ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Εκτίμηση στάσης του σώματος με χρήση τεχνολογιών υπολογιστικής όρασης και βαθιάς μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χατζηστάμου Δημήτριος

Επιβλέπων: Κουτσούρης Δημήτριος-Διονύσιος
Καθηγητής Ε.Μ.Π

Συνεπιβλέπουσα: Δρ. Πετροπούλου Ουρανία
ΕΔΙΠ Α΄ Ε.Μ.Π

Αθήνα, Οκτώβριος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Εκτίμηση στάσης σώματος με χρήση τεχνολογιών υπολογιστικής όρασης και βαθιάς μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χατζηστάμου Χ. Δημήτριος

Επιβλέπων: Κουτσούρης Δημήτριος-Διονύσιος
Καθηγητής Ε.Μ.Π

Συνεπιβλέπουσα: Δρ. Πετροπούλου Ουρανία
ΕΔΙΠ Α΄ Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την ..... Οκτώβριου 2020.

.....	.....	.....
Δημήτριος-Διονύσιος Κουτσούρης Καθηγητής Ε.Μ.Π	Παναγιώτης Τσανάκας Καθηγητής Ε.Μ.Π	Γιώργος Ματσόπουλος Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2020

(Υπογραφή)

.....
Χατζηστάμου Χ. Δημήτριος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

ΠΕΡΙΛΗΨΗ

Τις τελευταίες δεκαετίες έχει παρατηρηθεί ιδιαίτερη ανάπτυξη στην έρευνα των τεχνικών της υπολογιστικής όρασης (computer vision) ενώ και η ραγδαία εξέλιξη στον τομέα της υπολογιστικής ισχύος έχει δώσει μεγάλη ώθηση στον τομέα της βαθιάς μάθησης (deep learning). Ένα αντικείμενο της όρασης υπολογιστών που χρησιμοποιεί και τεχνικές βαθιάς μάθησης είναι και εκτίμηση της στάσης του σώματος (body detection), που είναι ένα υποσύνολο της ανάλυσης της εικόνας.

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η διερεύνηση και η αξιολόγηση των μεθόδων εκτίμησης της στάσης του ανθρώπινου σώματος με χρήση της τεχνολογίας υπολογιστικής όρασης και βαθιάς όρασης.

Αρχικά γίνεται μια εισαγωγή στις βασικές έννοιες με τις οποίες ασχολείται η παρούσα διπλωματική εργασία. Στην συνέχεια μέσα από την μελέτη πολλών εργασιών για την εκτίμηση της ανθρώπινης στάσης με χρήση τεχνολογίας υπολογιστικής όρασης γίνεται μια ιστορική αναδρομή στην εξέλιξη των μεθόδων της εκτίμησης της στάσης μέσω τεχνολογίας οράσεως από τα πρώιμα στάδια έως σήμερα. Στην συνέχεια ερχόμαστε σε πιο σύγχρονες μεθόδους που χρησιμοποιούν την τεχνολογία της βαθιάς μάθησης. Πρώτα γίνεται μια παρουσίαση και ανάλυση των μεθόδων βαθιάς μάθησης για την εκτίμηση της ανθρώπινης στάσης ενός ατόμου και στην συνέχεια για την εκτίμηση της ανθρώπινης στάσης πολλαπλών ατόμων.

Τέλος γίνεται μια παρουσίαση της απόδοσης των μεθόδων βαθιάς μάθησης που παρουσιάσαμε στην παρούσα διπλωματική εργασία, μια σύγκριση των αποδόσεων τους και αναφέρονται προτάσεις για μελλοντική έρευνα σε αυτόν τον τομέα.

Λέξεις κλειδιά

Τεχνητή Νοημοσύνη, Υπολογιστική Όραση, Βαθιά Μάθηση , Εκτίμηση Στάσης Ανθρώπινου Σώματος, Εκτίμηση Στάσης Ενός Ατόμου, Εκτίμηση Στάσης Πολλών Ατόμων

ABSTRACT

Over the last few decades, research into computer vision techniques has grown considerably, and rapid development in computing power have boosted the field of deep learning. An individual field of computer vision that utilizes deep learning techniques is body detection, a sector of image analysis.

The aim of this diploma thesis is to investigate and evaluate methods of detection the posture of human body using computer vision and deep learning technology.

First, an introduction was made to the basic meanings of this dissertation. Then, through the study of many works on the detection of human posture using computer vision technology, a historical flashback was made to the evolution of the methods for human body detection from the early stages through the present day. Then we came to more modern methods that use the technology of deep learning. First there was a presentation and analysis of deep learning methods for a single-person human posture detection and then for multi-person human posture detection.

Finally, there was a presentation of the performance of the deep learning methods which we had presented in this diploma thesis, a comparison of their performance and proposals for future research in this area was mentioned.

Key words

Artificial Intelligence, Computer Vision, Deep Learning, Body Posture Detection, Single-Person Body Detection, Multi-Person Body Detection

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα προπτυχιακή διπλωματική εργασία εκπονήθηκε κατά το ακαδημαϊκό έτος 2019-2020 στα πλαίσια του τομέα Ηλεκτρικής Ισχύος της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα Καθηγητή μου κ. Δημήτριο-Διονύσιο Κουτσούρη που μου έδωσε την ευκαιρία να διεκπεραιώσω αυτό το τόσο ενδιαφέρον θέμα. Ιδιαίτερα, θα ήθελα να ευχαριστήσω την Διδάκτορα στον Τομέα Συστημάτων Μετάδοσης Πληροφορίας και Τεχνολογίας Υλικών κ. Πετροπούλου Ουρανία για την πολύτιμη βοήθεια που μου παρείχε και για την άψογη συνεργασία μας κατά την διάρκεια της εκπόνησης της παρούσας εργασίας. Ήταν πάντα άμεσα διαθέσιμη να απαντήσει στις απορίες μου και να με καθοδηγήσει. Χάρη στην πολύτιμη βοήθεια της κατάφερα να ολοκληρώσω αυτήν την εργασία.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ στην οικογένεια μου για την στήριξη που μου προσέφερε ιδιαίτερα τις δύσκολες περιόδους και στους φίλους μου που με συνόδευσαν σε αυτό το όμορφο ταξίδι.

Χατζηστάμου Δημήτριος

Αθήνα, Οκτώβριος 2020

Στους γονείς μου και στον Βαγγέλη,

που δεν βρίσκεται πια κοντά μας

Πίνακας περιεχομένων

ΠΕΡΙΕΧΟΜΕΝΑ ΕΙΚΟΝΩΝ

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Η διείσδυση της τεχνητής νοημοσύνης στα συστήματα υγείας

Μια τεράστια ποικιλία τεχνικών τεχνητής νοημοσύνης έχουν αναπτυχθεί σε συγκεκριμένα προβλήματα υγειονομικής περίθαλψης κατά την διάρκεια των τελευταίων τριάντα χρόνων με ποικίλα επίπεδα επιτυχίας, ενώ υπάρχει μια έλλειψη συστηματικής αντιστοίχισης των δυνατοτήτων της τεχνητής νοημοσύνης, με το εύρος ευκαιριών εφαρμογής της.

Ο ψηφιακός μετασχηματισμός στην υγειονομική περίθαλψη επιδιώκει λύσεις ώστε να κάνει την υγειονομική περίθαλψη ασφαλέστερη, πιο προσιτή και με μεγαλύτερη πρόσβαση, γεγονός που οδήγησε ώστε να είναι ένας ταχέως αναπτυσσόμενος τομέας έρευνας. Οι ειδικοί των επαγγελματιών της υγείας έχουν καταστήσει την υγειονομική περίθαλψη μια από τις πιο κύριες περιοχές στις οποίες οι τεχνολογίες τεχνητής νοημοσύνης έχουν εισχωρήσει ,με διαφορετικά επίπεδα επιτυχίας. Οι πρώτες προσπάθειες να μιμηθούν την ανθρώπινη διαγνωστική συμπεριφορά μέσω ειδικών συστημάτων που χρησιμοποιούν τους κανόνες και την ασαφή λογική έγιναν με βάση την γνώση συστημάτων με πλαίσια, σενάρια για διάφορα συστήματα συλλογισμού καθώς και επεξεργασία φυσικής γλώσσας με βάση τεχνικές ερωτήσεων-απαντήσεων. Η αυξανόμενη ψηφιοποίηση των ιατρικών δεδομένων των ασθενών και η βελτιωμένη υιοθέτηση της ψηφιακής απεικόνισης και των ιατρικών συσκευών έχουν οδηγήσει στην υιοθέτηση τεχνολογιών που χρησιμοποιούν μεγάλους όγκους δεδομένων για την μηχανική μάθηση και τα βαθιά νευρωνικά δίκτυα. Συνολικά η πλειοψηφία των προσπαθειών έχει είτε καθοδηγηθεί από την τεχνολογία στην ανάπτυξη συγκεκριμένων τεχνολογιών τεχνητής νοημοσύνης, είτε οδηγείται από τα δεδομένα με στόχο ένα μεγάλο ιατρικό ζήτημα.

Το μόνο σίγουρο είναι πως η εφαρμογή της τεχνητής νοημοσύνης προσφέρει πολλές προκλήσεις και ευκαιρίες για την δημιουργία τεχνολογιών που θα βελτιώσουν άρδην την υγειονομική περίθαλψη. Στην παρούσα διπλωματική εργασία εξετάζεται η πρόσφατη βιβλιογραφία σχετικά με την εφαρμογή τεχνικών υπολογιστικής οράσεως και βαθιάς μάθησης στον τομέα της εκτίμησης της στάσης του ανθρώπινου σώματος, ένα πεδίο το οποίο χρίζει ποικίλων εφαρμογών στον τομέα της ιατρικής. Με βάση την ανάλυση των εργασιών, προτείνονται οι μέθοδοι βαθιάς μάθησης που θα μπορούσαν να είναι το όχημα που μέσω των πολλών βιοιατρικών δεδομένων θα οδηγήσει σε μηχανήματα και εφαρμογές ακριβούς προσέγγισης της στάσης του ανθρώπινου

σώματος. Επίσης στην παρουσίαση των μεθόδων σημειώνονται και ορισμένοι περιορισμοί και ανάγκες για μια βελτιωμένη ανάπτυξη των μεθόδων και των εφαρμογών τους, ώστε να οδηγηθούμε σε μεγαλύτερα ποσοστά απόδοσης τους μέχρι την καθολική εφαρμογής τους σε διάφορα ιατρικά πεδία [1].

1.2 Η μέθοδος της εκτίμηση της στάσης του σώματος

Οι παραγωγικές προσεγγίσεις (γνωστές και ως προσεγγίσεις που βασίζονται σε μοντέλα ή από πάνω προς τα κάτω) προσεγγίζουν ένα ανθρώπινο μοντέλο με την παρατήρηση μιας εικόνας. Ειδικότερα, οι παραγωγικές προσεγγίσεις δημιουργούν ένα μοντέλο του ανθρώπινου σώματος μέσα από ένα σύνολο σχολιασμένων εικόνων. Αυτό το μοντέλο αποτελείται από ένα ΄΄μαντείο΄΄ το οποίο μπορεί να δημιουργηθεί έτσι ώστε να παράγει πιθανές διαμορφώσεις του ανθρώπινου σώματος μέσω ενός συνόλου παραμέτρων, το οποίο υποδείκνυε τους βαθμούς ελευθερίας. Μόλις δημιουργηθεί ένα μοντέλο, τότε ΄΄ταιριάζεται΄΄ σε μια δοσμένη εικόνα, έτσι ώστε το αποτέλεσμα του μοντέλου να είναι πιο κοντά στο αντικείμενο της εικόνας. Έτσι, το ταίριασμα αποτελείται από την εύρεση των παραμέτρων που δίνουν ή παράγουν, το μοντέλο που περιγράφει με καλύτερο τρόπο το ανθρώπινο σώμα στην εικόνα εισόδου. Το ανθρώπινο μοντέλο συνεχώς αναβαθμίζεται σύμφωνα με την παρακολούθηση της εικόνας χρησιμοποιώντας τοπική βελτιστοποίηση ή στοχαστική έρευνα. Γενικά, οι μεθόδοι σύνθεσης μπορούν να υποδιαιρεθούν σε ολιστικές ή με βάση τα τμήματα, ανάλογα με το εάν το μοντέλο διαμορφώνει το ανθρώπινο σώμα ως σύνολο ή με την βοήθεια των μερών από τα οποία αποτελείται. Παρά τα εντυπωσιακά αποτελέσματα της, η παραγωγικές μεθόδοι είναι συνήθως επιρρεπείς να παγιδευτούν σε τοπικά ελάχιστα, καθιστώντας αυτές εξαρτώμενες μιας καλής αρχικοποίησης .

Μια κατηγορία παραγωγικών μεθόδων που χρησιμοποιείται ευρέως είναι οι τελευταίες, βασισμένες σε μέρη εφαρμογές, στις οποίες η εμφάνιση ενός αντικειμένου αποσυντίθεται σε περιγράμματα τοπικών μερών, τα οποία ακολούθως είναι αναγκασμένα να ανήκουν σε ένα εύλογο παράδειγμα στην τάξη του αντικειμένου. Αυτό είναι, η μοντελοποίηση του αντικειμένου τοπικά, και τα μέρη είναι τότε από κοινού αναγκασμένα να μοιάζουν σαν ένα εύλογο ανθρώπινο σώμα. Σε μερικές εργασίες, αυτοί οι περιορισμοί μπορούν να θεωρηθούν ως ελατήρια που συνδέουν ζεύγη περιγραμμάτων. Στο πρόβλημα της εκτίμησης της στάσης του σώματος τα ελατήρια μπορούν να θεωρηθούν ως κοινές αρθρώσεις του μοντέλου. Με αυτή την έννοια, οι αρθρώσεις επιβάλλουν κινηματικούς περιορισμούς στο μοντέλο καθώς κατασκευάζονται σύμφωνα με την σκελετική δομή του ανθρώπου. Αυτό ωφελεί τη μέθοδο με την έννοια ότι μπορεί να διαχειριστεί σοβαρές αποκλείσεις με περιορισμένο τρόπο.

Ένα από τα έργα, που αποτέλεσαν ακρογωνιαίο λίθο των βασισμένων σε μέρη μοντέλων, βασίζεται σε εικονογραφικές δομές. Οι εικονογραφικές δομές εισήχθησαν από τους Fischler και Elschlager, 1973 ως τρόπος να αναπαρασταθούν τα αντικείμενα στις εικόνες, μέσω της τοπικής εμφάνισης των τμημάτων του και της σύνδεσης κάθε τμήματος με τα γειτονικά του, προκειμένου να επιτευχθεί μια αποτελεσματική προσαρμογή του προτύπου. Η κύρια ιδέα των εικονογραφημένων δομών είναι να περιγράψουν το ανθρώπινο σώμα ως μια συλλογή άκαμπτων προτύπων, κάθε ένα από τα οποία περιγράφει ένα συγκεκριμένο τμήμα του αντικειμένου, και έχει δυνητικά τη

μορφή ενός δομημένου μη κατευθυνόμενου γραφικού μοντέλου, όπως απεικονίζεται στην εικόνα 1.2. Η δομή του γραφήματος επιτρέπει τη μοντελοποίηση των μερών ανεξάρτητα με σεβασμό στους γείτονες του. Τα τοπικά περιγράμματα δίνουν μια βαθμολογία που αντιστοιχεί σε λκφκκφκφκ, ενώ οι αλληλεπιδράσεις διαμορφώνονται μέσω δυνητικών δυνατοτήτων. Οι δυνατότητες της ενεργητικής ενεργειακής αποδοτικότητας μειώνονται απέναντι στη επίτευξη του στόχου. Ενώ στην εργασία των Fischler και Elschlager το 1973, η ελαχιστοποίηση γίνεται μέσω τοπικής αναζήτησης, ο Felzenszwalb και ο Huttenlocher το 2005, πρότειναν μια στατιστική προσέγγιση που οδηγεί σε αποτελεσματική βελτιστοποίηση. Η συνάρτηση κόστους εμπεριέχει μια παράμετρο που σχετίζεται με την εμφάνιση και μια που σχετίζεται με την διαμόρφωση της στάσης. Η ελαχιστοποίηση πραγματοποιείται ταυτόχρονα σε σχέση με τα δύο.

Εικόνα 1: Μοντέλο εικονογραφικών δομών, Πηγή: Pictorial Structures for Object Recognition

Σε αυτή τη γραμμή, παρουσιάστηκαν μερικές εξελιγμένες εναλλακτικές μορφές των μοντέλων εικονογραφικών δομών όπως το μοντέλο με μίξη, το ιεραρχικό, το πολυμορφικό μοντέλο ή το μοντέλο δυνατής εμφάνισης. Τα μοντέλα με μίξη (Yang και Ramanan 2011) κωδικοποιούν την άρθρωση με τη χρήση πολλών, μικρών, μη προσανατολισμένων προτύπων, αντί για παρατεταμένες και περιστρεφόμενες πλατφόρμες (Felzenszwalb και Huttenlocher 2005). Αυτά τα μοντέλα, τα οποία σχηματίζονται ως δομή δέντρου, προσεγγίζουν μικρές καμπύλες και γωνίες του ανθρώπινου σώματος με τη μετάφραση μικρών προτύπων που συνδέονται με ελατήρια. Τα ιεραρχικά μοντέλα (Tian, Zitnick και Narasimhan 2012) καθιστούν δυνατή την καταγραφή των σχέσεων μεταξύ τους, εκτός από τις τοπικές πληροφορίες, με την εισαγωγή ιεραρχικών δέντρων. Συγκεκριμένα, οι ιεραρχικές δομές ενσωματώνουν λανθάνοντες κόμβους και μίξη μερών για την αποτελεσματική δημιουργία συμπερασμάτων.

Στη συνέχεια, τα πολυτροπικά δομημένα μοντέλα (MODEC) (Sapp και Taskar 2013) εισάγουν ένα σύνολο βοηθητικών μεταβλητών, γνωστόν ως λειτουργίες, οι οποίοι σχηματίζουν ένα γράφημα δέντρου μαζί με τη θέση των αρθρώσεων. Αυτοί οι τρόποι προορίζονται να μοντελοποιήσουν ένα υποτοπικό σύνολο αρθρώσεων, έτσι ώστε η συμπερίληψη σε ολόκληρο το σύνολο των υπομονάδων να είναι εφικτή για όλο το σώμα. Στην εργασία των Sapp και Taskar το 2013, οι υπομονάδες είναι το αριστερό και το δεξί μέρος του σώματος αντίστοιχα, όπου το καθένα λαμβάνει ένα από τα 32 πιθανά χαρακτηριστικά(δηλ. το αριστερό και το δεξί μέρος του σώματος παρουσιάζουν αρθρώσεις συμβατές μεταξύ τους). Για να επιταχυνθεί η συμπερίληψη,

μια αλληλουχία φίλτρων σε ολόκληρο το σώμα, ώστε να απομακρύνει απίθανες διαμορφώσεις της κατάστασης, έτσι ώστε να γίνονται συμπέρασμα στις υπόλοιπες λειτουργίες. Η συνολική βαθμολογία για τις κοινές χωρικές τοποθεσίες και υπομονάδες σχηματίζουν ένα τυχαίων συνθηκών πεδίο (Lafferty, McCallum και Pereira 2001) δέντρο-γράφημα, έτσι ώστε να μπορεί να εφαρμοστεί ακριβές συμπέρασμα από την μετάδοση των μηνυμάτων.

Μια άλλη επέκταση του μοντέλου εικονογραφικών δομών (Felzenszwab και Huttenlocher 2005) είναι τα μοντέλα ΄΄ισχυρής εμφάνισης΄΄ που εισήχθησαν στην εργασία (Pischulin et al. 2013). Σε αυτή τους την δουλειά επεκτείνουν διαδοχικά τις εικονογραφικές δομές για να βελτιώσουν τις υποθέσεις του τμήματος του σώματος. Στην αρχή, η απλοποιημένη δομή του μοντέλου αναδεικνύεται με ευέλικτο μοντέλο, το οποίο περιλαμβάνει ενσωματωμένους ανιχνευτές μεμονομένων εξαρτημάτων. Αυτή η δομή έχει ισχυρότερες αναπαραστάσεις τοπικής εμφάνισης σε σύγκριση με το βασικό μοντέλο PS. Στη συνέχεια (Pischulin et al. 2013) αντικατέστησαν ανιχνευτές μεμονωμένων εξαρτημάτων με μίγματα ανιχνευτών μερών για την επίτευξη καλύτερων αποτελεσμάτων. Τέλος χρησιμοποίησαν το μοντέλο το οποίο είναι ανεφοδιασμένο με μείγματα ανιχνευτών μερών μαζί με τα semi-global αντικείμενα. Το τελικό μοντέλο συνδυάζει την τοπική εμφάνιση (λόγω του μοντέλου με τ μίγματα ανιχνευτών μερών) και τις μεσαίου επιπέδου αναπαραστάσεις (λόγω ενσωμάτωσης των θέσεων) οδηγώντας σε σύγχρονες επιδόσεις με μεθόδους που βασίζονται σε μερικές μεθόδους [2,3,4,5,6].

1.3 Όραση υπολογιστών

Το αντικείμενο της όρασης υπολογιστών αποσκοπεί σε δύο στόχους. Από την σκοπιά της βιολογικής επιστήμης, στόχος της όρασης των ηλεκτρονικών υπολογιστών είναι να βρούμε υπολογιστικά μοντέλα του ανθρώπινου οπτικού συστήματος. Από την σκοπιά της μηχανικής θεωρείται ότι η όραση των υπολογιστών αποσκοπεί στην κατασκευή αυτόνομων συστημάτων που θα μπορούσαν να εκτελέσουν ορισμένα από τα καθήκοντα που το ανθρώπινο οπτικό σύστημα μπορεί να εκτελέσει (και να το ξεπεράσει σε πολλές περιπτώσεις). Πολλές εργασίες της όρασης ηλεκτρονικών υπολογιστών σχετίζονται με την εξαγωγή πληροφοριών 3D και χρονικών από χρονικά μεταβαλλόμενα δεδομένα 2D όπως λαμβάνεται από μια ή περισσότερες τηλεοπτικές κάμερες, και γενικότερα την κατανόηση τέτοιων δυναμικών σκηνών.

Φυσικά οι δύο αυτοί στόχοι είναι στενά συνδεδεμένοι. Οι ιδιότητες και τα χαρακτηριστικά του ανθρώπινου οπτικού συστήματος συχνά δίνουν έμπνευση στους μηχανικούς που σχεδιάζουν συστήματα οράσεως υπολογιστών. Αντίθετα οι αλγόριθμοι της όρασης υπολογιστών μπορούν να προσφέρουν πληροφορίες για το πώς το ανθρώπινο οπτικό σύστημα δρα.

Είναι κοινώς αποδεκτό ότι ο πατέρας της όρασης υπολογιστών είναι ο Larry Roberts, ο οποίος στο διδακτορικό του στο Massachusetts Institute of Technology (1960) εξέτασέ τις δυνατότητες εξαγωγής 3D γεωμετρικών πληροφοριών, από 2D προοπτικές προβολές μπλοκ (polyhedral) [1]. Πολλοί ερευνητές, στο MIT και αλλού, που ασχολήθηκαν με τον τομέα της τεχνητής νοημοσύνης ακολούθησαν αυτό το έργο και μελέτησαν την όραση υπολογιστών στο πλαίσιο των προοπτικών προβολών μπλοκ.

Αργότερα οι ερευνητές συνειδητοποίησαν ότι ήταν απαραίτητο να αντιμετωπιστούν εικόνες από τον πραγματικό κόσμο και χρειάζονται έρευνες στα λεγόμενα καθήκοντα όρασης ΄΄χαμηλού επιπέδου΄΄ όπως η ανίχνευση άκρων και η κατάτμηση. Ένα σημαντικό ορόσημο ήταν το πλαίσιο που πρότεινε ο David Marr (γύρω στο 1978) στο MIT, ο οποίος υιοθέτησε μια προσέγγιση από την βάση προς την κατεύθυνση της κατανόησης της σκηνής.

Οι αλγόριθμοι επεξεργασίας εικόνας χαμηλού επιπέδου εφαρμόζονται σε 2D εικόνες για να αποκτήσουν το αρχικό σκίτσο (τμήματα κατευθυνόμενων άκρων κλπ), από τα οποία προκύπτει ένα σκίτσο 2,5D της σκηνής χρησιμοποιώντας διόφθαλμο στερεό. Τέλος χρησιμοποιούνται τεχνικές υψηλού επιπέδου (διαρθρωτική ανάλυση, a priori γνώση) για την λήψη 3D παραστάσεων μοντέλων των αντικειμένων στην σκηνή.

Παρόλα αυτά, πιο πρόσφατα αρκετοί ερευνητές στον τομέα της όρασης υπολογιστών συνειδητοποίησαν μερικούς από τους περιορισμούς του πρότυπου του Marr και υποστήριξαν μια πιο ανομοιογενή και ετερογενή προσέγγιση. Βασικά το πρόγραμμα του Marr είναι εξαιρετικά δύσκολο να πραγματοποιηθεί, αλλά το πιο βασικό είναι οι πως οι ερευνητές συνειδητοποίησαν πως στις περισσότερες εφαρμογές στην όραση υπολογιστών δεν είναι απαραίτητο να αποκτήσουμε πλήρη μοντέλα 3D αντικειμένων. Για παράδειγμα, στην αυτόνομη πλοήγηση οχημάτων με χρήση υπολογιστή, ενδέχεται να είναι απαραίτητο να μάθετε μόνο αν ένα αντικείμενο κινείται μακριά από και προς το όχημα σας, αλλά όχι την ακριβή κίνηση 3D του αντικειμένου. Αυτό το νέο πρότυπο ονομάζεται μερικές φορές ΄΄Purposive Vision΄΄ που υποδηλώνει ότι οι αλγόριθμοι πρέπει να καθοδηγούνται από στόχους και σε πολλές περιπτώσεις να είναι ποιοτικοί [3]. Ένας από τους κύριους υποστηρικτές αυτού του νέου παραδείγματος είναι ο Γιάννης Αλοίμονος, του Πανεπιστημίου του Maryland.

Κοιτάζοντας την ιστορία της όρασης των υπολογιστών, είναι σημαντικό να σημειωθεί ότι λόγω του ευρέως φάσματος δυνητικών εφαρμογών, η τάση ήταν η συγχώνευση της όρασης υπολογιστών με άλλα στενά συνδεδεμένα πεδία. Αυτά περιλαμβάνουν:

- Επεξεργασία εικόνας: οι πρώτες εικόνες πρέπει να επεξεργασθούν πριν από την περαιτέρω ανάλυση.

- Φωτογραμμετρία: οι κάμερές που χρησιμοποιούνται για την απεικόνιση πρέπει να βαθμονομηθούν. Ο προσδιορισμός του αντικειμένου σε 3D είναι σημαντικό τόσο για την όραση όσο και για την φωτογραμμετρία.

- Τα γραφικά υπολογιστών: η 3D μοντελοποίηση είναι κεντρική τόσο για την όραση του υπολογιστή όσο και για τα γραφικά του υπολογιστή. Οι συναρπαστικές εφαρμογές χρειάζονται ταυτόχρονα την όραση και τα γραφικά των υπολογιστών.

Το πεδίο έρευνας της όρασης υπολογιστών είναι γνωστό πως είναι δύσκολο. Σχεδόν κανένα πρόβλημα έρευνας δεν έχει λυθεί ικανοποιητικά. Ένας βασικός λόγος για την δυσκολία του συγκεκριμένου αντικειμένου είναι ότι το ανθρώπινο οπτικό σύστημα είναι απλά πάρα πολύ καλό για πολλά καθήκοντα (π.χ. αναγνώριση προσώπου), έτσι ώστε τα συστήματα οράσεως υπολογιστών να υποφέρουν σε σύγκριση με το ανθρώπινο. Ένας άνθρωπος μπορεί να αναγνωρίσει πρόσωπα κάτω από κάθε είδους

παραλλαγές στον φωτισμό, έκφραση κλπ. Στις περισσότερες περιπτώσεις δεν έχουμε καμία δυσκολία να αναγνωρίσουμε ένα φίλο σε μια φωτογραφία που τραβήξαμε ακόμα και πριν πολλά χρόνια. Επίσης φαίνεται ότι δεν υπάρχει όριο για το πόσα πρόσωπα μπορούμε να αποθηκεύσουμε στους εγκεφάλους μας για μελλοντική αναγνώριση. Δεν υπάρχει όμως καμία ελπίδα για την οικοδόμηση ενός αυτόνομου συστήματος με τέτοια αστρική εκτέλεση.

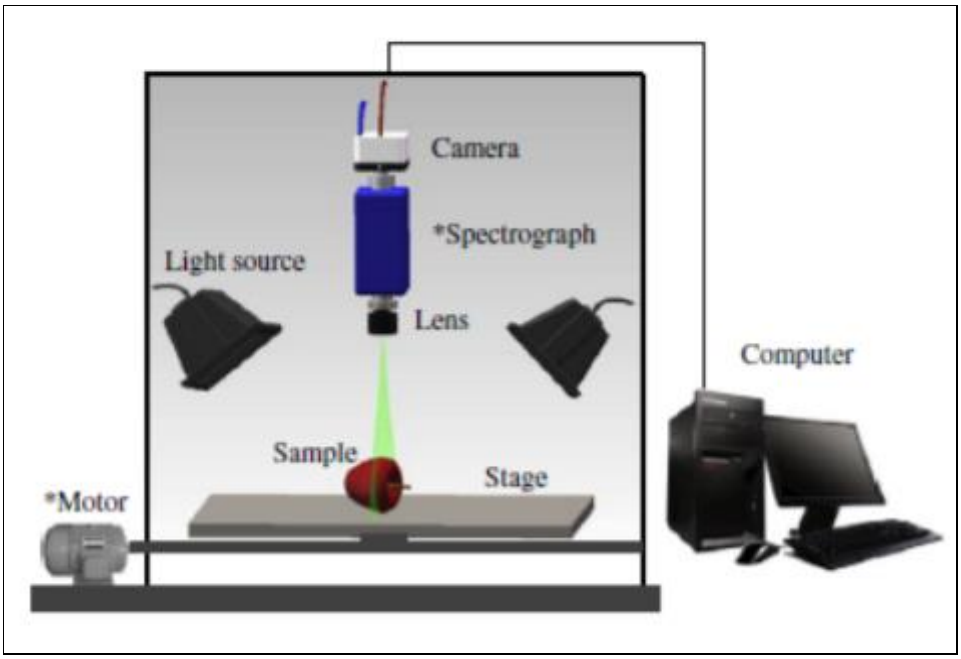
Δύο σημαντικές δυσκολίες στον τομέα της πληροφορικής μπορούν να εντοπιστούν:

1. Πως αποστάζουμε και αντιπροσωπεύουμε την τεράστια ποσότητα ανθρώπινης γνώσης σε ένα υπολογιστή με τέτοιο τρόπο ώστε η ανάκτηση να είναι εύκολη;

2. Πως πραγματοποιούμε (τόσο στο υλικό όσο και στο λογισμικό) το τεράστιο εύρος των υπολογισμών που συχνά απαιτείται κατά τέτοιο τρόπο ώστε η εργασία (όπως η αναγνώριση προσώπου) να μπορεί να γίνει σε πραγματικό χρόνο;

Οι παρελθούσες και οι παρούσες εφαρμογές της οράσεως των υπολογιστών περιλαμβάνουν: αυτόνομη πλοήγηση, ρομποτική και βιομηχανικές επιθεωρήσεις. Στην καλύτερη περίπτωση τα αποτελέσματα ήταν μικτά(αποκλείουμε τις εφαρμογές βιομηχανικής επιθεώρησης που περιλαμβάνουν μόνο επεξεργασία εικόνας 2D και αναγνώριση μοτίβου). Η κύρια δυσκολία είναι ότι στους αλγόριθμους της όρασης των υπολογιστών είναι σχεδόν όλα εύθραυστα. Ένας αλγόριθμος μπορεί να λειτουργήσει σε ορισμένες περιπτώσεις και σε άλλες όχι. Σύμφωνα με την γνώμη αρκετών ερευνητών για να είναι μια εφαρμογή όρασης υπολογιστών επιτυχής θα πρέπει να ικανοποιεί δύο κριτήρια: 1) πιθανότητα ανθρώπινης αλληλεπίδρασης. 2) Συγχώρεση (δηλαδή να είναι ανεκτά μερικά λάθη). Πρέπει επίσης να τονιστεί ότι πολλές εφαρμογές της όρασης των υπολογιστών θα πρέπει να συνδυαστούν με άλλες μορφές(όπως ο ήχος) για την επίτευξη των στόχων.

Η όραση υπολογιστών είναι ένα πεδίο με πάνω από 30 χρόνια ζωής. Αν και ως ερευνητικό πεδίο προσφέρει πολλά προκλητικά και συναρπαστικά προβλήματα, από πλευράς επιτυχημένων εφαρμογών μηχανικής είναι μάλλον απογοητευτικό. Ωστόσο, πιο πρόσφατα, εμφανίστηκαν αρκετές πολύ συναρπαστικές εφαρμογές όπου πιστεύεται πως η όραση των υπολογιστών μπορεί να συμβάλλει σημαντικά [7]



Εικόνα 2: Σύστημα όρασης υπολογιστή, Πηγή: Sciencedirect.com

1.4 *Νευρωνικά δίκτυα και βαθιά μάθηση*

Το έργο στο επιστημονικό πεδίο των νευρωνικών δικτύων βασίστηκε, από τις απαρχές, στο γεγονός ότι ο ανθρώπινος εγκέφαλος εκτελεί τους υπολογισμούς με εντελώς διαφορετικό τρόπο από τον συμβατικό ψηφιακό υπολογιστή. Ο εγκέφαλός είναι ένας εξαιρετικά πολύπλοκος, μη γραμμικός, παράλληλος επεξεργαστής. Έχει την δυνατότητα να οργανώνει τα δομικά του στοιχεία, γνωστά ως νευρώνες, με τρόπο ώστε να εκτελούν συγκεκριμένους υπολογισμούς (π.χ. αναγνώριση προτύπων, αντίληψη και έλεγχος κίνησης) με ταχύτητα πολλαπλάσια από αυτή του γρηγορότερου ψηφιακού υπολογιστή που υπάρχει σήμερα.

Ο χαρακτηρισμός ενός νευρωνικού συστήματος ως ΄΄εξελισσόμενο΄΄ είναι συνώνυμος με την έννοια της πλαστικότητας: αυτή δίνει στο νευρικό σύστημα τη δυνατότητα να προσαρμόζεται ανάλογα με το περιβάλλον του. Και ακριβώς όπως είναι ζωτική για την λειτουργία των νευρώνων ως μονάδες επεξεργασίας πληροφοριών στον ανθρώπινο εγκέφαλο, είναι εξίσου σημαντική για τα νευρωνικά δίκτυα που αποτελούνται από τεχνητούς νευρώνες. Στην πλέον γενική του μορφή, ένα νευρωνικό δίκτυο είναι μια μηχανή σχεδιασμένη να προσομοιώνει τον τρόπο με τον οποίο ο εγκέφαλος εκτελεί μια συγκεκριμένη εργασία ή λειτουργία.

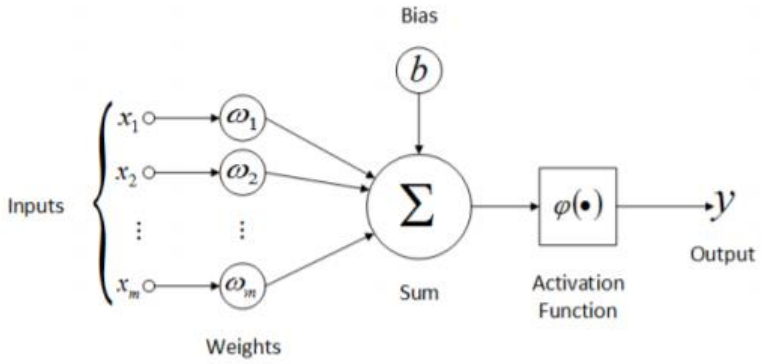
Ένα τεχνητό νευρωνικό δίκτυο είναι ένας μεγάλος παράλληλος επεξεργαστής με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας και έχει από την φύση του τη δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση. Μοιάζει με τον ανθρώπινο εγκέφαλο σε δύο σημεία:

1. Το δίκτυο προσλαμβάνει την γνώση από το περιβάλλον του, μέσω μιας διαδικασίας μάθησης.

2. Η ισχύς των συνδέσεων μεταξύ των νευρώνων, που αποκαλείται συνοπτικό βάρος, χρησιμοποιείται για την αποθήκευση της γνώσης που αποκτιέται.

Η διαδικασία μέσω της οποίας επιτυγχάνεται η μάθηση αποκαλείται αλγόριθμος μάθησης και η λειτουργία του είναι να τροποποιεί τα συνοπτικά βάρη του δικτύου με τον κατάλληλο τρόπο για την επίτευξη του επιθυμητού στόχου.

Το νευρωνικό δίκτυο οφείλει την υπολογιστική του ισχύ κατά πρώτον στην παράλληλη, κατανεμημένη δομή του και κατά δεύτερον στην ικανότητα του να μαθαίνει και ως εκ τούτου να γενικεύει. Ο όρος γενίκευση αναφέρεται στην παραγωγή, από το νευρωνικό δίκτυο, λογικών εξόδων για εισόδους τις οποίες δεν έχει συναντήσει κατά την διάρκεια της εκπαίδευσης του. Αυτές οι δύο δυνατότητες δίνουν στα νευρωνικά δίκτυα την ικανότητα να βρίσκουν καλές προσεγγιστικές λύσεις σε πολύπλοκα προβλήματα, τα οποία είναι όμως μη επιδεκτικά σε λύσεις.



Εικόνα 3: Δομή τεχνητού νευρώνα, Πηγή: sciencedirect.com

Ένας νευρώνας είναι μια μονάδα επεξεργασίας πληροφορίας, η οποία είναι θεμελιώδης για την λειτουργία του νευρωνικού δικτύου. Το ακόλουθο διάγραμμα παρουσιάζει το μοντέλο ενός νευρώνα που αποτελεί την βάση για την σχεδίασης μιας μεγάλης οικογένειας νευρωνικών δικτύων.

Τα τρία βασικά στοιχεία αυτού του νευρώνα είναι:

1. Ένα σύνολο συνάψεων (ή διασυνδέσεων), κάθε μια εκ των οποίων χαρακτηρίζεται από το δικό της βάρος ή δύναμη. Συγκεκριμένα ένα σήμα x_j στην είσοδο της σύναψης j που συνδέεται με τον νευρώνα k πολλαπλασιάζεται επί το συναπτικό βάρος w_{kj}. Ο πρώτος δείκτης στο w_{kj} αναφέρεται στο εν λόγω νευρώνα και ο δεύτερος δείκτης αναφέρεται στο άκρο εισόδου της σύναψης στην οποία αναφέρεται το βάρος. Ανόμοια με το βάρος μια σύναψης στον ανθρώπινο εγκέφαλο, το συναπτικό βάρος ενός τεχνητού νευρώνα μπορεί να λαμβάνει και αρνητικές και θετικές τιμές.
2. Έναν αθροιστή για την άθροιση των σημάτων εισόδου, σταθμισμένων από τα αντίστοιχα συναπτικά βάρη του νευρώνα.
3. Μία συνάρτηση ενεργοποίησης για τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα. Η συνάρτηση ενεργοποίησης αναφέρεται επίσης ως συνάρτηση περιορισμού, επειδή περιορίζει το επιτρεπτό εύρος πλάτους του σήματος εξόδου σε κάποια πεπερασμένη τιμή. Τυπικά, το κανονικοποιημένο εύρος τιμών πλάτους της εξόδου ενός νευρώνα γράφεται ως κλειστό διάστημα,

με την μορφή [0 , 1] ή [-1,1]. Οι πιο συνηθισμένες συναρτήσεις ενεργοποίησης είναι η βηματική, η συνάρτηση προσήμου, η ταυτοτική συνάρτηση (στην περίπτωση γραμμικών νευρώνων) και η σιγμοειδής συνάρτηση.

Το μοντέλο του νευρώνα της εικόνας 2 περιλαμβάνει επίσης μια εξωτερικά εφαρμοζόμενη πόλωση, η οποία συμβολίζεται ως b_k .Η πόλωση έχει ως αποτέλεσμα την αύξηση ή την μείωση της δικτυακής διέγερσης της συνάρτησης ενεργοποίησης, ανάλογα με το εάν είναι θετική ή αρνητική, αντίστοιχα. Με μαθηματικούς όρους ο νευρώνας k που απεικονίζεται στο διάγραμμα μπορεί να περιγράφει με το ακόλουθο ζεύγος εξισώσεων:

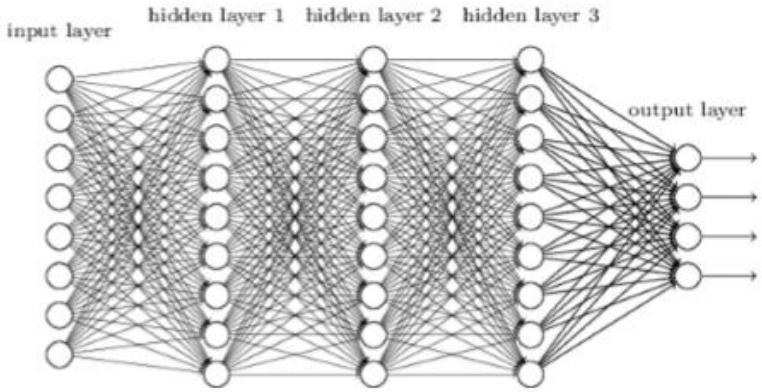
$$u_k = \sum_{j=1}^{m} w_{kj} \times x_j \quad (1.1)$$

$$y_k = \Phi(u_k + b_k) \quad (1.2)$$

Ως μηχανική μάθηση ορίζεται το ευρύ επιστημονικό πεδίο της τεχνητής νοημοσύνης, το οποίο δίνει τη δυνατότητα στους υπολογιστές να ΄΄μαθαίνουν΄΄ με δεδομένα και να βελτιώνονται με το πέρασμα του χρόνου με αυτόματο τρόπο, ενώ τροφοδοτούνται με πληροφορία. Η μηχανική μάθηση διακρίνεται σε τρείς κατηγορίες:

- Επιβλεπόμενη μάθηση (Supervised learning): Το υπολογιστικό σύστημα δέχεται τα δεδομένα εισόδου μαζί με τις ετικέτες του, δηλαδή τα επιθυμητά χαρακτηριστικά, και ο στόχος είναι να μάθει έναν γενικό κανόνα προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα.
- Μη επιβλεπόμενη μάθηση (Unsupervised learning): Στο σύστημα δεν παρέχονται μαζί με τα δεδομένα εισόδου ετικέτες, δηλαδή επιπρόσθετες πληροφορίες, και καλείται να βρει τη δομή των δεδομένων. Η μη επιτηρούμενη μάθηση μπορεί να χρησιμοποιηθεί για την ανακάλυψη κρυμμένων μοτίβων σε δεδομένα ή ως μέσο για την εύρεση κάποιου χαρακτηριστικού.
- Ενισχυτική μάθηση (Reinforcement learning): Ουσιαστικά πρόκειται για συνδυασμό των προηγούμενων δύο, καθώς το σύστημα λαμβάνει δεδομένα εισόδου χωρίς ετικέτες και προσπαθεί να εξάγει χαρακτηριστικά. Σε περίπτωση που το κάνει με επιτυχία ΄΄επιβραβεύεται΄΄ ενώ αντίθετα δέχεται ΄΄τιμωρία΄΄. Στόχος είναι η μεγιστοποίηση της ανταμοιβής.

Εφόσον έχει οριστεί η έννοια του τεχνητού νευρωνικού δικτύου, καθίσταται δυνατό να αποτυπωθεί η έννοια της βαθιάς μηχανικής μάθησης. Η βαθιά μηχανική μάθηση είναι η διαδικασία εφαρμογής ΄΄μαθησιακών διεργασιών΄΄ σε νευρωνικά δίκτυα πολλών επιπέδων. Αποτελεί κομμάτι μιας ευρύτερης οικογένειας μεθόδων μηχανικής μάθησης βασιζόμενων στην αναπαράσταση δεδομένων, σε αντίθεση σε αλγόριθμους επικεντρωμένους σε υπολογιστικές εργασίες. Οι αρχιτεκτονικές βαθιάς μηχανικής μάθησης έχουν εφαρμογή σε πολλά πεδία της τεχνητής νοημοσύνης, όπως η όραση των υπολογιστών, η αναγνώριση φωνής και η βιοιατρική[8,9].



Εικόνα 4: Βαθύ νευρωνικό δίκτυο, Πηγή: tipsandtrics.com

ΚΕΦΑΛΑΙΟ 2

Ομαδοποίηση και επισκόπηση των διαφορετικών μεθόδων για την εκτίμηση της στάσης του σώματος με χρήση τεχνολογίας οράσεως υπολογιστών

2.1 Εισαγωγή

Η εκτίμηση της στάσης του ανθρώπινου σώματος αποτελεί το επίκεντρο αυτής της διπλωματικής εργασίας. Μια μεγάλη ποικιλία ερευνητικών μελετών έχει εφαρμοστεί σχετικά με την εκτίμηση της ανθρώπινης στάσης, λόγω των πολλαπλών εφαρμογών που βασίζονται στην ανάλυση της ανθρώπινης στάσης σε εικόνες και βίντεο. Υπάρχουν σύμφωνα με την υπάρχουσα βιβλιογραφία τρείς διαφορετικές κατηγορίες: οι πρώιμες μέθοδοι, οι βασισμένες στην δομή εικονογραφικές μέθοδοι και τα βαθιά συνελικτικά νευρωνικά δίκτυα. Αυτές οι τρείς μέθοδοι συζητούνται στις επόμενες υποενότητες.

2.2 Εκτίμηση στάσης από πάνω προς τα κάτω

Οι προσεγγίσεις από πάνω προς τα κάτω ταιριάζουν με ένα άμεσο μοντέλο παρατήρησης της εικόνας. Το άμεσο μοντέλο σημαίνει ότι a priori το ανθρώπινο μοντέλο χρησιμοποιείται ως μοντέλο που αναπαριστά το παρατηρούμενο μοντέλο. Αυτό το ανθρώπινο μοντέλο στην συνέχεια ενημερώνεται συνεχώς από τις παρατηρήσεις. Ως εκ τούτου, παρέχει οποιαδήποτε επιθυμητή πληροφορία ανά πάσα στιγμή. Τα μοντέλα που χρησιμοποιούνται στις μεθόδους βασισμένες σε άμεσα μοντέλα είναι γενικά πολύ λεπτομερή. Ταυτοποιούνται ρητά σε ένα πρόγραμμα υπολογιστή και χρησιμοποιούνται εντατικά κατά την παρατήρηση. Ένα από τα κυριότερα οφέλη από την εισαγωγή ενός ανθρώπινου μοντέλου είναι η ικανότητα να χειριστούμε την απόφραξη από διάφορους κινηματικούς περιορισμούς. Ένας αριθμός αρθρώσεων και ράβδων χρησιμοποιείται για να αναπαραστήσει ένα άμεσο ανθρώπινο μοντέλο και αυτές οι αρθρώσεις συνδέονται με τις ράβδους [10,11].

2.2.1 Οι πρώιμες μέθοδοι

Οι περισσότεροι από τους πρώτους αλγορίθμους εκτίμησης της στάσης του σώματος αντιμετώπισαν το πρόβλημα με μια ρητή γεωμετρική αναπαράσταση του ανθρώπινου σχήματος και κινηματικής δομής για την ανασυγκρότηση της στάσης. Το ανθρώπινο μοντέλο αναπαρίσταται συγκεκριμένα από ένα χώρο κατάστασης όπου ο κάθε άξονας αντιπροσωπεύει ένα βαθμό ελευθερίας μια άρθρωσης. Ως εκ τούτου, ένα σημείο εκφράζει μια στάση στον χώρο κατάστασης που αντιστοιχεί στα σημεία της εικόνα. Το πρόβλημα είναι το πως θα χρησιμοποιήσετε την αναπαράσταση της κατάστασης και πως να συσχετίσετε τα δεδομένα της εικόνας με τα δεδομένα της στάσης. Μια γενική προσέγγιση αντιμετωπίζει αυτό το πρόβλημα χρησιμοποιώντας μια μεθοδολογία ανάλυσης-σύνθεσης για την βελτιστοποίηση της ομοιότητας μεταξύ της προβολής του μοντέλου και των δεδομένων της παρατηρούμενης εικόνας [12]. Έτσι, αυτή η μεθοδολογία που βασίζεται στο μοντέλο ανάλυσης-σύνθεση περιέχει δύο μέρη: το πρώτο είναι η πρόβλεψη της στάσης και τότε, το προβλεπόμενο μοντέλο χρησιμοποιείται για την σύγκριση μεταξύ της προβολής του μοντέλου και των παρατηρούμενων εικόνων.

