Αναγνώριση χειρονομίας σε βίντεο

Η Μεταπτυχιακή Διπλωματική Εργασία

υποβάλλεται στην ορισθείσα
από τη Συνέλευση
του Τμήματος Μηχανικών Η/Υ και Πληροφορικής
Εξεταστική Επιτροπή

από τον

Δημήτριο Έξαρχο

ως μέρος των υποχρεώσεων για την απόκτηση του

ΔΙΠΛΩΜΑΤΟΣ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗ ΜΗΧΑΝΙΚΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΜΕ ΕΙΔΙΚΕΥΣΗ
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΚΑΙ ΜΗΧΑΝΙΚΗ ΔΕΔΟΜΕΝΩΝ

Πανεπιστήμιο Ιωαννίνων
Πολυτεχνική Σχολή
Ιωάννινα 2024

Εξεταστική επιτροπή:

- **Λυσίμαχος-Παύλος Κόντης,** Καθηγητής Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων (Επιβλέπων)
- **Κωνσταντίνος Παρσόπουλος,** Καθηγητής Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
- **Χριστόφορος Νίκου,** Καθηγητής Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων

ΑΦΙΕΡΩΣΗ

Στην οικογένειά μου.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις ευχαριστίες μου προς τον καθηγητή Κ. Κόντη και την καθηγήτρια κ. Τσιλιγιάννη, οι οποίοι, συνέβαλαν σημαντικά στην επιτυχή ολοκλήρωση αυτής της διατριβής.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΕΡΙΛΗΨΗ

Δημήτριος Έξαρχος, Δ.Μ.Σ. στη Μηχανική Δεδομένων και Υπολογιστικών Συστημάτων, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων, Φεβρουάριος 2024
Αναγνώριση χειρονομίας σε βίντεο
Επιβλέπων: Λυσίμαχος-Παύλος Κόντης Καθηγητής

Οι επιτυχημένες προσπάθειες στην έρευνα αναγνώρισης χειρονομιών κατά τις τελευταίες δύο δεκαετίες άνοιξαν τον δρόμο για συστήματα αλληλεπίδρασης ανθρώπου-υπολογιστή. Ανεπίλυτες προκλήσεις όπως η αξιόπιστη ταυτοποίηση της φάσης μιας χειρονομίας, την ευαισθησία της σε μεταβολές μεγέθους, σχήματος και ταχύτητας, καθώς και θέματα εμπόδισης της χειρονομίας, διατηρούν την έρευνα για την αναγνώριση χειρονομιών εξαιρετικά ενεργή. Παρέχουμε μια επισκόπηση των αλγορίθμων αναγνώρισης χειρονομιών βασισμένων σε οπτικά συστήματα που αναφέρθηκαν τα τελευταία χρόνια, καθώς τονίζουμε ιδιαίτερα τις τεχνικές που αφορούν το τρισδιάστατο συνελικτικό δίκτυο. Η εργασία επίσης ανασκοπεί 23 δημόσιες και διαθέσιμες βάσεις δεδομένων χειρονομιών και παρέχει συνδέσμους για τη λήψη τους.

EXTENDED ABSTRACT

Dimitrios Exarchos, M.Sc. in Data and Computer Systems Engineering, Department of Computer Science and Engineering, School of Engineering, University of Ioannina, Greece, February 2024
Hand Gesture Recognition
Advisor: Lysimachos-Pavlos Kontis, Professor

The field of hand gesture recognition has witnessed significant advancements in the last two decades, marking substantial progress towards more intuitive human-computer interaction systems. Despite these advancements, the domain faces ongoing challenges such as the precise identification of gesturing phases, the accommodation of variations in gesture size, shape, and speed, and the difficulties presented by occlusion. This paper reviews the landscape of vision-based hand gesture recognition algorithms, highlighting contributions made from the late 1990s through the recent years, with a special focus on the emergence and impact of three-dimensional convolutional networks (3D CNNs) on gesture recognition.

Three-dimensional convolutional networks have revolutionized the way spatial-temporal data is processed, offering robust frameworks for capturing the dynamic nature of gestures through both spatial and temporal dimensions simultaneously. These models excel in learning from depth and motion information, addressing some of the traditional challenges in gesture recognition by providing superior invariance to scale, viewpoint, and lighting conditions. The adaptability of 3D CNNs to various gesture recognition scenarios, including sign language interpretation, medical rehabilitation, and interactive gaming, underscores their potential in enhancing gesture-based interfaces.

Furthermore, this paper presents an exhaustive review of 23 publicly available hand gesture databases, which are critical for the development, testing, and benchmarking of gesture recognition algorithms. These databases encompass a wide range of gestures, from simple hand movements to complex sign languages, captured under diverse

conditions to reflect real-world scenarios. The availability of such comprehensive datasets, along with their download links, is instrumental for researchers and practitioners in the field, facilitating the development of more accurate and generalizable gesture recognition models.

In conclusion, while significant progress has been made in hand gesture recognition, the field continues to evolve rapidly, driven by advances in computational models like 3D CNNs and the growing availability of diverse gesture databases. The challenges of gesture phase identification, variability in gesture appearance, and occlusion remain pertinent areas for future research. Addressing these challenges will not only enhance the robustness of gesture recognition systems but also broaden their applicability in creating more natural and intuitive human-computer interfaces.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Στόχοι

Η μη λεκτική επικοινωνία, η οποία περιλαμβάνει την επικοινωνία μέσω των χειρονομιών, των σωματικών στάσεων και των προσωπικών εκφράσεων, αποτελεί περίπου τα δύο τρίτα όλης της επικοινωνίας ανάμεσα στους ανθρώπους [1]. Οι χειρονομίες είναι μία από τους πιο κοινούς τρόπους επικοινωνίας μέσω της σωματικής γλώσσας και χρησιμοποιούνται για επικοινωνία και αλληλεπίδραση. Ενώ το υπόλοιπο σώμα δηλώνει μια πιο γενική συναισθηματική κατάσταση, οι χειρονομίες μπορούν να έχουν συγκεκριμένο γλωσσικό περιεχόμενο [2]. Λόγω της ταχύτητας και της εκφραστικότητας στην αλληλεπίδραση, οι χειρονομίες χρησιμοποιούνται ευρέως σε γλώσσες νοηματικής και σε συστήματα αλληλεπίδρασης ανθρώπου-υπολογιστή.

Ένας συνεχής στόχος στον σχεδιασμό δι' επαφών ανθρώπου-μηχανής είναι να επιτρέψει την αποτελεσματική και συναρπαστική αλληλεπίδραση. Για παράδειγμα, τα συστήματα αναγνώρισης χειρονομιών βασισμένα σε οπτικά συστήματα (HGR) μπορούν να επιτρέψουν την αλληλεπίδραση χωρίς επαφή σε αποστειρωμένα περιβάλλοντα όπως χειρουργικές αίθουσες νοσοκομείων, ή να παρέχουν ελκυστικές εφαρμογές στους

χώρους της ψυχαγωγίας και των παιχνιδιών. Ωστόσο, το HGR δεν είναι τόσο αξιόπιστο όσο η παραδοσιακή αλληλεπίδραση με πληκτρολόγιο ή το ποντίκι. Ζητήματα όπως η ευαισθησία στις μεταβολές του μεγέθους και της ταχύτητας, η κακή απόδοση απέναντι σε περίπλοκα φόντα και στις διαφορετικές συνθήκες φωτισμού, καθώς και η αξιοπιστία της ανίχνευση φάσης της χειρονομίας, έχουν περιορίσει τη χρήση των χειρονομιών ως έμπιστη μορφή αλληλεπίδρασης στον σχεδιασμό δι' επαφών.

1.2 Ταξινόμηση των Χειρονομιών

Υπάρχουν πολλοί τρόποι για να κατηγοριοποιηθούν οι χειρονομίες, με βάση τα παρατηρήσιμα χαρακτηριστικά και με βάση την ερμηνεία. Στην πρώτη κατηγορία, οι χειρονομίες ταξινομούνται με βάση τις χρονικές σχέσεις, σε δύο τύπους: στατικές και δυναμικές χειρονομίες (Εικ.1). Οι στατικές χειρονομίες (ή αλλιώς στάσεις/θέσεις χεριών) είναι εκείνες στις οποίες η θέση του χεριού δεν αλλάζει κατά τη διάρκεια της χειρονομίας. Οι στατικές χειρονομίες βασίζονται κυρίως στο σχήμα και τις γωνίες κάμψης των δακτύλων. Στις δυναμικές χειρονομίες, η θέση του χεριού αλλάζει συνεχώς σε σχέση με τον χρόνο. Οι δυναμικές χειρονομίες γενικά έχουν τρεις φάσεις κίνησης: προετοιμασία, κύρια κίνηση και φάση ανάκλησης [3]. Το μήνυμα σε μια δυναμική χειρονομία περιέχεται κυρίως στη φάσης της κύριας κίνησης. Οι δυναμικές χειρονομίες βασίζονται, εκτός από το σχήμα και τις γωνίες κάμψης των δακτύλων και στις τροχιές και στους προσανατολισμούς του χεριού.

Στη δεύτερη κατηγορία, οι χειρονομίες ταξινομούνται με βάση το τι ερμηνεύουν. Για παράδειγμα οι συμβολικές, οι επεξηγηματικές, οι ρυθμιστικές, οι συναισθηματικής έκφρασης, και οι προσαρμογείς [4,5] είναι οι τυπικές κατηγορίες για την περιγραφή των χειρονομιών. Οι συμβολικές (επίσης χαρακτηρισμένα ως αυτόνομες χειρονομίες) είναι χειρονομίες που μπορούν να υποκαταστήσουν τα λεκτικά λόγια (για παράδειγμα, δείχνοντας τον αντίχειρα προς τα πάνω αντί να πει κανείς ότι όλα είναι εντάξει). Οι επεξηγηματικές είναι χειρονομίες που χρησιμοποιούνται για να εξηγήσουν λεκτικές λέξεις (για παράδειγμα, δίνοντας οδηγίες δείχνοντας). Οι ρυθμιστικές υποστηρίζουν την αλληλεπίδραση και την επικοινωνία μεταξύ ομιλητή και ακροατή (για παράδειγμα, σηκώνοντας το χέρι για να γίνει διαχείριση στην σειρά της ομιλίας). Οι συναισθηματικής έκφρασης είναι προσωπικές εκφράσεις, οι οποίες όταν συνδυάζονται με στάσεις αντανακλούν την ένταση μιας χειρονομίες (για παράδειγμα, κοιτώντας ένα αντικείμενο και μετακινώντας το σώμα προς τα πίσω αντανακλά το συναίσθημα του φόβου). Οι προσαρμογείς είναι

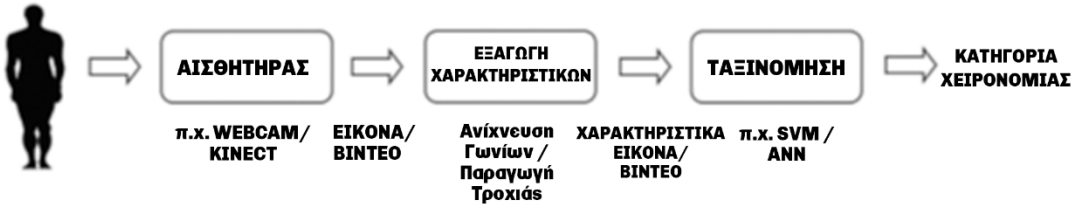
χειρονομίες που χρησιμοποιούνται κάποια στιγμή για προσωπική ευκολία, αλλά έχουν γίνει συνήθεια (για παράδειγμα, επανατοποθετώντας τα γυαλιά στην θέση τους σε μια τεταμένη κατάσταση).

Εικόνα 1:Ταξινόμηση των χειρονομιών βάσει της χρονικής τους φύσης. Οι στατικές χειρονομίες είναι ανεξάρτητες του χρόνου, ενώ οι δυναμικές χειρονομίες εξαρτώνται από αυτόν.

1.3 Αναγνώριση Χειρονομιών

Η Εικ. 2 δείχνει το διάγραμμα ενός τυπικού οπτικού συστήματος αναγνώρισης χειρονομιών. Ο αισθητήρας στα οπτικά συστήματα είναι μια κάμερα. Ο Berman και άλλοι [6] εξέτασαν διάφορους αισθητήρες που χρησιμοποιούνται σε συστήματα αναγνώρισης χειρονομιών και παρείχαν μια εκτενή ανάλυση της ενσωμάτωσης των αισθητήρων στα συστήματα αναγνώρισης χειρονομιών και την επίδρασή τους στην απόδοση του οπτικού συστήματος. Με βάση την εξαγωγή των χαρακτηριστικών, τα συστήματα αναγνώρισης χειρονομιών χωρίζονται σε δύο κατηγορίες: μεθόδους βασισμένες στην εμφάνιση και

μεθόδους βασισμένες σε τρισδιάστατα (3D) μοντέλα χεριού. Οι μέθοδοι βασισμένες στην εμφάνιση χρησιμοποιούν χαρακτηριστικά από εικόνες που τις εκπαιδεύουν ώστε να μοντελοποιήσουν την οπτική εμφάνιση, και να συγκρίνουν αυτές τις παραμέτρους με τα χαρακτηριστικά της εικόνας για αξιολόγηση. Οι μέθοδοι βασισμένες σε τρισδιάστατα μοντέλα χεριού βασίζονται σε ένα τρισδιάστατο κινηματικό μοντέλο, εκτιμώντας τις γωνιακές και γραμμικές παραμέτρους του μοντέλου.



Εικόνα 2: Διαδικασία αναγνώρισης χειρονομιών

1.4 Έρευνα και Αξιολόγηση των Τεχνικών Αναγνώρισης Χειρονομιών

Η μελέτη μας βασίζεται σε προηγούμενες προσπάθειες για τη διερεύνηση του τομέα της HGR (ή αλλιώς Human Gesture Recognition). Ο Mitra και άλλοι [7] παρείχαν μια έρευνα διαφόρων μεθόδων αναγνώρισης χειρονομιών, καλύπτοντας χειρονομίες χεριών και μπράτσων, χειρονομίες του κεφαλιού και του προσώπου, καθώς και σωματικές χειρονομίες. Οι μέθοδοι HGR που ερευνήθηκαν στην έρευνα ήταν περιορισμένοι σε Κρυφά Μοντέλα Μαρκόβ (HMMs ή Hidden Markov Models), αλγορίθμους φιλτραρίσματος σωματιδίων και συμπύκνωσης, και Τεχνητά Νευρωνικά Δίκτυα (ANNs ή Artificial Neural Netoworks). Οι μέθοδοι μοντελοποίησης χεριού και της απεικόνισης του χεριού σε 3D για εκτίμηση στάσης εξετάζονται στο [8] (αγνοώντας τα σχήματα ταξινόμησης χειρονομιών). Μια ανάλυση των νοηματικών γλωσσών, των γραμματικών διαδικασιών σε νοηματικές χειρονομίες και ζητήματα που σχετίζονται με την αυτόματη αναγνώριση νοηματικών γλωσσών συζητιούνται στο [9]. Η πιο πρόσφατη από τις παραπάνω εργασίες [8]

κάλυψε τις εξελίξεις μέχρι το έτος 2005. Η ανασκόπηση κατέληξε στο συμπέρασμα ότι οι μέθοδοι που μελετήθηκαν είναι πειραματικοί και η χρήση τους περιορίζεται σε εργαστηριακά περιβάλλοντα.

Αυτή η εργασία, ανασκοπεί σε πρόσφατες εργασίες στο HGR με επίκεντρο τις εξελίξεις των τελευταίων 16 ετών. Αλγόριθμοι που χρησιμοποιούν παραδοσιακές κάμερες RGB (Κεφάλαιο 2) καθώς και οι νεότερες κάμερες RGB-D (Κεφάλαιο 3) εξετάζονται, κάνοντας την ανασκόπηση μοναδική. Οι μέθοδοι HGR κατηγοριοποιούνται και αναλύονται βάσει της τεχνικής που χρησιμοποιούν για την ταξινόμηση των χειρονομιών. Πραγματοποιείται μια ποσοτική σύγκριση των αλγορίθμων HGR βάσει διαφορετικών χαρακτηριστικών που πηγάζουν από τους αλγορίθμους και της πειραματικής μεθοδολογίας που ακολουθήθηκε στη δοκιμή του κάθε αλγορίθμου. Παρέχεται επίσης μια ανασκόπηση των διαθέσιμων βάσεων δεδομένων των χειρονομιών (Κεφάλαιο 4) και μια συζήτηση για την έρευνα πάνω στην αναγνώριση χειρονομιών (Κεφάλαιο 5). Ελπίζουμε ότι αυτή η ανασκόπηση είναι επίκαιρη, δεδομένου των αυξανόμενων ερευνητικών προσπαθειών και της επέκτασης της αγοράς πάνω σε συστήματα αλληλεπίδρασης με χειρονομίες.

ΚΕΦΑΛΑΙΟ 2

ΑΝΑΓΝΩΡΙΣΗ ΧΕΙΡΟΝΟΜΙΩΝ: ΜΕΘΟΔΟΙ ΣΕ ΑΙΣΘΗΤΗΡΕΣ RGB

Οι τεχνικές που χρησιμοποιούνται για τη δυναμική HGR μπορούν να κατηγοριοποιηθούν ως (α) HMM [10–23] και άλλες στατιστικές μεθόδους [24–31], (β) ANN [32–34] και άλλες μεθόδους βασισμένες σε μηχανική μάθηση [35,36], (γ) μεθόδους βασισμένες σε Ιδιοχώρο [37,38], (δ) Μεθόδους Προσαρμογής Καμπύλης [39], και (ε) Δυναμικό Προγραμματισμό [40]/Δυναμική Προσαρμογή Χρόνου [41,42] (Εικ.3).

2.1 Διαδικασία Υποβολής Τελικού Αντίτυπου Διατριβής

Οι μέθοδοι HMM αποτελούν την πιο ευρέως χρησιμοποιημένη τεχνική HGR. Το HMM είναι ένα στατιστικό μοντέλο στο οποίο το σύστημα που μοντελοποιείται υποθέτετε ότι είναι μια διαδικασία Μαρκόβ με άγνωστες παραμέτρους. Το HMM αναπαριστά τη στατιστική συμπεριφορά μιας ακολουθίας παρατηρήσιμων συμβόλων χρησιμοποιώντας ένα δίκτυο κρυφών επιπέδων με πιθανότητες μετάβασης και εξόδου. Το HMM μπορεί να χρησιμοποιηθεί για αναγνώριση μοτίβων για τις χειρονομίες εφόσον οι κρυφοί παράμετροι προσδιοριστούν χρησιμοποιώντας τα παρατηρήσιμα δεδομένα.

Οι μέθοδοι αναγνώρισης δυναμικών χειρονομιών βασισμένες σε HMM χρησιμοποιούν κυρίως τα χρονικά και τα χωρικά χαρακτηριστικά των εισαγόμενων εικόνων. Ο Chen

και άλλοι [14] χρησιμοποίησαν τον δείκτη Fourier και την ανάλυση κίνησης βασισμένη σε οπτική ροή για να χαρακτηρίσουν τα χωρικά και χρονικά χαρακτηριστικά αντίστοιχα. Ο αλγόριθμος εξάγει το σχήμα του χεριού από περίπλοκα παρασκήνια παρακολουθώντας το χέρι σε πραγματικό χρόνο. Οι αλγόριθμοι βασισμένοι σε HMM αναγνωρίζουν το καλύτερο μοντέλο χειρονομίας με βάση την πιθανότητα εμφάνισης ενός μοτίβου. Οι παραλλαγές στη χειρονομία από ένα αναφορικό μοτίβο μειώνουν την πιθανότητα της χειρονομίας με το μοντέλο.

Εικόνα 3: Ταξινόμηση των τεχνικών αναγνώρισης χειρονομιών που εξετάστηκαν.

2.2 Διαμόρφωση Κειμένου

Ο Lee και ο Kim [10] παρουσίασαν μια έννοια στο HMM μοντέλο που σχετίζεται με την τοποθέτηση ενός ορίου για το φιλτράρισμα μοτίβων με λιγότερη πιθανότητα. Η κατεύθυνση της κίνησης του χεριού χρησιμοποιείται για να αναπαραστήσει τις χωροχρονικές ακολουθίες των χειρονομιών. Η μέθοδος ανιχνεύει αξιόπιστα το τελικό σημείο μιας χειρονομίας και βρίσκει το αρχικό σημείο με αναδρομή.

