{\mathcal{N}}{}^{-1}$ (Κεφ. 7.1.4) για να ανακτήσουμε τις 3Δ συντεταγμένες των δεικτών και της πόζας στις αρχικές φυσικές τους διαστάσεις.

Σχήμα 7.4: Χρησιμοποιώντας μια ρύθμιση πολλαπλών προβολών με έναν μικρό αριθμό χωροχρονικά συσχετισμένων καμερών βάθους υπέρυθρης ακτινοβολίας που τοποθετούνται περιμετρικά γύρω από ένα υποκείμενο με οπισθοανακλαστικούς δείκτες προσαρτημένους στο σώμα, καταγράφουμε τις κινήσεις του σώματος. Εντοπίζουμε τους δείκτες εκμεταλλευόμενοι τις έντονες ανταυακλάσεις των δεικτών που προκαλούνται στις εικόνες υπερύθρων και την ανίχνευση βάθους των αισθητήρων. Αποδίδοντας τους 3Δ δείκτες μετά από κανονικοποίηση σε δύο αντίθετες εικόνες βάθους, εκπαιδεύουμε ένα FCN για την από κοινού και διαδοχική πρόβλεψη των θερμικών χαρτών των δεικτών και των αρθρώσεων, αποκωδικοποιώντας στη συνέχεια με μια νέα πλήρως διαφορροποιήσιμη μονάδα τους 3Δ δείκτες και τις θέσεις των αρθρώσεων. Κατά τον χρόνο εκτέλεσης, διεξάγουμε μία ενιαία τροφοδότηση σε δύο υπερ-στάδια, όπου τα πρώτα στάδια εντοπίζουν τους δείκτες (πρώτο υπερ-στάδιο) και τα τελευταία (δεύτερο υπερ-στάδιο) εκτιμούν την πόζα του σώματος (απεικονίζουμε το παράδειγμα δύο όψεων της έννοιας της εισόδου/επιβλεψης πολλαπλών όψεων για λόγους απλότητας).

7.2.1 Μεταβατικά Δίκτυα από Δείκτες σε Ανθρώπινη Πόζα

Αρχιτεκτονική δικτύων

Όσον αφορά τη σχεδίαση των αρχιτεκτονικών των νευρωνικών δικτύων, βασιζόμαστε σε μια πολύ συγκεκριμένη προσέγγιση. Προτείνουμε FCN αρχιτεκτονικές πολλαπλών σταδίων με ομαλή κλιμάκωση από τις προβλέψεις χαρτών θερμότητας δεικτών σε χάρτες θερμότητας αρθρώσεων που οδηγούν σε μοντέλα με καλύτερη απόδοση και αποτελεσματικότερο σχεδιασμό, όπως αποδεικνύεται και συζητείται στο Κεφ. 7.3.3. Δεδομένου ότι ο βασική πληροφορία που καθορίζει την τελική πόζα είναι η χωρική κατανομή των 3Δ οπτικών δεικτών, σχεδιάζουμε τα δίκτυά μας να προβλέπουν/αποθορυβοποιούν τις 3Δ συντεταγμένες τους από τα πρώτα στάδια, υπολογίζοντας τις 3Δ συντεταγμένες των αρθρώσεων από τα τελευταία. Με αυτόν τον τρόπο, οι θέσεις των δεικτών διορθώνονται πριν από τον εντοπισμό των συντεταγμένων των αρθρώσεων, με αποτέλεσμα να λαμβάνουμε πιο εύστοχες και αξιόπιστες προβλέψεις.

Σχεδιάζουμε τις αρχιτεκτονικές των μοντέλων βασιζόμενοι σε αποτελεσματικά/σύγχρονα δίκτυα πρόβλεψης θερμικών χαρτών που έχουμε προαναφέρει σε προηγούμενα κεφάλαια, όπως τα Convolutional Pose Machines (CPM) [46], Stacked Hourglass (SH) [97] και ένα πιο πρόσφατο, το HRNet [117]. Προβλέπουμε χάρτες θερμότητας διπλής όψης τροφοδοτώντας τα δίκτυα με τους κανονικοποιημένους χάρτες βαθους $\mathcal{D}_{front}$ και $\mathcal{D}_{back}$. Ως εκ τούτου, εφαρμόζουμε τα μοντέλα μας και στις δύο όψεις και τα αποτελέσματα συγχωνεύονται για να καταλήξουμε σε μια τελική πρόβλεψη που εποπτεύεται από κοινού για τους δύο χάρτες εισόδου.

Ακολουθούμε τον ίδιο σχεδιασμό αρχιτεκτονικής για όλα τα δίκτυα, όπως απεικονίζεται στο Σχήμα 7.8. Αρχικά, τροφοδοτούμε κάθε χάρτη βάθους $\mathcal{D}_v, v = \{front, back\}$ σε ένα αρχικό δίκτυο προεπεξεργασίας για την εξαγωγή ενός χάρτη χαρακτηριστικών $\mathbf{F}_v$ ($\mathbf{F}$ για λόγους συντομίας). Σχεδιάζουμε ένα δίκτυο $2K$ σταδίων, $K \in \mathbb{N}$, και το χωρίζουμε σε δύο υπερ-στάδια που αποτελούνται από K στάδια το καθένα. Το πρώτο υπερ-στάδιο προβλέπει τους θερμικούς χάρτες δεικτών $\bar{\mathbf{H}}_M$ και το δεύτερο αρθρώσεων $\bar{\mathbf{H}}_J$, που προκύπτουν ως αθροίσματα των ενδιάμεσων θερμικών χαρτών $\mathbf{H}_{S,s_t}$ που προβλέπει κάθε στάδιο $s_t \in \{1, 2, ..., 2K\}$ κάθε υπερ-σταδίου $S \in \{M, J\}$. Ο χάρτης χαρακτηριστικών $\mathbf{F}$ συγχωνεύεται με τους $\mathbf{H}_{S,s_t}, s_t < 2K$ σε κάθε στάδιο για να τροφοδοτήσει το επόμενο.

Αντί να επιβλέπουμε τους θερμικούς χάρτες όπως αρχικά προτάθηκε από τους συγγραφείς των δικτύων (ανά στάδιο ενδιάμεση επίβλεψη στα CPM και SH και επίβλεψη θερμικού χάρτη τελευταίου σταδίου στο HRNet), επιβλέπουμε μόνο τους αθροισμένους θερμικούς χάρτες $\bar{\mathbf{H}}_M$ και $\bar{\mathbf{H}}_J$ με τις ground truth συντεταγμένες και τους ground truth θερμικούς χάρτες των αντίστοιχων αρθρώσεων. Όπως αναλύεται περαιτέρω στο Κεφ. 7.3.3, αυτό το σχήμα συνάθροισης συγκλίνει ταχύτερα και ευνοεί τη σταδιακή μετάβαση από τις πυκνές παρατηρήσεις στην αραιή παλινδρόμηση δεικτών και αρθρώσεων, όπως επίσης χρησιμοποιείται και επαληθεύεται και σε άλλες ερευνητικές εργασίες [118]. Λεπτομέρειες των προτεινόμενων αρχιτεκτονικών παρουσιάζονται παρακάτω στο Κεφ. 7.3.3.

Πρόβλεψη χάρτη θερμότητας

Έστω $\mathbf{H}^k_{s_t}, s_t \in \{1, 2, ..., 2K\}$ ο k-οστός χάρτης θερμότητας πριν από την εφαρμογή της συνάρτησης *Softmax* στο στάδιο s_t. Αρχικά, πραγματοποιείται η άθροιση των χαρτών θερ-

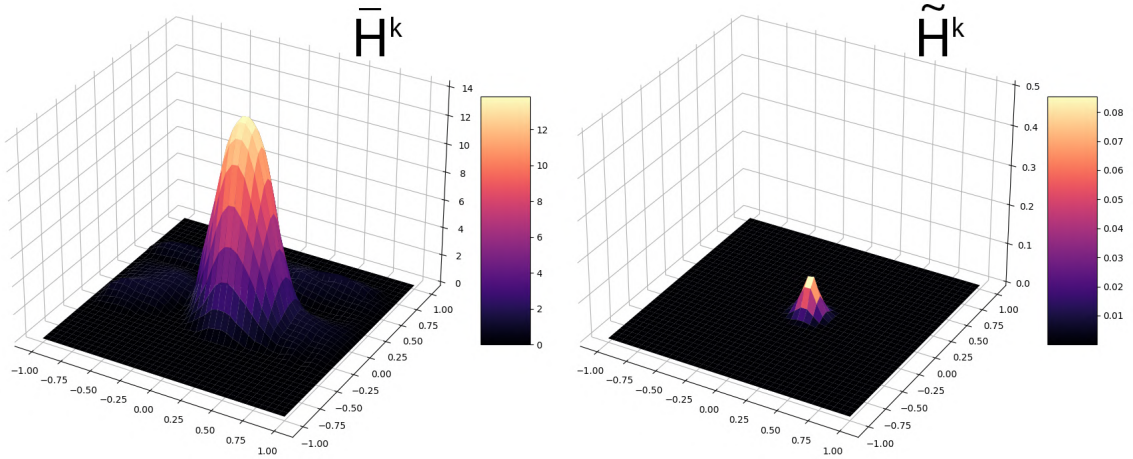
μότητας των σταδίων:

$$\bar{\mathbf{H}}^k = \sum_{s_t=f}^{l} \mathbf{H}_{s_t}^k \tag{7.3}$$

Συνολικά, εκτελούμε δύο αθροίσεις, μία για την παλινδρόμηση των δεικτών από το στάδιο $f = 1$ έως το στάδιο $l = K$ και μία για την παλινδρόμηση των αρθρώσεων από το στάδιο $f = K + 1$ έως το στάδιο $l = 2K$. Με την εφαρμογή της *Softmax* σε κάθε ένα από τους αθροισμένους χάρτες θερμότητας $\bar{\mathbf{H}}^k$ προκύπτουν οι χάρτες $\tilde{\mathbf{H}}^k$:

$$\tilde{\mathbf{H}}^k(\mathbf{p}) = \frac{e^{\bar{\mathbf{H}}^k(\mathbf{p})}}{\sum_{\mathbf{p}\in\Omega_k} e^{\bar{\mathbf{H}}^k(\mathbf{p})}} \tag{7.4}$$

όπου $\mathbf{p}$ συμβολίζει ένα εικονοστοιχείο θερμικού χάρτη και Ω_k το χωρικό πεδίο των θερμικών χαρτών. Παραδείγματα των θερμικών χαρτών $\bar{\mathbf{H}}$ και $\tilde{\mathbf{H}}$ απεικονίζονται στο Σχήμα 7.5.



Σχήμα 7.5: *Επιφανειακή απεικόνιση των προβλεπόμενων θερμικών χαρτών πριν ($\bar{\boldsymbol{H}}^k$) και μετά το Softmax ($\tilde{\boldsymbol{H}}^k$). Αφήνουμε το δίκτυο να προβλέπει θερμικούς χάρτες που ικανοποιούν και τις δύο εργασίες, δηλαδή ο μέσος όρος των τιμών $\bar{\boldsymbol{H}}^k$ να είναι ίσος με τη ζ-συντεταγμένη του τρισδιάστατου σημείου κλειδιού, ενώ μετά το Σοφτμαξ, η θερμοχάρτα $\tilde{\boldsymbol{H}}^k$ να προσεγγίζει μια γκαουσιανή κατανομή για την εκτίμηση της ξψ-συντεταγμένης.*

7.2.2 Χωρική Παλινδρόμηση Πολλαπλών Προβολών

Κάνουμε παλινδρόμηση $M = 53$ κανονικοποιημένων συντεταγμένων δεικτών $\hat{\mathbf{X}}_M \in \mathbb{R}^{M\times 3}$ και $J = 19$ κανονικοποιημένων συντεταγμένων αρθρώσεων $\hat{\mathbf{X}}_J \in \mathbb{R}^{J\times 3}$ με την αποκωδικοποίηση των θερμικών χαρτών $\bar{\mathbf{H}}$ και $\tilde{\mathbf{H}}$ που προβλέπονται από κάθε αντίστοιχο υπερ-στάδιο.

3Δ Κέντρο μάζας με zMean

Συνεισφέρουμε στην παλινδρόμηση 3Δ συντεταγμένων προτείνοντας μια νέα τεχνική, το κέντρο μάζας στον 3Δ χώρο (*CoM3D*) με την εισαγωγή ενός πλήρως διαφοροποιήσιμου επιπέδου, του *zMean*, συνδυάζοντας το με τη γνωστή τεχνική αποκωδικοποίησης συντεταγμένων

CoM από θερμικούς χάρτες. Η ισχυρισμός μας είναι ότι τα πλήρως συνελικτικά δίκτυα μπορούν να μάθουν να προβλέπουν τις κατανομές των θερμικών χαρτών σε ποικίλα εύρη τιμών μέσω ενός επιπλέον επιπέδου χωρικής πληροφορίας που επιτρέπει την κωδικοποίηση της τρίτης διάστασης, του βάθους.

Έτσι, εκτιμούμε τις 3Δ συντεταγμένες με την εισαγωγή του $CoM3D$ που συνδυάζει δύο πλήρως διαφοροποιήσιμα επίπεδα, τα $zMean$ και CoM, ($zMean + CoM = CoM3D$), με το $zMean$ να είναι η κύρια συνεισφορά μας. Ενώ για το CoM ακολουθούμε την τυπική διαδικασία [71, 72, 73] για την αποκωδικοποίηση των συντεταγμένων x και y, το κίνητρο πίσω από το $zMean$ είναι να εκμεταλλευτούμε έναν επιπλέον βαθμό ελευθερίας που επιτρέπει η εφαρμογή της $Softmax$ υπό την άμεση εποπτεία του χάρτη θερμότητας, προκειμένου να περιορίσουμε επιπλέον τον μέσο όρο του $\bar{\mathbf{H}}^k$ να προσεγγίζει τη συντεταγμένη z, οδηγώντας σε μια συμπαγή κωδικοποίηση 3Δ συντεταγμένων. Αναλυτικότερα, έστω (x_k, y_k, z_k) που υποδηλώνει τις προβλεπόμενες κανονικοποιημένες 3Δ συντεταγμένες για τον δείκτη ή την άρθρωση k, με το k να βρίσκεται είτε στο $\{1, ..., M\}$ είτε στο $\{1, ..., J\}$, αντίστοιχα. Στη συνέχεια, υπολογίζουμε το z_k ως εξής:

$$z_k = \frac{1}{N_x N_y} \sum_{\mathbf{p} \in \Omega_k} \bar{\mathbf{H}}^k(\mathbf{p}) \tag{7.5}$$

και τα $(x, y)_k$ από:

$$(x, y)_k = (\frac{1}{N_x}, \frac{1}{N_y}) \circ \sum_{\mathbf{p} \in \Omega} \tilde{\mathbf{H}}^k(\mathbf{p}) \cdot \mathbf{p} \tag{7.6}$$

με $N_x = 40$, $N_y = 40$ το πλάτος και ύψος κάθε θερμικού χάρτη, όπως έχει σχεδιαστεί για τις αρχιτεκτονικές των δικτύων μας, και $\circ$ που δηλώνει τον πολλαπλασιασμό κατά στοιχείο.

Εποπτεία πολλαπλών προβολών

Τέλος, για δύο αντίθετες όψεις απεικόνισης, διεξάγουμε από κοινού εποπτεία διπλής όψης εκτιμώντας ένα 3Δ σημείο ανά διπλή είσοδο περιστρέφοντας την κανονικοποιημένη πρόβλεψη συντεταγμένων $(x_{k,back}, y_{k,back}, z_{k,back})$ για την $\mathcal{D}_{back}$ κατά 180° γύρω από τον άξονα Y και τον μέσο όρο της με την κανονικοποιημένη πρόβλεψη συντεταγμένων $(x_{k,front}, y_{k,front}, z_{k,front})$ για την $\mathcal{D}_{front}$ κατά:

$$(\hat{x_k}, \hat{y_k}, \hat{z_k}) = \begin{pmatrix} \frac{1}{2}(x_{k,front} + (1 - x_{k,back})), \\ \frac{1}{2}(y_{k,front} + y_{k,back}), \\ \frac{1}{2}(z_{k,front} + (1 - z_{k,back})) \end{pmatrix}. \tag{7.7}$$

Με αυτόν τον τρόπο, προσεγγίζουμε κάθε μεμονωμένη 3Δ συντεταγμένη από δύο αντίθετες πλευρές που καλύπτουν τον 3Δ όγκο όπου περιέχεται το ανθρώπινο σώμα, εκτιμώντας τις κανονικοποιημένες συντεταγμένες $\hat{\mathbf{X}}_M$ και $\hat{\mathbf{X}}_J$ του δείκτη και της άρθρωσης, αντίστοιχα. Με άλλα λόγια, το μοντέλο μας μαθαίνει να προβλέπει θερμικούς χάρτες των οποίων η μέση τιμή προσεγγίζει την κανονικοποιημένη ground truth συντεταγμένη "z", ενώ το κανονικοποιημένο 2Δ κέντρο μάζας τους, μετά την εφαρμογή της $Softmax$, προσεγγίζει τις κανονικοποιημένες ground truth συντεταγμένες "x" και "y".

7.2.3 Συναρτήσεις Απώλειας

Κατά τη διάρκεια της εκπαίδευσης, επιβλέπουμε από κοινού τις συντεταγμένες $\hat{\mathbf{X}}_M$ και $\hat{\mathbf{X}}_J$ με τις ground truth συντεταγμένες, $\hat{\mathbf{M}}_{gt}$ και $\hat{\mathbf{J}}_{gt}$, αντίστοιχα. Από τη μία, σε αντίθεση με το DSNT [72], αντί να χρησιμοποιήσουμε την απώλεια ευκλείδειας απόστασης στο 3Δ, χρησιμοποιούμε την απώλεια Wing, $\mathcal{L}_{wing}$ [119], η οποία οδηγεί σε ευνοϊκότερη σύγκλιση. Η Wing αποτελεί μια συνάρτηση απωλειών η οποία συμπεριφέρεται ως λογαριθμική συνάρτηση για μικρά σφάλματα, ενώ για μεγαλύτερα ως L1. Η συνάρτηση απώλειας $\mathcal{L}_{wing}$ ορίζεται από τη σχέση:

$$\mathcal{L}_{wing}(x) = \begin{cases} w \ln(1 + |x|/\epsilon) & if |x| < w \\ |x| - C & otherwise \end{cases} \tag{7.8}$$

όπου το μη αρνητικό w θέτει το εύρος του μη γραμμικού μέρους σε $(-w, w)$, το ϵ περιορίζει την καμπυλότητα της μη γραμμικής περιοχής και $C = w - w \ln(1 + w/\epsilon)$ είναι μια σταθερή τιμή που συνδέει με ομαλό τρόπο τα τμηματικά καθορισμένα γραμμικά και μη γραμμικά μέρη της συνάρτησης. Η είσοδος x στο $\mathcal{L}_{wing}$ είναι η 3Δ ευκλείδεια απόσταση μεταξύ των προβλεπόμενων και των πραγματικών σημείων ενδιαφέροντος.

Από την άλλη, παρόμοια με το DSNT, επιβλέπουμε άμεσα και τη δομή του θερμικού χάρτη, καθώς η χωρική βελτιστοποίηση σε επίπεδο εικονοστοιχείου ενισχύει την εκπαίδευση του μοντέλου, βελτιώνοντας την απόδοσή του. Επιβάλλουμε κανονικοποίηση στο χάρτη θερμότητας για να τον οδηγήσουμε άμεσα προς συγκεκριμένη κατανομή. Πιο συγκεκριμένα, αναγκάζουμε τους χάρτες θερμότητας να προσεγγίζουν 2Δ γκαουσιανές κατανομές ελαχιστοποιώντας την απόκλιση μεταξύ των εκτιμώμενων χαρτών και των ground truth γκαουσιανών κατανομών με κέντρο τις 2Δ ορθογραφικές προβολές $\mathbf{p}_{gt}$ των κανονικοποιημένων ground truth 3Δ σημείων. Ο όρος εξομάλυνσης της κατανομής ορίζεται ως εξής:

$$\mathcal{L}_D(\tilde{\mathbf{H}}, \mathbf{p}_{gt}) = D(\tilde{H} || \mathcal{N}(\mathbf{p}_{gt}, \sigma^2 \mathbf{I}_2)) \tag{7.9}$$

όπου $D(\cdot||\cdot)$ είναι η απόκλιση Jensen-Shannon [120].