Σαφώς ο προβλεπόμενος χώρος κατάστασης περιγράφει ένα μεγάλο αριθμό πιθανών στάσεων το οποίο το καθιστά παράλογο να ταιριάζει με τα παρατηρούμενα δεδομένα. Έτσι, η ιδέα της εισαγωγής περιορισμών χρησιμοποιείται για να κλαδεύσουν τον χώρο κατάστασης. Στο πρωτοποριακό έργο [12,13], ο Hogg εισήγαγε μια προσέγγιση για τον προσδιορισμό των 3D θέσεων και στάσεων κινούμενων ατόμων από 2D εικόνες. Ο Moeslund και ο Granum [11] χώρισαν άμεσα τον χώρο κατάστασης σε νόμιμες και παράνομες περιοχές.

Μια άλλη μέθοδος για την μείωση του χώρου κατάστασης είναι η χρήση μιας γνωστής κυκλικής κίνησης(π.χ. τρέξιμο και περπάτημα). Στο [15], εξετάζεται το άτομο που βαδίζει παράλληλα με το επίπεδο της εικόνας. Όλες οι παράμετροι στάσης υπολογίζονται χρησιμοποιώντας ένα κυκλικό μοντέλο κίνησης του ατόμου που βαδίζει. Αυτό είναι μια αποτελεσματικό κλάδεμά για την κυκλική κίνηση. Ο Ong και ο Gong χαρτογράφησαν τα εκπαιδευτικά δεδομένα στο χώρο κατάστασης και χρησιμοποίησαν ένα ιεραρχικό Κύριο Στοιχείο Ανάλυσης (PCA) για την εξαγωγή ενός δευτερεύοντος χώρου για την εκτίμηση του βαθμού ασάφειας στις 2D ενδείξεις. Στο [16] ο Pavlovic πήγε αυτήν την ιδέα ένα βήμα παραπέρα μαθαίνοντας δυναμικά μοντέλα από τις παρατηρούμενες τροχιές του χώρου κατάστασης. Πιο αποτελεσματικά, ο Moeslund και ο Granum [14] μείωσαν την διαστάσεις του χώρου κατάστασης, αναπαριστώντας το ανθρώπινο μοντέλο με διαφορετικούς βαθμούς ελευθερίας ενός δομικού μοντέλου.

Όσο αφορά την σύγκριση μεταξύ του μοντέλου προβολής και των παρατηρούμενων εικόνων, ο Hogg [12,13] πρότεινε την χρήση αφαιρούμενης εικόνας για να αποκτήσει τις άκρες ενός ανθρώπου και συνέκρινε τα άκρα από την εικόνα και το ανθρώπινο μοντέλο. Ένα πιο εξελιγμένο σύστημα [15] συνδύασε τμήματα άκρων με ένα συγκεκριμένο μοντέλο κίνησης για να αποκτήσουν ένα πιο ισχυρό αποτέλεσμα. Ο Wachter και ο Nagel [17] χρησιμοποίησαν τόσο τις άκρες όσο και τις περιφερειακές πληροφορίες στην αντιστοίχιση της εικόνας και του ανθρώπινου μοντέλου.

Η σιλουέτα είναι δεδομένα με βάση την περιοχή και έχει το πλεονέκτημα έναντι των άκρων της ύπαρξης ισχυρό έως θόρυβο. Ο Kameda [18] υπολογίστε την ομοιότητα της

σιλουέτας μεταξύ της εικόνας και το ανθρώπινου μοντέλου. Στο έργο του Hu [19], η ομοιότητα της σιλουέτας υπολογίζεται με μια τοπική στρατηγική αγώνα που βασίζεται στο θετικό και αρνητικά αποτελέσματα αντιστοίχισης. Εφαρμόζουν επίσης γενετικό αλγόριθμο για βελτίωση αποτελέσματα που ταιριάζουν. Επιπλέον, το φίλτρο Kalman προσανατολισμένο στη δομή είναι δεσμευμένο στο μεγάλη μορφολογική κλίμακα για τη βελτίωση της ακρίβειας που ταιριάζει. Σε [20-22], το περίγραμμα χρησιμοποιείται για τον υπολογισμό της ομοιότητας μεταξύ της εικόνας και των δεδομένων του ανθρώπινου μοντέλου.

2.2.2 Εκτίμηση στάσης μονής όψης

Η ανασυγκρότηση της ανθρώπινης στάσης από μονόφθαλμες αλληλουχίες εικόνας είναι σημαντική και αποτελεί ένα ερευνητικό πεδίο πρόκλησης με πολλές εφαρμογές. Σε μια μονόφθαλμη εκτίμηση της ανθρώπινης στάσης, οι κινηματικοί περιορισμοί χρησιμοποιούνται συνήθως στο ανθρώπινο άμεσο μοντέλο [23,24]. Στο έργο των Wachter και Nagel [2.13] το εκτεταμένο φίλτρο Kalman χρησιμοποιείται για την εκτίμηση της ανθρώπινης στάσης με κινηματικούς περιορισμούς. Ο Sminchisescu και ο Triggs [25] έχουν διερευνήσει την εφαρμογή της στοχαστικής δειγματοληψίας για την εκτίμηση της μονόφθαλμης ανθρώπινης στάσης. Χρησιμοποιούν ένα μετρημένο ισχυρό κόστος που συνδυάζει μια ισχυρά εξαγόμενη οπτική ροή, ενέργεια άκρου, όρια κίνησης και προγενέστερα μοντέλα για την αντιστοίχιση εικόνας. Η δειγματοληψία με φουσκωμένο συντελεστή κλίμακας εισάγεται για να καθοδηγήσει τα σωματίδια και να μειώσει τα λανθασμένα τοπικά ελάχιστα. Στην περαιτέρω έρευνα [26], χρησιμοποιούν απλό κινηματικό συλλογισμό για την απαρίθμηση των πιθανών κινήσεων προς τα εμπρός/ προς τα πίσω που προκαλούν οπτικές αμφισημίες.

Πιθανοτικές προσεγγίσεις που χρησιμοποιούν μέρη του ανθρώπινου σώματος μαζί με ανθρώπινη κινηματική έχουν επίσης διερευνηθεί για την εκτίμηση της μονόφθαλμης ανθρώπινης στάσης. Στο [27] χάρτες προτάσεων εισάγονται για να αντιπροσωπεύουν την εκτιμώμενη πιθανότητα των μερών του σώματος σε 3D στάσεις χώρου με ένα ρητό 3D μοντέλο. Μια προσέγγιση με αλυσίδα Markov Chain Monte Carlo (MCMC) βάσεων δεδομένων χρησιμοποιείται για την έρευνα του ανθρώπινου χώρου πόζας. Η MCMC εφαρμόστηκε για την εκτίμηση 3D στάσεων από μεμονωμένες εικόνες αθλητών σε μια ποικιλία σύνθετων στάσεων, αλλά εξακολουθεί να υποφέρει από υψηλό υπολογιστικό κόστος. Οι Moeslund και οι άλλοι [28,29] απασχολούν διαδοχικά MCMC βάσεις δεδομένων για την εκτίμηση της ανθρώπινης στάσης. Ένας ανιχνευτής μερών χρησιμοποιείται για τον εντοπισμό της θέσης του χεριού στην εικόνα. Αυτή η εκτίμηση εφαρμόζεται για να διορθώσει την πρόβλεψη και την μείωση του αριθμού των σωματιδίων. Οι Navaratnam και οι άλλοι [30] προτείνουν ένα ιεραρχικό κινηματικό μοντέλο μερικής βάσης για την εκτίμηση της στάσης του άνω σώματος. Το ανθρώπινο σώμα αντιμετωπίζεται ως μια συλλογή μερών που συνδέονται σε μια κινηματική αλυσίδα. Οι κινηματικοί περιορισμοί σε μια συλλογή συνδεδεμένων τμημάτων παρουσιάζονται ιεραρχικά.

Η εκτίμηση της ανθρώπινης στάσης με το μοντέλο από πάνω προς τα κάτω μονής όψης υποφέρει από συσσώρευση λαθών. Σε περίπτωση ασάφειας, όπως η αυτό-απόφραξη, υπάρχει πού μεγάλη πιθανότητα να επιλέξει λάθος στάση. Αυτά τα σφάλματα καθιστούν δύσκολη την ανάκτηση της στάσης.

2.2.3 Εκτίμηση στάσης πολλαπλών όψεων

Η επανακατασκευή της ανθρώπινης στάσης από πολλές εικόνες προβολής είναι πιο αποτελεσματική λύση για πολύπλοκες κινήσεις. Αυτό χρησιμοποιείται για να ξεπεραστούν τα προβλήματα εκτίμησης στάσης της μονής όψης. Χρησιμοποιούνται ντετερμινιστικές μέθοδοι διαβάθμισης κλίσης για την εκτίμηση της ανθρώπινης στάσης σε πολλές σκηνές προβολής. Στο έργο Delamarre και Faugeras [31] εκτιμάται η κίνηση ενός αρθρωτού αντικειμένου σε δύο ή περισσότερες σταθερές κάμερες λαμβάνοντας υπόψιν την ποιότητα των εικόνων σε όλες τις προβολές. Επιπλέον οι φυσικές δυνάμεις εφαρμόζονται σε κάθε μέρος του σώματος σε κινηματικό 3D ανθρώπινο μοντέλο. Αυτές οι δυνάμεις καθοδηγούν την ελαχιστοποίηση των διαφορών μεταξύ των εικόνων και προβλεπόμενου τρισδιάστατου μοντέλου. Πολλές εργασίες εργάστηκαν μια μεθοδολογία ανάλυσης-σύνθεσης ντετερμινιστικής διαβάθμισης κλίσης για εφαρμογές πιο σύνθετων κινήσεων. Στο έργο των Plankers και Fua [32] χρησιμοποίησαν στερεό και ενδείξεις σκιαγραφίας για να χειριστούν περίπλοκες κινήσεις που περιέχουν αυτό-απόφραξη. Ένας κοινός περιορισμός των εφαρμογών διαβάθμισης κλίσης είναι η χρήση εκτίμησης στάσης με βάση την διανομή Gauss. Επομένως περιορίζεται στην μονοτροπική κατανομή πιθανότητας. Στην πράξη η εκτίμηση στάσης είναι συνήθως πολυτροπικό και μη Γκαουζιανό πρόβλημα. Για να επιτύχουν μια πιο στιβαρή παρακολούθηση, χρησιμοποιούνται στρατηγικές στοχαστικής δειγματοληψίας για την αναζήτηση του χώρου κατάστασης στάσης.

Το φίλτρο σωματιδίων είναι μια στοχαστική τεχνική για την εκτίμηση και την παρακολούθηση στάσης. Το φιλτράρισμα σωματιδίων [33,34] είναι μία από τις πιο κοινές προσεγγίσεις για την παρακολούθηση της ανθρώπινης κίνησης, που χρησιμοποιείσαι την στάση στο τρέχον πλαίσιο και ένα δυναμικό μοντέλο για να προβλέψει την επόμενη στάση. Το φίλτρο σωματιδίου (PF) χρησιμοποιεί πολλαπλές προβλέψεις, που λαμβάνονται με την λήψη δειγμάτων της στάσης και της θέσης πριν και στην συνέχεια διαδίδοντας τα χρησιμοποιώντας το δυναμικό μοντέλο συγκρίνοντας τα με τα παρατηρούμενα δεδομένα της εικόνας και υπολογίζοντας την πιθανότητα. Η προηγούμενη στάση είναι συνήθως αρκετά διασκορπισμένη αλλά η συνάρτηση πιθανότητας του δυναμικού μοντέλου μπορεί να είναι πολύ υψηλή, περιλαμβάνοντας πολλά τοπικά μέγιστα τα οποία είναι δύσκολο να τα υπολογίσεις με λεπτομέρεια. Η κύρια δυσκολία με την εφαρμογή των φίλτρων σωματιδίων είναι η υψηλή διάσταση του χώρου κατάστασης στην εκτίμηση στην στάσης. Έτσι, ο αριθμός των σωματιδίων αυξάνεται εκθετικά με την διάσταση. Στο έργο των MacCormick και Isard [35] εισήγαγαν την τεχνική της δειγματοληψίας διχοτόμησης για να μειώσει την διάσταση του χώρου κατάστασης για αποτελεσματική εκτίμηση 2D στάσης των αρθρωτών αντικειμένων. Υπάρχουν δύο χαρακτηριστικά της δειγματοληψίας διχοτόμησης στο πεδίο των αρθρωτών αντικειμένων: το πρώτο είναι ο αριθμός των δειγμάτων που αφιερώνονται σε κάθε διαμελισμό μπορεί να ποικίλει για σημαντικές υπολογιστικές προόδους και δεύτερον ο αριθμός των εκτιμήσεων πιθανότητας μπορεί να μειωθεί στο μισό εκφράζοντας τις πιθανότητες ως μια εύκολή υπολογιστική συνάρτηση. Επιπλέον, είναι σύστημα αυτό-αρχικοποίησης και πραγματικού χρόνου και δείχνει την ευρωστία και την ακρίβεια για πιο πολύπλοκες διαδραστικές εργασίες. Ωστόσο, αυτή η εφαρμογή εφαρμόζεται μόνο στην παρακολούθηση χεριών και δεν επεκτείνεται στη συνολική εκτίμηση της στάσης του σώματος.
Το φίλτρο ανόπτησης σωματιδίων (APF) προτείνεται στο έργο του Deutscher και των υπολοίπων [36,37] και χρησιμοποιείται για την καταγραφή της ανθρώπινης κίνησης σε

ένα σύστημα πολλαπλών καμερών. Συνδυάζουν μια προσομοιωμένη ανόπτηση με φίλτρο σωματιδίων που φαίνεται να είναι αποτελεσματική στην αναζήτηση των υψηλών-διαστάσεων διαμορφωμένων χώρων στη αρθρωτή εκτίμηση στάσης και παρακολούθηση της κίνησης του σώματος. Αυτή η εφαρμογή χρησιμοποιεί μια αρχή συνέχισης για την σταδιακή εισαγωγή επιρροής των στενών κορυφών (peak) στην συνάρτηση καταλληλότητας. Το παραδοσιακό φίλτρο σωματιδίων έχει το πρόβλημα ότι μπορεί εύκολα να αποσπάται από τοπικά μέγιστα. Στο φίλτρο ανόπτησης σωματιδίων, το αραιό σύνολο σωματιδίων μπορεί να κινηθεί σταδιακά προς το γενικό μέγιστο χωρίς να αποσπάται από τα τοπικά μέγιστα. Επιπλέον, οι βελτιώνει και επεκτείνει το APF με δύο τρόπους. Πρώτον, εφαρμόζεται ένας μηχανισμός στο χώρο αναζήτησης για να επιτευχθεί μια απαλή διχοτόμηση. Προτείνουν ένα μέσο για να κάνουν το βήμα διάχυσης στο APF προσαρμοστικό κατά την ανόπτηση. Αυτό μπορεί να οδηγήσει σε αυτό που μπορεί να ερμηνευτεί ως μια απαλή στρατηγική ιεραρχικής αναζήτησης που χωρίζει αυτόματα το χώρο αναζήτησης και ως εκ τούτου οδηγεί σε περαιτέρω αύξηση στην αποτελεσματικότητα. Το δεύτερο είναι ότι εισάγουν ένα τελεστή διασταύρωσης (παρόμοιο με αυτόν που βρίσκεται στους γενετικούς αλγόριθμους) στη δομή του φίλτρου σωματιδίων. Δείχνουν ότι αυτός ο τελεστής βελτιώνει την ικανότητα των ανιχνευτών να αναζητήσουν τους χώρους διαμόρφωσης των αρθρωτών αντικειμένων.

Ο Sigal και ο Balan [38] στο έργο τους παρουσίασαν το σύνολο δεδομένων HumanEva για ποσοτική αξιολόγηση των ανταγωνιστικών μεθόδων της αρθρωτής ανθρώπινης στάσης. Το HumanEva είναι ένα τυπικό σημείο αναφοράς για την εκτίμηση ανθρώπινης στάσης πολλαπλών προβολών 3D στο εργαστήριο. Αυτό το σύνολο δεδομένων αποτελείται από το HumanEva-I και το HumanEva-II από ένα σύνολο ακολουθιών πολλαπλών προβολών. Το σύνολο δεδομένων περιέχει βάδισμα, τζόνκινκ, χειρονομίες χεριών, ρίψη και πιάσιμο μπάλας και στυλ κινήσεων πυγμαχίας από τρία διαφορετικά θέματα. Επίσης παρουσίασαν μια βασική μεθοδολογική γραμμή για την παρακολούθηση αρθρωτών αντικειμένων. Ένα σχετικά τυπικό πλαίσιο Bayesian χρησιμοποιείται για την βελτιστοποίηση με τη μορφή διαδοχικής σπουδαιότητας επαναπροσδιορισμού και APF. Συνδυάζουν την συνάρτηση πιθανότητας με βάση την ακμή, την συνάρτηση πιθανότητας βασισμένη στην σκιαγραφία και την αμφίδρομη συνάρτηση πιθανότητας βασισμένη στην σκιαγραφία μαζί στην μεταγενέστερη αναπαράσταση. Το φίλτρο APF έχει ευρέως χρησιμοποιηθεί για την αρθρωτή παρακολούθηση της ανθρώπινης κίνησης λόγω της ικανότητας του να εκτιμά ακριβώς τα στατιστικά των πολυτροπικών και μη Γκαουζιανών διαδικασιών. Ωστόσο, η απόδοση των φίλτρων ανόπτησης σωματιδίων μειώνεται όταν ο ρυθμός καρέ είναι χαμηλότερος ή η κίνηση είναι γρήγορη.

Υπάρχουν μερικές εργασίες που συνδυάζουν στοχαστική αναζήτηση με διαβάθμιση κλίσης για την εκτίμηση τοπικού μέρους για την ανάκτηση της κίνησης ολόκληρου του σώματος. Το έργο του Carranza και των υπολοίπων [39] αποδεικνύει την εκτίμηση στάσης ολόκληρου του σώματος πολλαπλών όψεων συνδυάζοντας μια ντετερμινιστική αναζήτηση πλέγματος με διαβάθμιση κλίσης. Για κάθε μέρος του σώματος, μια αναζήτηση πλέγματος βρίσκει πρώτα το σύνολο των έγκυρων θέσεων ελαχιστοποιώντας την επικάλυψη μεταξύ των παρατηρούμενων 2D σχημάτων και των προβολών του μοντέλου. Μια συνάρτηση καταλληλότητας χρησιμοποιείται για την αξιολόγηση των έγκυρων στάσεων για να βρει το καλύτερη στάση. Στην συνέχεια, αυτή η καλύτερη στάση βελτιώνεται με βελτιστοποίηση της κλίση διαβάθμισης. Παρόλο που η μέθοδος τους δεν απαιτεί συγκεκριμένη 3D ανακατασκευή, ένα ακριβές

μοντέλο σώματος και τμηματοποίηση του ατόμου στα διαφορετικά σημεία προβολής είναι ζωτικής σημασίας για την επίτευξη μιας πολύ σημαντικής μέτρησης. Σε σχετική εργασία του Kehl και των υπολοίπων [40] προτείνεται μια Stochastic Meta Descent (SMD) με στοχαστική δειγματοληψία για την εκτίμηση στάσης ολόκληρου του σώματος μέσω πολλαπλών προβολών. Εισάγουν την βελτιστοποίηση SMD η οποία επιτρέπει την αποφυγή της προσέγγισης σύγκλισης σε τοπικά ελάχιστα.

2.2.4 Προτεραιότητες κίνησης

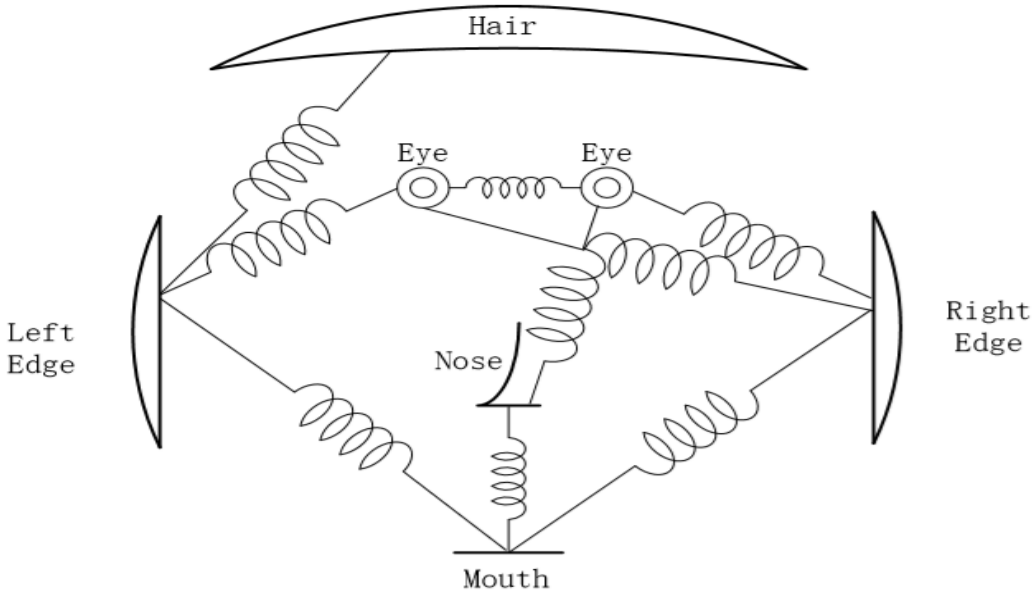
Υπάρχουν πολλές έρευνες που επικεντρώνονται στα μοντέλα προτεραιότητας κίνησης τα οποία προέρχονται από τη προπόνηση δεδομένων από μονή ή πολλαπλή όψη. Τα περισσότερα στατιστικά μοντέλα κίνησης μπορούν μόνο να χρησιμοποιηθούν για συγκεκριμένες κινήσεις με(περπάτημα και τζόνινκ) με καθορισμένους περιορισμούς. Όταν θεωρείται μόνο μια κατηγορία κινήσεων, οι προτεραιότητες κίνησης μπορούν να βοηθήσουν για να βελτιώσουν την απόδοση στην εκτίμηση της στάσης [2.30].

Στην εργασία Sidenblath και των υπολοίπων [42-44] συνδυάζεται η στοχαστική δειγματοληψία με ένα δυνατή μάθηση προτεραιότητα κίνησης από συγκεκριμένων κινήσεων. Τα δείγματα διαδίδονται σε ένα πλαίσιο φίλτρου σωματιδίων με την δυναμική δείγματος. Μια εφαρμογή βασισμένη σε αντίγραφο χρησιμοποιείται στην εργασία [44] όπου τα παραδείγματα κίνησης έχουν βρεθεί να δείχνουν πιθανές κινήσεις κατευθύνσεων. Στην εργασία [45], η ανθρώπινη εμφάνιση και η είκονα προτεραιοτήτων κίνησης συνδυάζονται μαζί για να μοντελοποίησουν την πιθανότητα από την παρακολούθηση διαφόρων εικόνων για μια δοσμένη κίνηση. Αυτές οι μέθοδοι απασχολούν μια μεθοδολογία ανάλυσης από σύνθεση στην αναδημιουργία ανθρώπινης κίνησης. Ομοίως, μια ιεραρχική ανάλυση θεμελιωδών κύριων συστατικών (PCA) χρησιμοποιείται στην κωδικοποίηση της γεωμετρίας και της κινηματικής και το μοντέλο κρυμμένων Markov (HMM) χρησιμοποιείται στην αναπαράσταση της ανθρώπινης δυναμικής για την εκτίμηση της στάσης μονής όψης [46]. Ο Agarwal και ο Triggs [47] στο έργο τους δημιούργησαν ένα σύμπλεγμα των εκπαιδευόμενων δεδομένων τους σε στάσεις σώματος με παρόμοιες δυναμικές για πιο γενικές κινήσεις (περπάτημα και τρέξιμο). Η δουλειά τους αποδεικνύει ότι αυτό το ισχυρό προηγούμενο ανθρώπινων κινήσεων επιτρέπει την 2D εκτίμηση στάσης για κινήσεις που κινούνται γρήγορα.

Υπάρχουν μερικές έρευνες που έχουν διευρύνει την χρήση προηγούμενων κινήσεων για 3D ανασυγκρότηση κίνησης. Στην εργασία του Howe και των υπολοίπων [48] χρησιμοποιούνται αποσπάσματα κίνησης από μια βάση δεδομένων για να αποσπάσουν 3D στάσεις από εντοπισμένα χαρακτηριστικά γνωρίσματα εικόνων από απλές κινήσεις. Από μια ακολουθία 2D στάσεων, η 3D κίνηση ανασυγκροτείται βρίσκοντας τον χάρτη (MAP) με εκτίμηση των ακολουθιών μικρών κινήσεων. Ο Sigal και οι υπόλοιποι στο έργο τους [49] υιοθέτησαν τους ανιχνευτές άκρων και κεφαλιών που ενσωματώνονται στο εκπαιδευμένο πρότυπο κινήσεων για να συμπεράνει την ανθρώπινη στάση του περπατήματος με μονή όψη, με αυτόματη προετοιμασία. Η εκτίμηση της ανθρώπινης στάσης και κίνησης επιλύεται με μη παραμετρική διάδοση πεποιθήσεων μέσω στοχαστικής δειγματοληψίας πάνω από ένα βρόχου γράφημα. Η εργασία των Urtasum και Fua [50] χρησιμοποιεί προσωρινά μοντέλα κίνησης από ακολουθίες καταγραφημένων δεδομένων κίνησης. Η 3D ανθρώπινη κίνηση αναδημιουργείται

χρησιμοποιώντας ένα ντετερμισνιστικό σχέδιο βελτιστοποίησης. Ταιριάζουν σε μοντέλα πλήρους σώματος σε στερεοφωνικά δεδομένα για περπάτημα και τρέξιμο. Επιπλέον, χρησιμοποιώντας μια βάση δεδομένων πολλαπλών δραστηριοτήτων, το παραμετρικό μοντέλο κίνησης χρησιμοποιείται στην συνέχεια για τον περιορισμό των κινήσεων με μεταβλητή ταχύτητα από στέρεοφωνικό (stereo). Στο έργο του Urtasun και των υπολοίπων [51], χρησιμοποιεί ένα Γκαουσιανό μοντέλο διεργασίας λανθάνοντος μεταβλητού (GPLVM) για να μαθαίνονται προηγούμενα μοντέλα συγκεκριμένων κινήσεων, όπως το γκολφ ή το περπάτημα από ακολουθίες εικόνων μονής-όψης. Το μοντέλο GPLVM δημιουργεί ομαλές αντιστοιχίσεις μεταξύ του χώρου στάσης και του λανθάνοντα χώρου, που είναι χρήσιμο για την χρήση της κλίσης καθόδου για την βελτίωση της εκτίμησης της ανθρώπινης στάσης. Σε μια μεταγενέστερη εργασία [52,53] ένα Γκαουσιανό μοντέλο δυναμικής διαδικασίας (GPDM) μαθαίνεται ένα δυναμικό πρότυπο κινήσεων στο λανθάνον διάστημα από τα δεδομένα κατάρτισης. Η εργασία των Moon και Pavlovic [54] έχει ερευνήσει την επίδραση της ειδικής δυναμικής στην ενσωμάτωση της εκτίμησης της ανθρώπινης κίνησης των 3D αρθρωτών αντικειμένων σε ακολουθίες εικόνων μονής-όψης.

Εν ολίγης, η εισαγωγή μια συγκεκριμένης κίνησης του ανθρώπινου σώματος έχει επιτύχει την εκτίμηση των πολύπλοκων κινήσεων από πολλαπλές-όψεις ή μονές όψεις. Εντούτοις, υπάρχουν δύο κύρια μειονεκτήματα της εκτίμησης από πάνω προς τα κάτω. Το πρώτο είναι το γεγονός ότι η μη αυτόματη προετοιμασία στο πρώτο καρέ μια ακολουθίας βίντεο είναι απαραίτητή, δεδομένου ότι η αρχική εκτίμηση λαμβάνεται συχνά από το προηγούμενο πλαίσιο. Ένα άλλο μειονέκτημα είναι το υψηλό υπολογιστικό κόστος της απόδοσης προς τα εμπρός του μοντέλου του ανθρώπινου σώματος (3D ή 2D) και τον υπολογισμό της ομοιότητας μεταξύ της προβολής του ανθρώπινου μοντέλου και των παρατηρούμενων εικόνων.



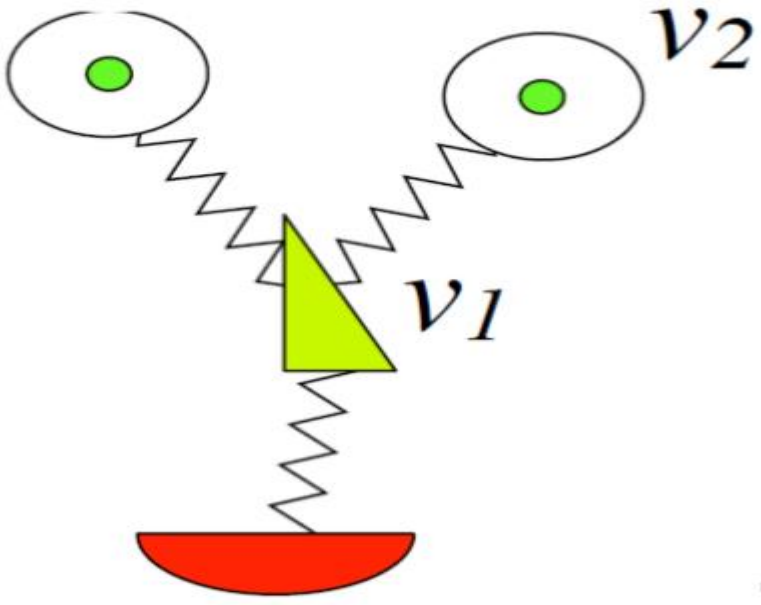
Εικόνα 5: Σχηματική αναπαράσταση του μοντέλου προσώπου, υποδεικνύοντας τα στοιχεία και τους δεσμούς τους, Πηγή: researchgate.com

2.3 Εκτίμηση από την βάση προς τα επάνω

Οι προσεγγίσεις εκτίμησης από πάνω προς τα κάτω χαρακτηρίζονται από την ανίχνευση τμημάτων του ανθρώπινου σώματος και στην συνέχεια την συναρμολόγηση αυτών σε μια δομή ανθρώπινου σώματος. Τα μέρη του σώματος περιγράφονται συνήθως από 2D πρότυπα και βρίσκονται από ανιχνευτές τμημάτων. Οι προσεγγίσεις από κάτω προς τα πάνω έχουν το πλεονέκτημα ότι δεν απαιτείται χειροκίνητη προετοιμασία και δεν απαιτείται συγκεκριμένο μοντέλο πριν. Έτσι, οι μέθοδοι εκτίμησης από την βάση προς τα επάνω έχουν μικρότερο όριο στην εφαρμογή τους και είναι πιο ισχυρές στις ταχείες κινήσεις. Κατά τις τελευταίες δεκαετίες, όλο και περισσότεροι ερευνητές εστιάζουν στις προσεγγίσεις από την βάση προς τα επάνω [55-60]. Μεταξύ αυτών των μεθόδων, οι μέθοδοι που είναι βασισμένοι στην εικονογραφική δομή είναι οι πιο επιτυχημένες τεχνικές για την εκτίμηση της στάσης με την προσέγγιση από την βάση προς τα επάνω.

2.3.1 Εικονογραφική δομή

Η εικογραφική δομή είναι η μέθοδος μοντελοποίησης ενός παρατηρούμενου αντικειμένου από μια συλλογή τμημάτων τοποθετημένων σε παραμορφώσιμη ρύθμιση παραμέτρων. Τα μοντέλα εμφάνισης χρησιμοποιούνται για να μοντελοποιήσουν κάθε μέρος ξεχωριστά και η παραμορφώσιμη διαμόρφωση αντιπροσωπεύεται από τις ελατήριες συνδέσεις μεταξύ των τμημάτων. Το πρόβλημα της αντιστοίχισης μια εικονογραφικής δομής με μια εικόνα είναι στην εύρεση της καλύτερης τοποθέτησης των τμημάτων σε μια παρατηρούμενη εικόνα, όπου η ποιότητα μια τοποθέτησης εξαρτάται τόσο από το πόσο καλά κάθε μέρος ταιριάζει με την εικόνα και από το πόσο καλά οι τοποθετήσεις συμφωνούν με την παραμορφώσιμη διαμόρφωση.



Εικόνα 6: Απλοποίηση της αρχικής σχηματικής αναπαράστασης σε μια δομή δέντρου, Πηγή: researchgate.com

Το 1973 οι εικονογραφικές δομές (PS) εισάγονται από το έργο του Fischler και του Elschlager [59]. Προτείνουν την σχηματική αναπαράσταση του μοντέλου του προσώπου στο πλαίσιο των εικονογραφικών δομών. Αυτή η αναπαράσταση που απλοποιεί το μεταφραστικό πρόβλημα ότι τα στοιχεία (κομμάτια εικόνων, τοπικοί πίνακες αξιολόγησης κ.λ.π) και οι σχεσιακές μορφές (ελατήρια) είναι δισδιάστατες και όχι μονοδιάστατες. Τα μέρη του σώματος ενός ανθρώπου διαμορφώνεται ως υπό όρους τυχαίο πεδίο (CRP). Όπως φαίνεται στην εικόνα 5, τα μέρη ενός προσώπου περιγράφονται σε αυτή την σχηματική αναπαράσταση, συμπεριλαμβανομένων (μαλλιά, δεξιά άκρη, αριστερή άκρη, μύτη, στόμα και δύο μάτια). Τα μέρη του προσώπου συνδέονται με ΄΄ελατήρια΄΄. Αυτά τα ΄΄ελατήρια΄΄ που συνδέουν τα άκαμπτα μέρη του προσώπου χρησιμεύουν τόσο για να περιορίσουν την σχετική κίνηση και να μετρήσουν το ΄΄κόστος΄΄ της κίνησης από το πόσο είναι τεντωμένο. Αυτές οι σαν ελατήρια συνδέσεις μεταξύ ζευγαριών μερών του προσώπου αντιπροσωπεύουν την παραμορφώσιμη διαμόρφωση του προσώπου. Τα γενικά μοντέλα εμφάνισης χρησιμοποιούνται για τα μέρη του προσώπου. Επιπλέον, αναπτύσσεται μια δυναμική προσέγγιση προγραμματισμού που εκμεταλλεύεται την αποσύνθεση για να μειώσει δραστικά τις υπολογιστές απαιτήσεις.

Ένας φυσικός τρόπος για να εκφράσουν ένα τέτοιο μοντέλο PS είναι από την άποψη ενός μη κατευθυνόμενη G=(V,E), όπου V=u_1, … ,u_n είναι ένα σύνολο κόμβων n τμημάτων και υπάρχει μια άκρη (v_1,v_j) ε Ε για κάθε ζεύγος που συνδέεται με τα μέρη v_i και v_j. Μια περίπτωση του παρατηρούμενου θέματος δίνεται από μια διαμόρφωση L = (l_1, … , l_n). Κάθε l_i δηλώνει την θέση (τοποθεσία) του μέρους v_i στην παρατηρούμενη εικόνα. Η θέση του κάθε μέρους είναι δυνατόν να καθορίσει την θέση του ή πιο σύνθετες παραμέτρους. Στην εργασία [59], το πρόβλημα της αντιστοίχισης ενός μοντέλου PS με μια παρατηρούμενη εικόνα ορίζεται ως η ελαχιστοποίηση μια συνάρτησης ενέργειας (κόστους). Για κάθε μέρος v_i, μια αντίστοιχη λειτουργία κόστους m_i(I , l_i) μετρά το βαθμό της αναντιστοιχίας, όταν αυτό το μέρος τοποθετείται στην θέση l_i στην εικόνα I. Για κάθε ζευγάρι συνδεδεμένων μερών (v_i , v_j), υπάρχει μια συνάρτηση κόστους d_{ij}(l_i,l_j) πού μετρά το βαθμό παραμόρφωσης του μοντέλου όταν το μέρος v_i βρίσκεται στο l_i και το μέρος v_j βρίσκεται στο l_j στην εικόνα. Ο στόχος είναι να βρεθεί η καλύτερη αντιστοιχισμένη διαμόρφωση, όπως μετράτε από την λειτουργία κόστους παραμόρφωσης d_{ij}(l_i,l_j). Αυτή η καλύτερη αντιστοιχία μπορεί να εκφραστεί ως:

$$L^* = arg\min_L \Big(\sum_{(v_i,v_j)\epsilon E} d_{ij}\,(l_i, l_j) + \sum_{v_i \in V} m_i(I, l_i) \quad (2.1)$$

Αυτό είναι ένα πρόβλημα ελαχιστοποίησης που είναι αρκετά γενικό και εμφανίζεται σε πολλούς τομείς της όρασης των υπολογιστών. Γενικά το κόστος παραμόρφωσης είναι μίνι μια λειτουργία της σχετικής θέσης μεταξύ δύο συνδεδεμένων μερών.

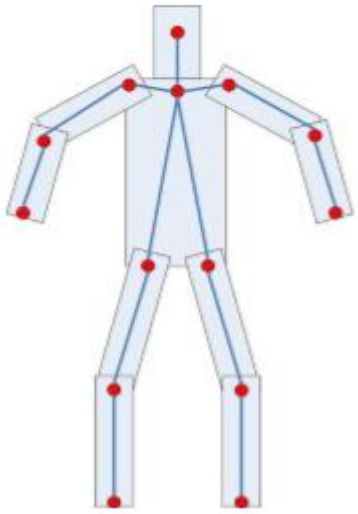
Στο έργο των Felzenszwable και Huttenlocher [57,60], προτείνεται ένας αλγόριθμος για το πρόβλημα της ελαχιστοποίησης στης ενέργειας σε μια εικονογραφική δομή. Επεκτείνουν επίσης αυτό το μοντέλο που είναι βασισμένο στα μέρη (τμήματα) μέρος σε πολλά αντικείμενα, συμπεριλαμβανομένων προσώπων, ανθρώπων και ζώων. Οι συνδέσεις μεταξύ των μερών συνήθως θεωρούνται ότι σχηματίζουν μια δομή δέντρων , η οποία επιτρέπει αποτελεσματικά συμπεράσματα και τον χρόνο της δοκιμής. Όπως

φαίνεται στην εικόνα 6, μια δομή δέντρου χρησιμοποιείται για να αντιπροσωπεύσει το ανθρώπινο αντικείμενο. Το πρόβλημα είναι να βρούμε την καλύτερη θέση l_2 για κάθε v_2 που αντιστοιχεί σε v_1. Αυτό μπορεί να λυθεί με την αφαίρεση του v_2, και επαναλαμβάνοντας με μικρότερο δέντρο, μέχρι μόνο ένα μέρος. Επιπλέον, αποδεικνύουν ότι ο περιορισμός σε μια δομή δέντρου επιτρέπει την χρήση τυποποιημένων τεχνικών δυναμικού προγραμματισμού και ο περιορισμός στην μορφή του κάθε συνδεδεμένου ζεύγους μερών επιτρέπει τη χρήση των μετασχηματισμών απόστασης.

2.3.2 Εικονογραφικές μέθοδοι βασισμένες στην δομή για την εκτίμηση στάσης υπάρχουσες εικόνες

Το εικονογραφικό μοντέλο δομής για ένα παρατηρούμενο αντικείμενο δίνεται από μια συλλογή τμημάτων με συνδέσεις μεταξύ ορισμένων ζευγών αντικειμένων(μερών). Πιο συγκεκριμένα, για τα αρθρωτά ανθρώπινα αντικείμενα, τα μέρη μπορούν να χωριστούν στον κορμό, τα χέρια, το κεφάλι και τα πόδια του ανθρώπου. Στο μοντέλο PS, ο απαιτούμενος αριθμός των μερών του ανθρώπινου σώματος εξαρτάται από την εφαρμογή και την απαιτούμενη ακρίβεια. Για παράδειγμα, ένα αρθρωτό μοντέλο ανθρώπινου σώματος με 14 μέρη είναι σε θέση να παρέχει πιο ακριβή αποτελέσματα σε σχέση με ένα μοντέλο με 6 μέρη του σώματος. Το εικονογραφικό μοντέλο βασισμένο στην ανθρώπινη στάση απεικονίζεται στην εικόνα 7.

Ακολουθώντας τις εργασίες [57,59], πολλοί ερευνητές εστιάζουν στα μοντέλα που είναι βασισμένα στην εικογραφική δομή για την εκτίμηση της στάσης του σώματος υπάρχουσες εικόνες. Στην εργασία [61], η εικογραφικές δομές επεκτείνονται με συσχετίσεις μεταξύ των μερών του ανθρώπινου σώματος σε μια εικόνα. Για τα κινούμενα αντικείμενα, χρησιμοποιούνται συσχετίσεις μεταξύ του άνω βραχίονα και του ποδιού για πιο ισχυρή εκτίμηση των ανθρώπινων στάσεων. Ο Ronfard και οι υπόλοιποι [62] χρησιμοποιούν την ιδέα των εικονογραφικών δομών αλλά αντικατέστησαν τους απλούς ανιχνευτές αντικειμένων (μερών) από αποκλειστικούς ανιχνευτές που το μοντέλο μάθησης της εμφάνισης για κάθε μέρος χρησιμοποιεί τις μηχανές διανυσμάτων υποστήριξης (SVM).



Εικόνα 7: Το μοντέλο ανθρώπινου σώματος βασισμένο στην εικονογραφική δομή, Πηγή: semanticscholar.org

Ο ανιχνευτής Dalal-Triggs [63] χρησιμοποιεί ένα μόνο φίλτρο στο ιστόγραμμα των προσανατολισμένων κλίσεων (HOG) προσφέροντας στην αναπαράσταση της ανίχνευσης του ανθρώπου. Αυτός ο ανιχνευτής χρησιμοποιεί μια προσέγγιση συρόμενου παραθύρου, όπου ένα φίλτρο εφαρμόζεται σε όλες τις θέσεις και κλίμακες σε μια παρατηρούμενη εικόνα. Μετά την εισαγωγή της περιγραφής HOG, πολλοί ερευνητές χρησιμοποιούν HOG για να οικοδομήσουν τα μοντέλα τους εμφάνισης για την ανίχνευση του ανθρώπου και την εκτίμηση της στάσης του.

Το μοντέλο παραμορφώσιμων μερών [64,65] είναι μια μέθοδος βασισμένη στο εικονογραφικό πλαίσιο δομών. Αρχικά προτάθηκε για την ανίχνευση αντικειμένων. Μετά από πολλή εργασία οι μελετητές επέκτειναν το μοντέλο παραμορφώσιμων μερών και για την εκτίμηση της ανθρώπινης στάσης [66,72,67-71]. Όπως περιγράφεται στις εργασίες [64,65], ένα μοντέλο παραμορφώσιμου μέρους ορίζεται από ένα φίλτρο ΄΄ρίζας΄΄ συν ένα σύνολο φίλτρων μερών και σχετικών μοντέλων παραμόρφωσης. Η βαθμολογία αυτού του μοντέλου παραμορφώσιμων μερών (τμημάτων) σε μια συγκεκριμένη θέση και κλίμακα μέσα σε μια εικόνα είναι ίση με την βαθμολογία του φίλτρου ρίζα στην δεδομένη θέση συν το άθροισμα πάνω από τις μέγιστες βαθμολογίες των φίλτρων μερών (τμημάτων) στην θέση του και μείον ένα κόστος παραμόρφωσης μεταξύ κάθε ζεύγους τμημάτων. Οι βαθμολογίες ρίζας και οι βαθμολογίες φίλτρου τμημάτων καθορίζονται από το προϊόν κουκίδας μεταξύ ενός φίλτρου και ένα υποπαράθυρο μιας πυραμίδας προσφοράς HOG που υπολογίζεται από μια εικόνα εισόδου. Η πυραμίδα προσφοράς χρησιμοποιεί δυνατότητες υψηλότερης ανάλυσης για την επίτευξη υψηλής απόδοσης αναγνώρισης. Ακολουθώντας την εξίσωσή (2.1), το μοντέλο παραμορφώσιμου μέρους μπορεί να περιγράφει ως εξής:

$$score(l_o,\ldots,l_n) = \sum_{i=0}^{n} a_i \times \varphi(I,l_i) - \sum_{ij\epsilon E} Y_{ij} \times \psi(l_i,l_j) + b \quad (2.2)$$

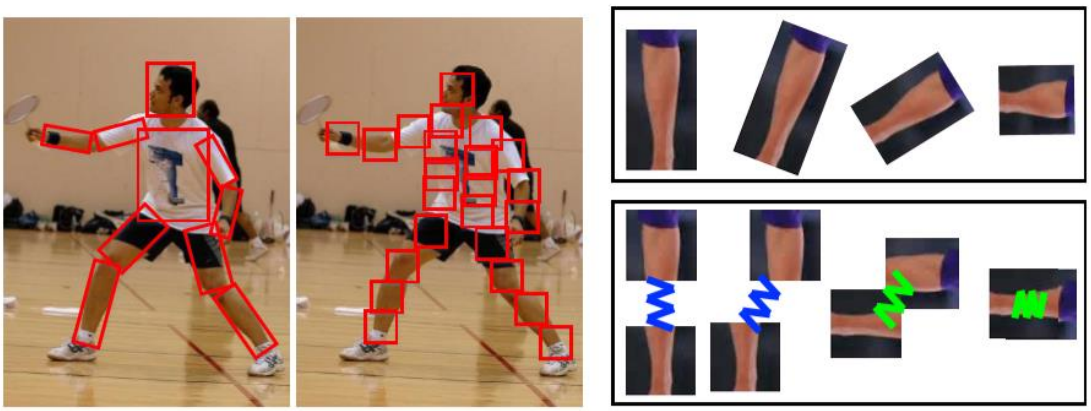
όπου $\varphi(I,l_i)$ είναι ένα διάνυσμα χαρακτηριστικό που εξάγεται από την θέση l_i για το μέρος i στην εικόνα I. $\psi(l_i,l_j) = [x_i - x_j, (x_i-x_j)^2, y_i-y_j, (y_i-y_j)^2]^T$ είναι οι σχετικές θέσεις μεταξύ του μέρους I και του μέρους j. E είναι ένα σύνολο συνδέσεων μεταξύ δύο

διαφορετικών τμημάτων. $A_ι$ και Y_{ij} είναι διανύσματα του μοντέλου των παραμέτρων. Η βαθμολογία μπορεί να εκφραστεί με όρους προϊόντος κουκίδας, β×Φ(I,L), ανάμεσα σε ένα μοντέλο παραμέτρων β και ενός διανύσματος προσφοράς:

$$\beta = (\alpha_0, \ldots, \alpha_n, \ldots, Y_{ij}, \ldots, b \ (2.3)$$

$$\Phi(I, L) = (\varphi(I, l_0), \ldots, \varphi(I, l_n), \ldots, -\psi(l_i, l_j), \ldots, 1) \ (2.4)$$

Αυτό απεικονίζει μια σύνδεση μεταξύ του μοντέλου παραμορφώσιμου μέρους και των γραμμικών ταξινομητών. Οι παράμετροι του μοντέλου εκπαιδεύονται με το διάνυσμα υποστήριξης μηχανών (SVM). Οι καλύτερες δυνατές θέσεις βελτιστοποιούνται με την μεγιστοποιώντας βαθμολογία της συνάρτηση αποτελέσματος (l_0, … ,l_n). Επιπλέον ορίζεται ένα μοντέλο μείγματος για την βελτίωση της απόδοσης. Ο αριθμός μείγματος $m_i \in \{1, \ldots, M\}$ ορίζεται για το μέρος i. Όπως και στην περίπτωση ενός μόνο μοντέλου, η συνάρτηση αποτελέσματος ενός μοντέλου μείγματος μπορεί επίσης να εκφραστεί με ένα προϊόν κουκίδας μεταξύ ενός διανύσματος των παραμέτρων του μοντέλου β και του διανύσματος Φ(I,L).