Το HMM βασίζεται σε ομογενείς αλυσίδες Μαρκόβ, καθώς η δυναμική του συστήματος καθορίζεται μόνο από πιθανότητες μετάβασης ανεξάρτητες από τον χρόνο. Ο Marcel και άλλοι [15] πρότειναν μια επέκταση του HMM, το λεγόμενο Input/Output Hidden

Markov Model (IOHMM), για την HGR. Το IOHMM βασίζεται σε μη ομογενείς αλυσίδες Μαρκόβ όπου οι πιθανότητες εξόδου και μετάβασης εξαρτώνται από την είσοδο. Το IOHMM μαθαίνει να αντιστοιχίζει τις ακολουθίες εισόδου, τις παρατηρήσεις, τις ακολουθίες εξόδου και τις κατηγορίες χειρονομιών για όλα τα δεδομένα χρησιμοποιώντας επoπτευόμενη διακριτική μάθηση. Σε σύγκριση με τα HMM, το IOHMM είναι μια διακριτική προσέγγιση καθώς μοντελοποιεί άμεσα τις μεταγενέστερες πιθανότητες. Η μελέτη στο [15] περιορίστηκε σε δύο κλάσεις. Ο Just και άλλοι [13] επέκτειναν τη μελέτη για την αναγνώριση χειρονομιών με ένα και δύο χέρια και παρείχαν σύγκριση του HMM και του IOHMM. Πειράματα που διεξήχθησαν σε μεγαλύτερες βάσεις δεδομένων, κυμαινόμενες από 7 έως 16 κλάσεις χειρονομιών, έχουν συμπεράνει ότι το HMM έχει καλύτερη απόδοση από το IOHMM για μεγάλο αριθμό κατηγοριών.

Στην έρευνα [11], συνδυάζονται διάφορα χαρακτηριστικά όπως, η τοποθεσία της χειρονομίας, η γωνία της και ταχύτητας του χεριού της για την υλοποίηση ενός HMM για την HGR. Η τοποθεσία του χεριού προσδιορίζεται μέσω ανάλυσης του χρώματος του δέρματος και παρακολουθείται συνδέοντας το κέντρο των κινούμενων περιοχών του χεριού. Η έρευνα συγκρίνει τη χρησιμότητα των τριών χαρακτηριστικών, της τοποθεσίας, της γωνίας και της ταχύτητας, και καταλήγει στο συμπέρασμα ότι τα γωνιακά χαρακτηριστικά είναι τα πιο αποτελεσματικά, έχοντας καλύτερη διακριτική δύναμη. Τα χαρακτηριστικά τοποθεσίας και ταχύτητας κατατάσσονται δεύτερα και τρίτα αντίστοιχα. Μια παρόμοια υλοποίηση του HMM που χρησιμοποιεί τα γωνιακά χαρακτηριστικά της κίνησης κατά μήκος της τροχιάς του κέντρου του χεριού παρέχεται στην [16].

Ο Ramamoorthy και άλλοι ανέπτυξαν ένα σύστημα HGR συνδυάζοντας ένα σχήμα με τα χρονικά χαρακτηριστικά από το HMM μαζί με ένα σύστημα αναγνώρισης στατικών σχημάτων [12]. Χρησιμοποίησαν έναν φίλτρο Kalman βασισμένο σε ανιχνευτή περιγράμματος χεριού, ο οποίος παρέχει χρονικά χαρακτηριστικά της χειρονομίας. Οι χειρονομίες αναγνωρίζονται χρησιμοποιώντας έναν ταξινομητή βασισμένο στην διακριτική ανάλυση του περιγράμματος της χειρονομίας. Αυτοί οι συμβολικοί δείκτες των χειρονομιών χρησιμοποιούνται για την εκπαίδευση του HMM. Το σύστημα μπορεί αξιόπιστα να αναγνωρίσει δυναμικές χειρονομίες παρά την κίνηση και τις διακριτές αλλαγές στις στάσεις του χεριού. Επίσης, ο αλγόριθμος έχει την ικανότητα να αναγνωρίσει τα αρχικά και τελικά σημεία των χειρονομιών μέσα σε μια ακολουθία.

Οι αλγόριθμοι αναγνώρισης δυναμικών χειρονομιών χρησιμοποιούν ένα αναδρομικό σχήμα εντοπισμού που αρχικά ανιχνεύει το τελικό σημείο μιας χειρονομίας και στη συνέχεια επιστρέφει στο αρχικό σημείο. Ο Kim και άλλοι [17] πρότειναν μια εναλλακτική

μέθοδο, ένα σχήμα εντοπισμού προς τα εμπρός, το οποίο εκτελεί τμηματοποίηση και αναγνώριση της χειρονομίας ταυτόχρονα. Τα αρχικά και τελικά σημεία των χειρονομιών ανιχνεύονται από τα μηδενικά σημεία της διαφορικής πιθανότητας του σήματος. Ένα σετ από χαρακτηριστικά βασισμένα σε 3D αρθρώσεις εξάγεται με μια τεχνική αντιστοίχισης που συσχετίζει τα 2D δεδομένα του σχήματος στα 3D δεδομένα αρθρώσεων. Οι χειρονομίες ταξινομούνται με ψηφοφορία πλειοψηφίας χρησιμοποιώντας ένα αθροιστικό HMM.

Ο Davis και ο Shah [31] αποσυνέθεσαν τις χειρονομίες σε τέσσερις διακριτές φάσεις που συμβαίνουν με μια σταθερή σειρά και ανέπτυξαν ένα μοντέλο Μηχανισμού Πεπερασμένων Καταστάσεων (FSM ή Finite State Machine) για αναγνώριση. Η χρονική υπογραφή της κίνησης του χεριού εξάγεται και η χειρονομία μοντελοποιείται χρησιμοποιώντας έναν FSM στην [29]. Η έννοια της κινητικής ενέργειας χρησιμοποιείται για να εκτιμήσει την κυρίαρχη κίνηση από μια σειρά εικόνων. Ο Hong και άλλοι [30] χρησιμοποίησαν τις 2D θέσεις των κέντρων των κεφαλιών και των χεριών των ανθρώπων για να αναπτύξουν τον FSM. Ένα δυναμικό μοντέλο που βασίζεται σε ένα Bayesian δίκτυο προτείνεται στο [24] για την αναγνώριση τόσο ακίνητων όσο και ασταμάτητων στο χρόνο χειρονομιών. Τα χαρακτηριστικά που χρησιμοποιούνται είναι κατεύθυνση για την κλάση της χειρονομίας και προσδιορίζονται από την κίνηση του χεριού, την θέση σχέσης μεταξύ των δύο χεριών, και την θέση σχέσης μεταξύ προσώπου και χεριών.

Ο Chen και άλλοι [25] πρότειναν μια διπλού επιπέδου προσέγγιση, στατιστικής και συντακτικής ανάλυσης για την αναγνώριση στατικών και δυναμικών χειρονομιών αντίστοιχα. Το πρώτο επίπεδο, η στατιστική ανάλυση, βασίζεται σε χαρακτηριστικά τύπου Haar και στον αλγόριθμο μάθησης AdaBoost. Το δεύτερο επίπεδο, η συντακτική ανάλυση, βασίζεται σε μια στοχαστική αναλυτική γραμματική χωρίς περιεχόμενο (stochastic context-free grammar ή SCFG). Τα χαρακτηριστικά του τύπου Haar περιγράφουν αποτελεσματικά το μοτίβο στάσης του χεριού και ο αλγόριθμος AdaBoost δημιουργεί έναν ισχυρό ταξινομητή συνδυάζοντας μια σειρά από αδύναμους ταξινομητές. Οι στάσεις που ανιχνεύονται από το πρώτο επίπεδο μετατρέπονται σε μια ακολουθία τερματικών συμβολοσειρών σύμφωνα με τη γραμματική, στο δεύτερο στάδιο.

2.3 Τεχνικές Βασισμένες Σε Μηχανική Μάθηση

Οι Pavlo Molchanov και άλλοι [44] στο άρθρο τους παρουσιάζουν ένα Αναδρομικό 3D Συνελικτικό Νευρωνικό Δίκτυο (R3DCNN) που σχεδιάστηκε για την ταυτόχρονη ανίχνευση και κατάταξη δυναμικών χειρονομιών από πολυμορφικά δεδομένα. Το μοντέλο

εκμεταλλεύεται την Connectionist Temporal Classification (CTC) για εκπαίδευση, επιτρέποντας την πρόβλεψη των κλάσεων από χειρονομίες που βρίσκονται σε ζωντανή ροή σε μη τμηματοποιημένες ροές εισόδου. Η αρχιτεκτονική αποτελείται από ένα βαθύ 3D-CNN για την εξαγωγή χωροχρονικών χαρακτηριστικών, ένα αναδρομικό επίπεδο για την χρονική μοντελοποίηση, και ένα επίπεδο softmax για την πρόβλεψη πιθανοτήτων υπό συνθήκη κλάσης. Το 3D-CNN προ-εκπαιδεύεται στο σύνολο δεδομένων Sport-1M και λεπτομερώς ρυθμίζεται σε ένα νέο πολυμορφικό σύνολο δεδομένων δυναμικών χειρονομιών που παρουσιάζεται στο άρθρο. Το πλήρες μοντέλο R3DCNN εκπαιδεύεται από την αρχή με ανάδραση-μέσω-χρόνου (BPTT), χρησιμοποιώντας και τις δύο συναρτήσεις κόστους, την negative log-likelihood και την CTC. Εφαρμόζονται τεχνικές, όπως η μείωση βάρους, το drop-out, και το feature map drop-out για να μειωθεί η υπερεκπαίδευση. Για να αυξηθεί η ποικιλία στα παραδείγματα εκπαίδευσης, χρησιμοποιούνται τεχνικές επαύξησης όπως τυχαία χωρική περιστροφή, κλιμάκωση, χρονική κλιμάκωση, και jittering. Το μοντέλο αυτό συνδυάζει τις πιθανότητες που προκύπτουν από διαφορετικούς τύπους εισόδων (modalities), όπως είναι οι εικόνες βάθους, χρώματος και στερεοσκοπικού υπέρυθρου, για να βελτιώσει την ακρίβεια της αναγνώρισης των χειρονομιών. Κάθε τύπος εισόδου επεξεργάζεται από ένα ξεχωριστό νευρωνικό δίκτυο που έχει σχεδιαστεί ειδικά για να διαχειρίζεται αυτή τη συγκεκριμένη είσοδο. Αυτά τα δίκτυα παράγουν κάθε φορά πιθανότητες που δείχνουν πόσο πιθανό είναι μια συγκεκριμένη χειρονομία να ανήκει σε κάθε κατηγορία που έχουν εκπαιδευτεί να αναγνωρίζουν. Το μοντέλο συνδυάζει τις πιθανότητες υπό συνθήκη κλάσης που εκτιμώνται από δίκτυα ειδικά για κάθε τύπο εισόδου για αποτελεσματική συγχώνευση πολλαπλών τύπων εισόδου. Αυτή η προσέγγιση εκμεταλλεύεται τα πλεονεκτήματα κάθε διαφορετική είσοδο και αντισταθμίζει τις αδυναμίες τους, οδηγώντας σε πιο αξιόπιστη και ακριβή αναγνώριση χειρονομιών μέσω της αποτελεσματικής συγχώνευσης των πληροφοριών από όλες τις διαθέσιμες πηγές. Σε ένα σύνολο δεδομένων δυναμικών χειρονομιών που καταγράφηκαν με αισθητήρες βάθους, χρώματος και στερεοσκοπικού IR, το μοντέλο επιτυγχάνει ακρίβεια 83.8%, πλησιάζοντας την ανθρώπινη ακρίβεια του 88.4%. Το μοντέλο επιδεικνύει ανώτερη επίδοση σε διάφορους τύπους εισόδου και συνδυασμούς αυτών, με τα καλύτερα αποτελέσματα να επιτυγχάνονται με τη συγχώνευση όλων των τύπων εισόδου. Το μοντέλο R3DCNN υπερτερεί των ανταγωνιστικών αλγορίθμων προηγμένης τεχνολογίας στα σύνολα δεδομένων όπως το SKIG και το ChaLearn2014. Η σχεδίαση του μοντέλου επιτρέπει την πρόωρη

ανίχνευση χειρονομιών, επιτυγχάνοντας μηδενική ή αρνητική καθυστέρηση, ζωτικής σημασίας για την ανταποκρισιμότητα μεταξύ των δι' επαφών και του χρήστη. Η ένταξη ενός αναδρομικού επιπέδου με ένα 3D-CNN και η χρήση της CTC για την online κατηγοριοποίηση χειρονομιών αποτελούν νέες θετικές συνεισφορές που αντιμετωπίζουν τις προκλήσεις της αναγνώρισης δυναμικών χειρονομιών σε πραγματικά συστήματα. Το προτεινόμενο μοντέλο είναι ιδιαίτερα κατάλληλο για εφαρμογές αλληλεπίδρασης ανθρώπου-υπολογιστή όπου είναι ουσιώδης η ταχεία και ακριβής αναγνώριση δυναμικών χειρονομιών.

Ο Duan και άλλοι [47] στην μελέτη τους παρουσιάζει ένα πρωτοποριακό τρόπο για την αναγνώριση απομονωμένων χειρονομιών, ενσωματώνοντας μια τεχνική σύνθεσης πολλαπλών εισόδων που αξιοποιεί τις ακολουθίες βίντεο RGB και βάθους. Τα κύρια στοιχεία αυτού του πλαισίου είναι το Δίκτυο Συναίνεσης Δύο Ροών (2SCVN ή και Two Stream Consensus Voting Network) και η Ροή Σύνθετης Ανάλυσης Βάθους 3D (3DDSN ή και 3D Depth-Saliency Conv-Net stream). Το 2SCVN σχεδιάστηκε για να αντλεί και να εκμεταλλεύεται τόσο τις βραχυπρόθεσμες όσο και τις μακροπρόθεσμες πληροφορίες από τις δομές των ακολουθιών RGB, χρησιμοποιώντας ένα μηχανισμό ψηφοφορίας συναίνεσης για την συγκέντρωση προβλέψεων από χωρικές και χρονικές ροές, μειώνοντας την αβεβαιότητα και τη διακύμανση στις τελικές προβλέψεις του μοντέλου. Η 3DDSN επικεντρώνεται στην ταυτοποίηση των χαρακτηριστικών από μικρές κινήσεις χρησιμοποιώντας τις εικόνες βάθους και το saliency, στοχεύοντας να αντιμετωπίσει την απώλεια τρισδιάστατης δομικής πληροφορίας και τις παρεμβολές από θορύβους φόντου. Η μεθοδολογία ξεκινάει από την αναπαράσταση του βίντεο ως είσοδο με διαφορετικές μορφές (RGB, βάθος, saliency και πεδία οπτικής ροής), τα οποία επεξεργάζονται μέσω των αντίστοιχων δικτύων (2SCVN για RGB και οπτική ροή· 3DDSN για βάθος και saliency). Η τελική απόφαση αναγνώρισης χειρονομιών λαμβάνεται συγχωνεύοντας τις εξόδους από 2SCVN και 3DDSN, συνδυάζοντας τα πλεονεκτήματά τους για να επιτύχουν ακριβή ταξινόμηση. Τα αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος υπερτερεί σημαντικά των υπαρχόντων μεθόδων στο σημείο αναφοράς σύνυλο δεδομένων ChaLearn IsoGD, βελτιώνοντας την πρώτη θέση κατά μεγάλο περιθώριο (10,29%) και επιτυγχάνοντας το καλύτερο αποτέλεσμα στο σετ δεδομένων RGBD-HuDaAct (96,74%). Εκτενείς πειραματισμοί και ποιοτικές αναλύσεις καταδεικνύουν την αποτελεσματικότητα του προτεινόμενου πλαισίου. Η σύνθεση πολλαπλών μορφών, ιδίως η συμπερίληψη πληροφοριών βάθους και saliency, έπαιξε καίριο ρόλο στη βελτίωση της ακρίβειας αναγνώρισης. Το μοντέλο δεν

ξεχωρίζει μόνο στην αναγνώριση χειρονομιών αλλά επίσης δείχνει υποσχόμενα αποτελέσματα για άλλες εργασίες αναγνώρισης μέσα στο βίντεο, υποδεικνύοντας την αποτελεσματικότητα και τη δυνατότητα γενίκευση.

Η παρακάτω μελέτη του Li και άλλων [52] παρουσιάζει μια μέθοδο αναγνώρισης χειρονομιών σε βίντεο χρησιμοποιώντας δεδομένα RGB-D, με την εφαρμογή ενός αισθητήρα Kinect για την ταυτόχρονη καταγραφή των εικόνων σε RGB και σε βάθους μορφή. Η κύρια στρατηγική της μεθοδολογίας τους εστιάζει στη χρήση ενός 3D Δικτύου Συνέλιξης (μοντέλο C3D) για την εκμάθηση χωροχρονικών χαρακτηριστικών από τα δεδομένα βίντεο. Η μεθοδολογία αποτελείται από αρκετά βασικά βήματα:

- Προ-επεξεργασία: Τα εισαγόμενα βίντεο μετατρέπονται σε βίντεο 32 καρέ για να εξασφαλιστεί ομοιομορφία στο σύνολο δεδομένων και για να διατηρηθούν λεπτομέρειες κίνησης που είναι κρίσιμες για τη διάκριση παρόμοιων χειρονομιών.
- Εξαγωγή Χαρακτηριστικών: Το μοντέλο C3D χρησιμοποιείται για την εξαγωγή χαρακτηριστικών τόσο από τα βίντεο RGB όσο και από τα βίντεο βάθους ξεχωριστά. Το μοντέλο C3D επιλέγεται λόγω της ικανότητάς του να εξάγει χρονικές πληροφορίες μέσω της 3D συνέλιξης και pooling, καθιστώντας την είσοδο κατάλληλη για οπτικοακουστικό υλικό.
- Συγχώνευση Χαρακτηριστικών: Η μελέτη παρουσιάζει δύο στρατηγικές για τη σύνθεση των χαρακτηριστικών που εξάγονται από δεδομένα RGB και βάθους στο πλαίσιο της αναγνώρισης χειρονομιών σε βίντεο:
- Μέσος Όρος Χαρακτηριστικών (Averaging the features): Αυτή η προσέγγιση αφορά τον υπολογισμό του μέσου όρου των χαρακτηριστικών που προέρχονται τόσο από τα RGB όσο και από τα βάθους δεδομένα, προκειμένου να συνθέσει ένα νέο σύνολο χαρακτηριστικών για την ταξινόμηση.
- Συγχώνευση για Δημιουργία Χαρακτηριστικού Διανύσματος Υψηλότερης Διάστασης (Integrating them to create a higher-dimensional feature vector): Η δεύτερη μέθοδος συνδυάζει τα χαρακτηριστικά από τα RGB και βάθους δεδομένα μέσω της ενσωμάτωσής τους σε ένα διάνυσμα υψηλότερης διάστασης, προσδοκώντας σε μια πιο πλούσια αναπαράσταση για την ταξινόμηση.

Η προσέγγιση με τον μέσο όρο των χαρακτηριστικών πέτυχε ακρίβεια 49,0% στο υποσύνολο επικύρωσης της βάσης δεδομένων ChaLearn LAP IsoGD.

Η προσέγγιση με την ενσωμάτωση των χαρακτηριστικών σε ένα διάνυσμα υψηλότερης διάστασης πέτυχε ελαφρώς καλύτερη ακρίβεια 49,2% στο ίδιο υποσύνολο. Και οι δύο προσεγγίσεις συνέβαλαν θετικά στην βελτίωση της απόδοσης σε σύγκριση με τη χρήση μόνο των αρχικών χαρακτηριστικών, με τη μέθοδο ενσωμάτωσης να εμφανίζει ελαφρώς καλύτερα αποτελέσματα.

Οι ίδιοι συγγραφείς στο επόμενο άρθρο [51] χρησιμοποιούν πάλι δεδομένα RGB-D (δηλαδή, δεδομένα RGB και βάθους που λαμβάνονται ταυτόχρονα μέσω ενός αισθητήρα Kinect). Χρησιμοποιούν το μοντέλο C3D (3D Συνελικτικό Νευρωνικό Δίκτυο) για την εξαγωγή χωροχρονικών χαρακτηριστικών από δεδομένα βίντεο. Εισάγει τη θεωρία saliency για τη βελτίωση της εξαγωγής χαρακτηριστικών εστιάζοντας σε περιοχές σχετικές με τις χειρονομίες και μειώνοντας τον θόρυβο φόντου. Χρησιμοποιούν ανάλυση διακριτικής συσχέτισης (discriminant correlation analysis) για τη σύνθεση χαρακτηριστικών, ενσωματώνοντας χαρακτηριστικά από δεδομένα RGB, βάθους και saliency για τη βελτίωση της απόδοσης αναγνώρισης. Τέλος, έχουν έναν γραμμικό ταξινομητή SVM για την τελική ταξινόμηση των χειρονομιών. Η μεθοδολογία τους γίνεται ως εξής:
Προεπεξεργασία Δεδομένων: Μετατρέπει τα εισαγόμενα δεδομένα RGB-D σε βίντεο 32 καρέ βάσει στατιστικής ανάλυσης για να διατηρήσει περισσότερες πληροφορίες χειρονομίας.