Τέλος, η συνολική συνάρτηση απώλειας που χρησιμοποιείται για την επίβλεψη του μοντέλου κατά την εκπαίδευση ορίζεται ως εξής:

$$\mathcal{L}_{total} = \lambda_1(\mathcal{L}_{wing,M} + \mathcal{L}_{wing,J}) + \lambda_2(\mathcal{L}_{D,front} + \mathcal{L}_{D,back}), \tag{7.10}$$

όπου λ_1, λ_2 είναι υπερ-παράμετροι που σταθμίζουν τις απώλειες κατανομής των συντεταγμένων και του θερμικού χάρτη, αντίστοιχα. Στο Σχήμα 7.6 απεικονίζεται ένα συνοπτικό διάγραμμα που απεικονίζει τα διάφορα επιμέρους τμήματα της μεθόδου.

Σχήμα 7.6: Τροφοδοτούμε το πρώτο υπερ-στάδιο δεικτών με τους κανονικοποιημένου χάρτες βάθους πολλαπλών όψεων, και διαδοχικά, το δεύτερο στάδιο πόζας, προβλέποντας τους χάρτες θερμότητας των δεικτών και των αρθρώσεων $\bar{\boldsymbol{H}}$, αντίστοιχα, με αποτέλεσμα να προκύπτουν οι χάρτες $\tilde{\boldsymbol{H}}$ μετά από την εφαρμογή της Softmax. Εφαρμόζοντας το CoM3D, αποκωδικοποιούμε τις συντεταγμένες z από το $\bar{\boldsymbol{H}}$ με το zMean και τις συντεταγμένες x,y από το $\tilde{\boldsymbol{H}}$ με το CoM. Αναπροσαρμόζοντας με αυτόν τον τρόπο τις 3Δ συντεταγμένες από κάθε προβολή, τις συγχωνεύουμε στο τελικό στάδιο. Επιβλέπουμε και τις δύο προβλέψεις $\tilde{\boldsymbol{H}}$ και x, y, z δεικτών και αρθρώσεων με $\mathcal{L}_D$ και $\mathcal{L}_{wing}$ αντίστοιχα, εκπαιδεύοντας ενιαία έτσι όλο το δίκτυο.

7.3 Πειραματική αξιολόγηση

Σε αυτή την ενότητα παρουσιάζουμε τα πειράματα που πραγματοποιήσαμε για την αξιολόγηση της μεθόδου μας. Στην ενότητα 7.3.1, παρουσιάζουμε το σύνολο δεδομένων που δημιουργήθηκε σύμφωνα με τα βήματα προεπεξεργασίας που περιγράψαμε στην ενότητα 7.1 και χρησιμοποιήθηκε για την εκπαίδευση, την επικύρωση και τη δοκιμή του μοντέλου μας. Στη συνέχεια, αναλύουμε τη μεθοδολογία αξιολόγησης που ακολουθήσαμε παρουσιάζοντας τις μετρικές που χρησιμοποιήσαμε (Κεφ. 7.3.2), τις state-of-the-art μεθόδους που συγκρίναμε με τη δική μας (Κεφ. 7.3.2) και τις λεπτομέρειες υλοποίησης για την εκτέλεση των πειραμάτων (Κεφ. 7.3.2). Στο Κεφ. 7.3.3, παρουσιάζουμε και συζητάμε ποσοτικά και ποιοτικά πειραματικά αποτελέσματα, δίνοντας πληροφορίες σχετικά με την απόδοση του μοντέλου μας. Τέλος, μια τελική μελέτη δίνει στοιχεία σχετικά με την αναγκαιότητα και τον αντίκτυπο των συνεισφορών μας στο Κεφ. 7.3.5.

7.3.1 Σύνολο δεδομένων

Λαμβάνοντας υπόψη τα βήματα συλλογής και προεπεξεργασίας των δεδομένων που περιγράφονται στην ενότητα 7.1, δημιουργούμε ένα σύνολο 12.197 δειγμάτων από 11 δραστηριότητες ενός ατόμου. Χωρίζουμε τα υποκείμενα $S1, S2, S3, S4$ σε δύο ζευγάρια ανδρών-γυναικών χρησιμοποιώντας τα δεδομένα του πρώτου ζευγαριού ($S3$ και $S4$) που εκτελούν 7 από τις 11 δραστηριότητες για εκπαίδευση (*running*, *basketball_ dribbling*, *sitting_ down*, *object_ dropping_ n_ picking*, *stretching_ n_ talking*, *watching_ scary_ movie* και *inflight_ safety_ announcement*) και τα δεδομένα από το δεύτερο ($S1$ και $S2$) εκτελώντας τα υπόλοιπα για επικύρωση (*jumping_ jack* και *bending*) και δοκιμή (*punching_ n_ kicking* και *sitting_ on_ a_ stool*). Χωρίσαμε το σύνολο δεδομένων μας με αυτόν τον τρόπο για να αξιολογήσουμε τα μοντέλα σε αθέατα υποκείμενα με διαφορετικές δομές σώματος και αθέατες δραστηριότητες, παρέχοντας αξιόπιστα και δίκαια συμπεράσματα όσον αφορά την απόδοσή τους. Τα σύνολα δεδομένων εκπαίδευσης, επικύρωσης και δοκιμής αποτελούνται από 8165, 1990 και 2042 δείγματα αντίστοιχα, όπως παρουσιάζονται στον Πίνακα 7.1.

7.3.2 Μεθοδολογία

Μετρικές

Μετράμε τα σφάλματα των μεθόδων σε φυσικές διαστάσεις εφαρμόζοντας τον αντίστροφο μετασχηματισμό κλίμακας και μετατόπισης $\mathcal{T}_{\mathcal{N}}{}^{-1}$, όπως περιγράφεται στο κεφάλαιο 7.1, με τη χρήση των μετρικών που ακολουθούν:

- Αξιολογούμε την ακρίβεια της παλινδρόμησης της 3Δ πόζας χρησιμοποιώντας το ευρέως χρησιμοποιούμενο μέσο σφάλμα ανά θέση άρθρωσης ($\mathrm{M_{PJPE}}$) [121] και το μέσο σφάλμα ανά θέση δείκτη ($\mathrm{M_{PMPE}}$) όταν είναι εφαρμόσιμο, δηλαδή όταν οι δείκτες προβλέπονται από τα μοντέλα.

- Με παρόμοιο τρόπο, μετράμε επίσης το μέσο τετραγωνικό σφάλμα ανά άρθρωση και θέση δείκτη ($\mathrm{RMS_{PJPE}}$ και $\mathrm{RMS_{PMPE}}$), τα οποία αποτελούν παραλλαγές των $\mathrm{M_{PJPE}}$

Πίνακας 7.1: *Αριθμός, δραστηριότητες και υποκείμενα συνόλου δεδομένων εκπαίδευσης, επαλήθευσης και δοκιμής.*

activity	**samples**	**set**	**subjects**
running	639	train	{S3,S4}
basketball_dribbling	666		
sitting_down	1,205		
object_dropping_n_picking	754		
stretching_n_talking	1,201		
watching_scary_movie	826		
in-flight_safety_announc.	2,874		
total train samples	***8,165***		
jumping_jack	692	val	{S1,S2}
bending	851		
total val samples	***1,543***		
punching_n_kicking	930	test	{S1,S2}
sitting_on_a_stool	1,112		
total test samples	***2,042***		

και M_{PMPE}, αντίστοιχα, με βάση το μέσο τετραγωνικό σφάλμα (RMSE) αντί του μέσου απόλυτου σφάλματος (MAE), όπως συζητήθηκε στο Κεφ.5.

- Χρησιμοποιούμε τη μέση μέση ακρίβεια (mAP) χρησιμοποιώντας τη μετρική Percentage of Correct Keypoints 3D (PCK3D) [89] σε ένα εύρος κατωφλίων σφάλματος α_{3D}.

- Πέραν από τις σημειακές μετρικές στον 3Δ χώρο, παρουσιάζουμε αποτελέσματα που υπολογίζονται με τη συγχώνευση των δεδομένων πόζας (κατεύθυνση των οστών προς τα εμπρός) και του προσανατολισμού που ορίζουν οι διάφορες ομάδες δεικτών που αντιστοιχούν στις αρθρώσεις, παρέχοντας μια επιπλέον μετρική σε σχέση με τη σταθερότητα των προβλέψεων και τις δυνατότητες που παρέχει η ταυτόχρονη παλινδρόμηση δεικτών και αρθρώσεων. Εξετάζουμε το μέσο όρο και τη ρίζα του τετραγωνικού μέσου όρο ανά άρθρωση των γωνιακών σφαλμάτων, M_{PJAE} και RMS_{PJAE}, μετρώντας τη γωνία θ σε μοίρες, μεταξύ των προβλεπόμενων και των ground truth περιστροφών των αρθρώσεων ως εξής:

$$\theta = cos^{-1}(2\langle\, \hat{q}_{j,gt}, \hat{q}_j\rangle^2 - 1) \tag{7.11}$$

όπου $\langle\, \hat{q}_{j,gt}, \hat{q}\rangle$ δηλώνει το εσωτερικό γινόμενο μεταξύ $\hat{q}_{j,gt}$ και $\hat{q}_j$ της άρθρωσης j, όπου q είναι η περιστροφή εκφρασμένη σε quaternion.

Αξίζει να σημειωθεί ότι για λόγους ορθής σύγκρισης με άλλες μεθόδους, χρησιμοποιούμε 17 από τις 19 συνολικά αρθρώσεις της εκτιμώμενης πόζας για την αξιολόγηση, εξαιρουμένων των δακτύλων των ποδιών.

Σύγκριση με σύγχρονες μεθόδους

Λόγω της έλλειψης δημόσιων μεθόδων που στοχεύουν στο συγκεκριμένο πρόβλημα, δηλαδή την παλινδρόμηση δεικτών και πόζας από θορυβώδη δεδομένα οπτικών δεικτών, εντοπίστη-

καν σχετικές μέθοδοι και εκπαιδεύτηκαν εκ νέου για να προσαρμοστούν στο δικό μας σύνολο δεδομένων, προσφέροντας έγκυρες συγκρίσεις. Διαχωρίζουμε τις μεθόδους με βάση τα δεδομένα εισόδου σε μεθόδους καταγραφής κίνησης μέσω δεικτών (*marker-based*) και μεθόδους χωρίς τη χρήση δεικτών ή οποιασδήποτε άλλης προσάρτησης (*markerless*). Για τις πρώτες η είσοδος είναι τα δεδομένα $\hat{\mathbf{M}}_r$, δηλαδή το κανονικοποιημένο αραιό νέφος των ανιχνευμένων δεικτών, ενώ για τις δεύτερες, χρησιμοποιούμε τις χωροχρονικά ευθυγραμμισμένες έγχρωμες υπέρυθρες εικόνες πολλαπλών όψεων μαζί με τις ενδογενείς και εξωγενείς παραμέτρους της κάμερας.

Αναλυτικότερα, συγκρίνουμε το μοντέλο μας με δύο marker-based μεθόδους, μια που βασίζεται σε νευρωνικά συνελικτικά δίκτυα γράφων (Graph Convolutional Neural Networks - GCN) σχεδιασμένη για εκτίμηση 3Δ πόζας χεριού από έγχρωμη εικόνα (HOPE) [9], προσαρμοσμένη στην εργασία μας, και μια μέθοδο άμεσης 3Δ παλινδρόμησης που χρησιμοποιείται για την σήμανση δεικτών σε πραγματικό χρόνο (OML) [10], προσαρμοσμένη για την ταυτόχρονη παλινδρόμηση των 3Δ συντεταγμένων των δεικτών-αρθρώσεων με την τροφοδοσία ενός και δύο χαρτών βάθους. Επιπλέον, αξιολογούμε τρεις markerless μεθόδους, δύο top-down εκτίμησης πόζας από χωροχρονικά συσχετισμένες έγχρωμες εικόνες πολλαπλών όψεων διαφορίσιμης τριγωνοποίησης (LT) [6] (όπως αναλύθηκε στο Κεφ. 5) και μία bottom-up βασισμένη σε γράφους (4DA) [11] από χωροχρονικά συσχετισμένες έγχρωμες εικόνες πολλαπλών προβολών.

Μέθοδοι που βασίζονται σε δείκτες (marker-based): Το *HOPE* είναι ένα 'ελαφρύ' μοντέλο που έχει σχεδιαστεί για την από κοινού εκτίμηση της πόζας του χεριού και ενός αντικειμένου στον 2Δ και 3Δ χώρο μέσω δύο κλιμακωτών νευρωνικών GCN δικτύων. Προσαρμόζουμε το πρώτο για την εκτίμηση των 2Δ συντεταγμένων των αρθρώσεων στους ορθογραφικούς χάρτες βάθους, ακολουθούμενο από το δεύτερο, το Adaptive Graph-U-Net [122], για τη μετατροπή των 2Δ σε 3Δ συντεταγμένες. Τροποποιούμε επίσης την είσοδο στην ανάλυση του χάρτη βάθους μας ($1 \times 160 \times 160$), αντί για τις έγχρωμες εικόνες της αρχικής εργασίας, και εκπαιδεύουμε το μοντέλο από την αρχή με αρχικοποίηση βαρών *Xavier* [123].

Όπως και στο HOPE, προσαρμόζουμε το *OML* στη νέα ανάλυση του χάρτη βάθους εισόδου (160×160 αντί για 52×52 της αρχικής εργασίας), αρχικοποιώντας τα βάρη του μοντέλου με αρχικοποίηση *Xavier*. Επιπλέον, προσαρμόζουμε την έξοδο του δικτύου ώστε να προβλέπει ένα διάνυσμα με $M = 53$ και $J = 19$ 3Δ θέσεις, δηλαδή τον στόχο της εργασίας μας, ενώ παρουσιάζουμε μια επιπλέον παραλλαγή της προσέγγισης που τροφοδοτείται με δεδομένα βάθους πολλαπλών όψεων παρόμοια με το DeMoCap.

Μέθοδοι χωρίς δείκτες (markerless): Όσον αφορά τις μεθόδους *LT*, η $\text{LT}_{(\text{alg.})}$ που βασίζεται σε αλγεβρική τριγωνοποίηση με μαθησιακά βάρη εμπιστοσύνης μεταξύ των εκτιμήσεων των κάμερας συζητήθηκε εκτενώς στο Κεφ. 5, ενώ η $\text{LT}_{(\text{vol})}$ αποτελεί μια προσέγγιση ογκομετρικού τριγωνοποίησης που βασίζεται σε πυκνή γεωμετρική συνάθροιση προβλέψεων 2Δ θερμικών χαρτών από πολλαπλές οπτικές γωνίες. Η είσοδος που χρησιμοποιείται για την εκπαίδευση αυτών των μοντέλων είναι ένα καρέ N χωροχρονικά συσχετισμένων έγχρωμων εικόνων μαζί με τις αντίστοιχες πόζες των καμερών και τις εσωτερικές τους εγγενείς παραμέτρους. Στοχεύοντας σε μια δίκαιη σύγκριση μεταξύ των μεθόδων LT και DeMoCap, επανεκπαιδεύουμε τα μοντέλα LT στο δικό μας σύνολο δεδομένων αρχικοποιώντας με τα προεκπαιδευμένα βάρη λόγω των διαφορών μεταξύ των συνόλων δεδομένων.

Οι LT μέθοδοι χρησιμοποιούν το μοντέλο βαθιάς μάθησης Mask R-CNN 2D [124] με ResNet-152 [125] για την πρόβλεψη των ανθρώπινων ορίων στις εικόνες (οριοθέτηση), ωστόσο στα πειράματα μας χρησιμοποιούμε απευθείας τα ground truth όρια για να αποφύγουμε την επανεκπαίδευσή του. Χρησιμοποιούμε τα βάρη των προ-εκπαιδευμένων μοντέλων [1] που εκπαιδεύτηκαν στο σύνολο δεδομένων Human3.6 [126] με αρχικό μοντέλο που προ-εκπαιδεύτηκε στο σύνολο δεδομένων COCO [85] και τα εκπαιδεύουμε περαιτέρω στο έγχρωμο υπέρυθρο σύνολο δεδομένων μας για 10 εποχές υιοθετώντας το πλαίσιο εκπαίδευσης (υπερ-παραμέτρους και άλλες οδηγίες) που προτείνουν οι συγγραφείς. Χρησιμοποιούμε τα έγχρωμα δεδομένα υπέρυθρης ακτινοβολίας για την επανεκπαίδευση των μοντέλων αντί για τους αραιούς χάρτες βάθους των δεικτών μιας και θα απαιτούσαν εκπαίδευση των μοντέλων από την αρχή λόγω της εντελώς διαφορετικής φύσης των δεδομένων. Αξίζει να σημειωθεί ότι οι προβλέψεις που λαμβάνονται από τη μέθοδο $\text{LT}_{(\text{alg.})}$ χρησιμοποιούνται για την προσέγγιση $\text{LT}_{(\text{vol.})}$. Για την εκπαίδευση, χρησιμοποιούμε την επίσημη υλοποίηση και τις οδηγίες που παρέχονται από τους ίδιους τους συγγραφείς[2].

Η μέθοδος *4DA* λαμβάνει υπόψη τη χρονική συσχέτιση μεταξύ διαδοχικών καρέ. Με τη χρήση του OpenPose [1], από κάθε μεμονωμένη προβολή ανακτώνται χάρτες θερμότητας των αρθρώσεων και χάρτες εμπιστοσύνης σύνδεσης μεταξύ τους (πεδία σύνδεσης των μερών). Συνδυάζοντας τα παραπάνω μεταξύ δύο διαδοχικών καρέ, κατασκευάζεται ένας 4Δ γράφος με ακμές ανά όψη που συνδέουν συσχετισμένα μέρη του σώματος, ακμές αντιστοίχισης ανά όψη που συνδέουν το ίδιο μέρος του σώματος μεταξύ των διαφόρων προβολών και ακμές χρονικής συσχέτισης για την αντιστοίχιση των ανιχνευμένων 3Δ σημείων σε προηγούμενο καρέ με νέες 2Δ ανιχνεύσεις στο τρέχον. Με την ίδια πρακτική που ακολουθήσαμε για το LT, επανεκπαιδεύουμε το μοντέλο για 10 εποχές, αρχικοποιώντας το με τα βάρη του προ-εκπαιδευμένου μοντέλου OpenPose[3].

Λεπτομέρειες υλοποίησης DeMoCap

Εκπαιδεύουμε το δίκτυό μας για 200 εποχές χρησιμοποιώντας τον βελτιστοποιητή *Adam* [127] με αρχικό ρυθμό μάθησης ίσο με $1e-4$, ενώ εφαρμόζουμε γραμμική κυλιόμενη πτώση του ρυθμού ίση με 0.95 κάθε 4 εποχές. Το μέγεθος της παρτίδας δειγμάτων είναι 16 και η τυπική απόκλιση του χάρτη θερμότητας κατά τη διάρκεια της εποπτείας είναι $\sigma = 1.0$. Τα βάρη λ των απωλειών $\mathcal{L}_{wing}$ και $\mathcal{L}_D$, όπως ορίζονται στην εξίσωση 7.10, ορίζονται σε $\lambda_1 = 2$ και $\lambda_2 = 1$, ενώ οι παράμετροι για την $\mathcal{L}_{wing}$ ορίζονται σε $w = 10$ και $\epsilon = 2$.

Το μοντέλο υλοποιείται με PyTorch [128] και moai [129], ενώ τα πειράματα εκτελούνται σε συμβατικό υπολογιστή με 1 κάρτα γραφικών NVIDIA GTX 1080 Ti 12 GB RAM, και επεξεργαστή Intel i7(R), με ίδιο seed για όλα τα πειράματα για δίκαιη σύγκριση και αναπαραγωγιμότητα. Ο κώδικας και το σύνολο δεδομένων είναι δημόσια διαθέσιμα [4].

[1] Τα μοντέλα που ήταν δημόσια διαθέσιμα στο επίσημο αποθετήριο των συγγραφέων κατά τη στιγμή της παρούσας εργασίας.

[2] https://github.com/karfly/learnable-triangulation-pytorch/tree/9d1a26ea893a513bdff55f30ecbfd2ca8217bf5d

[3] https://github.com/CMU-Perceptual-Computing-Lab/openpose/tree/1f1aa9c59fe59c90cca685b724f4f97f76137224

[4] https://github.com/tofis/democap

7.3.3 Πειραματικά αποτελέσματα

Το σύνολο επικύρωσης χρησιμοποιήθηκε για τη ρύθμιση των υπερ-παραμέτρων εκπαίδευσης, ενώ η αξιολόγηση στο σύνολο δοκιμής πραγματοποιήθηκε μετά την επιλογή των τελικών μοντέλων επιλογής. Παρουσιάζουμε τα αποτελέσματα και στα δύο σύνολα, δείχνοντας τη διακύμανση της ακρίβειας εξαγωγής συμπερασμάτων μεταξύ τους για τα διάφορα μοντέλα, υποδεικνύοντας τη δυνατότητα γενίκευσής τους.