(α) (β)

Εικόνα 8: (α) Αριστερά είναι η αναγνώριση της στάσης με το κλασσικό μοντέλο εικονογραφικής δομής και δεξιά είναι η εκτίμηση με το μοντέλο εύκαμπτου μείγματος μερών (β) Στην κορυφή είναι ένα μόνο μέρος που έχει διαφορετικό προσανατολισμό και κλίμακα στο κλασσικό μοντέλο. Στο κάτω κομμάτι είναι ένα μικρό τμήμα (μέρος) με την μετάφραση μεγάλων τμημάτων που συνδέονται με ένα ελατήριο, Πηγή: researchgate.net

Ενώ το αντικείμενο ορίζεται με συστατικά μείγματος, το διάνυσμα των παραμέτρων του μοντέλου μείγματος β είναι η συνένωση των διανυσμάτων των παραμέτρων του μοντέλου για κάθε στοιχείο. Το διάνυσμα Φ(I,L) είναι αραιό, με μη μηδενικές εισόδους που ορίζονται από το Φ(I,L΄).

$$\beta = (\beta_1, \ldots, \beta_m)(2.5)$$

$$\Phi(I, L) = (0, \ldots, 0, \Phi(I, L'), 0, \ldots, 0)(2.6)$$

Τα τελευταία χρόνια, το μοντέλο εύκαμπτων μειγμάτων μερών (FMP) [66] είναι μια από τις πιο επιτυχημένες μεθόδους για την ανίχνευση της αρθρωτής ανθρώπινης στάσης σε στατικές εικόνες. Αυτό το μοντέλο βασίζεται στο μοντέλο παραμορφώσιμων μερών. Σε αντίθεση με τα παραδοσιακά μοντέλα, χρησιμοποιούν ένατων εικονογραφικών δομών με μικρά, μη προσανατολισμένα μέρη. Τα μέρη τους αντιστοιχήσαν τα μεσαία και τελικά σημεία κάθε άκρου. Τα μέρη μοντελοποιήθηκαν ως μείγμα ώστε να αποτυπώσουν τους προσανατολισμούς των άκρων. Όπως φαίνεται και στην εικόνα 8, τα μικρότερα υποδείγματα χρησιμοποιούνται στο μοντέλο FMP που είναι πιο ευέλικτο για να αντιπροσωπεύει κάθε μέρος του σώματος. Όλα τα μέρη του σώματος σχετίζονται σε μια δομή δέντρου, και μπορεί να βελτιστοποιηθεί αποτελεσματικά με δυναμικό προγραμματισμό. Το αποτελεσματικό μοντέλο με βάση τα διακριτά μέρη [64] χρησιμοποιείται για την εκμάθηση όλων των παραμέτρων, συμπεριλαμβανομένης της τοπικής εμφάνισης, των σχέσεων συν-εμφάνισης και των χωρικών σχέσεων (με βάση στην δομημένη SVM).

Η εργασία του Tian και των υπολοίπων [73], γίνεται επέκταση της FMP στην χωρική ιεραρχία του μοντέλου μείγματος για την εκτίμηση της ανθρώπινης στάσης. Αυτό το μοντέλο χρησιμοποιεί έναν εκθετικό αριθμό στάσεων με μια συμπαγή αναπαράσταση μείγματος σε κάθε μέρος. Δοκιμάζουν τον τύπο στάσης από το μοντέλο μαθήσεως, απασχολώντας ένα μοντέλο λανθάνοντος δέντρου, ως κόμβους της ρίζας για να χειριστεί την γεωμετρική παραμόρφωση. Στην εργασία των Park και Ramanan [74], προτείνεται ένας αλγόριθμος N-best για την δημιουργία ενός συνόλου από N υποψήφιους υψηλής-βαθμολογίας. Η ενιαία-καλύτερη στάση υπολογίζεται από τους από τους N-καλύτερους υποψήφιους από το δυναμικό προγραμματισμό. Αποδεικνύουν ότι οι τοπικές ασάφειες μπορούν να τελειοποιηθούν με την προτεινόμενη προσέγγιση τους. Οι Wang και η Li [69] χρησιμοποιήσαν τα μέρη του σώματος μεσαίου επιπέδου στο λανθάνον μοντέλο δέντρων τους και να προτείνει έναν αλγόριθμό για την αυτόματη εκμάθηση του δέντρου για να συνδεθούν όλα τα μέρη. Το μοντέλο τους περιέχει 14 μονά μέρη και 10 συνδυασμένα. Το συνδυασμένο μοντέλο χρησιμοποιείται για να έχει πιο αποτελεσματικό μοντέλο εμφάνισης. Αποδεικνύουν ότι το μοντέλο τους έχει καλές επιδόσεις στην εκτίμηση της ανθρώπινης στάσης. Έχουν προταθεί χαρακτηριστικά βάσει περιγράμματος για την εκτίμηση της αρθρωτής στάσης [75], σε μια προσπάθεια να επιλυθούν ορισμένες από τις καταστάσεις σύγχυσης στο παρασκήνιο.

Ο περιορισμός των ιεραρχικών Poselet μοντέλων δημιουργεί ανιχνευτές μη άκαμπτων εσφαλμένων άκρων δεδομένου ότι είναι μεταβλητές στην όψη. Ένα από τα κίνητρα για την χρήση ενός μικρού άκαμπτου μοντέλου με βάση τα μέρη είναι ότι επιτρέπει την ομαλοποίηση πάνω από την διαμόρφωση για κάθε μέρος. Λόγω των αλλαγών φωτισμού, των ενδυμάτων και του σχήματος σωμάτων, ένα πρότυπο εμφάνισης πρέπει να καταγράψει την παραλλαγή την μεταβολή στο χώρο της πιθανής εμφάνισης. Ωστόσο, η προσέγγιση αυτή οδηγεί σε μοντέλα εμφάνισης που αντιπροσωπεύουν κατά προσέγγιση παράλληλες ακμές ή κωνικούς κυλίνδρους που οδηγούν σε εσφαλμένες θετικές ανιχνεύσεις. Εάν ήταν διαθέσιμα αρκετά δεδομένα εικόνας, θα μπορούσε κανείς να ελπίζει να δημιουργήσει μοντέλα εμφάνισης για εκτίμηση της στάσης με τουλάχιστον δύο συνδεδεμένα μέρη. Αυτή η συνδεδεμένη εμφάνιση με βάση τα μέρη είναι πολύ πιο εμφανής από την εμφάνιση ενός μόνο μέρους. Έτσι το μοντέλο εμφάνισης έχει περισσότερο πλαίσιο. Αυτό είναι το κίνητρο της προσέγγισης Poselet στο έργο των Bourdev και Malik [76].

Ένα Poselet είναι ένας ανιχνευτής εκπαιδευμένος για μια συγκεκριμένη διαμόρφωση και εμφάνιση των άκαμπτων τμημάτων ή μεγάλων τμημάτων των ανθρώπινων σωμάτων (π.χ. κορμός + βραχίονας). Στην εργασία του Wang και των υπολοίπων [67] προτείνεται το ιεραρχικό Poselet μοντέλο για την εκτίμηση της στάσης από ένα loopy αλγόριθμο μετάδοσης της πεποίθησης. Γεφυρώνουν το χάσμα μεταξύ των μεθόδων με βάση τα μέρη και των μεθόδων με βάση παραδείγματα. Οι μέθοδοι που βασίζονται σε παραδείγματα αναζητούν εικόνες με παρόμοιες διαμορφώσεις ολόκληρου του σώματος. Ο περιορισμός των προσεγγίσεων βασισμένες στα παραδείγματα είναι ότι δεν μπορούν να χειριστούν μια δοκιμαστική εικόνα στην οποία τα πόδια είναι παρόμοια με μια εικόνα κατάρτισης και οι ώμοι είναι παρόμοιοι με άλλη εικόνα κατάρτισης. Η προσέγγιση μπορεί να χρησιμοποιηθεί για την 3D εκτίμηση της ανθρώπινης στάσης με μεμονωμένες εικόνες [76], την ανίχνευση του ανθρώπου [77] και την ταξινόμηση των χαρακτηριστικών [78]. Στο έργο του Pishchulin και των υπολοίπων [79] όρισαν ένα μοντέλο δένδρου στο οποίο οι μοναδιαίοι και οι προς ζεύγος όροι εξαρτώνται από τα Poselets στοιχεία. Αυτό το μοντέλο υπό όρους ορίζεται από όλα τα μέρη του σώματος που είναι συνδεδεμένα a-priori, αλλά το οποίο γίνεται ένα εισελκόμενο μοντέλο PS. Τα Poselets χρησιμεύουν ως μια αναπαράσταση μεσαίου επιπέδου που κωδικοποιεί από κοινού την άρθρωση πολλών ανθρώπινων μερών του σώματος σε μια παρατηρούμενη εικόνα.

Παρόμοια, οι μέθοδοι στην εργασία [69] χρησιμοποιούν τα μέρη του σώματος μεσαίου επιπέδου στο λανθάνον μοντέλο δέντρων τους και να προτείνει ένα αλγόριθμο για την αυτόματη εκμάθηση του δέντρου για να συνδέσουν όλα τα μέρη. Συνδυάζουν Poselets και μικρά μέρη του σώματος μαζί στο μοντέλο δομής δέντρου. Εδώ τα Poselets χρησιμοποιούνται για να χειριστούν την μεγάλη διακύμανση στην εμφάνιση. Αποδεικνύουν ότι το προτεινόμενο μοντέλο τους έχει καλές επιδόσεις σε διάφορα σύνολα δεδομένων. Μια πολύ πρόσφατη εργασία [80] βελτιώνει το μοντέλο με περισσότερα επίπεδα αντικειμένων και επιτυγχάνει καλές επιδόσεις. Αντί να εκτελούν συμπεράσματα σε ένα μαθημένο γραφικό μοντέλο, χτίζουν μια ιεραρχική μηχανή συμπερασμάτων για την αρθρωτή εκτίμηση της ανθρώπινης στάσης. Αυτή η μέθοδος είναι ένας αλγόριθμος διαδοχικής πρόβλεψης που μιμείται την μηχανική της διαβίβασης μηνυμάτων για την πρόβλεψη μιας εμπιστοσύνης για κάθε μεταβλητή. Ωστόσο, θα πρέπει να σημειωθεί ότι οι μέθοδοι που βασίζονται στο Poselet έχουν τον περιορισμό του επαρκούς αριθμού των δεδομένων εκπαίδευσης. Συνεπώς, είναι απίθανο η προσέγγιση Poselet να είναι αποτελεσματική χωρίς ένα σημαντικό αριθμό από δεδομένα εκπαίδευσης.

Πολλά πρόσφατα έργα εισήγαγαν επίσης τμήματα υψηλότερου επιπέδου σε ιεραρχικά μοντέλα για την εκτίμηση της στάσης. Το κίνητρο αυτής της προσέγγισης είναι να συνδυαστούν τα οφέλη τόσο των εφαρμογών με βάση τα μέρη όσο και της προσέγγισης Poselet πολλαπλών τμημάτων. Οι περισσότερες ιεραρχικές μέθοδοι περιλαμβάνουν ένα ολόκληρο ανιχνευτή ατόμων στην ρίζα τους και τα μεμονωμένα μέρη στα φύλλα τους. Οι πρώτες εργασίες σε ιεραρχικά μοντέλα για την 2D ανάλυση του ανθρώπου είναι το γράφημα AND/OR [81]. Καθορίζουν τα μοντέλα εμφάνισης σε υπό-μέρη των τμημάτων του σώματος και να θέσει όλα τα μικρά κομμάτια μαζί στο ιεραρχικό μοντέλο. Στην εργασία του Wang και των υπολοίπων [67] προτείνουν μια προσέγγιση ιεραρχικών Poselets με βάση το μοντέλο εικογραφικής δομής. Τέτοιες προσεγγίσεις οδηγούν σε αυξημένη απόδοση κατά την εκτίμηση των κάτω άκρων, αλλά δεν αντιμετωπίζουν τις στάσεις που περιλαμβάνονται στα δεδομένα κατάρτισης. Στην εργασία [76] χρησιμοποιούν μέρη μεγάλης κλίμακας που μπορούν να ενσωματωθούν

σε μια ιεραρχική, χονδροειδή έως λεπτή αναπαράσταση. Το μοντέλο τους επιτυγχάνει μια ισορροπία μεταξύ της πολυπλοκότητας του μοντέλου και του πλούτου του μοντέλου, μοιράζοντας μοντέλα εμφάνισης τύπων τμημάτων και αποσυνθέτοντας πολύπλοκες στάσεις σε σχέσεις ζεύγους. Ο Duan και οι υπόλοιποι [82] προτείνουν ένα ιεραρχικό σύνθετο μοντέλο μέσω μιας διαδικασίας μεγιστοποίησης για κοινή μάθηση. Στην εργασία [73] χρησιμοποιείται ένα μοντέλο λανθάνοντος δέντρου ως κόμβους ρίζας για να χειριστεί την γεωμετρική παραμόρφωση. Προτείνουν ένα ιεραρχικό χωρικό μοντέλο που μπορεί να καταγράψει ένα εκθετικό αριθμό στάσεων με την δειγματοληψία ενός μίγματος στάσεων από το εκπαιδευόμενο μοντέλο.

2.3.3 Μέθοδοι βασισμένες στην εικογραφική δομή για την εκτίμηση στάσης σε βίντεο

Το μοντέλο εικονογραφικής δομής χρησιμοποιείται ευρέως για την εκτίμηση της ανθρώπινης στάσης σε βίντεο. Σε σύγκριση με την εκτίμηση της στάσης σε εικόνες, το χρονικό στοιχείο των βίντεο παρέχει ένα πρόσθετο ερέθισμα για την εκτίμηση της στάσης, καθώς υπάρχουν ισχυρές εξαρτήσεις των θέσεων των μερών του ανθρώπινου σώματος μεταξύ χρονικά κοντά καρέ βίντεο.

Το έργο strike-a-pose [83] αναζητά τουλάχιστον ένα καρέ στην ακολουθία βίντεο για ένα προκαθορισμένο χαρακτηριστικό στάσης, που είναι πιο εύκολο να ανιχνευθεί από μια γενική στάση. Με βάση αυτή την ιδέα, χτίζουν ένα ατομικό-συγκεκριμένο μοντέλο εμφάνισης για την εκτίμηση της ανθρώπινης στάσης. Ο Eicher και ο Ferrari [58] παρουσιάζουν καλύτερα μοντέλα εμφάνισης για την εικονογραφική δομή. Δείχνουν ότι ορισμένα μέρη έχουν μάλλον σταθερή θέση στα πλαίσια αναφοράς και τα μοντέλα εμφάνισης των διαφορετικών μερών είναι στατιστικά σχετικά. Για παράδειγμα, τα χαμηλά μέρη του χεριού ενός ατόμου είναι χρωματισμένα είτε σαν τον κορμό είτε σαν το πρόσωπο. Μόνο σπάνια έχουν ένα εντελώς διαφορετικό χρώμα. Έτσι, η εμφάνιση ορισμένων τμημάτων μπορεί να προβλεφθεί από την εμφάνιση άλλων μερών. Μαθαίνουν μια θέση προγενέστερη των αντικειμένων με σεβασμό στην αναφορά του πλαισίου και ένας μηχανισμός μεταφοράς εμφάνισης των διαφορετικών αντικειμένων από τα δεδομένα εκπαίδευσης. Αυτές οι υποδείξεις αξιοποιούνται για να δημιουργήσουν μοντέλα εμφάνισης για τα μέρη του σώματος.

Στην εργασία [84], προτείνεται μια μέθοδος για την εκτίμηση στάσης του άνω μέρους του σώματος. Η προσέγγιση αυτή χρησιμοποιείται για την εκτίμηση της στάσης του άνω μέρους του σώματος σε ανεξέλεγκτες εικόνες, χωρίς προηγούμενη γνώση του φόντου, των ενδυμάτων ή την θέση και την κλίμακα των τμημάτων του ανθρώπινου σώματος σε μια εικόνα ή βίντεο. Ένας γενικός ανιχνευτής του άνω μέρους του σώματος χρησιμοποιείται για να περιορίσει την θέση και την εμφάνιση των τμημάτων του ανθρώπινου σώματος σε μια εικόνα. Αυτός ο ανιχνευτής του άνω μέρους του σώματος εκπαιδεύεται με την χρήση ενός συρόμενου μηχανισμού παραθύρων που ακολουθείται από τη μη μέγιστη καταστολή. Ο Sapp και ο Taskar [70] προτείνουν ένα πολυτροπικό σε γενικό (global) επίπεδο μοντέλο και χρησιμοποιούν 32 καταστάσεις στάσεων για να μοντελοποιήσουν το πλευρικό μέρος του σώματος. Αυτό το μοντέλο είναι εκπαιδευμένο τόσο σε μεγάλο πεδίο όσο και σε τοπικά ερεθίσματα σε επίπεδο τμημάτων. Χρησιμοποιούν ένα δομημένο επικαλυπτόμενο μοντέλο επιλογής βημάτων το οποίο ελέγχει την αντιστάθμιση μεταξύ της ταχύτητας και της ακρίβειας.

Η οπτική ροή είναι μια άλλη μέθοδος που χρησιμοποιείται ως υπόδειξη είτε για ανίχνευση μέρους του σώματος είτε για την διάδοση στάσης από καρέ σε καρέ. Στην εργασία του Sapp και των υπολοίπων [85] εισάγεται η οπτική ροή ως χαρακτηριστικό για να εντοπιστούν τα περιγράμματα στο προσκήνιο. Κάθε υπομοντέλο στην καθορισμένη δομή δέντρου τους ακολουθεί μια ενιαία ένωση μέσω του χρόνου. Τα περιγράμματα στο προσκήνιο ενσωματώνονται καλά με την μέθοδο της εκτίμησης στάσης τους. Στην εργασία του Φραγκιαδάκη [86] έχουμε εκμετάλλευση της οπτικής ροής για την κατάτμηση της στάσης. Συνδυάζουν ένα τραχύ κομμάτι-συνετό(piece-wise) affine με αξιόπιστες αντιστοιχίες pixel από την οπτική ροή. Μια λεπτή οπτική ροή χρησιμοποιείται για να ακολουθήσει τους αγκώνες και να κινήσει τα άκρα της αλληλουχιών άρθρωσης. Αυτή η ΄΄αρθρωτή΄΄ ροή μπορεί να ακολουθήσει με ακρίβεια τα αρθρωτά αντικείμενα (ανθρώπινο σώμα) με μεγάλες περιστροφές ή μικτές μετατοπίσεις των άκαμπτων αντικειμένων. Στην εργασία του Cherian και των υπολοίπων [87], παρουσιάζεται μια μέθοδος για την εκτίμηση της αρθρωτής ανθρώπινης στάσης στα βίντεο, η οποία βασίζεται επίσης στην οπτική ροή. Η οπτική ροή χρησιμοποιείται για την επέκταση του ευέλικτου μοντέλου μείγματος αντικειμένων [66] σε μια μόνο εικόνα. Πρώτα ένα σύνολο υποψηφίων στάσεων παράγεται σε κάθε καρέ με μια οπτική ροή βασισμένη στην μέθοδο για την εκτίμηση της ανθρώπινης στάσης του σώματος. Κατόπιν υπολογίζουν τις Κ καλύτερες στάσεις [74], σε κάθε καρέ για να λάβουν ένα ποικίλο σύνολο υποψήφιων στάσεων. Επιπλέον αποσυνθέτουν την Κ καλύτερη υποψήφια ανθρώπινη στάση σε άκρα και τα παρακολουθούν για να δημιουργήσουν ακολουθίες σώματος-αντικειμένων. Τέλος, η πλήρης στάση ανασυντίθεται με την ανάμειξη αυτών των ακολουθιών αντικειμένων.

2.4 Βαθύ συνελικτικό νευρωνικό δίκτυο για την εκτίμηση της στάσης

Κατά την διάρκεια των τελευταίων ετών, οι τεχνικές βαθιάς μάθησης έχουν σημειώσει τεράστια πρόοδο στην όραση των υπολογιστών. Τα βαθιά συνελικτικά νευρωνικά δίκτυα (DCNNs) είναι ένα είδος αυτών των τεχνικών στο πλαίσιο της βαθιάς μάθησης και έχει γίνει η μέθοδος επιλογής σε πολλούς τομείς της όρασης του υπολογιστή. Τα DCCNs έχουν δείξει εξαιρετικές επιδόσεις σε εργασίες οπτικής ταξινόμησης και πιο πρόσφατα στον εντοπισμό αντικειμένων.

2.4.1 Βαθύ συνελικτικό νευρωνικό δίκτυο στην όραση υπολογιστή

Σε πρόσφατες έρευνες όρασης υπολογιστών, το CNN είναι μια πολύ δημοφιλής προσέγγιση βαθιάς μάθησης. Στην εργασία [88], τα πολλαπλά επίπεδα αναπαράστασης έχουν μάθει να διαμορφώνουν πολύπλοκες μη γραμμικές σχέσεις. Το DCNN έχει επιδείξει εξαιρετική απόδοση για την ταξινόμηση εικόνων εργασίας [89-91]. Πιο πρόσφατα οι αρχιτεκτονικές του CNN εφαρμόστηκαν με επιτυχία για τον εντοπισμό και την διάγνωση αντικειμένων [92,93]. Ο Long και οι υπόλοιποι στην εργασία τους [109] παρουσίασαν πλήρως συνελικτικά δίκτυα, τα οποία επιτρέπουν την ανά pixel πρόβλεψη, ‘όπως η σημασιολογική κατάτμηση. Στην εργασία [92], το DetectorNet αντιμετωπίζει το πρόβλημα της ανίχνευσης αντικειμένων και προτείνει μια διαδικασία

συμπεράσματος πολλαπλών οπών για την παραγωγή ανιχνεύσεων αντικειμένων υψηλής ανάλυσης. Το OverFeat [93] παράγει πυκνά, πολλαπλής κλίμακας χαρακτηριστικά CNN για την ταξινόμηση αντικειμένων, τον εντοπισμό και την εκτίμηση από μια εικόνα εξετάζοντας κάθε συρόμενο παράθυρο.

Ο Gireshick και οι υπόλοιποι [94] πρότειναν την μέθοδο R-CNN με την εφαρμογή των υψηλής χωρητικότητας συνελικτικών νευρωνικών δικτύων στις εφαρμογές των περιοχών από την βάση προς τα επάνω [95] για τον εντοπισμό και την κατάτμηση αντικειμένων. Η συγκεκριμένη μέθοδος ξεπερνάει σε επιδόσεις την μέθοδο OverFeat και βελτιώνει την απόδοση περισσότερο από 30% σε σχέση με τις τελευταίες τεχνολογίες σε PASCAL VOC 2012. Στην εργασία [96] εγκρίνεται το R-CNN [94] για να εντοπίσει διάφορα τμήματα και να επαληθεύσει ότι η χρήση των προτάσεων της περιοχής μπορεί να βοηθήσει στον εντοπισμό μικρότερων τμημάτων. Με βάση αυτό το R-CNN έχει αποδειχθεί αποτελεσματικό για αναγνώριση ΄΄λεπτών-κόκκων΄΄. Στην εργασία [96] ο He και οι υπόλοιποι πρότειναν μια ομαδοποίηση χωρικής πυραμίδας συγκέντρωσης σε DCNN για οπτική αναγνώριση. Αυτή η δομή δικτύου μπορεί να δημιουργήσει μια αναπαράσταση σταθερού μήκους, ανεξάρτητα από την κλίμακα των εικόνων εισόδου. Τα αποτελέσματα του πειράματος δείχνουν ότι η προτεινόμενη μέθοδος τους είναι πολύ αποτελεσματική στα καθήκοντα ταξινόμησης και ανίχνευσης. Ωστόσο, η μέθοδος αυτή δεν λαμβάνει υπόψη τις πολύπλοκες σχέσεις ανάμεσα σε διαφορετικά αντικείμενα (μέρη) και δεν εφαρμόζεται στην εκτίμηση της ανθρώπινης θέσης.

2.4.2 Το βαθύ συνελικτικό δίκτυο για την εκτίμηση της στάσης του σώματος

Για την εκτίμηση της στάσης, οι αλγόριθμοι με τις καλύτερες επιδόσεις σήμερα [98-100] βασίζονται στα βαθιά συνελικτικά δίκτυα. Υπήρχαν ορισμένα πρώτα παραδείγματα χρήσης των νευρωνικών δικτύων συνέλιξης για συγκρίσεις ρων στάσεων [101]. Πιο πρόσφατα, στην εργασία του Toshev και των υπολοίπων [99] ανέπτυξαν την μέθοδο DeepPose που είναι ένας χείμαρρος με βάση το CNN κοινών οπισθοδρομήσεων εφαρμόζεται για την καταγραφή του περιβάλλοντος και της αιτίας σχετικά με την ανθρώπινη στάση σε μια ολιστική μόδα. Τα δίκτυα DeepPose χρησιμοποιούν μια πλήρη εικόνα ως είσοδο και διαμορφώνει τις μεθόδους χωρίς κανένα ρητό χαρακτηριστικό αναπαράστασης ή ανιχνευτών τμημάτων. Στην εργασία [102] ο Jain και οι υπόλοιποι, εισάγει μια πολυστρωματική CNN αρχιτεκτονική και συνδυάζει χαρακτηριστικά χαμηλού επιπέδου με ένα υψηλότερου επιπέδου αδύναμο χωρικό μοντέλου για την βελτίωση της απόδοσης. Μετά από το [102], ο Thompson και οι υπόλοιποι [103] προσπάθησαν να συνδυάσουν ένα ανιχνευτή μέρους CNN με ένα χωρικό μοντέλο με βάση τα μέρη σε μια ενοποιημένη εκμάθηση του πλαισίου. Αυτή η μέθοδος μπορεί να αυξήσει σημαντικά την απόδοση της εκτίμησης της στάσης. Στην εργασία [98] ο Chen και ο Alan καθορίζουν ένα γραφικό μοντέλο με εξαρτώμενο από την εικόνα ζεύγος σχέσεων για την εκτίμηση της ανθρώπινης στάσης. Σε αυτό το μοντέλο το CNN χρησιμοποιείται για να μάθει τις υπό όρους πιθανότητες για την παρουσία αντικειμένων και τις χωρικές σχέσεις μεταξύ των μερών. Πρόσφατα, ο Fan και οι υπόλοιποι [104] προτείνουν διπλής πηγής βαθιά συνελικτικά νευρωνικά δίκτυα να ενταχθούν στην εμφάνιση μέρους του σώματος και την ολιστική προβολή του κάθε μέρους του σώματος για περισσότερο ακριβή εκτίμηση της ανθρώπινης στάσης.

Οι χρονικές πληροφορίες στα βίντεο χρησιμοποιήθηκαν αρχικά με DCNN για την αναγνώριση δράσης [105], όπου μια αρχιτεκτονική DCCN δύο ροών ενσωματώνει χωρικά και χρονικά δίκτυα. Εδώ η χρονική πληροφορία είναι οπτική ροή που χρησιμοποιείται ως λειτουργία εισόδου σε αυτό το δίκτυο. Ακολουθώντας αυτήν την εργασία, οι ερευνητές [106,107] διευρύνουν την χρήση των χρονικών πληροφορίων για την εκτίμηση την στάση του άνω μέρους του σώματος ή του συνολικού σώματος σε ένα βίντεο. Η οπτική ροή ή τα RGB χαρακτηριστικά υπολογίζονται από κοντινά καρέ στο δίκτυο και οι κοινές θέσεις εντοπίζονται στο τρέχον καρέ. Πιο πρόσφατα η Pfister και οι υπόλοιποι [108], προτείνει μια μέθοδο για την εκτίμηση της στάσης σε βίντεο που είναι δυνατό να χρησιμοποιήσει την εφαρμογή σε πολλαπλά πλαίσια. Το πλήρη συνελικτικό χωρικό δίκτυο προβλέπει θερμικό χάρτη εμπιστοσύνης για κάθε ένωση σωμάτων σε αυτά τα καρέ. Αποδεικνύουν ότι οι θερμικοί χάρτες των θέσεων από γειτονικά καρέ μπορεί να στρεβλωθεί και να ευθυγραμμιστεί με την χρήση οπτικής ροής από το τρέχον καρέ.

ΚΕΦΑΛΑΙΟ 3

Κυριότερες μέθοδοι εκτίμησης της στάσης του σώματος ενός ατόμου με χρήση της τεχνολογίας βαθιάς μάθησης

3.1 Η μέθοδος deep pose

3.1.1 Η μοντελοποίηση του προβλήματος

Σε αυτή την μέθοδο αξιοποιούνται οι πρόσφατες εξελίξεις στο τομέα της βαθιάς μάθησης και προτείνεται ένας νέος αλγόριθμος που βασίζεται στο βαθύ νευρωνικό δίκτυο (DNN). Τα δίκτυα DNN έχουν επιδείξει εξαιρετικές επιδόσεις σε εργασίες οπτικής ταξινόμησης [110] και πρόσφατα κατά την εντοπισμό αντικειμένων [111,112]. Διαμορφώνεται η εκτίμηση της στάσης του σώματος ως πρόβλημα κοινής παλινδρόμησης και δείχνεται πως προσαρμόζεται το πρόβλημα επιτυχώς στις ρυθμίσεις του δικτύου DNN. Η θέση κάθε άρθρωσης του σώματος οπισθοχωρεί στην χρήση ως είσοδος της πλήρους εικόνας και ενός 7 επιπέδων γενικού συνελεκτικού δικτύου DNN. Υπάρχουν δύο πλεονεκτήματα αυτής της διαμόρφωσης. Πρώτον, το DNN είναι σε θέση να αποτυπώνει το συνολικό πλαίσιο κάθε άρθρωσης του σώματος. Κάθε παλινδρόμηση της άρθρωσης χρησιμοποιεί την πλήρη εικόνα ως σήμα. Δεύτερον η προσέγγιση είναι ουσιαστικά απλούστερη να διαμορφωθεί από τις μεθόδους που βασίζονται σε γραφικά μοντέλα. Δεν χρειάζεται να σχεδιαστούν ρητά τα χαρακτηριστικά γνωρίσματα και οι ανιχνευτές για τα διάφορα μέρη. Αντίθετα δείχνεται ότι με αυτή την μέθοδο ένα γενικό συνελικτικό DNN μπορεί να εκπαιδυτεί για αθτό το πρόβλημα.

Χρησιμοποιείται ο ακόλουθος συμβολισμός. Για να εκφράσουμε μια στάση, κωδικοποιούμε τις θέσεις k όλων των αρθρώσεων του σώματος σε ένα διάνυσμα στάσης που το ορίζουμε ως $y = (\dots, y_i^T, \dots)^T, i \epsilon \{1, \dots, k\}$, όπου το y_i περιέχει τις συντεταγμένες x και y της i^{th} άρθρωσης. Μια εικόνα με ετικέτα σημειώνεται με (x,y), όπου το x αντιπροσωπεύει τα δεδομένα της εικόνας και το y το έδαφος αληθείας του διανύσματος στάσης. Επιπλέον, δεδομένου ότι οι κοινές συντεταγμένες είναι σε απόλυτη συντεταγμένες απόλυτης εικόνας, αποδεικνύεται επωφελής για την κανονικοποίηση αυτών w.r.t a box b που οριοθετεί το ανθρώπινο σώμα ή τα μέρη του. Σε ασήμαντη περίπτωση, το πλαίσιο μπορεί να υποδηλώνει την πλήρη εικόνα. Ένα τέτοιο πλάισιο ορίζεται από το κέντρο του $b_c \epsilon \mathbb{R}^2$ καθώς και το πλάτος b_w και το ύψος

b_h: b = (b_c, b_w, b_h). Τότε η άρθρωση y_i μπορεί να μεταφραστεί από το κέντρο του πλαισίου και κλιμακούμενο από το μέγεθος του πλαισίου το οποίο αναφέρεται ως κανονικοποίηση από το b:

$$N(y_i;\ b) = \begin{pmatrix} 1/b_w & 0 \\ 0 & 1/b_h \end{pmatrix} (y_i\ -\ b_c) \quad (3.1) -$$

Επιπλέον, μπορούμε να εφαρμόσουμε την ίδια κανονικοποίηση στα στοιχεία του διανύσματος στάσης $N(y;\ b) = (\dots, N(y_i; b)^T, \dots)^T$ με αποτέλεσμα ένα κανονικοποιημένο διάνυσμα στάσης. Τέλος, με μια μικρή κατάχρηση σημειογραφίας, χρησιμοποιούμε N(x; b) για να δείξουμε μια περικοπή της εικόνας x από το πλαίσιο οριοθέτησης b, το οποίο κανονικοποιεί εκ των πραγμάτων την εικόνα από το πλαίσιο. Για συντομογραφία δηλώνουμε N(·) την κανονικοποίηση με το b να είναι το πλήρες πλαίσιο της εικόνας.

3.1.1 Η εκτίμηση της στάσης του σώματος ως παλινδρόμηση με βάση το DNN

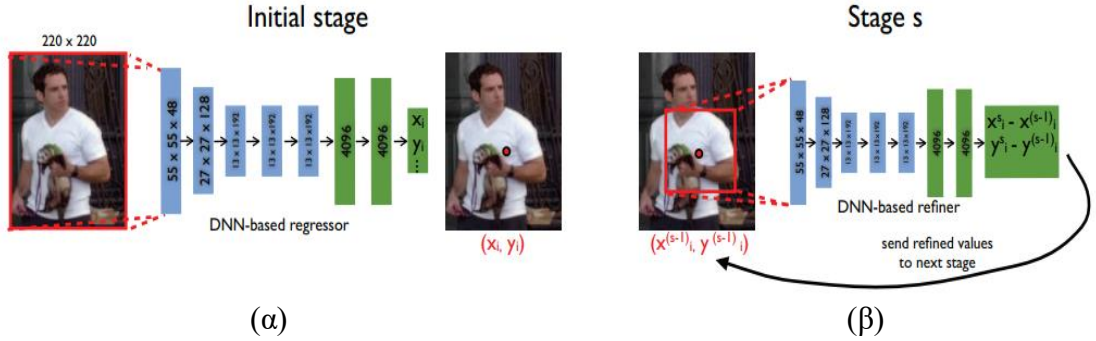
Σε αυτή την μέθοδο, αντιμετωπίζεται το πρόβλημα της εκτίμησης της στάσης του σώματος ως παλινδρόμηση, όπου εκπαιδεύεται και χρησιμοποιείται η συνάρτηση ψ(x;θ) ∈ $\mathbb{R}^{2k}$, η οποία για μια εικόνα x παλινδρομεί σε ένα κανονικοποιημένο διάνυσμα στάσης, όπου θ εκφράζει τις παραμέτρους του μοντέλου. Έτσι, χρησιμοποιώντας τον κανονικοποιημένο μετασχηματισμό από την εξίσωση (3.1) η πρόβλεψη στάσης y^* σε απόλυτες συντεταγμένες εικόνας διαβάζεται:

$$y^* = N^{-1}(\psi(N(x); \theta)\ \ (3.2)$$

Παρά την απλή διατύπωση, η δύναμη και η πολυπλοκότητα της μεθόδου είναι στο ψ, το οποίο βασίζεται σε ένα βαθύ νευρωνικό δίκτυο (DNN). Ένα τέτοιο συνελικτικό δίκτυο αποτελείται από διάφορα επίπεδα, το καθένα από τα οποία είναι ένας γραμμικός μετασχηματισμός ακολουθούμενος από ένα μη γραμμικό. Το πρώτο στρώμα παίρνει ως είσοδο μια εικόνα προκαθορισμένου μεγέθους και έχει μέγεθος ίσο με τρείς φορές τον αριθμό των pixel των καναλιών χρώματος. Το τελευταίο επίπεδο παράγει τις τιμές ΄΄στόχους΄΄ της παλινδρόμησης, στην περίπτωση μας 2k συντεταγμένες αρθρώσεων.

Η αρχιτεκτονική του ψ στην μέθοδο βασίζεται στην εργασία του Krizhensky και των υπολοίπων [110] για την ταξινόμηση εικόνας δεδομένου ότι έχει εμφανίσει εξαιρετικά αποτελέσματα στον εντοπισμό αντικειμένων [111].Με λίγα λόγια, το δίκτυο αποτελείται από 7 επίπεδα (εικόνα 2α). Υποδεικνύεται C ένα συνελεκτικό επίπεδο, LRN ένα επίπεδο κανονικοποίησης τοπικής απόκρισης, P ένα επίπεδο ομαδοποίησης και F ένα πλήρως συνδεδεμένο επίπεδο. Μόνο τα C και F επίπεδα περιέχουν επίκτητες παραμέτρους, ενώ οι υπόλοιποι παραμέτροι είναι ελεύθερες. Και τα δύο επίπεδα C,F αποτελούνται από ένα γραμμικό μετασχηματισμό ακολουθούμενο από ένα μη γραμμικό, ο οποίος στην περίπτωση μας είναι μια ανακαθορισμένη γραμμική μονάδα. Για τα C επίπεδα, το μέγεθος ορίζεται ως πλάτος × ύψος × βάθος, όπου οι δύο πρώτες διαστάσεις έχουν χωρική σημασία, ενώ το βάθος ορίζει τον αριθμό των φίλτρων. Το μέγεθος του φίλτρου για τα δύο πρώτα C επίπεδα είναι 11×11 και 5×5 και για τα υπόλοιπα τρία είναι 3×3. Η ομαδοποίηση εφαρμόζεται μετά από τρία επίπεδα και συμβάλλει στην αύξηση των επιδόσεων παρά την μείωση της ανάλυσης. Η είσοδος στο

δίκτυο είναι μια εικόνα 220×220 , η οποία μέσω του βηματισμού 4 τροφοδοτείται στο δίκτυο. Ο συνολικός αριθμός των παραμέτρων στο παραπάνω μοντέλο είναι 40M. Για περισσότερες λεπτομέρειες μπορείτε να διαβάσετε την εργασία [110].



(α) (β)

Εικόνα 9: (α) Σχηματική προβολή της παλινδρόμησης στάσης με βάση τα DNN. Οπτικοποιούμε τα επίπεδα του δικτύου με αντίστοιχες διαστάσεις τους, όπου τα συνελικτικά επίπεδα είναι με μπλε χρώμα, ενώ τα πλήρως συνδεδεμένα είναι με πράσινο χρώμα. (β) Στο στάδιο s, εφαρμόζεται μια τελειοποίηση της παλινδρόμισης σε μία δευτερεύουσα εικόνα για να βελτιώσει μια πρόβλεψη από το προηγούμενο στάδιο, Πηγή: deepai.org

Η χρήση μιας γενικής αρχιτεκτονικής DNN υποκινείται από τα εξαιρετικά αποτελέσματα της τόσο στην ταξινόμηση όσο και στον εντοπισμό του προβλήματος. Επιπλέον, ένα τέτοιο μοντέλο είναι πραγματικά ολιστικό. Η τελική εκτίμηση της θέσης άρθρωσης της βασίζεται σε μια σύνθετη μη γραμμική μετατροπή της πλήρους εικόνας. Επιπλέον, η χρήση ενός DNN αποκλείει την ανάγκη σχεδιασμού ενός συγκεκριμένου ειδικού μοντέλου για την εκτίμηση της στάσης. Αντί αυτού, ένα τέτοιο μοντέλο και τα χαρακτηριστικά τους μαθαίνονται από τα δεδομένα. Αν και η απώλεια της παλινδρόμησης δεν διαμορφώνει τις σαφείς αλληλεπιδράσεις μεταξύ των αρθρώσεων. Αυτές καταγράφονται έμμεσα από όλα τα κρυμμένα 7 επίπεδα.

Στην εκπαίδευση η διαφορά με την εργασία [110] είναι στις απώλειες. Αντί μιας απώλειας ταξινόμησης, εκπαιδεύεται μια γραμμική παλινδρόμηση πάνω από το τελευταίο επίπεδο δικτύου για να προβλέψει ένα διάνυσμα στάσης ελαχιστοποιώντας την L2 απόσταση μεταξύ της πρόβλεψης και του πραγματικού διανύσματος στάσης. Δεδομένου ότι το έδαφος διανύσματος πραγματικής στάσης ορίζεται σε απόλυτες συντεταγμένες εικόνας και οι στάσεις διαφέρουν σε μέγεθος από εικόνα σε εικόνα, ομαλοποιούμε το εκπαιδευτικό μας σύνολο D χρησιμοποιώντας την ομαλοποίηση από την εξίσωση (3.1):

$$D_N = \{(N(x), N(y))|(x, y)\ \epsilon\ D\} \quad (3.3)$$

Στην συνέχεια, η απώλεια L2 για την απόκτηση βέλτιστων παραμέτρων δικτύου διαβάζει:

$$\underset{\theta}{argmin} \sum_{(x,y)\in D_N} \sum_{i=1}^{k} \|y_i - \psi_i(x;\theta)\|_2^2 \quad (3.4)$$

Για λόγους σαφήνειας γράφονται οι βελτιστοποιήσεις πάνω από μεμονωμένες αρθρώσεις. Πρέπει ν σημειωθεί ότι ο ανωτέρω στόχος μπορεί να είναι χρήσιμος ακόμα και αν σε μερικές εικόνες δεν επισημαίνονται όλες οι αρθρώσεις. Σε αυτήν την περίπτωση, οι αντίστοιχοι όροι αθροιστικά πρέπει να παραληφθούν.

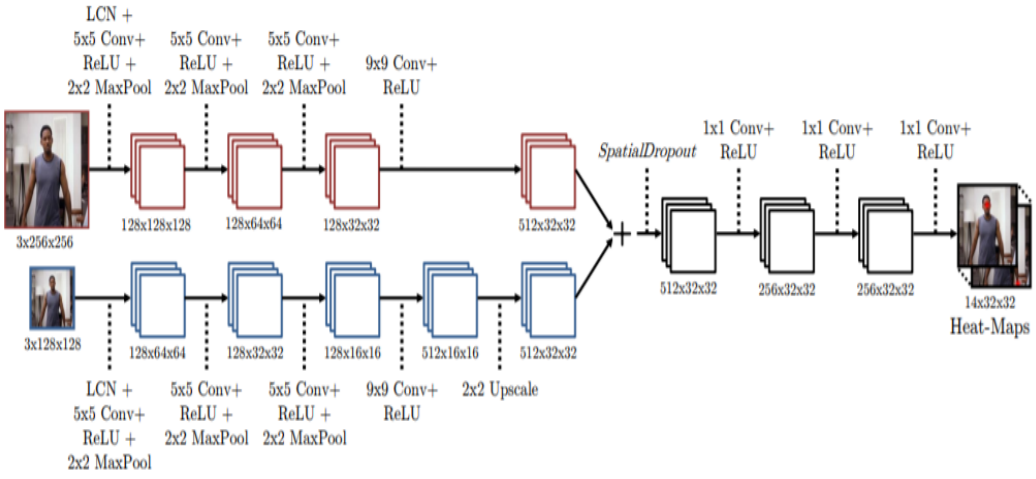
Οι παραπάνω παράμετροι θ βελτιστοποιούνται για την χρήση του Backpropagation σε μια κατανεμημένη διαδικτυακή υλοποίηση. Για κάθε μίνι-παρτίδα τουη μεγέθους 128, υπολογίζονται οι προσαρμοστικές ενημερώσεις κλίσεων [113]. Το ποσοστό μάθησης, ως η πιο σημαντική παράμετρος, έχει οριστεί 0,0005. Δεδομένου ότι το μοντέλο έχει μεγάλο αριθμό παραμέτρων και ότι τα χρησιμοποιημένα σύνολα δεδομένων είναι σχετικά μικρά μεγέθη, αυξάνονται τα δεδομένα χρησιμοποιώντας μεγάλο αριθμό τυχαίων μεταφρασμένων αποκομμάτων εικόνων, αριστερές/δεξιές αναστροφές καθώς η συστηματοποίηση DropOut για τα επίπεδα F έχει οριστεί σε 0,6.

3.2 Μοντέλο παλινδρόμησης χάρτη θερμότητας

Το μοντέλο αυτό είναι εμπνευσμένο από το έργο του Tompson και των υπολοίπων [114]. Χρησιμοποιείται μια αρχιτεκτονική ConvNet πολλαπλής ανάλυσης (εικόνα 10) για την υλοποίηση ενός ανιχνευτή συρόμενων παραθύρων με επικαλυπτόμενα περιβάλλοντα για την παραγωγή μιας χονδροειδούς εξόδου χάρτη θερμότητας. Δεδομένου ότι η δουλεία τους είναι μια επέκταση του μοντέλου τους, θα παρουσιαστεί μια επισκόπηση της αρχιτεκτονικής και θα εξηγηθούν οι επεκτάσεις στο μοντέλο τους.

3.2.1 Το μοντέλο χονδροειδούς παλιμδρόμησης χάρτη θερμότητας

Το χονδροειδές μοντέλο παλινδρόμησης χάρτη θερμότητας λαμβάνει ως είσοδο μια RGB Γκαουσιανή πυραμίδα 3 επιπέδων (στην εικόνα 10 εμφανίζονται 2 επίπεδα για συντομία) και παράγει ένα χάρτη θερμότητας για κάθε άρθωση που περιγράφει την πιθανότητα ανά pixel για την εν λόγω άρθρωση που εμφανίζεται σε κάθε χωρική θέση εξόδου. Χρησιμοποιούμε ανάλυση εικόνων εισόδου 320×240 και 256×256 pixels για τα σύνολα δεδομένων FLIC [115] και MPII [116] αντίστοιχα. Το πρώτο επίπεδο του δικτύου είναι ένα επίπεδο τοπικής-αντίθεσης-κανονικοποίησης (LCN) με το ίδιο φίλτρο kernel σε κάθε μια από τις τρείς τράπεζες επίλυσης.



Εικόνα 10: Ανιχνευτής συρόμενων παραθύρων πολλαπλής-ανάλυσης με επικαλυπτόμενα περιβάλλοντα (μοντέλο που χρησιμοποιείται σε σύνολο δεδομένων FLIC), Πηγή: groundai.com

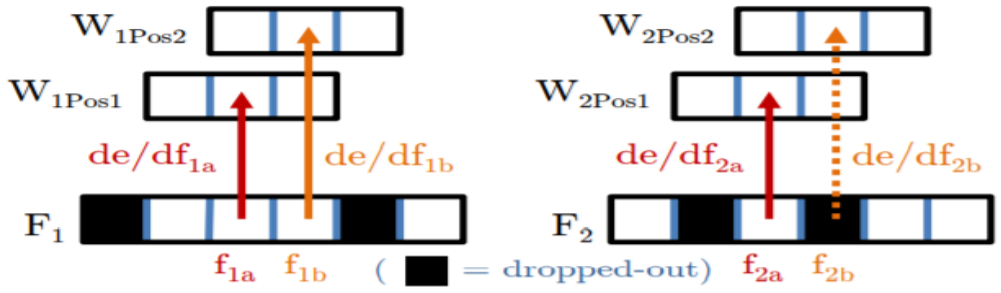
Κάθε εικόνα LCN εισάγεται στην συνέχεια σε ένα 7 σταδίων πολλαπλών φάσεων συνελικτικό δίκτυο (11 στάδια για το MPII μοντέλο συνόλου δεδομένων). Λόγω της παρουσίας ομαδοποίησης η έξοδος χάρτη θερμότητας είναι σε χαμηλότερη ανάλυση από την εικόνα εισόδου. Θα πρέπει να σημειωθεί ότι τα τελευταία 4 στάδια (ή 3 στάδια για το μοντέλο συνόλου δεδομένων MPII) προσομοιώνουν αποτελεσματικά ένα πλήρως συνδεδεμένο δίκτυο για μια στοχευμένη είσοδο μεγέθους μπαλώματος (το οποίο είναι συνήθως πολύ μικρότερο πλαίσιο από την εικόνα εισόδου).

3.2.2 Χωρική εγκατάλειψη

Βελτιώνεται το μοντέλο [114] με την προσθήκη ενός επιπροσθέτου επιπέδου εγκατάλειψης πριν από το πρώτο 1×1 συνελικτικό επίπεδο στην εικόνα 10. Ο ρόλος του επιπέδου εγκατάλειψης είναι να βελτιώσει την γενική απόδοση της μεθόδου με την παρεμπόδιση των ενεργοποιήσεων από το να γίνουν έντονα συσχετιζόμενες [117], η οποία με την σειρά της οδηγεί σε υπέρ-εκπαίδευση. Στην τυπική υλοποίηση εγκατάλειψης, οι ενεργοποιήσεις δικτύου ΄΄dropped-out΄΄ (μηδενίζοντας την ενεργοποίηση για τον εν λόγω νευρώνα) κατά την διάρκεια της εκπαίδευσης με την ανεξάρτητη πιθανότητα pdrop. Κατά τον χρόνο της δοκιμής όλες οι ενεργοποιήσεις χρησιμοποιούντα, αλλά ένα κέρδος 1-pdrop πολλαπλασιάζεται στις ενεργοποιήσεις των νευρώνων για να ληφθεί υπόψη για την αύξηση στην αναμενόμενη πόλωση.