- Εξαγωγή Χαρακτηριστικών: Εξάγει χαρακτηριστικά ξεχωριστά από δεδομένα RGB, βάθους και παραγόμενα saliency βίντεο χρησιμοποιώντας το μοντέλο C3D. Δημιουργούνται saliency βίντεο για να εξαλειφθούν παράγοντες μη σχετικοί με τις χειρονομίες όπως το φόντο και τα ρούχα.
- Σύνθεση Χαρακτηριστικών: Συνδυάζει χαρακτηριστικά από διάφορες πηγές δεδομένων (RGB, βάθος, saliency) χρησιμοποιώντας ανάλυση διακριτικής συσχέτισης για να μεγιστοποιήσει τα συμπληρωματικά τους πλεονεκτήματα.
- Ταξινόμηση: Εφαρμόζει έναν ταξινομητή SVM γραμμικού τύπου στο ενοποιημένο σύνολο χαρακτηριστικών για την αναγνώριση χειρονομιών.

Το μοντέλο δείχνει αποτελεσματικότητα σε αναγνώριση χειρονομιών βασισμένη σε μεγάλης κλίμακας βίντεο, καταλαμβάνοντας την πρώτη θέση στην πρόκληση ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge. Έδειξε βελτιώσεις στην ακρίβεια μέσω της προσαρμοστικής ενοποίησης καρέ, της εισαγωγής saliency δεδομένων και του σχήματος σύνθεσης χαρακτηριστικών. Η προτεινόμενη μέθοδος ξεπέρασε τη βασική

μέθοδο και άλλες τεχνικές κορυφαίας τεχνολογίας στην ακρίβεια αναγνώρισης χειρονομιών. Η προσθήκη των saliency δεδομένων και μια πιο ολοκληρωμένη στρατηγική ενοποίησης καρέ συνέβαλαν σημαντικά στην απόδοση του μοντέλου, καταγράφοντας πιο λεπτομερείς πληροφορίες χειρονομίας και μειώνοντας θορύβους.

Στην μελέτη [54] παρουσιάζεται μια μέθοδος αναγνώρισης χειρονομιών που ενσωματώνει δεδομένα RGB, βάθους και οπτικής ροής για την αναγνώριση χειρονομιών τα οποία βρίσκονται σε βίντεο μεγάλης κλίμακας. Η μέθοδος χρησιμοποιεί το δίκτυο ResC3D, εκμεταλλευόμενη τα πλεονεκτήματα των μοντέλων Res-Net και C3D για την εξαγωγή χωροχρονικών χαρακτηριστικών.

Κύρια στοιχεία του μοντέλου περιλαμβάνουν:

- Βελτίωση Βίντεο: Χρησιμοποιεί τη θεωρία Retinex και φίλτρο διάμεσου (median) για τα βίντεο RGB και βάθους προκειμένου να μειώσει τις διακυμάνσεις φωτισμού και τον θόρυβο.
- Σχήμα Ενοποιημένων Βαρυσταθμισμένων Πλαισίων (Weighted Frame Unification Scheme): Χρησιμοποιεί έναν "μηχανισμό προσοχής σημαντικού καρέ" για να επιλέξει τα πιο αντιπροσωπευτικά καρέ βάσει της έντασης κίνησης, διασφαλίζοντας τη διατήρηση σημαντικών πληροφοριών κίνησης.
- Μοντέλο ResC3D: Υιοθετεί ένα μοντέλο ResC3D για τη μάθηση και εξαγωγή χαρακτηριστικών από multi-modal δεδομένα, συνδυάζοντας τα οφέλη των βαθιών νευρωνικών δικτύων (Deep Residual Network) με την 3D συνέλιξη.
- Σχήμα Σύνθεσης Βασισμένο σε Στατιστική Ανάλυση: Τα χαρακτηριστικά από διαφορετικές modalities συγχωνεύονται χρησιμοποιώντας την κανονική συσχέτιση ανάλυσης (canonical correlation analysis) για να μεγιστοποιηθεί η συσχέτιση μεταξύ των διαφορετικών λόγω modality καρέ και να ενσωματωθούν αποτελεσματικά τα χαρακτηριστικά.

Η μεθοδολογία περιλαμβάνει τα εξής βήματα:

- Βελτίωση Βίντεο: Για την κανονικοποίηση του φωτισμού και την απομάκρυνση θορύβου από δεδομένα RGB και βάθους.
- Βαρυσταθμισμένη Ενοποίηση Πλαισίων (Weighted Frame Unification): Για τη δειγματοληψία βίντεο βάσει έντασης κίνησης, διατηρώντας τις πληροφορίες κίνησης.

- Εξαγωγή Χαρακτηριστικών από διαφορετικού τύπου δεδομένα: Χρησιμοποιώντας το μοντέλο ResC3D για την εξαγωγή χαρακτηριστικών από δεδομένα RGB, βάθους και οπτικής ροής.
- Σύνθεση Χαρακτηριστικών: Εφαρμογή κανονικής συσχέτισης ανάλυσης για τη σύνθεση χαρακτηριστικών από διαφορετικές modalities.
- Ταξινόμηση: Χρήση ενός γραμμικού ταξινομητή SVM για την απόκτηση του τελικού αποτελέσματος αναγνώρισης βάσει των συγχωνευμένων χαρακτηριστικών.

Η μέθοδος πέτυχε σημαντικές βελτιώσεις στην ακρίβεια αναγνώρισης, αποδεικνύοντας την αποτελεσματικότητά της στη διαχείριση μεγάλων βιντεοσκοπημένων συνόλων δεδομένων για την αναγνώριση χειρονομιών. Επιτεύχθηκε ακρίβεια 67.71% στο σετ δοκιμών του συνόλου δεδομένων Chalearn LAP IsoGD, κερδίζοντας την πρώτη θέση στην πρόκληση αναγνώρισης απομονωμένων χειρονομιών. Η σύνθεση δεδομένων από διαφορετικούς τύπους και η χρήση του δικτύου ResC3D για την εξαγωγή χαρακτηριστικών τους ήταν καθοριστική στην ενίσχυση της απόδοσης αναγνώρισης. Η μέθοδος αντιμετώπισε αποτελεσματικά παράγοντες μη συσχετιζόμενους με τις χειρονομίες, όπως ο φωτισμός και ο θόρυβος, ενώ το σχήμα ενοποίησης βαρυσταθμισμένων πλαισίων και η στατιστική ανάλυση βασισμένη στη σύνθεση συνέβαλαν στην ακριβή και αξιόπιστη αναγνώριση χειρονομιών.

Το παρακάτω άρθρο [64] παρουσιάζει μια προηγμένη προσέγγιση για τη συνεχή αναγνώριση χειρονομιών, ενσωματώνοντας βαθιά νευρωνικά δίκτυα για την αντιμετώπιση τις προκλήσεις της τμηματοποίησης αναγνώρισης των ακολουθιών που περιέχουν χειρονομίες. Η προτεινόμενη λύση αποτελείται από δύο κύρια δίκτυα: ένα δίκτυο τμηματοποίησης των χειρονομιών και ένα δίκτυο αναγνώρισης χειρονομιών. Το δίκτυο αυτό είναι σχεδιασμένο για να αναγνωρίζει τα σημεία έναρξης και λήξης των χειρονομιών μέσα σε συνεχείς ακολουθίες χειρονομιών, διαχωρίζοντας αποτελεσματικά το βίντεο σε απομονωμένα κομμάτια. Η διαδικασία τμηματοποίησης χρησιμοποιεί ένα δίκτυο Res3D με χρονική διαστολή, για να αξιοποιηθεί η πληροφορία του περιβάλλοντος ως προς τον χρόνο χωρίς απώλεια ανάλυσης. Αυτή η μέθοδος επιτρέπει πιο ακριβή ανίχνευση των εικόνων που περιλαμβάνουν μια χειρονομία, χρησιμοποιώντας μια squared hinge συνάρτηση κόστους. Αυτή η συνάρτηση εφαρμόζει διαφορετικά βάρη στις εικόνες που βρίσκονται στο πλαίσιο της προβολής της χειρονομίας σε σχέση με αυτές στις οποίες ακόμα δεν υπάρχει κάποια χειρονομία. Μετά τη τμηματοποίηση, το δίκτυο αναγνώρισης ταξινομεί

τις απομονωμένες ακολουθίες των χειρονομιών. Αυτό το δίκτυο χρησιμοποιεί μια καινοτόμα αρχιτεκτονική με την ονομασία "3DCNN-ConvLSTM-2DCNN", συνδυάζοντας τα πλεονεκτήματα των 3D Συνελικτικών Νευρωνικών Δικτύων (3DCNN) για την εκμάθηση βραχυπρόθεσμων χωροχρονικών χαρακτηριστικών, του Συνελικτικού Μακροπρόθεσμου Βραχυπρόθεσμου Μνήμης (ConvLSTM) για τη σύλληψη μακροπρόθεσμων χωροχρονικών χαρτών χαρακτηριστικών, και του 2D CNN (ειδικά του MobileNet) για την τελική ταξινόμηση βάσει των εκμαθημένων feature maps (ή activation maps= το επίπεδο που περιέχει όλη την πληροφορία λίγο πριν αυτή περάσει στο τελευταίο επίπεδο ταξινόμησης). Αυτή η προσέγγιση επιτρέπει την αποτελεσματική εκμάθηση των χαρακτηριστικών της χειρονομίας σε διάφορα σημεία της ακολουθίας και ανεξάρτητα με την ποικιλομορφία της εικόνας. Η μεθοδολογία δοκιμάστηκε σε διάφορα σύνολα δεδομένων, συμπεριλαμβανομένων του ChaLearn LAP Continuous Gesture Dataset (ConGD), του Montalbano Gesture Recognition Dataset και του Jester Dataset-V1, επιδεικνύοντας κορυφαία απόδοση στη συνεχή αναγνώριση χειρονομιών. Το δίκτυο τμηματοποίησης χώρισε αποτελεσματικά τις συνεχείς ακολουθίες χειρονομιών σε απομονωμένες περιπτώσεις, και το δίκτυο αναγνώρισης πέτυχε υψηλή ακρίβεια στην ταξινόμηση αυτών των χειρονομιών. Τα πειράματα απέδειξαν την ανωτερότητα της προτεινόμενης μεθόδου έναντι των τωρινών τεχνικών, επισημαίνοντας ιδιαίτερα τα οφέλη του δικτύου Res3D με χρονική διάταση για τη διαίρεση και της αρχιτεκτονικής "3DCNN-ConvLSTM-2DCNN" για την αναγνώριση. Η συγχώνευση αυτών των δικτύων αντιμετωπίζει τις προκλήσεις της συνεχούς αναγνώρισης χειρονομιών, συμπεριλαμβανομένων της μεταβλητότητας των μηκών των ακολουθιών και της ανάγκης για ακριβή χρονική διαίρεση.

Στην εργασία [52] παρουσιάζεται μια καινοτόμα προσέγγιση για τη συνεχή αναγνώριση χειρονομιών, εκμεταλλευόμενη τα βαθιά νευρωνικά δίκτυα για να αντιμετωπίσει τις προκλήσεις που συνδέονται με την αναγνώριση χειρονομιών από ζωντανές ροές. Το μοντέλο αποτελείται από δύο κύρια στοιχεία: ένα δίκτυο τμηματοποίησης χειρονομιών και ένα δίκτυο αναγνώρισης χειρονομιών.

Δίκτυο Τμηματοποίησης Χειρονομιών: Αυτό το δίκτυο βασίζεται σε ένα τροποποιημένο δίκτυο Res3D, ενσωματώνοντας temporal dilation για να αξιοποιήσει καλύτερα τις χρονοχωρικές πληροφορίες για την τμηματοποίηση των χειρονομιών. Το δίκτυο ταξινομεί κάθε καρέ σε δύο κατηγορίες: boundaries και non-boundaries, χρησιμοποιώντας για

συνάρτηση απώλειας της hinge loss, ώστε να αντιμετωπίσει το πρόβλημα της εξισορρόπησης των βαρών ανάμεσα από τις δύο αυτές κατηγορίες.

Δίκτυο Αναγνώρισης Χειρονομιών: Μετά την τμηματοποίηση, αυτό το δίκτυο χρησιμοποιεί μια μοναδική αρχιτεκτονική "3DCNN-ConvLSTM-2DCNN" για την αναγνώριση απομονωμένων ακολουθιών με χειρονομίες. Αυτή η αρχιτεκτονική συνδυάζει τα πλεονεκτήματα των 3D Συνελικτικών Νευρωνικών Δικτύων (3DCNN) για την εκμάθηση βραχυπρόθεσμων χρονοχωρικών χαρακτηριστικών, του Συνελικτικού Μακροχρόνιου Βραχυπρόθεσμου Μνήμης (ConvLSTM) για την αποτύπωση μακροπρόθεσμων χρονοχωρικών μοτίβων και του 2D CNN (ειδικά του MobileNet) για την ταξινόμηση των χειρονομιών βάσει των εκμαθημένων χρονοχωρικών χαρακτηριστικών τους. Αυτός ο συνδυασμός έχει σχεδιαστεί ειδικά για να χειριστεί χειρονομίες διαφόρων μηκών χωρίς να απαιτεί μια σταθερού μήκους ακολουθία ως είσοδο. Τα Πειραματικά αποτελέσματα σε σύνολα δεδομένων όπως το ChaLearn LAP Continuous Gesture Dataset (ConGD), το Montalbano Gesture Recognition Dataset και το Jester Dataset-V1 επιβεβαιώνουν την αποτελεσματικότητα των προτεινόμενων δικτύων. Η προσέγγιση επιτυγχάνει κορυφαίες επιδόσεις στη συνεχή αναγνώριση χειρονομιών, παρουσιάζοντας τη δυνατότητα του μοντέλου να αναγνωρίζει με ακρίβεια χειρονομίες από ακολουθίες βίντεο. Η έρευνα υποδηλώνει μια σημαντική πρόοδο στην αλληλεπίδραση ανθρώπου-υπολογιστή, ενδεχομένως ενισχύοντας εφαρμογές στην εικονική πραγματικότητα, την κατανόηση της νοηματικής γλώσσας και άλλους τομείς που απαιτούν περίπλοκη αναγνώριση χειρονομιών.

Στην επόμενη αναφορά [44] ο Cao και άλλοι παρουσιάζουν ένα καινοτόμο μοντέλο μάθησης της αναγνώρισης χειρονομιών εγωκεντρικής οπτικής, ειδικά σχεδιασμένο για να αντιμετωπίζει τις προκλήσεις των βίντεο που περιέχουν υλικό πρώτου-προσώπου (first-person view). Το μοντέλο συνδυάζει τα 3D Συνελικτικά Νευρωνικά Δίκτυα (CNNs) με τα Αναδρομικά Νευρωνικά Δίκτυα (RNNs), χρησιμοποιώντας συγκεκριμένα τα δίκτυα Long Short-Term Memory (LSTM), για την αποτελεσματική επεξεργασία ακολουθιών βίντεο για την αναγνώριση χειρονομιών. Ένα ξεχωριστό χαρακτηριστικό του μοντέλου είναι η εισαγωγή μιας Μονάδας Χωροχρονικής Μετασχηματισμού (STTM ή Spatiotemporal Transformer Module), εμπνευσμένη έννοια από τον χώρο του χωρικού μετασχηματιστή, η οποία μετασχηματίζει τους 3D χάρτες χαρακτηριστικών (feature maps) σε ένα canonical view και στις δύο χωρό-χρονικές διαστάσεις. Αυτή η μονάδα περιλαμβάνει ένα τοπικό δίκτυο πρόβλεψης των παραμέτρων του μετασχηματισμού, δηλαδή αναλύει το

βίντεο εισόδου και αποφασίζει αν πρέπει να υπάρξει κάποιος μετασχηματισμός σε διάφορες περιοχές μιας εικόνας, ώστε να γίνει πιο σταθερή η οπτική γωνία, έναν grid generator ο οποίος δείχνει στην μονάδα σε ποια περιοχή της αρχικής εικόνας να εστιάσει ώστε να αντλήσει σημαντικές πληροφορίες και τέλος σε έναν sampler οποίος εφαρμόζει το sampling grid στην περιοχή της χειρονομίας κάνοντας interpolation με τα χαρακτηριστικά της χειρονομίας στην περιοχή όπου συμβαίνει. Ως αποτέλεσμα, παρόλο που ενδέχεται να κουνηθεί η κάμερα που βρίσκεται στο κεφάλι του ανθρώπου, τα σημαντικά χαρακτηριστικά της εικόνας παραμένουν σταθερά μετά τον μετασχηματισμό και επομένως κάνουν την δουλειά του μοντέλου πιο εύκολη για αναγνώριση της χειρονομίας. Η μεθοδολογία επικεντρώνεται στην αντιμετώπιση των προκλήσεων της εγωκεντρικής κίνησης και της μερικής ορατότητας που είναι ουσιώδεις στα βίντεο πρώτου προσώπου. Η προτεινόμενη STTM, με την ικανότητά της για μετασχηματισμό ομογραφίας 3D, διορθώνει τις παραμορφώσεις που προκαλούνται από τις κινήσεις του κεφαλιού χωρίς να ανιχνεύει ρητά το χέρι και να εκτιμά την κίνηση του κεφαλιού. Αυτό επιτυγχάνεται προβλέποντας τις παραμέτρους μετασχηματισμού με βάση τα προηγούμενα καρέ, ενσωματώνοντας πληροφορίες μακροπρόθεσμης διάρκειας μέσω αναδρομικών συνδέσεων. Η προσέγγιση διευκολύνει την επεξεργασία βίντεο με σημαντική κίνηση του κεφαλιού και των χεριών που είναι μερικώς ή εντελώς εκτός του πεδίου θέασης. Ολόκληρο το πλαίσιο, συμπεριλαμβανομένου του αναδρομικού STTM (RSTTM), έχει σχεδιαστεί για να ενσωματωθεί σε ένα 3D CNN σε οποιοδήποτε σημείο μεταξύ των συνελικτικών επιπέδων, διασφαλίζοντας ένα σύστημα που μαθαίνει από την αρχή μέχρι το τέλος. Για να επικυρώσουν τη μέθοδό τους, οι συγγραφείς παρουσιάζουν το σύνολο δεδομένων EgoGesture, το οποίο ισχυρίζονται ότι είναι το μεγαλύτερο σύνολο δεδομένων χειρονομιών εγωκεντρικής οπτικής μέχρι σήμερα, περιλαμβάνοντας πάνω από 24,000 δείγματα χειρονομιών και 3 εκατομμύρια καρέ σε διάφορες σκηνές κλειστού ή ανοιχτού χώρου. Αυτό το σύνολο δεδομένων, με το μέγεθος, την ποικιλία και τα ρεαλιστικά δεδομένα του, στοχεύει να παρέχει ένα αξιόπιστο πεδίο δοκιμών για την ταξινόμηση και την ανίχνευση χειρονομιών με συνεχείς ακολουθίες χειρονομιών, προωθώντας την έρευνα στην αναγνώριση εγωκεντρικών χειρονομιών. Το μοντέλο πετυχαίνει εξαιρετική απόδοση σε σύγκριση με τις κορυφαίες μεθόδους, επιτυγχάνοντας σημαντικές βελτιώσεις στην ακρίβεια, ιδιαίτερα σε σενάρια που περιλαμβάνουν εγωκεντρική κίνηση (π.χ. περπάτημα). Η ενσωμάτωση της STTM όχι μόνο ενισχύει την ικανότητα του μοντέλου να αντιμετωπίζει χωρικές και χρονικές

παραλλαγές αλλά επίσης βελτιώνει σημαντικά την αναγνωριστική ακρίβεια των μπερδεμένων κλάσεων των χειρονομιών. Η ανάλυση αποκαλύπτει περαιτέρω ότι γίνεται καλύτερη εκμάθηση των χαρακτηριστικών του βαθύ νευρωνικού δικτύου, σε αντίθεση με τα hand-crafted features (δηλαδή τα χαρακτηριστικά που εξάγονται από κάποιον άνθρωπο ή αλλιώς και εμπειρικά και θεωρούνται σημαντικά) προσφέροντας μεγαλύτερη ανθεκτικότητα σε αλλαγές στον φωτισμό και των γενικών κινήσεων. Επιπλέον, η συγχώνευση των εικόνων RGB και βάθους οδηγεί σε περαιτέρω βελτιώσεις της απόδοσης, υπογραμμίζοντας την αποτελεσματικότητα του προτεινόμενου μοντέλου στη χρήση διαφορετικών τύπων δεδομένων ως είσοδο για την αναγνώριση χειρονομιών.