Αξιολόγηση με διάφορες FCN Αρχιτεκτονικές Νευρωνικών Δικτύων

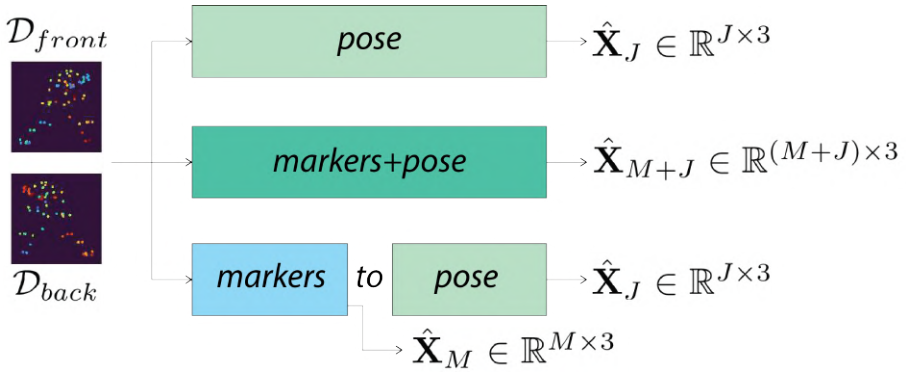
Για όλες τις αρχιτεκτονικές, ακολουθούμε την ίδια προσέγγιση. Όλα τα δίκτυά μας αποτελούνται από 2 υπερ-στάδια, δηλαδή ομάδες σταδίων όπως ορίζονται στις αρχιτεκτονικές FCN πολλαπλών σταδίων και πολλαπλών κλάδων. Το πρώτο υπερ-στάδιο προβλέπει τους θερμικούς χάρτες $\mathbf{H}_{M,1\ldots K}$ τους οποίους αθροίζουμε στους τελικού χάρτες δεικτών $\bar{\mathbf{H}}_M$, ενώ το δεύτερο υπερ-στάδιο προβλέπει τους $\mathbf{H}_{J,K+1\ldots 2K}$ οι οποίοι συναθροίζονται στους τελικούς θερμικούς χάρτες αρθρώσεων $\bar{\mathbf{H}}_J$. Οι προβλέψεις κάθε σταδίου και τα αντίστοιχα χαρακτηριστικά συνενώνονται για να τροφοδοτήσουν κάθε επόμενο στάδιο. Οι διάφορες αρχιτεκτονικές σχεδιάζονται στο Σχήμα 7.8. Συζητάμε λεπτομέρειες για κάθε δίκτυο ξεχωριστά παρακάτω.

- **Convolutional Pose Machines (CPM).** Ακολουθώντας την προτότυπη εργασία [46], ορίζουμε συνολικά 6 στάδια, τα οποία διαχωρίζονται σε δύο υπερ-στάδια των 3 σταδίων το καθένα. Ωστόσο, μειώνουμε τον αριθμό των επιπέδων *MaxPooling* σε 2 αντί για 3 αφαιρώντας το τρίτο. Αυτό έχει ως αποτέλεσμα έναν χάρτη χαρακτηριστικών υψηλότερης ανάλυσης $\mathbf{F}$, (δηλ. 2Δ χωρικού μεγέθους 40×40) που οδηγεί σε αυξημένη ανάλυση του χάρτη θερμότητας, παρόμοια με τα υπόλοιπα δίκτυα. Στη συνέχεια, ακολουθούμε την αρχιτεκτονική όπως είχε αρχικά προταθεί στο δίκτυο CPM, δηλαδή το πρώτο στάδιο αποτελείται από ένα 9×9 ακολουθούμενο από δύο 1×1 συνελικτικά στρώματα, ενώ κάθε επόμενο στάδιο αποτελείται από 5 συνελικτικά στρώματα ($3 \cdot 11 \times 11 - 2 \cdot 1 \times 1$). Όλα τα στάδια τροφοδοτούνται με τη συνένωση του $\mathbf{F}$ και της εξόδου του προηγούμενου σταδίου, εκτός από το *Stage1* το οποίο τροφοδοτείται μόνο με το $\mathbf{F}$.

- **Stacked Hourglass (SH).** Σχεδιάζουμε και υλοποιούμε μια αρχιτεκτονική Stacked Hourglass [97] 8 σταδίων με βάση τις υπερ-παραμέτρους της αρχικής εργασίας, επιλέγοντας 4 στάδια ανά υπερ-στάδιο. Ξεκινώντας από μια μονάδα προ-επεξεργασίας, εξάγεται ένας χάρτης χαρακτηριστικών $\mathbf{F}$, τον οποίο συνενώνουμε με τις ενδιάμεσες εξόδους του χάρτη χαρακτηριστικών κάθε σταδίου. Χρησιμοποιούμε μονάδες Hourglass με βάθος ίσο με 2, καταλήγοντας σε θερμικούς χάρτες 2Δ χωρικού μεγέθους ίσου με 40×40.

- **High Resolution Network (HRNet).** Ακολουθώντας την αρχική εργασία [117], σχεδιάζουμε ένα δίκτυο βασισμένο στο HRNet, στοιβάζοντας δύο αρχιτεκτονικές HRNet 4 σταδίων, μία ανά υπερ-στάδιο. Λόγω των πολλών σταδίων και κλάδων του HRNet, επιλέγουμε να υλοποιήσουμε το δεύτερο υπερ-στάδιο ως ένα δεύτερο HRNet 4 σταδίων αντί για ένα μοντέλο 8 σταδίων καταλήγοντας αντίστοιχα σε 8 κλάδους. Παρομοίως με τις προηγούμενες δύο αρχιτεκτονικές, τροφοδοτούμε το δεύτερο υπερ-στάδιο με τους

χάρτες χαρακτηριστικών $\mathbf{F}$ συνενωμένους με τους εκτιμώμενους χάρτες θερμότητας. Η ρύθμιση κάθε μονάδας υπερ-σταδίου είναι παρόμοια με αυτή που προτάθηκε στην αρχική εργασία, όπου το δεύτερο, το τρίτο και το τέταρτο στάδιο κάθε υπερ-σταδίου αποτελούνται από 1, 4, 3 μπλοκ ανταλλαγής χαρακτηριστικών, αντίστοιχα.



Σχήμα 7.7: *Σχεδιάζουμε όλες τις αρχιτεκτονικές FCN σε 3 παραλλαγές για να αξιολογήσουμε την σημασία των μεταβατικών δικτύων. Τροφοδοτούμε τα $\mathcal{D}_{front}$ και $\mathcal{D}_{back}$ σε όλες τις παραλλαγές με **pose**, **markers+pose** και **markers-to-pose** αποδίδοντας μόνο αρθρώσεις ($\hat{\boldsymbol{X}}_J \in \mathbb{R}^{J\times 3}$), ταυτόχρονα δείκτες και αρθρώσεις ($\hat{\boldsymbol{X}}_{M+J} \in \mathbb{R}^{(M+J)\times 3}$), και διαδοχικά δείκτες ($\hat{\boldsymbol{X}}_M \in \mathbb{R}^{M\times 3}$) και αρθρώσεις ($\hat{\boldsymbol{X}}_J \in \mathbb{R}^{J\times 3}$), αντίστοιχα.*

Αρχικά, αξιολογούμε το DeMoCap εφαρμόζοντας το μεταβατικό μας σχεδιασμό σε διάφορες FCN αρχιτεκτονικές, όπως αναλύεται στην ενότητα 7.2.1. Ο λόγος είναι διττός: i) για να επικυρώσουμε τον ισχυρισμό μας ότι οι μεταβατική σχεδίαση (markers-to-pose) αποδίδει καλύτερα ανεξάρτητα από την αρχιτεκτονική που χρησιμοποιείται για την πρόβλεψη των χαρτών θερμότητας και ii) για να επιλέξουμε το μοντέλο με τις καλύτερες επιδόσεις για τη σύγκριση με τις SoA μεθόδους.

Εκπαιδεύουμε το DeMoCap υπό το μεταβατικό σχεδιασμό με τις αρχιτεκτονικές CPM, SH και HRNet σε τρεις παραλλαγές το καθένα:

- ***pose***: Εκπαιδεύουμε τα μοντέλα σε ένα ενιαίο πλαίσιο, παλινδρομώντας και επιβλέποντας μόνο τους χάρτες θερμότητας και τις συντεταγμένες των αρθρώσεων με τους δείκτες να απουσιάζουν, εστιάζοντας έτσι σε μία μόνο εργασία, την πρόβλεψη των 3Δ θέσεων των αρθρώσεων $\hat{\mathbf{X}}_J \in \mathbb{R}^{J\times 3}$.

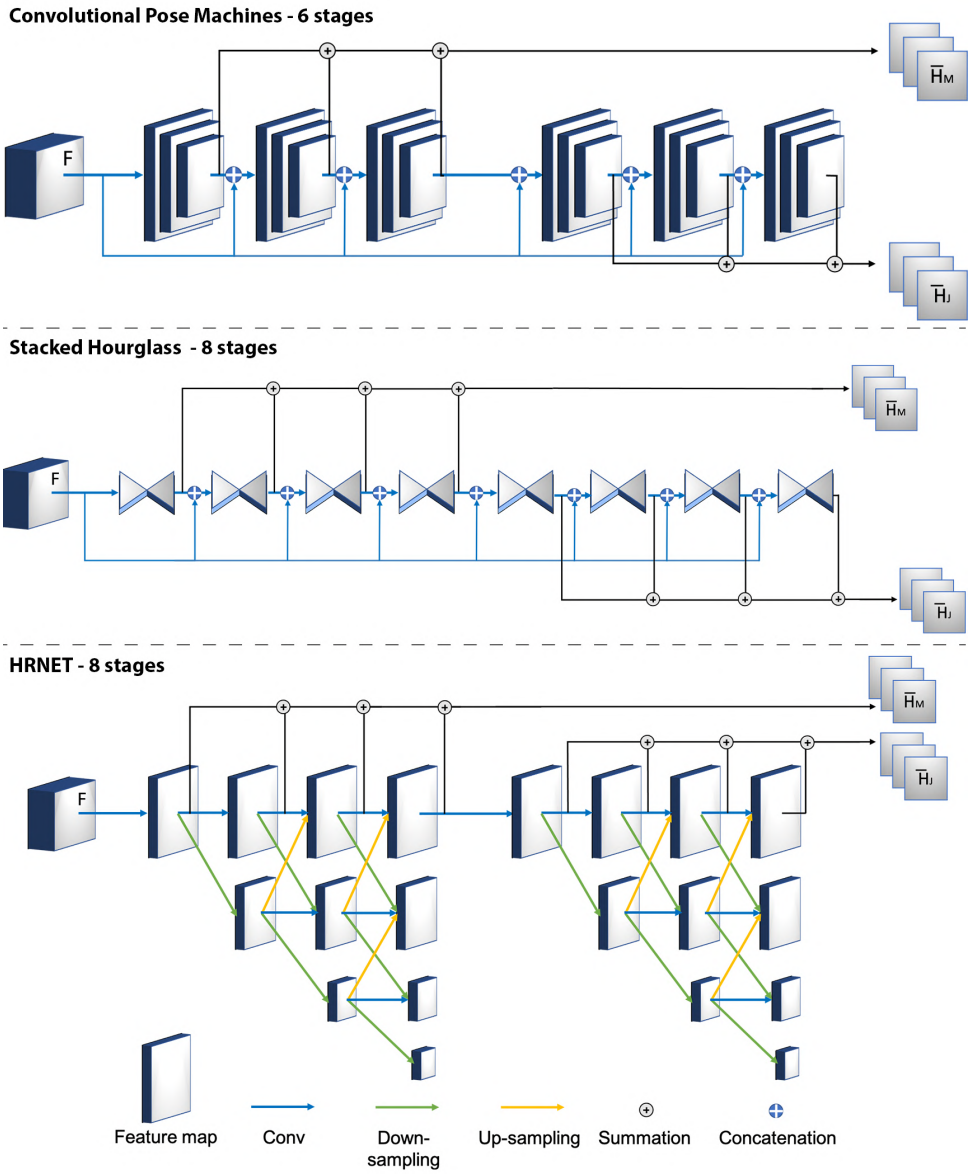
- ***markers+pose***: Ομοίως με την παραλλαγή *pose*, εκπαιδεύουμε τα μοντέλα σε μια ενιαία λογική, ωστόσο εκτιμούμε και επιβλέπουμε τόσο τους χάρτες θερμότητας και τις συντεταγμένες των δεικτών όσο και των αρθρώσεων, $\hat{\mathbf{X}}_{M+J} \in \mathbb{R}^{(M+J)\times 3}$.

- ***markers-to-pose***: Εκπαιδεύουμε το DeMoCap στη βασική του σχεδίαση όπου το πρώτο υπερ-στάδιο προβλέπει τους δείκτες και το δεύτερο την πόζα, με αποτέλεσμα να προκύπτουν διαδοχικά οι χάρτες $\hat{\mathbf{X}}_M \in \mathbb{R}^{M\times 3}$ και $\hat{\mathbf{X}}_J \in \mathbb{R}^{J\times 3}$, αντίστοιχα.

Τα αποτελέσματα αυτών των πειραμάτων απεικονίζονται στον Πίνακα 7.2. Η σταδιακή προσέγγιση *markers-to-pose* επιτυγχάνει υψηλότερη απόδοση για όλα τα μοντέλα στο κύριο στόχο μας, δηλαδή την εκτίμηση της πόζας. Ωστόσο, σε ορισμένα πειράματα, οι δείκτες

εντοπίζονται με μεγαλύτερη ακρίβεια από την παραλλαγή *markers+pose*. Για την ταυτόχρονη εκτίμηση δεικτών και πόζας, η οποία επιτρέπει επίσης την εκτίμηση της 3Δ περιστροφής, οι παράμετροι του δικτύου είναι βέλτιστοι για την προσέγγιση *markers-to-pose*, καθώς, για την *markers+pose*, οι θερμικές χάρτες δεικτών και αρθρώσεων προβλέπονται σε όλα τα στάδια του δικτύου, κάνοντας το δίκτυο μεγαλύτερο και λιγότερο απαδοτικό.

Όσον αφορά τα μοντέλα, αν και το SH αποδίδει αξιοσημείωτα στο σύνολο δοκιμών ξεπερνώντας το HRNet, θεωρούμε το τελευταίο ως αρχιτεκτονική επιλογής για τα υπόλοιπα πειράματα, καθώς οι ενδείξεις μας βασίστηκαν στα πειραματικά αποτελέσματα που ανακτήθηκαν στο σύνολο επικύρωσης.



Σχήμα 7.8: ***Μεταβατικές FCN Αρχιτεκτονικές πολλαπλών σταδίων.*** *Παρουσιάζουμε τις αρχιτεκτονικές που χρησιμοποιήσαμε για την εκπαίδευση του DeMoCap. Σε όλες τους, ακολουθούμε την ίδια γενική ιδέα όπου τα πρώτα στάδια προβλέπουν τους θερμικούς χάρτες των δεικτών ανά στάδιο* s*,* $\boldsymbol{H}_{M,1...s}$*, και τους αθροισμένους* $\bar{\boldsymbol{H}}_M$*, ενώ τα τελευταία στάδια προβλέπουν τους* $\boldsymbol{H}_{J,1...s}$ *και τους* $\bar{\boldsymbol{H}}_J$*. Οι προβλέψεις κάθε σταδίου και οι χάρτες χαρακτηριστικών* $\boldsymbol{F}$ *συνενώνονται για να τροφοδοτήσουν κάθε επόμενο στάδιο.*

Πίνακας 7.2: *Αποτελέσμα M_{PJPE}, RMS_{PJPE}, M_{PMPE}, RMS_{PMPE}, mAP_{50mm}, M_{PJAE} και RMS_{PJAE} μεταξύ του DeMoCap με CPM, SH και HRNet σε 3 παραλλαγές το καθένα, pose, marker+pose και markers-to-pose.*

DeMoCap \ ***Metrics*** (mm/°)	**Set**	$\mathbf{M_{PJPE}}\downarrow$	$\mathbf{RMS_{PJPE}}\downarrow$	$\mathbf{M_{PMPE}}\downarrow$	$\mathbf{RMS_{PMPE}}\downarrow$	$\mathbf{mAP_{50mm}}\uparrow$	$\mathbf{M_{PJAE}}\downarrow$	$\mathbf{RMS_{PJAE}}\downarrow$
CPM-6-{pose}		41.45	53.96	-	-	81.72%	-	-
CPM-6-{marker+pose}		40.05	50.00	53.21	65.06	79.33%	20.71	24.37
CPM-6-{markers-to-pose}		38.72	48.26	51.46	62.89	82.31%	20.10	23.60
SH-8-{pose}		38.60	49.67	-	-	82.76%	-	-
SH-8-{marker+pose}	val	38.55	51.68	42.30	54.44	85.35%	17.70	22.41
SH-8-{markers-to-pose}		36.94	51.42	44.84	58.70	87.49%	18.19	23.55
HRNet-8-{pose}		34.44	42.97	-	-	87.88%	-	-
HRNet-8-{marker+pose}		34.48	43.05	**40.42**	**49.57**	89.50%	**16.91**	**20.23**
HRNet-8-{markers-to-pose}		**33.83**	**42.65**	42.33	51.74	**90.41%**	18.66	22.47
CPM-6-{pose}		45.96	58.75	-	-	83.45%	-	-
CPM-6-{marker+pose}		46.28	58.21	62.83	77.91	83.20%	19.18	24.41
CPM-6-{markers-to-pose}		45.14	57.13	61.14	76.11	84.97%	18.22	23.32
SH-8-{pose}		40.40	51.05	-	-	88.39%	-	-
SH-8-{marker+pose}	test	39.02	49.90	51.13	64.46	87.86%	15.81	21.11
SH-8-{markers-to-pose}		**38.45**	**48.31**	**47.10**	**59.21**	**89.32%**	**15.08**	**20.00**
HRNet-8-{pose}		40.99	53.05	-	-	87.51%	-	-
HRNet-8-{marker+pose}		40.89	52.42	51.88	65.47	87.37%	19.19	22.18
HRNet-8-{markers-to-pose}		40.04	51.69	52.92	66.49	88.05%	19.73	26.18

Ποσοτική Ανάλυση

Τα μοντέλα 4DA και LT για εκτίμηση 3Δ πόζας από έγχρωμες εικόνες χωρίς δείκτες είναι αποτελεσματικά και δείχνουν την ικανότητά τους να εκτιμούν 3Δ πόζες από πολλαπλές χωροχρονικά συσχετισμένες προβολές, όντας σχετικά σταθερά και ανθεκτικά στις αποκρύψεις και την περιορισμένη θέαση ανά προβολή. Όπως παρουσιάζεται στην αρχική ερευνητική εργασία [6], η ογκομετρική εκδοχή LT αποδίδει καλύτερα και στο σύνολο δεδομένων μας, παρουσιάζοντας μεγαλύτερη ακρίβεια από την αλγεβρική.