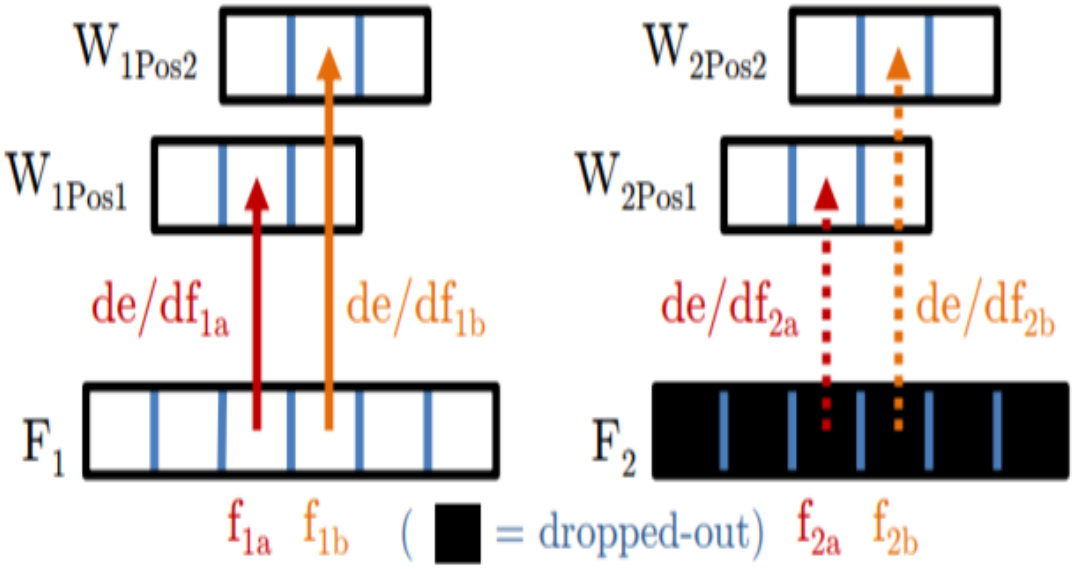
Στα αρχικά πειράματα, διαπιστώνεται ότι η εφαρμογή του επιπέδου εγκατάλειψης (όπου κάθε ενεργοποίηση χάρτη λειτουργίας συνέλιξης είναι ΄΄dropped-out΄΄ ανεξάρτητα) πριν το 1×1 συνελικτικό επίπεδο γενικώς αυξάνεται ο χρόνος εκπαίδευσης αλλά δεν εμποδίζει την υπερ-εκπαίδευση. Δεδομένου ότι το δίκτυο μας είναι πλήρως συνελικτικό και οι φυσικές εικόνες εμφανίζουν ισχυρή χωρική αντιστοιχία, η λειτουργία ενεργοποίησης χάρτη είναι επίσης αντίστοιχα ισχυρή και σε αυτό το πρότυπο ρύθμισης η εγκατάλειψη αποτυγχάνει.

Η τυπική εγκατάλειψη στην έξοδο μιας 1D συνέλιξης απεικονίζεται στην εικόνα 11. Οι δύο πρώτες γραμμές pixels αντιπροσωπεύουν τους πυρήνες συνέλιξης για τους χάρτες δυνατοτήτων 1 και 2, και η κάτω γραμμή αντιπροσωπεύει τις δυνατότητες εξόδου του προηγούμενου επιπέδου. Κατά την διάρκεια τηε πίσω διάδοσης, το κεντρικό pixel του W_2 πυρήνα λαμβάνει συνεισφορά διαβάθμισης τόσο από το f_{2a} και f_{2b} καθώς ο πυρήνας συνέλιξης W_2 μεταφράζεται μέσω την χαρακτηριστική εισόδου F_2. Σε αυτό το παράδειγμα, το F_{2B} απορρίφθηκε τυχαία (έτσι η ενεργοποίηση ορίστηκε στο μηδέν), ενώ το f_{2a} όχι. Δεδομένου ότι τα F_2 και F_1 είναι οι έξοδοι ενός συνελικτικού επιπέδου αναμένεται f_{2a} και f_{2b} να είναι έντονα συσχετιζόμενα: δηλαδή $f_{2a} \approx f_{2b}$ και $de/df_{2a} \approx de/df_{2b}$ (όπου e είναι η συνάρτηση σφάλματος για την ελαχιστοποίηση). Ενώ η συνεισφορά κλίσης από το f_{2b} είναι μηδέν, η έντονα συσχετιζόμενη κλίση f_{2a} παραμένει. Στην ουσία, το πραγματικό ποσοστό μάθησης κλιμακώνεται από την πιθανότητα εγκατάλειψης p, αλλά η ανεξαρτησία δεν βελτιώνεται.



Εικόνα 11: Τυπική εγκατάλειψη μετά από ένα επίπεδο συνέλιξης 1D, Πηγή: reaserchgate.net

Αντί αυτού διαμορφώνουμε μια νέα μέθοδο εγκατάλειψης, την οποία αποκαλούμε SpatialDropout. Για μια δεδομένη λειτουργία συνέλιξης τανυστή του μεγέθους n_{feats}×height×width, εκτελούμε μόνο n_{feats} δοκιμές εγκατάλειψης και επεκτείνουμε την τιμή εγκατάλειψης σε ολόκληρη την δυνατότητα του χάρτη. Επομένως, τα γειτονικά pixels στον χαρακτηριστικό χάρτη dropped-out είναι είτε όλα 0 (dropped-out) ή όλα ενεργά όπως απεικονίζεται στην εικόνα 12. Βρέθηκε ότι αυτή η τροποποιημένη υλοποίηση εγκατάλειψης βελτιώνει τις επιδόσεις, ειδικά στο FLIC σύνολο δεδομένων, όπου το μέγεθος του συνόλου εκπαίδευσης είναι μικρό.



Εικόνα 12: SpatialDropout μετά από ένα επίπεδο συνέλιξης 1D, Πηγή: researchgate.com

3.2.3 Εκπαίδευση και αύξηση δεδομένων

Εκπαιδεύουμε το μοντέλο της εικόνας 10 ελαχιστοποιώντας την απόσταση σφάλμα-ελαχίστων τετραγώνων (MSE) του προβλεπόμενου χάρτη θερμότητας σε ένα χάρτη θερμότητας στόχου. Ο στόχος είναι ένα 2D Gaussian σταθερής διακύμανσης (σ ≈ 1.5 pixels) στο κέντρο της αλήθειας-εδάφους (x,y) κοινής θέσης. Η συνάρτηση αντικειμένου είναι:

$$E_1 = \frac{1}{N} \sum_{j=1}^{N} \sum_{xy} \left\| H'_j(x,y) - H_j(x,y) \right\|^2 \quad (3.5)$$

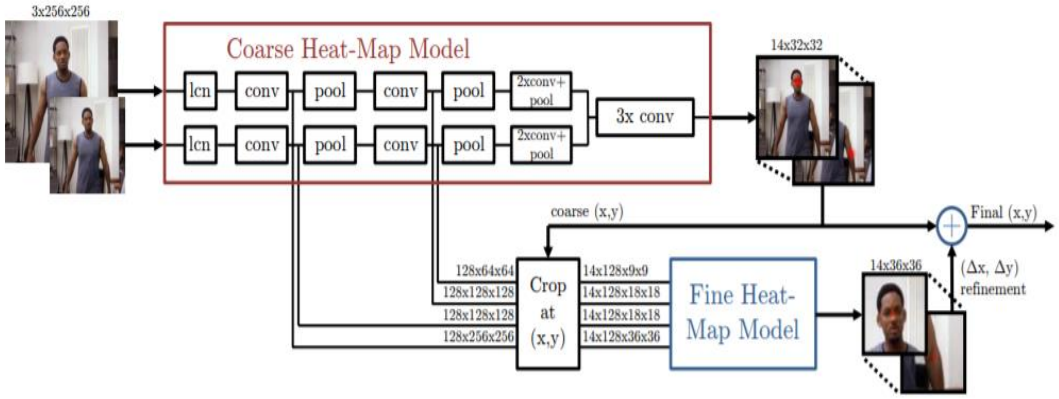
Όπου H'$_j$ και H$_j$ είναι η προβλεπόμενη και αλήθεια εδάφους χαρτών θερμότητας αντίστοιχα για την άρθρωση jth.

Κατά την διάρκεια της προπόνησης, κάθε εικονίδιο εισόδου περιστρέφεται τυχαία (r ∈ [-20° , +20°], σε κλίμακα (s ∈ [0.5 , 1.5]) και αναστράφηκε (με πιθανότητα 0.5) προκειμένου να βελτιωθεί η γενική επίδοση στο σύνολο επαλήθευσης. Σημειώστε ότι αυτό ακολουθεί το ίδιο πρωτόκολλο κατάρτισης όπως στο [114].

Πολλές εικόνες περιέχουν πολλά άτομα, ενώ μόνο ένα άτομο είναι δυνατόν να σχολιαστεί. Για να καταστεί δυνατή η εξαγωγή του στόχου σχολιασμού κατά την διάρκεια της δοκιμής, τόσο τα σύνολα δεδομένων FLIC όσο και το MPII περιλαμβάνουν κατά προσέγγιση θέσεις κορμού. Από την στιγμή που ο ανιχνευτής συρόμενων παραθύρων θα ανιχνεύσει όλες τις παρουσίες αρθρώσεων αδιακρίτως σε ένα μόνο καρέ, τότε ενσωματώνονται αυτές οι πληροφορίες κορμού εφαρμόζοντας το χωρικό μοντέλο που βασίζεται στο MRF του Tompson και των υπολοίπων [114], το οποίο διατυπώνει μια δομή δέντρου MRF σε χωρικές τοποθεσίες με τυχαία μεταβλητή για κάθε άρθρωση. Οι πιο πιθανές τοποθεσίες αρθρώσεων προκύπτουν (χρησιμοποιώντας μήνυμα) δεδομένων των θορύβων των διανομών εισόδου από το ConvNet. Η θέση της αλήθειας-εδάφους είναι συνδεδεμένη με τις 14 προβλεπόμενες αρθρώσεις από την έξοδο ConvNet και αυτές οι 15 θέσεις αρθρώσεων εισάγονται τότε στο MRF. Σε αυτήν την ρύθμιση, το βήμα συμπεράσματος MRF θα μάθει να μετριάζει τις ενεργοποιήσεις αρθρώσεων από ανθρώπους για τους οποίους ο κορμός εδάφους-αλήθειας δεν είναι ανατομικά εφαρμόσιμη, με αποτέλεσμα να "επιλέγει" το σωστό πρόσωπο για την επισήμανση. Οι ενδιαφερόμενοι αναγνώστες θα πρέπει να ανατρέξουν στην εργασία [115] για περισσότερες λεπτομέρειες.

3.2.4 Λεπτό μοντέλο παλινδρόμησης χάρτη θερμότητας

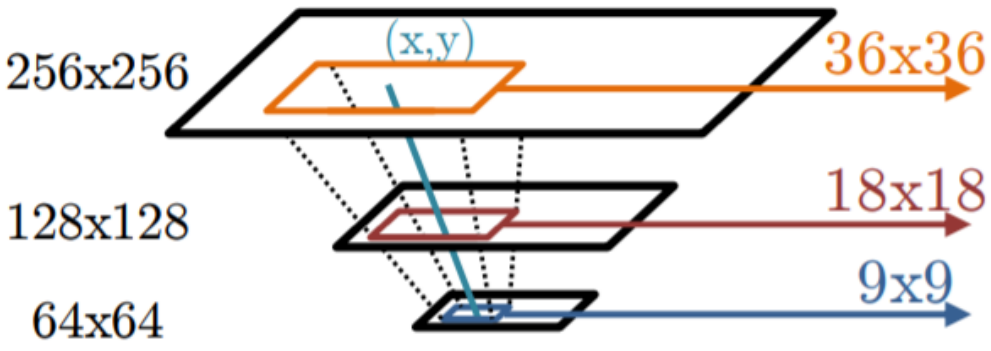
Η πλήρης αρχιτεκτονική του συστήματος εμφανίζεται στην εικόνα 13. Αποτελείται από το μοντέλο του χάρτη θερμότητας με βάση τα τμήματα για χονδροειδή εντοπισμό, μια ενότητα για την δειγματοληψία και την περικοπή των χαρακτηριστικών συνέλιξης σε μια συγκεκριμένη περιοχή (x,y)για κάθε άρθρωση, καθώς και ένα πρόσθετο μοντέλο για μικρορύθμιση.



Εικόνα 13: Επισκόπηση της επικαλυπτόμενης αρχιτεκτονικής, Πηγή: researchgate.net

Το κοινό συμπέρασμα από μια εικόνα εισόδου έχει ως εξής: μεταβιβάζουμε προς τα εμπρός (FPROP) μέσω του μοντέλου χονδροειδούς χάρτη θερμότητας, στην συνέχεια συμπεραίνουμε όλες τις θέσεις αρθρώσεων (x,y) από την μέγιστη τιμή του χάρτη θερμότητας σε κάθε άρθρωση. Στην συνέχεια χρησιμοποιούμε αυτή την χονδροειδή θέση (x,y) για την δειγματοληψία και την περικοπή των πρώτων δύο επιπέδων συνέλιξης (για όλες τις τράπεζες επίλυσης) σε κάθε μια από τις θέσεις αρθρώσεων. Μετέπειτα γίνεται FPROP αυτά τα χαρακτηριστικά μέσω ενός λεπτού μοντέλου χάρτη θερμότητας για την παραγωγή ενός (Δ_χ, Δ_ψ) μετατόπισης εντός του υπό-παραθύρου της περικοπής. Τέλος, προσθέτουμε την βελτίωση της θέσης στην χονδροειδή θέση για την παραγωγή ενός τελικού εντοπισμού (x,y) για κάθε άρθρωση.

Η εικόνα 14 εμφανίζει την λειτουργικότητα της μονάδας περικοπής για μία μόνο άρθρωση. Εμείς απλά περικόπτουμε από ένα παράθυρο με κέντρο την χονδροειδή θέση άρθρωσης (x, y) σε κάθε χάρτη ανάλυσης των χαρακτηριστικών, ωστόσο το κάνουμε διατηρώντας το αναδυόμενο μέγεθος του παραθύρου σταθερό με την κλιμάκωση της περιοχής περικοπής σε κάθε επίπεδο υψηλής ανάλυσης. Σημειώνεται ότι η πίσω διάδοση (BPROP) μέσω αυτής μονάδας από τη λειτουργία εισόδου στην λειτουργία εξόδου είναι τετριμμένη. Οι διαβαθμίσεις εξόδου από τις περικομμένες εικόνες προστίθενται απλά στις διαβαθμίσεις εξόδου των επιπέδων συνέλιξης στο μοντέλο χονδροειδούς χάρτη θερμότητας στις θέσεις pixel του δείγματος.

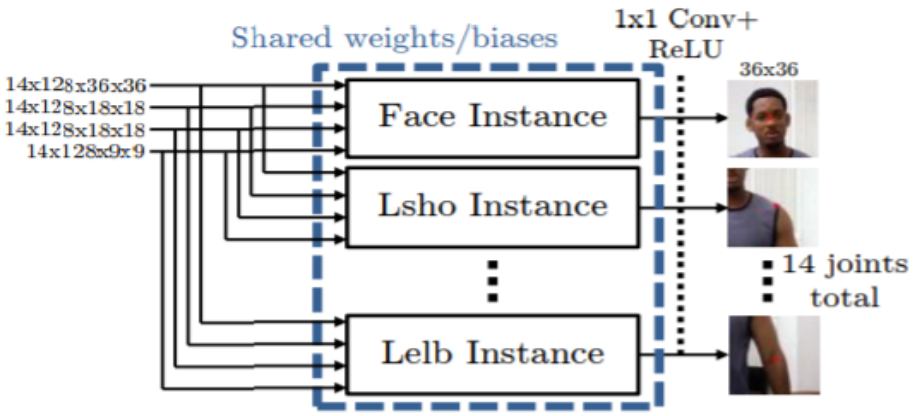


Εικόνα 14: Λειτουργικότητα μονάδας περικοπής για μια μόνο άρθρωση, Πηγή: researchgate.net

Το λεπτό μοντέλο χάρτη θερμότητας είναι ένα σιαμαίο δίκτυο [118] 7 περιπτώσεων (14 για το σύνολο δεδομένων MPII), όπου τα βάρη και οι τάσεις κάθε ενότητας μοιράζονται (π.χ. αναπαραγωγή σε όλες τις περιπτώσεις και ενημέρωση μαζί κατά την διάρκεια BPROP). Έτσι η θέση του δείγματος για κάθε άρθρωση είναι διαφορετική, τα χαρακτηριστικά της συνέλιξης δεν μοιράζονται τον ίδιο χωρικό πλαίσιο και έτσι η συνελικτικά υπόδίκτυα πρέπει να εφαρμόζονται σε κάθε άρθρωση ανεξάρτητα. Ωστόσο, χρησιμοποιείται η κοινή χρήση παραμέτρων μεταξύ κάθε ένα από τα 7 υπό-δίκτυα και τότε εκτελείται μια 1×1 συνέλιξη, χωρίς κοινή χρήση βαρών, ώστε να εξάγουμε ένα λεπτομερώς-αναλυμένο χάρτη θερμότητας για κάθε άρθρωση. Ο σκοπός αυτού του τελευταίου επιπέδου είναι να εκτελέσει την τελική ανίχνευση για κάθε άρθρωση.

Σημειώνεται ότι ενδέχεται να εκτελεστούν περιττοί υπολογισμοί στο δίκτυο σιαμαίων. Εάν δύο περικοπόμενα υπό-παράθυρα επικαλύπτονται και δεδομένου ότι τα βάρη της συνέλιξης είναι κοινά, η ίδια συνέλιξη ίσως εφαρμοστεί πολλές φορές στις ίδιες χωρικές θέσεις. Ωστόσο, δια πιστώνεται στην πράξη ότι αυτό συμβαίνει σπάνια. Οι αρθρώσεις είναι σπάνια κοινά τοποθετημένες και το χωρικό μέγεθος του πλαισίου επιλέγεται έτσι ώστε να υπάρχει μικρή επικάλυψη ανάμεσα στις περικοπόμενες υπό-περιοχές (σημειώνεται ότι το πλαίσιο των περικοπόμενων εικόνων που φαίνεται στις εικόνες 4 και 8 είναι υπερβολικά για λόγους σαφήνειας).

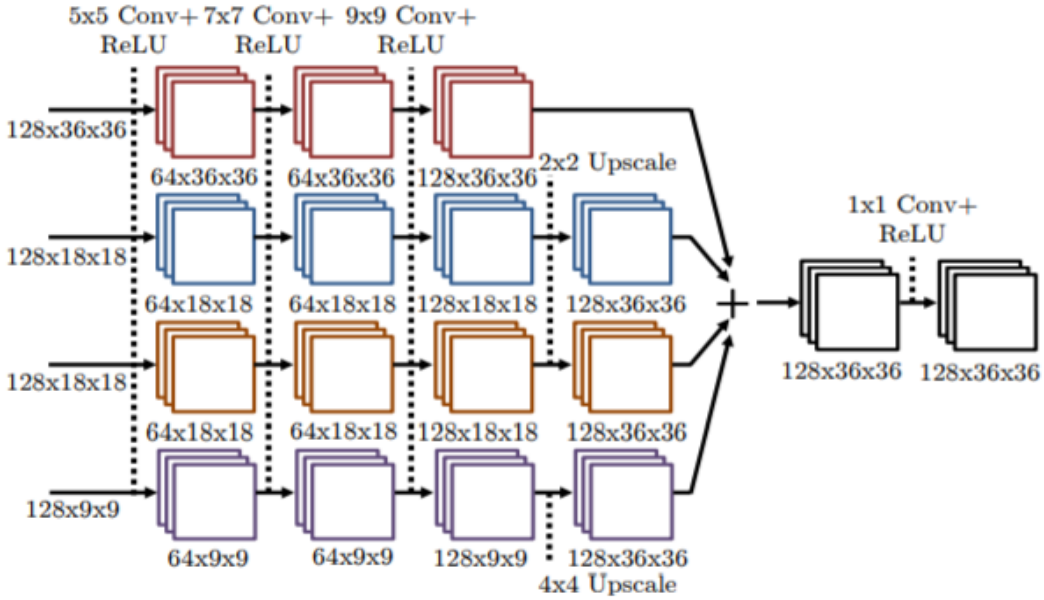
Κάθε περίπτωση του υπό-δικτύου στην εικόνα 7 είναι ένα ConvNet 4 επιπέδων, όπως φαίνεται και στην εικόνα 8. Δεδομένου ότι οι εικόνες εισόδου είναι διαφορετικών αναλύσεων και προέρχονται από ποικίλα βάθη στο χονδροειδές μοντέλο του χάρτη θερμότητας, αντιμετωπίζονται τα χαρακτηριστικά εισόδου ως ξεχωριστές τράπεζες ανάλυσης και εφαρμόζεται μια παρόμοια στρατηγική αντιμετώπισής όπως περιεγράφηκε και παραπάνω. Έτσι εφαρμόζεται το ίδιο μέγεθος συνελίξεων σε κάθε τράπεζα, αναβαθμίζονται τα χαμηλής ανάλυσης χαρακτηριστικά για να τα φέρουν σε κανονική ανάλυση. Επιπλέον γίνονται οι ενεργοποιήσεις σε κάθε χάρτη δυνατοτήτων και μετά εφαρμόζεται 1×1 συνελίξεις στα χαρακτηριστικά εξόδου.



Εικόνα 15: Μοντέλο λεπτού χάρτη θερμότητας- Σιαμαίο δίκτυο 14 αρθρώσεων, Πηγή: researchgate.net

Θα πρέπει να σημειωθεί ότι αυτή η επικαλυπτόμενη αρχιτεκτονική μπορεί να επεκταθεί περαιτέρω όσο είναι δυνατόν να έχουμε πολλά επίπεδα κλιμάκωσης το καθένα με όλο και μικρότερη ομαδοποίηση. Ωστόσο στην πράξη έχει βρεθεί ότι μόνο ένα επίπεδο

παρέχει επαρκή ακρίβεια, και ιδίως εντός του επιπέδου του θορύβου σήματος στην βάση δεδομένων FLIC.



Εικόνα 16: Το λεπτό δίκτυο χάρτη θερμότητας για μια μόνο άρθρωση, Πηγή: researchgate.net

Πριν από την κοινή εκπαίδευση, πρώτα προ-προπονείται το χονδροειδές μοντέλο χάρτη θερμότητας ελαχιστοποιώντας την εξίσωση 3.5. Στην συνέχεια διατηρούνται οι παράμετροι του χονδροειδούς μοντέλου σταθερές και εκπαιδεύεται το λεπτό μοντέλο χάρτη θερμότητας ελαχιστοποιώντας την εξίσωση:

$$E_2 = \frac{1}{N} \sum_{j=1}^{N} \sum_{xy} \left\| G'_j(x,y) - G_j(x,y) \right\|^2 \quad (3.6)$$

Όπου G' και G είναι το σύνολο των προβλεπόμενων και των εδαφών αληθείας των χαρτών θερμότητας αντίστοιχα για το λεπτό μοντέλο του χάρτη θερμότητας. Τέλος εκπαιδεύονται από κοινού και τα δύο μοντέλα χάρτη θερμότητάς ελαχιστοποιώντας το $E_3 = E_1 + \lambda E_2$.Όπου το λ είναι μια σταθερά που χρησιμοποιείται για την αντιστάθμιση της σχετικής σημασίας και των δύο επιμέρους καθηκόντων. Αντιμετωπίζουμε το λ ως άλλη μια υπερ-παράμετρο του δικτύου και επιλέγεται για την βελτιστοποίηση της απόδοσης σε σχέση με το σύνολο επαλήθευσης (χρησιμοποιούμε λ=0,1).

Ιδανικά, μια πιο άμεση λειτουργία βελτιστοποίησης θα προσπαθήσει να μετρήσει το argmax των δύο χαρτών θερμότητας και ως εκ τούτου άμεσα ελαχιστοποιείται η τελική (x , y) πρόβλεψη. Ωστόσο, δεδομένου ότι η συνάρτηση argmax δεν είναι διαφορίσιμη αντ' αυτού αναδιατυπώνεται το πρόβλημα ως μια παλινδρόμηση σε ένα σύνολο στόχων χαρτών θερμότητας και ελαχιστοποιούμε την απόσταση από αυτούς τους χάρτες θερμότητας.

3.3 Μοντέλο με επαναληπτική ανάδραση σφάλματος

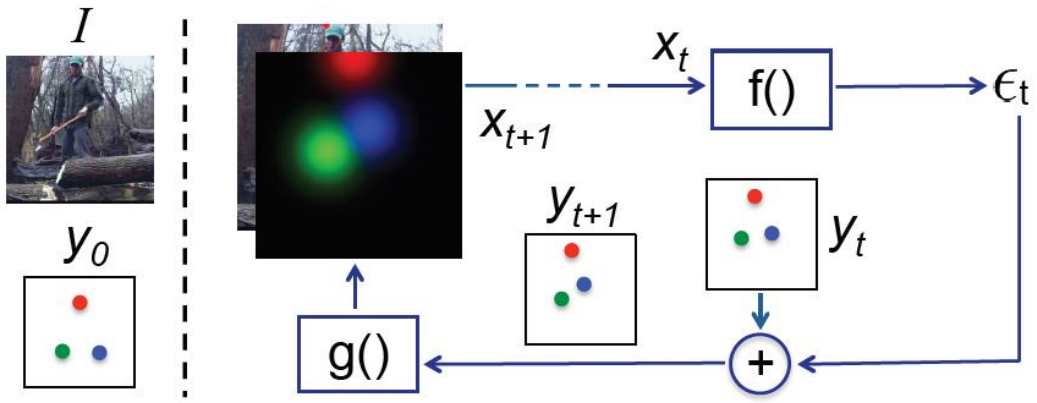
Τα δίκτυα συνέλιξης (ConvNets) [119] αναπαριστούν εικόνες χρησιμοποιώντας μια ιεραρχία χαρακτηριστικών πολλαπλών επιπέδων και εμπνέονται από την δομή και την λειτουργικότητα της οπτικής οδού του εγκεφάλου [120,121]. Ο υπολογισμός των χαρακτηριστικών γνωρισμάτων σε αυτά τα δίκτυα είναι καθαρά feedforward, σε αντίθεση με το ανθρώπινο οπτικό σύστημα όπου οι συνδέσεις ανάδρασης αφθονούν [122,123,124]. Η ανάδραση μπορεί να χρησιμοποιηθεί για τη διαμόρφωση και την εξαγωγή ειδικών χαρακτηριστικών στα αρχικά επίπεδα ώστε να μοντελοποιήσει χρονικά και χωρικά περιβάλλοντα (π.χ. προετοιμασία) για την εκμετάλλευση προηγούμενων γνώσεων σχετικά με το σχήμα για την τμηματοποίηση και την 3D αντίληψη ή απλά για την καθοδήγηση της οπτικής προσοχής στις περιοχές της εικόνας που σχετίζονται με το έργο που είναι υπό εξέταση.

Η κύρια συνεισφορά της μεθόδου αυτής είναι η δημιουργία ενός γενικού πλαισίου για την μοντελοποίηση πλούσιων δομών τόσο στα διαστήματα είσοδού όσο και έξοδού με την εκμάθηση απαγωγέων ιεραρχικών χαρακτηριστικών πέρα από το κοινό χώρο. Αυτό επιτυγχάνεται ενσωματώνοντας μια από πάνω προς τα κάτω ανάδραση αντί να προσπαθούμε να προβλέψουμε άμεσα τον στόχο εξόδων, όπως στην διαδικασία feedback, προβλέπεται το λάθος με την τρέχουσα εκτίμηση τους και να το διορθώνει μετά επαναληπτικά. Το πλαίσιο αυτό καλείται επαναληπτική ανατροφοδότηση σφάλματος ή IEF.

Στο IEF, ένα μοντέλο feedforward f λειτουργεί στον επαυξημένο χώρο εισόδου που δημιουργείται από την συνένωση (δηλώνεται με $\oplus$) της εικόνας RGB I με οπτική αναπαράσταση g της εκτιμώμενης εξόδου y_t για να προβλέψει μια "διόρθωση" (ϵ_t) που φέρνει το y_t πιο κοντά στο έδαφος αληθείας εξόδου y. Το σήμα διόρθωσης t εφαρμόζεται στην τρέχουσα έξοδο t για την δημιουργία του y_{t+1} και αυτό μετατρέπεται σε μια οπτική αναπαράσταση από το g, το οποίο στοιβάζεται με την εικόνα για να παράγει νέες εισόδους $x_{t+1} = I \oplus g(y_t)$ για f, και ούτω καθεξής επαναληπτικά. Αυτή η διαδικασία έχει προετοιμαστεί με μια εικασία της εξόδου (y_0) και επαναλαμβάνεται μέχρι ένα προκαθορισμένο κριτήριο τερματισμού. Το μοντέλο είναι εκπαιδευμένο να παράγει οριοθετημένες διορθώσεις σε κάθε επανάληψη, π.χ. $\|e_t\|_2 < L$. Το κίνητρο για την τροποποίηση του y_t από μια οριακή τιμή είναι ότι ο χώρος του x_t είναι συνήθως ιδιαίτερα μη γραμμικός και ως εκ τούτου οι τοπικές διορθώσεις θα πρέπει να είναι ευκολότερες στην εκμάθηση. Η λειτουργία του μοντέλου μπορεί να περιγράφει μαθηματικά από τις ακόλουθες εξισώσεις:

$$e_t = f(x_t) \quad (3.7)$$
$$y_{t+1} = y_t + \varepsilon_t \quad (3.8)$$
$$x_{t+1} = I \oplus g(y_{t+1}) \quad (3.9)$$

όπου οι συναρτήσεις f και g έχουν πρόσθετες παραμέτρους που έχουν διδαχθεί Θ_f και Θ_g αντίστοιχα. Αν και έχει χρησιμοποιηθεί το προβλεπόμενο σφάλμα για την πρόσθετη τροποποίηση y_t στην εξίσωση 3.8, γενικά το y_{t+1} μπορεί να είναι αποτέλεσμα μια αυθαίρετης μη γραμμικής συνάρτησης που λειτουργεί σε y_t, ε_t.



Εικόνα 17: Εφαρμογή επαναληπτικής ανάδρασης σφάλματος για εκτίμηση της ανθρώπινης στάσης, Πηγή: arxiv-vanity.com

Στο τρέχον παράδειγμα της εκτίμησης της ανθρώπινης στάσης, y_t είναι ένα διάνυσμα των ρετινοτοπικών θέσεων όλων των σημείων αναφοράς που είναι μεμονωμένα και τα οποία αντιστοιχίζονται από το g σε χάρτες θερμότητας (π.χ K χάρτες θερμότητας για K σημεία αναφοράς). Οι χάρτες θερμότητας στοιβάζονται μαζί με την εικόνα και εισέρχονται ως είσοδο στο f (βλέπε στο σχήμα 1 για μια επισκόπηση). Η συνάρτηση ΄΄απόδοσης΄΄ g σε αυτή την συγκεκριμένη περίπτωση δεν μαθαίνεται αλλά αντί αυτού διαμορφώνεται ως μια 2D Gaussian έχοντας σταθερή τυπική απόκλιση και με κέντρο την θέση του σημείου κλειδιού. Διαισθητικά, αυτοί οι χάρτες θερμότητας κωδικοποιούν την τρέχουσα πεποίθηση στις θέσεις των σημείων κλειδιών στο επίπεδο εικόνας και ως εκ τούτου αποτελούν μια φυσική αναπαράσταση για χαρακτηριστικά μάθησης πάνω από το κοινό χώρο των διαμορφώσεων του σώματος και της RGB εικόνας.

Η διάσταση των εισόδων στο f είναι H × W × (K + 3), όπου H, W αναπαριστούν το ύψος και το πλάτος της εικόνας και το (K + 3) αντιστοιχούν σε K σημεία αναφοράς και το και στα 3 κανάλια χρώματος της εικόνας. Μοντελοποιείται η f με ένα ConvNet με παραμέτρους Θ_f (π.χ. βάρη ConvNet). Όπως το ConvNet παίρνει $I \oplus g(y_t)$ ως εισόδους, έχει την δυνατότητα να μαθαίνει χαρακτηριστικά στο κοινό χώρο εισόδου-εξόδου.

3.3.1 Η εκμάθηση της μεθόδου

Για να συμπεράνουμε το έδαφος αληθείας της εξόδου (y), η συγκεκριμένη μέθοδος βελτιώνει επαναληπτικά την τρέχουσα έξοδο y(t). Σε κάθε επανάληψη, η f προβλέπει μια διόρθωση (ε_t) η οποία βελτιώνει τοπικά την τρέχουσα έξοδο. Σημειώνεται ότι εκπαιδεύεται το μοντέλο για να προβλέψει οριακές διορθώσεις, αλλά δεν επιβάλλονται τέτοιοι περιορισμοί κατά την διάρκεια της δοκιμής. Οι παράμετροι (Θ_f, Θ_g) των συναρτήσεων f και g στο μοντέλο, μαθαίνουν με την βελτιστοποίηση της εξίσωσης (3.10):

$$\min_{\Theta_f \Theta_g} \sum_{t=1}^{T} (\varepsilon_t, e(y, y_t)) \quad (3.10)$$

Όπου, ε_t και e(y, y_t) προβλέπονται και στοχεύουν σε οριακές διορθώσεις αντίστοιχα. Η συνάρτηση h είναι μια μέτρηση της απόστασης, όπως μια τετραγωνική απώλεια. Τ είναι ο αριθμός των βημάτων διόρθωσης που λαμβάνονται από το μοντέλο. Το Τ μπορεί είτε να επιλεγεί για να είναι μια σταθερά ή γενικότερα να είναι μια συνάρτηση του t (δηλαδή, όρος τερματισμού).

Algorithm 1 Learning Iterative Error Feedback with Fixed Path Consolidation

procedure FPC-LEARN
  Initialize y_0
  $E \leftarrow \{\}$
  for $t \leftarrow 1$ to (T_{steps}) **do**
    for all training examples (I, y) **do**
      $\epsilon_t \leftarrow e(y, y_t)$
    end for
    $E \leftarrow E \cup \epsilon_t$
    for $j \leftarrow 1$ to N **do**
      Update Θ_f and Θ_g with SGD, using loss h and target corrections E
    end for
  end for
end procedure

Η λειτουργία κόστους βελτιστοποιείται , χρησιμοποιώντας στοχαστική διαβάθμιση καθόδου (SGD) με κάθε βήμα διόρθωσης να είναι ανεξάρτητο παράδειγμα κατάρτισης. Το σύνολο της εκπαίδευσης αυξάνεται προοδευτικά: ξεκινάει με την εκμάθηση των ανάλογων δειγμάτων στο πρώτο βήμα για Ν φορές, στην συνέχεια προστίθενται τα ανάλογα δείγματα που αντιστοιχούν στο δεύτερο βήμα και εκπαιδεύει άλλες Ν φορές και ούτω καθεξής, έτσι ώστε τα αρχικά βήματα να βελτιστοποιηθούν περισσότερο.

Με την υπόθεση ότι μόνο η έξοδος εδάφους αληθείας (y) παράγεται στον χρόνο εκπαίδευσης, δεν είναι σαφές ποιοι θα πρέπει να είναι οι ενδιάμεσοι στόχοι (y_t). Η πιο απλή στρατηγική, η οποία χρησιμοποιείται, είναι να προκαθορίσουμε y_t για κάθε επανάληψη χρησιμοποιώντας ένα σύνολο σταθερών διορθώσεων e(y, y_t) ξεκινώντας από το y_0, λαμβάνοντας τα (y_0, y_1, ..y). Η συνολική διαδικασία εκμάθησης ονομάζεται σταθεροποίηση σταθερής διαδρομής (FPC) και περιγράφεται στο αλγόριθμο 1 (algorithm 1).

Οι διορθώσεις που οροθετούνται από τον στόχο για κάθε επανάληψη υπολογίζονται χρησιμοποιώντας μια συνάρτηση e(y, y_t), η οποία μπορεί να πάρει διαφορετικές μορφές για διαφορετικά προβλήματα . Εάν για παράδειγμα, η έξοδος είναι 1D, τότε e(y, y_t) = max(sign(y-y_t)×a, y-y_t) θα συνεπάγεται ότι ο οριοθετημένος ΄΄στόχος΄΄ σφάλματος θα διορθώσει το y_t από μια μέγιστη ποσότητα α προς την κατεύθυνση του y.

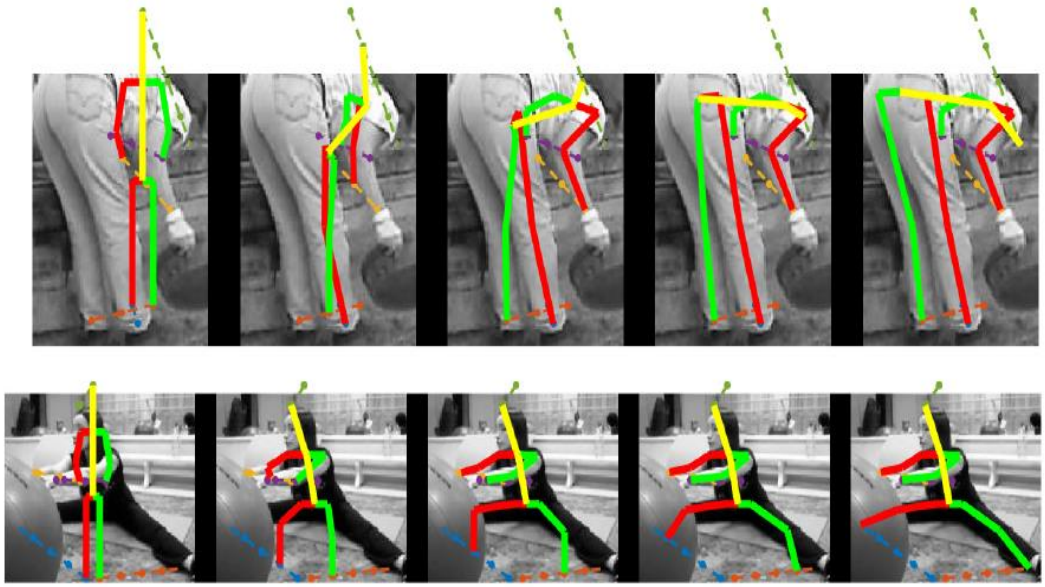
3.3.2 Εκμάθηση της μεθόδου για την εκτίμηση της στάσης του σώματος

Η ανθρώπινη στάση αναπαραστάθηκε από ένα σύνολο 2D θέσεων κλειδιών y: $\{y^k \in \mathbb{R}^2, k \in [1, K]\}$, όπου το K είναι ο αριθμός των σημείων κλειδιών και το y^k υποδηλώνει το k^{th} σημείο κλειδί. Η προβλεπόμενη θέση των σημείων κλειδιών στην t^{th} επανάληψη έχει y_t: $\{ y_t^k, k \in [1, k]\}$. Η απόδοση των y_t ως χαρτών θερμότητας συνδεδεμένα με την εικόνα δίνεται ως είσοδος σε ένα ConvNet. Το ConvNet εκπαιδεύτηκε να προβλέψει μια ακολουθία ‘’οριοθετούμενων’’ διορθώσεων για κάθε σημείο κλειδί (ε_t^k). Οι διορθώσεις χρησιμοποιήθηκαν για την επαναληπτική βελτίωση των θέσεων των σημείων κλειδιών.

Το $u = y^k - y_t^k$ και το αντίστοιχο διάνυσμα μονάδας είναι $\hat{u} = \frac{u}{\|u\|_2}$. Τότε ο στόχος ‘’οριοθετημένης’’ διόρθωσης για την t^{th} επανάληψη και k^{th} σημείο κλειδί υπολογίστηκε ως:

$$e(y^k, y_t^k) = min(L, \|u\|) \cdot \hat{u} \quad (3.11)$$

όπου το L υποδηλώνει την μέγιστη μετατόπιση για κάθε θέση σημείου κλειδιού. Μια ενδιαφέρουσα ιδιότητα αυτής της συνάρτησης είναι ότι είναι σταθερή ενώ ένα βασικό σημείο κλειδί απέχει μακριά από το έδαφος αληθείας και ποικίλει μόνο σε κλίμακα όταν είναι πιο κοντά από το L στο έδαφος αληθείας. Αυτό απλοποιεί το πρόβλημα εκμάθησης: δίνεται μια εικόνα και μια σταθερή αρχική στάση και το μοντέλο πρέπει απλά να προβλέψει μια σταθερή κατεύθυνση προς την οποία θα μετακινηθούν τα σημεία κλειδιά και για να ‘’επιβραδύνει’’ την κίνηση προς αυτή την κατεύθυνση, όταν το σημείο κλειδί έρχεται κοντά στο έδαφος αληθείας. Δείτε την εικόνα 18 για εικογράφιση της μεθόδου.

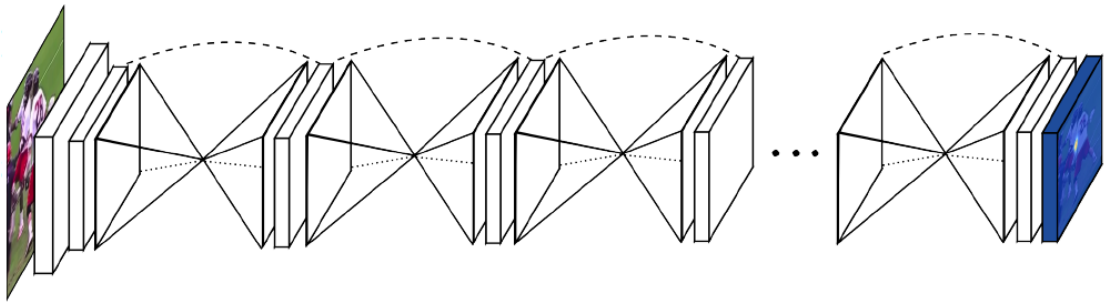


Εικόνα 18: Στην εκτίμηση της ανθρώπινης στάσης που τρέχει στο παράδειγμα μας, η ακολουθία των διορθώσεων ε_t κινεί τα σημεία κλειδιά κατά μήκος των γραμμών στην εικόνα, ξεκινώντας από μια αρχική μέση στάση y_0 (αριστερά), σε όλη την διαδρομή προς το έδαφος αληθείας θέτουν y (δεξιά),εδώ φαίνεται για δύο διαφοτερετικές εικόνες, Πηγή: semanticscholar.org

3.4 Δίκτυο στοιβασμένης κλεψύδρας για την εκτίμηση της στάσης του σώματος

Μια άλλη μέθοδος για την πρόβλεψη της ανθρώπινης στάσης είναι ο σχεδιασμός του δικτύου με την μέθοδο ‘’στοιβασμένης κλεψύδρας’’. Το δίκτυο λαμβάνει και ενοποιεί πληροφορίες σε όλες τις κλίμακες της εικόνας. Αναφέρεται στο σχεδιασμό ως κλεψύδρα λόγω της απεικόνισής των βημάτων της ομαδοποίησης και των επακόλουθων δειγματοληψειών που χρησιμοποιούνται για να πάρουμε την τελική έξοδο του δικτύου. Όπως πολλές συνελικτικές προσεγγίσεις που παράγουν pixel-wise εξόδους, το δίκτυο κλεψύδρας ομαδοποιεί κάτω σε πολύ χαμηλή ανάλυση, και στην συνέχεια upsamples και συνδυάζει χαρακτηριστικά σε πολλαπλά αποτελέσματα [125,126]. Από την άλλη πλευρά, η κλεψύδρα διαφέρει από προηγούμενους σχεδιασμούς κυρίως στην πιο συμμετρική της τοπολογία.

Στην συνέχεια επεκτείνεται σε μια κλεψύδρα τοποθετώντας διαδοχικά πολλαπλές μονάδες κλεψυδρών μαζί από άκρη σε άκρη. Αυτό επιτρέπει επαναλαμβανόμενες από κάτω προς τα πάνω, από πάνω προς τα κάτω συμπεράσματα σε διάφορες κλίμακες. Σε συνδυασμό με την χρήση ενδιάμεσης εποπτείας, το επαναλαμβανόμενο αμφίδρομο συμπέρασμα είναι ζωτικής σημασίας για τις επιδόσεις του δικτύου.



Εικόνα 19: Το δίκτυο για την εκτίμηση της στάσης του σώματος αποτελείται από πολλαπλές στοιβασμένες Κλεψύδρες, Πηγή: paperswithcode.com

Η μονάδα κλεψύδρας διαφέρει από τα προηγούμενούς σχεδιασμούς δικτύων κυρίως επειδή έχει πιο συμμετρική κατανομή της ικανότητα μεταξύ της επεξεργασίας από κάτω προς τα πάνω (από υψηλές αναλύσεις έως χαμηλές αναλύσεις) και της επεξεργασίας από πάνω προς κάτω (από χαμηλές σε υψηλές αναλύσεις). Για παράδειγμα, τα πλήρως συνελίκτικά δίκτυα [126] και οι ολιστικά ένθετες αρχιτεκτονικές [127] είναι και οι δύο βαριές σε επεξεργασίες από κάτω προς τα πάνω, αλλά ελαφριές στις από πάνω προς τα κάτω επεξεργασίες, οι οποίες αποτελούνται μόνο από ένα (σταθμισμένη) συγχώνευση προβλέψεων σε πολλαπλές κλίμακες. Τα πλήρως συνελικτικά δίκτυα επίσης εκπαιδεύονται σε πολλαπλά στάδια.

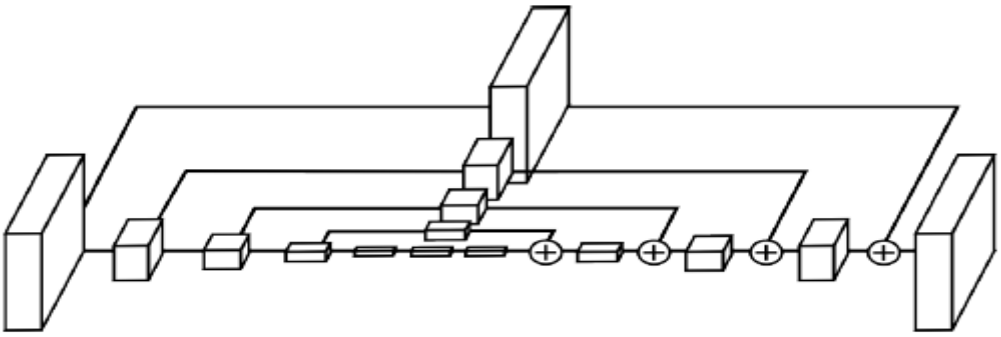
3.4.1 Η αρχιτεκτονική του δικτύου

Ο σχεδιασμός της κλεψύδρας έχει ως κίνητρο την ανάγκη λήψης πληροφοριών σε κάθε κλίμακα. Ενώ τα τοπικά στοιχεία είναι απαραίτητα για τον εντοπισμό χαρακτηριστικών όπως τα πρόσωπα και τα χέρια, μια τελική εκτίμηση ης στάσης του σώματος απαιτεί συνεκτική κατανόηση όλου του σώματος. Ο προσανατολισμός του ατόμου, η διάταξη των άκρων του και οι σχέσεις των παρακείμενων αρθρώσεων είναι μεταξύ των πολλών ενδείξεων που αναγνωρίζονται καλύτερα σε διαφορετικές κλίμακες στην εικόνα. Η κλεψύδρα είναι ένα απλό, ελάχιστο σχέδιο που έχει την ικανότητα να συλλαμβάνει όλα αυτά τα χαρακτηριστικά και τα ‘’φέρνει’’ μαζί για την παραγωγή pixel-wise προβλέψεων.