Στην έρευνα [50], εξετάζεται η αναγνώριση δυναμικών χειρονομιών με τη χρήση δικτύων Συνέλιξης Νευρωνικών Δικτύων (CNN) 2D και 3D. Τα CNN 2D αποδείχθηκαν πιο αποδοτικά από τα 3D στη μάθηση χρονικών αναπαραστάσεων, ειδικά για βίντεο με μικρές κινήσεις, όπως οι δυναμικές χειρονομίες. Τα βίντεο χειρονομιών μετατράπηκαν σε ένα 2D σχέδιο πλακιδίων με χρονικά ταξινομημένα και μη επικαλυπτόμενα καρέ. Προτάθηκαν δύο τύποι συγχώνευσης: μία πρώιμη για τον συνδυασμό βάθους και χρωματικών λειτουργικών μονάδων και μία αργή για τη συγχώνευση προβλέψεων των CNN 2D και 3D. Λόγω του περιορισμένου αριθμού δειγμάτων στα σύνολα δεδομένων χειρονομιών, εφαρμόστηκαν μέθοδοι επαύξησης δεδομένων σε δύο ομάδες, με τη μία να αλλάζει τα εικονοστοιχεία και την άλλη απλώς να τα μετακινεί. Τα αποτελέσματα επιβεβαίωσαν ότι η προτεινόμενη μέθοδος χρήσης των CNN 2D ξεπερνά τα CNN 3D, όπως τα C3D. Η τελική δομή που αποτελείται από δύο στάδια και δύο ροές βελτίωσε την ακρίβεια αναγνώρισης στα σύνολα δεδομένων Cambridge HGD και VIVA HGD. Μια φυσική κατεύθυνση για μελλοντική έρευνα είναι η εξέταση νέων δικτύων CNN 2D που έχουν προταθεί πρόσφατα, όπως τα Dense-Net και inception-v4.

Το άρθρο [48] παρουσιάζει ένα πολυδιάστατο πλαίσιο βασισμένο στην βαθιά μάθηση του νευρωνικού δικτύου με όνομα MultiD-CNN για την αναγνώριση ανθρώπινων χειρονομιών σε βίντεο. Επικεντρώνεται στην ανάλυση και στην ερμηνεία των ορατών ανθρώπινων συμπεριφορών μέσω της υπολογιστικής όρασης. Αυτό το πλαίσιο συνδυάζει τα πλεονεκτήματα των δεδομένων RGB και βάθους (RGB-D) για τη βελτίωση των συστημάτων αλληλεπίδρασης ανθρώπινου-υπολογιστικού (HCI), ειδικά στις εργασίες αναγνώρισης χειρονομιών. Το MultiD-CNN αποτελείται από δύο βασικά υπο-δίκτυα: το Δίκτυο Χρωμάτων-Βάθους 3D (3D-CDCN) και το Δίκτυο Αναπαράστασης Κίνησης 2D (2D-

MRCN). Το 3D-CDCN σχεδιάστηκε για να μαθαίνει ταυτόχρονα χωροχρονικά χαρακτηριστικά από και τισδύο ροές RGB και βάθους χρησιμοποιώντας 3D Συνελικτικά Residual Δίκτυα (3D ResNets) και ένα δίκτυο Συνελικτικής Μακροπρόθεσμης Μνήμης (ConvLSTM). Αυτό το υπο-δίκτυο στοχεύει στη λεπτομερειακή σύλληψη των χωροχρονικών παραλλαγών των χειρονομιών σε μια τρισδιάστατη δομή, αντιμετωπίζοντας προκλήσεις όπως η παραλλαγή της οπτικής γωνίας και τα έντονα παρασκήνια. Το 2D-MRCN επικεντρώνεται στην αναγνώριση χειρονομιών από τις κινήσεις τους. Επεξεργάζεται βελτιωμένες Εικόνες Ιστορικού Κίνησης (iMHI) και βελτιωμένους Χάρτες Κίνησης Βάθους (iDMM), οι οποίες είναι στατικές εικόνες που συσσωρεύουν πληροφορίες κίνησης σε διαδοχικές ακολουθίες βίντεο. Αυτή η προσέγγιση στοχεύει να μειώσει την επιρροή των παραγόντων που δεν σχετίζονται με τη χειρονομία μαθαίνοντας αποτελεσματικά χωροχρονικά χαρακτηριστικά σε μια δισδιάστατη δομή. Η μεθοδολογία περιλαμβάνει επίσης ολοκληρωμένες στρατηγικές συγχώνευσης σε διάφορα επίπεδα (επίπεδα χαρακτηριστικών και ταξινόμησης) για να συνδυάσει τις εξόδους διαφόρων συστατικών εντός του MultiD-CNN, αξιοποιώντας τα συμπληρωματικά πλεονεκτήματα κάθε υπο-δικτύου και των διαφορετικών τύπων δεδομένων. Ο προτεινόμενος τρόπος αναγνώρισης χειρονομιών και γενικότερα ανθρώπινων κινήσεων MultiD-CNN αξιολογήθηκε σε τέσσερα απαιτητικά σύνολα δεδομένων: Chalearn LAP IsoGD, Shefeld Kinect Gesture (SKIG), NATOPS χειρονομία, και SBU Kinect. Το μοντέλο πέτυχε κορυφαίες επιδόσεις σε αυτά τα σύνολα δεδομένων, επιδεικνύοντας την αποτελεσματικότητά του και τη γενική εφαρμογή του σε προβλήματα ταξινόμησης βίντεο πέρα από την αναγνώριση χειρονομιών. Για το IsoGD, το MultiD-CNN υπερέβη τις υπάρχουσες μεθόδους, επιδεικνύοντας την αντοχή του έναντι περίπλοκων φόντων και την ικανότητά του να διαχωρίζει διάφορους τύπους χειρονομιών. Στο σύνολο δεδομένων SKIG, η αρχιτεκτονική πέτυχε σχεδόν τέλεια ακρίβεια ταξινόμησης, υπογραμμίζοντας την ικανότητά του να χειρίζεται παραλλαγές των χειρονομιών του χεριού υπό διαφορετικές συνθήκες. Τα σύνολα δεδομένων NATOPS και SBU επιβεβαίωσαν περαιτέρω την ανωτερότητα της αρχιτεκτονικής αυτής στην αναγνώριση χειρονομιών του άνω μέρους του σώματος και των αλληλεπιδράσεων δύο ατόμων, αντίστοιχα, με σημαντικά υψηλότερη ακρίβεια σε σύγκριση με τις ανταγωνιστικές προσεγγίσεις. Συνοπτικά, το MultiD-CNN ενσωματώνει αποτελεσματικά χωρικές και χρονικές πληροφορίες από βίντεο RGB-D μέσω μιας νέας συνδυασμένης χρήσης μοντέλων βαθιάς

μάθησης, αντιμετωπίζοντας τους περιορισμούς των υπαρχόντων συστημάτων αναγνώρισης χειρονομιών και θέτοντας νέα πρότυπα για εφαρμογές HCI.

Το άρθρο του Lu και άλλων [53] παρουσιάζει μια τροποποιημένη αρχιτεκτονική δικτύου Convolutional 3D (C3D) προσαρμοσμένη για την Αναγνώριση Χειρονομιών μέσω Μάθησης Ενός Παραδείγματος (OSLHGR) από ακολουθίες με εικόνες βάθους. Αυτή η αρχιτεκτονική αποτελεί ενίσχυση του δικτύου C3D, το οποίο είναι γνωστό για την αποτελεσματικότητά του στη μάθηση χωροχρονικών χαρακτηριστικών για εργασίες βασισμένες σε βίντεο. Οι τροποποιήσεις περιλαμβάνουν την προσαρμογή του δικτύου για την αποτελεσματική διαχείριση σεναρίων μάθησης ενός παραδείγματος και την ενσωμάτωση μηχανισμών όπως η Ευκλείδεια απόσταση για τη διάκριση μεταξύ θετικών και αρνητικών παραδειγμάτων. Το δίκτυο επίσης χρησιμοποιεί μεταφορά μάθησης, όπου προ-εκπαιδευμένοι παράμετροι μοντέλου σε μεγάλα σύνολα δεδομένων χρησιμοποιούνται και λεπτομερώς ρυθμίζονται με περιορισμένο αριθμό παραδειγμάτων ανά κλάση για την προσαρμογή σε νέες χειρονομίες. Στην μεθοδολογία γίνονται τα παρακάτω βήματα:

- Προ-επεξεργασία Δεδομένων: Η προσέγγιση ξεκινά με την προ-επεξεργασία των δεδομένων για να εξασφαλιστεί η ομοιομορφία στον αριθμό των καρέ, η αλλαγή μεγέθους των καρέ σε σταθερό μέγεθος, και ενίσχυση των δεδομένων για αυξημένη γενίκευση στην εκπαίδευση.
- Μεταφορά Βαρών και Λεπτομερής Ρύθμιση: Το μοντέλο εκμεταλλεύεται τη μεταφορά των βαρών από ένα βασικό δίκτυο που έχει εκπαιδευτεί σε ένα μεγάλο σύνολο δεδομένων, μεταφέροντας τα μαθημένα χαρακτηριστικά σε ένα στοχευμένο δίκτυο. Αυτό το στοχευμένο δίκτυο στη συνέχεια λεπτομερώς ρυθμίζεται σε ένα νέο, μικρότερο σύνολο δεδομένων που προορίζεται για την αναγνώριση χειρονομιών, με έμφαση στην προσαρμογή σε σενάρια μάθησης ενός παραδείγματος.
- Εξαγωγή Χαρακτηριστικών και Ταξινόμηση: Το δίκτυο σχεδιάζεται για την εξαγωγή discriminative χωρο-χρονικών χαρακτηριστικών από ακολουθίες βίντεο και χρησιμοποιεί για ταξινομητή μια συνάρτηση Softmax για την αναγνώριση χειρονομιών. Για τη μάθηση ενός παραδείγματος, η μέθοδος περιλαμβάνει λεπτομερή ρύθμιση με πολύ λίγα παραδείγματα και χρησιμοποιεί ένα μηχανισμό διάκρισης για να χειριστεί αποτελεσματικά νέες κλάσεις χειρονομιών.

Το τροποποιημένο δίκτυο C3D πέτυχε ανώτερη απόδοση στις εργασίες αναγνώρισης χειρονομιών, επιτυγχάνοντας state-of-the-art αποτελέσματα στο διαθέσιμο σύνολο

δεδομένων VIVA και SKIG, ενώ η αξιολόγηση γίνεται σε ένα νέο σύνολο δεδομένων που εισήγαγαν οι συγγραφείς, που αναφέρεται ως σύνολο δεδομένων BSG. Η προσέγγιση ήταν αποτελεσματική σε σενάρια μάθησης ενός παραδείγματος, όπου το δίκτυο μπορούσε να μάθει από έναν ελάχιστο αριθμό παραδειγμάτων ανά κλάση, επιδεικνύοντας την προσαρμοστικότητά του σε νέες χειρονομίες με περιορισμένα δεδομένα. Το multi-modal fusion framework, το οποίο συνδυάζει το τροποποιημένο C3D με έναν γραμμικό ταξινομητή SVM, ενίσχυσε περαιτέρω την ακρίβεια αναγνώρισης, δείχνοντας το όφελος της ένταξης διαφορετικών μονταλιτών και ταξινομητών για βελτιωμένη απόδοση. Το άρθρο επισημαίνει επίσης την αποτελεσματικότητα της συνεχώς fine-tuning τεχνικής, δηλαδή της τεχνικής βελτιστοποίησης των βαρών του δικτύου και της μεταφοράς βαρών από ένα γενικό μοντέλο σε ένα πιο ειδικό μοντέλο για τη σημαντική μείωση του κινδύνου υπερεκπαίδευσης, ιδίως όταν αντιμετωπίζουμε μικρό αριθμό δειγμάτων εκπαίδευσης.

Το άρθρο [46] παρουσιάζει ένα νέο βαθύ νευρωνικό residual1 δίκτυο προσοχής 3D, με την ονομασία Res3ATN, για την αναγνώριση χειρονομιών από βίντεο. Το μοντέλο εμπνέεται από την επιτυχία των δικτύων προσοχής και των βαθιών residual δικτύων, υλοποιώντας τα για να επικεντρωθεί σε σημαντικά κομμάτια των καρέ για την αναγνώριση χειρονομιών. Το Res3ATN χρησιμοποιεί 3D συνελικτικά νευρωνικά δίκτυα (CNNs) για να εντοπίσει και τις χωρικές και τις χρονικές σχέσεις στα δεδομένα βίντεο. Διαθέτει στοιβαγμένα μπλοκ προσοχής που αλλάζουν προσαρμοστικά ανάλογα με το βάθος του δικτύου, επιτρέποντας στο μοντέλο να επικεντρώνεται σε σημαντικά χαρακτηριστικά για την ταξινόμηση χειρονομιών. Το δίκτυο ενσωματώνει residual blocks για ευκολότερη βελτιστοποίηση των βαθύτερων δικτύων, ενισχύοντας την ικανότητα μάθησης και την ακρίβεια. Η μεθοδολογία περιλαμβάνει τη χρήση 3D CNNs μέσα σε ένα πλαίσιο residual προσοχής για την επεξεργασία καρέ, ώστε να γίνει η αναγνώριση των χειρονομιών. Η αρχιτεκτονική του δικτύου περιλαμβάνει residual blocks για την εκμάθηση βαθιών χαρακτηριστικών και μπλοκ προσοχής για την έμφαση σε σημαντικά χαρακτηριστικά. Το μοντέλο εκπαιδεύτηκε από την αρχή σε τρεις διαφορετικές βάσεις δεδομένων: EgoGesture, Jester και NVIDIA Dynamic Hand Gesture dataset. Χρησιμοποιήθηκαν τεχνικές επαύξησης δεδομένων για να ενισχυθεί η ποικιλία των δεδομένων εκπαίδευσης και να βελτιωθεί η γενίκευση του μοντέλου. Το μοντέλο αξιολογήθηκε χρησιμοποιώντας

[1] Η residual ή υπολειπόμενη μάθηση περιλαμβάνει την προσθήκη συντομεύσεων που παρακάμπτουν έναν ή περισσότερους φίλτρα σε ένα νευρωνικό δίκτυο.

μετρικές όπως η ακρίβεια top-1 και top-5 σε διαφορετικές βάσεις δεδομένων. Τα αποτελέσματα και τα ευρήματα δείχνουν ότι το Res3ATN υπερέχει άλλων συγκρινόμενων δικτύων (C3D, ResNet-10, ResNext-101) σε εργασίες αναγνώρισης χειρονομιών σε όλες τις τρεις βάσεις δεδομένων. Η χρήση πολλαπλών μπλοκ προσοχής βρέθηκε να είναι ωφέλιμη, με το μοντέλο να επιτυγχάνει καλύτερη ακρίβεια με αυξημένο αριθμό μπλοκ προσοχής. Η θέση των μπλοκ προσοχής μέσα στο δίκτυο επηρέασε επίσης την απόδοση, με ορισμένες εκδοχές να παράγουν καλύτερα αποτελέσματα. Το έγγραφο προτείνει ότι περαιτέρω βελτιώσεις θα μπορούσαν να γίνουν με την αύξηση του μεγέθους των εικόνων ως είσοδο, χρησιμοποιώντας μεγαλύτερα μεγέθη δέσμης(batch) κατά την εκπαίδευση και εφαρμόζοντας προ-εκπαιδευμένα δίκτυα σε συναφείς εργασίες για μεταφορά μάθησης. Η έρευνα συμβάλλει στην προώθηση της τεχνολογίας αναγνώρισης χειρονομιών, με πιθανές εφαρμογές σε διάφορους τομείς όπως τα αυτοματοποιημένα συστήματα, η ερμηνεία της νοηματικής γλώσσας και η αλληλεπίδραση ανθρώπου-υπολογιστή.

Το άρθρο των Zhang και άλλων [61] παρουσιάζει μια καινοτόμο μέθοδο αναγνώρισης δυναμικών χειρονομιών χρησιμοποιώντας δύο βασικά μοντέλα: ένα μοντέλο 3D Dense-Net και ένα μοντέλο 2D αναπαράστασης κίνησης CNN βασισμένο στο VGGNet. Η κύρια καινοτομία βρίσκεται στον συνδυασμό αυτών των μοντέλων για την ενίσχυση της απόδοσης αναγνώρισης χειρονομιών μέσω της πιο αποτελεσματικής καταγραφής των χωρικών και χρονικών χαρακτηριστικών. Το μοντέλο 3D Dense-Net επεκτείνει την αρχιτεκτονική Dense-Net σε 3D, επιτρέποντας τη μάθηση χωροχρονικών χαρακτηριστικών απευθείας από ακολουθίες με εικόνες RGB. Αυτό το μοντέλο χρησιμοποιεί πυκνές συνδέσεις (dense connections) για την αντιμετώπιση του προβλήματος εξαφάνισης του gradient και τη βελτίωση της διάδοσης των χαρακτηριστικών ανά επίπεδο, καθιστώντας το ικανό να αποτυπώνει τις δυναμικές πληροφορίες των χειρονομιών. Το μοντέλο 2D αναπαράστασης κίνησης CNN χρησιμοποιεί τις τεχνικές Motion History Image (MHI) και ψευδοχρώματος(pseudo-coloring) για να δημιουργήσει μια γενικευμένη αναπαράσταση κίνησης, η οποία στη συνέχεια επεξεργάζεται από ένα τροποποιημένο 2D CNN βασισμένο στο VGG-Net για την εξαγωγή χαρακτηριστικών χειρονομιών. Αυτή η προσέγγιση στοχεύει στην ενίσχυση της ικανότητας του μοντέλου να κάνει την διάκριση μεταξύ διαφορετικών χειρονομιών, εστιάζοντας στις πληροφορίες κίνησης. Η μεθοδολογία περιλαμβάνει την αναπαράσταση κίνησης μέσω των τεχνικών MHI και ψευδοχρωματισμού για την κωδικοποίηση της διαδοχικής κίνησης μιας ακολουθίας σε μία εικόνα, βελτιώνοντας

την ικανότητα του μοντέλου να καταγράφει τα ουσιώδη χαρακτηριστικά της κίνησης ενώ μειώνει ασήμαντους παράγοντες όπως ο θόρυβος φόντου. Εξερευνά διάφορες στρατηγικές συγχώνευσης στα επίπεδα των χαρακτηριστικών και των αποφάσεων, ώστε να συγχωνεύσει τις εξόδους των μοντέλων 3D Dense-Net και 2D CNN, στοχεύοντας στην αξιοποίηση των συμπληρωματικών δυνατοτήτων κάθε μοντέλου για τη βελτίωση της συνολικής απόδοσης αναγνώρισης. Η μεθοδολογία επιβεβαιώθηκε σε δύο δημόσια σύνολα δεδομένων χειρονομιών, VIVA και UTD-MHAD, αποδεικνύοντας την αποτελεσματικότητά της σε διάφορους τύπους χειρονομιών και σε διάφορες συνθήκες. Η προτεινόμενη μέθοδος επιτυγχάνει ακρίβεια 89.1% στο σύνολο δεδομένων VIVA και 89.5% στο UTD-MHAD, υπερβαίνοντας πολλές άλλες μεθόδους. Αυτή η επιτυχία αποδίδεται στην καινοτόμο χρήση του 3D Dense-Net για την καταγραφή των χωροχρονικών χαρακτηριστικών και την αποτελεσματική τεχνική αναπαράστασης κίνησης. Το άρθρο καταλήγει ότι η προτεινόμενη μέθοδος, μέσω του καινοτόμου σχεδιασμού μοντέλου και της στρατηγικής σύνθεσης, παρέχει μια γενικευμένη λύση για την αναγνώριση δυναμικών χειρονομιών. Η ικανότητα της μεθόδου να ενσωματώνει αποτελεσματικά χωροχρονικές πληροφορίες και χαρακτηριστικά κίνησης την ξεχωρίζει από τις υπάρχουσες προσεγγίσεις. Ωστόσο, αναγνωρίστηκαν περιορισμοί όπως η ευαισθησία της κίνησης σε σχέση με την κάμερα, με μελλοντικές εργασίες να στοχεύουν στην ανάπτυξη πιο ευέλικτων και ανεξάρτητων από την οπτική γωνία μεθόδων αναγνώρισης χειρονομιών.