Η μέθοδός μας αποδίδει πιο αξιόπιστες και ακριβείς προβλέψεις από τις 4DA και LT σε όλες τις μετρικές σε σχέση με τα συνολικά αποτελέσματα παρουσιάζοντας χαμηλότερο $\mathrm{M_{PJPE}}$, μεγαλύτερο $\mathrm{mAP_{50mm}}$ και χαμηλότερο $\mathrm{RMS_{PJPE}}$, παρά το γεγονός ότι τα μοντέλα αυτά έχουν εκπαιδευτεί με πολύ μεγαλύτερα σε αριθμό δειγμάτων και υποκειμένων και γενικότερα ποικιλομορφίας σύνολα δεδομένων. Αυτό οφείλεται στις σημαντικές διαφορές του DeMoCap έναντι των μεθόδων πολλαπλών όψεων που λειτουργούν με πυκνά οπτικά δεδομένα. Το DeMoCap εκπαιδεύεται να προβλέπει τις θέσεις των δεικτών και την πόζα από αραιά 3Δ δεδομένα δεικτών, όπου η είσοδος σχετίζεται με και αφορά αποκλειστικά την πόζα του ανθρώπινου σώματος στον 3Δ χώρο. Δεν υπάρχει κανένας περαιτέρω πλεονασμός στην πληροφορία που να περιλαμβάνει για παράδειγμα κάποιο μη σχετικό φόντο, ρουχισμό των υποκειμένων (υφάσματα ή χρώματα), συνθήκες φωτισμού ή οποιαδήποτε άλλη πληροφορία, όπως συμβαίνει με τα μοντέλα βαθιάς μάθησης που εκπαιδεύονται σε πυκνές οπτικές ροές. Με άλλα λόγια, τα δεδομένα εισόδου του DeMoCap δεν είναι ευαίσθητα δημιουργώντας πόλωση στο εκπαιδευόμενο μοντέλο, όπως συμβαίνει με τα δεδομένα χρώματος, βάθους ή οποιασδήποτε άλλης πυκνής οπτικής ροής. Συν τοις άλλοις, το DeMoCap εκπαιδεύεται σε αμιγώς 3Δ δεδομένα, ενώ τα LT και 4DA βασίζονται στη συγχώνευση πολλαπλών 2Δ ανιχνεύσεων, αποδυναμώνοντας τον εντοπισμό των τελικών 3Δ βασικών σημείων. Παρά την προσπάθεια μείωσης των βαρών των λανθασμένων 2Δ προβλέψεων με τεχνικές μαθησιακής στάθμισης ή συσχέτισης γραφημάτων πριν από την τελική 3Δ συγχώνευση, τα σφάλματα δεν μπορούν να εξαλειφθούν πλήρως, οδηγώντας σε λανθασμένες εκτιμήσεις όταν τα μέρη του σώματος δεν εντοπίζονται εντός του οπτικού πεδίου τουλάχιστον μιας από τις κάμερες ή αποκρύπτονται με αποτέλεσμα λανθασμένες προβλέψεις. Για το DeMoCap, η απόκρυψη ενός δείκτη σε μία από τις προβολές δεν θεωρείται αρκετή για να οδηγήσει το μοντέλο σε αποτυχία. Η ανίχνευση ενός δείκτη από μία τουλάχιστον κάμερα βάθους, μια εξαιρετικά πιθανή περίπτωση όταν μικρός αριθμός αισθητήρων βάθους καταγράφει την κίνηση, μπορεί να είναι αρκετή για να οδηγήσει το μοντέλο σε ακριβή πρόβλεψη δεδομένης της χαμηλής διακύμανσης της 3Δ εισόδου.

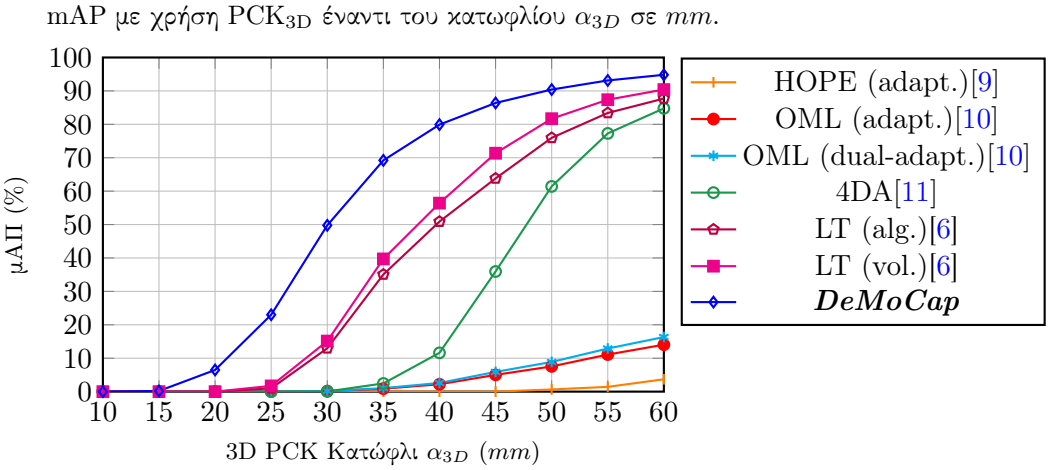
Παρά τις προσπάθειές μας να βελτιώσουμε την αποτελεσματικότητα των μοντέλων HOPE και OML, η απόδοσή τους είναι σχετικά ανεπαρκής, καθώς δυσκολεύονται να αναχθούν με ακρίβεια οι 3Δ συντεταγμένες σε δείγματα από αθέατα άτομα και δραστηριότητες. Οι χαμηλές επιδόσεις τους θα μπορούσαν ενδεχομένως να εξηγηθούν από τη χρήση της άμεσης παλινδρόμησης και το σχετικά μικρό σύνολο δεδομένων εκπαίδευσης που διαθέτουμε, σε αντίθεση με το μέγεθος του συνόλου δεδομένων των αρχικών εργασιών και τη χρήση προ-εκπαιδευμένων μοντέλων που δεν ήταν εφαρμόσιμα στα δεδομένα χαρτών βάθους που χρησιμοποιούμε ως είσοδο. Επιπλέον, η παρούσα πρόκληση είναι συγκριτικά πιο απαιτητική, καθώς εκτιμούμε $M + J = 72$ 3Δ συντεταγμένες σε σύγκριση με το HOPE και το OML που έχουν σχεδιαστεί

Πίνακας 7.3: *M_{PJPE}, RMS_{PJPE}, M_{PMPE}, RMS_{PMPE}, mAP_{50mm}, M_{PJAE} και RMS_{PJAE} μεταξύ των HOPE [9], OML [10], LT [6], 4DA [11] και της μεθόδου μας. Για λόγους σαφήνειας, χρησιμοποιούμε το **C** για δεδομένα χρώματος και **M** για δεδομένα δεικτών υποδεικνύοντας μεθόδους χωρίς δείκτες και με δείκτες, αντίστοιχα.*

Method \ ***Metrics*** (mm/°)	**Set-In**	$\mathbf{M_{PJPE}}\downarrow$	$\mathbf{RMS_{PJPE}}\downarrow$	$\mathbf{M_{PMPE}}\downarrow$	$\mathbf{RMS_{PMPE}}\downarrow$	$\mathbf{mAP_{50mm}}\uparrow$	$\mathbf{M_{PJAE}}\downarrow$	$\mathbf{RMS_{PJAE}}\downarrow$
4DA [11]	val-C	51.34	63.19	-	-	62.15%	-	-
$LT_{(alg.)}$[6]		46.91	54.70	-	-	71.55%	-	-
$LT_{(vol.)}$[6]		43.76	49.66	-	-	83.60%	-	-
HOPE $[9]_{(adapt.)}$	val-M	124.14	144.42	-	-	0.0%	-	-
$OML_{(adapt.)}$[10]		113.36	137.61	129.92	151.49	7.45%	38.02	46.17
$OML_{(dual\text{-}adapt.)}$[10]		108.66	131.19	124.15	146.34	8.88%	34.09	42.33
DeMoCap		**33.83**	**42.65**	**42.33**	**51.74**	**90.41%**	**18.66**	**22.47**
4DA [11]	test-C	50.36	76.26	-	-	61.41%	-	-
$LT_{(alg.)}$[6]		49.57	70.31	-	-	75.95%	-	-
$LT_{(vol.)}$[6]		46.69	64.91	-	-	81.68%	-	-
HOPE $[9]_{(adapt.)}$	test-M	115.50	136.54	-	-	3.57%	-	-
$OML_{(adapt.)}$[10]		97.71	121.36	122.52	138.19	21.64%	28.40	35.50
$OML_{(dual\text{-}adapt.)}$[10]		93.70	111.95	112.39	131.96	18.51%	27.13	33.13
DeMoCap		**40.04**	**51.69**	**52.92**	**66.49**	**88.05%**	**19.73**	**26.18**

για την παλινδρόμηση λιγότερων από 30 3Δ σημείων. Παρ' όλα αυτά, αξίζει να σημειωθεί ότι το εκπαιδευμένο OML μοντέλο διπλής προβολής παρουσιάζει σχετικά καλύτερα αποτελέσματα από το μοντέλο μονής προβολής, αναδεικνύοντας τις δυνατότητες της εποπτείας πολλαπλών όψεων.

Μια πιο ολοκληρωμένη ανάλυση όσον αφορά την απόδοση των μεθόδων στο σύνολο δοκιμών απεικονίζεται στο Σχήμα 7.9, όπου συγκρίνουμε τις μεθόδους με τη μετρική mAP με PCK3D έναντι του κατωφλίου α_{3D} σε mm. Η προτεινόμενη μέθοδος υπερτερεί των υπόλοιπων με υψηλότερο mAP έναντι όλου του εύρους των κατωφλίων α_{3D}, ενώ για $\alpha_{3D} = 35mm$, το mAP είναι ήδη $\approx 70\%$. Οι δύο OML μέθοδοι δεν είναι συγκρίσιμες με τις μεθόδους 4DA, LT και DeMoCap για $\alpha_{3D} \in [10, 60]$ σε mm. Τέλος, και τα δύο LT μοντέλα παρουσιάζουν μεγαλύτερη ακρίβεια από το 4DA.



Σχήμα 7.9: *Σύγκριση της γραφικής παράστασης μεταξύ HOPE[9], OML[10], LT[6], 4DA[11] και DeMoCap με χρήση της μετρικής mAP PCK3D έναντι του κατωφλίου α_{3D} σε mm στο σύνολο δοκιμής. Τα αποτελέσματά μας επιτυγχάνουν υψηλές αποδόσεις σε χαμηλά κατώφλια α_{3D} επιδεικνύοντας την αποτελεσματικότητα της μεθόδου μας.*

Αποτελέσματα ανά άρθρωση. Πέραν από την παρουσίαση των συνολικών αποτελεσμάτων, αξιολογούμε την απόδοση των μεθόδων σε ανάλυση ανά άρθρωση του ανθρώπινου σώματος στον Πίνακα 7.4, όπου το DeMoCap υπερτερεί των συγκρινόμενων μεθόδων στις περισσότερες αρθρώσεις του σώματος.

Η ανάλυση αυτή μας επιτρέπει να αξιολογήσουμε τη συνέπεια των μοντέλων στην εκτίμηση των διαφόρων αρθρώσεων ξεχωριστά. Το σκεπτικό πίσω από αυτή την ανάλυση είναι ότι, παραδοσιακά στην εκτίμηση της ανθρώπινης πόζας και κίνησης, το επίπεδο δυσκολίας για τον εντοπισμό των αρθρώσεων αυξάνεται σταδιακά καθώς μετακινούμαστε από τις αρθρώσεις του κορμού του σώματος, δηλαδή τους γοφούς, τη σπονδυλική στήλη, τους ώμους, τον αυχένα, προς το κεφάλι και τις τελικές αρθρώσεις των άκρων, δηλαδή τους αστραγάλους και τους καρπούς. Οι τελευταίοι, ως τελικοί κόμβοι μιας αρθρωτής δομής, του ανθρώπινου σώματος, κινούνται πιο ελεύθερα από το υπόλοιπο σώμα παρουσιάζοντας μεγάλη απόκλιση όσον αφορά τη τοποθέτησή τους ως προς το υπόλοιπο σώμα. Παρ' όλα αυτά, θεωρούμε σημαντικό για μια μέθοδο που στοχεύει στην καταγραφή της ανθρώπινης κίνησης να παρουσιάζει ευρωστία και συνέπεια σε όλες τις εκτιμήσεις των αρθρώσεων.

Πίνακας 7.4: *Σφάλμα 3Δ ευκλείδειας απόστασης ανά άρθρωση σε mm. Οι αρθρώσεις σε* ***bold-italic*** *υποδεικνύουν τις διμερείς αρθρώσεις για τις οποίες παρουσιάζεται το μέσο σφάλμα. Για λόγους σαφήνειας,* ***C*** *για έγχρωμες εικόνες και* ***M*** *για δεδομένα δεικτών υποδηλώνουν τις μεθόδους χωρίς δείκτες και με δείκτες, αντίστοιχα.*

Method \ ***Joints*** (mm)	**Set-In**	**Head**	**Neck**	***Shoulders***	***Elbows***	***Wrists***	**Pelvis**	***Spines***	***Hips***	***Knees***	***Ankles***
4DA [11]	val-C	64.22	27.95	49.54	77.89	97.01	25.47	-	34.00	34.25	33.54
$LT_{(alg.)}$[6]		32.82	**25.57**	41.08	56.98	68.43	36.12	33.37	45.26	53.22	53.16
$LT_{(vol.)}$[6]		32.68	25.77	39.43	51.84	61.99	34.82	33.04	42.29	48.54	48.17
$HOPE_{(adapt.)}$[9]	val-M	103.64	85.76	91.71	155.68	161.11	119.87	82.49	132.99	154.47	122.11
$OML_{(adapt.)}$[10]		81.76	61.79	88.71	151.48	232.93	91.60	84.38	93.01	98.50	97.00
$OML_{(dual\text{-}adapt.)}$[10]		81.35	60.58	83.50	148.87	219.25	83.27	81.55	90.51	92.70	94.61
DeMoCap		**26.34**	27.06	**35.18**	**35.44**	**40.38**	**26.78**	**30.55**	**28.63**	**35.12**	**42.18**
4DA [11]	test-C	62.25	25.67	50.49	46.31	51.71	38.85	-	47.20	54.82	65.85
$LT_{(alg.)}$[6]		28.96	20.93	32.63	53.21	65.67	34.51	30.59	54.96	55.69	86.36
$LT_{(vol.)}$[6]		**28.53**	20.69	31.54	50.23	61.00	**33.18**	**30.21**	51.39	52.15	79.14
$HOPE\ [9]_{(adapt.)}$	test-M	98.14	80.23	98.05	128.29	149.24	72.66	68.00	76.08	172.01	164.61
$OML_{(adapt.)}$[10]		45.35	51.22	80.85	132.28	147.66	73.87	63.56	73.92	117.46	129.59
$OML_{(dual\text{-}adapt.)}$[10]		45.12	50.22	76.38	130.09	138.75	67.15	61.42	71.94	110.25	126.40
DeMoCap		30.09	**19.30**	**23.28**	**28.94**	**45.25**	49.38	31.39	**27.85**	**32.95**	**52.62**

Στα πειράματά μας, η ίδια πρόκληση/δυσκολία εντοπίζεται σε όλες τις μεθόδους, ωστόσο συγκριτικά τα σφάλματα μεταξύ των αρθρώσεων του άκρου και του κορμού για τις HOPE, OML, 4DA και LT παρουσιάζουν μεγαλύτερη διακύμανση από την προτεινόμενη μέθοδο, πράγμα που σημαίνει ότι εκτιμά την πόζα με πιο ισορροπημένη ακρίβεια όσον αφορά τις αρθρώσεις του σώματος από ό,τι οι άλλες προσεγγίσεις. Οι απότομες κινήσεις των μερών του σώματος όταν εκτελούνται ταχείες δραστηριότητες προκαλούν θόλωση στην εικόνα οδηγώντας τα μοντέλα που βασίζονται σε οπτικά δεδομένα σε λανθασμένες εκτιμήσεις. Για να ξεπεραστεί αυτή η πρόκληση, το DeMoCap έχει σχεδιαστεί ρητά για να εκτελεί την εκτίμηση της πόζας σε δύο φάσεις. Η αρχική θορυβώδης και ελλιπής είσοδος δεικτών βελτιώνεται/αποκαθίσταται σε μια πρώτη φάση, οδηγώντας την εκτίμηση της πόζας σε ένα μεταγενέστερο στάδιο με βάση τα βελτιωμένα δεδομένα δεικτών. Η αποθορυβοποίηση των δεικτών επιτρέπει στο μοντέλο να εκτελεί με μεγαλύτερη ακρίβεια τις εκτιμήσεις των αρθρώσεων στο τελευταίο στάδιο, οι οποίες σχετίζονται μόνο με τις θέσεις των δεικτών, χωρίς καμία άλλη συσχέτιση με την αρχική θολή έγχρωμη εικόνα. Λαμβάνουμε αξιοσημείωτα χαμηλότερα σφάλματα για τις αρθρώσεις *Wrists* και *Ankles* στον Πίνακα 7.4, ειδικά στο σύνολο δοκιμών όπου κατά τη διάρκεια της εντελώς ελεύθερα εκτελούμενης δραστηριότητας *punching_n_kicking*, πολλά μέρη του σώματος βρίσκονται εκτός του οπτικού πεδίου των καμερών για αρκετά καρέ.

Αποτελέσματα ανά δραστηριότητα. Στον πίνακα 7.5, παρουσιάζουμε και συζητάμε τα αποτελέσματα του μοντέλου αναλύοντάς τα ανά δράση για να αξιολογήσουμε την απόδοση του μοντέλου σε ακολουθίες διαφορετικών κινήσεως και πόζας. Συμπεριλαμβανομένων των δραστηριοτήτων τόσο των συνόλων επικύρωσης όσο και των συνόλων δοκιμής, παρουσιάζουμε τα MPJPE αποτελέσματα για 4 δραστηριότητες: *jumping_jack*, *bending*, *punching_n_kicking* και *sitting_on_a_stool*. Παρά τη λεπτομερή βελτιστοποίηση των μοντέλων στο σύνολο επικύρωσης, και τα δύο σύνολα μοιράζονται τα ίδια χαρακτηριστικά λαμβάνοντας υπόψη ότι και τα δύο περιλαμβάνουν αθέατα υποκείμενα και δραστηριότητες σε σχέση με το σύνολο εκπαίδευσης. Ως εκ τούτου, το επίπεδο δυσκολίας για τη σύλληψη αυτών των κινήσεων καθορίζεται κυρίως από τις αντικειμενικές προκλήσεις κάθε συγκεκριμένης εκτέλεσης/δραστηριότητας, όπως η πολυπλοκότητα και η ταχύτητά τους, έχοντας ως αποτέλεσμα αρκετές αποκρύψεις ή ελλιπή δεδομένα, ή κινήσεις μερών του σώματος εκτός του πεδίου θέασης των καμερών. Αυτό αποδεικνύεται από τα χαμηλότερα σφάλματα στην ενέργεια *sitting_on_a_stool* η οποία παρουσιάζει χαμηλότερα σφάλματα από τις ενέργειες του συνόλου επικύρωσης.

Για τη δραστηριότητα *punching_n_kicking*, οι ηθοποιοί κλήθηκαν να γρονθοκοπήσουν και να κλοτσήσουν μπροστά στις κάμερες χωρίς καθοδήγηση ή άλλους περιορισμούς. Αυτό είχε ως αποτέλεσμα την καταγραφή εξαιρετικά απαιτητικών δεδομένων λόγω των γρήγορων και ελεύθερων κινήσεων με θορυβώδη δείγματα λόγω θόλωσης, μερικής απόκρυψης ή μερών του σώματος εκτός του οπτικού πεδίου των καμερών. Αυτό παρατηρείται στην ανάλυση ανά δραστηριότητα, όπου τα αποτελέσματα όλων των μεθόδων για τη συγκεκριμένη, ελεύθερη δραστηριότητα παρουσιάζουν τα υψηλότερα σφάλματα.