Το δίκτυο πρέπει να διαθέτει κάποιο μηχανισμό για την αποτελεσματική επεξεργασία και τη σύνδεση χαρακτηριστικών σε όλες τις κλίμακες. Ορισμένες προσεγγίσεις αντιμετωπίζουν αυτό το θέμα με την χρήση αγωγών διοχέτευσης που επεξεργάζονται την εικόνα ανεξάρτητα σε πολλαπλές αναλύσεις και συνδυάζουν δυνατότητες αργότερα στο δίκτυο [125,128]. Αντί αυτού επιλέγεται να χρησιμοποιείτε ένας μοναδικός αγωγός διοχέτευσης με επίπεδα παράλειψης για τη διατήρηση χωρικών πληροφοριών σε κάθε ανάλυση. Το δίκτυο φτάνει στη χαμηλότερη ανάλυση 4×4 pixel, επιτρέποντας μικρότερα χωρικά που θα εφαρμοστούν και που θα συγκρίνουν δυνατότητες σε ολόκληρο το χώρο της εικόνας.

Η κλεψύδρα έχει ρυθμιστεί ως εξής: Τα συνελικτικά και τα μέγιστα επίπεδα συγκέντρωσης που χρησιμοποιούνται για την επεξεργασία χαρακτηριστικών σε πολύ χαμηλή ανάλυση. Σε κάθε βήμα μέγιστης συγκέντρωσης, το δίκτυο branches off και εφαρμόζει περισσότερες συνελίξεις στην αρχική προ-συγκεντρωτική ανάλυση. Αφού επιτευχθεί η χαμηλότερη ανάλυση, το δίκτυο ξεκινάει την από πάνω προς τα κάτω ακολουθία της αναδειγματοληψίας και του συνδυασμού των χαρακτηριστικών σε όλες τις κλίμακες. Για να συγκεντρωθούν πληροφορίες σε δύο παρακείμενα αποτελέσματα, ακολουθείτε η διαδικασία που περιγράφεται από τον Tompson και τους υπολοίπους [125] και κάνει αναδειγματολήψία στους κοντινότερος γείτονες της κατώτερης ανάλυσης ακολουθούμενη από μια προσθήκη elementwise των δύο συνόλων χαρακτηριστικών. Η αρχιτεκτονική της κλεψύδρας είναι συμμετρική, έτσι λοιπόν για κάθε επίπεδο που υπάρχει στο δρόμο προς τα κάτω υπάρχει και ένα αντίστοιχο επίπεδο που ανεβαίνει.

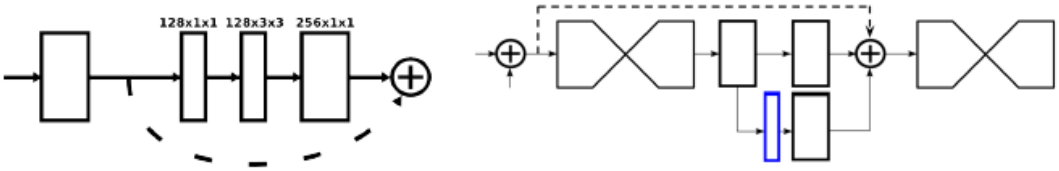
Μετά την επίτευξή της ανάλυσης εξόδου του δικτύου, δυο διαδοχικοί γύροι 1×1 συνελίξεων εφαρμόζονται για την παραγωγή των προβλέψεων του τελικού δικτύου. Η έξοδος του δικτύου είναι ένα σύνολο χαρτών θερμότητας όπου για ένα δεδομένο χάρτη θερμότητας το δίκτυο προβλέπει την πιθανότητα παρουσίας μια άρθρωσης σε κάθε pixel. Η πλήρης δομή (εκτός από τα τελικά επίπεδα 1×1) απεικονίζεται στη εικόνα 20.



Εικόνα 20: Απεικόνιση μια δομής ''κλεψύδρας''. Κάθε κουτί στην εικόνα αντιστοιχεί σε residual μονάδα όπως φαίνεται στην εικόνα 21, Πηγή: arxiv.org

3.4.2 Υλοποίηση επιπέδου

Ενώ διατηρείται το συνολικό σχήμα κλεψύδρας, εξακολουθεί να υπάρχει ευελιξία στην συγκεκριμένη εφαρμογή των επιπέδων. Διαφορετικές επιλογές μπορεί να έχουν μέτριο αντίκτυπο σχετικά με τις επιδόσεις και την κατάρτιση του δικτύου. Διερευνώνται διάφορες επιλογές για το σχεδιασμό επιπέδου στο δίκτυο μας. Οι πρόσφατες εργασίες έδειξαν την αξία της μείωσης των βημάτων με 1×1 συνελίξεις, καθώς και τα πλεονεκτήματα της χρήσης διαδοχικών μικρότερων φίλτρων για να καταγράψουν ένα μεγαλύτερο χωρικό περιβάλλον. [129,130] Για παράδειγμα, μπορεί κανείς να αντικαταστήσει ένα 5×5 φίλτρο με δύο ξεχωριστά 3×3 φίλτρα. Δοκιμάστηκε ο συνολικός σχεδιασμός του δικτύου μας, με εναλλαγή σε διαφορετικές μονάδες επίπεδου που βασίζονται εκτός λειτουργίας αυτών των πληροφορίων. Βιώνεται μια αύξηση της απόδοσης του δικτύου μετά την μετάβαση από την πρότυπη συνέλιξη επιπέδων με μεγάλα φίλτρα και χωρίς βήματα μείωσης σε νεότερες μεθόδους αυτή του He και των υπολοίπων [130] και τα σχέδια βασισμένα στην ΄΄έναρξη΄΄ [129]. Μετά την αρχική βελτίωση των επιδόσεων με αυτούς τους τύπους σχεδίων, πρόσθετες εξερευνήσεις και τροποποιήσεις στα επίπεδα δεν έκαναν πολλά για να ενισχύσουν περαιτέρω την απόδοση ή τον χρόνο εκπαίδευσης.



Εικόνα 21: (Αριστερά): Εναπομένουσα μονάδα που χρησιμοποιείτε σε όλο το δίκτυο (Δεξιά): Απεικόνιση της ενδιάμεσης διαδικασίας εποπτείας. Το δίκτυο χωρίζει και παράγει ένα σύνολο χαρτών θερμότητας (περιγράφεται με μπλεμχρώμα) όπου μπορεί να εφαρμοστεί η απώλεια, Πηγή: arxiv.org

Ο τελικός σχεδιασμός κάνει εκτεταμένη χρήση των υπολειμματικών ενοτήτων. Φίλτρα μεγαλύτερα από 3×3 δεν χρησιμοποιούνται ποτέ και η συμφόρηση περιορίζει τον συνολικό αριθμό παραμέτρων σε κάθε επίπεδο που περιορίζει την συνολική χρήση μνήμης. Η μονάδα που χρησιμοποιείται στο δίκτυο μας εμφανίζεται στην εικόνα 21. Για να τεθεί αυτό το πλαίσιο του πλήρους σχεδιασμού του δικτύου, κάθε πλαίσιο στην εικόνα 20 αντιπροσωπεύει μια μόνο εναπομένουσα μονάδα.

Η λειτουργία με την πλήρη ανάλυση εισόδου 256×256 απαιτεί ένα σημαντικό ποσό απαιτεί ένα σημαντικό ποσό μνήμης GPU, έτσι ώστε η υψηλότερη ανάλυση της κλεψύδρας (και έτσι η τελική ανάλυση εξόδου) είναι 64×64. Αυτό δεν επιδρά την ικανότητα του δικτύου να παράγει ακριβείς κοινές προβλέψεις. Το πλήρες δίκτυο ξεκινά ένα 7×7 συνελικτικό επίπεδο με διασκελισμό 2, ακολουθούμενη από μια υπολειμματική μονάδα και ένα γύρο μέγιστης συγκέντρωσης για να μειωθεί η ανάλυση από 256 σε 64. Δύο επόμενες εναπομένουσες ενότητες προηγούνται της κλεψύδρας που φαίνεται στην εικόνα 20. Σε ολόκληρη την κλεψύδρα όλες οι εναπομένουσες ενότητες παράγουν έξοδο 256 χαρακτηριστικών.

3.4.3 Στοιβασμένη κλεψύδρα με ενδιάμεση εποπτεία

Η αρχιτεκτονική του δικτύου πηγαίνει ένα βήμα περαιτέρω με την στοίβαξη πολλαπλών κλεψυδρών τέλος με τέλος, τροφοδοτώντας έτσι την έξοδο του ενός ως είσοδο του επόμενου. Αυτό παρέχει το δίκτυο με ένα μηχανισμό για επαναλαμβανόμενα από κάτω προς τα πάνω, από πάνω προς τα κάτω επιτρεπόμενα συμπεράσματα για την επαναξιολόγηση των αρχικών εκτιμήσεων και χαρακτηριστικών σε ολόκληρη την εικόνα. Το κλειδί αυτής της προσέγγισης είναι η πρόβλεψη ενδιάμεσων χαρτών θερμότητας στα οποία μπορεί να εφαρμοστεί μια απώλεια. Οι προβλέψεις δημιουργούνται μετά την διέλευση από κάθε κλεψύδρα όπου το δίκτυο είχε την ευκαιρία να επεξεργαστεί τα χαρακτηριστικά και στα δύο τοπικά και καθολικά πλαίσια. Οι επόμενες ενότητες κλεψύδρας επιτρέπουν σε αυτά τα χαρακτηριστικά υψηλού επιπέδου να υποβληθούν ξανά σε επεξεργασία για την περαιτέρω αξιολόγηση και επανεκτίμηση των χωρικών σχέσεων υψηλότερης τάξης. Αυτό είναι παρόμοιο με άλλες μεθόδους εκτίμησης της στάσης του σώματος που έχουν επιδείξει ισχυρή απόδοση με πολλαπλά επαναληπτικά στάδια και ενδιάμεση επίβλεψη [128,131,132].

Εξετάζονται τα όρια της εφαρμογής ενδιάμεσης εποπτείας με την χρήση μόνο μια ενιαίας μονάδας κλεψύδρας. Οι περισσότερες διαταγές υψηλών χαρακτηριστικών παρουσιάζονται μόνο σε χαμηλότερες αναλύσεις εκτός απο το τέλος, όταν πραγματοποιείται η αναδειγματοληψία. Εάν η εποπτεία παρέχεται μετά την από την στιγμή που το δίκτυο πραγματοποιεί αναδειγματοληψία τότε δεν υπάρχει τρόπος να επαναξιολογηθούν αυτές τα χαρακτηριστικά σε σχέση με τα υπόλοιπα σε ένα ευρύτερο καθολικό πλαίσιο. Εάν θέλουν το δίκτυο να βελτιώσει καλύτερα τις προβλέψεις του, τότε αυτές οι προβλέψεις δεν μπορούν να αξιολογηθούν αποκλειστικά σε τοπική κλίμακα. Η σχέση με άλλες κοινές προβλέψεις, καθώς και το γενικό πλαίσιο και η κατανόηση της πλήρους εικόνας είναι κρίσιμη. Η εφαρμογή της εποπτείας νωρίτερα από στον αγωγό διοχέτευσης πριν από την συγκέντρωση είναι μια δυνατότητα, αλλά σε αυτό το σημείο τα χαρακτηριστικά σε ένα δεδομένο pixel είναι το αποτέλεσμα της επεξεργασίας ενός σχετικά τοπικού δεκτικού τομέα και έτσι αγνοούν τις κρίσιμες καθολικές υποδείξεις.

Επαναλαμβανόμενα από κάτω προς τα πάνω, από πάνω προς τα κάτω συμπεράσματα με στοιβασμένες κλεψύδρες ανακουφίζουν από τις παραπάνω ανησυχίες. Οι τοπικές και οι καθολικές υποδείξεις ενσωματώνονται σε κάθε μονάδα κλεψύδρας και ζητώντας

από το δίκτυο να παράγει τις πρώτες προβλέψεις απαιτεί να έχει μια κατανόηση υψηλού επιπέδου της εικόνας, ενώ μόνο εν μέρη μέσω του πλήρους δικτύου. Τα επόμενα στάδια της επεξεργασίας από κάτω προς τα πάνω, από κάτω προς τα πάνω γίνεται επανεξέταση αυτών των χαρακτηριστικών.

Αυτή η προσέγγιση είναι πολύ σημαντική για να μετακινηθεί μεταξύ των κλιμάκων επειδή η διατήρηση της χωρικής θέσης των χαρακτηριστικών γνωρισμάτων είναι ουσιαστικής σημασίας για να κάνει το τελικό βήμα εντοπισμού. Η ακριβής θέση μια άρθρωσης αποτελεί απαραίτητο βήμα για τις αποφάσεις που λαμβάνονται από το δίκτυο. Με ένα δομημένο πρόβλημα όπως η εκτίμησης της στάσης του σώματος, η έξοδος είναι μια αλληλεπίδραση πολλών διαφορετικών χαρακτηριστικών γνωρισμάτων που πρέπει να έρθουν μαζί για να σχηματίσουν μια συνεκτική κατανόηση της σκηνής. Αντιφατικά αποδεικτικά στοιχεία και η ανατομική είναι μεγάλα δώρα που μας δείχνουν ότι κατά μήκος της γραμμής έγινε κάποιο λάθος και με την μετάβαση πέρα-δώθε το δίκτυο μπορεί να διατηρήσει ακριβείς τοπικές πληροφορίες, ενώ εξετάζει στην συνέχεια και την συνολική συνοχή των χαρακτηριστικών.

Στην συνέχεια επανεντάσσονται οι ενδιάμεσες προβλέψεις πίσω στο χώρο χαρακτηριστικών γνωρισμάτων με την χαρτογράφηση τους σε ένα μεγαλύτερο αριθμό καναλιών με μια πρόσθετη συνέλιξη 1×1. Αυτά προστίθενται πίσω στα ενδιάμεσα χαρακτηριστικά γνωρίσματα από την κλεψύδρα μαζί με την παραγωγή χαρακτηριστικών γνωρισμάτων από το προηγούμενο στάδιο κλεψύδρας (απεικονίζεται στην εικόνα 21). Η έξοδος που προκύπτει χρησιμεύει άμεσα ως είσοδος για την ακολουθούμενη μονάδα κλεψύδρας που δημιουργεί ένα άλλο σύνολο προβλέψεων. Στο τελικό σχεδιασμό του δικτύου, χρησιμοποιούνται 8 κλεψύδρες. Είναι σημαντικό να σημειωθεί ότι τα βάρη δεν μοιράζονται σε όλες τις μονάδες της κλεψύδρας και μια απώλεια εφαρμόζεται στις προβλέψεις όλων των κλεψυδρών χρησιμοποιώντας το ίδιο έδαφος αληθείας.

3.5 Συνελικτίκες μηχανές στάσης (CPM)

Στην ενότητα αυτή παρουσιάζεται η μέθοδος των συνελικτικών μηχανών στάσης (CPM) για την εκτίμηση της στάσης του ανθρώπινου σώματος. Τα CPMs κληρινομούν τα οφέλη της αρχιτεκτονικής της μηχανής στάσης [133] και τα συνδιάζει με τα πλεονεκτήματα που παρέχονται από τις συνελικτικές αρχιτεκτονικές: την δυνατότητα να μάθουν χαρακτηριστικές αναπαραστάσεις και για την εικόνα και για το χωρικό πλαίσιο απευθείας από τα δεδομένα, την διαφορετική αρχιτεκτονική που επιτρέπει την καθολική κοινή κατάρτιση και την ικανότητα να μπορούν να χειριστούν αποτελεσματικά μεγάλα εκπαιδευόμενα σύνολα δεδομένων.

Τα CPMs αποτελούνται από μια ακολουθία συνελικτικών δικτύων τα οποία παράγουν επανειλημμένα 2D χάρτες πεποιθήσεων για την θέση του κάθε μέρους. Σε κάθε ενός CPM, χρησιμοποιούνται χαρακτηριστικά εικόνας και χάρτες πεποιθήσεων που παρήχθησαν από το προηγούμενο στάδιο ως είσοδος. Οι χάρτες πεποιθήσεων παρέχουν στο επόμενο στάδιο μια εκφραστική μη παραμετρική κωδικοποίηση της χωρικής αβεβαιότητας θέσης για κάθε μέρος, επιτρέποντας στο CPM να μάθει πλούσια εξαρτόμενα από την εικόνα χωρικά μοντέλα των σχέσεων μεταξύ των μερών. Αντί να

αναλύουν ρητά τέτοιους χάρτες πεποιθήσεων είτε να χρησιμοποιούνται γραφικά μοντέλα [134,135,136] ή εξειδικευμένα βήματα μετά την επεξεργασία [135,137], μαθαίνονται συνελικτικά δίκτυα που λειτουργούν άμεσα με ενδιάμεσους χάρτες πεποιθήσεων και μαθαίνουν σιωπηρά εικονό-εξαρτώμενα χωρικά μοντέλα των σχέσεων μεταξύ των μερών. Η συνολική προτεινόμενη πολυβάθμια αρχιτεκτονική είναι πλήρως διαφορίσιμη και ως εκ τούτου μπορεί να εκπαιδευτεί.

3.5.1 Μηχανές στάσης

Δηλώνεται η θέση των pixel του p-th ανατομικού σημείου ενδιαφέροντος (το οποίο αναφέρεται ως μέρος), $Y_p \in \mathbb{Z} \subset \mathbb{R}^2$, όπου $\mathbb{Z}$ είναι το σύνολο όλων των (u, v) θέσεων σε μια εικόνα. Ο στόχος είναι να προβλέψουμε τις θέσεις της εικόνας $Y = (Y_1, \dots, Y_P)$ για όλα τα P μέρη. Μια μηχανή στάσης [133] (δες τις εικόνες 22α 22β) αποτελείται από ακολουθίες πολύ-κλάσεων προβλέψεων, $g_t(\cdot)$, που εκπαιδεύονται για να προβλέψουν την θέση κάθε τμήματος σε κάθε επίπεδο της ιεραρχίας. Σε κάθε επίπεδο $t \in \{1 \dots T\}$, οι ταξινομητές g_t προβλέπουν τις πεποιθήσεις για την αντιστοίχιση μιας θέσης σε κάθε μέρος $Y_p = z$, $\forall z \in \mathbb{Z}$, με βάση τα χαρακτηριστικά που εξάγονται από την εικόνα στην τοποθεσία z δηλώνεται από $x_z \in \mathbb{R}^d$ και συμφραζόμενες πληροφορίες από τον προηγούμενο ταξινομητη στην γειτονιά γυρώ από κάθε Y_p το στάδιο t. Ένας ταξινομητή στο πρώτο στάδιο t=1, ως εκ τούτου παράγει τις ακόλουθες τιμές πεποιθήσεων:

$$g_1(x_z) \rightarrow b_1^p(Y_p = z)_{p \in 0 \dots P} \quad (3.11)$$

Όπου $b_1^p(Y_p = z)$ είναι η βαθμολογία που προβλέπεται από τον ταξινομητη g_1 για την αντιστοίχιση του p^{th} μέρους στο πρώτο στάδιο στη θέση της εικόνας z. Εκπροσωπούμε όλες τις πεποιθήσεις του μέρους p που αξιολογείται σε κάθε θέση $z = (u,v)^T$ στην εικόνα ως $b_t^p \in \mathbb{R}^{w \times h}$, όπου w και h είναι το πλάτος και το ύψος της εικόνας, αντίστοιχα. Αυτό είναι :

$$b_t^p[u, v] = b_t^p(Y_p = z) \quad (3.12)$$

Για λογούς ευκολίας, υποδεικνύεται η συλλογή χαρτών πεποιθήσεων για όλα τα μέρη όπως $b_t \in \mathbb{R}^{w \times h \times (P+1)}$ (P μέρη συν ένα για φόντο).

Στα επόμενα στάδια, ο ταξινομητής προβλέπει την πεποίθηση για την ανάθεση μια θέσης σε κάθε μέρος $Y_p = z$, $\forall z \in \mathbb{Z}$, βασισμένη στα (3.11) χαρακτηριστικά των δεδομένων της εικόνας $x_z^t \in \mathbb{R}^d$ ξανά και (3.12) συμφραζόμενες πληροφορίες από τον ταξινομητή στην γειτονία γύρω από κάθε Y_p:

$$g_t(x'_z, \psi_t(z, b_{t-1})) \rightarrow b_t^p(Y_p = z)_{p \in 0 \dots P+1} \quad (3.13)$$

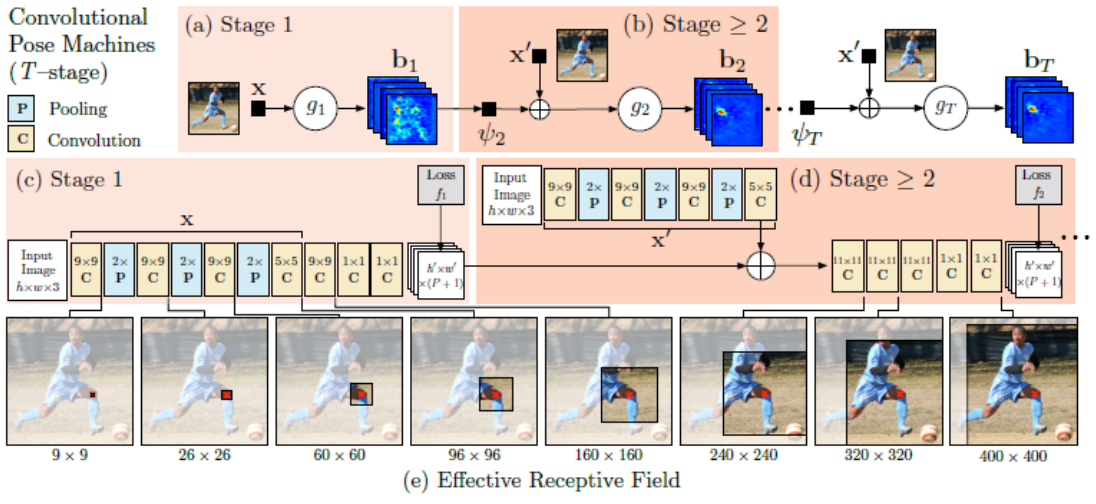
Όπου $\psi_{t>1}(\cdot)$ είναι μια χαρτογράφηση από τις πεποιθήσεις b_{t-1} στο πλαίσιο χαρακτηριστικών. Σε κάθε στάδιο, οι υπολογιζόμενες πεποιθήσεις παρέχουν μια ολοένα και πιο εκλεπτυσμένη εκτίμηση για την θέση κάθε μέρους. Σημειώστε ότι επιτρέπουμε τα χαρακτηριστικά εικόνας x'_z για το επόμενο στάδιο να είναι διαφορετικό

από το χαρακτηριστικό της εικόνας που χρησιμοποιείται στο πρώτο στάδιο x. Η μηχανή στάσης που προτείνεται στο [133] χρησιμοποιείται ενισχυμένες τυχαίες δομές για την πρόβλεψη του ($\{g_t\}$), σταθερά χειροποίητα χαρακτηριστικά εικόνας σε όλα τα στάδια ($x' = x$) και σταθερό χειροποίητο περιβάλλον χαρτών δυνατοτήτων ($\psi_t(\cdot)$) για να καταγράψει το χωρικό πλαίσιο σε όλα τα στάδια.

3.5.2 Μέθοδος συνελικτικών μηχανών στάσης

Δείχνεται πως η πρόβλεψη και ο υπολογισμός χαρακτηριστικών γνωρισμάτων μια εικόνας μηχανής στάσης μπορεί να αντικατασταθεί από μια βαθιά συνελικτική αρχιτεκτονική που επιτρέπει τόσο τις αναπαραστάσεις εικόνων όσο και συμφραζόμενων χαρακτηριστικών που πρέπει να μαθευτούν απευθείας από τα δεδομένα. Οι συνελικτικές αρχιτεκτονικές έχουν επίσης το πλεονέκτημα ότι είναι εντελως διαφορισιμες, επιτρέποντας έτσι να τεθεί τέλος με τέλος κοινή εκπαίδευση όλων των σταδίων ενός CPM. Περιγράφεται το σχέδιο για ένα CPM που συνδυάζει τα πλεονεκτήματα των βαθιών συνελικτικών αρχιτεκτονικών με την έμμεση χωρική μοντελοποίηση που παρέχεται από το πλαίσιο μηχανών στάσης.

Το πρώτο στάδιο μιας συνελικτικής μηχανής στάσης προβλέπει τις πεποιθήσεις του μέρους από μόνο τοπικά στοιχεία της εικόνας. Η εικόνα 22c δείχνει την δομή του δικτύου που χρησιμοποιείται για τον εντοπισμό μέρους από την στοιχεία τοπικής εικόνας που χρησιμοποιούν ένα βαθύ συνελικτικό δίκτυο. Τα στοιχεία είναι τοπικά επειδή το δεκτικό πεδίο του πρώτου σταδίου του δικτύου περιορίζεται σε μια μικρή ενημερωμένη έκδοση κώδικά γύρω από την θέση pixel της εξόδου. Χρησιμοποιείται μια δομή δικτύου που αποτελείται από 5 συνελικτικά επίπεδα ακολουθούμενα από 2 1×1 συνελικτικά δίκτυα, που έχουν ως αποτέλεσμα μια πλήρη συνελικτική αρχιτεκτονική [138]. Στην πράξη, για να επιτύχουμε κάποια ακρίβεια, ομαλοποιούμε την είσοδο εικόνων περικοπής στο μέγεθος 368×368 και το δεκτικό πεδίο του δικτύου που φαίνεται παραπάνω είναι 160×160 pixel. Το δίκτυο μπορεί να θεωρηθεί σαν να σέρνουμε ένα βαθύ δίκτυο σε μια εικόνα και να οπισθοδρομούμε από τα τοπικά αποδεικτικά στοιχεία κάθε ενημερωμένης έκδοσης κώδικα εικόνας 160×160 σε ένα P+1 μεγέθους διάνυσμα εξόδου το οποίο αναπαριστά μια βαθμολογία για κάθε μέρος σε αυτή την θέση της εικόνας.



(e) Effective Receptive Field

Εικόνα 22: Αρχιτεκτονικοί και δεκτικοί τομείς των CPMs. Παρουσιάζουμε μια συνελικτική αρχιτεκτονική και δεκτικά πεδία σε όλα τα επίπεδα για ένα CPM με οποιαδήποτε στάδια T. Η μηχανή στάσης [29] εμφανίζεται σε ένθετα (a) και (b) και τα αντίστοιχα δίκτυα συνέλιξης εμφανίζονται σε ένθετα (c) και (d). Τα ένθετα (a) και (c) εμφανίζουν την αρχιτεκτονική που λειτουργεί μόνο με βάση τα στοιχεία εικόνας στο πρώτο στάδιο. Ένθετα (b) και (d) δείχνουν την αρχιτεκτονική για τα επόμενα στάδια, τα οποία λειτουργούν τόσο σε αποδεικτικά στοιχεία εικόνας, καθώς και σε χάρτες πεποιθήσεων από τα προηγούμενα στάδια. Οι αρχιτεκτονικές στα στοιχεία (b) και (d) επαναλαμβάνονται για όλα τα επόμενα στάδια (2 έως T). Το δίκτυο εποπτεύεται τοπικά μετά από κάθε στάδιο χρησιμοποιώντας ένα ενδιάμεσο επίπεδο απώλειας που αποτρέπει την εξαφάνιση των κλίσεων κατά τη διάρκεια της εκπαίδευσης. Παρακάτω στο ένθετο (e) δείχνουμε το αποτελεσματικό δεκτικό πεδίο σε μια εικόνα (με κέντρο στο αριστερό γόνατο) της αρχιτεκτονικής, όπου το μεγάλο δεκτικό πεδίο επιτρέπει στο μοντέλο να συλλάβει μεγάλης εμβέλειας χωρικές εξαρτήσεις, όπως αυτές μεταξύ του κεφαλιού και των γονάτων. (Καλύτερη προβολή με χρώμα.)

3.5.3 Εκπαίδευση στις συνελικτικές μηχανές στάσης

Ο σχεδιασμός περιγράφεται παραπάνω για τα αποτελέσματα μιας μηχανής στάσης σε μια βαθιά αρχιτεκτονική τα οποία μπορούν να έχουν ένα μεγάλο αριθμό επιπέδων. Η εκπαίδευση ενός τέτοιου δικτύου με πολλά στρώματα μπορεί να είναι επιρρεπής στο πρόβλημα της εξαφάνισης κλίσεων [135,140,141] όπου όπως παρατηρήθηκε από τον Bradley [140] και τον Bengio και τους υπολοίπους [141], το μέγεθος των εκ των πίσω αναπαραχθέντων κλίσεων μειώνεται σε δύναμη με τον αριθμό των ενδιάμεσων επιπέδων μεταξύ του επιπέδου εξόδου και του επιπέδου εισόδου.

Ευτυχώς, το πλαίσιο διαδοχικής πρόβλεψης της μηχανής στάσης παρέχει μια φυσική προσέγγιση για την εκπαίδευση της βαθιάς αρχιτεκτονικής που αντιμετωπίζει αυτό το πρόβλημα. Κάθε στάδιο της μηχανής στάσης εκπαιδεύεται για να παράγει επανειλημμένα τους χάρτες πεποιθήσεως για τις θέσεις καθενός από τα μέρη. Ενθαρρύνεται το δίκτυο να καταλήξει επανειλημμένα σε μια τέτοια εκπροσώπηση, καθορίζοντας μια λειτουργία απώλειας στην παραγωγή κάθε σταδίου t που ελαχιστοποιεί την l_2 απόσταση ανάμεσα των προβλεπόμενων και των ιδανικών χαρτών πεποιθήσεων για κάθε μέρος. Ο ιδανικός χάρτης πεποίθησης για ένα μέρος p γράφεται ως $b_*^p(Y_p = z)$, η οποία δημιουργείται τοποθετώντας Gaussian κορυφές στις θέσεις εδάφους αληθείας για κάθε μέρος του σώματος p. Η συνάρτηση κόστους που στοχεύετέ να ελαχιστοποιηθεί στην έξοδο κάθε σταδίου κάθε επιπέδου δίνεται από:

$$f_t = \sum_{p=1}^{P+1} \sum_{z \in \mathbb{Z}} \left\| b_t^p(z) - b_*^p(z) \right\|_2^2 \quad (3.14)$$

Ο συνολικός στόχο για την πλήρη αρχιτεκτονική επιτυγχάνεται προσθέτοντας τις απώλειες σε κάθε στάδιο και δίνεται από την σχέση:

$$\mathcal{F} = \sum_{t=1}^{T} f_t \quad (3.15)$$

Χρησιμοποιείται η τυποποιημένη στοχαστική κλίση για να εκπαιδευτούν από κοινού όλα τα στάδια T στο δίκτυο. Για να μοιραστούν τα χαρακτηριστικά εικόνας x΄ σε όλα τα επόμενα στάδια, μοιράζονται τα βάρη των αντίστοιχων συνελικτικών επιπέδων (εικόνα 22) σε όλα τα στάδια $t \geq 2$.

ΚΕΦΑΛΑΙΟ 4

Κυριότερες μέθοδοι εκτίμησης της στάσης του σώματος πολλαπλών ατόμων με χρήση της τεχνολογίας βαθιάς μάθησης

4.1 Η μέθοδος Openpose

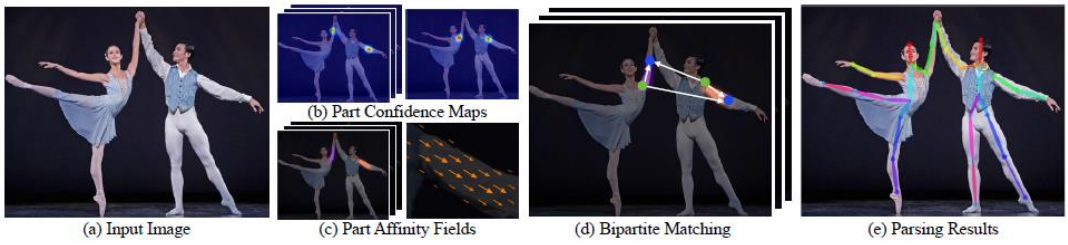
Σε αυτή την εργασία, παρουσιάζεται μια αποτελεσματική μέθοδος για εκτίμηση της στάσης πολλαπλών ατόμων με ανταγωνιστική απόδοση σε πολλά δημόσια κριτήρια αξιολόγησης. Παρουσιάζεται η πρώτη αναπαράσταση των αποτελεσμάτων συσχέτισης μέσω πεδίων συνάφειας μερών σώματος (Part Affinity Fields), ενός συνόλου 2D διανυσματικών πεδίων που κωδικοποιούν τη θέση και τον προσανατολισμό των άκρων πάνω από τον τομέα της εικόνας (image domain). Δείχνεται ότι η ταυτόχρονη εξαγωγή συμπερασμάτων από αυτές τις από κάτω προς τα πάνω αναπαραστάσεις της ανίχνευσης και της συσχέτισης κωδικοποιεί επαρκές καθολικό πλαίσιο για μια «άπληστη ανάλυση» (greedy parse) για την επίτευξη αποτελεσμάτων υψηλής ποιότητας με πολύ μικρότερο υπολογιστικό κόστος.

Για την εκτίμηση της στάσης πολλαπλών ατόμων, οι περισσότερες προσεγγίσεις [142], [43], [144], [145], [146], [147], [148], [149], [150] έχουν χρησιμοποιήσει μια στρατηγική από πάνω προς τα κάτω που πρώτα ανιχνεύει ανθρώπους και στη συνέχεια εκτιμά τη στάση κάθε ατόμου ανεξάρτητα σε κάθε περιοχή που εντοπίστηκε. Αν και αυτή η στρατηγική καθιστά τις τεχνικές που αναπτύχθηκαν για την περίπτωση ενός ατόμου άμεσα εφαρμόσιμες, όχι μόνο πάσχει από την έγκαιρη δέσμευση για την ανίχνευση ατόμου, αλλά επίσης αποτυγχάνει να συλλάβει τις χωρικές εξαρτήσεις μεταξύ διαφορετικών ατόμων που απαιτούν καθολικά συμπεράσματα. Ορισμένες προσεγγίσεις έχουν αρχίσει να λαμβάνουν υπόψη τις εξαρτήσεις μεταξύ ατόμων. Οι Eichner et al. [151] επέκτειναν τις εικονογραφικές δομές λαμβάνοντας υπόψη ένα σύνολο αλληλεπιδρώντων ατόμων και η ταξινόμηση του βάθους, αλλά έχοντας την απαίτηση ένας ακόμη ανιχνευτής ατόμων να εκκινήσει τις υποθέσεις ανίχνευσης. Οι Pishchulin et al. [152] πρότεινε μια προσέγγιση από κάτω προς τα πάνω, η οποία χαρακτηρίζει από κοινού τους υποψήφιους ανίχνευσης μερών και τους συσχετίζει με μεμονωμένα άτομα, με βαθμολογίες ανά ζεύγη βάσει ανάλυσης παλινδρόμησης από τις χωρικές μετατοπίσεις των ανιχνευόμενων μερών. Αυτή η προσέγγιση δεν βασίζεται σε ανιχνεύσεις ατόμων, ωστόσο, η επίλυση του προτεινόμενου ακέραιου γραμμικού προγραμματισμού επί του πλήρως συνδεδεμένου γραφήματος είναι ένα πρόβλημα NP-hard και, επομένως, ο μέσος χρόνος επεξεργασίας για μία μόνο εικόνα είναι πάνω από

μία ώρα. Οι Insafutdinov et al. [153] βασίστηκαν στα προηγούμενα [152] αλλά χρησιμοποίησαν ισχυρότερους ανιχνευτές μερών με βάση το ResNet [154] και βαθμολογίες ανά ζεύγη που εξαρτώνται από εικόνες, βελτιώνοντας σε μεγάλο βαθμό τον χρόνο εκτέλεσης με μια σταδιακή προσέγγιση βελτιστοποίησης, αλλά η μέθοδος διαρκεί αρκετά λεπτά ανά εικόνα, με όριο 150 το πολύ προτάσεων μερών. Οι παραστάσεις ανά ζεύγη που χρησιμοποιούνται στο [153], οι οποίες είναι διανύσματα μετατόπισης μεταξύ κάθε ζεύγους μερών του σώματος, είναι δύσκολο να αναλυθούν παλινδρομικά με ακρίβεια και, επομένως, απαιτείται ξεχωριστή λογιστική ανάλυση παλινδρόμησης για να μετατραπούν τα ανά ζεύγη χαρακτηριστικά σε βαθμολογία πιθανότητας.

4.1.1 Ανάλυση της μεθόδου

Η εικόνα 23 απεικονίζει τη συνολική διοχέτευση της μεθόδου. Το σύστημα λαμβάνει, ως εισαγόμενο δεδομένο, μια έγχρωμη εικόνα μεγέθους $w \times h$ (εικόνα 23α) και παράγει τις 2D θέσεις των ανατομικών σημείων-κλειδιών για κάθε άτομο στην εικόνα (εικόνα 23ε). Πρώτον, ένα τροφοδοτικό δίκτυο προβλέπει ένα σύνολο 2D χαρτών αξιοπιστίας **S** των θέσεων των μερών του σώματος (εικόνα 23β) και ένα σύνολο 2D διανυσματικών πεδίων L των πεδίων συνάφειας μερών (PAFs), τα οποία κωδικοποιούν τον βαθμό συσχέτισης μεταξύ των μερών (εικόνα 23γ). Το σύνολο $S = (S_1, S_2, ..., S_J)$ έχει χάρτες αξιοπιστίας J, έναν ανά μέρος, όπου $S_j \in \mathbb{R}^{w \times h}$, $j \in \{1…J\}$. Το σύνολο $L = (L_1, L_2, ..., L_C)$ έχει διανυσματικά πεδία C, ένα ανά άκρο, όπου $L_c \in \mathbb{R}^{w \times h \times 2}$, $c \in \{1…C\}$. Αναφέρονται σε ζεύγη μερών ως άκρα για ευκρίνεια, αλλά μερικά ζεύγη δεν είναι ανθρώπινα άκρα (π.χ. το πρόσωπο). Κάθε θέση εικόνας στο L_c κωδικοποιεί ένα 2D διάνυσμα. Τέλος, οι χάρτες αξιοπιστίας και τα PAFs αναλύονται βάσει «άπληστης» συναγωγής συμπερασμάτων (εικόνα 23δ) για την εξαγωγή των 2D σημείων για όλα τα άτομα στην εικόνα.

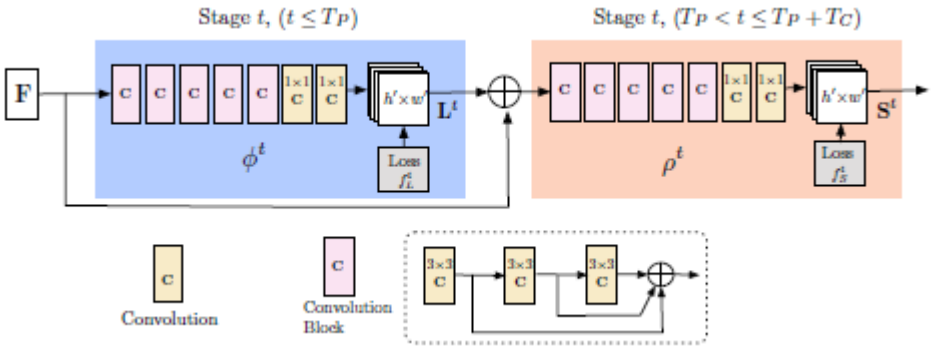


Εικόνα 23: Συνολικός αγωγός διοχέτευσης της μεθόδου, Πηγή: deeplearning.vn

4.1.2 Αρχιτεκτονική του δικτύου

Η αρχιτεκτονική, που φαίνεται στην εικόνα 24, προβλέπει επαναλαμβανόμενα πεδία συνάφειας που κωδικοποιούν τη συσχέτιση μεταξύ των μερών, που εμφανίζονται με μπλε χρώμα και χάρτες αξιοπιστίας ανίχνευσης σε μπεζ χρώμα. Η επαναλαμβανόμενη αρχιτεκτονική πρόβλεψης, κατά το [155], βελτιώνει τις προβλέψεις σε διαδοχικά στάδια, $t \in \{1,…,T\}$, με ενδιάμεση επίβλεψη σε κάθε στάδιο.

Το βάθος δικτύου αυξάνεται σε σχέση με το [156]. Στην αρχική προσέγγιση, η αρχιτεκτονική του δικτύου περιελάβανε αρκετά συνελικτικά επίπεδα 7×7. Στο τρέχον μοντέλο, το δεκτικό πεδίο διατηρείται ενώ ο υπολογισμός μειώνεται, αντικαθιστώντας κάθε συνελικτικό πυρήνα 7×7 με 3 διαδοχικούς 3×3 πυρήνες. Ενώ ο αριθμός των λειτουργιών για το πρώτο είναι $2 \times 7^2 - 1 = 97$, για το δεύτερο, είναι μόνο 51. Επιπλέον, η έξοδος κάθε ενός από τους 3 συνελικτικούς πυρήνες συνδέεται σειριακά, ακολουθώντας μια προσέγγιση παρόμοια με το DenseNet [157]. Ο αριθμός των επιπέδων μη-γραμμικότητας τριπλασιάζεται και το δίκτυο μπορεί να διατηρήσει τα χαρακτηριστικά τόσο του χαμηλότερου επιπέδου όσο και του υψηλότερου επιπέδου.



Εικόνα 24: Η αρχιτεκτονική του πολυεπίπεδου CNN, Πηγή: sementicscholar.org

4.1.3 Ταυτόχρονη ανίχνευση και συσχέτιση

Η εικόνα αναλύεται από ένα συνελικτικό νευρωνικό δίκτυο (CNN) (το οποίο έχει εκκινήσει από τα πρώτα 10 στρώματα του VGG-19 [158] και έχει προσαρμοστεί με ακρίβεια), δημιουργώντας ένα σύνολο χαρτών χαρακτηριστικών F που εισάγονται στο πρώτο στάδιο. Σε αυτό το στάδιο, το δίκτυο παράγει ένα σύνολο πεδίων συνάφειας μερών (PAFs) $L^1 = \varphi^1$ (F), όπου το φ^1 αναφέρεται στα CNN για εξαγωγή συμπερασμάτων στο στάδιο 1. Σε κάθε επόμενο στάδιο, οι προβλέψεις από το προηγούμενο στάδιο και το τα αρχικά χαρακτηριστικά εικόνας F συνδέονται σειριακά και χρησιμοποιούνται για την παραγωγή βελτιστοποιημένων προβλέψεων,

$$L^t = \varphi^t(F, L^{t-1}), \forall\, 2 \leq t \leq T_p \quad (4.1)$$

όπου το φ^t αναφέρεται στα CNN προς εξαγωγή συμπερασμάτων στο Στάδιο t, και το T_P στον αριθμό των συνολικών σταδίων PAF. Μετά τις επαναλήψεις T_P, η διαδικασία επαναλαμβάνεται για την ανίχνευση χαρτών αξιοπιστίας, ξεκινώντας από την πιο πρόσφατη πρόβλεψη PAF,

$$S^{T_p} = p^t(F, L^{T_p}), \forall t = T_p \quad (4.2)$$
$$S^t = p^t(F, L^{T_p}, S^{t-1}), \forall\, T_p < t \leq T_p + T_c \quad (4.3)$$

όπου το ρ^t αναφέρεται στα CNN προς εξαγωγή συμπερασμάτων στο στάδιο t, και το T_C στον αριθμό των συνολικών σταδίων των χαρτών αξιοπιστίας.

Αυτή η προσέγγιση διαφέρει από το [155], όπου και οι δύο κλάδοι του PAF και του χάρτη αξιοπιστίας βελτιστοποιήθηκαν σε κάθε στάδιο. Ως εκ τούτου, ο υπολογισμός ανά στάδιο μειώνεται κατά το ήμισυ. Παρατηρούμε εμπειρικά ότι οι βελτιστοποιημένες προβλέψεις πεδίου συνάφειας βελτιώνουν τα αποτελέσματα του χάρτη αξιοπιστίας, ενώ το αντίστροφο δεν ισχύει. Διαισθητικά, αν κοιτάξει κανείς την έξοδο καναλιού PAF, μπορεί να μαντέψει κανείς τις θέσεις των μερών του σώματος. Ωστόσο, εάν δουν ένα σωρό από μέρη του σώματος χωρίς άλλες πληροφορίες, δεν μπορούν να τα αναλύσουν σε διαφορετικά άτομα.

Η εικόνα 25 δείχνει τη βελτίωση των πεδίων συνάφειας σε διάφορα στάδια. Τα αποτελέσματα του χάρτη αξιοπιστίας προβλέπονται πάνω από

Εικόνα 25: PAFs του δεξιού πήχη σε διάφορα στάδια. Παρόλο που υπάρχει σύγχυση μεταξύ αριστερών και δεξιών μερών του σώματος και των άκρων σε πρώιμα στάδια, οι εκτιμήσεις βελτιώνονται όλο και περισσότερο μέσω της εξαγωγής καθολικών συμπερασμάτων σε μεταγενέστερα στάδια**, Πηγή: semanticscholar.org**

Η εικόνα 25 δείχνει τη βελτίωση των πεδίων συνάφειας σε διάφορα στάδια. Τα αποτελέσματα του χάρτη αξιοπιστίας προβλέπονται πάνω από τις πιο πρόσφατες και πιο βελτιωμένες προβλέψεις PAF, και έχουν ως αποτέλεσμα μια σχεδόν ανεπαίσθητη διαφορά στα στάδια του χάρτη αξιοπιστίας. Για να καθοδηγηθεί το δίκτυο να προβλέψει επαναλαμβανόμενα τα PAF των μερών του σώματος στον πρώτο κλάδο και τους χάρτες αξιοπιστίας στον δεύτερο κλάδο, εφαρμόζεται μια λειτουργία απώλειας στο τέλος κάθε σταδίου. Χρησιμοποιείται μια απώλεια L_2 μεταξύ των εκτιμώμενων προβλέψεων και των πραγματικών χαρτών και πεδίων. Εδώ, σταθμίζονται οι λειτουργίες απώλειας χωρικά για να αντιμετωπιστεί ένα πρακτικό ζήτημα: το γεγονός ότι ορισμένα σύνολα δεδομένων δεν χαρακτηρίζουν πλήρως όλα τα άτομα. Συγκεκριμένα, η λειτουργία απώλειας του κλάδου PAF στο στάδιο t_i και η λειτουργία απώλειας του κλάδου χάρτη αξιοπιστίας στο στάδιο t_k είναι:

$$f_L^{t_1} = \sum_{c=1}^{c} \sum_{p} W(p) \cdot \left\| L_c^{t_1}(p) - L_c^*(p) \right\|_2^2 \quad (4.4)$$

$$f_s^{t_k} = \sum_{j=1}^{j} \sum_{p} W(p) \cdot \left\| S_j^{t_k}(p) - S_j^*(p) \right\|_2^2 \quad (4.5)$$

όπου L_c^* είναι το πραγματικό PAF, το S_j^* είναι ο πραγματικός χάρτης αξιοπιστίας μέρους και το W είναι μια δυαδική μάσκα με W (p) = 0, όταν λείπει ο σχολιασμός στο pixel p. Η μάσκα χρησιμοποιείται για να αποφευχθεί η ποινικοποίηση των πραγματικών θετικών προβλέψεων κατά τη διάρκεια της εκπαίδευσης. Η ενδιάμεση επίβλεψη σε

κάθε στάδιο αντιμετωπίζει το πρόβλημα της εξαφάνισης των διανυσμάτων κλίσης (Vanishing Gradient Problem) αναπληρώνοντας την κλίση περιοδικά [155]. Ο γενικός στόχος είναι:

$$f = \sum_{t=1}^{T_p} f_L^t + \sum_{t=T_p+1}^{T_p+T_c} f_s^t \quad (4.6)$$

4.1.4 Χάρτες αξιοπιστίας και ανίχνευσης μερών

Για την αξιολόγηση του f_S στην εξίσωση (4.6) κατά τη διάρκεια της εκπαίδευσης, δημιουργούνται οι πραγματικούς χάρτες αξιοπιστίας S^* από τα σχολιασμένα 2D σημεία-κλειδιά. Κάθε χάρτης αξιοπιστίας αποτελεί μια 2D αναπαράσταση της πεποίθησης ότι ένα συγκεκριμένο μέρος του σώματος μπορεί να βρίσκεται σε οποιοδήποτε δεδομένο pixel. Ιδανικά, εάν ένα μοναδικό άτομο εμφανίζεται στην εικόνα, σε κάθε χάρτη αξιοπιστίας θα πρέπει να υπάρχει μία μοναδική κορυφή, εάν το αντίστοιχο μέρος είναι ορατό. Εάν υπάρχουν πολλά άτομα στην εικόνα, θα πρέπει να υπάρχει μια κορυφή που να αντιστοιχεί σε κάθε ορατό μέρος j για κάθε άτομο k.