Η εργασία [62] παρουσιάζει ένα καινοτόμο μοντέλο αναγνώρισης χειρονομιών που συνδυάζει το 3D-DenseNet και τα Δίκτυα Χρονικής Συνέλιξης (TCNs) με έναν ενισχυμένο μηχανισμό προσοχής μέσω των μετασχηματισμένων Δικτύων Συμπίεσης-Ενθάρρυνσης (SENets ή και Squeeze-and-Excitation Networks). Το μοντέλο δομείται για να αντιμετωπίσει δύο κύριες προκλήσεις στην αναγνώριση χειρονομιών: την εξαγωγή βραχυπρόθεσμων χωροχρονικών χαρακτηριστικών και την ταξινόμηση μακροπρόθεσμων πληροφοριών ακολουθίας αποτελεσματικά. Το 3D-DenseNet χρησιμοποιείται για την εξαγωγή βραχυπρόθεσμων χωροχρονικών χαρακτηριστικών από βίντεο κλιπ, επικεντρώνοντας στα χωρικά και χρονικά χαρακτηριστικά σε τοπικό επίπεδο (local temporal pooling) αντιμετωπίζοντας την αντίληψη ότι μια εικόνα από μόνη της δεν μεταφέρει αρκετή πληροφορία. Τα Δίκτυα Χρονικής Συνέλιξης (TCNs) αντικαθιστούν τα RNNs για την ανάλυση πληροφοριών της ακολουθίας, προσφέροντας σαφήνεια και απλότητα στη διαχείρισή τους. Τα Δίκτυα Συμπίεσης-Ενθάρρυνσης (SE-Nets) βελτιώνονται και

εφαρμόζονται εντός των TCNs για να ενισχύσουν την ικανότητα του δικτύου στην εξαγωγή χρονικών χαρακτηριστικών μοντελοποιώντας τις αλληλεξαρτήσεις μεταξύ των feature maps των συνελικτικών χαρακτηριστικών. Η μεθοδολογία περιλαμβάνει βήματα προ-επεξεργασίας όπως την αύξηση δεδομένων (αντίστροφη σειρά καρέ, οριζόντια αντιστροφή) και την κανονικοποίηση των δεδομένων για να διασφαλιστούν ομοιόμορφες διαστάσεις εισόδου για το μοντέλο. Το μοντέλο αξιολογήθηκε σε δύο διαθέσιμα δημόσια δυναμικά σετ δεδομένων χειρονομιών χεριών (VIVA και NVGesture), αποδεικνύοντας την αποτελεσματικότητά του στην εξαγωγή τόσο βραχυπρόθεσμων όσο και μακροπρόθεσμων χωροχρονικών πληροφοριών για την αναγνώριση χειρονομιών. Επίτευξε την υψηλότερη ακρίβεια στο σετ δεδομένων VIVA, ξεπερνώντας τις κορυφαίες μεθόδους κατά 5,46%. Επίσης, παρουσίασε ανταγωνιστικά αποτελέσματα στο σετ δεδομένων NVGesture, με ιδιαίτερες βελτιώσεις στην ακρίβεια για τα RGB και οπτικά ροής modalities.

Η παρούσα έρευνα [50] αναδεικνύει τη σημασία της αναγνώρισης χειρονομιών ως καινοτόμου προσέγγισης για την αλληλεπίδραση ανθρώπου-υπολογιστή, εστιάζοντας στην πρόκληση της αναγνώρισης δυναμικών χειρονομιών. Οι δυναμικές χειρονομίες, που εκτελούνται συνεχώς στον χρόνο, αντιμετωπίζουν πολλαπλές προκλήσεις όπως ο συνωστισμός στο παρασκήνιο, τα γρήγορα και μικρά κινήματα των χεριών και των δακτύλων, οι διαφορετικές συνθήκες φωτισμού και ο μεγάλος αριθμός χειρονομιών. Η έρευνα εξετάζει δύο κύριες κατηγορίες μεθόδων αναγνώρισης: τις φορετές συσκευές και τις τεχνικές υπολογιστικής όρασης. Παρόλο που οι φορετές συσκευές βελτιώνουν τη διαίρεση και αναγνώριση των χεριών, η έλλειψη ευκολίας έχει περιορίσει τη δημοτικότητά τους. Αντιθέτως, οι μέθοδοι βασισμένοι στην όραση έχουν αναπτυχθεί εκτενώς, χρησιμοποιώντας διάφορα διανύσματα χαρακτηριστικών και αλγορίθμους μηχανικής μάθησης, όπως τα HMM, τα τεχνητά νευρωνικά δίκτυα και τα SVM, για την αναγνώριση χειρονομιών. Η επαναστατική συνεισφορά των Συνελικτικών Νευρωνικών Δικτύων (CNN) στην επεξεργασία εικόνων και την αναγνώριση χειρονομιών τονίζεται, με την ικανότητά τους να εξάγουν σημαντικά χαρακτηριστικά μέσω συνέλιξης και μέγιστης συγκέντρωσης. Το Google-Net και άλλα επιφανειακά CNN όπως το Alex-Net αναδεικνύονται για την αποδοτική τους χρήση στην ταχύτητα και την ακρίβεια ταξινόμησης. Η προτεινόμενη προσέγγιση της έρευνας εστιάζει στη σύγκριση των 2D και 3D CNN για τη μάθηση των χρονικών αναπαραστάσεων χειρονομιών, αναδεικνύοντας τις προκλήσεις και τις λύσεις στην

αναγνώριση δυναμικών χειρονομιών. Η εργασία προτείνει μια καινοτόμο τεχνική συγχώνευσης δεδομένων χρώματος και βάθους, βελτιώνοντας την ακρίβεια και μειώνοντας τους χρόνους εκπαίδευσης και αναγνώρισης σε σύγκριση με τις παραδοσιακές μεθόδους. Οι κύριες συνεισφορές της έρευνας περιλαμβάνουν την απόδειξη ότι τα 2D CNN μπορούν να υπερτερούν των 3D CNN σε εργασίες ταξινόμησης βίντεο χειρονομιών με καλύτερη απόδοση και λιγότερη πολυπλοκότητα χρόνου, καθώς και την πρωτοποριακή αναπαράσταση βίντεο μέσω της συγχώνευσης των μορφών χρώματος και βάθους, καταστέλλοντας την κατάλληλη για εκπαίδευση με τακτικά 2D CNN. Αυτή η έρευνα ανοίγει νέους δρόμους για την αναγνώριση χειρονομιών, προσφέροντας προηγμένες λύσεις σε προκλήσεις που είχαν παραμείνει ανεπίλυτες μέχρι σήμερα, και ενδέχεται να επηρεάσει σημαντικά τις μελλοντικές εφαρμογές στον σχεδιασμό δι' επαφών ανθρώπου-υπολογιστή.

Το άρθρο των Yu και άλλων [60] παρουσιάζει μια αρχιτεκτονική με διπλή ροή δεδομένων ως είσοδο σε συνελικτικό νευρωνικό δίκτυο, σχεδιασμένη για την αποτελεσματική ένωση των modality εικόνας και οπτικής ροής για την αναγνώριση χειρονομιών. Αυτή η αρχιτεκτονική συνδυάζει προσεκτικά την εμφάνιση και τις πληροφορίες κίνησης των χειρονομιών. Η αρχιτεκτονική χρησιμοποιεί 3D CNNs για να ενσωματώσει πληροφορίες χρονικής εξάρτησης, επιτρέποντας την επεξεργασία δεδομένων από βίντεο καταγράφοντας χαρακτηριστικά από τις χωρικές και χρονικές διαστάσεις. Προτείνεται μια μέθοδος επιλεκτικής σύντηξης χαρακτηριστικών για να επιλεγούν καλύτερα χαρακτηριστικά από τα στατικά καρέ και την οπτική ροή, τονίζοντας την αλληλεπίδραση μεταξύ αυτών των στοιχείων αντί για την απλή συνένωσή τους. Το μοντέλο χρησιμοποιεί επίσης έναν μηχανισμό προσοχής διπλής ροής δεδομένων ως είσοδο μέσα στην αρχιτεκτονική για την εξαγωγή χαμηλού επιπέδου χωρικών σημασιολογικών και κινητικών πληροφοριών από τα αρχικά καρέ RGB και τα καρέ οπτικής ροής. Χρησιμοποιείται ένα βαθύ μοντέλο CNN, εμπνευσμένο από το μοντέλο Res-Net, για την απόκτηση υψηλού επιπέδου αναπαραστάσεων κάθε τμήματος βίντεο, ενισχύοντας την ικανότητα του μοντέλου να μαθαίνει διακριτικά χαρακτηριστικά από τους πολλαπλούς τύπους εισόδων. Η μεθοδολογία περιλαμβάνει προ-επεξεργασία δεδομένων, με τη χρήση του αλγορίθμου Brox για τον υπολογισμό οριζόντιων και κατακόρυφων καρέ οπτικής ροής, ταιριάζοντας αυτά τα καρέ με τα αντίστοιχα των RGB εικόνων. Επίσης, υπάρχει σύντηξη πληροφοριών από εικόνες RGB και οπτικής ροής που χρησιμοποιούνται ως είσοδο μέσω μιας επιλεκτικής διαδικασίας με τη χρήση βαρών προσοχής και την εφαρμογή ενός βαθύ μοντέλου CNN

βασισμένου στην αρχιτεκτονική Res-Net για την εξαγωγή υψηλού επιπέδου χαρακτηριστικών. Οι πληροφορίες των χαρακτηριστικών από τα δεδομένα συνδυάζονται και κατατάσσονται σε αντίστοιχες κατηγορίες χειρονομιών με τη χρήση ενός επιπέδου softmax. Ο σκοπός και η χρήση του μοντέλου είναι να βελτιώσει την ακρίβεια αναγνώρισης χειρονομιών μέσω της προσεκτικής σύντηξης χωρικών και κινητικών πληροφοριών από στατικές εικόνες και εικόνες οπτικών ορών(optical flow images). Αντιμετωπίζει προκλήσεις όπως οι μεταβαλλόμενες συνθήκες φωτισμού, τα περίπλοκα παρασκήνια και οι ατομικές διαφορές στην εκτέλεση των χειρονομιών. Αυτή η προσέγγιση είναι ιδιαίτερα κατάλληλη για εφαρμογές σε αναγνώριση νοηματικής γλώσσας, έξυπνη οδήγηση, εικονική πραγματικότητα και άλλους τομείς αλληλεπίδρασης ανθρώπου-υπολογιστή. Τα αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος επιτυγχάνει σημαντική βελτίωση στην ακρίβεια αναγνώρισης σε σύγκριση με τις υπάρχουσες μεθόδους, επιδεικνύοντας την αποτελεσματικότητα της επιλεκτικής σύντηξης χαρακτηριστικών και της ολοκλήρωσης χωρικών και χρονικών πληροφοριών. Το άρθρο αναφέρει ότι επιτεύχθηκε μια ακρίβεια top-1 95,77% και μια ακρίβεια top-5 99,74% χρησιμοποιώντας 10 τμήματα για κάθε βίντεο στο σύνολο δεδομένων Jester, υπογραμμίζοντας την αποδοτικότητα του μοντέλου στην καταγραφή διακριτικών χαρακτηριστικών που είναι απαραίτητα για την ακριβή αναγνώριση χειρονομιών.

Η μελέτη [58] παρουσιάζει μια καινοτόμο προσέγγιση για την αναγνώριση δυναμικών χειρονομιών στο σύνολο δεδομένων της Ινδικής Νοηματικής Γλώσσας (ISL), και βασίζεται σε ένα τρισδιάστατο Συνελικτικό Δίκτυο (3D-CNN). Η μεθοδολογία αξιοποιεί τα πλεονεκτήματα των 3D-CNN για την επεξεργασία χωροχρονικών χαρακτηριστικών, αποτυπώνοντας και τις στατικές και τις δυναμικές πτυχές των χειρονομιών. Η προτεινόμενη αρχιτεκτονική δικτύου ενσωματώνει πολλαπλά στάδια τρισδιάστατης συνέλιξης και max pooling για την εξαγωγή χαρακτηριστικών από βιντεοσκοπημένες ακολουθίες χειρονομιών. Τα χαρακτηριστικά αυτά στη συνέχεια μετατρέπονται σε έναν μονοδιάστατο διάνυσμα και περνούν μέσα από ένα Multi-Layer Perceptron (MLP) με ένα στρώμα softmax για την ταξινόμηση των 20 διαφορετικών χειρονομιών ISL. Αυτό το μοντέλο διακρίνεται για την εστίασή του στις δυναμικές χειρονομίες και την ικανότητά του να διαχειρίζεται τη μεταβλητότητα στην εκτέλεση χειρονομιών από διαφορετικά άτομα. Το σύνολο δεδομένων περιλαμβάνει 20 χειρονομίες από το πρότυπο ISL, τα οποία ηχογραφήθηκαν από δέκα υποκείμενα υπό διάφορες συνθήκες φόντου, φωτισμού και

προσανατολισμού. Αυτό οδήγησε σε 2400 βιντεοσκοπημένα δείγματα, τα οποία περαιτέρω πολλαπλασιάστηκαν μέσω κατακόρυφης αναστροφής για ένα πιο γενικευμένο προς χρήση σύνολο δεδομένων. Το σύνολο δεδομένων αποσκοπούσε να αντανακλά πραγματικές συνθήκες ενσωματώνοντας μια ποικιλία συμμετεχόντων και διαφορετικά περιβάλλοντα εγγραφής, ενισχύοντας την γενίκευση του μοντέλου. Το προτεινόμενο μοντέλο εκπαιδεύτηκε και επικυρώθηκε σε αυτό το εκτεταμένο σύνολο δεδομένων, εμφανίζοντας ενθαρρυντικά αποτελέσματα ως προς την συνολική ακρίβεια, την ακρίβεια συγκεκριμένης κλάσης, την ανάκλαση και τα σκορ F1. Συγκεκριμένα, το μοντέλο επιτυγχάνει ακρίβεια εκπαίδευσης 99.67% και ακρίβεια επικύρωσης 88% σε 100 εποχές. Αυτές οι μετρήσεις απόδοσης υποδηλώνουν την αποτελεσματικότητα του μοντέλου στην αναγνώριση χειρονομιών ISL, με ιδιαίτερες δυνατότητες στην ταυτοποίηση χειρονομιών που χρησιμοποιούνται συχνά στην ινδική κοινότητα.

Το άρθρο [57] προτείνει ένα σύστημα διπλής ροής εισόδου σε ένα βαθύ νευρωνικό δίκτυο για αναγνώριση χειρονομιών και κινήσεων, ενσωματώνοντας τόσο τα χωρικά όσο και τα χρονικά δεδομένα. Η πρώτη ροή χρησιμοποιεί ένα Τρισδιάστατο Συνελικτικό Νευρωνικό Δίκτυο (3D-CNN) για την καταγραφή χωρικών και χρονικών πληροφοριών από βίντεο χειρονομιών. Το δεύτερο ρεύμα χρησιμοποιεί ένα Δισδιάστατο Συνελικτικό Νευρωνικό Δίκτυο (2D-CNN), όπου η είσοδος είναι μια εικόνα Προτύπου Κίνησης Καθοδηγούμενη από Οπτική Ροή (OFMT ή Optical Flow Guided Motion Template). Αυτή η καινοτόμος προσέγγιση συνδυάζει την οπτική ροή και τα πρότυπα κίνησης για να δημιουργήσει μια υβριδική αναπαράσταση που ενισχύει την ανάλυση χρονικών δεδομένων. Η σύνθεση των εξόδων από και τα δύο ρεύματα επιτυγχάνεται με τη χρήση μιας μεθόδου βασισμένης στην πιθανότητα, με στόχο την αξιοποίηση των πληροφοριών τόσο από τα βίντεο RGB όσο και από τα πρότυπα κίνησης για την ενίσχυση της συνολικής απόδοσης. Η μεθοδολογία περιλαμβάνει την προετοιμασία και προ-επεξεργασία δεδομένων, χρησιμοποιώντας δύο κύρια σύνολα δεδομένων: το Palm's Graffiti Digits και ένα σύνολο δεδομένων που συλλέχθηκε εσωτερικά. Εφαρμόζονται διάφορες στρατηγικές ενίσχυσης δεδομένων για την αύξηση της ποικιλομορφίας και την πρόληψη της υπερεκπαίδευσης. Για την εξαγωγή χαρακτηριστικών, το OFMT συνδυάζει την οπτική ροή με πρότυπα κίνησης, μειώνοντας τον θόρυβο του παρασκηνίου και αποτυπώνοντας με ακρίβεια τα περιγράμματα των χειρονομιών. Καθορίζονται κανόνες σύνθεσης για να συνδυαστούν προβλέψεις από και τα δύο ρεύματα, μέσω παραμέτρων που προέρχονται από πρακτική εμπειρία

ώστε να επιτευχθεί η δυνατή σύνθεση των χαρακτηριστικών. Το μοντέλο δοκιμάστηκε στα σύνολα δεδομένων Palm's Graffiti Digits και σε ένα εσωτερικό σύνολο δεδομένων, επιδεικνύοντας εντυπωσιακές βελτιώσεις στην απόδοση σε σύγκριση με τις παραδοσιακές μεθόδους. Το μοντέλο 2D-CNN επιτυγχάνει υψηλότερη ακρίβεια με την ενίσχυση δεδομένων, ενώ το μοντέλο 3D-CNN αποδίδει καλά στην αποτύπωση χρονικών πτυχών των χειρονομιών. Η μέθοδος σύνθεσης παρείχε σημαντική ενίσχυση, οδηγώντας σε ακρίβεια 99,20%, υπερβαίνοντας τις υπάρχουσες μεθόδους και αποδεικνύοντας την αποτελεσματικότητα του συνδυασμού των χωρικών και χρονικών εξόδων των ροών.

Το άρθρο [59] προτείνει ένα μοντέλο βαθιάς εκμάθησης που έχει σχεδιαστεί ειδικά για την αναγνώριση δυναμικών χειρονομιών σε βίντεο συνεχούς ροής. Το καινοτόμο μοντέλο αποτελείται από ένα 3D Συνελικτικό Νευρωνικό Δίκτυο (3D-CNN) και από ένα Δίκτυο Μακροχρόνιας Μνήμης (LSTM). Αυτή η αρχιτεκτονική είναι ιδιαίτερα καινοτόμα καθώς ενσωματώνει τις δυνατότητες μάθησης χωρικών πληροφοριών των 3D-CNNs με τις ικανότητες μάθησης χρονικών πληροφοριών των LSTMs. Το μοντέλο δομείται έτσι ώστε να χρησιμοποιεί πρώτα το 3D-CNN για να μαθαίνει χωρικές πληροφορίες από διαδοχικά καρέ βίντεο, δημιουργώντας χάρτες χαρακτηριστικών. Αυτοί οι χάρτες χαρακτηριστικών μετατρέπονται στη συνέχεια σε διανύσματα και τροφοδοτούνται στο δίκτυο LSTM, το οποίο έχει σχεδιαστεί για να μαθαίνει χρονικές πληροφορίες και να ταξινομεί τις χειρονομίες. Για την αντιμετώπιση των προκλήσεων της υπερ-εκπαίδευσης, το μοντέλο χρησιμοποιεί τεχνικές κανονικοποίησης L2 και κανονικοποίησης παρτίδας (batch normalization). Η μεθοδολογία αποτελείται από αρκετά βήματα: φόρτωση δεδομένων, εμπλουτισμό δεδομένων, εκπαίδευση και δοκιμή. Για τον εμπλουτισμό δεδομένων, εφαρμόζονται αφινικές (affine) μετασχηματισμοί, κανονικοποιήσεις αντίθεσης και προσθετικός θόρυβος Γκαουσιανής κατανομής για να βελτιωθεί η γενικευτική ικανότητα μάθησης του μοντέλου. Η εκπαίδευση και η επικύρωση του μοντέλου πραγματοποιήθηκαν σε ένα υποσύνολο του σετ δεδομένων 20BN-jester, το οποίο περιλαμβάνει 148.092 βιντεοκλίπ δυναμικών χειρονομιών από διάφορες κατηγορίες. Λόγω περιορισμών υπολογιστικής ισχύος, η μελέτη επικεντρώθηκε σε 15 κατηγορίες από τις δυνατές 27, χρησιμοποιώντας 2000 τυχαία βίντεο ανά κατηγορία για εκπαίδευση και επικύρωση. Το προτεινόμενο μοντέλο πέτυχε ικανοποιητικά αποτελέσματα στην αναγνώριση δυναμικών χειρονομιών. Επιτυγχάνει ακρίβεια εκπαίδευσης 99%, ακρίβεια επικύρωσης 97,5% και ακρίβεια δοκιμής 97% σε δεδομένα που δεν έχει δει. Αυτή η απόδοση είναι αξιοσημείωτα

ανώτερη σε σύγκριση με τα βασικά μοντέλα που συζητήθηκαν, όπως το MobileNet-V2 + LSTM. Η υψηλή ακρίβεια του μοντέλου αποδίδεται εν μέρει στις εκτενείς τεχνικές εμπλουτισμού δεδομένων και στα συνδυασμένα πλεονεκτήματα των 3D-CNN και LSTM για την εκμετάλλευση των οπτικοακουστικών δεδομένων. Ο πίνακας σύγχυσης από τη φάση δοκιμής απεικονίζει την υψηλή ακρίβεια και ανάκληση του μοντέλου σε διάφορες κατηγορίες χειρονομιών, δείχνοντας την αξιοπιστία και την ικανότητα γενίκευσής του σε πραγματικές εφαρμογές.