Πίνακας 7.5: *M_{PJPE} αποτελέσματα ανά δραστηριότητα για τα σύνολα επικύρωσης και δοκιμής, παρουσιάζοντας την απόδοση των μοντέλων σε διάφορες ενέργειες. Για λόγους σαφήνειας, **C** για έγχρωμες εικόνες και **M** για δεδομένα δεικτών υποδηλώνουν τις μεθόδους χωρίς δείκτες και με δείκτες, αντίστοιχα.*

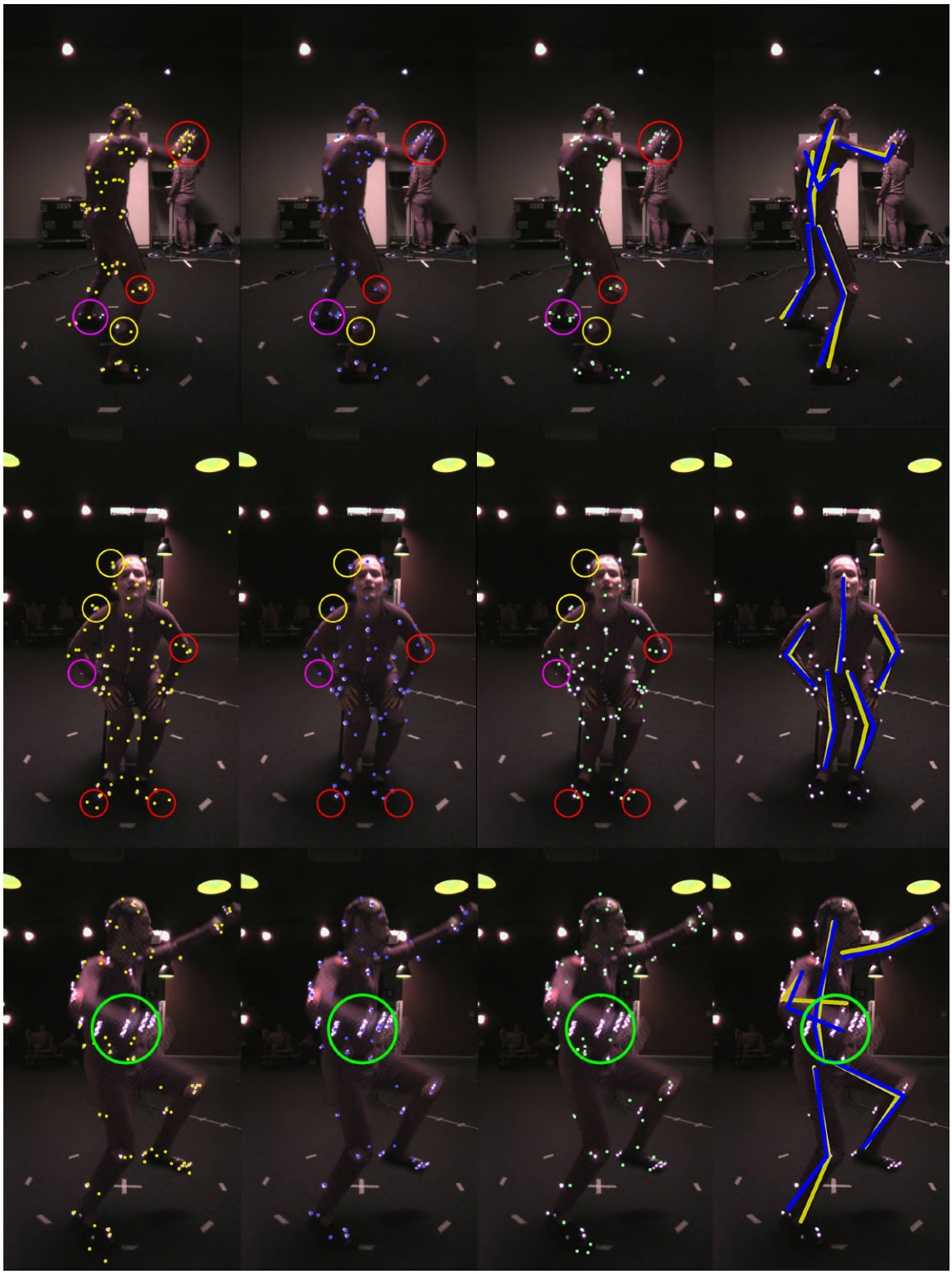
Method \ Action	**In**	**Jumping Jack**$_{\text{val}}$	**Bending**$_{\text{val}}$	**Punching & Kicking**$_{\text{test}}$	**Sitting on stool**$_{\text{test}}$
4DA [11]	C	51.17	51.52	61.46	37.67
LT$_{\text{(alg.)}}$[6]		44.18	48.63	64.36	34.83
LT$_{\text{(vol.)}}$[6]		41.47	45.22	59.74	33.71
HOPE [9]$_{\text{(adapt.)}}$	M	103.44	143.03	147.82	88.63
OML$_{\text{(adapt.)}}$[10]		101.12	123.87	131.76	69.29
OML$_{\text{(dual-adapt.)}}$[10]		97.34	119.03	124.96	67.71
DeMoCap		**28.74**	**38.47**	**57.19**	**25.77**

7.3.4 Ποιοτική Ανάλυση

Σε αυτή την ενότητα, παρουσιάζουμε και συζητάμε ποιοτικά αποτελέσματα, όπως απεικονίζονται στο Σχήμα 7.10. Προβάλλουμε τις 3Δ συντεταγμένες των δεικτών και της πόζας στις υπέρυθρες προβολές για να τις συσχετίσουμε με τις πραγματικές κινήσεις των ηθοποιών. Για λόγους σαφήνειας, απεικονίζουμε ξεχωριστά την ακατέργαστη θορυβώδη είσοδο (D415 δείκτες), τα ground truth δεδομένα και τους εκτιμώμενους δείκτες. Οι ground truth και οι εκτιμώμενες πόζες απεικονίζονται μαζί για να διευκολύνουμε την οπτική σύγκριση μεταξύ τους.

Η ύπαρξη των υπέρυθρων εικόνων στο φόντο, μας επιτρέπει να επισημάνουμε τις προκλήσεις που επιλύουμε με το μοντέλο μας. Συγκεκριμένα, υποδεικνύουμε τις διορθώσεις των δεικτών στη θορυβώδη είσοδο και την ακριβή πρόβλεψη της πόζας. Η είσοδος σημειακών δεικτών που καταγράφεται με το χαμηλού κόστους σύστημα D415 είναι ιδιαίτερα θορυβώδης. Αυτό οφείλεται στον 3Δ εντοπισμό των δεικτών από κάθε αισθητήρα βάθους ξεχωριστά, οι οποίοι, σε συνδυασμό με τα συστηματικά σφάλματα υπολογισμού των χαρτών βάθους, οδηγούν σε απομακρυσμένα μεταξύ τους 3Δ σημεία που, παρόλα αυτά, αντιπροσωπεύουν την 3Δ θέση του ίδιου δείκτη. Συγκρίνοντας τις εκτιμώμενες θέσεις των δεικτών σε σχέση με την αρχική είσοδο, επισημαίνουμε την αποθορυβοποίηση που επιτυγχάνει το προτεινόμενο μοντέλο. Η συγχώνευση των παρατηρήσεων των δεικτών επιτυγχάνεται με την αυτόματη ομαδοποίηση των θορυβωδών 3Δ σημείων που έχουν καταγραφεί από διαφορετικούς αισθητήρες (Σχ. 7.10, κίτρινοι κύκλοι). Επιπρόσθετα, οι δείκτες 'φαντάσματα' που προκύπτουν από λανθασμένη ανίχνευση κατά την καταγραφή αγνοούνται (Εικ. 7.10, κόκκινοι κύκλοι). Οι μη θεατοί δείκτες, είτε αποκρυπτόμενοι είτε μη ανιχνευθέντες, ανακτώνται (Εικ. 7.10, κύκλοι ματζέντα). Επίσης, η θόλωση της εικόνας αντιμετωπίζεται αποτελεσματικά για την εξάλειψη των σφαλμάτων εκτίμησης των συντεταγμένων των αρθρώσεων (Εικ. 7.10, πράσινοι κύκλοι). Με άλλα λόγια, εντοπίζουμε και επισημαίνουμε με ακρίβεια τους οπτικούς δείκτες που καταγράφονται με αισθητήρες χαμηλού κόστους που μας παρέχουν δεδομένα με υψηλό θόρυβο. Λόγω της διακριτής εξαγωγής αποτελεσμάτων σε κάθε μεμονωμένο καρέ, επιλύεται η εναλλαγή δεικτών που είναι συχνή στα παραδοσιακά συστήματα καταγραφής κίνησης. Το μοντέλο μας παρέχει άμεσα στιγμιαία εκτίμηση 3Δ πόζας από επισημασμένους 3Δ δείκτες χωρίς τη χρήση οποιουδήποτε ανθρωποειδούς προτύπου ή γνώσης για τη δομή του σώματος, πράγμα που σημαίνει ότι δεν υπάρχει καμία προηγούμενη πληροφορία όσον αφορά τα μήκη των οστών και τη σχετική θέση των αρθρώσεων, αντίθετα, το μοντέλο προβλέπει τη στάση στο φυσικό χώρο και στα φυσικά μεγέθη σε κάθε εκτίμηση μεμονωμένα. Παρ' όλα αυτά, είναι σημαντικό να τονιστεί ότι τα μοντέλα DeMoCap εκπαιδεύονται με βάση μια συγκεκριμένη τοποθέτηση των δεικτών, πράγμα που σημαίνει ότι διαφορετική τοποθέτηση μπορεί να οδηγήσει το μοντέλο σε λανθασμένες προβλέψεις.

Στο Σχήμα 7.11, παρουσιάζουμε περισσότερα ποιοτικά αποτελέσματα με τη χρήση 3Δ οπτικοποιήσεων. Απεικονίζουμε 20 δείγματα από το σύνολο δοκιμών, απεικονίζοντας τα ground truth δεδομένα με μπλε χρώμα και τις προβλέψεις μας με κίτρινο, συμπεριλαμβανομένων των περιπτώσεων αποτυχίας στην τελευταία σειρά.



Σχήμα 7.10: *Στις τρεις πρώτες στήλες, απεικονίζουμε τους θορυβώδεις (κίτρινο), ground truth (μπλε) και προβλεπόμενους (πράσινο) δείκτες προβαλλόμενους σε σε μία μόνο από τις υπέρυθρες προβολή του καρέ πολλαπλών προβολών (διαφορετικό καρέ ανά σειρά). Στην τέταρτη στήλη απεικονίζονται οι προβλεπόμενες (κίτρινο) και ground truth (μπλε) πόζες. Οι κόκκινοι, μωβ και κίτρινοι κύκλοι υποδεικνύουν την αποθορυβοποίηση των δεικτών 'φαντασμάτων', την ανάκτηση των ελλιπών δεικτών και τα σφάλματα του αισθητήρα βάθους, αντίστοιχα. Οι πράσινοι κύκλοι υποδεικνύουν τα προβλήματα θόλωσης που αντιμετωπίζει επιτυχώς το μοντέλο.*

Σχήμα 7.11: *Ποιοτικά αποτελέσματα της ground truth (μπλε) και της προβλεπόμενης (κίτρινο) πόζας στο 3Δ χώρο. Το μοντέλο μας εκτιμά 3Δ πόζες συγκρίσιμες με τις ground truth. Στην τελευταία σειρά απεικονίζονται περιπτώσεις αποτυχίας, όταν οι πόζες των υποκειμένων είναι εξαιρετικά απαιτητικές.*

7.3.5 Μελέτη συνεισφορών

Διεξήγαμε και συζητάμε μια εκτεταμένη μελέτη συνεισφορών για να επικυρώσουμε το σχεδιασμό της προτεινόμενης μεθόδου. Αντικαθιστούμε, αφαιρούμε ή ρυθμίζουμε διαφορετικά μία μόνο συνεισφορά της προσέγγισής μας ανά πείραμα, παρουσιάζοντας τη βαρύτητά της στον

αναγνώστη ξεχωριστά. Αναλυτικότερα:

1. αντικαθιστούμε το νεοεισαχθέν, πλήρως διαφοροποιήσιμη *CoM3D* (Κεφ. 7.2.2) με άλλο state-of-the-art επιτελική επίπεδο, το integral 3D regression module [71],
2. εξετάζουμε την είσοδο/επίβλεψη μονής έναντι διπλής όψης (Κεφ. 7.2.2),
3. εξετάζουμε την είσοδο/επίβλεψη τετραπλής έναντι διπλής όψης (Κεφ. 7.2.2),
4. εξετάζουμε την πόλωση εξαιτίας κβαντισμού κατά την απόδοση μεταξύ υψηλής (Κεφ. 7.1.4) και χαμηλής ανάλυσης της εισόδου, δηλ. των χαρτών βάθους,
5. αφαιρούμε τη χρήση της επαύξησης των δεδομένων (Κεφ. 7.1.4),
6. αφαιρούμε τη χρήση της κανονικοποίησης των δεδομένων (Κεφ. 7.1.4),
7. αφαιρούμε τη χρήση ενδιάμεσης συνάθροισης των χαρτών θερμότητας (Κεφ. 7.2.1) έναντι της χρήσης αποκλειστικά του τελευταίου σταδίου πρόβλεψης χαρτών θερμότητας,
8. εξετάζουμε την απόδοση του μοντέλου μας σε δεδομένα δεικτών που έχουν καταγραφεί μόνο από 3 κάμερες,
9. εξετάζουμε την απόδοση του μοντέλου μας σε δεδομένα δεικτών που έχουν καταγραφεί μόνο από 2 αντιδιαμετρικά τοποθετημένες κάμερες.

Στη συνέχεια της ενότητας ακολουθούμε την ίδια αρίθμηση, όπως και στα αποτελέσματα που παρουσιάζονται στον πίνακα 7.6.

1. Integral 3D regression *vs.* CoM3D. Μια από τις κύριες συνεισφορές της μεθόδου μας είναι η εισαγωγή της πλήρως διαφοροποιήσιμης μονάδας *CoM3D* για την 3Δ παλινδρόμηση που περιλαμβάνει ένα επίπεδο *zMean* ακολουθούμενο από *Softmax* και ένα επίπεδο *CoM*. Η κύρια διαφορά σε σχέση με άλλες προσεγγίσεις χωρικής παλινδρόμησης είναι η χρήση του *zMean* για την παλινδρόμηση με συντεταγμένες z. Για να αξιολογήσουμε την αξία του, εκπαιδεύουμε το μοντέλο μας αντικαθιστώντας τη μονάδα *CoM3D* με την ολοκληρωτική παλινδρόμηση πόζας που προτείνεται από την [71]. Όπως φαίνεται στον πίνακα 7.6, η χρήση της ολοκληρωτικής μονάδας παλινδρόμησης πόζας δεν είναι αρκετά αποτελεσματική στην εργασία μας με αποτέλεσμα να έχουμε μεγαλύτερα σφάλματα από το αρχικό μοντέλο ($83.68mm$ και $89.32mm$ έναντι $33.83mm$ και $40.04mm$ $\mathbf{M}_{\text{PJPE}}$ στα σύνολα επικύρωσης και δοκιμής, αντίστοιχα), ενώ ο χρόνος εξαγωγής συμπερασμάτων αυξάνεται αισθητά.

Αριθμός προβολών απόδοσης. Η είσοδος του μοντέλου επιλογής είναι ένα ζεύγος χαρτών βάθους που προκύπτουν από την απόδοση των 3Δ θέσεων των ανακλαστικών δεικτών από δύο αντίθετες προβολές. Εννοιολογικά, εκμεταλλευόμαστε την 3Δ πληροφορία μας που μας επιτρέπει να 'δημιουργήσουμε' πολυάριθμες 2Δ εισόδους από ένα μόνο 3Δ δείγμα, υποστηρίζοντας ότι η είσοδος και η εποπτεία πολλαπλών προβολών οδηγεί σε υψηλότερης ακρίβειας αποτελέσματα. Για να αξιολογήσουμε αυτόν τον ισχυρισμό, διεξάγουμε δύο πειράματα ρυθμίζοντας τον αριθμό των προβολών απόδοσης.

2. 1 *vs.* 2 χάρτες βάθους εισόδου. Δημιουργούμε μόνο ένα χάρτη βάθους και εκπαιδεύουμε το δίκτυο με είσοδο μονής όψης. Όπως απεικονίζεται στον Πίνακα 7.6 (No.

2), αυτό το μοντέλο παρουσιάζει χαμηλότερη απόδοση σε σύγκριση με την προτεινόμενη προσέγγιση διπλής όψης σε όλες τις μετρικές, επιβεβαιώνοντας ότι η τεχνική πολλαπλών όψεων οδηγεί το μοντέλο σε πιο ακριβείς και αξιόπιστες προβλέψεις.

3. 4 *vs.* 2 χάρτες βάθους εισόδου. Επιπλέον, διεξάγεται ένα ακόμη πείραμα με αυξημένο αριθμό προβολών απόδοσης του οποίου τα αποτελέσματα φαίνονται στον Πίνακα 7.6 (No. 3). Εκπαιδεύουμε και αξιολογούμε ένα μοντέλο με είσοδο 4 όψεων, το οποίο παρουσιάζει υψηλότερη απόδοση από το προτεινόμενο αρχικό μοντέλο διπλής όψης. Αναλυτικότερα, το μοντέλο αποδίδει παρόμοια στο σύνολο επικύρωσης, παρουσιάζοντας ωστόσο σαφή υπεροχή στο σύνολο δοκιμών σε όλες τις μετρικές, περίπου $3mm$ απόλυτη βελτίωση σε όλες τις μετρικές σφάλματος που βασίζονται στην ευκλείδεια απόσταση, 1.72% $\text{mAP}_{50\text{mm}}$ και 3.67° και 5.41° M_{PJAE} και RMS_{PJAE}, αντίστοιχα, δεδομένης της υψηλότερης αξιοπιστίας των αποτελεσμάτων όταν εκπαιδεύονται σε είσοδο πολλαπλών προβολών και βασίζονται σε μεγαλύτερο αριθμό εκτιμήσεων πολλαπλών όψεων.

Το συμπέρασμα για αυτά τα πειράματα είναι τριπλό, πρώτον, παρά την αραιή κατανομή των δεικτών, εξακολουθούν να υπάρχουν ασάφειες μίας όψης σε δύσκολες και πολύπλοκες στάσεις του σώματος, με αποτέλεσμα τοπικά πυκνά υποσύνολα και πιθανές αποκρύψεις δεικτών, τις οποίες η απόδοση πολλαπλών όψεων μπορεί να ξεπεράσει. Δεύτερον, αυξάνοντας τον αριθμό των προβολών απεικόνισης, βελτιώνεται η αξιοπιστία και η ακρίβεια των προβλέψεων, επιβεβαιώνοντας τους ισχυρισμούς μας όσον αφορά τη συμβολή της εποπτείας πολλαπλών προβολών. Τέλος, η αύξηση του αριθμού των εισόδων βάθους απόδοσης αυξάνει γραμμικά την υπολογιστική πολυπλοκότητα και το κόστος απόδοσης του μοντέλου. Θεωρούμε ότι η χρήση δύο αντίθετων απεικονίσεων βάθους είναι ιδανική ως συμβιβασμός μεταξύ αποτελεσματικότητας και αποδοτικότητας, δεδομένων των συγκρίσιμων αποτελεσμάτων τους.

4. Υψηλή *vs.* χαμηλή ανάλυση απόδοσης εισόδου. Αναπτύσσουμε το DeMoCap θέτοντας ένα αμιγώς 3Δ πρόβλημα ως ένα πρόβλημα 3Δ παλινδρόμησης από πολλαπλές εισόδους 2.5Δ (βάθους), επιτρέποντας τη χρήση 2Δ πλήρως συνελικτικών αρχιτεκτονικών με αποδεδειγμένη και αξιοσημείωτη αποτελεσματικότητα σε εργασίες εντοπισμού σημείων. Εξαιτίας αυτού, υποχρεούμαστε να υποστούμε το κόστος της κωδικοποίησης των 3Δ μη κβαντισμένων δεδομένων σε κβαντισμένα 2Δ πλέγματα, οδηγώντας σε κάποιου βαθμού απώλεια πληροφορίας. Συγκεκριμένα, οι κατασκευασμένοι χάρτες βάθους που βασίζονται σε κβαντισμένες συντεταγμένες δεν είναι απολύτως ακριβείς, οδηγώντας σε μη βέλτιστη εποπτεία και υποβαθμισμένη απόδοση του μοντέλου, όπως αναλύσαμε στο Κεφ. 6. Περιορίζουμε το σφάλμα κβαντισμού και την απώλεια πληροφορίας με την απόδοση των χαρτών βάθους σε υψηλή ανάλυση εικονοστοιχείων, 800×800 συγκεκριμένα, και τη γραμμική κλιμάκωση τους στην ανάλυση εισόδου μας, δηλαδή 160×160.

Στον πίνακα 7.6 (No. 4), εκπαιδεύουμε και αξιολογούμε το μοντέλο με δεδομένα που αποδίδονται απευθείας σε χαμηλή ανάλυση, κωδικοποιώντας έτσι υψηλότερα σφάλματα κβαντισμού. Αυτό το μοντέλο παρουσιάζει χαμηλότερη απόδοση, επικυρώνοντας την άποψη ότι η απόδοση χαμηλής ανάλυσης οδηγεί σε πολωμένη παλινδρόμηση συντεταγμένων. Αξίζει να σημειωθεί ότι τα σφάλματα δεν είναι εξαιρετικά υψηλότερα, υποθέτοντας ότι η εποπτεία πολλαπλών προβολών μπορεί να χειριστεί καλύτερα αυτή την πόλωση, καθώς και τα χαρακτηριστικά της 3Δ παλινδρόμησης συντεταγμένων σε επίπεδο υπο-εικονοστοιχείου της μονάδας αποκωδικοποίησης συντεταγμένων χαρτών θερμότητας, *CoM3D*.