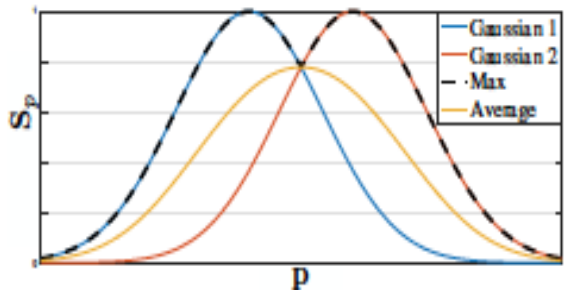
Δημιουργούμε πρώτα μεμονωμένους χάρτες αξιοπιστίας $S_{j,k}^*$για κάθε άτομο k. Έστω η πραγματική θέση του μέρους j του σώματος για το άτομο k στην εικόνα. Η τιμή στη θέση ορίζεται ως:

$$S_{j,k}^*(p) = exp\left(-\frac{\left\|p - x_{j,k}\right\|_2^2}{\sigma^2}\right) \quad (4.7)$$

όπου το σ ελέγχει την εξάπλωση της κορυφής. Ο πραγματικός χάρτης αξιοπιστίας που προβλέπεται από το δίκτυο είναι μία συνάθροιση των μεμονωμένων χαρτών αξιοπιστίας μέσω ενός μέγιστου χειριστή:

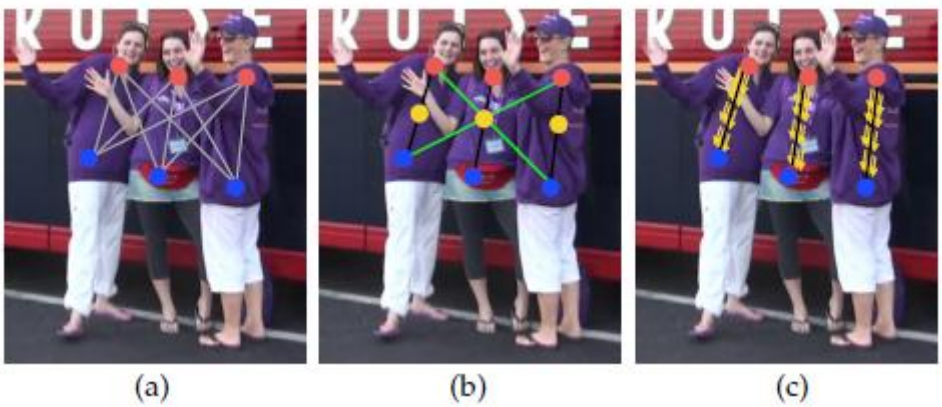
$$S_j^*(p) = \max_k S_{j,k}^*(p) \quad (4.8)$$

Λαμβάνεται το μέγιστο των χαρτών αξιοπιστίας αντί του μέσου όρου, έτσι ώστε η ακρίβεια των κοντινών κορυφών να παραμένει διακριτή. Κατά τη διάρκεια της δοκιμής, προβλέπονται χάρτες αξιοπιστίας και λαμβάνονται υποψήφια σημεία μερών του σώματος εκτελώντας τη διαδικασία non-maximum suppression (NMS).



4.1.5 Πεδία συνάφειας μερών για την ανίχνευση μερών

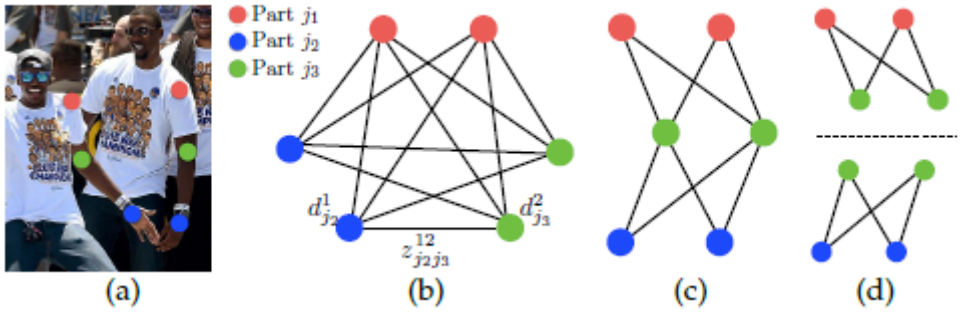
Λαμβάνοντας υπόψη ένα σύνολο ανιχνευμένων μερών του σώματος (εμφανίζονται ως κόκκινα και μπλε σημεία στην εικόνα 26α), στην συνέχεια θα αναλυθεί πώς συναρμολογούνται για να σχηματιστούν οι στάσεις ολόκληρου του σώματος ενός άγνωστου αριθμού ατόμων. Χρειάζεται ένα μέτρο αξιοπιστίας του συσχετισμού για κάθε ζεύγος ανιχνεύσεων μερών του σώματος, δηλαδή βεβαιότητας ότι ανήκουν στο ίδιο άτομο. Ένας πιθανός τρόπος μέτρησης της συσχέτισης είναι η ανίχνευση ενός επιπλέον μέσου σημείου μεταξύ κάθε ζεύγους τμημάτων ενός άκρου και ο έλεγχος της συχνότητάς του μεταξύ των υποψήφιων ανιχνεύσεων μερών, όπως φαίνεται στην εικόνα 26β. Ωστόσο, όταν οι άνθρωποι συνωστίζονται, όντας επιρρεπείς σε αυτό, αυτά τα μεσαία σημεία είναι πιθανό να οδηγήσουν σε λανθασμένες συσχετίσεις (εμφανίζονται ως πράσινες γραμμές στην εικόνα 26β). Τέτοιες λανθασμένες συσχετίσεις προκύπτουν λόγω δύο περιορισμών στην αναπαράσταση: (1) κωδικοποιεί μόνο τη θέση και όχι τον προσανατολισμό κάθε άκρου, (2) μειώνει την περιοχή στήριξης ενός άκρου σε ένα μόνο σημείο.



Εικόνα 26: : Στρατηγικές συσχέτισης μερών. (α) Τα υποψήφια σημεία ανίχνευσης μέρους του σώματος (κόκκινες και μπλε κουκκίδες) για δύο τύπους μερών του σώματος και όλα τα υποψήφια σημεία σύνδεσης (γκρίζες γραμμές). (β) Τα αποτελέσματα της σύνδεσης χρησιμοποιούν την αναπαράσταση του μέσου σημείου (κίτρινες κουκκίδες): σωστές συνδέσεις (μαύρες γραμμές) και λανθασμένες συνδέσεις (πράσινες γραμμές) που ικανοποιούν επίσης τον περιορισμό συχνότητας. (γ) Τα αποτελέσματα με χρήση PAF (κίτρινα βέλη). Με την κωδικοποίηση της θέσης και του προσανατολισμού πάνω στην στήριγμα του άκρου, τα PAF εξαλείφουν τις εσφαλμένες συσχετίσεις, **Πηγή: medium.com**

Τα πεδία συνάφειας μερών (PAF) αντιμετωπίζουν αυτούς τους περιορισμούς. Διατηρούν τόσο τις πληροφορίες θέσης και όσο και του προσανατολισμού σε όλη την περιοχή στήριξης του άκρου (όπως φαίνεται στην εικόνα. 26γ). Κάθε PAF είναι ένα διανυσματικό πεδίο 2D για κάθε άκρο. Για κάθε pixel στην περιοχή που ανήκει σε ένα συγκεκριμένο άκρο, ένα διάνυσμα 2D κωδικοποιεί την κατεύθυνση που δείχνει από το ένα μέρος του άκρου στο άλλο. Κάθε τύπος άκρου έχει ένα αντίστοιχο PAF που ενώνει τα δύο συνδεδεμένα μέρη του σώματος.

Για την αξιολόγηση του f_L στην εξίσωση (4.6) κατά τη διάρκεια της εκπαίδευσης, ορίζουμε το πραγματικό PAF, $L^*_{c,k}$, σε ένα σημείο p της εικόνας ως



Εικόνα. 27: Αντιπαραβολή γραφημάτων. (α) Αρχική εικόνα με ανιχνεύσεις μερών. (β) *K*-μερές γράφημα. (γ) Δενδροειδής δομή. (δ) Ένα σύνολο διμερών γραφημάτων, Πηγή: medium.com

$$L^*_{c,k}(p) = \begin{cases} v & ,\varepsilon\acute{\alpha}\nu\ p\ \sigma\tau o\ \acute{\alpha}\kappa\rho o\ c,k \\ 0 & ,\alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Εδώ $v = (x_{j2,k} - x_{j1,k})/\left\|x_{j2,k} - x_{j1,k}\right\|_2$ είναι το μοναδιαίο διάνυσμα προς την κατεύθυνση του άκρου. Το σύνολο των σημείων στο άκρο ορίζεται ως εκείνα που βρίσκονται εντός ορίου απόστασης του ευθύγραμμου τμήματος, δηλαδή εκείνα τα σημεία p για τα οποία:

$$0 \le v \cdot (p - x_{j1,k}) \le l_{c,k}\ \kappa\alpha\iota\ |v_\perp \cdot (p - x_{j1,t})| \le \sigma_l\ ,$$

όπου το πλάτος σ_l του άκρου $\mathbf{v}_\perp$είναι μια απόσταση σε pixel, το μήκος του άκρου είναι $l_{c,k} = \left\|x_{j2,k} - x_{j1,k}\right\|_2$και $\boldsymbol{v}_\perp$.είναι ένα διάνυσμα κάθετο στο v.

Το πραγματικό πεδίο συνάφειας μερών είναι ο μέσος όρος των πεδίων συνάφειας όλων των ατόμων στην εικόνα,

$$L^*_c(p) = \frac{1}{n_c(p)} \sum_k L^*_{c,k}\,(p) \quad (4.10)$$

όπου $n_c(p)$ είναι ο αριθμός των μη μηδενικών διανυσμάτων στο σημείο p σε όλα τα άτομα k.

Κατά τη διάρκεια της δοκιμής, μετριέται η συσχέτιση μεταξύ ανιχνεύσεων των υποψήφιων μερών υπολογίζοντας το επικαμπύλιο ολοκλήρωμα επί του αντίστοιχου PAF κατά μήκος του ευθύγραμμου τμήματος που συνδέει τις θέσεις των υποψήφιων μερών. Με άλλα λόγια, μετριέται η ευθυγράμμιση του προβλεπόμενου PAF με το υποψήφιο άκρο που θα σχηματιστεί συνδέοντας τα ανιχνευμένα μέρη του σώματος. Συγκεκριμένα, για δύο θέσεις d_{j1} και d_{j2} υποψήφιων μερών, δοκιμάζεται το προβλεπόμενο πεδίο συνάφειας μερών L_c κατά μήκος του ευθύγραμμου τμήματος για να μετρήσουμε την αξιοπιστία του συσχετισμού τους:

$$E = \int_{u=0}^{u=1} L_c(p(u)) \cdot \frac{d_{j2} - d_{j1}}{\left\|d_{j2} - d_{j1}\right\|_2} du \quad (4.11)$$

όπου το p(u) υπολογίζει με παρεμβολή τη θέση των δύο μερών του σώματος d_{j1} και d_{j2},

$$p(u) = (1-u) \cdot d_{j1} + u \cdot d_{j2} \ (4.12)$$

Στην πράξη, υπολογίζεται το ολοκλήρωμα με δειγματοληψία αθροίζοντας ομοιόμορφα διατεταγμένες στο χώρο τιμές του u.

4.1.6 Ανάλυση πολλαπλών ατόμων με χρήση PAFs

Εφαρμόζεται η διαδικασία non-maximum suppression (NMS) στους χάρτες αξιοπιστίας ανίχνευσης για να ληφθεί ένα διακριτό σύνολο υποψήφιων θέσεων των μερών. Για κάθε μέρος, ενδέχεται να έχουμε αρκετές υποψήφιες θέσεις, λόγω πολλαπλών ατόμων στην εικόνα ή ψευδών θετικών αποτελεσμάτων (FP) (εικόνα 27β). Αυτά τα υποψήφια μέρη ορίζουν ένα μεγάλο σύνολο πιθανών άκρων. Βαθμολογείται κάθε υποψήφιο άκρο χρησιμοποιώντας τον υπολογισμό επικαμπύλιου ολοκληρώματος στο PAF, που ορίζεται στην εξίσωση (4.11). Το πρόβλημα της εύρεσης της βέλτιστης ανάλυσης σχετίζεται με μια αντιστοίχιση *K*-διαστάσεων, πρόβλημα που είναι γνωστό ότι είναι NP-Hard [159] (εικόνα 27γ). Στην παρούσα εργασία, παρουσιάζεται μια «άπληστη» (greedy) χαλάρωση που παράγει με συνέπεια υψηλής ποιότητας αντιστοιχίσεις. Εικάζεται ότι ο λόγος είναι ότι οι βαθμολογίες συσχέτισης ανά ζεύγη κωδικοποιούν σιωπηρά το καθολικό πλαίσιο, λόγω του μεγάλου δεκτικού πεδίου του δικτύου PAF.

Επισήμως, λαμβάνεται πρώτα ένα σύνολο υποψηφίων D_J για ανίχνευση μερών του σώματος για πολλαπλά άτομα, όπου $D_J = \{d_j^m : για\ j \in \{1 \dots J\}, m \in \{1 \dots N_j\}\}$, όπου N_j είναι ο αριθμός των υποψηφίων θέσεων του μέρους j και $d_j^m \in \mathbb{R}^2$ είναι η θέση του υποψηφίου της m-ης ανίχνευσης του μέρους j σώματος. Αυτοί οι υποψήφιοι ανίχνευσης μερών πρέπει ακόμη να συσχετιστούν με άλλα μέρη του ίδιου ατόμου. Με άλλα λόγια, πρέπει να βρεθούν τα ζεύγη ανιχνεύσεων μερών που είναι στην πραγματικότητα συνδεδεμένα άκρα. Ορίζεται μια μεταβλητη $z_{j1j2}^{mn} \in \{0,1\}$ για να δειχθεί εάν δύο υποψήφιοι ανίχνευσης d_{j1}^m *και* d_{j2}^n είναι συνδεδεμένοι και ο στόχος είναι να βρούμε τη βέλτιστη ανάθεση για το σύνολο όλων των πιθανών συνδέσεων, $Z = \{z_{j1j2}^{mn} : για\ j_1, j_2 \in \{1, \dots, J\}, m \in \{1, \dots, N_{j1}\}, n \in \{1, \dots, N_{j2}\}$.

Εάν ληφθεί υπόψη ένα μοναδικό ζεύγος μερών j_1 και j_2 (π.χ., λαιμό και δεξί ισχίο) για το c-το άκρο, η εύρεση της βέλτιστης συσχέτισης περιορίζεται σε ένα πρόβλημα αντιστοίχισης διμερούς γραφήματος μέγιστης βαρύτητας [159]. Αυτή η περίπτωση φαίνεται στην εικόνα 26β. Σε αυτό το πρόβλημα αντιστοίχισης γραφήματος, οι κόμβοι του γραφήματος είναι οι υποψήφιοι ανίχνευσης μερών του σώματος D_{J1} και $D_{J2,}$ και οι άκρες είναι όλες οι πιθανές συνδέσεις μεταξύ των ζευγών των υποψηφίων ανίχνευσης. Επιπλέον, κάθε άκρο σταθμίζεται κατά την εξίσωση (4.11) - το άθροισμα συνάφειας μέρους. Μια αντιστοίχιση σε ένα διμερές γράφημα είναι ένα υποσύνολο των άκρων που έχουν επιλεγεί με τέτοιο τρόπο ώστε κανένα άκρο να μην μοιράζεται έναν κόμβο. Στόχος είναι να βρεθεί μια αντιστοίχιση με τη μέγιστη βαρύτητα για τις επιλεγμένες άκρες,

$$\max_{z_c} E_c = \max_{z_c} \sum_{m \in D_{J1}} \sum_{n \in D_{J2}} E_{mn} \cdot z_{j1j2}^{mn} \quad (4.13)$$

$$\forall\, m \in D_{j1}, \sum_{n \in D_{j2}} z_{j1j2}^{mn} \leq 1 \quad (4.14)$$

$$\forall\, n \in D_{j2}, \sum_{m \in Dj1} z_{j1j2}^{mn} \leq 1 \quad (4.15)$$

όπου το E_c είναι η συνολική βαρύτητα της αντιστοίχισης από τον τύπο c άκρου, το Z_c είναι το υποσύνολο του Z για τον τύπο c άκρου, και το E_{mn} είναι η συνάφεια μέρους μεταξύ των μερών d_{j1}^m και d_{j2}^n που ορίζεται στην εξίσωση (4.11). Οι εξισώσεις (4.14) και (4.15) επιβάλλουν ότι δύο άκρα δεν μοιράζονται έναν κόμβο, δηλαδή, δύο άκρα του ίδιου τύπου (π.χ. αριστερός πήχης) δεν μοιράζονται ένα μέρος. Μπορούμε να χρησιμοποιήσουμε τον ουγγρικό αλγόριθμο [160] για να έχουμε τη βέλτιστη αντιστοίχιση.

Όσον αφορά την εύρεση ολόκληρης της στάσης του σώματος πολλαπλών ατόμων, ο προσδιορισμός του Z είναι ένα πρόβλημα αντιστοίχισης K-διαστάσεων. Αυτό το πρόβλημα είναι NP-Hard [159] και υπάρχουν πολλές χαλαρώσεις. Στην εργασία αυτή, προστίθενται δύο χαλαρώσεις στη βελτιστοποίηση, ειδικές για τον τομέα. Πρώτον, επιλέγεται ένας ελάχιστος αριθμός άκρων για να λάβουμε μια ανεπτυγμένη δενδροειδή δομή ανθρώπινης στάσης αντί να χρησιμοποιηθεί το πλήρες γράφημα, όπως φαίνεται στην εικόνα. 27γ. Δεύτερον, αναλύεται περαιτέρω το πρόβλημα αντιστοίχισης σε ένα σύνολο υποπροβλημάτων διμερούς αντιστοίχισης και προσδιορίζεται η αντιστοίχιση σε γειτονικούς κόμβους δέντρων ανεξάρτητα, όπως φαίνεται στην εικόνα 27δ.

(α) Αρχική ανάλυση ατόμου (β) Πλεονάζουσα ανάλυση PAF

Εικόνα 28: Σημασία των πλεοναζουσών συνδέσεων PAF. (α) Δύο διαφορετικά άτομα συγχωνεύονται λανθασμένα λόγω λανθασμένης σύνδεσης του λαιμού με τη μύτη. (β) Η υψηλότερη αξιοπιστία της σύνδεσης του δεξιού αυτιού-ώμου αποφεύγει τη λανθασμένη σύνδεση μύτης-λαιμού, Πηγή: arxiv.org

Με αυτές τις δύο χαλαρώσεις, η βελτιστοποίηση αναλύεται απλώς ως εξής:

$$\max_{z} E = \sum_{c=1}^{c} \max_{z\ c} E_c \quad (4.16)$$

Λαμβάνονται, λοιπόν, οι υποψήφιοι σύνδεσης άκρων για κάθε τύπο άκρου ανεξάρτητα, χρησιμοποιώντας τις εξισώσεις. (4.13)-(4.15). Με όλους τους υποψηφίους σύνδεσης άκρων, μπορούν να συναρμολογήσουν οι συνδέσεις που μοιράζονται τα ίδια τα υποψήφια μέρη ανίχνευσης σε στάσεις πλήρους σώματος πολλαπλών ατόμων. Το σχήμα βελτιστοποίησης που προτείνεται επί δενδροειδούς δομής είναι κατά πολύ γρηγορότερο από τη βελτιστοποίηση επί πλήρως συνδεδεμένου γραφήματος [152], [153].

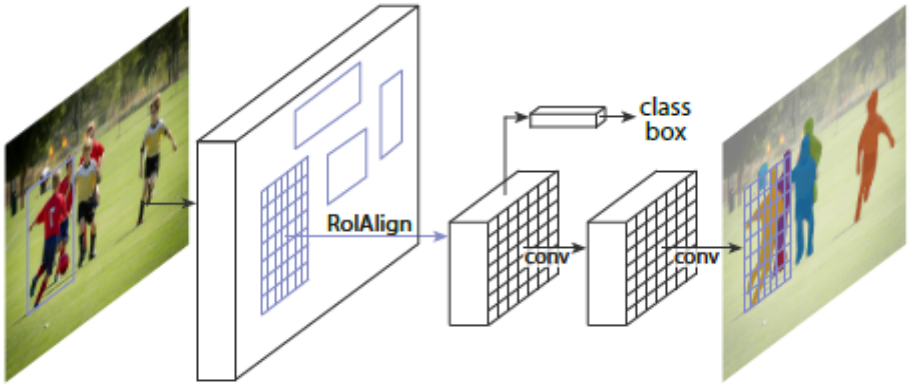
Το τρέχον μοντέλο ενσωματώνει επίσης πλεονάζουσες συνδέσεις PAF (π.χ. μεταξύ αυτιών και ώμων, καρπών και ώμων κλπ.). Αυτή η πλεονασματικότητα βελτιώνει ιδιαίτερα την ακρίβεια σε εικόνες συνωστισμού, όπως φαίνεται στην εικόνα. 28. Για να χειριστούν αυτές τις πλεονάζουσες συνδέσεις, τροποποιούν ελαφρώς τον αλγόριθμο ανάλυσης πολλαπλών ατόμων. Ενώ η αρχική προσέγγιση ξεκίνησε από ένα πρωταρχικό συστατικό, ο αλγόριθμός ταξινομεί όλες τις πιθανές συνδέσεις κατά ζεύγη βάσει της βαθμολογίας PAF τους. Εάν μια σύνδεση προσπαθεί να συνδέσει 2 μέρη του σώματος που έχουν ήδη ανατεθεί σε διαφορετικά άτομα, ο αλγόριθμος αναγνωρίζει ότι αυτό θα έρχεται σε αντίθεση με μια σύνδεση PAF με μεγαλύτερη αξιοπιστία, και στη συνέχεια αγνοείται η τρέχουσα σύνδεση

4.2 Η μέθοδος Mask R-CNN

Η μέθοδος που ονομάζεται MASK R-CNN, επεκτείνει το Faster R-CNN [161] προσθέτοντας ένα κλάδο για την πρόβλεψη των μασκών κατάτμησης σε κάθε Περιοχή Ενδιαφέροντος (RoI), παράλληλα με τον υπάρχοντα κλάδο για την ταξινόμηση και την κουτιών παλινδρόμησης (εικόνα 1). Ο κλάδος της μάσκας είναι ένα μικρό FCN που εφαρμόζεται σε κάθε RoI, προβλέποντας μια μάσκα κατάτμησης με τρόπο pixel-to-pixel. Η μάσκα R-CNN είναι απλή στην εφαρμογή και στην εκπαίδευση δεδομένου του Faster R-CNN πλαισίου, το οποίο διευκολύνει ένα ευρύ φάσμα ευέλικτων σχεδίων αρχιτεκτονικής. Επιπλέον, ο κλάδος της μάσκας προσθέτει μόνο μια μικρή υπολογιστική επιβάρυνση, επιτρέπει ένα γρήγορο σύστημα και γρήγορο πειραματισμό.

Αρχικά η μάσκα R-CNN είναι μια διαισθητική επέκταση της Faster R-CNN, αλλά την κατασκευή του κλάδου μάσκας, αλλά η κατασκευή του κλάδου μάσκας σωστά είναι ζωτικής σημασίας για να έχουμε καλά αποτελέσματα. Το πιο σημαντικό είναι, ότι το Fast R-CNN δεν έχει σχεδιαστεί για αντιστοίχιση pixel σε pixel μεταξύ εισόδου και εξόδου του δικτύου. Αυτό είναι εμφανές στο πως το RoIPool [162,163], η de facto για την παρακολούθηση σε περιπτώσεις, εκτελεί χονδροειδή χωρική ποσοτικοποίηση για την εξαγωγή των χαρακτηριστικών. Για να διορθώσουμε και την ευθυγράμμιση, προτείνουμε ένα απλό, χωρίς επίπεδο κβάντισης, που καλείται RoIAlign, που διατηρεί πιστά τις ακριβείς χωρικές θέσεις. Παρά το ότι είναι μια φαινομενικά μικρή αλλαγή, το RoIAlign έχει μεγάλο αντίκτυπο: βελτιώνει την ακρίβεια της μάσκας από 10% έως 50%, γεγονός που οδηγεί σε μεγαλύτερα κέρδη κάτω από αυστηρότερές τοπικές μετρήσεις. Δεύτερον, βρέθηκε ότι είναι απαραίτητο να αποσυνδέσουμε την μάσκα και την πρόβλεψη κλάσης: προβλέπεται μια δυαδική μάσκα για κάθε κλάση ανεξάρτητα, χωρίς ανταγωνισμό μεταξύ των κλάσεων, και βασίζεται στον κλάδο ταξινόμησης RoI του δικτύου για να προβλέψουμε την κατηγορία. Αντίθετα τα FCNs συνήθως εκτελούν

κατηγοριοποίηση πολλαπλών κατηγοριών ανά pixel και με βάση τα πειράματα λειτουργεί άσχημα για παραδείγματα κατάτμησης.



Εικόνα 29: Το πλαίσιο Mask R-CNN για παράδειγμα κατάτμησης, Πηγή: reaserchgate.net

Η μάσκα R-CNN είναι εννοιολογικά απλή: Η Faster R-CNN έχει δύο εξόδους για κάθε υποψήφιο αντικείμενο, μια ετικέτα τάξης και μια μετατόπιση κουτιού οριοθέτησης. Σε αυτά προσθέτουμε ένα τρίτο κλάδο που παράγει την μάσκα αντικειμένου. Η μάσκα R-CNN είναι έτσι μια φυσική και διαισθητική ιδέα. Αλλά η πρόσθετη μάσκα εξόδου είναι ξεχωριστή από τις εξόδους τάξης και κιβωτίου, που απαιτούν την εξαγωγή πολύ λεπτότερης χωρικής διάταξης ενός αντικειμένου. Στην συνέχεια, εισάγονται τα βασικά στοιχεία της μάσκας R-CNN, συμπεριλαμβανομένης της ευθυγράμμισης pixel-to-pixel, η οποία είναι το κυριότερο κομμάτι που λείπει από το Fast/Faster R-CNN.

Αρχικά εξετάζεται εν συντομία ο ανιχνευτής Faster R-CNN [161]. Ο Faster R-CNN αποτελείται από δύο στάδια. Το πρώτο στάδιο, που ονομάζεται Δίκτυο Προτάσεων Περιφέρειας (RPN), προτείνει πλαίσια οριοθέτησης υποψηφίων αντικειμένων. Το δεύτερο στάδιο, το οποίο είναι στην ουσία το Fast R-CNN [163], εξάγει τα χαρακτηριστικά γνωρίσματα χρησιμοποιώντας το RoIPool από κάθε υποψήφιο πλαίσιο και εκτελεί ταξινόμηση και παλινδρόμηση πλαισίου οριοθέτησης. Τα χαρακτηριστικά χρησιμοποιούνται και από τα δύο στάδια μπορούν να μοιραστούν για ταχύτερα συμπεράσματα.

Στην συνέχεια προχωράμε με το ανιχνευτή Mask R-CNN. Η μάσκα R-CNN υιοθετεί τα δύο ίδια στάδια διαδικασιών, με πανομοιότυπο πρώτο στάδιο (το οποίο είναι RPN). Στο δεύτερο στάδιο, παράλληλα με την πρόβλεψη της μετατόπισης κλάσης και κιβωτίου, το Mask R-CNN εξάγει επίσης μια δυαδική μάσκα για κάθε απόδοση επένδυσης. Αυτό έρχεται σε αντίθεση με τα πιο πρόσφατα συστήματα, όπου η ταξινόμηση εξαρτάται από τις προβλέψεις της μάσκας. (π.χ. [164,165,166]). Η προσέγγιση ακολουθεί το πνεύμα του Fast R-CNN [163] που εφαρμόζει την ταξινόμηση και την παλινδρόμηση πλαισίων οριοθέτησης παράλληλα (η οποία αποδείχθηκε ότι απλοποιεί σε μεγάλο βαθμό την πολυβάθμια γραμμή του αρχικού R-CNN [164]).

Τυπικά, κατά την διάρκεια της εκπαίδευσης, ορίζουμε μια απώλεια πολλαπλών εργασιών σε κάθε δείγμα RoI ως $L = L_{cls} + L_{box} + L_{mask}$. Η απώλεια ταξινόμησης L_{cls}

και η απώλεια πλαισίου οριοθέτησης L_{box} είναι πανομοιότυπές με αυτές που ορίζονται στο [163]. Ο κλάδος της μάσκας έχει Km^2 εξόδους διαστάσεων για κάθε (RoI), η οποία κωδικοποιεί K δυαδικές μάσκες ανάλυσης m×m, μια για κάθε από τις K κλάσεις. Σε αυτό εφαρμόζουμε εα σιγμοειδές ανά pixel και ορίζουμε το L_{mask} ως την μέση απώλεια δυαδικής διασταυρωμένης εντροπίας. Για μια απόδοση επένδυσης που σχετίζεται με την τάξη του εδάφους αληθείας k, το L_{mask} ορίζεται μόνο στη k-th μάσκα (οι έξοδοι των υπολοίπων μασκών δεν υπολογίζονται στις απώλειες).

Ο ορισμός για το L_{mask} επιτρέπει στο δίκτυο να παράγει μάσκες για κάθε κατηγορία χωρίς ανταγωνισμό μεταξύ των τάξεων. Βασίζεται στο ειδικό κλάδο ταξινόμησης για να προβλεφθεί η ετικέτα κλάσης που χρησιμοποιείται για την επιλογή της μάσκας εξόδου. Αυτό αποσυνδέει την μάσκα και την πρόβλεψη κλάσης. Αυτό διαφέρει από την κοινή πρακτική κατά την εφαρμογή FCNs [168] σε σημασιολογική κατάτμηση, η οποία συνήθως χρησιμοποιεί ένα softmax αν pixel και μια απώλεια πολυωνυμικής διασταυρωμένης εντροπίας. Φαίνεται από πειράματα που έχουν γίνει ότι αυτό είναι το κλειδί για καλά αποτελέσματα κατάτμησης.

Στην συνέχεια αναλύεται η αναπαράσταση της μάσκας. Η μάσκα κωδικοποιεί την χωρική διάταξη ενός αντικειμένου εισόδου. Έτσι, σε αντίθεση με τις ετικέτες κλάσης ή τις μετατοπίσεις κιβωτίων, που αναπόφευκτα συμπτύχθηκαν σε διανύσματα μικρής εξόδου από πλήρως συνδεδεμένα (fc) επίπεδα, η εξαγωγή της χωρικής δομής των μασκών μπορεί να αντιμετωπιστεί φυσικά από την pixel σε pixel ανταπόκριση που παρέχεται από τις συνελίξεις.

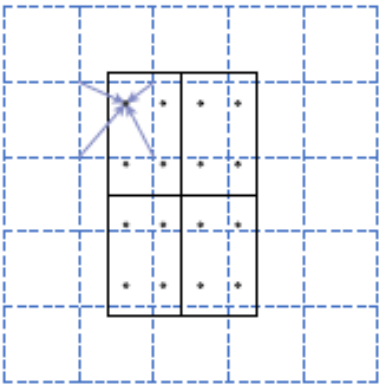
Συγκεκριμένα, προβλέπεται μια μάσκα m×m για κάθε RoI που χρησιμοποιείται σε ένα FCN [168]. Αυτό επιτρέπει σε κάθε επίπεδο στον κλάδο της μάσκας για να διατηρηθεί η ρητή m×m χωρική διάταξη του αντικειμένου χωρίς να το καταρρέει σε μια διανυσματική αναπαράσταση που στερείται χωρικών διαστάσεων. Σε αντίθεση με τις προηγούμενες μεθόδους που καταφύγουν σε επίπεδα fc για την αναπαράσταση μάσκας [164,169,165], η πλήρης συνελικτική αναπαράσταση απαιτεί λιγότερες παραμέτρους και είναι πιο ακριβής, όπως αποδεικνύεται από τα πειράματα.

Αυτή η pixel σε pixel συμπεριφορά χρειάζεται τα χαρακτηριστικά RoI, οι οποίοι οι ίδιοι είναι μικροί χάρτες χαρακτηριστικών γνωρισμάτων, για να ευθυγραμμιστούν καλά ώστε να διατηρήσουν πιστά τη ρητή χωρική αντιστοιχία ανά pixel. Αυτό οδήγησε ώστε να αναπτυχθεί το ακόλουθο στρώμα RoIAlign που παίζει βασικό ρόλο στην πρόβλεψη της μάσκας.

RoIPool [163] είναι μια τυπική λειτουργία για την εξαγωγή ενός μικρού χάρτη χαρακτηριστικών (π.χ. 77) για κάθε RoI. Η RoIPool αφού πρώτα κβαντίζει ένα κυμαινόμενο αριθμό RoI στην διακριτή κοκκοποίηση του χάρτη χαρακτηριστικών γνωρισμάτων, αυτή η κβαντισμένη RoI υποδιαιρείται έπειτα στους χωρικούς κάδους που είναι οι ήδη κβαντισμένοι και τέλος οι τιμές των χαρακτηριστικών που καλύπτονται από κάθε κάδο συγκεντρώνονται (συνήθως με μέγιστη ομαδοποίηση). Η κβάντιση πραγματοποιείται, π.χ. σε συνεχή συντεταγμένη x μέσω υπολογιστών [x=16], όπου 16 είναι ένας διασκελισμός χαρακτηριστικού χάρτη και [·] στρογγυλοποιεί. Ομοίως, κβαντισμός εκτελείται κατά την διαίρεση σε κάδους (π.χ 7×7). Αυτές οι ποσοτικοποιήσεις εισάγουν αποκλίσεις μεταξύ του RoI και των εξαγόμενων χαρακτηριστικών. Αν και αυτό μπορεί να μην επηρεάσει την ταξινόμηση, η οποία είναι

ισχυρή σε μικρές μεταφράσεις, έχει μεγάλη αρνητική επίδραση στην πρόβλεψη μασκών ακριβείας pixel.

Για να αντιμετωπιστεί αυτό, προτείνεται ένα επίπεδο RoIAlign που αφαιρεί την σκληρή κβάντιση του RoIPool , ευθυγραμμίζοντας κατάλληλα τα εξαγόμενα χαρακτηριστικά γνωρίσματα με την είσοδο. Η προτεινόμενη αλλαγή είναι απλή: αποφεύγεται οποιαδήποτε κβάντιση των ορίων RoI ή των κάδων (π.χ χρησιμοποιείται x/16 αντί για [x/16]). Χρησιμοποιείται διγραμμική παρεμβολή [170] για να υπολογίσουμε τις ακριβείς τιμές των χαρακτηριστικών εισόδου σε τέσσερις θέσεις που λαμβάνονται τακτικά δείγματα σε κάθε κάδο απόδοσης και συγκεντρώνεται το αποτέλεσμα (χρησιμοποιώντας max ή μέσο όρο), εικόνα 24. Σημειώνεται ότι τα αποτελέσματα δεν είναι ευαίσθητα στις ακριβείς θέσεις δειγματοληψίας ή στον αριθμό των σημείων δειγματοληψίας, εφόσον δεν πραγματοποιείται κβάντιση.



Εικόνα 30: RoIAlign: Το διακεκομμένο πλέγμα αντιπροσωπεύει ένα χάρτη δυνατοτήτων, οι συμπαγείς γραμμές ένα RoI (με 22 κάδους σε αυτό το παράδειγμα) και οι τελείες τα 4 σημεία δειγματοληψίας σε κάθε θέση κάδου. **Πηγή: quora.com**

4.3 Η μέθοδος RMPE

4.3.1 Εκτίμηση της ανθρώπινης στάσης πολλαπλών ατόμων

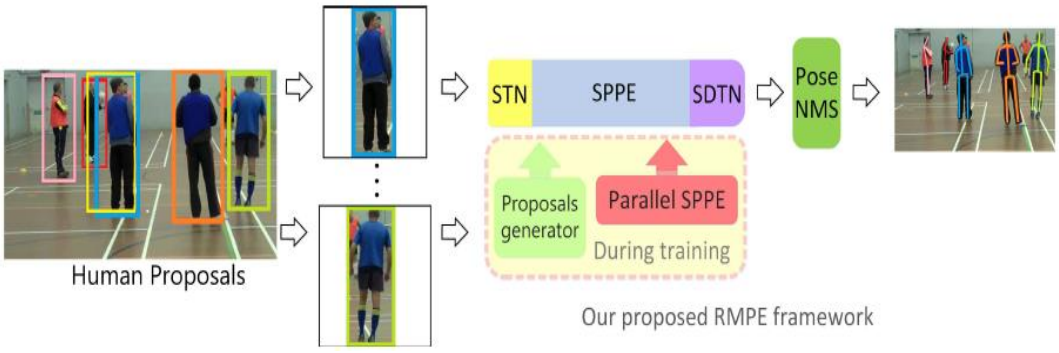
Γίνεται μια επισκόπηση αντιπροσωπευτικών εργασιών πάνω στο πλαίσιο που είναι βασισμένο σε μέρη [171, 172, 173, 174, 175]. Οι Chen et al. παρουσίασαν μια προσέγγιση για την ανάλυση των ατόμων που περιστοιχίζονται από εμπόδια σε μεγάλο βαθμό από ένα γραφικό μοντέλο που διαμορφώνει τους ανθρώπους ως ευέλικτες συνθέσεις των μερών του σώματος [181]. Οι Gkiox et al. χρησιμοποίησαν μικροστάσεις k-poselets για να ανιχνεύσουν από κοινού τους ανθρώπους και να προβλέψουν τις θέσεις των ανθρώπινων στάσεων [172]. Ο τελικός εντοπισμός στάσης προβλέπεται από έναν σταθμισμένο μέσο όρο όλων των ενεργοποιημένων μικροστάσεων. Οι Pishchulin et al. πρότειναν το DeepCut να ανιχνεύσει πρώτα όλα τα μέρη του σώματος και, στη συνέχεια, να ορίσει ετικέτες και να συναρμολογήσει αυτά τα μέρη μέσω ολοκληρωμένου γραμμικού προγραμματισμού [174]. Ένας ισχυρότερος ανιχνευτής μερών του σώματος που βασίζεται στο ResNet [176] και μια καλύτερη

στρατηγική αύξησης της βελτιστοποίησης προτείνεται από τους Insafutdinov et al. [175]. Ενώ οι μέθοδοι που βασίστηκαν σε μέρη έχουν δείξει καλή απόδοση, οι ανιχνευτές μερών του σώματός τους μπορεί να είναι ευάλωτοι, αφού λαμβάνονται υπόψη μόνο μικρές τοπικές περιοχές.

Η παρουσίαση που γίνεται ακολουθεί το πλαίσιο δύο βημάτων [177, 172]. Στην εργασία, χρησιμοποιείται μια μέθοδος SPPE με βάση το CNN για να υπολογιστούν οι στάσεις, ενώ οι Pishchulin et al. [177] χρησιμοποίησαν συμβατικά μοντέλα εικονογραφικής δομής για τον υπολογισμό στάσης. Συγκεκριμένα, οι Insafutdinov et al. [175] προτείνουν μια παρόμοια ομοχειρία δύο βημάτων που χρησιμοποιεί το Faster R-CNN ως ανθρώπινο ανιχνευτή τους και ένα μοναδιαίο DeeperCut ως υπολογιστή των στάσεών τους. Με την ανάπτυξη της ανίχνευσης αντικειμένων και υπολογισμού στάσης ενός ατόμου, το πλαίσιο δύο βημάτων μπορεί να επιτύχει περαιτέρω πρόοδο στην απόδοσή του. Η μέθοδος στοχεύει στην επίλυση του προβλήματος της ατελούς αναγνώρισης του ανθρώπου στο πλαίσιο δύο βημάτων για τη μεγιστοποίηση της ισχύος του SPPE.

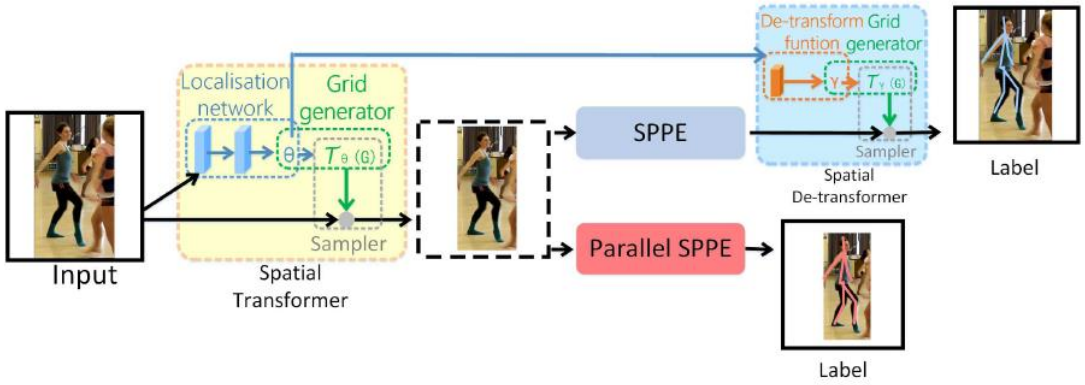
4.3.2 Περιφερειακός υπολογισμός της στάσης πολλαπλών ατόμων

Η ομοχειρία του προτεινόμενου RMPE μας απεικονίζεται στην εικόνα 31. Τα πλαίσια οριοθέτησης ανθρώπου που λαμβάνονται από τον ανιχνευτή ανθρώπου τροφοδοτούνται στη μονάδα «Symmetric STN + SPPE» και οι προτάσεις στάσεων δημιουργούνται αυτόματα. Οι παραγόμενες προτάσεις στάσεων τελειοποιούνται από παραμετρικό Pose NMS για να ληφθούν οι εκτιμώμενες ανθρώπινες στάσεις. Κατά τη διάρκεια της εκπαίδευσης, παρουσιάζεται το «Parallel SPPE» για να αποφύγουμε τοπικά ελάχιστα και να αξιοποιήσουμε περαιτέρω τη δύναμη του SSTN. Για να αυξήσουμε τα υπάρχοντα δείγματα εκπαίδευσης, έχει σχεδιαστεί μια γεννήτρια προτάσεων με καθοδήγηση στάσεων (PGPG). Στις ακόλουθες ενότητες, παρουσιάζονται τα τρία βασικά στοιχεία του πλαισίου της ενότητας.



Εικόνα 31: Αγωγός διοχέτευσης πλαισίου RMPE, Πηγή: openaccess.thecvf.com

Οι ανθρώπινες προτάσεις που παρέχονται από ανθρώπινους ανιχνευτές δεν είναι κατάλληλες για το SPPE. Αυτό συμβαίνει επειδή το SPPE είναι ειδικά εκπαιδευμένο σε εικόνες μεμονωμένου ατόμου και είναι πολύ ευαίσθητο σε σφάλματα εντοπισμού. Έχει αποδειχθεί ότι η μικρή μετάφραση ή περικοπή ανθρώπινων προτάσεων μπορεί να επηρεάσει σημαντικά την απόδοση του SPPE [178]. Το συμμετρικό μας STN + παράλληλο SPPE παρουσιάστηκε για να ενισχύσει το SPPE όταν του δόθηκαν ατελείς ανθρώπινες προτάσεις. Η μονάδα του SSTN και παράλληλου SPPE δίνεται στην εικόνα 32.



Εικόνα 32: Μια απεικόνιση της συμμετρικής αρχιτεκτονικής STN και της στρατηγικής εκπαίδευσης με παράλληλο SPPE, Πηγή: openaccess.thecvf.com

Αρχικά εξετάζεται η μονάδα του STN και του SDTN. Το δίκτυο χωρικών μετασχηματιστών [179] (STN) επέδειξε εξαιρετική απόδοση στην αυτόματη επιλογή περιοχής ενδιαφέροντος. Στην παρούσα εργασία, χρησιμοποιείται το STN για να εξαγάγουμε υψηλής ποιότητας κυρίαρχες ανθρώπινες προτάσεις. Μαθηματικά, το STN εκτελεί ομοπαράλληλο μετασχηματισμό 2D που μπορεί να εκφραστεί ως

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = [\theta_1 \quad \theta_2 \quad \theta_3] \cdot \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (4.17)$$

όπου θ_1, θ_2 και θ_3 είναι διανύσματα στο $\mathbb{R}^2$. Τα $\{x_i^s, y_i^s\}$ και $\{x_i^t y_i^t\}$ είναι οι συντεταγμένες πριν και μετά τον μετασχηματισμό, αντίστοιχα. Μετά το SPPE, η προκύπτουσα στάση χαρτογραφείται στην αρχική εικόνα της ανθρώπινης πρότασης. Φυσικά, απαιτείται ένα χωρικό δίκτυο μετασχηματιστών (SDTN) για την αναδιατύπωση της εκτιμώμενης ανθρώπινης στάσης πίσω στην αρχική συντεταγμένη εικόνας. Το SDTN υπολογίζει το γ τον από μετασχηματισμό και δημιουργεί πλέγματα με βάση το γ:

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = [\gamma_1 \quad \gamma_2 \quad \gamma_3] \cdot \begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} \quad (4.18)$$

Δεδομένου ότι το SDTN είναι μια αντίστροφη διαδικασία του STN, μπορούμε να έχουμε τα ακόλουθα:

$$[\gamma_1 \quad \gamma_2] = [\theta_1 \quad \theta_2]^{-1} \text{ (4.19)}$$

$$[\gamma_3] = -1 \times [\gamma_1 \quad \gamma_2]\theta_3 \text{ (4.20)}$$

Για αντίστροφη μετάδοση μέσω SDTN, το $\frac{\partial J(W,b)}{\partial \theta}$ μπορεί να προκύψει ως

$$\frac{\partial J(W,b)}{\partial [\theta_1 \quad \theta_2]} = \frac{\partial J(W,b)}{\partial [\gamma_1 \quad \gamma_2]} \times \frac{\partial [\gamma_1 \quad \gamma_2]}{\partial [\theta_1 \quad \theta_2]} + \frac{\partial J(W,b)}{\partial \gamma_3} \times \frac{\partial \gamma_3}{\partial [\gamma_1 \quad \gamma_2]} \times \frac{\partial [\gamma_1 \quad \gamma_2]}{\partial [\theta_1 \quad \theta_2]} \text{ (4.21)}$$

όσον αφορά τα θ_1 και θ_2, και

$$\frac{\partial J(W,b)}{\partial \theta_3} = \frac{\partial J(W,b)}{\partial \gamma_3} \times \frac{\partial \gamma_3}{\partial \theta_3} \text{ (4.22)}$$

σε σχέση με το θ_3. Τα $\frac{\partial [\gamma_1 \quad \gamma_2]}{\partial [\theta_1 \quad \theta_2]}$ και $\frac{\partial \gamma_3}{\partial \theta_3}$ μπορούν να προκύψουν από τις εξισώσεις (4.19) και (4.20), αντίστοιχα.

Αφού εξαγάγουμε υψηλής ποιότητας κυρίαρχες περιοχές ανθρώπινων προτάσεων, μπορούμε να χρησιμοποιήσουμε το SPPE εμπορίου για ακριβή υπολογισμό στάσης. Στην εκπαίδευσή μας, το SSTN είναι βελτιστοποιημένο μαζί με το SPPE μας.

Στην συνέχεια εξετάζεται η λειτουργία του παράλληλου SPPE. Για να βοηθηθεί περαιτέρω το STN να εξαγάγει καλές περιοχές όπου κυριαρχούν οι άνθρωποι, προσθέτουμε έναν παράλληλο κλάδο SPPE στη φράση εκπαίδευσης. Αυτός ο κλάδος μοιράζεται το ίδιο STN με το αρχικό SPPE, αλλά ο χωρικός μετασχηματιστής (SDTN) παραλείπεται. Η ετικέτα ανθρώπινης στάσης αυτού του κλάδου ορίζεται ως κεντραρισμένη. Για να γίνει πιο συγκεκριμένο, η έξοδος αυτού του κλάδου SPPE συγκρίνεται άμεσα με τις ετικέτες των κεντρικά τοποθετημένων πραγματικών στάσεων. Παγώνονται όλα τα επίπεδα αυτού του παράλληλου SPPE κατά τη διάρκεια της εκπαίδευσης. Τα βάρη αυτού του κλάδου είναι σταθερά και σκοπός του είναι να αναπαράγει κεντρικά τοποθετημένα σφάλματα στάσης πίσω στη μονάδα STN. Εάν η εξαχθείσα στάση του STN δεν βρίσκεται στο κέντρο, ο παράλληλος κλάδος θα αναπαράγει μεγάλα σφάλματα. Με αυτόν τον τρόπο, μπορεί να βοηθήθεί το STN να επικεντρωθεί στη σωστή περιοχή και να εξαγάγει υψηλής ποιότητας περιοχές όπου κυριαρχούν οι άνθρωποι. Στη φάση δοκιμής, το παράλληλο SPPE απορρίπτεται. Η αποτελεσματικότητα του παράλληλου SPPE επαληθεύεται και από τα πειράματά.