Η επόμενη μελέτη [63] τονίζει ότι η αναγνώριση των χειρονομιών είναι μια τεχνολογία κλειδί για την επικοινωνία μεταξύ ανθρώπων με ακουστικές και ομιλητικές δυσκολίες, καθώς και για τη διευκόλυνση της αλληλεπίδρασης ανθρώπου-υπολογιστή. Εξετάζεται η ανάγκη για αποδοτική και ακριβή αναγνώριση χειρονομιών λόγω της αυξανόμενης ζήτησης για διερμηνευτικές υπηρεσίες σε άτομα με προβλήματα ακοής, καθώς και την επέκταση των εφαρμογών και των συσκευών χωρίς φυσική επαφή. Η έρευνα παρουσιάζει μια συγκριτική ανάλυση των πρόσφατων μελετών που εστιάζουν στην αναγνώριση και ταξινόμηση των χειρονομιών με την υποστήριξη διαφόρων τεχνολογιών, όπως οπτικές και μη-οπτικές προσεγγίσεις. Ειδικότερα, εξετάζεται η εφαρμογή αισθητήρων, γαντιών, λέιζερ, καθώς και συμβατικών και 3D καμερών για την αποτελεσματική ανίχνευση και αναγνώριση των χειρονομιών. Μεταξύ των κυριότερων συνεισφορών της μελέτης είναι η παροχή μιας επισκόπησης των υπαρχόντων γλωσσών των κινήσεων και των σχετικών εφαρμογών, καθώς και μιας ειδικής μελέτης στην Καζακική Γλώσσα. Επιπρόσθετα, η έρευνα αναδεικνύει τις προκλήσεις που εντοπίζονται στην αναγνώριση των χειρονομιών, όπως οι επιδράσεις του φωτισμού, του φόντου, του χρώματος του δέρματος, της απόστασης και της θέσης, καθώς και η κατεύθυνση του χεριού. Στο πλαίσιο της μελέτης αναλύονται διάφορες μέθοδοι και τεχνικές εξαγωγής χαρακτηριστικών, όπως οι δείκτες Fourier, PCA, καθώς και η χρήση του Ιστογράμματος Κατευθυνόμενων Ακμών (HOG) για την αποτελεσματική επεξεργασία εικόνων. Επίσης, εξετάζονται διάφοροι ταξινομητές, όπως τα Τεχνητά Νευρωνικά Δίκτυα (ANN), τα Συνελικτικά Νευρωνικά Δίκτυα (CNN και 3DCNN), η Ανάλυση Γραμμικών Διακρίσεων (LDA), οι Μηχανές Διανυσματικής Υποστήριξης (SVM), και άλλοι. Η εργασία καταλήγει στην ανάγκη για περαιτέρω έρευνα και ανάπτυξη εργαλείων λογισμικού για την αναγνώριση των χειρονομιών, ιδιαίτερα στο πεδίο της Καζακικής Γλώσσας των Κινήσεων. Τονίζεται η σημασία της ανάπτυξης αλγορίθμων και μεθόδων που θα είναι σε θέση να αντιμετωπίσουν τις προκλήσεις στην αναγνώριση χειρονομιών και να προσφέρουν ακριβείς και αποδοτικές λύσεις.

2.4 Μέθοδοι Βασισμένοι στον Ιδιοχώρο

Ο Patwardhan και ο Roy [37] πρότειναν ένα πλαίσιο βασισμένο στον χώρο Eigenspace για τη μοντελοποίηση δυναμικών χειρονομιών που περιλαμβάνουν και πληροφορίες σχήματος και πληροφορίες κίνησης. Οι μέθοδοι βασισμένες σε χαρακτηριστικά περιλαμβάνουν ένα ξεχωριστό και χρονοβόρο βήμα ανίχνευσης χαρακτηριστικών, το οποίο αποφεύγεται σε αυτόν τον αλγόριθμο. Ο αλγόριθμος είναι ανεκτικός σε παραμορφώσεις του σχήματος του χεριού: περιστροφή, μεταφορά, κλίμακα και διατμηματική μετατόπιση.

2.5 Προσαρμογή Καμπύλης

Ο Shin και άλλοι [39] πρότειναν μια γεωμετρική μέθοδο χρησιμοποιώντας καμπύλες Bezier για την ανάλυση και ταξινόμηση δυναμικών χειρονομιών. Οι χειρονομίες αναγνωρίζονται με την προσαρμογή της καμπύλης στην τρισδιάστατη κίνηση του χεριού. Η ταχύτητα της χειρονομίας ενσωματώνεται στον αλγόριθμο για να επιτρέψει την καλύτερη και πιο ακριβή αναγνώριση της χειρονομίας από τροχιές που έχουν μεταβολές στην ταχύτητα.

2.6 Δυναμικός Προγραμματισμούς και Δυναμική Προσαρμογή Χρόνου

Ο Kuremoto και άλλοι [40] πρότειναν μια μέθοδο αναγνώρισης χειρονομιών βασισμένη σε one-pass δυναμικό προγραμματισμό. Ένα σύστημα εξαγωγής χαρακτηριστικών βιολογικά εμπνευσμένο και βασισμένο στο μοντέλο retina-V1 που πρότειναν ο Tohyama και ο Fukushima [43] εκτιμά την κίνηση του χεριού. Οι χειρονομίες θεωρούνται ως συνδυασμοί προτύπων απλών κινήσεων. Οι κινήσεις χρησιμοποιούνται για να συνθέσουν ένα σύνολο από 40 πρότυπα χειρονομιών.

Η Δυναμική Προσαρμογή Χρόνου (Dynamic Time Wraping ή DTW), είναι μια εφαρμογή του δυναμικού προγραμματισμού, η οποία έχει χρησιμοποιηθεί ευρέως στην αναγνώριση απομονωμένων χειρονομιών. Ο Andrea Corradini [41] πρότεινε μια προσέγγιση βασισμένη σε πρότυπα με DTW για τη χρονική ευθυγράμμιση και κανονικοποίηση υπολογίζοντας μια χρονική μετατροπή μεταξύ των δύο σημάτων για να ταιριάξουν. Ο

Lichtenauer και άλλοι [42] πρότειναν Statical DTW (SDTW) για τη χρονική προσαρμογή και δύο ταξινομητές, δηλαδή τους διακριτικούς ανιχνευτές χαρακτηριστικών (Combined Discriminative feature Detectors ή CDFDs) και την τετραγωνική ταξινόμηση στα διακριτικά χαρακτηριστικά μέσω fisher mapping (Q-DFFM), για ταξινόμηση. Οι ταξινομητές έδειξαν ότι υπερτερούν του HMM και του SDTW.

Μια σύνοψη και σύγκριση των χαρακτηριστικών των αλγορίθμων αναγνώρισης χειρονομιών που ανασκοπούνται σε αυτή την ενότητα παρέχονται στους Πίνακες 1 και 2.

Πίνακας 1: Βλέπουμε τις μεθόδους αναγνώρισης χειρονομιών, χαρακτηριστικά που χρησιμοποιούνται, μέθοδοι ταξινόμησης, και αναφερόμενες εφαρμογές.

Εργασία	Χαρακτηριστικά που χρησιμοποιήθηκαν	Μέθοδος Ταξινόμησης	Εφαρμογή
[10]	Κατεύθυνση της κίνησης του χεριού	HMM (Κρυφό Μοντέλο Μαρκόβ)	Περιήγηση εντολών σε παρουσίαση Power-Point(R)
[11]	Θέση, γωνία και ταχύτητα χεριού	HMM (Κρυφό Μοντέλο Μαρκόβ)	HCI - αναγνώριση αλφαριθμητικών χαρακτήρων και γραφικών στοιχείων
[14]	Δείκτης Fourier /οπτική ροή	HMM (Κρυφό Μοντέλο Μαρκόβ)	Ταϊβανέζικη νοηματική γλώσσα
[12]	Σχήμα και κίνηση χεριού	HMM (Κρυφό Μοντέλο Μαρκόβ)	Εξ αποστάσεως ελέγχου ρομπότ
[17]	Δεδομένα τρισδιάστατης αρθρώσεως	Αθροιστικό HMM	Έλεγχος φώτων και κουρτινών σε έξυπνο σπίτι
[13]	3D τροχιά, μετατόπιση χεριού, χρώμα και σχήμα του 'blob' χεριού	HMM και IOHMM (Εισόδου/Εξόδου Κρυφό Μοντέλο Μαρκόβ)	Διάδραση-παιχνίδι, χειρισμός
[25]	Χαρακτηριστικά του Haar	Στατιστική/συντακτική ανάλυση	Μη καθορισμένο
[24]	Χαρακτηριστικά κατεύθυνσης	DBN (Δυναμικό Bayesian Δίκτυο)	Έλεγχος media player
[39]	Τρισδιάστατη τροχιά κίνησης	Προσαρμογή καμπύλης	Πλοήγηση των οπτικοποιημένων δεδομένων βιοπληροφορικής 3D
[37]	Σχήμα/τροχιά χεριού	Προβλεπτικός eigen-tracker	Έλεγχος αναπαραγωγής ήχου
[32]	Δισδιάστατο πεδίο/τροχιά κίνησης	NN (Νευρωνικό Δίκτυο)	Αμερικανική νοηματική γλώσσα

[34]	Δείκτες Fourier (σχήμα του 'blob' χεριού)	RBF, HMM, και RNN (Επαναλαμβανόμενο Νευρωνικό Δίκτυο)	Χειρισμός αντικειμένων στο περιβάλλον εργασίας των Windows
[29]	Κίνηση χεριού (κίνηση της ενέργειας)	FSM (Πεπερασμένο Σύστημα Καταστάσεων)	HRI (Αλληλεπίδραση Ανθρώπου-Ρομπότ)
[42]	Τρισδιάστατα χαρακτηριστικά της κίνησης του χεριού	CDFD και Q-DFFM (Τετραγωνική Ταξινόμηση σε Διακριτικά Χαρακτηριστικά Αντιστοίχισης Fisher)	Ολλανδική νοηματική γλώσσα

Περιγραφή: HMM-hidden Markov model, IOHMM-input/output hidden markov model, HCI-human computer interaction, DBN-dynamic Bayesian network, NN-neural network, RBF-radial basis function, RNN-recurrent neural networks, FSM-finite state machines, HRI-human robot interaction, CDFD-combined discriminative feature detectors,and QDFFM-quadratic classification on discriminative features fisher mapping.

Work	Accuracy	Class	Subject	Sample	UI	Spot	BG	Light	Scale	Light	Extensibility	CV	Data
[10]	93.14	10	8	6.2	0	1	0	0	0	0	0	0	0
[11]	93.25	48	20	5	0	1	0	0	1	1	0	0	0
[14]	93.60	20	20	3	0	0	1	0	1	0	0	0	0
[12]	81.71	5	5	14	0	1	1	1	0	1	0	0	0
[17]	95.42	8	1	60	0	1	0	0	0	0	0	0	0
[13]	75 και 98	16 και 7	20 και 7	50 και 10	0	0	0	0	0	0	0	0	1
[25]	87.21	4	1	25	0	1	0	0	1	1	0	0	0
[24]	99.59	10	7	1	0	1	0	0	0	0	0	1	0
[39]	97.9	10	4	2.38	0	0	1	0	0	1	0	0	0
[37]	100	8	1	2	0	0	0	0	1	0	0	0	0

[32]	96.21	40	1	7.6	0	1	1		0	1	0	1	0
[34]	91.9	14	1	21.07	0	1	0	0	1	0	0	0	0
[29]	Δεν αναφ.	5	1	1	0	0	0	0	0	0	0	0	0
[42]	92.3	120	75	15	1	1	0	1	0	0	0	1	0

Πίνακας 2: χαρακτηριστικά των αλγορίθμων και η πειραματική μεθοδολογία που υιοθετήθηκε κατά τη δοκιμή των αλγορίθμων (η λίστα στο τέλος παρέχει περιγραφή των τίτλων των στηλών). Τα χαρακτηριστικά στη στήλη 6 και μετά είναι δυαδικά, , όπου το 1 αντιπροσωπεύει τη συμμόρφωση της εργασίας με το χαρακτηριστικό, ενώ το 0 αντιπροσωπεύει τη μη συμμόρφωση.

Περιγραφή: Ακρίβεια-Η αναγνωριστική ακρίβεια του αλγορίθμου σε ποσοστό, Κλάση-Ο αριθμός των κατηγοριών χειρονομιών που ο αλγόριθμος αναγνωρίζει, Θέμα-Ο αριθμός των ατόμων στο σετ δοκιμών, Δείγμα-Ο αριθμός δειγμάτων δοκιμής ανά κλάση ανά άτομο, Ανεξαρτησία Χρήστη-Εάν ο αλγόριθμος δοκιμάστηκε με διαφορετικά άτομα από αυτά που χρησιμοποιήθηκαν για εκπαίδευση, Ανίχνευση-Εάν ο αλγόριθμος μπορεί να αναγνωρίσει χειρονομίες, Φόντο-Πολύπλοκο ή απλό φόντο, με το 1 για πολύπλοκο, Θόρυβος-Η παρουσία άλλων ανθρώπων στο φόντο, Κλίμακα-Εάν ο αλγόριθμος λαμβάνει υπόψη τις διαφορές στο μέγεθος των χειρονομιών, Φως-Εάν λαμβάνονται υπόψη οι διακυμάνσεις στο φωτισμό, Επεκτασιμότητα-Υποστήριξη για online ή offline μάθηση, με το 1 για online, Cross Validation-Εάν ο αλγόριθμος αξιολογήθηκε με Cross Validation, Δεδομένα-Εάν το σετ δεδομένων είναι δημόσιο ή ιδιωτικό, με το 1 για δημόσιο.

ΚΕΦΑΛΑΙΟ 3

ΠΡΟΣΦΑΤΕΣ ΤΑΣΕΙΣ ΣΤΗΝ ΑΝΑΓΝΩΡΙΣΗ ΧΕΙΡΟΝΟΜΙΩΝ: ΜΕΘΟΔΟΙ ΒΑΣΙΣΜΕΝΟΙ ΣΕ ΑΙΣΘΗΤΗΡΕΣ RGB-D

Οι κάμερες βάθους χρησιμοποιούνται στην υπολογιστική όραση πολλά χρόνια. Ωστόσο, η εφαρμογή των καμερών βάθους ήταν περιορισμένη λόγω της υψηλής τιμής τους και της κακής ποιότητάς τους. Η κυκλοφορία της χαμηλού κόστους χρωματικής-βάθους (RGB-D) κάμερας Kinect [65,66] από την Microsoft έχει δημιουργήσει μια επανάσταση στην αναγνώριση χειρονομιών παρέχοντας εικόνες υψηλής ποιότητας βάθους, αντιμετωπίζοντας ζητήματα όπως τα περίπλοκα φόντα και τις μεταβολές φωτισμού. Η συσκευή υπολογίζει έναν τρισδιάστατο χάρτη της σκηνής χρησιμοποιώντας έναν συνδυασμό από RGB και IR κάμερα. Πρόσφατα, ο Han και άλλοι [67] παρείχαν μια ανασκόπηση για το πώς η κάμερα Kinect είναι χρήσιμη στην αντιμετώπιση των θεμελιωδών προβλημάτων στην υπολογιστική όραση. Οι αισθητήρες όπως το Microsoft Kinect(R) και το ASUS Xtion PRO LIVE(R) παρέχουν αξιόπιστη παρακολούθηση των στάσεων του ανθρώπινου σώματος σε σενάρια παιχνιδιών. Βασισμένα στην παρακολούθηση, αυτές οι συσκευές παρέχουν χαρακτηριστικά όπως οι συντεταγμένες ενός σκελετικού μοντέλου, τα οποία χρησιμοποιούνται για την αναγνώριση χειρονομιών. Τα σκελετικά δεδομένα από αυτούς τους αισθητήρες RGB-D μπορούν να μετατραπούν σε σημαντικά και υψηλού επιπέδου χαρακτηριστικά, και επρόκειτο να αναπτυχθούν αλγόριθμοι για την καλύτερη ταξινόμηση των χειρονομιών. Η αναγνώριση των χειρονομιών του χεριού είναι ιδιαίτερα προκλητική λόγω της περίπλοκης αρθρωτής δομής και της σχετικά μικρής περιοχής του χεριού πάνω στην εικόνα. Επιπλέον, ένας αξιόπιστος αλγόριθμος αναγνώρισης χειρονομιών του χεριού πρέπει να έχει καλή κατανόηση του μεγέθους και της ταχύτητα της χειρονομίας, καθώς και τον προσανατολισμό της. Ο Rafael και άλλοι [68] αξιολόγησαν την

επιρροή της πληροφορίας βάθους στη διαδικασία αναγνώρισης χειρονομιών και κατέληξαν στο συμπέρασμα ότι η χρήση σιλουέτας βάθους αυξάνει σημαντικά την ακρίβεια αναγνώρισης. Ο Dominio και άλλοι [69] πρότειναν έναν αλγόριθμο για τον συνδυασμό πολλαπλών δεικτών βασισμένων στο βάθος για την αναγνώριση χειρονομιών του χεριού. Οι κάμερες RGB-D χρησιμοποιούνται κυρίως για την αναγνώριση στάσεων ολόκληρου του ανθρώπινου σώματος [66,70–73], καθώς αυτές οι κάμερες παρέχουν παρακολούθηση του σκελετού του σώματος. Αυτή η ενότητα του άρθρου ανασκοπεί αλγόριθμους αναγνώρισης χειρονομιών του χεριού βασισμένους σε κάμερες RGB-D, ταξινομώντας τη σχετική βιβλιογραφία σε δύο κατηγορίες, (α) προσεγγίσεις βασισμένες στο Kinect, και (β) άλλες προσεγγίσεις βασισμένες σε αισθητήρες RGB-D.

3.1 Μέθοδοι Βασισμένοι στην Κάμερα Kinect

Ο Wu και άλλοι [76] πρότειναν ένα σύστημα για την εκμάθηση χειρονομιών από μόνο ένα παράδειγμα μάθησης ανά κατηγορία, δηλαδή τη μέθοδο One-shot-learning. Τα χαρακτηριστικά εξάγονται βασισμένα στην τεχνική Extended-Motion-History-Image (Extended-MHI) και οι χειρονομίες ταξινομούνται υπολογίζοντας τον μέγιστο συντελεστή συσχέτισης. Οι εικόνες Motion History Images (MHI) [105] χρησιμοποιούνται για να αναπαραστήσουν τις κινήσεις ενός αντικειμένου σε ένα βίντεο. Όλα τα καρέ σε μια ακολουθία βίντεο προβάλλονται σε μία εικόνα κατά μήκος του χρονικού άξονα, για να αποτυπώσουν τις χρονικές πληροφορίες της ακολουθίας. Το extended-MHI προτείνεται για τη βελτίωση της απόδοσης του MHI αντισταθμίζοντας τις μη κινούμενες περιοχές και τις επαναλαμβανόμενες ενέργειες. Ο αλγόριθμος Multi-view Spectral Embedding (MSE) χρησιμοποιείται για τον συνδυασμό των δεδομένων RGB και βάθους με έναν φυσικά σημαντικό τρόπο. Ο αλγόριθμος MSE ανακαλύπτει την εγγενή σχέση μεταξύ των χαρακτηριστικών RGB και βάθους, βελτιώνοντας το ποσοστό αναγνώρισης του αλγορίθμου.

Ο Lui [77,84] πρότεινε έναν αλγόριθμο αναγνώρισης χειρονομιών βασισμένο σε ένα μη γραμμικό πλαίσιο παλινδρόμησης για πολλές περιπτώσεις. Η υποκείμενη γεωμετρία και μια μέθοδος ελαχίστων τετραγώνων χρησιμοποιούνται για την ανάπτυξη του αλγορίθμου. Η παλινδρόμηση ελαχίστων τετραγώνων διατυπώνεται ως μια σύνθετη συνάρτηση, λαμβάνοντας υπόψη γεωμετρικές ιδιότητες. Ο Gallo και άλλοι [75] πρότειναν ένα σύστημα αναγνώρισης χειρονομιών βασισμένο στο Kinect με εφαρμογή στην εξερεύνηση ιατρικών πληροφοριών από μία εικόνας. Διάφορες χειρονομίες για λειτουργίες όπως ζουμ, κίνηση, εξαγωγή περιοχής ενδιαφέροντος, περιστροφή και μεταφορά ιατρικών

εικόνων αναγνωρίζονται μέσω τοπολογικής ανάλυσης της περιοχής του χεριού. Μετρικές ευκλείδειας απόστασης και συνδιακυμάνσεις μιας λογαριθμικής-Ευκλείδειας μετρικής χρησιμοποιούνται ως χαρακτηριστικά στο [78]. Οι χειρονομίες ταξινομούνται χρησιμοποιώντας τον ταξινομητή του πλησιέστερου γείτονα.

Μια καινοτόμα προσέγγιση one-shot-learning για την αναγνώριση χειρονομιών από εικόνες βάθους κίνησης βασισμένη στην αντιστοίχιση προτύπων παρουσιάζεται στο [85]. Η μέθοδος βασίζεται στον υπολογισμό χωρικών-χρονικών δεικτών από το βίντεο, η οποία μετράει την ομοιότητα μιας χειρονομίας σε ένα λεξικό. Ο ταξινομητής βασίζεται στον συντελεστή συσχέτισης από την τυπική απόκλιση του μετασχηματισμού Fourier της εικόνας και του MHI.