5. Χωρίς *vs.* με επαύξηση δεδομένων. Για να καταδείξουμε τη συμβολή της 3Δ περιστροφικής επαύξησης των δεδομένων μας (Κεφ. 7.1.4), εκπαιδεύουμε το μοντέλο μας αποκλείοντάς τη (Πίνακας 7.6 (No. 5). Η είσοδος του αραιού νέφους σημείων μας παρέχει το προνόμιο να το περιστρέψουμε πριν την προβολή, αυξάνοντας σε σημαντικό βαθμό το αριθμό των νέων διαφορετικών δειγμάτων χαρτών βάθους κατά την εκπαίδευση αποδίδοντας σαφώς καλύτερα αποτέλεσμα όπως φαίνεται στα αποτελέσματα, μια σημαντική διαφορά σε σύγκριση με τους περιορισμούς της ψευδο-περιστροφικής επαύξησης που εφαρμόζεται σε πυκνά 2Δ δεδομένα.

6. Χωρίς *vs.* με κανονικοποίηση δεδομένων. Στον πίνακα 7.6 (No. 6), αξιολογούμε τη συμβολή του ογκομετρικού μετασχηματισμού κανονικοποίησης κλίμακας και μετατόπισης $\mathcal{T}_N$ που εκτελούμε στο νέφος σημείων δεικτών ώστε να καταλαμβάνει ίσο όγκο στο κανονικοποιημένο ογκομετρικό πλέγμα για όλα τα δείγματα. Παρατηρούμε ότι αυτή η κανονικοποίηση ενισχύει την απόδοση του μοντέλου μας σε όλες τις μετρικές. Η εκτίμησή μας είναι ότι αυτός ο μετασχηματισμός οδηγεί σε υψηλή διακύμανση σε σχέση με τις δομές του ανθρώπινου σώματος, επιτρέποντας στο μοντέλο να μάθει μια πιο ελεγχόμενη και συνεπή αναπαράσταση ανεξάρτητα από τις φυσικές διαστάσεις της 3Δ πόζας.

7. Χωρίς *vs.* με άθροιση χαρτών θερμότητας. Στην προσέγγισή μας, αντί να επιβλέπουμε τους θερμικούς χάρτες όπως είχε αρχικά προταθεί για τις αρχιτεκτονικές στις οποίες στηριζόμαστε, επιβλέπουμε και μετα-επεξεργαζόμαστε μόνο τους αθροισμένους χάρτες θερμότητας. Η συνάθροιση οδηγεί το μοντέλο μας σε ταχύτερη σύγκλιση και ελαφρώς καλύτερα αποτελέσματα, ειδικά για την εκτίμηση των δεικτών, όπως φαίνεται στον πίνακα 7.6 (No. 7), σε σύγκριση με την επίβλεψη μόνο του τελευταίου σταδίου (HRNetV1), όπως προτείνεται για το HRNet στην αντίστοιχη εργασία [117].

Αριθμός αισθητήρων βάθους. Τέλος, διεξάγουμε πειράματα του μοντέλου μας στα σύνολα επικύρωσης και δοκιμής λαμβάνοντας υπόψη τις παρατηρήσεις δεικτών μόνο από 3 και 2 αισθητήρες (Πίνακας 7.6 (No. 8 και No. 9)), αντίστοιχα, αντί της πλήρους διάταξης 4 αισθητήρων με την οποία εκπαιδεύσαμε το DeMoCap. Παρουσιάζουμε αυτό το πείραμα για να αξιολογήσουμε την πόλωση του μοντέλου μας στο σύνολο εκπαίδευσης και τις δυνατότητες γενίκευσης, καθώς και την ευαισθησία του στη μείωση των αισθητήρων/προβολών.

8. 3 *vs.* 4 αισθητήρες βάθους. Όπως ήταν αναμενόμενο, η ακρίβεια είναι χαμηλότερη σε σχέση με τα δεδομένα που καταγράφονται με 4 αισθητήρες, ωστόσο, τα αποτελέσματα εξακολουθούν να είναι καλύτερα από τις συγκρινόμενες μεθόδους, παρουσιάζοντας χαμηλότερη σφάλματα $\mathrm{M_{PJPE}}$ και $\mathrm{RMS_{PJPE}}$ και υψηλότερη ακρίβεια $\mathrm{mAP_{50mm}}$ ($46.81mm$, $53.50mm$ και 77.17% στο σύνολο επικύρωσης και $47.58mm$, $60.02mm$ και $81.88mm$ στο σύνολο δοκιμών, αντίστοιχα).

9. 2 *vs.* 4 αισθητήρες βάθους. Στην αξιολόγηση με 2 αισθητήρες, η απόδοση μειώνεται περαιτέρω, ωστόσο τα αποτελέσματα μπορούν να θεωρηθούν υπολογίσιμα δεδομένης της έλλειψης πληροφορίας. Το συμπέρασμά μας από αυτό το πείραμα είναι ότι το DeMoCap ακολουθεί μια αναμενόμενη αναλογία/εξάρτηση από τον αριθμό των αισθητήρων που καταγράφουν τους δείκτες, όπως τα περισσότερα συστήματα καταγραφής κίνησης.

Πίνακας 7.6: ***Αποτελέσματα ανάλυσης συνεισφορών.*** *Αφαιρούμε μία προς μία τις συνεισφορές του μοντέλου μας για να αναδείξουμε την αποτελεσματικότητα και την αναγκαιότητά τους.*

	Method \ Metrics (mm / °)	**Set**	$\mathbf{M_{PJPE}}\downarrow$	$\mathbf{RMS_{PJPE}}\downarrow$	$\mathbf{M_{PMPE}}\downarrow$	$\mathbf{RMS_{PMPE}}\downarrow$	$\mathbf{mAP_{50mm}}\uparrow$	$\mathbf{M_{PJAE}}\downarrow$	$\mathbf{RMS_{PJAE}}\downarrow$
#1.	Integral 3D regression *vs.* CoM3D	val	83.68	96.02	109.77	121.34	24.53%	29.88	33.66
#2.	1- *vs.* 2-view depth input		38.04	46.48	48.67	59.85	85.55%	19.57	22.83
#3.	4- *vs.* 2-view depth input		**33.81**	42.86	44.01	55.30	89.70%	19.48	25.62
#4.	High- *vs.* low-resolution rendering		34.91	43.71	43.69	53.44	89.44%	19.49	23.86
#5.	W/o *vs.* w/ data augmentation		64.82	90.94	85.53	115.81	50.42%	35.44	46.02
#6.	W/o *vs.* w/ data normalization		38.11	47.53	67.99	78.07	87.04%	20.56	23.75
#7.	W/o *vs.* w/ heatmap aggregation		34.73	43.25	48.52	57.74	88.33%	**17.48**	**20.49**
#8.	3 *vs.* 4 capturing depth sensors		46.81	53.50	57.90	66.40	77.17%	22.79	27.22
#9.	2 *vs.* 4 capturing depth sensors		59.86	87.50	73.73	100.88	64.66%	28.08	39.08
	DeMoCap		33.83	**42.65**	**42.33**	**51.74**	**90.41%**	18.66	22.47
#1.	Integral 3D regression *vs.* CoM3D	test	89.32	100.09	116.42	126.91	22.14%	30.89	36.29
#2.	1- *vs.* 2-view depth input		41.87	54.17	56.71	71.44	86.73%	18.98	24.65
#3.	4- *vs.* 2-view depth input		**37.13**	**48.14**	**49.64**	**63.22**	**89.72%**	**16.06**	**20.77**
#4.	High- *vs.* low-resolution rendering		41.50	53.01	53.49	67.06	87.76%	17.29	22.52
#5.	W/o *vs.* w/ data augmentation		63.12	79.83	80.45	100.82	61.36%	24.08	31.93
#6.	W/o *vs.* w/ data normalization		44.19	56.22	72.34	91.28	86.73%	18.64	23.86
#7.	W/o *vs.* w/ heatmap aggregation		40.56	51.97	58.09	72.21	87.76%	16.91	22.10
#8.	3 *vs.* 4 capturing depth sensors		47.58	60.02	61.33	76.30	81.88%	19.25	25.35
#9.	2 *vs.* 4 capturing depth sensors		59.03	74.46	76.73	95.21	67.97%	24.26	33.09
	DeMoCap		40.04	51.69	52.92	66.49	88.05%	19.73	26.18

7.3.6 Μελέτη σε καθαρά δεδομένα δεικτών

Στο σύνολο δεδομένων του DeMoCap

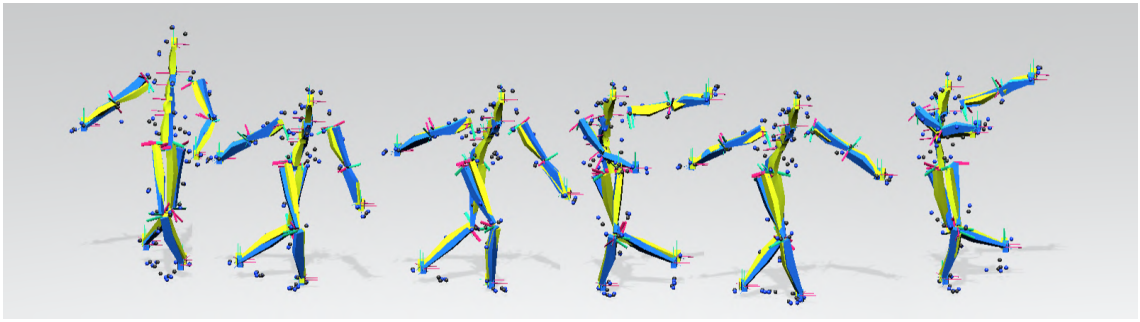
Αξιολογούμε περαιτέρω το μοντέλο μας με την εκπαίδευση και την αξιολόγησή του χρησιμοποιώντας ως είσοδο τα μετα-επεξεργασμένα, καθαρά δεδομένα δεικτών από τη VICON που χρησιμοποιήθηκαν ως ground-truth, με χρήση των ίδιων υπερ-παραμέτρων για την εκπαίδευση. Αυτά τα πειράματα αναδεικνύουν τη συμπεριφορά του μοντέλου σε ιδανικές συνθήκες, όπου τα δεδομένα δεικτών είναι απολύτως αποθορυβοποιημένα και υψηλής ακρίβειας, χωρίς το θόρυβο που υπάρχει στα αρχικά οπτικά δεδομένα δεικτών, είτε αυτά έχουν ληφθεί με αισθητήρες βάθους χαμηλού κόστους είτε με συστήματα καταγραφής κίνησης υψηλών προδιαγραφών πριν από τη μετα-επεξεργασία. Στον Πίνακα 7.7, παρουσιάζουμε τα αποτελέσματα του DeMoCap και του DeMoCap που εκπαιδεύτηκε με δεδομένα VICON ($\text{DeMoCap}_{\text{vicon}}$), τα οποία αξιολογήθηκαν τόσο σε καθαρά δεδομένα VICON όσο και σε θορυβώδη δεδομένα από αισθητήρες βάθους καταναλωτικής ποιότητας (RS).

Όπως αναμενόταν, το $\text{DeMoCap}_{\text{vicon}}$ επιτυγχάνει σημαντικά υψηλή ακρίβεια τόσο στα σύνολα επικύρωσης όσο και στα σύνολα δοκιμής VICON, επιτυγχάνοντας M_{PJPE} και M_{PMPE} χαμηλότερα από $3cm$ και $\text{mAP}_{50\text{mm}}$ 99.81% και 94.07% σε κάθε σύνολο, αντίστοιχα. Από την άλλη, το $\text{DeMoCap}_{\text{vicon}}$ παρουσιάζει χαμηλή απόδοση στα θορυβώδη δεδομένα RS υπερβαίνοντας τα $6cm$ για τα σφάλματα απόλυτης ευκλείδειας απόστασης και η μέση ακρίβεια $\text{mAP}_{50\text{mm}}$ είναι χαμηλότερη από 56%, χαμηλότερη ακόμη και από το DeMoCap που αξιολογήθηκε σε δεδομένα μόνο 2 αισθητήρων, προφανώς δεδομένης της ανομοιότητας των συνόλων αξιολόγησης σε σύγκριση με το σύνολο εκπαίδευσης.

Αποτελέσματα ιδιαίτερου ενδιαφέροντος παρουσιάζει το μοντέλο μας όταν αξιολογείται στο καθαρό σύνολο επικύρωσης και δοκιμής της VICON. Το DeMoCap επιδεικνύει σημαντικά καλύτερες επιδόσεις στα δεδομένα VICON από ό,τι στα δεδομένα RS, αν και εκπαιδεύτηκε στα τελευταία, επιτρέποντάς μας να θεωρήσουμε ότι το μοντέλο παρουσιάζει καλή γενίκευση, χωρίς πόλωση στο συστηματικό θόρυβο της κάμερας, τις πόζες ή τις εσωτερικές της παραμέτρους. Το μοντέλο επιτυγχάνει υψηλή ακρίβεια σε θορυβώδη και καθαρά δεδομένα, δείχνοντας ότι η αύξηση της ακρίβειας καταγραφής των δεικτών οδηγεί σε πιο ακριβείς εκτιμήσεις.

Πίνακας 7.7: ***Αποτελέσματα του DeMoCap με καθαρά δεδομένα δεικτών.*** *Εκπαιδεύουμε το DeMoCap με καθαρά δεδομένα δεικτών για να αξιολογήσουμε την απόδοση των μοντέλων σε διάφορους συνδυασμούς μεταξύ συνόλων εκπαίδευσης και επικύρωσης/ελέγχου.*

Method \ Metrics (mm)	**Set**	$\mathbf{M}_{\text{PJPE}}$	$\mathbf{RMS}_{\text{PJPE}}$	$\mathbf{M}_{\text{PMPE}}$	$\mathbf{RMS}_{\text{PMPE}}$	$\mathbf{mAP}_{\mathbf{50mm}}$	$\mathbf{M}_{\text{PJAE}}$	$\mathbf{RMS}_{\text{PJAE}}$
DeMoCap on VICON data	val	29.64	34.58	36.61	42.95	96.11%	16.28	18.60
$\text{DeMoCap}_{\text{vicon}}$ on RS data		72.76	108.94	94.75	135.48	45.11%	31.42	41.00
$\text{DeMoCap}_{\text{vicon}}$ on VICON data		**21.04**	**24.24**	**19.09**	**24.49**	**99.81%**	**11.18**	**13.52**
DeMoCap		33.83	42.65	42.33	51.74	90.41%	18.66	22.47
DeMoCap on VICON data	test	37.28	47.63	47.80	59.79	89.86%	14.52	19.40
$\text{DeMoCap}_{\text{vicon}}$ on RS data		62.81	87.50	90.78	122.57	55.09%	28.34	35.88
$\text{DeMoCap}_{\text{vicon}}$ on VICON data		**25.16**	**32.98**	**27.44**	**37.30**	**94.07%**	**10.04**	**15.28**
DeMoCap		40.04	51.69	52.92	66.49	88.05%	19.73	26.18



Σχήμα 7.12: *Ποιοτικά αποτελέσματα διαφόρων καρέ που απεικονίζονται στην ίδια σκηνή (κίνηση) από την ακολουθία DanceTurns002 του SFU Dataset [12], σε εντελώς αθέατες δομές σώματος και δραστηριότητες. Οι κίτρινες πόζες αντιπροσωπεύουν τις προβλέψεις του DeMoCap, ενώ οι μπλε τα ground-truth δεδομένα του συνόλου δεδομένων.*

Πίνακας 7.8: *Αποτελέσματα των μοντέλων στα καθαρά δεδομένα καταγραφής κίνησης του συνόλου δεδομένων SFU [12].*

Model \ Metrics (mm)	$\mathbf{M}_{\mathrm{PJPE}}$	$\mathbf{RMS}_{\mathrm{PJPE}}$	$\mathbf{mAP}_{\mathbf{50mm}}$
DeMoCap	58.28	69.38	48.95%
$\mathrm{DeMoCap}_{\mathrm{vicon}}$	**45.28**	**54.95**	**75.46%**

Στο σύνολο δεδομένων SFU

Τέλος, αξιολογούμε την απόδοση των μοντέλων μας, DeMoCap και $\mathrm{DeMoCap}_{\mathrm{vicon}}$ σε ένα δημόσιο σύνολο δεδομένων με σχετικά παρόμοια προσάρτηση δεικτών και δομή πόζας με 53 δείκτες και 30 αρθρώσεις, τη βάση δεδομένων καταγραφής κίνησης SFU Motion Capture Database [12]. Ενδεικτικά, στα πειράματά μας, περιλαμβάνουμε δύο απαιτητικές δραστηριότητες, *DanceTurns002*[5] και *HopOverObstacle001*[6]. Τα ποσοτικά αποτελέσματα για 575 δείγματα συνολικά παρουσιάζονται στον πίνακα 7.8. Οπτικά, τα μοντέλα παρουσιάζουν συγκρίσιμα αποτελέσματα, όπως απεικονίζεται στο Σχήμα 7.12, αριθμητικά όμως, μόνο το $\mathrm{DeMoCap}_{\mathrm{vicon}}$ επιτυγχάνει σχετικά υψηλή ακρίβεια, ενώ για το DeMoCap, αποδεικνύεται πιο απαιτητικό. Αξίζει να τονιστούν οι χωρικές μετατοπίσεις που υπάρχουν μεταξύ των διαφόρων δομών σώματος μεταξύ διαφορετικών συνόλων δεδομένων, οι οποίες εισάγουν ένα σταθερό σφάλμα στις μετρήσεις, όπως αναλύεται στην ενότητα 7.4.

7.4 Συμπεράσματα

Σε αυτή την ενότητα, παρουσιάζουμε μια σύνοψη των παρατηρήσεών μας, συζητώντας τα πλεονεκτήματα της προτεινόμενης μεθόδου καταγραφής κίνησης. Εξ όσων γνωρίζουμε, το DeMoCap είναι η πρώτη μέθοδος υπολογιστικής όρασης που επιτρέπει τη χρήση εξοπλισμού χαμηλού κόστους για την καταγραφή κίνησης με δείκτες επιτυγχάνοντας συγκρίσιμα αποτελέσματα με τα συστήματα καταγραφής υψηλής τεχνολογίας. Η DeMoCap είναι μία από τις πρωτοποριακές μεθόδους καταγραφής κίνησης βαθιάς μάθησης με δείκτες που επιτρέπει την ενιαία παλινδρόμηση της πόζας από ένα αραιό σύνολο 3Δ σημείων. Η μέθοδος αποδίδει κα-

[5] http://mocap.cs.sfu.ca/index154af.html?id=0018_DanceTurns002.bvh
[6] http://mocap.cs.sfu.ca/index1fe61.html?id=0015_HopOverObstacle001.bvh

λύτερα από τις πρόσφατες σύγχρονες μεθόδους βασισμένες σε έγχρωμες εικόνες (π.χ. LT και 4DA) παρά τη χρήση εξαιρετικά λανθασμένων εκτιμήσεων βάθους από αισθητήρες χαμηλού κόστους (σφάλμα βάθους μεγαλύτερο από $3cm$ σε απόσταση $1.5m$ από την κάμερα), ενώ το μέσο σφάλμα ανά άρθρωση πέφτει κάτω από 2.5 cm όταν εκπαιδεύεται και αξιολογείται σε καθαρά δεδομένα. Το DeMoCap γενικεύεται καλά ακόμη και με τη χρήση χαμηλού αριθμού καμερών (2 ή 3 αισθητήρες, Κεφ. 7.3.5), επιδεικνύοντας αυξημένη σταθερότητα σε σύγκριση με μεθόδους που βασίζονται σε δυνητικά 'αδύναμη' ανίχνευση (π.χ. 2Δ ανιχνευτές πόζας). Το μοντέλο οδηγείται στην απόρριψη λανθασμένων ανιχνεύσεων και στην ανάκτηση εκλιπόντων δεικτών στο πρώτο στάδιο (αποθορυβοποίηση δεικτών), επιτρέποντας την εκτίμηση της πόζας από πρωτύτερα αποθορυβοποιημένη πληροφορία.