Το παράλληλο SPPE μπορεί να θεωρηθεί ως κανονικοποιητής κατά τη διάρκεια της εκπαίδευσης. Βοηθά να αποφευχθεί μια κακή λύση (τοπικό ελάχιστο) όπου το STN δεν μετατρέπει τη στάση στο κέντρο των ανθρωπίνων περιοχών. Η πιθανότητα επίτευξης ενός τοπικού ελάχιστου αυξάνεται επειδή το αντιστάθμισμα από το SDTN θα κάνει το δίκτυο να δημιουργήσει λιγότερα σφάλματα. Αυτά τα σφάλματα είναι απαραίτητα για την εκπαίδευση του STN. Με το παράλληλο SPPE, το STN εκπαιδεύεται για να μετακινήσει τον άνθρωπο στο κέντρο της εξαγόμενης περιοχής για να διευκολύνει την ακριβή εκτίμηση στάσης από το SPPE.

Μπορεί να φαίνεται διαισθητικό να αντικατασταθεί το παράλληλο SPPE με μια απώλεια παλινδρόμησης κεντρικά τοποθετημένων στάσεων στην έξοδο του SPPE

(πριν από το SDTN). Ωστόσο, αυτή η προσέγγιση θα υποβαθμίσει την απόδοση του συστήματός μας. Αν και το STN μπορεί εν μέρει να μετασχηματίσει την είσοδο, είναι αδύνατο να τοποθετηθεί τέλεια το άτομο στην ίδια θέση με την ετικέτα. Η διαφορά στο χώρο συντεταγμένων μεταξύ της εισόδου και της ετικέτας του SPPE θα επηρεάσει σε μεγάλο βαθμό την ικανότητά του να μάθει τον υπολογισμό στάσης. Αυτό θα προκαλέσει τη μείωση της απόδοσης του βασικού κλάδου SPPE. Έτσι, για να διασφαλιστεί ότι τόσο το STN όσο και το SPPE μπορούν να αξιοποιήσουν πλήρως τη δική τους ισχύ, ένα παράλληλο SPPE με βάρη σε διακοπή είναι απαραίτητο για το πλαίσιο. Το παράλληλο SPPE παράγει πάντα μεγάλα σφάλματα για μη κεντραρισμένες στάσεις, για να ωθήσει το STN να παράγει κεντραρισμένη στάση, χωρίς να επηρεάζεται η απόδοση του κύριου κλάδου SPPE.

4.3.3 Παραμετρική στάση NMS

Οι ανιχνευτές ανθρώπου δημιουργούν αναπόφευκτα πλεονάζουσες ανιχνεύσεις, οι οποίες με τη σειρά τους παράγουν πλεονάζουσες εκτιμήσεις στάσεων. Επομένως, απαιτείται η διαδικασία non-maximum suppression (NMS) των στάσεων για την εξάλειψη της πλεονασματικότητας. Οι προηγούμενες μέθοδοι [180, 171] είτε δεν είναι αποτελεσματικές είτε όχι αρκετά ακριβείς. Προτείνεται λοιπόν μια παραμετρική μέθοδο NMS των στάσεων. Παρόμοια με την προηγούμενη υποενότητα, η στάση P_i, με αρθρώσεις m δίνεται ως $\{\langle k_i^1, c_i^1\rangle, \ldots, \langle k_i^m, c_i^m\rangle\}$, όπου τα k_i^j και c_i^j είναι η j$^{\text{th}}$ θέση και η βαθμολογία αξιοπιστίας των αρθρώσεων αντίστοιχα.

Στην συνέχεια εξετάζεται το σύστημα NMS. Επανεξετάζοντας τη διαδικασία NMS των στάσεων ως εξής: πρώτον, η πιο αξιόπιστη στάση επιλέγεται ως αναφορά, και μερικές στάσεις κοντά σε αυτή εξαλείφονται εφαρμόζοντας το *κριτήριο εξάλειψης*. Αυτή η διαδικασία επαναλαμβάνεται στις υπόλοιπες στάσεις που έχουν τεθεί έως ότου εξαλειφθούν οι πλεονάζουσες και αναφέρονται μόνο οι μοναδικές.

Το κριτήριο εξάλειψης είναι το εξής: πρέπει να καθορίστεί η ομοιότητα στάσης για να εξαλείφθούν εκείνες που είναι πολύ κοντά και πολύ παρόμοιες μεταξύ τους. Ορίζεται μια μέθοδο μέτρησης απόστασης στάσης $d(P_i, P_j|\Lambda)$ για τη μέτρηση της ομοιότητας στάσης και ένα όριο η ως κριτήριο εξάλειψης, όπου το Λ είναι ένα σύνολο παραμέτρων της συνάρτησης $d(\cdot)$. Το κριτήριο εξάλειψης μπορεί να γραφτεί ως εξής:

$$f(P_i, P_j|\Lambda, \eta) = 1[d(P_i, P_j|\Lambda, \lambda) \leq \eta] \quad (4.23)$$

Εάν το $d(\cdot)$ είναι μικρότερο από το η, το αποτέλεσμα του f$(\cdot)$ πρέπει να είναι 1, πράγμα που υποδηλώνει ότι η στάση P_i πρέπει να εξαλειφθεί λόγω πλεονασματικότητας σε σχέση με τη στάση αναφοράς P_j.

Τώρα, παρουσιάζεται τη συνάρτηση απόστασης$d_{pose}(P_i, P_j)$. Υποθετικά το πλαίσιο για το P_i είναι B_i. Στη συνέχεια, ορίζουμε μια ελαστική συνάρτηση αντιστοίχισης :

$$K_{Sim}(P_i, P_j|\sigma_1) = \begin{cases} \sum_n \tanh\frac{c_i^n}{\sigma_1} \cdot \tanh\frac{c_j^n}{\sigma_1} & , \text{εάν το } k_j^n \text{ είναι εντός του } B(k_i^n) \\ 0 & , \quad \text{αλλίως} \end{cases} \quad (4.24)$$

όπου το $B(k_i^n)$ είναι ένα κέντρο πλαισίου στο k_i^n, και κάθε διάσταση του $B(k_i^n)$ είναι το 1/10 του αρχικού πλαισίου B_i. Η λειτουργία tanh αποβάλλει τις στάσεις με βαθμολογίες χαμηλής αξιοπιστίας φιλτράροντάς τες. Όταν δύο αντίστοιχες αρθρώσεις έχουν υψηλές βαθμολογίες αξιοπιστίας, το αποτέλεσμα θα είναι κοντά στο 1. Αυτή η απόσταση μετρά ελαστικά τον αριθμό των αρθρώσεων που αντιστοιχίζονται μεταξύ των στάσεων. Λαμβάνεται επίσης υπόψη η χωρική απόσταση μεταξύ των μερών, που μπορεί να γραφτεί ως

$$H_{Sim}(P_i, P_j|\sigma_2) = \sum_n exp[-\frac{\left(k_i^n - k_j^n\right)^2}{\sigma_2}] \quad (4.25)$$

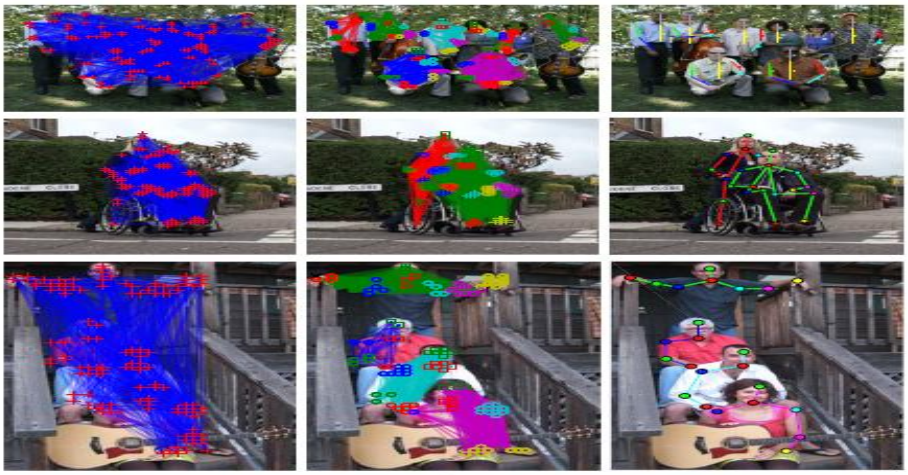
Συνδυάζοντας τις εξισώσεις. (4.24) και (4.25), η συνάρτηση τελικής απόστασης μπορεί να γραφτεί ως

$$d(P_i, P_j|\Lambda) = K_{Sim}(P_i, P_j|\sigma_1) \; + \; \lambda \cdot H_{Sim}(P_i, P_j|\sigma_2) \; (4.26)$$

όπου λ είναι ένα βάρος που εξισορροπεί τις δύο αποστάσεις και $\Lambda = \{\sigma_1, \sigma_2, \lambda\}$.

4.4 Η μέθοδος Deepcut

Σε αυτήν την ενότητα, το πρόβλημα της εκτίμησης αρθρωτών στάσεων ενός άγνωστου αριθμού ατόμων σε μια εικόνα εμφανίζεται ως πρόβλημα βελτιστοποίησης. Ο στόχος αυτής της διατύπωσης είναι να εκθέσουμε από κοινού τρία προβλήματα: 1. Η επιλογή ενός υποσυνόλου μερών του σώματος από ένα σύνολο *D υποψηφίων μερών του σώματος*, εκτιμώμενη από μια εικόνα όπως περιγράφεται στην Ενότητα 4 και απεικονίζεται ως κόμβοι ενός γραφήματος στην εικόνα 33(α). 2. Ο *ορισμός ετικέτας* (labeling) σε κάθε επιλεγμένο μέρος του σώματος βάσει μιας από τις κατηγορίες *C μερών του σώματος*, π.χ. «βραχίονας», «πόδι», «κορμός», όπως απεικονίζεται στην εικόνα 33(γ). 3. Ο *διαχωρισμός* των μερών του σώματος που ανήκουν στο ίδιο άτομο, όπως απεικονίζεται στην εικόνα 33(β).



Εικόνα 33: Επισκόπηση της μεθόδου DeepCut, Πηγή: arxiv-vanity.com

4.4.1 Εφικτές Λύσεις

Κωδικοποιούνται οι ετικέτες των τριών προβλημάτων από κοινού μέσω τριπλών (*x*, *y*, *z*) δυαδικών τυχαίων μεταβλητών με πεδία ορισμού $x \in \{0,1\}^{D\times C}, y \in \{0,1\}^{\binom{D}{2}}$ και $z \in \{0,1\}^{\binom{D}{2}\times C^2}$. Εδώ, $x_{dc} = 1$ σημαίνει ότι το υποψήφιο μέρος *d* του σώματος είναι της κλάσης c, $y_{dd'} = 1$ σημαίνει ότι τα υποψήφια μέρη d του σώματος και d΄ ανήκουν στο ίδιο άτομο, και οι $z_{dd'cc'}$ είναι βοηθητικές μεταβλητές για τη συσχέτιση x και y μέσω $z_{dd'cc'} = x_{dc}x_{d'c'}y_{dd'}$. Έτσι, το $z_{dd'cc'} = 1$ σημαίνει ότι το υποψήφιο μέρος *d* του σώματος είναι της κλάσης c ($x_{dc} = 1$), το υποψήφιο μέρος d΄ του σώματος είναι της κλάσης c΄ ($x_{d'c'} = 1$) και τα υποψήφια μέρη d και d΄ του σώματος ανήκουν στο ίδιο άτομο ($y_{dd'} = 1$).

Προκειμένου να περιοριστούν οι ετικέτες 01 (*x*, *y*, *z*) σε άρτια καθορισμένες αρθρωτές στάσεις ενός ή περισσοτέρων ατόμων, επιβάλλονται οι γραμμικές ανισότητες (4.27) - (4.29) που αναφέρονται παρακάτω. Εδώ, οι ανισότητες (1) εγγυώνται ότι κάθε μέρος του σώματος φέρει ετικέτα από το πολύ μία κατηγορία μερών του σώματος. (Εάν ορίζεται ετικέτα χωρίς κατηγορία μερών του σώματος, αποκρύπτεται). Οι ανισότητες (4.28) εγγυώνται ότι τα ξεχωριστά μέρη d και d΄ του σώματος ανήκουν στο ίδιο άτομο μόνο εάν δεν αποκρύπτεται ούτε το d ούτε το d΄. Οι ανισότητες (4.29) εγγυώνται, για οποιαδήποτε τρία ξεχωριστά μέρη d, d΄, d΄΄ του σώματος, ότι εάν τα d και d΄είναι τα ίδια άτομα (όπως υποδεικνύεται από $y_{dd'} = 1$), και τα d΄ και d΄΄ είναι το ίδιο άτομο (όπως υποδεικνύεται από $y_{d'd''} = 1$), τότε επίσης τα d και d΄΄ είναι το ίδιο άτομο ($y_{dd''} = 1$), δηλαδή, μεταβατικότητα [181]. Τέλος, οι ανισότητες (4.30) εγγυώνται, για οποιοδήποτε $dd' \in \binom{D}{2}$ και οποιοδήποτε $cc' \in C^2$ ότι $z_{dd'cc'} = x_{dc}x_{d'c'}y_{dd'}$. Αυτοί οι περιορισμοί επιτρέπουν να γράφεται μια αντικειμενική συνάρτηση ως γραμμική μορφή στο z που διαφορετικά θα γραφόταν ως κυβική μορφή στα x και y. Με το X_{DC} εννοείται το σύνολο όλων των μεταβλητών (*x*, *y*, *z*) που ικανοποιούν όλες τις ανισότητες, δηλαδή το σύνολο των εφικτών λύσεων.

$$\forall d \in D \forall cc' \in \binom{C}{2}: x_{dc} + x'_{dc'} \leq 1 \text{ (4.27))}$$

$$\forall dd' \in \binom{D}{2}: y_{dd'} \leq \sum_{c \in C} x_{dc}$$

$$y_{dd'} \leq \sum_{c \in C} x_{d'c} \quad (4.28)$$

$$\forall dd'd'' \in \binom{D}{3}: y_{d'd} + y_{d'd''} - 1 \leq y_{dd''} (4.29)$$

$$\forall dd' \in \binom{D}{2} \forall cc' \in C^2: x_{dc} + x_{d'c'} + y_{dd'} - 2 \leq z_{dd'cc'}$$

$$z_{dd'cc'} \leq x_{dc}$$
$$z_{dd'cc'} \leq x_{d'c'}$$
$$z_{dd'cc'} \leq y_{dd'} \quad (4.30)$$

Όταν υπάρχει το πολύ ένα άτομο σε μια εικόνα, περιορίζονται περαιτέρω οι εφικτές λύσεις σε μια σαφώς καθορισμένη στάση ενός μοναδικού ατόμου. Αυτό επιτυγχάνεται με μια πρόσθετη κατηγορία ανισοτήτων που εγγυώνται, για δύο ξεχωριστά μέρη του σώματος που δεν αποκρύπτονται, ότι πρέπει να ομαδοποιηθούν μαζί:

$$\forall dd' \in \binom{D}{2} \forall cc' \in C^2: x_{dc} + x_{d'c'} - 1 \leq y_{dd'} \text{ (4.31)}$$

4.4.2 Αντικειμενική συνάρτηση

Για κάθε ζεύγος $(d, c) \in D \times C$, υπολογίζεται μια πιθανότητα $p_{dc} \in [0,1]$ του μέρους d του σώματος που ανήκει στην κατηγορία c. Στο πλαίσιο των CRF, αυτές οι πιθανότητες ονομάζονται part unaries.

Για κάθε $dd' \in \binom{D}{2}$ και κάθε $cc' \in C^2$, θεωρούμε πιθανότητα $p_{dd'cc'} \in (0,1)$ της υπό όρους πιθανότητας των d και d΄ που ανήκουν στο ίδιο άτομο, δεδομένου ότι τα d και d΄ είναι μέρη του σώματος των κατηγοριών c και c΄, αντίστοιχα. Για $c \neq c'$, αυτές οι πιθανότητες $p_{dd'cc'}$ είναι οι όροι ανά ζεύγη σε ένα γραφικό μοντέλο του ανθρώπινου σώματος. Σε αντίθεση με το κλασικό μοντέλο εικονογραφικών δομών, το μοντέλο αυτό επιτρέπει ένα πλήρως συνδεδεμένο γράφημα, όπου κάθε μέρος του σώματος συνδέεται με όλα τα άλλα μέρη σε ολόκληρο το σύνολο D με έναν όρο ανά ζεύγος. Για c = c΄, $p_{dd'cc'}$ είναι η πιθανότητα των υποψηφίων μερών d και d΄ που αντιπροσωπεύουν το ίδιο μέρος του ίδιου ατόμου. Αυτό διευκολύνει την ομαδοποίηση υποψηφίων πολλαπλών μερών του ίδιου μέρους του ίδιου ατόμου και μια *απωθητική* ιδιότητα που αποτρέπει τους κοντινούς υποψηφίους μερών του ίδιου τύπου να συσχετίζονται με διαφορετικά άτομα. Το πρόβλημα βελτιστοποίησης που ονομάζουμε πρόβλημα κατάτμησης και ορισμού ετικέτας υποσυνόλου είναι το ILP που ελαχιστοποιείται στο σύνολο των εφικτών λύσεων X_{DC}:

$$\min_{(x,y,z) \in X_{dc}} \langle \alpha, x \rangle + \langle \beta, z \rangle \text{ (4.32)}$$

όπου χρησιμοποιήσαμε τις συντομογραφημένες σημειώσεις

$$\alpha_{dc} \coloneqq \log 1 - pdcpdc \quad (4.33)$$

$$\beta_{dd'cc'} := \log \frac{1 - p_{dd'cc'}}{p_{dd'cc'}} \quad (4.34)$$

$$\langle \alpha, x \rangle \coloneqq \sum_{d \in D} \sum_{c \in C} \alpha_{dc} x_{dc} \quad (4.35)$$

$$\langle \beta, z \rangle := \sum_{dd' \in \binom{D}{2}} \sum_{c,c' \in C} \beta_{dd'cc'} z_{dd'cc'} \quad (4.36)$$

Ο στόχος (4.32) - (4.36) είναι η εκτίμηση MAP ενός μέτρου πιθανότητας των κοινών ανιχνεύσεων x και των ομαδοποιήσεων y, z των μερών του σώματος, όπου οι προηγούμενες πιθανότητες p_{dc} και $p_{dd'cc'}$ υπολογίζονται ανεξάρτητα από τα δεδομένα και η πιθανότητα είναι μια θετική σταθερά, εφόσον (x, y, z) ικανοποιεί (4.27) - (4.30) και είναι 0, σε κάθε περίπτωση. Η ακριβής μορφή (4.32) - (4.36) λαμβάνεται όταν ελαχιστοποιείται ο αρνητικός λογάριθμος αυτού του μέτρου πιθανότητας.

4.4.3. Βελτιστοποίηση

Προκειμένου να επιτευχθούν εφικτές λύσεις του ILP (4.32) με εγγυημένα όρια, διαχωρίζονται οι ανισότητες (4.27) - (4.31) στο δίκτυο βρόχων branch-and-cut του υπερσύγχρονου ILP solver Gurobi. Πιο συγκεκριμένα, επιλύεται μια ακολουθία χαλάρωσης του προβλήματος (4.32), ξεκινώντας από το (ασήμαντο) μη περιορισμένο πρόβλημα. Κάθε πρόβλημα επιλύεται χρησιμοποιώντας τις περικοπές που προτείνει ο Gurobi. Μόλις βρεθεί μια ακέραια εφικτή λύση, εντοπίζονται παραβιάσεις ανισοτήτων (4.27) - (4.31), εάν υπάρχουν, με την αναζήτηση κατά πλάτος (BFS), τις προστίθενται στην ομάδα περιορισμών και επιλύεται ξανά η ενισχυμένη χαλάρωση. Μόλις βρεθεί μια ακέραια λύση που ικανοποιεί όλες τις ανισότητες, μαζί με ένα κατώτατο όριο που βεβαιώνει ένα κενό βελτιστοποίησης κάτω από 1%, σταματάει η διαδικασία.

4.4.4 Πιθανότητες κατά ζεύγη

Εδώ περιγράφεται ο υπολογισμός των όρων κατά ζεύγη. Προσδιορίζονται ζευγάρια $f_{dd'}$ για τη μεταβλητή $z_{dd'cc'}$. Κάθε ανίχνευση μέρους d περιλαμβάνει τις πιθανότητες $f_{p_{dc}}$, τη θέση του (x_d, y_d), κλίμακα h_d και συντεταγμένες του πλαισίου περιορισμού Bd. Δεδομένων δύο ανιχνεύσεων d και d′, και τα αντίστοιχα χαρακτηριστικά $\left(f_{p_{dc}}, x_d, y_d, h_d, B_d\right)$ και $\left(f_{p_{d'c}}, x_{d'}, y_{d'}, h_{d'}, B_{d'}\right)$, ορίζονται δύο σύνολα βοηθητικών μεταβλητών για το $z_{dd'cc'}$, ένα σύνολο για c = c′ (ομαδοποίηση στην ίδια κατηγορία μερών του σώματος) και ένα για $c \neq c'$ (ορισμός ετικέτας σε δύο κατηγορίες μερών του σώματος) . Αυτά τα χαρακτηριστικά αποτυπώνουν την εγγύτητα, την κινηματική σχέση και την ομοιότητα εμφάνισης μεταξύ των μερών του σώματος.

Η ίδια κατηγορία μερών του σώματος ($c \neq c'$). Δύο ανιχνεύσεις που δηλώνουν το ίδιο μέρος του σώματος του ίδιου ατόμου θα πρέπει να βρίσκονται πολύ κοντά η μία στην άλλη. Εισάγονται οι ακόλουθες βοηθητικές μεταβλητές που αποτυπώνουν τις σχέσεις χώρου:

$\Delta x = |x_d - x_{d'}|/\bar{h}, \Delta y = |y_d - y_{d'}|/\bar{h}, \Delta h = |h_d - h_{d'}|/\bar{h}$, IOUnion, IOMin, IOMax. Τα τελευταία τρία αποτελούν διασταυρώσεις πάνω από την ένωση / ελάχιστο / μέγιστο των δύο πλαισίων ανίχνευσης, αντίστοιχα, και $\bar{h} = (h_d + h_{d'})/2$.

Στη μη-γραμμική χαρτογράφηση αυξάνεται η αναπαράσταση χαρακτηριστικών προσθέτοντας τετραγωνικούς και εκθετικούς όρους. Το τελευταίο χαρακτηριστικό ζεύγους $f_{dd'}$ για τη μεταβλητή $z_{dd'cc'}$ είναι $(\Delta x, \Delta y, \Delta h, IOUnion, IOMin, IOMax, (\Delta x)^2, \ldots, (IOMax)^2, \exp(-\Delta x), \ldots, \exp(-IOMax))$.

Για δύο διαφορετικές κατηγορίες μερών του σώματος (c≠c'). Κωδικοποιούνται οι περιορισμοί κινηματικού σώματος στο χαρακτηριστικό ζεύγους εισάγοντας τις βοηθητικές μεταβλητές Sdd0 και Rdd0, όπου $S_{dd'}$ και $R_{dd'}$,είναι η Ευκλείδεια απόσταση και η γωνία μεταξύ δύο ανιχνεύσεων, αντίστοιχα. Για να καταγράψουμε την κοινή κατανομή των $S_{dd'}$ και $R_{dd'}$, αντί να χρησιμοποιήσουμε τα $S_{dd'}$ και $R_{dd'}$, απευθείας, χρησιμοποιούμε την εκ των υστέρων πιθανότητα $p(z_{dd'cc'} = 1|S_{dd'}, R_{dd'})$ ως χαρακτηριστικό ζεύγους για το $z_{dd'cc'}$ για την κωδικοποίηση των γεωμετρικών σχέσεων μεταξύ της κατηγορίας μερών του σώματος c και c΄. Πιο συγκεκριμένα, έχοντας υπόψη ότι η προηγούμενη πιθανότητα $p(z_{dd'cc'} = 1) = p(z_{dd'cc'} = 0) = 0.5$, η εκ των υστέρων πιθανότητα ανίχνευσης d και d΄ έχει την ετικέτα μέρους του σώματος c και c΄, δηλαδή $z_{dd'cc'} = 1$, και είναι:

$$p(z_{dd'cc'} = 1|S_{dd'}, R_{dd'}) = \frac{p(S_{dd'}, R_{dd'|z_{dd'cc'}=1})}{p(S_{dd'}, R_{dd'}|z_{dd'cc'=1)} + p(S_{dd'}, R_{dd'}|z_{dd'cc'=0)}}$$

όπου $p(S_{dd'}, R_{dd'|z_{dd'cc'}=1}$ λαμβάνεται πραγματοποιώντας ένα κανονικοποιημένο 2D ιστόγραμμα των $S_{dd'}$ και $R_{dd'}$ από θετικά παραδείγματα εκπαίδευσης, κατ᾽ αναλογία με την αρνητική πιθανότητα $p(S_{dd'}, R_{dd'}|z_{dd'cc'=0})$. Οι συντελεστές α και β η αντικειμενική συνάρτηση (4.32) ορίζονται από τον λόγο πιθανότητας στο χώρο καταγραφής (4.33),(4.34). Εδώ περιγράφεται ο υπολογισμός της αντίστοιχης πυκνότητας πιθανότητας:

1) Για κάθε ζεύγος ανίχνευσης και κατηγορίας μερών, συγκεκριμένα για οποιαδήποτε $(d, c) \epsilon D \times C$, υπολογίζεται μια πιθανότητα $p_{dc} \in (0,1)$ της ανίχνευσης d που είναι μέρος του σώματος της κατηγορίας c.

2) Για κάθε συνδυασμό δύο διακριτών ανιχνεύσεων και δύο κατηγοριών μερών του σώματος, συγκεκριμένα για οποιοδήποτε $dd' \epsilon \binom{D}{2}$ και οποιοδήποτε $cc' \epsilon C^2$, υπολογίζουμε την πιθανότητα $p_{dd'cc'} \epsilon (0,1)$ των d και d΄ να ανήκουν στο ίδιο άτομο, ενώ στο μεταξύ τα d και d΄ είναι μέρη του σώματος των κατηγοριών c και c΄, αντίστοιχα.

Για την μάθηση. δεδομένων των χαρακτηριστικών $f_{dd'}$και μιας συνάρτησης Gaussian εκ τω προτέρων πιθανότητας $\boldsymbol{p(\theta_{cc'}) = N(0, \sigma^2)}$ στις παραμέτρους, το λογιστικό μοντέλο είναι

$$p(z_{dd'} = 1|f_{dd'}, \theta_{cc'}) = \frac{1}{1 + exp(-\langle \theta_{cc'}, f_{dd'} \rangle} \quad (4.37)$$

Οι $(|c| \times (|c|+1))/2$ παράμετροι υπολογίζονται με χρήση ML.
Εξαγωγή συμπερασμάτων. Δεδομένων δύο ανιχνεύσεων d και d΄, οι συντελεστές α_{dc} για x_{dc} και $\alpha_{d'c}$ για $x_{d'c}$ λαμβάνονται βάσει της εξίσωσης (4.33), ο συντελεστής $\beta_{dd'cc'}$ για $z_{dd'cc'}$ έχει τη μορφή:

$$\beta_{dd'cc'} = \log\frac{1-p_{dd'cc'}}{p_{dd'cc'}} = -\langle f_{dd'}, \theta_{\alpha'}\rangle \quad (4.38)$$

Οι παράμετροι μοντέλου $\theta_{cc'}$ μαθαίνονται χρησιμοποιώντας λογιστική παλινδρόμηση[182].

ΚΕΦΑΛΑΙΟ 5

Συμπεράσματα και επόμενα βήματα

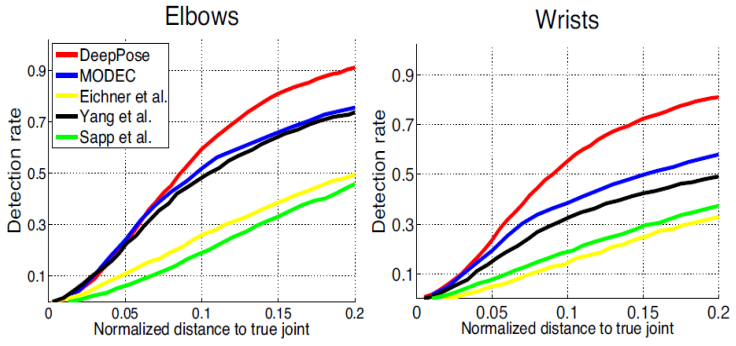
5.1 Συμπεράσματα

Οι μέθοδοι που επιλέχθηκαν στην παρούσα διπλωματική εργασία αποτελούν κατά κύριο λόγο τεχνολογία αιχμής και έχουν επιτύχει υψηλές αποδόσεις σε σχέση με άλλες αρχιτεκτονικές.

Οι μέθοδοι έχουν δοκιμαστεί σε πειράματα και έχουν χρησιμοποιήσει για την λειτουργία τους γνωστές βάσεις δεδομένων όπως η FLIC (με 4000 εικόνες εκπαίδευσης και 1000 εικόνες δοκιμής από καρέ του σινεμά), η LSP (με 11000 εικόνες εκπαίδευσης και 1000 εικόνες δοκιμής από καρέ αθλητικών γεγονότων) και η πολύ απαιτητική MPII(με 28000 εικόνες εκπαίδευσης και 11700 εικόνες δοκιμής).

Αρχικά θα γίνει μια αξιολόγηση των μεθόδων αναγνώρισης της στάσης του σώματος ενός ατόμου. Η αξιολόγηση γίνεται χρησιμοποιώντας το τυπικό ποσοστό σωστών σημείων κλειδιών (Percentage of Correct Keypoints-PCK), μέτρηση που αναφέρει το ποσοστό των ανιχνεύσεων που εμπίπτουν σε μια κανονικοποιημένη απόσταση του εδάφους αληθείας. Για την βάση δεδομένων FLIC η απόσταση κανονικοποιείται κατά μέγεθος κορμού για την βάση δεδομένων MPII κατά ένα κλάσμα του μεγέθους της κεφαλής (αναφέρεται ως PCKh).

1. Το δίκτυο DeepPose χρησιμοποίησε για τις δοκιμές τις βάσεις δεδομένων FLIC και LSP.



Σύγκριση PCK στην βάση δεδομένων FLIC [183]

Τα αποτελέσματα για την βάση δεδομένων FLIC φαίνονται στο γράφημα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων FLIC φτάνουν 92% PCKh@0.2 στον αγκώνα και 82% στον καρπό. Οπότε παρατηρούμε καθαρή υπεροχή από τις άλλες εφαρμογές

Method	Arm		Leg		Ave.
	Upper	Lower	Upper	Lower	
DeepPose-st1	0.5	0.27	0.74	0.65	0.54
DeepPose-st2	**0.56**	0.36	**0.78**	0.70	0.60
DeepPose-st3	**0.56**	**0.38**	0.77	**0.71**	**0.61**
Dantone et al.	0.45	0.25	0.65	0.61	0.49
Tian et al.	0.52	0.33	0.70	0.60	0.56
Johnson et al.	0.54	**0.38**	0.75	0.66	0.58
Wang et al.	**0.565**	0.37	0.76	0.68	0.59
Pishchulin	0.49	0.32	0.74	0.70	0.56

Πίνακας αποτελεσμάτων LSP(PCKh@0.5) [183]

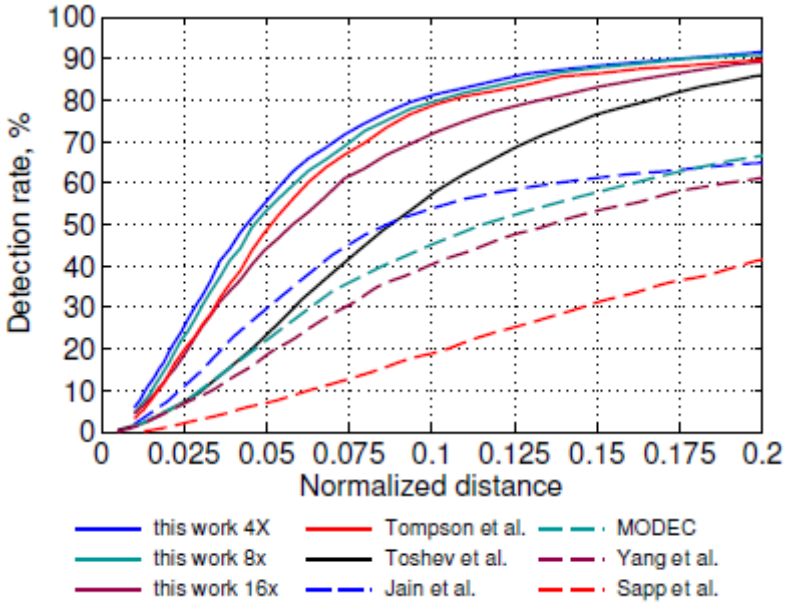
Τα αποτελέσματα για την βάση δεδομένων LSP φαίνονται στον πίνακα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων LSP έχουν μια μέση τιμή απόδοσης στα άκρα απόδοσης 61% PCKh@0.5. Βλέπουμε ότι υπάρχει μια καθαρή υπεροχή έναντι των άλλων περιοχών, ειδικά πετυχαίνοντας καλύτερες εκτιμήσεις για τα πόδια. Αξίζει να σημειωθεί ότι ενώ οι άλλες προσεγγίσεις παρουσιάζουν πλεονεκτήματα για συγκεκριμένα άκρα, καμία δεν κυριαρχεί σε όλα τα άκρα. Αντίθετα η μέθοδος DeepPose δείχνει ισχυρά αποτελέσματα για όλα τα προκλητικά άκρα.

Η μέθοδος Deeppose αποτυπώνει πλήρως κάθε άρθρωση του σώματος και είναι απλούστερη από μεθόδους που χρησιμοποιούν γραφικά μοντέλα. Ένα άλλο αξιοσημείωτο είναι ότι αποδίδει μεγαλύτερα κέρδη σε σχέση με άλλες αρχιτεκτονικές στις περιοχές χαμηλής ακρίβειας.

2. Το δίκτυο παλινδρόμησης χάρτη θερμότητας χρησιμοποίησε για τις δοκιμές τις βάσεις δεδομένων FLIC και MPII.

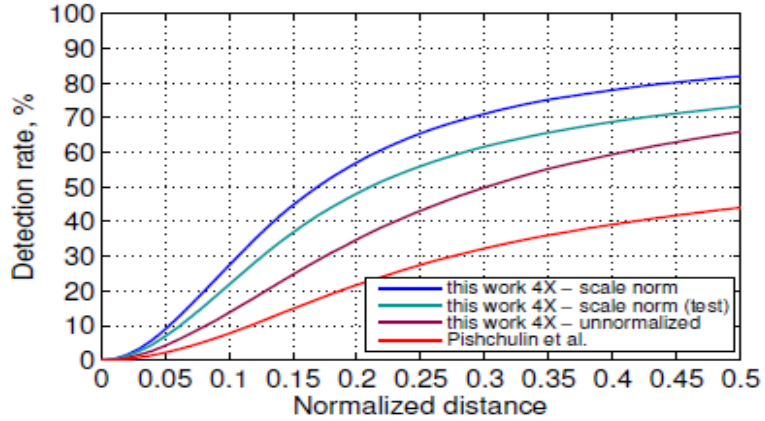
	Head	Shoulder	Elbow	Wrist
Yang et al.	-	-	22.6	15.3
Sapp et al.	-	-	6.4	7.9
Eichner et al.	-	-	11.1	5.2
MODEC et al.	-	-	28.0	22.3
Toshev et al.	-	-	25.2	26.4
Jain et al.	-	42.6	24.1	22.3
Tompson et al.	90.7	70.4	50.2	55.4
This work 4x	**92.6**	73.0	**57.1**	**60.4**
This work 8x	92.1	**75.8**	55.6	56.6
This work 16x	91.6	73.0	47.7	45.5

Πίνακας αποτελεσμάτων FLIC (PCKh 0.05) [184]



Σύγκριση PCK στην βάση δεδομένων FLIC (μέσο PCK για αγκώνα και καρπό) [184]

Τα αποτελέσματα για την βάση δεδομένων FLIC φαίνονται στο γράφημα και στον πίνακα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων FLIC φτάνουν στο 91% απόδοση μέσο όρο για αγκώνα και καρπό PCKh@0.2. Το μοντέλο μας ξεπερνά τα προηγούμενα αποτελέσματα τελευταίας τεχνολογίας από τον Tompson για μεγάλες αποστάσεις, λόγω της χρήσης του SpatialDropout. Στην περιοχή υψηλής ακρίβειας, το κλιμακωμένο δίκτυο είναι ικανό να ξεπεράσει όλες τις μεθόδους τελευταίας τεχνολογίας, μάλιστα με σημαντικό περιθώριο.



Σύγκριση PCK στην βάση δεδομένων MPII (μέσο PCK για όλα τα σημεία) [184]

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Upper Body	Full Body
Gkioxari et al.	-	36.3	26.1	15.3	-	-	-	25.9	-
Sapp & Taskar	-	38.0	26.3	19.3	-	-	-	27.9	-
Yang & Ramanan	73.2	56.2	41.3	32.1	36.2	33.2	34.5	43.2	44.5
Pishchulin et al.	74.2	49.0	40.8	34.1	36.5	34.4	35.1	41.3	44.0
This work - scale normalized	**96.1**	**91.9**	**83.9**	**77.8**	**80.9**	**72.3**	**64.8**	**84.5**	**82.0**
This work - scale normalized (test only)	93.5	87.5	75.5	67.8	68.3	60.3	51.7	77.0	73.3
This work - unnormalized	83.4	77.5	67.5	59.8	64.6	55.6	46.1	68.3	66.0

Πίνακας αποτελεσμάτων MPII (PCKh@0.5) [184]

Τα αποτελέσματα για την βάση δεδομένων MPII φαίνονται στον πίνακα στο γράφημα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων MPII όπως βλέπουμε και πίνακα φτάνει για όλο το σώμα 82% PCKh@0.5 , ποσοστό που ξεπερνά όλες τις υπάρχουσες μεθόδους τελευταίας τεχνολογίας .

Παρατηρήσαμε ότι όσο αυξάνει η συγκέντρωση της cascade τόσο η βελτίωση της απόδοσης είναι αισθητή. Ένα ακόμα συμπέρασμα ήταν ότι σε εργασίες εντοπισμού όπως το ανθρώπινο σώμα αποτελούν εκτιμήσεις που συχνά απαιτούν υψηλό βαθμό χωρικής ακρίβειας. Σε αυτή την εργασία είδαμε πως η ακρίβεια που χάνεται λόγω της συγκέντρωσης σε παραδοσιακές αρχιτεκτονικές ConvNet, ανακτάται αποτελεσματικά καθώς ανακτώνται τα υπολογιστικά οφέλη της συγκέντρωσης. Τέλος παρουσιάστηκε μια νέα επικαλυπτόμενη αρχιτεκτονική που συνδυάζει τα λεπτά με τα χονδροειδή δίκτυα τα οποία πέτυχαν όπως είδαμε νέας τελευταίας τεχνολογίας αποτελέσματα για τις βάσεις δεδομένων FLIC και MPII.

3. Το μοντέλο επαναληπτικής ανάδρασης σφάλματος χρησιμοποίησε για τις δοκιμές τις βάσεις δεδομένων MPII και LSP.

	Torso	Upper Leg	Lower Leg	Upper Arm	Forearm	Head	Total
Pishchulin et al.	88.9	64.0	58.1	45.5	35.1	85.1	58.0
Tompson et al.	90.3	70.4	61.1	63.0	51.2	83.7	66.6
Fan et al.	95.4	77.7	69.8	62.8	49.1	86.6	70.1
Chen and Yuille	96.0	77.2	72.2	69.7	58.1	85.6	73.6
IEF	95.3	81.8	73.3	66.7	51.0	84.4	72.5

Πίνακας αποτελεσμάτων LSP (PCKh@0.5) [185]

Τα αποτελέσματα για την βάση δεδομένων LSP φαίνονται στο πίνακα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων φτάνει συνολικά μέση απόδοση 72.5% PCKh@0.5 ,απόδοση που όπως φαίνεται και από τον πίνακα είναι αισθητά καλύτερη από την απόδοση των ανταγωνιστικών μεθόδων.

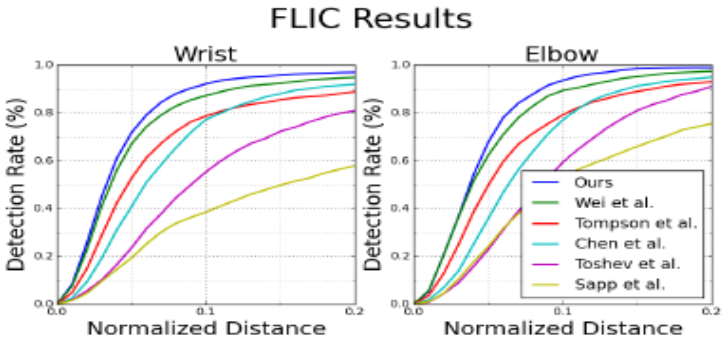
	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	UBody	FBody
Yang & Ramanan	73.2	56.2	41.3	32.1	36.2	33.2	34.5	43.2	44.5
Pischulin et al	74.2	49.0	40.8	34.1	36.5	34.4	35.1	41.3	44.0
Tompson et al.	96.1	91.9	83.9	77.8	80.9	72.3	64.8	84.5	82.0
IEF	95.7	91.6	81.5	72.4	82.7	73.1	66.4	82.0	81.3
Tompson et al.	83.4	77.5	67.5	59.8	64.6	55.6	46.1	68.3	66.0
IEF	95.5	91.6	81.5	72.4	82.7	73.1	66.9	81.9	81.3

Πίνακας αποτελεσμάτων MPII (PCKh@0.5) [185]

Τα αποτελέσματα στην βάση δεδομένων MPII φτάνει συνολικά μέση απόδοση 81.3% PCKh@0.5 ,απόδοση που όπως φαίνεται και από τον πίνακα είναι αισθητά καλύτερη από την απόδοση των ανταγωνιστικών μεθόδων. Ειδικά σε πιο ρεαλιστική ρύθμιση των άγνωστων πληροφοριών κλίμακας η απόδοση της μεθόδου μας είναι 81.3% και του κύριου ανταγωνιστή της (μέθοδος Tompson) να ανέρχεται σε 66%, γεγονός που δείχνει ότι η μέθοδος μας είναι πού πιο αποδοτική και μέθοδος αιχμής.

Σημαντικό είναι ότι παρατηρήθηκε πως επιτύγχανέ καλύτερα αποτελέσματα όχι στην άμεση παλινδρόμηση στις θέσεις των σημείων κλειδιών (άμεση πρόβλεψη) αλλά με παλινδρόμηση στις θέσεις κλειδιά του εδάφους αληθείας (αντί για οριοθετημένες διορθώσεις σε κάθε επανάληψη), ξεκινώντας με την στάση στην προηγούμενη επανάληψη.

4. <u>Το δίκτυο στοιβασμένης κλεψύδρας χρησιμοποίησε για τις δοκιμές τις βάσεις δεδομένων FLIC και MPII</u>.

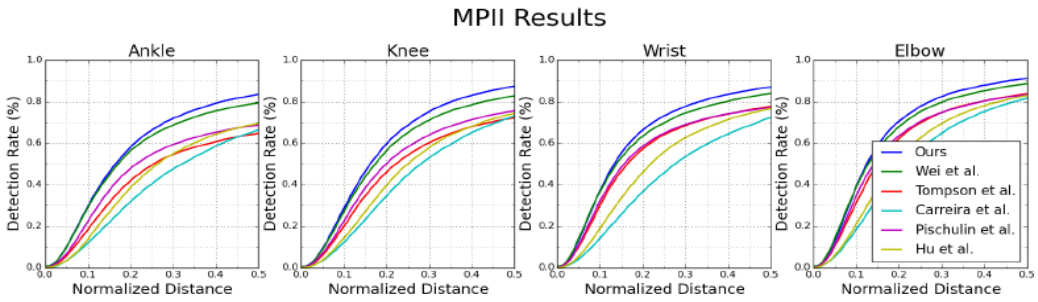


Σύγκριση PCK στην βάση δεδομένων FLIC [186]

	Elbow	Wrist
Sapp et al.	76.5	59.1
Toshev et al.	92.3	82.0
Tompson et al.	93.1	89.0
Chen et al.	95.3	92.4
Wei et al.	97.6	95.0
Our model	**99.0**	**97.0**

Πίνακας αποτελεσμάτων FLIC (PCKh@0.2) [186]

Τα αποτελέσματα για την βάση δεδομένων FLIC φαίνονται στο γράφημα και στο πίνακα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων FLIC φτάνουν το 99% PCKh@0.2 ακρίβεια στον αγκώνα και 97% στον καρπό. Είναι σημαντικό να σημειωθεί ότι αυτά τα αποτελέσματα είναι με βάση τον παρατηρητή , γεγονός που συνάγει με τον τρόπο τον οποίο άλλοι έχουν αξιολογήσει την έξοδο που παρήγαγαν στον FLIC.



Σύγκριση PCK στην βάση δεδομένων MPII [186]

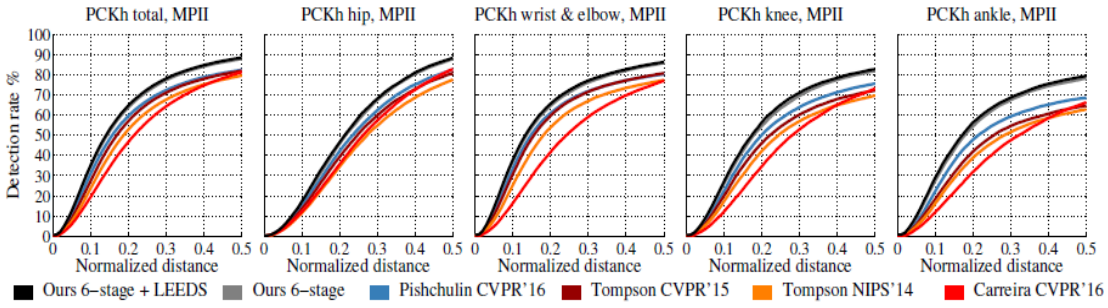
	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Tompson et al.	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira et al.	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Pishchulin et al.	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Hu et al.	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Wei et al.	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Our model	**98.2**	**96.3**	**91.2**	**87.1**	**90.1**	**87.4**	**83.6**	**90.9**

Πίνακας αποτελεσμάτων MPII (PCKh 0.5) [186]

Στα αποτελέσματα για την βάση δεδομένων MPII πέτυχαν αποτελέσματα τελευταίας τεχνολογίας σε όλες τις αρθρώσεις. Τα αποτελέσματα για την βάση δεδομένων MPII φαίνονται στο γράφημα και στο πίνακα παραπάνω. Στα δύσκολα σημεία όπως ο καρπός, οι αγκώνες, τα γόνατα και οι αστράγαλοι βελτιώνονται τα πιο πρόσφατα αποτελέσματα τελευταίας τεχνολογίας κατά μέσο όρο 3.5% (PCKh@0.5) με ένα μέσο ποσοστό σφάλματος 12.8% κάτω από 16.3%. Η τελική ακρίβεια αγκώνα είναι 91.2% και η ακρίβεια καρπού 87.1%.