Ένας αλγόριθμος για την ανίχνευση και αναγνώριση χειρονομιών του χεριού μέσω του συνδυασμού του DTW (Dynamic Time Warping) με εκτιμήσεις πιθανοτήτων προτείνεται στο [87]. Ο αλγόριθμος έχει γενίκευση σε σχέση με την θέση και τον προσανατολισμό του προσώπου που κάνει τη χειρονομία και της ταχύτητας της χειρονομίας. Οι Cheng και άλλοι [88,89] πρότειναν αλγορίθμους βασισμένους στο DTW για την αναγνώριση τρισδιάστατων χειρονομιών του χεριού. Ένα παραμετροποιημένο παράθυρο αναζήτησης εισάγεται στον πίνακα κόστους της παραδοσιακής προσέγγισης DTW για να ανιχνεύσει την αρχή και το τέλος συγκεκριμένων χειρονομιών από μια ατελείωτη ακολουθία χειρονομιών.

Ένας άλλος αλγόριθμος για αναγνώριση χειρονομιών με τη μέθοδο one-shot learning από δεδομένα RGB-D προτείνεται από τον Wan και άλλους [86]. Χρησιμοποιείται μια νέα χωροχρονική αναπαράσταση χαρακτηριστικών ονομαζόμενη 3D enhanced motion scale-invariant feature transform (3DEMoSIFT). Το νέο σύνολο χαρακτηριστικών είναι ανεξάρτητο από την κλίμακα και την περιστροφή καθώς συνδυάζει δεδομένα RGB-D. Μια μέθοδος αραιής κωδικοποίησης με την ονομασία simulation orthogonal matching pursuit (SOMP) εφαρμόζεται για να αναπαραστήσει κάθε χαρακτηριστικό μέσω ενός μικρού αριθμού κωδικοποιημένων λέξεων με γραμμικό συνδυασμό.

3.2 Μέθοδοι βασισμένοι σε RGB Κάμερες

Ο Holte και άλλοι [73] χρησιμοποίησαν μια κάμερα έντασης-βάθους (CSEM Swissranger SR-2) για να αναπτύξουν έναν αλγόριθμο αναγνώρισης χειρονομιών ανεξάρτητο από την οπτική γωνία. Σε αντίθεση με την καθιερωμένη μέθοδο αναγνώρισης βασισμένη σε

τροχιές, οι χειρονομίες αναγνωρίζονται με βάση τα βασικά στοιχεία κίνησης στα τρισδιάστατα δεδομένα. Τα πρωτογενή στοιχεία αναπαρίστανται με έναν τρόπο ανεξάρτητο από την οπτική γωνία χρησιμοποιώντας το harmonic shape context. Χρησιμοποιείται ένας ταξινομητής πιθανοτικής επεξεργασίας αποστάσεων για την ταξινόμηση. Ο αλγόριθμος έχει προσανατολιστική ανεξαρτησία, καθώς εκπαιδεύεται σε δεδομένα από μία οπτική γωνία και δοκιμάζεται σε δεδομένα από διαφορετική οπτική γωνία.

Στο [79], χρησιμοποιούνται πιθανοκρατικά δισδιάστατα πρότυπα που δημιουργούνται με βάση την τροχιά κίνησης του χεριού για την αναγνώριση δυναμικών χειρονομιών. Το πιθανοκρατικό πρότυπο λαμβάνει υπόψη διάφορες παραμορφώσεις της τροχιάς με διαφορετικές πιθανότητες. Ένας ταξινομητής Longest Common Subsequences (LCS) τροποποιείται σε ταξινομητή most probable longest common subsequence (MPLCS), για να μετρήσει την ομοιότητα μεταξύ του πιθανοκρατικού προτύπου και του δείγματος χειρονομίας του χεριού. Ο Erden και άλλοι [108] σχεδίασαν ένα σύστημα απομακρυσμένου ελέγχου βασισμένο σε χειρονομίες του χεριού που συνδυάζει αισθητήρες υπέρυθρων με μια κάμερα RGB.

ΚΕΦΑΛΑΙΟ 4

ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΔΥΝΑΜΙΚΩΝ ΧΕΙΡΟΝΟΜΙΩΝ

Ερευνητές από το Πανεπιστήμιο του Κέιμπριτζ και τη Microsoft Research διεξήγαγαν μια μελέτη [109] σχετικά με το πώς να αναπτύξουν παραδείγματα ώστε να γίνει καλύτερη η αντιπροσώπευση των δεδομένων των χειρονομιών για την εκπαίδευση αλγορίθμων μηχανικής μάθησης. Χρησιμοποίησαν δύο μετρικές, την ορθότητα και την κάλυψη, για να αξιολογήσουν πόσο καλά το σύνολο δεδομένων αντιπροσωπεύει τα δεδομένα του πραγματικού κόσμου. Η μετρική της ορθότητας αναφέρεται στην ομοιότητα των κινήσεων των χειρονομιών στα παραδείγματα σε σχέση με αυτές που ο χρήστης θέλει να εκτελέσουν. Δηλαδή, αν οι χειρονομίες που κάνει ο χρήστης είναι σωστές σε σχέση με το τι ήθελε να κάνει. Η μετρική της κάλυψης αναφέρεται στην πληρότητα του συνόλου δεδομένων στην αναπαράσταση φυσικών και πιθανών παραλλαγών των μοτίβων στις κινήσεις των χειρονομιών. Η κάλυψη σχετίζεται με την ελευθερία που δίνεται στο παράδειγμα. Ερευνήθηκε η πιο κατάλληλη σημειωτική μορφή οδηγιών και η σειρά τους για την επίτευξη της καλύτερης ορθότητας και κάλυψης, τόσο για το σύνολο δεδομένων όσο και για το σύστημα αναγνώρισης χειρονομιών που εκπαιδεύεται. Οι μορφές που ερευνήθηκαν περιλάμβαναν περιγραφικό κείμενο, στατική ακολουθία εικόνων και βίντεο. Το βίντεο ακολουθούμενο από κείμενο επιλέχθηκε ως η καλύτερη σειρά μορφών για να διευκολύνει τόσο την κατανόηση όσο και την ελευθερία των χειρονομιών στα παραδείγματα.

Οι τυποποιημένες βάσεις δεδομένων χειρονομιών του χεριού είναι απαραίτητες για την αξιόπιστη δοκιμή και σύγκριση αλγορίθμων αναγνώρισης χειρονομιών του χεριού. Η

διαθεσιμότητα βάσεων δεδομένων χειρονομιών του χεριού ήταν περιορισμένη μέχρι το έτος 2007 αλλά από τότε έχει αυξηθεί αρκετά. Αυτή η ενότητα παρέχει μια ανασκόπηση των διαθέσιμων στο κοινό συνόλων δεδομένων χειρονομιών του χεριού. Ο Πίνακας 3 καταγράφει βάσεις δεδομένων στάσεων και χειρονομιών του χεριού μαζί με τους διαδικτυακούς συνδέσμους για την λήψη τους. Ο Πίνακας 4 παρέχει πληροφορίες των βάσεων δεδομένων όπως ο αριθμός κλάσεων, παραδείγματα και τύποι δεδομένων που είναι διαθέσιμα. Οι εργασίες που χρησιμοποίησαν αυτά τα σύνολα δεδομένων περιλαμβάνονται επίσης για να πραγματοποιηθεί μια πιθανή συγκριτική μελέτη. Συνολικά, 23 σύνολα δεδομένων είναι διαθέσιμα κατά τον χρόνο δημοσίευσης αυτής της ανασκόπησης.

4.1 Βάσεις Δεδομένων με Στάσεις και Χειρονομίες του Χεριού από τους Sebastian-Marcel

Η βάση δεδομένων περιλαμβάνει τρία σύνολα δεδομένων στάσεων του χεριού, τη Βάση Δεδομένων Στατικής Στάσης Χεριού Jochen Triesch [64], τη Βάση Δεδομένων Στατικής Στάσης Χεριού Jochen Triesch II [63], και τη Βάση Δεδομένων Στατικής Στάσης Χεριού Sebastien Marcel [115], καθώς και μία βάση δεδομένων δυναμικών χειρονομιών του χεριού, τη Βάση Δεδομένων Δυναμικής Στάσης Χεριού Sebastien Marcel [15]. Τα σύνολα δεδομένων των στάσεων του χεριού περιλαμβάνουν απλά καθώς και περίπλοκα φόντα. Οι δυναμικές χειρονομίες περιλαμβάνουν διάφορες εντολές όπως Κλικ, Στοπ-αρπάξτε-οκ, Περιστροφή, και Όχι.

4.2 Βάση Δεδομένων Χειρονομιών του Cambridge

Αυτή η βάση δεδομένων περιέχει εικόνες στάσης των χεριών. Διαθέτει ακολουθίες στατικών εικόνων που αντιστοιχούν σε κινήσεις του χεριού, καθιστώντας την κατάλληλη για τη δοκιμή αλγορίθμων αναγνώρισης δυναμικών χειρονομιών [119]. Η βάση δεδομένων αποτελείται από χειρονομίες που ορίζονται από τρία αρχικά σχήματα του χεριού (επίπεδο, ανοιχτό, και σχήμα V) και 3 αρχικές κινήσεις (προς τα αριστερά, προς τα δεξιά, και συστολή). Ο στόχος για αυτή τη βάση δεδομένων είναι η ταξινόμηση των σχημάτων και των κινήσεων του χεριού ταυτόχρονα. Η βάση δεδομένων διαθέτει αρκετά μεγάλες διακυμάνσεις εντός κλάσης στην χωρική και χρονική ευθυγράμμιση των χειρονομιών.

4.3 Βάση Δεδομένων Χειρονομιών από τον Shen και άλλους

Η βάση δεδομένων είναι χρήσιμη στη δοκιμή τόσο των αλγορίθμων αναγνώρισης χειρονομιών όσο και στάσεων του χεριού, καθώς περιέχει μοτίβα κινήσεων και συγκεκριμένα σχήματα του χεριού [35]. Διαθέτει 10 κατηγορίες δυναμικών χειρονομιών του χεριού (π.χ. κίνηση δεξιά, κίνηση αριστερά, περιστροφή πάνω) που εκτελούνται με εφτά διαφορετικές στάσεις του χεριού (π.χ. αντίχειρας, γροθιά, όλα τα δάχτυλα εκτεταμένα), συνολικά 70 δείγματα χειρονομιών ανά παράδειγμα.

4.4 Βάση Δεδομένων Χειρονομιών Χειρισμού Αεροσκαφών NATOPS

Αυτή η βάση δεδομένων περιλαμβάνει 24 χειρονομίες του σώματος και του χεριού, επιλεγμένες από τα σήματα χειρισμού αεροσκαφών του NATOPS (Naval Air Training and Operating Procedures Standardization) [114]. Μια στερεοσκοπική κάμερα χρησιμοποιήθηκε για τη συλλογή των δεδομένων της βάσης. Αποτελείται από βίντεο με δεδομένα RGB και βάθους. Περιέχει επίσης τα εξαγόμενα σύνολα χαρακτηριστικών του σώματος και του χεριού σε μορφές Matlab και CSV.

4.5 Βάση Δεδομένων Χειρονομιών από τον Yoon και άλλους

Αυτή η βάση δεδομένων περιλαμβάνει 48 κατηγορίες αλφαβητικών χειρονομιών (αλφανουμερικοί χαρακτήρες και γραφικά στοιχεία) που έχουν καταγραφεί από 20 άτομα, 10 φορές για κάθε χειρονομία [11]. Η βάση δεδομένων περιέχει ακολουθίες συντεταγμένων x-y που αντιπροσωπεύουν χειρονομίες.

4.6 Βάση Δεδομένων Interact Play του Sebastian Marcel

Η βάση δεδομένων περιέχει τρισδιάστατες τροχιές από τμηματοποιημένες χειρονομίες του χεριού, συμπεριλαμβανομένου και των συντεταγμένων του κεφαλιού και του κορμού [13, 113]. Κάθε τροχιά αποθηκεύεται ως αρχείο κειμένου στη βάση δεδομένων. Η βάση δεδομένων περιλαμβάνει τόσο χειρονομίες με ένα χέρι (όπως στοπ, δείξε αριστερά, δείξε δεξιά) όσο και με δύο χεριά (όπως κολύμβηση, πέταγμα, χειροκρότημα). Οι

τροχιές των χειρονομιών περιέχουν τρισδιάστατες συντεταγμένες του κέντρου του κεφαλιού, των δύο χεριών και του κορμού.

4.7 Βάση Δεδομένων Χειρονομιών Keck

Η βάση δεδομένων χειρονομιών περιλαμβάνει 14 δυναμικές χειρονομίες, που είναι υποσύνολα στρατιωτικών σημάτων (όπως στρίψε αριστερά, πήγαινε πίσω, και επιτάχυνε) [118]. Η βάση δεδομένων διαιρείται σε δύο μέρη, σε σετ εκπαίδευσης και σε σετ δοκιμών. Το σετ εκπαίδευσης καταγράφεται χρησιμοποιώντας μια σταθερή κάμερα με το πρόσωπο να φαίνεται απέναντι σε ένα απλό και στατικό φόντο. Το σετ δοκιμών καταγράφεται από μια κινούμενη κάμερα, με την παρουσία φόντου με θορύβους και άλλα κινούμενα αντικείμενα.

4.8 Βάση Δεδομένων Χειρονομιών Κίνησης 6D

Η βάση δεδομένων χειρονομιών κίνησης 6D (6DMG) παρέχει ένα εκτενές σύνολο δεδομένων κίνησης χειρονομιών, συμπεριλαμβάνοντας τη θέση, τον προσανατολισμό, την επιτάχυνση και την γωνιακή ταχύτητα [112]. Τα δεδομένα αποθηκεύονται σε μια αρχική δυαδική μορφή και η βάση δεδομένων συνοδεύεται από δείγματα προγραμμάτων C++ για την πρόσβαση και την οπτικοποίηση των δεδομένων.

4.9 Δεδομένα Χειρονομιών ChaLearn

Η βάση δεδομένων χειρονομιών κίνησης 6D (6DMG) παρέχει ένα εκτενές σύνολο δεδομένων κίνησης χειρονομιών, συμπεριλαμβάνοντας τη θέση, τον προσανατολισμό, την επιτάχυνση και την γωνιακή ταχύτητα [112]. Τα δεδομένα αποθηκεύονται σε μια αρχική δυαδική μορφή και η βάση δεδομένων συνοδεύεται από δείγματα προγραμμάτων C++ για την πρόσβαση και την οπτικοποίηση των δεδομένων.

4.10 Δεδομένα Χειρονομιών Διαφορετικών Τύπων ChaLearn

Σε σύγκριση με τα δεδομένα χειρονομιών ChaLearn, η δοκιμή χρησιμοποιώντας τα multi-modal δεδομένα χειρονομιών ChaLearn [111] είναι πιο δύσκολη. Τα multi-modal

δεδομένα χειρονομιών ChaLearn περιλαμβάνουν την καταγραφή συνεχών ακολουθιών, την παρουσία χειρονομιών που υπάρχουν για να αποσπούν και να δυσκολεύουν, σχετικά μεγάλο αριθμό κατηγοριών, μακροσκελείς ακολουθίες χειρονομιών και χειρονομίες από διάφορους χρήστες. Πολλές λειτουργίες παρέχονται στη βάση δεδομένων, συμπεριλαμβανομένου του ήχου, των εικόνων RGB, των εικόνων βάθους, τα περιγράμματα των χρηστών και του σκελετικού μοντέλου των χρηστών.

4.11 Δεδομένα Χειρονομιών Διαφορετικών Τύπων ChAirGest

Αυτά τα δεδομένα αποκτήθηκαν με τη χρήση μιας κάμερας Kinect και τεσσάρων αδρανειακών μονάδων κίνησης που είναι προσαρτημένες στο δεξί βραχίονα και στον λαιμό των χρηστών. Οι χειρονομίες ξεκινούν από 3 διαφορετικές στάσεις ξεκούρασης και καταγράφονται σε 2 διαφορετικές συνθήκες φωτισμού [116].

4.12 Σύνολο Δεδομένων Χειρονομιών Sheffield Kinect (SKIG)

Το σύνολο δεδομένων SKIG [117] περιλαμβάνει 10 κατηγορίες χειρονομιών του χεριού, που έχουν καταγραφεί από 6 υποκείμενα χρησιμοποιώντας κάμερες RGB και Kinect. Τα δεδομένα έχουν καταγραφεί με 3 διαφορετικά φόντα (ξύλινη σανίδα, λευκό χαρτί, και χαρτί με χαρακτήρες) και 2 συνθήκες φωτισμού (φως και σκοτάδι).

4.13 Σύνολο Δεδομένων Χειρονομιών MSRC-12 Kinect

Το Microsoft Research Cambridge-12 (MSRC-12) είναι ένα σύνολο δεδομένων 12 κλάσεων δυναμικών χειρονομιών που έχει καταγραφεί χρησιμοποιώντας τα σκελετικά δεδομένα από Kinect [109]. Το σύνολο δεδομένων αποτελείται από ακολουθίες ανθρώπινων κινήσεων, αναπαριστώντας τη θέση τμημάτων του σώματος (20 σκελετικές αρθρώσεις). Περιλαμβάνει 594 ακολουθίες και 719.359 καρέ.

4.14 Σύνολο Δεδομένων nvGesture

Το σύνολο δεδομένων της NVIDIA [55] που ονομάζεται και nvGesture dataset, περιέχει 25 διαφορετικές χειρονομίες. Όλες οι χειρονομίες καταγράφθηκαν από

διαφορετικούς αισθητήρες. Οι καταγραφές ήταν συνεχείς και μπορούμε να χρησιμοποιήσουμε σε οποιοδήποτε μοντέλο μέχρι και 1532 δυναμικές χειρονομίες. Ο χώρος που έγινε η καταγραφή τους ήταν μέσα σε έναν προσομοιωτή αυτοκινήτου με φωτεινό και σκοτεινό τεχνητό φωτισμό. Συνολικά βοήθησαν 20 υποκείμενα για να δημιουργηθεί. Για την καταγραφή της μπροστινής όψης της χειρονομίας χρησιμοποιήθηκε ο SoftKinetic DS325 αισθητήρας, ενώ τοποθετημένος από πάνω από το υποκείμενο ήταν ένας DUO 3D αισθητήρας. Κατέγραψαν και ασπρόμαυρες και χρωματιστές εικόνες. Κάθε ακολουθία χειρονομίας αποτελείται από 30 καρέ ανά δευτερόλεπτο με εικόνα μεγέθους 320 επί 240 εικονοστοιχείων.

4.15 Σύνολο Δεδομένων ChaLearn IsoGD και ConGD

Τα δεδομένα του συνόλου δεδομένων CGD σχεδιάστηκαν για την τεχνική «one-shot learning», που σημαίνει εκπαίδευση με ένα παράδειγμα. Μόνο ένα παράδειγμα κάθε κλάσης υπάρχει σε κάθε batch και τα υπόλοιπα δεδομένα χρησιμοποιούνται για αξιολόγηση. Κάθε batch περιέχει 100 χειρονομίες. Στο συγκεκριμένο σύνολο δεδομένων ένα Lexicon ορίζεται ως ένα εύρος χειρονομιών οι οποίες έχουν κάποιο κοινό θέμα, όπως για παράδειγμα χειρονομίες για κωφούς, υποθαλάσσιες χειρονομίες και άλλα. Τα δύο νέα σύνολα δεδομένων [154] είναι βγαλμένα από το ChaLearn Dataset και μερικές διαφορές. Το ChaLean LAP IsoGD περιέχει απομονωμένες χειρονομίες ενώ το ConGD περιέχει συνεχείς, όπου μια ακολουθία μπορεί να δείχνει πολλαπλές χειρονομίες κατά την διάρκεια της.

4.16 Σύνολο Δεδομένων LeapMotion-Gesture και Handicraft-Gesture

Τα δεδομένα του συνόλου δεδομένων CGD σχεδιάστηκαν για την τεχνική «one-shot learning», που σημαίνει εκπαίδευση με ένα παράδειγμα. Μόνο ένα παράδειγμα κάθε κλάσης υπάρχει σε κάθε batch και τα υπόλοιπα δεδομένα χρησιμοποιούνται για αξιολόγηση. Κάθε batch περιέχει 100 χειρονομίες. Στο συγκεκριμένο σύνολο δεδομένων ένα Lexicon ορίζεται ως ένα εύρος χειρονομιών οι οποίες έχουν κάποιο κοινό θέμα, όπως για παράδειγμα χειρονομίες για κωφούς, υποθαλάσσιες χειρονομίες και άλλα. Τα δύο νέα σύνολα δεδομένων [154] είναι βγαλμένα από το ChaLearn Dataset και μερικές διαφορές. Το ChaLean LAP IsoGD περιέχει απομονωμένες χειρονομίες ενώ το ConGD περιέχει συνεχείς, όπου μια ακολουθία μπορεί να δείχνει πολλαπλές χειρονομίες κατά την διάρκεια

της.Αυτά τα δύο σύνολα δεδομένων [149] χτίστηκαν μέσω του αισθητήρα βάθους LMC (Leap Motion Controller). Το πρώτο, περιέχει δώδεκα χειρονομίες-λέξεις από το ASL (American Sign Language) οι οποίες είναι το μπάνιο, μπλε, τελείωμα, πράσινο, πείνα, γάλα, παρελθόν, γουρούνι, και μαγαζί, όπου, το γράμμα j και z. Το δεύτερο σύνολο δεδομένων δημιουργήθηκε για να υπάρχει μέτρο σύγκρισης στα δεδομένα από τον αισθητήρα LMC και περιέχει δέκα χειρονομίες που σχετίζονται με ικανότητες αγγειοπλαστικής όπως τρύπημα, τσίμπημα, τράβηγμα, ξύσιμο, χαστούκι, πάτημα, κόψιμο, κύκλος, πάτημα κλειδού και κούρεμα. Το πρώτο σύνολο δεδομένων περιέχει 360 και το δεύτερο 300 ακολουθίες βάθους, ενώ ο ρυθμός των καρέ είναι 60.