Το DeMoCap επικεντρώνεται αποκλειστικά σε δεδομένα που αφορούν και επιλύουν την ανθρώπινη πόζα, δηλαδή στους δείκτες που είναι προσαρτημένοι στο σώμα, χωρίς άλλες παρεμβολές του περιβάλλοντος, όπως συμβαίνει με τα δεδομένα χρώματος ή τα πυκνά δεδομένα βάθους. Τέλος, το DeMoCap αποδίδει καλύτερα όταν μειώνεται ο θόρυβος (όπως αξιολογείται στο Κεφ. 7.3.6), παρά την ύπαρξη συστηματικού θορύβου των αισθητήρων βάθους στο σύνολο εκπαίδευσης, δείχνοντας ότι το μοντέλο, επηρεάζεται κατά κύριο λόγο μονάχα από την ποιότητα της παρακολούθησης των δεικτών, όπως όλα τα συστήματα καταγραφής κίνησης με δείκτες.

Κεφάλαιο 8

Συμπεράσματα και μελλοντικές κατευθύνσεις

8.1 Συμπεράσματα

Κατά την πορεία της παρούσας διατριβής και με την ερευνητική εμπειρία και μελέτη γύρω από θέματα και σύγχρονες μεθόδους υπολογιστικής όρασης και βαθιάς μάθησης επικεντρωμένες στην ανάλυση, καταγραφή και αναπαράσταση ανθρώπινων ενεργειών-κινήσεων, οι κύριες ερευνητικές συνεισφορές πραγματοποιήθηκαν στο πεδίο καταγραφής/ψηφιοποίησης ανθρώπινης κίνησης στον 3Δ χώρο με την ανάπτυξη νέων τεχνικών και μεθόδων βαθιάς μάθησης. Θέτοντας ως κύριο ερευνητικό στόχο την καταγραφή κίνησης με οπισθοανακλαστικούς δείκτες επαγγελματικών συστημάτων υψηλού κόστους, αναπτύχθηκαν καινοτόμες τεχνικές και μοντέλα βαθιάς μάθησης για την επίλυση υπαρχόντων προβλημάτων που εντοπίζονται και στα επαγγελματικά συστήματα που χρησιμοποιούν δείκτες, είτε νέων που προκύπτουν από τις ελαττωμένες δυνατότητες των εξοπλισμών χαμηλού κόστους. Συγκεκριμένα, η καταγραφή κίνησης με χρήση δεικτών αποτελεί το αξεπέραστο τεχνολογικό πρότυπο για καταγραφή και παρακολούθηση κίνησης υψηλής πιστότητας εδώ και δεκαετίες, ωστόσο, παρά την υψηλή ακρίβεια, η χρήση των συστημάτων αυτών είναι παγκοσμίως περιορισμένη, δεδομένου του υψηλού κόστους του εξοπλισμού, των αδειών λογισμικού και της συντήρησης, και άλλων παραγόντων.

Στην παρούσα διατριβή αναπτύχθηκαν νέες μέθοδοι που επιτρέπουν τη χρήση εξοπλισμού χαμηλού κόστους για την καταγραφή 3Δ κίνησης με δείκτες παράγοντας αποτελέσματα συγκρίσιμα με τα συστήματα υψηλού κόστους, στο βαθμό που το επιτρέπουν οι περιορισμοί εξοπλισμού και δεδομένων. Πρακτικά, η χρήση υπολογιστικής όρασης πολλαπλών προβολών, τεχνολογιών καταγραφής χαρτών βάθους και βαθιάς μάθησης συν-αξιοποιούνται για την ενίσχυση των αδυναμιών που φέρουν οι χαμηλού κόστους αισθητήρες βάθους που βρίσκει κανείς στο εμπόριο. Για τη μελέτη και περαιτέρω διερεύνηση του συγκεκριμένου ερευνητικού στόχου, δημιουργήθηκαν δύο μεγάλα σε όγκο και ποικιλία κινήσεων σύνολα δεδομένων, το DMC και το HUMAN4D, εκ των οποίων το πρώτο συμπεριλήφθηκε στην περιγραφή της μεθόδου DeepMoCap όπου και αξιοποιήθηκε, ενώ το δεύτερο παρουσιάστηκε πιο αναλυτικά στο Κεφ. 4. Από τη μία, αξιοποιώντας το σύνολο DMC, αναπτύχθηκε η μέθοδος DeepMoCap που παρουσιάστηκε στο Κεφ. 3, επιδεικνύοντας πολύ υποσχόμενα αποτελέσματα για τη χρήση αισθητήρων βάθους και βαθιάς μάθησης για καταγραφή ανθρώπινης κίνησης με δείκτες. Από την άλλη, μέσω των συμπερασμάτων που εξήχθησαν από την ανάπτυξη του DeepMoCap, δημιουργήθηκε και αξιοποιήθηκε το σύνολο δεδομένων HUMAN4D, που σταδιακά, μέσω διερε-

ύνησης σύγχρονων μεθόδων εκτίμησης πόζας/κίνησης (Κεφ. 5) και τεχνικών αναπαράστασης συντεταγμένων σε δομημένους N-διάστατους χώρους (Κεφ. 6), μας επέτρεψε την ανάπτυξη του DeMoCap, της μεθόδου που ολοκλήρωσε το στόχο αυτής της διατριβής (Κεφ. 7).

Στο πλαίσιο λοιπόν αυτό, πραγματοποιήθηκε μια σειρά από ερευνητικές συνεισφορές για την επίτευξη των παραπάνω:

- Εισάγαμε την πρώτη μέθοδο που χρησιμοποιεί πλήρως συνελικτικά νευρωνικά δίκτυα για τον αυτόματο εντοπισμό και την ταυτοποίηση 3Δ οπτικών δεδομένων πυκνά δεδομένα υπερύθρων και βάθους πολλαπλών προβολών (DeepMoCap).

- Επεκτείναμε την αρχιτεκτονική CPM [1] εισάγοντας την έννοια του χρόνου με την τροφοδότηση μιας δεύτερης εισόδου 3Δ οπτικής ροής και χρησιμοποιώντας 2Δ διανυσματικά πεδία [1] υπό το πρίσμα της χρονικής συνέχειας και συσχέτισης.

- Εισάγουμε το πρώτο μοντέλο βαθιάς μάθησης που χρησιμοποιεί πλήρως συνελικτικά νευρωνικά δίκτυα για την ταυτόχρονη παλινδρόμηση των 3Δ θέσεων οπτικών δεικτών και 3Δ πόζας από αραιά σύνολα 3Δ σημείων, τα οποία καταγράφονται με τη χρήση μιας χαμηλού κόστους διάταξης αισθητήρων βάθους πολλαπλών όψεων.

- Εισάγουμε μια νέα αναπαράσταση δεδομένων δεικτών ανεξάρτητη κλίμακας μεγέθους και μετατόπισης σε έναν κανονικοποιημένο 3Δ χώρο. Μέσω αυτής, το μοντέλο ξεπερνά την πόλωση στα σχετικά περιορισμένα σε αριθμό δεδομένα εκπαίδευσης.

- Εισάγουμε την έννοια της μεταβατικής εκπαίδευσης από την αναπαράσταση δεικτών σε 3Δ πόζα. Το μοντέλο μας οδηγείται στην εκμάθηση της υποκείμενης δομικής σχέσης μεταξύ του ανθρώπινου σώματος και των δεικτών, αποκωδικοποιώντας τις συσχετίσεις τους από τα αραιά και θορυβώδη νέφη 3Δ σημείων.

- Θέτουμε την εκτίμηση 3Δ σημείων ως κοινό στόχο 2Δ εντοπισμού και παλινδρόμησης εντός ενός κανονικοποιημένου 3Δ χώρου μας για την έμμεση κωδικοποίηση της 3ης διάστασης z με την εισαγωγή μιας νέας πλήρως διαφορίσιμης τεχνικής για 3Δ χωρική παλινδρόμηση.

Επιπρόσθετες σημαντικές συνεισφορές που πραγματοποιήθηκαν για την εξυπηρέτηση της διατριβής και την επίτευξη των παραπάνω είναι:

- Δημιουργήσαμε και δημοσιεύσαμε το σύνολο δεδομένων DMC αποτελούμενο από (i) χρωματισμένους χάρτες βάθους πολλαπλών προβολών και 3Δ οπτικής ροής με επισημάνσεις για τις 2Δ θέσεις των δεικτών και (ii) χωροχρονικά συσχετισμένους χάρτες βάθους πολλαπλών όψεων μαζί με πόζες από Kinect for Xbox One, αδρανειακά και ground truth δεδομένα κίνησης [84] (https://vcl.iti.gr/deepmocap/dataset).

- Δημιουργήσαμε και δημοσιεύσαμε το σύνολο 4Δ δεδομένων HUMAN4D που περιέχει ένα μεγάλο όγκο επισημασμένων χωροχρονικά συσχετισμένων δεδομένων χρώματος-υπερύθρων-βάθους πολλαπλών όψεων και δεδομένων κίνησης. Το HUMAN4D είναι το πρώτο σύνολο δεδομένων που παρέχει καρέ πολλαπλών προβολών με συγχρονισμού υλισμικού μαζί με δεδομένα καταγραφής κίνησης με δείκτες και ηχητικά δεδομένα.

- Διερευνούμε και αξιολογούμε σύγχρονες μεθόδους εκτίμησης 2Δ και 3Δ πόζας σε ακολουθίες δεδομένων μίας και πολλαπλών όψεων, όπως και αναπαράστασης συντεταγμένων με κύριο ενδιαφέρον στην κωδικοποίηση και αποκωδικοποίηση συντεταγμένων από θερμικούς χάρτες.

8.2 Μελλοντικές Ερευνητικές Κατευθύνσεις

Οι μέθοδοι και τα μοντέλα βαθιάς μάθησης που αναπτύχθηκαν έχουν περιθώρια βελτίωσης σε σύγκριση με τα επαγγελματικά συστήματα που βασίζονται σε δείκτες. Οι σημερινοί αισθητήρες καταναλωτικής ποιότητας που χρησιμοποιούμε για να μειώσουμε το υψηλό κόστος των εξειδικευμένων καμερών καταγραφής κίνησης είναι περιορισμένοι όσον αφορά τη συχνότητα λήψης ($30Hz$ έναντι $120/240Hz$ ή υψηλότερα) και το εύρος ανίχνευσης βάθους (έως 4 μέτρα διατηρώντας αποδεκτή ακρίβεια). Έτσι, σε αντίθεση με τις επαγγελματικές λύσεις που είναι εφαρμόσιμες και σε χώρους καταγραφής μεγάλης κλίμακας, π.χ. αρένες, γήπεδα ή στάδια, οι δυνατότητες των μοντέλων μας είναι περιορισμένες όσον αφορά τον όγκο του χώρου καταγραφής, τουλάχιστον με βάση τις υπάρχουσες τεχνολογίες ανίχνευσης βάθους και υπέρυθρης ακτινοβολίας. Επιπλέον, παρόμοια με όλα τα στατιστικά μοντέλα που βασίζονται σε δεδομένα, εκπαιδεύουμε σε ειδικά σύνολα δεδομένων με συγκεκριμένη προσάρτηση δεικτών στο σώμα για τη καταγραφή της ανθρώπινης κίνησης. Το γεγονός αυτό θα οδηγήσει σε λανθασμένες προβλέψεις κατά την εμφάνιση διαφορετικών προσαρτήσεων, απαιτώντας εκ νέου εκπαίδευση σε δεδομένα που έχουν ληφθεί με τις νέες ρυθμίσεις. Τα παραδοσιακά συστήματα, μολονότι επίσης απαιτούν νέες ρυθμίσεις για την ταυτοποίηση των δεικτών και την παρακολούθησή τους, μπορούν να εφαρμοστούν σε μια ποικιλία κινούμενων οντοτήτων, από ανθρώπους έως ζώα και αντικείμενα με μικρότερο 'κόστος' λόγω του πιο βραχύ χρόνου και της λιγότερης προσπάθειας που απαιτείται για την πραγματοποίηση της διαδικασίας, χωρίς την ανάγκη δημιουργίας νέων συνόλων δεδομένων. Με άλλα λόγια, για όλες τις λύσεις που βασίζονται σε δείκτες, η τοποθέτηση των δεικτών είναι ένα ισχυρό prior για τη λειτουργία τους, όμως, αυτό το prior είναι ακόμη ισχυρότερο για τις μεθόδους μας λόγω της μοντελοποίησης που βασίζεται στα δεδομένα. Τέλος, η προσέγγισή μας περιορίζεται στην παλινδρόμηση των δεικτών και της στάσης ενός και μόνο ατόμου στο χώρο καταγραφής. Αυτό οφείλεται στη χρήση της χωρικής παλινδρόμησης που μας περιορίζει στην παλινδρόμηση μιας μόνο συντεταγμένης ανά στρώμα λανθάνουσας θερμικής χαρτογράφησης.

Επικεντρώνουμε τη μελλοντική μας εργασία στο να ξεπεράσουμε τους προαναφερθέντες περιορισμούς της μεθόδου μας:

- Στόχος μας είναι να διεξάγουμε έρευνα σχετικά με αυτό το δύσκολο έργο, ώστε να καταστεί δυνατή η καταγραφή της κίνησης πολλών ατόμων. Ζωτικής σημασίας είναι αρχικά η εύρεση/δημιουργία κατάλληλων δεδομένων (που καλύπτεται μερικώς από το σύνολο HUMAN4D) και κατά δεύτερον η ανάπτυξη/χρήση κατάλληλων μεθόδων αποκωδικοποίησης συντεταγμένων από πολλαπλά σημεία στον ίδιο χάρτη θερμότητας, εξίσου αποτελεσματικών με τη χωρική παλινδρόμηση.

- Παρόλο που εφαρμόζουμε τις μεθόδους μας σε χρονικά συνεχή 3Δ δεδομένα, δεν έχουν σχεδιαστεί για να διατηρούν μακροχρόνια εσωτερική μνήμη για διαδοχική επεξεργασία

δεδομένων (μόνο βραχυχρόνια στο DeepMoCap). Από τη μία πλευρά, η διακριτή ανά καρέ εκτίμηση μας επιτρέπει να παραλείψουμε ζητήματα που μπορούν να προκαλέσουν οι τεχνικές παρακολούθησης, όπως η εναλλαγή δεικτών, ωστόσο, η συνεκτίμηση της χρονικής πληροφορίας μπορεί να οδηγήσει σε υψηλότερη ακρίβεια και ευρωστία στη σύλληψη κίνησης. Ως εκ τούτου, μια σημαντική ερευνητική κατεύθυνση θα ήταν η διερεύνηση τεχνικών που θα επιτρέψουν την εισαγωγή μακροχρόνιων και βραχυχρόνιων χρονικών χαρακτηριστικών για την ανάπτυξη πιο αποδοτικών και αποτελεσματικών μοντέλων βαθιάς μάθησης καταγραφής κίνησης.

- Τέλος, δεδομένου του εντοπισμού και της ταυτοποίησης των δεικτών, μπορούν να διερευνηθούν διαφορίσιμες τεχνικές αντίστροφης κινηματικής που θα οδηγήσουν σε επίλυση των θέσεων των αρθρώσεων οι οποίες θα επιτρέψουν την ενιαία εκτίμηση του προσανατολισμού των οστών.

Βιβλιογραφία

[1] Zhe Cao, Tomas Simon, Shih En Wei και Yaser Sheikh. *Realtime multi-person 2D pose estimation using Part Affinity Fields. CVPR*, τόμος 1, σελίδα 7, 2017.

[2] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara και Yaser Sheikh. *Panoptic studio: A massively multiview system for social motion capture. Proceedings of the IEEE International Conference on Computer Vision*, σελίδες 3334–3342, 2015.

[3] Google Inc. *TURBO Colormap.* https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html. Accessed: 2019-08-20.

[4] Zhe Cao, Tomas Simon, Shih En Wei και Yaser Sheikh. *Realtime multi-person 2d pose estimation using part affinity fields. Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 7291–7299, 2017.

[5] Hao Shu Fang, Shuqin Xie, Yu Wing Tai και Cewu Lu. *RMPE: Regional Multi-person Pose Estimation. ICCV*, 2017.

[6] Karim Iskakov, Egor Burkov, Victor Lempitsky και Yury Malkov. *Learnable triangulation of human pose. Proceedings of the IEEE International Conference on Computer Vision*, σελίδες 7718–7727, 2019.

[7] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz και Mubarak Shah. *Deep learning-based human pose estimation: A survey. arXiv preprint arXiv:2012.13392*, 2020.

[8] *VICON.* https://www.vicon.com/.

[9] Bardia Doosti, Shujon Naha, Majid Mirbagheri και David J Crandall. *Hope-net: A graph-based model for hand-object pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 6608–6617, 2020.

[10] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D Twigg και Kenrick Kin. *Online optical marker-based hand tracking with deep labels. ACM Transactions on Graphics (TOG)*, 37(4):166, 2018.

[11] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li και Yebin Liu. *4d association graph for realtime multi-person motion capture using multiple video cameras. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, σελίδες 1324–1333, 2020.

[12] KangKang Yin Goh Jing Ying. *SFU Motion Capture Database*. http://mocap.cs.sfu.ca/, 2011.

[13] Charlotte Dubosc, Geoffrey Gorisse, Olivier Christmann, Sylvain Fleury, Killian Poinsot και Simon Richir. *Impact of avatar facial anthropomorphism on body ownership, attractiveness and social presence in collaborative tasks in immersive virtual environments. Computers & Graphics*, 101:82–92, 2021.

[14] Lik Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo και Pan Hui. *All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. arXiv preprint arXiv:2110.05352*, 2021.

[15] Hongwei Yi, Chun Hao P Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies και Michael J Black. *Human-aware object placement for visual environment reconstruction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 3959–3970, 2022.

[16] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei και Michael J Black. *Putting people in their place: Monocular regression of 3d people in depth. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 13243–13252, 2022.

[17] Simon NB Gunkel, Hans M Stokking, Martin J Prins, Nandavan der Stap, Frank B ter Haar και Omar A Niamut. *Virtual Reality Conferencing: Multi-user immersive VR experiences on the web. Proceedings of the 9th ACM Multimedia Systems Conference*, σελίδες 498–501, 2018.

[18] Hyunae Lee, Timothy Hyungsoo Jung, M Claudiatom Dieck και Namho Chung. *Experiencing immersive virtual reality in museums. Information & Management*, 57(5):103229, 2020.

[19] Leonid Sigal, Michael Isard, Horst Haussecker και Michael J Black. *Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. International journal of computer vision*, 98(1):15–48, 2012.

[20] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans Peter Seidel, Weipeng Xu, Dan Casas και Christian Theobalt. *Vnect: Real-time 3D human pose estimation with a single RGB camera. ACM Transactions on Graphics (TOG)*, 36(4):44, 2017.

[21] Dario Pavllo, Thibault Porssut, Bruno Herbelin και Ronan Boulic. *Real-time finger tracking using active motion capture: a neural network approach robust to occlusions. Proceedings of the 11th Annual International Conference on Motion, Interaction, and Games*, σελίδες 1–10, 2018.

[22] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, WeiPeng Xu, Mohamed Elgharib, Pascal Fua, Hans Peter Seidel, Helge Rhodin, Gerard Pons-Moll και

Christian Theobalt. *Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. arXiv preprint arXiv:1907.00837*, 2019.

[23] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang και Lei Zhang. *Bottom-up Higher-Resolution Networks for Multi-Person Pose Estimation. arXiv preprint arXiv:1908.10357*, 2019.

[24] Riza Alp Guler και Iasonas Kokkinos. *Holopose: Holistic 3d human reconstruction in-the-wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, σελίδες 10884–10894, 2019.

[25] Nadine Rüegg, Christoph Lassner, Michael J Black και Konrad Schindler. *Chained Representation Cycling: Learning to Estimate 3D Human Pose and Shape by Cycling Between Representations. arXiv preprint arXiv:2001.01613*, 2020.

[26] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung και Li Fei-Fei. *Towards viewpoint invariant 3D human pose estimation. European Conference on Computer Vision*, σελίδες 160–177. Springer, 2016.

[27] Soonchan Park, Ju Yong Chang, Hyuk Jeong, Jae Ho Lee και Ji Young Park. *Accurate and efficient 3d human pose estimation algorithm using single depth images for pose analysis in golf. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, σελίδες 49–57, 2017.

[28] Angel Martínez-González, Michael Villamizar, Olivier Canévet και Jean Marc Odobez. *Real-time Convolutional Networks for Depth-based Human Pose Estimation. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, σελίδες 41–47, 2018.

[29] Angel Martínez-González, Michael Villamizar, Olivier Canévet και Jean Marc Odobez. *Investigating Depth Domain Adaptation for Efficient Human Pose Estimation. 2018 European Conference on Computer Vision - Workshops, ECCV 2018*, 2018.