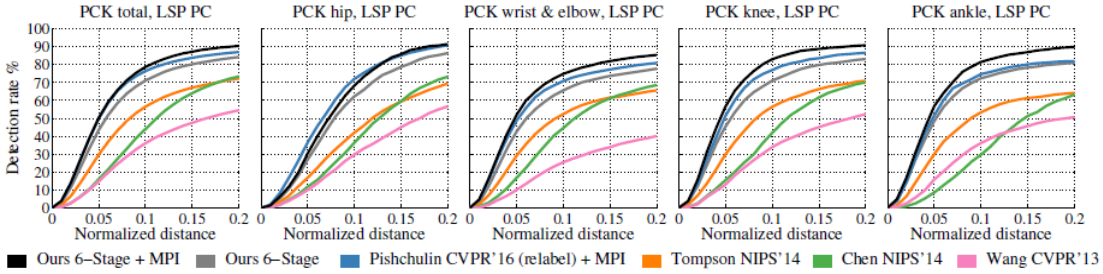
Οι δοκιμές μας έδειξαν ότι το δίκτυο χρειάζεται ένα ποικίλο και προκλητικό σύνολο με στάσεις (όπως το MPII), μαζί με ένα απλό μηχανισμό επαναξιολόγησης των αρχικών προβλέψεων. Επίσης είδαμε πως η ενδιάμεση εποπτεία είναι κρίσιμη για την εκπαίδευση του δικτύου και ότι υπάρχουν και περιπτώσεις που δεν τις χειρίζεται τέλεια το δίκτυο, αλλά συνολικά το σύστημα μας δείχνει ισχυρό σε μια ποικιλία προκλήσεων όπως η ύπαρξη πολλών ανθρώπων σε κοντινή απόσταση. Τέλος είναι σημαντικό να αναφέρουμε ότι η μέθοδος αυτή μπορεί να επεκταθεί και σε πολλούς ανθρώπους.

5. Οι συνελικτικές μηχανές στάσεις χρησιμοποίησαν για τις δοκιμές τις βάσεις δεδομένων MPII, LSP και FLIC.



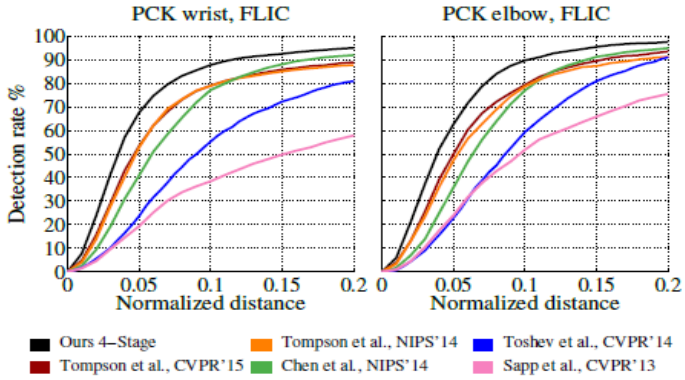
Πίνακας αποτελεσμάτων στην βάση δεδομένων MPII (PCK@0.5) [187]

Τα αποτελέσματα για την βάση δεδομένων MPII φαίνονται στο γράφημα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων MPII φτάνουν στο 87.95% PCKh@0.5 (88.52% όταν προσθέτουμε τα δεδομένα εκπαίδευσης LSP), η οποία είναι 6.11% υψηλότερη από τον πιο κοντινό ανταγωνιστή της και αξίζει να σημειωθεί ότι στον αστράγαλο (το πιο προκλητικό μέρος) η απόδοση μας είναι 78.28% (79.41% όταν προσθέτουμετα δεδομένα εκπαίδευσης LSP), το οποίο είναι 10.76% υψηλότερο από κοντινότερο ανταγωνιστή.



Πίνακας αποτελεσμάτων στην βάση δεδομένων LSP (PCK@0.2) [187]

Τα αποτελέσματα για την βάση δεδομένων LSP φαίνονται στο γράφημα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων LSP φτάνουν 84.32% (90.5% όταν προσθέτουμε τα δεδομένα εκπαίδευσης MPII), απόδοση υψηλότερη από κάθε ανταγωνιστή(μόνο η μέθοδος Pishchulin έρχεται κοντά με 83%).

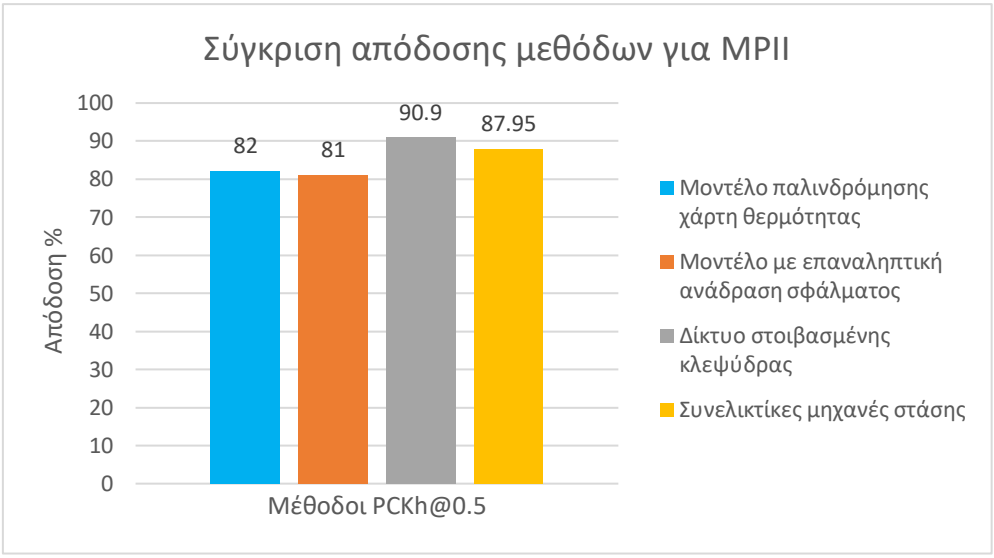


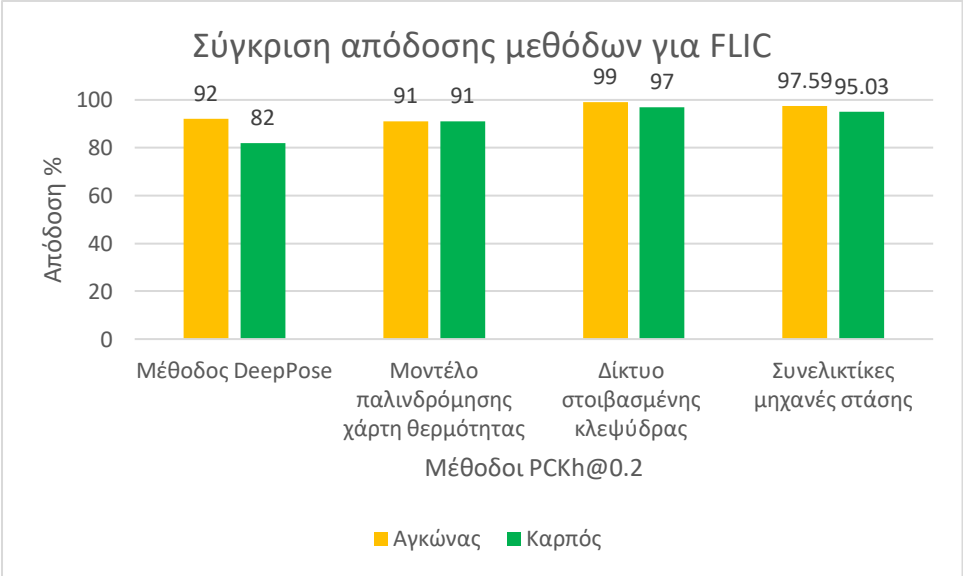
Πίνακας αποτελεσμάτων στην βάση δεδομένων FLIC (PCK@0.2) [187]

Τα αποτελέσματα για την βάση δεδομένων FLIC φαίνονται στο γράφημα παραπάνω. Τα αποτελέσματα και πάλι ξεπερνούν όλες τις προηγούμενες μεθόδους τεχνολογίας αιχμής με απόδοση 97.59% PCKh@0.2 στον αγκώνα και 95.03% στον καρπό. Στις περιοχές υψηλής ακρίβειας το πλεονέκτημα μας είναι ακόμα πιο σημαντικό: 14,8% στους καρπούς, 12,7% στους αγκώνες στο PCKh@0.05.

Οι μηχανές συνέλιξης παρέχουν μια αρχιτεκτονική από άκρο σε άκρο για την αντιμετώπιση δομημένων προβλημάτων πρόβλεψης στην όραση του υπολογιστή χωρίς την ανάγκη για γραφικό πρότυπο αναφοράς. Η αποτυχίες της κυρίως συμβαίνουν όταν είναι πολλοί άνθρωποι σε κοντινή απόσταση. Ο χειρισμός πολλών ατόμων σε μια μόνο από άκρο σε άκρο αρχιτεκτονική είναι επίσης ένα δύσκολο πρόβλημα και μια ενδιαφέρουσα περιοχή για μελλοντικές εργασίες.

Η σύγκριση των μεθόδων που αναλύσαμε στην παρούσα διπλωματική για την εκτίμηση της ανθρώπινης στάσης ενός ατόμου με χρήση των βάσεων δεδομένων MPII και FLIC φαίνονται στα παρακάτω γραφήματα:





Συνεχίζουμε με την αξιολόγηση των τεσσάρων μεθόδων εκτίμησης της στάσης του σώματος πολλαπλών ατόμων. Εδώ πέρα των βάσεων δεδομένων που αναπτύχθηκαν προηγουμένως, θα γίνει και χρήση της βάσης δεδομένων COCO (μια μεγάλη βάση δεδομένων με πάνω από 330000 εικόνες εκ των οποίων οι 200000 είναι επισημασμένες).

1. Το δίκτυο OpenPose χρησιμοποίησε για τις δοκιμές τις βάσεις δεδομένων COCO και MPII.

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	**mAP**	s/image
Subset of 288 images as in [1]									
Deepcut	73.4	71.8	57.9	39.9	56.7	44.0	32.0	54.1	57995
Iqbal et al.	70.0	65.2	56.4	46.1	52.7	47.9	44.5	54.7	10
DeeperCut	87.9	84.0	71.9	63.9	68.8	63.8	58.1	71.2	230
Newell et al.	91.5	87.2	75.9	65.4	72.2	67.0	62.1	74.5	-
ArtTrack	92.2	**91.3**	80.8	71.4	**79.1**	72.6	67.8	**79.3**	**0.005**
Fang et al.	89.3	88.1	80.7	75.5	73.7	**76.7**	**70.0**	79.1	-
Ours	**92.9**	**91.3**	**82.3**	72.6	76.0	70.9	66.8	79.0	**0.005**
Full testing set									
DeeperCut	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5	485
Iqbal et al.	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1	10
Levinko et al.	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6	-
ArtTrack	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3	**0.005**
Fang et al.	88.4	86.5	78.6	**70.4**	74.4	**73.0**	**65.8**	76.7	-
Newell et al.	**92.1**	89.3	78.9	69.8	76.2	71.6	64.7	77.5	-
Fieraru et al.	91.8	**89.5**	**80.4**	69.6	**77.3**	71.7	65.5	**78.0**	-
Ours (one scale)	89.0	84.9	74.9	64.2	71.0	65.6	58.1	72.5	0.005
Ours	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6	**0.005**

Πίνακας αποτελεσμάτων MPII (PCKh 0.5) [188]

Τα αποτελέσματα για την βάση δεδομένων MPII φαίνονται στον πίνακα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων MPII φτάνουν για το υποσύνολο 288 εικόνων το 79%PCKh@0.5 και υπερτερεί των προηγούμενων μεθόδων τεχνολογίας αιχμής κατά 8.5%. Για ολόκληρο το σύνολο δοκιμών MPII η απόδοσης μας φτάνει 75.6 % και υπερτερεί των προηγούμενων μεθόδων τεχνολογίας αιχμής με μεγάλο περιθώριο.

Team	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Top-Down Approaches					
Megvii	78.1	94.1	85.9	74.5	83.3
MRSA	76.5	92.4	84.0	73.0	82.7
The Sea Monsters*	75.9	92.1	83.0	71.7	82.1
Alpha-Pose	71.0	87.9	77.7	69.0	75.2
Mask R-CNN	69.2	90.4	76.0	64.9	76.3
Bottom-Up Approaches					
METU	70.5	87.7	77.2	66.1	77.3
TFMAN*	70.2	89.2	77.0	65.6	76.3
PersonLab	68.7	89.0	75.4	64.1	75.5
Associative Emb	65.5	86.8	72.3	60.6	72.6
Ours	64.2	86.2	70.1	61.0	68.8
Ours	61.8	84.9	67.5	57.1	68.2

Πίνακας αποτελεσμάτων COCO [188]

Τα αποτελέσματα για την βάση δεδομένων COCO φαίνονται στον πίνακα παραπάνω. Αξίζει να σημειωθεί ότι η μέθοδος μας έχει μεγαλύτερη πτώση σε ακρίβεια μόνο όταν εξετάζουμε άτομα με υψηλότερες κλίμακες.

Είδαμε πως σχεδιάστηκε μια αρχιτεκτονική που μαθαίνει από κοινού την ανίχνευση και την σύνδεση των τμημάτων. Επίσης δείχνετε πως ένας άπληστος αλγόριθμος ανάλυσης είναι επαρκής για την παραγωγή υψηλής ποιότητας αναλύσεων της στάσης του σώματος και διατηρεί την αποτελεσματικότητα του ανεξάρτητα από τον αριθμό των ατόμων. Αποδεικνύεται ότι η βελτίωση PAF είναι πολύ σημαντική και οδηγεί σε σημαντική αύξηση τόσο της απόδοσης του χρόνου εκτέλεσης όσο και της απόδοσης. Τέλος δείξαμε ότι συνδυάζοντας την εκτίμηση του σώματος και των ποδιών σε ένα

ενιαίο μοντέλο ενισχύει την ακρίβεια του κάθε στοιχείου ξεχωριστά και μειώνει τον χρόνο συμπεράσματος στην λειτουργία του δικτύου.

2. Η μέθοδος Mask R-CNN χρησιμοποίησε για τις δοκιμές τις βάσεις δεδομένων COCO και MPII.

Team	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Top-Down Approaches					
Megvii	78.1	94.1	85.9	74.5	83.3
MRSA	76.5	92.4	84.0	73.0	82.7
The Sea Monsters*	75.9	92.1	83.0	71.7	82.1
Alpha-Pose	71.0	87.9	77.7	69.0	75.2
Mask R-CNN	69.2	90.4	76.0	64.9	76.3

Πίνακας αποτελεσμάτων COCO [189]

Τα αποτελέσματα για την βάση δεδομένων COCO φαίνονται στον πίνακα παραπάνω. Βλέπουμε ότι τα ποσοστά μας είναι εξίσου υψηλά με μεθόδους τεχνολογίας αιχμής.

3. Η μέθοδος RMPE χρησιμοποίησε για τις δοκιμές τις βάσεις δεδομένων COCO και MPII.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
full testing set								
Iqbal&Gall, ECCVw16	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1
DeeperCut, ECCV16	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5
Levinkov *et al.*, CVPR17	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6
Insafutdinov *et al.*, CVPR17	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3
Cao *et al.*, CVPR17	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6
Newell & Deng, NIPS17	**92.1**	89.3	78.9	69.8	76.2	71.6	64.7	77.5
ours	88.4	86.5	78.6	70.4	74.4	73.0	65.8	76.7
ours++	91.3	**90.5**	**84.0**	**76.4**	**80.3**	**79.9**	**72.4**	**82.1**

Πίνακας αποτελεσμάτων MPII (PCKh 0.5) [190]

Τα αποτελέσματα για την βάση δεδομένων MPII φαίνονται στον πίνακα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων MPII είναι 76%PCKh@0.5 και μάλιστα με ένα μέσο όρο 72% στον εντοπισμό δύσκολων σημείων όπως καρπούς, αγκώνες, αστραγάλους και γόνατα, το οποίο είναι 3.3% μεγαλύτερο από αποτελέσματα προηγουμένων μεθόδων τεχνολογίας αιχμής. Με την χρησιμοποίηση ενός ισχυρότερου ανθρώπινου ανιχνευτή και ενός εκτιμητή θέσεως μπορούμε να επιτύχουμε 82% απόδοση, η οποία είναι 4.6% υψηλότερη από το προηγούμενο καλύτερο αποτέλεσμα.

Team	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
CMU-Pose	61.8	84.9	67.5	57.1	68.2
G-RMI	68.5	87.1	75.5	65.8	73.3
Mask R-CNN	63.1	87.3	68.7	57.8	71.4
Megvii	72.1	91.4	80.0	68.7	77.2
ours	61.8	83.7	69.8	58.6	67.6
ours++	**72.3**	89.2	79.1	68.0	**78.6**

Πίνακας αποτελεσμάτων COCO [190]

Τα αποτελέσματα για την βάση δεδομένων COCO φαίνονται στον πίνακα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων COCO, όπου η μέθοδος μας επιτυγχάνει αποδόσεις τελευταίας τεχνολογίας. Θα πρέπει να σημειωθεί ότι χωρίς συγκεκριμένο σχεδιασμό για το δίκτυο εκτίμησης στάσης, το πλαίσιο της εργασίας μας μπορεί να εκτελέσει στο ίδιο επίπεδο με την μέθοδο Mevgii, που προτείνει ένα νέο δίκτυο για την εκτίμηση της ανθρώπινης στάσης.

Αυτό που παρατηρήσαμε από τις μετρήσεις είναι ότι η RMPE υπερτερεί σημαντικά των μεθόδων αιχμής την εκτίμηση της ανθρώπινης στάσης πολλών ατόμων από την πλευρά της ακρίβειας και της αποτελεσματικότητας.

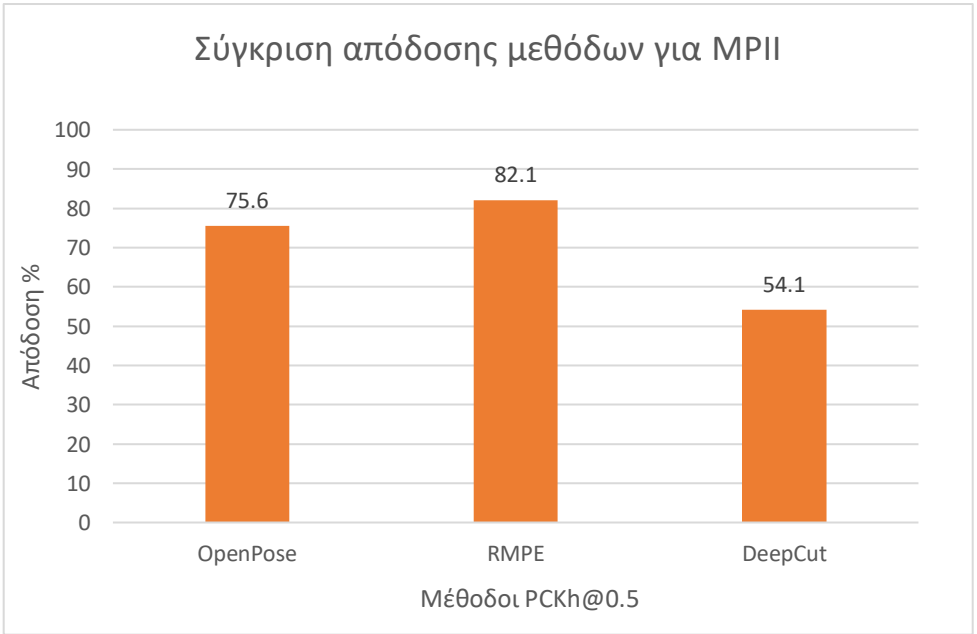
4. Η μέθοδος DeepCut χρησιμοποίησε για τις δοκιμές την βάση δεδομένων MPII.

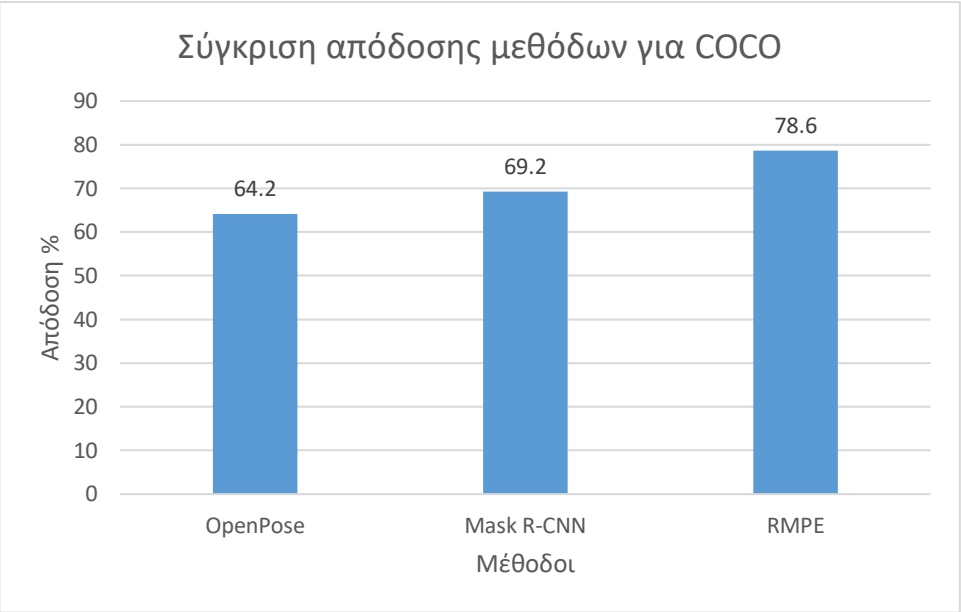
Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	UBody	FBody
AFR-CNN det ROI	71.1	65.8	49.8	34.0	47.7	36.6	20.6	55.2	47.1
AFR-CNN MP	71.8	67.8	54.9	38.1	52.0	41.2	30.4	58.2	51.4
AFR-CNN MP UB	75.2	71.0	56.4	39.6	-	-	-	60.5	-
Dense-CNN det ROI	77.2	71.8	55.9	42.1	53.8	39.9	27.4	61.8	53.2
Dense-CNN MP	73.4	71.8	57.9	39.9	**56.7**	**44.0**	**32.0**	60.7	**54.1**
Dense-CNN MP UB	**81.5**	**77.3**	**65.8**	**50.0**	-	-	-	**68.7**	-
AFR-CNN GT ROI	73.2	66.5	54.6	42.3	50.1	44.3	37.8	59.1	53.1
Dense-CNN GT ROI	78.1	74.1	62.2	52.0	56.9	48.7	46.1	66.6	60.2
Chen&Yuille SP GT ROI	65.0	34.2	22.0	15.7	19.2	15.8	14.2	34.2	27.1

Πίνακας αποτελεσμάτων MPII (PCKh 0.5) [191]

Τα αποτελέσματα για την βάση δεδομένων MPII φαίνονται στον πίνακα παραπάνω. Τα αποτελέσματα στην βάση δεδομένων MPII, φτάνουν 54.1% PCKh@0.5 που είναι εξίσου μεγάλο με άλλες τεχνολογίες αιχμής.

Η σύγκριση των μεθόδων που αναλύσαμε στην παρούσα διπλωματική για την εκτίμηση της ανθρώπινης στάσης πολλαπλών ατόμων με χρήση των βάσεων δεδομένων MPII και FLIC φαίνονται στα παρακάτω γραφήματα:





Όπως είπαμε και παραπάνω κατά την ανάλυση της μεθόδου RMPE, αλλά και σύμφωνα με τα γραφήματα όπου συγκρίνονται οι μέθοδοι που αναπτύχθηκαν στην παρούσα διπλωματική εργασία, βλέπουμε πως η μέθοδος RMPE υπερτερεί και μάλιστα σημαντικά των υπόλοιπων μεθόδων αιχμής τόσο στην χρήση της βάσεως δεδομένων MPII όσο και της COCO, τόσο από πλευρά ακρίβειας όσο και αποτελεσματικότητας. Αντίθετα η μέθοδος DeepCut δείχνει αδύναμη σε σχέση με τις υπόλοιπες. Τα χαμηλά ποσοστά απόδοσης των μεθόδων εκτίμησης της στάσης πολλαπλών ατόμων σε σχέση με τις μεθόδους ενός ατόμου, βασίζονται στο γεγονός ότι οι μέθοδοι αυτοί υλοποιούν ένα πρόβλημα πολύ μεγαλύτερης δυσκολίας και επίσης βρίσκονται σε σχετικά πρώιμο στάδιο σε σχέση με τις άλλες μεθόδους.

5.2 Μελλοντικά βήματα

Σε όλη αυτή την εργασία έχει αποδειχθεί ότι η χρήση πληροφοριών βάθους προκύπτει ως ένας φυσικός τρόπος για να ξεπεραστεί η απώλεια χωρικών πληροφοριών που συμβαίνει όταν εργαζόμαστε μόνο με στατικές μονόφθαλμες εικόνες RGB. Ωστόσο παρά τις βελτιώσεις των δικτύων για την ανίχνευση της ανθρώπινης στάσης του σώματος, εξακολουθεί να υπάρχει μεγάλο περιθώριο βελτίωσης της έρευνας. Το παρόν τμήμα της διπλωματικής εργασίας εξετάζει τις μελλοντικές εργασίες και προκλήσεις που παραμένουν ανοιχτές μετά από την μελέτη που έγινε για την παρούσα εργασία.

- Δεδομένα δικτύου: Ένα μέρος των μελλοντικών εργασιών θα μπορέσει να ασχοληθεί με την γεφύρωση του χάσματος μεταξύ των υφιστάμενων βάσεων δεδομένων και των συνόλων δεδομένων που περιέχουν ανθρώπους. Για το σκοπό αυτό, ένα μεγάλο σύνολο δεδομένων εικόνων RGB-D που περιέχουν ανθρώπους με τις συντεταγμένες των αρθρώσεων τους θα πρέπει να απελευθερωθούν. Αυτό το σύνολο δεδομένων θα μπορούσε να συλλεχθεί από άτομα σε εσωτερικούς και εξωτερικούς χώρους χρησιμοποιώντας το Kinect v2.

- Αρχιτεκτονική του δικτύου: Ένα άλλο κομμάτι προς μελέτη όπως προκύπτει από τις παρούσα διπλωματική εργασία θα ήταν κάποιες μικροαλλαγές στην σχεδίαση της αρχιτεκτονικής του δικτύου που θα οδηγούσαν σε βελτίωση τόσο της απόδοσης του δικτύου αλλά και του χρόνου εκτέλεσης.

Βιβλιογραφία

ΚΕΦΆΛΑΙΟ 1

[1] Pasi Tyrvainen, Minna Silvennoinen, Karlina Talvitie-Lamberg, Anniina Ala-Kitula, Reija Kuoremaki, "Identifying Opportunities for AI applications in Healthcare – Renewing the Natiinal Healthcare and Social Services", IEEE 6th International Conference on Serious Ganes and Applications for Health (SEGAH), 2018

[2] M.A Fiscler and R.A Elschalager " The Representation and Matching of Pictorial Structures", IEEE Transactions on Computers, 1973

[3] Felzenszwalb, Pedro F. and Daniel P. Huttenlocher, " Pictorial Structures for Object Recognition", Int'l Journal of Computer Vision, 2005

[4] Yang,Yi and Deva Ramanan " Articulated pose estimation with flexible mixtures-of-parts", IEEE Conf. on Computer Vision and Pattern Recognition", 2011

[5] Tian, Yuondong, C. Lawrence Zitnick, and Srinivasa G. Narasimhan, "Exploring the Spatial Hierarchy of Mixture Models for Human Pose Estimation", European Conf. on Computer Vision, 2012

[6] Sapp B. and Taskar B. ," MODEC: Multimodal Decomposable Models for Human Pose Estimation", IEEE Conf. on Computer Vision and Pattern Recognition, 2013

[7] T. S. Huang, "Computer Vision: Evolution and Promise", 5th International conference High Technology: Imaging and technology in Japan, 1996, page 13-20

[8] S. O. Haykin, Neural networks and Learning Machines

[9] J. Patterson and A. Gibson, Deep Learning: A practitioner's approach

ΚΕΦΆΛΑΙΟ 2

[10] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," Computer vision and image understanding, vol. 81, no. 3, pp. 231–268, 2001

[11] R. Poppe, "Vision-based human motion analysis: An overview," Computer vision and image understanding, vol. 108, no. 1, pp. 4–18, 2007

[12] D. Hogg, "Model-based vision: a program to see a walking person," Image and Vision computing, vol. 1, no. 1, pp. 5–20, 1983.

[13] D. C. Hogg, “Interpreting images of a known moving object,” Ph.D. dissertation, University of Sussex, 1984

[14] T. B. Moeslund and E. Granum, “3d human pose estimation using 2d-data and an alternative phase space representation,” Procedure Humans 2000, 2000.

[15] K. Rohr, “Human movement analysis based on explicit motion models,” in Motion-based recognition. Springer, 1997, pp. 171–198.

[16] V. Pavlović, J. M. Rehg, T.-J. Cham, and K. P. Murphy, “A dynamic bayesian network approach to figure tracking using learned dynamic models,” in Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 1. IEEE, 1999, pp. 94–101.

[17] D. M. Gavrila and L. S. Davis, “3-d model-based tracking of humans in action: a multi-view approach,” in Computer Vision and Pattern Recognition, 1996. Proceedings CVPR’96, 1996 IEEE Computer Society Conference on. IEEE, 1996, pp. 73–80.

[18] Y. Kameda, M. Minoh, and K. Ikeda, “Three dimensional pose estimation of an articulated object from its silhouette image,” in Asian Conference on Computer Vision, 1993, pp. 612–615

[19] C. Hu, Q. Yu, Y. Li, and S. Ma, “Extraction of parametric human model for posture recognition using genetic algorithm,” in Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE, 2000, pp. 518–523

[20] N. Jojic, J. Gu, H. C. Shen, and T. Huang, “3-d reconstruction of multipart self-occluding objects,” in Computer Vision-ACCV1998. Springer, 1997, pp. 455–462

[21] I. Kakadiaris and D. Metaxas, “Vision-based animation of digital humans,” in Computer Animation 98. Proceedings. IEEE, 1998, pp. 144–152.

[22] D. Meyer, J. Denzler, and H. Niemann, “Model based extraction of articulated objects in image sequences for gait analysis,” in Image Processing, 1997. Proceedings., International Conference on, vol. 3. IEEE, 1997, pp. 78–81.

[23] C. Bregler, J. Malik, and K. Pullen, “Twist based acquisition and tracking of animal and human kinematics,” International Journal of Computer Vision, vol. 56, no. 3, pp. 179–194, 2004

[24] S. Wachter and H.-H. Nagel, “Tracking of persons in monocular image sequences,” in Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE. IEEE, 1997, pp. 2–9.

[25] C. Sminchisescu and B. Triggs, “Estimating articulated human motion with covariance scaled sampling,” The International Journal of Robotics Research, vol. 22, no. 6, pp. 371–391, 2003.

[26] ——, “Kinematic jump processes for monocular 3d human tracking,” in Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 1. IEEE, 2003, pp. I–69

[27] M. W. Lee and I. Cohen, “Proposal maps driven mcmc for estimating human body pose in static images,” in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2. IEEE, 2004, pp. II–334.

[28] T. Moeslund, Pose estimating the human arm using kinematics and the sequential monte carlo framework. INTECH Open Access Publisher, 2005.

[29] T. B. Moeslund, C. B. Madsen, and E. Granum, “Modelling the 3d pose of a human arm and the shoulder complex utilising only two parameters,” Integrated Computer-Aided Engineering, vol. 12, no. 2, pp. 159–175, 2005.

[30] R. Navaratnam, A. Thayananthan, P. H. Torr, and R. Cipolla, “Hierarchical part-based human body pose estimation.” in BMVC, 2005.

[31] Q. Delamarre and O. Faugeras, “3d articulated models and multiview tracking with physical forces,” Computer Vision and Image Understanding, vol. 81, no. 3, pp. 328–357, 2001.

[32] P. Fua et al., “Articulated soft objects for multiview shape and motion capture,” IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 9, pp. 1182–1187, 2003.

[33] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking,” IEEE Transactions on Signal Processing, vol. 50, no. 2, pp. 174–188, 2002.

[34] A. Zhu, H. Snoussi, and A. Cherouat, “Articulated human motion tracking with foreground learning,” in Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European. IEEE, 2014, pp. 366–370.

[35] J. MacCormick and M. Isard, “Partitioned sampling, articulated objects, and interface-quality hand tracking,” in Computer Vision–ECCV 2000. Springer, 2000, pp. 3–19.

[36] J. Deutscher, A. Blake, and I. Reid, “Articulated body motion capture by annealed particle filtering,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2000, pp. 126–133.

[37] J. Deutscher and I. Reid, “Articulated body motion capture by stochastic search,” International Journal of Computer Vision, vol. 61, no. 2, pp. 185–205, 2005.

[38] L. Sigal, A. O. Balan, and M. J. Black, “Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated

human motion," International Journal of Computer Vision, vol. 87, no. 1-2, pp. 4–27, 2010.

[39] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," ACM Transactions on Graphics (TOG), vol. 22, no. 3, pp. 569–577, 2003.

[40] R. Kehl, M. Bray, and L. Van Gool, "Full body tracking from multiple views using stochastic sampling," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 2. IEEE, 2005, pp.129–136

[41] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in visionbased human motion capture and analysis," Computer vision and image understanding, vol. 104, no. 2, pp. 90–126, 2006.

[42] H. Sidenbladh and M. J. Black, "Learning image statistics for bayesian tracking," in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, vol. 2. IEEE, 2001, pp. 709–716.

[43] H. Sidenbladh, M. J. Black, and D. J. Fleet, "Stochastic tracking of 3d human figures using 2d image motion," in Computer Vision–ECCV 2000. Springer, 2000, pp. 702–718.

[44] H. Sidenbladh, M. J. Black, and L. Sigal, "Implicit probabilistic models of human motion for synthesis and tracking," in Computer Vision–ECCV 2002. Springer, 2002, pp. 784–800

[45] H. Sidenbladh and M. J. Black, "Learning the statistics of people in images and video," International Journal of Computer Vision, vol. 54, no. 1-3, pp. 183–209, 2003.

[46] I. Karaulova, P. M. Hall, and A. D. Marshall, "A hierarchical model of dynamics for tracking people with a single video camera." in BMVC, 2000, pp. 1–10.

[47] A. Agarwal and B. Triggs, "Tracking articulated motion with piecewise learned dynamical models," in European Conference on Computer Vision, vol. 3, 2004, pp. 54–65.

[48] N. R. Howe, M. E. Leventon, and W. T. Freeman, "Bayesian reconstruction of 3d human motion from single-camera video." in NIPS, vol. 99, 1999, pp. 820–6.

[49] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard, "Tracking looselimbed people," in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 1. IEEE, 2004, pp. I–421.

[50] R. Urtasun and P. Fua, "3d human body tracking using deterministic temporal motion models," in Computer Vision-ECCV 2004. Springer, 2004, pp. 92–

[51] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua, “Priors for people tracking from small training sets,” in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 1. IEEE, 2005, pp. 403–410.

[52] R. Urtasun, D. J. Fleet, and P. Fua, “3d people tracking with gaussian process dynamical models,” in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 1. IEEE, 2006, pp. 238–245.

[53] J. Wang, A. Hertzmann, and D. M. Blei, “Gaussian process dynamical models,” in Advances in neural information processing systems, 2005, pp. 1441–1448.

[54] K. Moon and V. Pavlović, “Impact of dynamics on subspace embedding and tracking of sequences,” in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 1. IEEE, 2006, pp. 198–205.

[55] G. Mori, X. Ren, A. Efros, J. Malik et al., “Recovering human body configurations: Combining segmentation and recognition,” in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2. IEEE, 2004, pp. II–326.

[56] D. A. Forsyth and M. M. Fleck, “Body plans,” in Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on. IEEE, 1997, pp. 678–683.

[57] P. F. Felzenszwalb and D. P. Huttenlocher, “Pictorial structures for object recognition,” International Journal of Computer Vision, vol. 61, no. 1, pp. 55–79, 2005.

[58] M. Eichner and V. Ferrari, “Better appearance models for pictorial structures,” in Proceedings of the British Machine Vision Conference (BMVC), September 2009.

[59] M. A. Fischler and R. A. Elschlager, “The representation and matching of pictorial structures,” IEEE Transactions on computers, no. 1, pp. 67–92, 1973.

[60] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient matching of pictorial structures,” in Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, vol. 2. IEEE, 2000, pp. 66–73.

[61] X. Lan and D. P. Huttenlocher, “Beyond trees: Common-factor models for 2d human pose recovery,” in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 1. IEEE, 2005, pp. 470–477.

[62] R. Ronfard, C. Schmid, and B. Triggs, “Learning to parse pictures of people,” in Computer Vision–ECCV 2002. Springer, 2002, pp. 700–714.

[63] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proceedings of IEEE Conference on Computer Vision and Pattern

Recognition (CVPR), vol. 1, June 2005, pp. 886–893 vol. 1.

[64] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

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

[66] Y. Yang and D. Ramanan, “Articulated pose estimation with flexible mixturesof-parts,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1385–1392.

[67] Y. Wang, D. Tran, and Z. Liao, “Learning hierarchical poselets for human parsing,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, ser. CVPR ’11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1705–1712.

[68] M. Sun and S. Savarese, “Articulated part-based model for joint object detection and pose estimation,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ser. ICCV ’11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 723–730.

[69] F. Wang and Y. Li, “Beyond physical connections: Tree models in human pose estimation,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 596–603.

[70] B. Sapp and B. Taskar, “Modec: Multimodal decomposable models for human pose estimation,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3674–3681.

[71] A. Zhu, H. Snoussi, and A. Cherouat, “Articulated pose estimation via multiple mixture parts model,” in Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on. IEEE, 2015, pp. 1–5.

[72] A. Zhu, H. Snoussi, T. Wang, and A. Cherouat, “Human pose estimation with multiple mixture parts model based on upper body categories,” Journal of Electronic Imaging, vol. 24, no. 4, p. 043021, 2015.

[73] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, “Exploring the spatial hierarchy of mixture models for human pose estimation,” in Proceedings of European Conference on Computer Vision (ECCV), 2012, pp. 256–269.

[74] D. Park and D. Ramanan, “N-best maximal decoders for part models,” in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 2627–2634.

[75] N. Ukita, “Articulated pose estimation with parts connectivity using discriminative local oriented contours,” in Proceedings of IEEE Conference

on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3154–3161.

[76] L. Bourdev and J. Malik, “Poselets: Body part detectors trained using 3d human pose annotations,” in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 1365–1372.

[77] L. Bourdev, S. Maji, T. Brox, and J. Malik, “Detecting people using mutually consistent poselet activations,” in Computer Vision–ECCV 2010. Springer, 2010, pp. 168–181.

[78] L. Bourdev, S. Maji, and J. Malik, in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 1543–1550.

[79] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, “Poselet conditioned pictorial structures,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2013, pp. 588–595.

[80] V. Ramakrishna, D. Munoz, M. Hebert, J. A. D. Bagnell, and Y. A. Sheikh, “Pose machines: Articulated pose estimation via inference machines,” in Proceedings of European Conference on Computer Vision (ECCV), no. CMU-RITR-, July 2014.

[81] Y. Chen, L. Zhu, C. Lin, H. Zhang, and A. L. Yuille, “Rapid inference on a novel and/or graph for object detection, segmentation and parsing,” in Advances in Neural Information Processing Systems, 2007, pp. 289–296.

[82] K. Duan, D. Batra, and D. J. Crandall, “A multi-layer composite model for human pose estimation.” in BMVC, 2012, pp. 1–11.

[83] D. Ramanan, D. A. Forsyth, and A. Zisserman, “Strike a pose: Tracking people by finding stylized poses,” in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 271–278.

[84] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, “2d articulated human pose estimation and retrieval in (almost) unconstrained still images,” International Journal of Computer Vision, vol. 99, 2012.

[85] B. Sapp, D. Weiss, and B. Taskar, “Parsing human motion with stretchable models,” in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, pp. 1281–1288.

[86] K. Fragkiadaki, H. Hu, and J. Shi, “Pose from flow and flow from pose,” in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013, pp. 2059–2066.

[87] A. Cherian, J. Mairal, K. Alahari, and C. Schmid, “Mixing body-part sequences for human pose estimation,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, ÉtatsUnis: IEEE, Jun. 2014.

[88] L. Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Handwritten digit recognition with a back-propagation network,” in Advances in Neural Information Processing Systems. Morgan Kaufmann, 1990, pp. 396–404.

[89] Y. LeCun, F. J. Huang, and L. Bottou, “Learning methods for generic object recognition with invariance to pose and lighting,” in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2, 2004, pp. II–97–104 Vol.2.

[90] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, “What is the best multi-stage architecture for object recognition?” in Proc. International Conference on Computer Vision (ICCV’09). IEEE, 2009.

[91] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” in Proceedings of the 26th Annual International Conference on Machine Learning, ser. ICML ’09. ACM, 2009, pp. 609–616.

[92] C. Szegedy, A. Toshev, and D. Erhan, “Deep neural networks for object detection,” in Advances in Neural Information Processing Systems 26, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., 2013, pp. 2553–2561.

[93] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” in International Conference on Learning Representations (ICLR 2014). CBLS, April 2014.

[94] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[95] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, “Selective search for object recognition,” vol. 104, no. 2. Springer US, 2013, pp. 154–171.

[96] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-based r-cnns for fine-grained category detection,” in Computer Vision ECCV 2014, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8689. Springer International Publishing, 2014, pp. 834–849.

[97] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” in Computer Vision–ECCV 2014. Springer, 2014, pp. 346–361.

[98] X. Chen and A. Yuille, “Articulated pose estimation by a graphical model with image dependent pairwise relations,” in Advances in Neural Information Processing Systems (NIPS), 2014.

[99] A. Toshev and C. Szegedy, “Deeppose: Human pose estimation via deep neural networks,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1653–1660.

[100] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, “Real-time continuous pose recovery of human hands using convolutional networks,” ACM Trans. Graph., vol. 33, no. 5, pp. 169:1–169:10, Sep. 2014.

[101] R. Fergus, G. Williams, I. Spiro, C. Bregler, and G. W. Taylor, “Pose-sensitive embedding by nonlinear nca regression,” in Advances in Neural Information Processing Systems, 2010, pp. 2280–2288.

[102] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, “Learning human pose estimation features with convolutional networks,” in International Conference on Learning Representations (ICLR), April 2014.

[103] J. J. Tompson, A. Jain, Y. Lecun, and C. Bregler, “Joint training of a convolutional network and a graphical model for human pose estimation,” in Advances
in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1799–1807.

[104] X. Fan, K. Zheng, Y. Lin, and S. Wang, “Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[105] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[106] A. Jain, J. Tompson, Y. LeCun, and C. Bregler, “Modeep: A deep learning framework using motion features for human pose estimation,” in Computer Vision–ACCV 2014. Springer, 2014, pp. 302–315.

[107] T. Pfister, K. Simonyan, J. Charles, and A. Zisserman, “Deep convolutional neural networks for efficient pose estimation in gesture videos,” in Computer Vision–ACCV 2014. Springer, 2015, pp. 538–552.

[108] T. Pfister, J. Charles, and A. Zisserman, “Flowing convnets for human pose estimation in videos,” arXiv preprint arXiv:1506.02897, 2015.

[109] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” arXiv preprint arXiv:1411.4038, 2014.

ΚΕΦΆΛΑΙΟ 3

[110] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[111] C. Szegedy, A. Toshev, and D. Erhan. Object detection via deep neural networks. In *NIPS 26*, 2013

[112] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[113] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *COLT*. ACL, 2010.

[114]J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Join training of a convolutional network and a graphical model for human pose estimation. NIPS, 2014.

[115] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In CVPR, 2013

[116] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. 2014.

[117] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

[118] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. S¨ackinger, and R. Shah. Signature verification using a siamese time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence, 1993

[119] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradientbased learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[120] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics,36(4):193–202, 1980

[121] P. Agrawal, D. Stansbury, J. Malik, and J. L. Gallant. Pixels to voxels: Modeling visual representation in the human brain. arXiv preprint arXiv:1407.5104, 2014

[122] D. J. Felleman and D. C. Van Essen. Distributed Hierarchical Processing in the Primate Cerebral Cortex. Cerebral Cortex, 1(1):1–47, Jan. 1991

[123] B. C. U. L. M. M. Kravitz DJ, Saleem KS. The ventral visual pathway: An expanded neural framework for the processing of object quality. volume 17(1), 2013.

[124] V. A. F. Lamme and P. R. Roelfaema. the distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences, 23:571, 2000

[125] Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems. (2014) 1799{1807

[126] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3431{3440

[127] Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1395{1403

[128] Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines.
Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on (2016)

[129] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015)

[130] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Computer Vision and Pattern Recognition, 2016. CVPR 2016. IEEE Conference on (2015)

[131] Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on (2016)

[132] P_ster, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1913{1921

[133] V. Ramakrishna, D. Munoz, M. Hebert, J. Bagnell, and Y. Sheikh. Pose Machines: Articulated Pose Estimation via Inference Machines. In ECCV, 2014.

[134] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. arXiv preprint arXiv:1511.06645, 2015.

[135] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In CVPR, 2015

[136] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In NIPS, 2014.

[137] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In CVPR, 2013.

[138] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[139] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 1994.

[140] D. Bradley. Learning In Modular Systems. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 2010.

[141] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In AISTATS, 2010.

ΚΕΦΆΛΑΙΟ 4

[142] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, "Mask r-cnn," in ICCV, 2017

[143] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multiperson pose estimation," in ICCV, 2017

[144] L. Pishchulin, A. Jain, M. Andriluka, T. Thorm¨ahlen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in CVPR, 2012.

[145] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "Using kposelets for detecting people and localizing their keypoints," in CVPR, 2014.

[146] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in ICCV, 2011.

[147] U. Iqbal and J. Gall, "Multi-person pose estimation with local jointto- person associations," in ECCV Workshop, 2016.

[148] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in CVPR, 2017.

[149] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in CVPR, 2018.

[150] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in ECCV, 2018.

[151] M. Eichner and V. Ferrari, "We are family: Joint pose estimation of multiple persons," in ECCV, 2010.

[152] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in CVPR, 2016.

[153] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in ECCV, 2016.

[154] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, 2016.

[155] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in CVPR, 2016.

[156] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in CVPR, 2017.

[157] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, “Densely connected convolutional networks,” in CVPR, 2017.

[158] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in ICLR, 2015.

[159] D. B. West et al., Introduction to graph theory. Prentice hall Upper Saddle River, 2001, vol. 2.

[160] H. W. Kuhn, “The hungarian method for the assignment problem,” in Naval research logistics quarterly. Wiley Online Library, 1955.

[161] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015

[162] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV. 2014.

[163] R. Girshick. Fast R-CNN. In ICCV, 2015

[164] P. O. Pinheiro, R. Collobert, and P. Dollar. Learning to segment object candidates. In NIPS, 2015

[165] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In CVPR, 2016.

[166] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In CVPR, 2017.

[167] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014

[168] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[169] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Doll´ar. Learning to refine object segments. In ECCV, 2016

[170] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In NIPS, 2015

[171] X. Chen and A. L. Yuille. Parsing occluded people by flexible compositions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3945–3954, 2015.

[172] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3582–3589, 2014

[173] J. G. Umar Iqbal. Multi-person pose estimation with local joint-to-person associations. In *European Conference on Computer Vision Workshops 2016 (ECCVW'16) – Workshop on Crowd Understanding (CUW'16)*, 2016.

[174] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[175] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In *European Conference on Computer Vision (ECCV)*, May 2016

[176] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2016

[177] L. Pishchulin, A. Jain, M. Andriluka, T. Thorm¨ahlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3178–3185, 2012.

[178] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *arXiv preprint arXiv:1603.06937*, 2016

[179] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Conference on Neural Information Processing Systems (NIPS)*, pages 2017–2025, 2015

[180] X. Burgos-Artizzu, D. Hall, P. Perona, and P. Dollar. Merging pose estimates across space and time. In *British Machine Vision Conference (BMVC)*, 2013.

[181] S. Chopra and M. Rao. The partition problem. Mathematical Programming, 59(1–3):87–115, 1993

[182] Leonid Pischulin, Eldar Insafutdinov, Siyu Tang. Deepcut: Joint Subset Partition and Labeling for Multi Person Pose Estimation, Stanford University

[183] Alexander Toshev, Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks, IEEE

[184] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, Christopher Bregler, Efficient Object Localization Using Convolutional Networks, arxiv.org

[185] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, Jitendra Malik, Human Pose Estimation with Iterative Error Feedback, arxiv.org

[186] Alejandro Newell, Kaiyu Yang, Jia Deng, Stacked Hourglass Networks for Human Pose Estimation, arxiv.org

[187] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh, Convolutional Pose Machines, arxiv.org

[188] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, arxiv.org

[189] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, Mask R-CNN, arxiv.org

[190] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, Cewu Lu, RMPE: Regional Multi-person Pose Estimation, arxi.org

[191] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, Bernt Schiele, DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation, arxiv.org