[30] Magnus Burenius, Josephine Sullivan και Stefan Carlsson. *3D pictorial structures for multiple view articulated pose estimation. Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 3618–3625, 2013.

[31] Ahmed Elhayek, Edilsonde Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele και Christian Theobalt. *Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, σελίδες 3810–3818, 2015.

[32] Helge Rhodin, Mathieu Salzmann και Pascal Fua. *Unsupervised geometry-aware representation for 3d human pose estimation. Proceedings of the European Conference on Computer Vision (ECCV)*, σελίδες 750–767, 2018.

[33] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang και Wenjun Zeng. *Cross view fusion for 3d human pose estimation. Proceedings of the IEEE International Conference on Computer Vision*, σελίδες 4342–4351, 2019.

[34] Hanyue Tu, Chunyu Wang και Wenjun Zeng. *Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, σελίδες 197–212. Springer, 2020.

[35] Walid Bekhtaoui, Ruhan Sa, Brian Teixeira, Vivek Singh, Klaus Kirchberg, Yao jen Chang και Ankur Kapoor. *View Invariant Human Body Detection and Pose Estimation from Multiple Depth Sensors. arXiv preprint arXiv:2005.04258*, 2020.

[36] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman και Andrew Blake. *Real-time human pose recognition in parts from single depth images. Computer Vision and Pattern Recognition (CVPR)*, σελίδες 1297–1304. Ieee, 2011.

[37] Kourosh Khoshelham και Sander Oude Elberink. *Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors*, 12(2):1437–1454, 2012.

[38] Pierre Plantard, Edouard Auvinet, Anne Sophie Le Pierres και Franck Multon. *Pose estimation with a kinect for ergonomic studies: Evaluation of the accuracy using a virtual mannequin. Sensors*, 15(1):1785–1803, 2015.

[39] Stylianos Asteriadis, Anargyros Chatzitofis, Dimitrios Zarpalas, Dimitrios S Alexiadis και Petros Daras. *Estimating human motion from multiple kinect sensors. Proc. of the 6th international conference on computer vision/computer graphics collaboration techniques and applications*, σελίδα 3. ACM, 2013.

[40] Iason Oikonomidis, Nikolaos Kyriazis και Antonis A Argyros. *Efficient model-based 3D tracking of hand articulations using Kinect. BmVC*, τόμος 1, σελίδα 3, 2011.

[41] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard και Thomas Brox. *3D Human Pose Estimation in RGBD Images for Robotic Task Learning. arXiv preprint arXiv:1803.02622*, 2018.

[42] Umer Rafi, Juergen Gall και Bastian Leibe. *A semantic occlusion model for human pose estimation from a single depth image. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[43] Hanbyul Joo, Tomas Simon και Yaser Sheikh. *Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, σελίδες 8320–8329, 2018.

[44] Zhenbao Liu, Jinxin Huang, Junwei Han, Shuhui Bu και Jianfeng Lv. *Human motion tracking by multiple RGBD cameras. IEEE Transactions on Circuits and Systems for Video Technology*, 27(9):2014–2027, 2017.

[45] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell και Yaser Sheikh. *Pose machines: Articulated pose estimation via inference machines. European Conference on Computer Vision*, σελίδες 33–47. Springer, 2014.

[46] Shih En Wei, Varun Ramakrishna, Takeo Kanade και Yaser Sheikh. *Convolutional Pose Machines. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, σελίδες 4724–4732, 2016.

[47] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka και Bernt Schiele. *Deepercut: A deeper, stronger, and faster multi-person pose estimation model. European Conference on Computer Vision*. Springer, 2016.

[48] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu και Pichao Wang. *Lifting transformer for 3d human pose estimation in video. arXiv preprint arXiv:2103.14304*, 2, 2021.

[49] Yizhuo Li, Miao Hao, Zonglin Di, Nitesh Bharadwaj Gundavarapu και Xiaolong Wang. *Test-time personalization with a transformer for human pose estimation. Advances in Neural Information Processing Systems*, 34:2583–2597, 2021.

[50] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang και Luc Van Gool. *Mhformer: Multi-hypothesis transformer for 3d human pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 13147–13156, 2022.

[51] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang και Wenming Yang. *Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Transactions on Multimedia*, 2022.

[52] Shih Yang Su, Frank Yu, Michael Zollhoefer και Helge Rhodin. *A-nerf: Surface-free human 3d pose refinement via neural rendering. arXiv preprint arXiv:2102.06199*, 2021.

[53] Hongyi Xu, Thiemo Alldieck και Cristian Sminchisescu. *H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. Advances in Neural Information Processing Systems*, 34:14955–14966, 2021.

[54] Guillaume Rochette, Chris Russell και Richard Bowden. *Human Pose Manipulation and Novel View Synthesis using Differentiable Rendering. 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, σελίδες 1–8. IEEE, 2021.

[55] Marco Carraro, Matteo Munaro, Jeff Burke και Emanuele Menegatti. *Real-time marker-less multi-person 3D pose estimation in RGB-Depth camera networks. arXiv preprint arXiv:1710.06235*, 2017.

[56] Alireza Shafaei και James J. Little. *Real-Time Human Motion Capture with Multiple Depth Cameras. Proc. of the 13th Conference on Computer and Robot Vision*. CIPPRS, 2016.

[57] Liang Shuai, Chao Li, Xiaohu Guo, Balakrishnan Prabhakaran και Jinxiang Chai. *Motion capture with ellipsoidal skeleton using multiple depth cameras. IEEE transactions on visualization and computer graphics*, 23(2):1085–1098, 2017.

[58] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll και Michael J Black. *SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG)*, 34(6):248, 2015.

[59] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero και Michael J Black. *Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. European conference on computer vision*, σελίδες 561–578. Springer, 2016.

[60] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas και Michael J Black. *Expressive body capture: 3d hands, face, and body from a single image. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, σελίδες 10975–10985, 2019.

[61] Ahmed AA Osman, Timo Bolkart και Michael J Black. *Star: Sparse trained articulated human body regressor. European Conference on Computer Vision*, σελίδες 598–613. Springer, 2020.

[62] Jie Song, Limin Wang, Luc Van Gool και Otmar Hilliges. *Thin-slicing network: A deep structured model for pose estimation in videos. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, τόμος 2, 2017.

[63] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang και Liang Lin. *LSTM Pose Machines. arXiv preprint arXiv:1712.06316*, 2017.

[64] Juan Luis Jimenez Bascones. *Cloud point labelling in optical motion capture systems.* Διδακτορική Διατριβή, Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2019.

[65] Simon Alexanderson, Carol O'Sullivan και Jonas Beskow. *Real-time labeling of non-rigid motion capture marker sets. Computers & graphics*, 69:59–67, 2017.

[66] Daniel Holden. *Robust solving of optical motion capture data by denoising. ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.

[67] Maksym Perepichka, Daniel Holden, Sudhir P Mudur και Tiberiu Popa. *Robust Marker Trajectory Repair for MOCAP using Kinematic Reference. Motion, Interaction and Games*, σελίδες 1–10. Ernst & Sohn, 2019.

[68] Alexander Toshev και Christian Szegedy. *Deeppose: Human pose estimation via deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 1653–1660, 2014.

[69] Jonathan J Tompson, Arjun Jain, Yann LeCun και Christoph Bregler. *Joint training of a convolutional network and a graphical model for human pose estimation. Advances in neural information processing systems*, σελίδες 1799–1807, 2014.

[70] Christopher Tensmeyer και Tony Martinez. *Robust Keypoint Detection. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, τόμος 5, σελίδες 1–7. IEEE, 2019.

[71] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang και Yichen Wei. *Integral human pose regression. Proceedings of the European Conference on Computer Vision (ECCV)*, σελίδες 529–545, 2018.

[72] Aiden Nibali, Zhen He, Stuart Morgan και Luke Prendergast. *Numerical coordinate regression with convolutional neural networks. arXiv preprint arXiv:1801.07372*, 2018.

[73] Diogo C Luvizon, Hedi Tabia και David Picard. *Human pose regression by combining indirect part detection and contextual information. Computers & Graphics*, 85:15–22, 2019.

[74] Catalin Ionescu, Dragos Papava, Vlad Olaru και Cristian Sminchisescu. *Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[75] Leonid Sigal, Alexandru O Balan και Michael J Black. *Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International journal of computer vision*, 87(1-2):4, 2010.

[76] Zhixuan Yu, Jae Shin Yoon, Prashanth Venkatesh, Jaesik Park, Jihun Yu και Hyun Soo Park. *Humbi 1.0: Human multiview behavioral imaging dataset. arXiv preprint arXiv:1812.00281*, 2018.

[77] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal και Ruzena Bajcsy. *Berkeley mhad: A comprehensive multimodal human action database. 2013 IEEE Workshop on Applications of Computer Vision (WACV)*, σελίδες 53–60. IEEE, 2013.

[78] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler και Bernt Schiele. *2D Human Pose Estimation: New Benchmark and State of the Art Analysis. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[79] PhaseSpace. http://www.phasespace.com.

[80] Dimitrios S Alexiadis, Anargyros Chatzitofis, Nikolaos Zioulis, Olga Zoidi, Georgios Louizis, Dimitrios Zarpalas και Petros Daras. *An integrated platform for live 3D human reconstruction and motion capturing. IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):798–813, 2017.

[81] Mariano Jaimez, Mohamed Souiai, Javier Gonzalez-Jimenez και Daniel Cremers. *A primal-dual framework for real-time dense RGB-D scene flow. International Conference on Robotics and Automation (ICRA)*, σελίδες 98–104. IEEE, 2015.

[82] Karen Simonyan και Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556*, 2014.

[83] Monique Paulich, Martin Schepers, Nina Rudigkeit και Giovanni Bellusci. *Xsens MTw Awinda: Miniature Wireless Inertial-Magnetic Motion Tracker for Highly Accurate 3D Kinematic Applications.*

[84] *PhaseSpace*. http://www.phasespace.com/.

[85] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár και C Lawrence Zitnick. *Microsoft coco: Common objects in context. European Conference on Computer Vision*, σελίδες 740–755. Springer, 2014.

[86] Charles K Chui, Guanrong Chen και others. *Kalman filtering*. Springer, 2017.

[87] Tobias Schubert, Alexis Gkogkidis, Tonio Ball και Wolfram Burgard. *Automatic initialization for skeleton tracking in optical motion capture. International Conference on Robotics and Automation (ICRA)*, σελίδες 734–739. IEEE, 2015.

[88] Victor Brian Zordan και Nicholas C Van Der Horst. *Mapping optical motion capture data to skeletal motion using a physical model. Proc. of the ACM SIGGRAPH-/Eurographics symposium on Computer animation*, σελίδες 245–250. Eurographics Association, 2003.

[89] Yi Yang και Deva Ramanan. *Articulated pose estimation with flexible mixtures-of-parts. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 1385–1392. IEEE, 2011.

[90] François Destelle, Amin Ahmadi, Noel E O'Connor, Kieran Moran, Anargyros Chatzitofis, Dimitrios Zarpalas και Petros Daras. *Low-cost accurate skeleton tracking based on fusion of kinect and wearable inertial sensors. European Signal Processing Conference (EUSIPCO)*, σελίδες 371–375. IEEE, 2014.

[91] Anders Grunnet-Jepsen, Paul Winer, Aki Takagi, John Sweetser, Kevin Zhao, Tri Khuong, Dan Nie και John Woodfill. *Using the RealSense D4xx depth sensors in multi-camera configurations. Santa Monica, CA, USA*, 2018.

[92] Vladimiros Sterzentsenko, Antonis Karakottas, Alexandros Papachristou, Nikolaos Zioulis, Alexandros Doumanoglou, Dimitrios Zarpalas και Petros Daras. *A low-cost, flexible and portable volumetric capturing system. 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, σελίδες 200–207. IEEE, 2018.

[93] Alexandros Papachristou, Nikolaos Zioulis, Dimitrios Zarpalas και Petros Daras. *Markerless structure-based multi-sensor calibration for free viewpoint video capture.* 2018.

[94] M Shamim Hossain και Ghulam Muhammad. *Emotion recognition using deep learning approach from audio–visual emotional big data. Information Fusion*, 49:69–78, 2019.

[95] Andrew Owens και Alexei A Efros. *Audio-visual scene analysis with self-supervised multisensory features. Proceedings of the European Conference on Computer Vision (ECCV)*, σελίδες 631–648, 2018.

[96] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo και Jonathan Huang. *Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. Proceedings of the IEEE International Conference on Computer Vision*, σελίδες 6301–6310, 2019.

[97] Alejandro Newell, Kaiyu Yang και Jia Deng. *Stacked hourglass networks for human pose estimation. European conference on computer vision*, σελίδες 483–499. Springer, 2016.

[98] Vasileios Belagiannis και Andrew Zisserman. *Recurrent human pose estimation. 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, σελίδες 468–475. IEEE, 2017.

[99] Miaopeng Li, Zimeng Zhou, Jie Li και Xinguo Liu. *Bottom-up pose estimation of multiple person with bounding box constraint. 2018 24th International Conference on Pattern Recognition (ICPR)*, σελίδες 115–120. IEEE, 2018.

[100] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou και Junsong Yuan. *A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. Proceedings of the IEEE International Conference on Computer Vision*, σελίδες 793–802, 2019.

[101] Piotr Szczuko. *Deep neural networks for human pose estimation from a very low resolution depth image. Multimedia Tools and Applications*, 78(20):29357–29377, 2019.

[102] Muhammed Kocabas, Salih Karagoz και Emre Akbas. *Self-supervised learning of 3d human pose using multi-view geometry. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, σελίδες 1077–1086, 2019.

[103] Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michelde Mathelin και Nicolas Padoy. *A multi-view RGB-D approach for human pose estimation in operating rooms. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, σελίδες 363–372. IEEE, 2017.

[104] Marco Carraro, Matteo Munaro, Jeff Burke και Emanuele Menegatti. *Real-time marker-less multi-person 3d pose estimation in rgb-depth camera networks. International Conference on Intelligent Autonomous Systems*, σελίδες 534–545. Springer, 2018.

[105] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang και Jingdong Wang. *Bottom-up human pose estimation via disentangled keypoint regression. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 14676–14686, 2021.

[106] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler και Bernt Schiele. *2d human pose estimation: New benchmark and state of the art analysis. Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, σελίδες 3686–3693, 2014.

[107] Yi Yang και Deva Ramanan. *Articulated human detection with flexible mixtures of parts. IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012.

[108] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu και Christian Theobalt. *Monocular 3D human pose estimation in the wild using improved CNN supervision. International Conference on 3D Vision (3DV)*, σελίδες 506–516. IEEE, 2017.

[109] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye και Ce Zhu. *Distribution-aware coordinate representation for human pose estimation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, σελίδες 7093–7102, 2020.

[110] Ke Sun, Bin Xiao, Dong Liu και Jingdong Wang. *Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 5693–5703, 2019.

[111] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen και Achintya Bhowmik. *Intel realsense stereoscopic depth cameras. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, σελίδες 1–10, 2017.

[112] Zhengyou Zhang. *Microsoft kinect sensor and its effect. IEEE multimedia*, 19(2):4–10, 2012.

[113] Andre Gaschler. *Real-time marker-based motion tracking: Application to kinematic model estimation of a humanoid robot. Thesis*, 2011.

[114] Richard Hartley και Andrew Zisserman. *Multiple view geometry in computer vision.* Cambridge university press, 2003.

[115] Gernot Riegler, Ali Osman Ulusoy και Andreas Geiger. *Octnet: Learning deep 3d representations at high resolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, σελίδες 3577–3586, 2017.

[116] Charles Ruizhongtai Qi, Li Yi, Hao Su και Leonidas J Guibas. *Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems*, σελίδες 5099–5108, 2017.

[117] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang και others. *Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence*, 2020.

[118] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin Ionut Popa και Cristian Sminchisescu. *Deep network for the integrated 3d sensing of multiple people in natural images. Advances in Neural Information Processing Systems*, 31:8410–8419, 2018.

[119] Zhen Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber και Xiao Jun Wu. *Wing loss for robust facial landmark localisation with convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, σελίδες 2235–2245, 2018.

[120] Bent Fuglede και Flemming Topsoe. *Jensen-Shannon divergence and Hilbert space embedding. International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, σελίδα 31. IEEE, 2004.

[121] Sijin Li, Weichen Zhang και Antoni B. Chan. *Maximum-Margin Structured Learning With Deep Networks for 3D Human Pose Estimation. Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[122] Hongyang Gao και Shuiwang Ji. *Graph u-nets. international conference on machine learning*, σελίδες 2083–2092. PMLR, 2019.

[123] Xavier Glorot και Yoshua Bengio. *Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics*, σελίδες 249–256, 2010.

[124] Kaiming He, Georgia Gkioxari, Piotr Dollár και Ross Girshick. *Mask r-cnn. Proceedings of the IEEE international conference on computer vision*, σελίδες 2961–2969, 2017.

[125] Kaiming He, Xiangyu Zhang, Shaoqing Ren και Jian Sun. *Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 770–778, 2016.

[126] Catalin Ionescu, Dragos Papava, Vlad Olaru και Cristian Sminchisescu. *Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[127] Diederik P Kingma και Jimmy Ba. *Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980*, 2014.

[128] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai και Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems 32*H. Wallach, H. Larochelle, A. Beygelzimer, F.d' Alché-Buc, E. Fox και R. Garnett, επιμελητές, σελίδες 8024–8035. Curran Associates, Inc., 2019.

[129] moai. *moai: Accelerating modern data-driven workflows*. https://github.com/ai-in-motion/moai, 2021.

Παραρτήματα

Παράρτημα Α΄

Δημοσιεύσεις

Α΄.1 Δημοσιεύσεις σε Διεθνή Περιοδικά

- **Chatzitofis, A.**, Zarpalas, D., Daras, P., Kollias, S. (2021). *DeMoCap: Low-Cost Marker-Based Motion Capture.* International Journal of Computer Vision, 129(12), 3338-3366. https://doi.org/10.1007/s11263-021-01526-z

- **Chatzitofis**, A., Saroglou, L., Boutis, P., Drakoulis, P., Zioulis, N., Subramanyam, S., ... , Kollias, S. , Daras, P. (2020). *HUMAN4D: A human-centric multimodal dataset for motions and immersive media.* IEEE Access, 8, 176241-176262. https://doi.org/10.1109/ACCESS.2020.3026276

- **Chatzitofis, A.**, Zarpalas, D., Kollias, S., Daras, P. (2019). *DeepMoCap: Deep optical motion capture using multiple depth sensors and retro-reflectors.* Sensors, 19(2), 282. https://doi.org/10.3390/s19020282

- Patrona, F., **Chatzitofis, A.**, Zarpalas, D., Daras, P. (2018). *Motion analysis: Action detection, recognition and evaluation based on motion capture data.* Pattern Recognition, 76, 612-622. https://doi.org/10.1016/j.patcog.2017.12.007

- Alexiadis, D. S., **Chatzitofis, A.**, Zioulis, N., Zoidi, O., Louizis, G., Zarpalas, D., Daras, P. (2016). *An integrated platform for live 3D human reconstruction and motion capturing.* IEEE Transactions on Circuits and Systems for Video Technology, 27(4), 798-813. https://doi.org/10.1109/TCSVT.2016.2576922

Α΄.2 Δημοσιεύσεις σε Διεθνή Συνέδρια

- Albanis, G. N., Zioulis, N., **Chatzitofis, A.**, Dimou, A., Zarpalas, D., Daras, P. (2021, January). *On end-to-end 6DOF object pose estimation and robustness to object scale.* In ML Reproducibility Challenge 2020.

- Sterzentsenko, V., Saroglou, L., **Chatzitofis, A.**, Thermos, S., Zioulis, N., Doumanoglou, A., ... , Daras, P. (2019). *Self-supervised deep depth denoising.* In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1242-1251). https://doi.org/10.1109/ICCV.2019.00133

- **Chatzitofis, A.**, Zarpalas, D., Daras, P. (2017, November). *A computerized system for real-time exercise performance monitoring and e-coaching using motion capture data.* In International Conference on Biomedical and Health Informatics (pp. 243-247). Springer, Singapore.

- **Chatzitofis, A.**, Zarpalas, D., Filos, D., Triantafyllidis, A., Chouvarda, I., Maglaveras, N., Daras, P. (2017, July). *Technological module for unsupervised, personalized cardiac rehabilitation exercising.* In 2017 IEEE 41st annual computer software and applications conference (COMPSAC) (Vol. 2, pp. 125-130). IEEE.