4.17 Σύνολο Δεδομένων DHG

Το σύνολο δεδομένων DHG [142] δημιουργήθηκε στα πλαίσια του διαγωνισμού SHREC2017-3D Shape Retrieval Contest. Τα δεδομένα καταγράφηκαν από την κάμερα βάθους Intel RealSense. Οι ακολουθίες είναι μεταβλητού μήκους και φτάνουν τον αριθμό των 2800, ενώ οι κλάσεις στις οποίες είναι χωρισμένο τα σύνολο δεδομένων είναι 14.

4.18 Σύνολο Δεδομένων EgoGesture

Το EgoGesture σύνολο δεδομένων [155] προσφέρει μια διαφορετική οπτική από τα υπόλοιπα καθώς η κάμερα στις ακολουθίες είναι σε πρώτο-πρόσωπο (first-person-view). Περιέχει περίπου 24.000 δείγματα. Υπάρχουν εικόνες και σε RGB και σε βάθος. Συνολικά περιέχει 83 κλάσεις οι οποίες καταγράφθηκαν σε Intel RealSense κάμερα. Ο ρυθμός των καρέ είναι 30 και οι εικόνες έχουν ανάλυση 640 επί 480.

4.19 Σύνολο Δεδομένων Jester

Το σύνολο δεδομένων Jester [150], είναι το μεγαλύτερο που έχει δημιουργηθεί και περιέχει τις δυναμικές χειρονομίες από 1300 διαφορετικούς ανθρώπους. Έχει 148.092 ακολουθίες με χειρονομίες 25 κλάσεις και 2 κλάσεις οι οποίες θα πρέπει να αναγνωριστούν ως μην μετρήσιμες καθώς δεν ανήκουν σε κάποια κατηγορία χειρονομίας αλλά είναι τυχαίες κινήσεις.

4.20 Σύνολο Δεδομένων IPN

Το IPN είναι επίσης ένα μεγάλο σύνολο δεδομένων [138] που έχει ως σκοπό να περιέχει μέσα εικόνες από τον πραγματικό κόσμο. Έχει 13 κλάσεις και όλες καταγράφθηκαν από ανθρώπους που καθόντουσαν μπροστά από την κάμερα του υπολογιστή τους. Συνολικά συλλέχθηκαν 4.218 RGB ακολουθίες. Οι ακολουθίες δεν έχουν σταθερό μήκος.

4.21 Σύνολο Δεδομένων SCUT-DHGA

Αυτό το σύνολο δεδομένων [148] περιέχει 29.160 ακολουθίες από RGB και εικόνες βάθους. Ο λόγος που δημιουργήθηκε είναι για τα βιομετρικά στοιχεία ελέγχου. Περιέχει 6 χειρονομίες οι οποίες καταγράφθηκαν σε απόσταση από 0.6 έως 0.8 μέτρα.

4.22 Σύνολο Δεδομένων GestureMNIST

Το GestureMNIST σύνολο δεδομένων [145], αποτελείται από 6 κλάσεις και κάθε κλάσεις έχει ακολουθίες που έχουν μέγεθος 12 καρέ. Ονομάζεται MNIST επειδή και αυτό έχει εικόνες με ανάλυση 28 επί 28 και τα δείγματα είναι περίπου 80.000.

4.23 Σύνολο Δεδομένων ArSL (Arabic Sign Language)

Αυτό το σύνολο δεδομένων [140] περιέχει αραβικές νοηματικές χειρονομίες. Έχει 23 κλάσεις και 150 δείγματα για κάθε χειρονομία. Η καταγραφή των εικόνων έγινε σε recording studio κάτω από ιδανικές συνθήκες φωτισμού.

Παρακάτω σας παραθέτω τους Πίνακες 3 και 4.

Πίνακας 3: Πληροφορίες των συνόλων δεδομένων και σύνδεσμοι για την χρήση τους

Αριθμός	Ονομασία, Έτος	Πηγή
1	ChaLearn gesture, 2011	http://gesture.chalearn.org/data

2	MSRC-12Kinectgesture, 2012	http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/
3	ChaLearn multi-modal gesture data, 2013	http://sunai.uoc.edu/chalearn/
4	6Dmotiongesturedatabase, 2011	2011 http://www.ece.gatech.edu/6DMG/6DMG.html
5	Sebastien Marcel interact play database, 2004	http://www.idiap.ch/resource/interactplay/
6	NATOPS aircraft handling signals database, 2011	http://groups.csail.mit.edu/mug/natops/
7	Sebastien Marcel hand posture and gesture datasets,2001	http://www.idiap.ch/resource/gestures/
8	Gesture dataset by Shenetal.,2012	http://users.eecs.northwestern.edu/~xsh835/GestureDataset.zip
9	Gesture dataset by Yoonetal.,2001	Available on e-mail request to yoonhs@etri.re.kr
10	ChAirGest multi-modal dataset, 2013	https://project.eia-fr.ch/chairgest/Pages/Download.aspx
11	Sheffield Kinect Gesture Dataset, 2013	http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm
12	Keck gesture dataset,2009	http://www.umiacs.umd.edu/~zhuolin/Keckgesturedataset.html
13	Cambridge hand gesture dataset,2007	http://www.iis.ee.ic.ac.uk/~tkkim/ges_db.htm
14	nvGesture, 2016	https://www.v7labs.com/open-datasets/nvgesture
15	ChaLearn Variations & ConGD, 2016	https://gesture.chalearn.org/2016-looking-at-people-cvpr-challenge/isogd-and-congd-datasets
16	LeapMotion-Gesture 3D & Handicraft-Gesture, 2016	https://www-intuidoc.irisa.fr/en/english-leap-motion-dynamic-hand-gesture-lmdhg-database/
17	DHG, 2017	http://www-rech.telecom-lille.fr/DHGdataset/
18	EgoGesture, 2018	https://nlpr.ia.ac.cn/iva/yfzhang/datasets/egogesture.html
19	Jester, 2019	https://developer.qualcomm.com/software/ai-datasets/jester
20	IPN, 2020	https://gibranbenitez.github.io/IPN_Hand/
21	SCUT-DHGA, 2021	https://github.com/SCUT-BIP-Lab/SCUT-DHGA
22	GestureMNIST, 2022	-
23	KSU-SSL, 2022	https://ieee-dataport.org/documents/ksu-arsl-arabic-sign-language

Πίνακας 4: S-Στατικές, D-δυναμικές. Γενικές πληροφορίες όπως ο αριθμός των κλάσεων, ο τύπος δεδομένων και πόσοι άνθρωποι βοήθησαν στην δημιουργία των συνόλων δεδομένων.

Αριθμός	Περιγραφή	S/D	Άρθρα
1	ChaLearn Gesture Challenge, 62.000 δείγματα	D	[82,90-92,100,127]
2	12 κλάσεις, 30 υποκείμενα, 6244 δείγματα	D	[126]
3	20 κλάσεις, 27 υποκείμενα, 13,858 δείγματα	D	[128]
4	20 κλάσεις, 28 υποκείμενα, 5600 δείγματα	D	[130]
5	16 κλάσεις, 22 υποκείμενα, 50 δείγματα/υποκείμενο	D	[13, 131]
6	24 κλάσεις, 20 υποκείμενα, 9600 δείγματα	S και D	[132]
7	Τρία σύνολα δεδομένων χειρονομιών: 10 ασπρόμαυρες, 12 με χρώμα, 6 ασπρόμαυρες, Ένα σύνολο δεδομένων για 4 κλάσεις	S και D	[63-65,133]
8	10 κλάσεις, 15 υποκείμενα, 1050 δείγματα	S και D	[35]
9	48 κλάσεις, 20 υποκείμενα, 9600 δείγματα	D	[11]
10	10 κλάσεις, 10 υποκείμενα, 1200 δείγματα, καταγραφή με Kinect και αδρανειακές μονάδες κίνησης	D	[134]
11	10 κλάσεις, 6 υποκείμενα, 2160 δείγματα, καταγραφή με Kinect και RGB κάμερες	D	[135]
12	14 κλάσεις, 3 υποκείμενα, 126 δείγματα εκπαίδευσης και 168 δείγματα δοκιμών10 κλάσεις, 1 υποκείμενο, 240 δείγματα, χρώμα και ασπρόμαυρο	D	[136]
13	9 κλάσεις, 2 υποκείμενα, 900 ακολουθίες εικόνων, με διαφορετικές συνθήκες φωτισμού	S και D	[137]
14	25 κλάσεις, 1532 ακολουθίες εικόνων, με RGB-D και υπέρυθρες	D	[55]
15	249 κλάσεις, 21 υποκείμενα, 47.933 ακολουθίες, με RGB-D εικόνες	D	[154]
16	13 +1 κλάσεις,21 υποκείμενα, 50 ακολουθίες	D	[149]
17	-	D	[142]
18	83 κλάσεις, 24.161 ακολουθίες, RGB-D	D	[155]
19	27 κλάσεις, 148.092 ακολουθίες	D	[150]
20	13 κλάσεις, 50 υποκείμενα, 4.000 ακολουθίες, RGB	S και D	[138]
21	6 κλάσεις, 29.160 ακολουθίες, RGB-D	D	[148]
22	-	D	[145]
23	80 κλάσεις, 40 υποκείμενα, 16.000 ακολουθίες, RGB	D	[140]

ΚΕΦΑΛΑΙΟ 5

ΣΥΜΠΕΡΑΣΜΑΤΑ

Παρά τις εξελίξεις στην αναγνώριση των χειρονομιών, υπάρχουν ακόμα ανεπίλυτες προκλήσεις στην αναγνώριση χειρονομιών. Αυτή η ενότητα ανασκοπεί συνοπτικά κάποια από τα ανεπίλυτα ζητήματα στον τομέα, παρέχει μια σύγκριση διαφόρων προσεγγίσεων και συζητά μερικές μελλοντικές κατευθύνσεις έρευνας.

5.1 Αναγνώριση των Επεξηγηματικών Χειρονομιών

Η αναγνώριση των επεξηγηματικών χειρονομιών αποτελεί πρόκληση καθώς η σημασία αυτών των χειρονομιών εξαρτάται από το πλαίσιο. Το πλαίσιο πρέπει να είναι αναγνωρίσιμο όπως και η επεξηγηματική χειρονομία. Από τις διάφορες επεξηγηματικές χειρονομίες, η χειρονομία στην οποία κάποιος δείχνει είναι πολύ χρήσιμη σε εφαρμογές όπως η εντολή σε κινητά ρομπότ. Η κατανόηση μιας χειρονομίας ενός ατόμου που δείχνει προς κάποια κατεύθυνση σε 3D δεδομένα περιλαμβάνει την ανίχνευση της χειρονομίας, τον εντοπισμό της θέσης του χεριού και την εξακρίβωση της κατεύθυνσης που δείχνει η χειρονομία. Η δυσκολία στην ακριβή εκτίμηση της κατεύθυνσης αυτής καθιστά την

αναγνώριση της χειρονομίας μια πρόκληση. Η κατεύθυνση που κοιτάζει το άτομο είναι χρήσιμη πληροφορία στην αναγνώριση μιας χειρονομίας. Για παράδειγμα, μια γραμμή που ενώνει το κέντρο των ματιών με την άκρη του δείκτη δακτύλου μπορεί να παρέχει μια εκτίμηση της κατεύθυνσης, η οποία με τη σειρά της μπορεί να χρησιμοποιηθεί για τον εντοπισμό του σημείου που δείχνει η χειρονομία [122, 123].

Η εκτίμηση της κατεύθυνσης της χειρονομίας με τη χρήση της γραμμής κεφαλιού-χεριού είναι αποτελεσματική όταν η χειρονομία που δείχνει ένα σημείο επεκτείνεται προς τα έξω και βρίσκεται στην επιφάνεια ενός φανταστικού ημισφαιρίου που έχει ως κέντρο τον ώμο [122, 123]. Παρόλο που αυτή η μέθοδος δεν είναι αποτελεσματική στην περίπτωση χειρονομιών κατά τις οποίες κάποιος κινεί μόνο τον πρόσθιο βραχίονα, μπορεί να κατανοήσει μικρές λεπτομέρειες της κίνησης του δείκτη μοντελοποιώντας τα κινητικά χαρακτηριστικά του πρόσθιου βραχίονα και του δείκτη [124]. Ο προσανατολισμός του κεφαλιού μπορεί να χρησιμοποιηθεί ως χαρακτηριστικό για τη βελτίωση της αναγνώρισης της χειρονομίας [126]. Σε σύγκριση, η εκτίμηση της διεύθυνσης της χειρονομίας με τη χρήση της γραμμής κεφαλιού-χεριού υπερτερεί αυτής που βασίζεται στον προσανατολισμό του πήχη. Ο Raheja και άλλοι [127] πρότειναν έναν αλγόριθμο για την ανίχνευση της τοποθεσίας της χειρονομίας που εστιάζει σε ένα σημείο, ο οποίος βασίζεται στις θέσεις του κεφαλιού, των ώμων και των αγκώνων. Η μέθοδος που πρότεινε ο Pateraki και άλλοι [128, 129] συνδυάζει τη στάση του προσώπου και τον προσανατολισμό του κεφαλιού με την κατεύθυνση του χεριού.

5.2 Προσεγγίσεις Βασισμένες στην Εμφάνιση και στην Μοντελοποίηση

Οι προσεγγίσεις βασισμένες στην εμφάνιση προσφέρουν καλύτερη απόδοση σε πραγματικό χρόνο σε σύγκριση με τις προσεγγίσεις που βασίζονται σε τρισδιάστατα μοντέλα του χεριού, καθώς η διαδικασία εξαγωγής χαρακτηριστικών από την εικόνα είναι ταχύτερη. Τα μοντέλα βασισμένα στην εμφάνιση οδηγούν τους αλγορίθμους σε καλύτερη υπολογιστική απόδοση και λειτουργούν καλά σε περιορισμένες συνθήκες. Ωστόσο τους λείπει η γενίκευση που είναι επιθυμητή για την ανθρώπου-υπολογιστή αλληλεπίδραση. Οι μέθοδοι βασισμένοι στην εμφάνιση χρησιμοποιούν κυρίως τα 2D δεδομένα του σχήματος του χεριού, τα οποία εξαρτώνται από τη γωνία προβολής. Η χρήση τέτοιων μεθόδων περιορίζεται από τις γωνίες προβολής. Μια ευρεία κλάση χειρονομιών θα μπορούσε να καλυφθεί σε προσεγγίσεις βασισμένες σε τρισδιάστατα μοντέλα, καθώς τα

μοντέλα προσφέρουν έναν τρόπο για εκτενή επεξεργασία των χειρονομιών του χεριού. Ωστόσο, τα 3D μοντέλα απαιτούν μεγάλη βάση δεδομένων εικόνων για να καλύψουν όλα τα χαρακτηριστικά σχήματα και τις παραλλαγές τους υπό διαφορετικές προβολές. Η σύγκριση της δοκιμαστικής εικόνας με όλα τα μοντέλα στη βάση δεδομένων είναι χρονοβόρα και υπολογιστικά δαπανηρή, περιορίζοντας τη χρήση των 3D μοντέλων για εφαρμογές σε πραγματικό χρόνο.

5.3 Χαρακτηριστικά

Η επιλεκτικότητα και η αμετάβλητη είναι δύο επιθυμητές και απαιτούμενες ιδιότητες για κάθε διαδικασία αναγνώρισης μοτίβων βασισμένη σε εικόνες. Οι προσεγγίσεις ως τώρα προσφέρουν καλή επιλεκτικότητα για μοτίβα σχήματος, αλλά τους λείπει η αμεταβλησία. Οι προσεγγίσεις βασισμένες σε ιστογράμματα έχουν την ιδιότητα της αμεταβλησίας. Ωστόσο, οι προσεγγίσεις αυτές λαμβάνουν υπόψη όλες τις πληροφορίες της εικόνας, κάτι που τις καθιστά ακατάλληλες για χρήση σε αναγνωρίσεις σχημάτων, όπως η αναγνώριση στάσεων του χεριού. Τα μοτίβα σχήματος-υφής που εξάγονται με τεχνικές εμπνευσμένες από τη βιολογία [130] παρέχουν χαρακτηριστικά που έχουν και επιλεκτικότητα και αμεταβλησία, και είναι χρήσιμα στην αναγνώριση στάσεων του χεριού [44].

Οι προσανατολισμοί και τα γωνιακά χαρακτηριστικά των χειρονομιών προσφέρουν καλύτερη αμεταβλησία σε σύγκριση με δεδομένα που δείχνουν μόνο θέση. Από την άλλη πλευρά, τα χαρακτηριστικά της θέσης είναι απλά και μπορούν να εξαχθούν με καλύτερη ακρίβεια. Τα χαρακτηριστικά βασισμένα στην υφή έχουν την ικανότητα να αποτυπώνουν καλύτερα τις χωρικές ιδιότητες της χειρονομίας σε σύγκριση με αυτά που αποτυπώνονται από χαρακτηριστικά όπως το χρώμα.

Οι αισθητήρες RGB-D επιτρέπουν την εξαγωγή αμετάβλητων χαρακτηριστικών παρά τα περίπλοκα παρασκήνια και τις μεταβολές σε κλίμακα, φωτισμό και οπτικές γωνίες. Τα ακριβή δεδομένα βάθους και η πληροφορία θέσης από αυτούς τους αισθητήρες επιταχύνουν την εξαγωγή των μοντέλων του χεριού, αυξάνοντας τη χρησιμότητα των προσεγγίσεων που βασίζονται σε μοντέλα.

5.4 Μέθοδοι Ταξινόμησης

Οι μέθοδοι βασισμένες σε Markov Hidden Models (HMM) είναι αποτελεσματικές και χρησιμοποιούνται ευρέως για την Αναγνώριση Χειρονομιών του Χεριού (Human Gesture Recognition ή HGR). Ωστόσο, οι προσεγγίσεις βασισμένες σε HMM απαιτούν έναν μεγάλο αριθμό δειγμάτων για εκπαίδευση και έχουν το μειονέκτημα μιας περίπλοκης διαδικασίας εκπαίδευσης. Το υπολογιστικό κόστος των αλγορίθμων βασισμένων σε HMM αυξάνεται με το λεξιλόγιο των χειρονομιών. Επιπλέον, η απόδοση των αλγορίθμων βασισμένων σε HMM μειώνεται όταν υπάρχουν παραλλαγές μεταξύ των συνθηκών εκπαίδευσης και δοκιμής. Η εύρεση των βέλτιστων συνόλων παραμέτρων και της καλύτερης τροχιάς της κίνησης για την χρονική τμηματοποίηση αποτελούν επιπλέον εμπόδια στη χρήση HMM.

Ο σχεδιασμός ενός Time-Delay Neural Network (TDNN) είναι ελκυστικός, καθώς η συμπαγής δομή του εξοικονομεί βάρη και καθιστά δυνατή την ανάπτυξη και ανίχνευση πιο γενικών χαρακτηριστικών. Η ιεραρχία των καθυστερήσεων στο TDNN βελτιστοποιεί αυτούς τους μηχανισμούς για την ανίχνευση των χαρακτηριστικών, αυξάνοντας το πεδίο εφαρμογής τους σε κάθε επίπεδο. Η χρήση των χρονικών χαρακτηριστικών στο επίπεδο εξόδου καθιστά το δίκτυο αμετάβλητο στις μετατοπίσεις (αδιαφορία για την ακριβή θέση του χεριού). Ο συνολικός αριθμός των βαρών στο δίκτυο είναι σχετικά μικρός, καθώς μόνο ένα μικρό παράθυρο του εισαγόμενου μοτίβου τροφοδοτείται στο TDNN σε κάθε στιγμιότυπο. Αυτό βοηθά στη μείωση του χρόνου εκπαίδευσης.

Οι αλγόριθμοι βασισμένοι σε γράφους έχουν το μειονέκτημα της υψηλής υπολογιστικής πολυπλοκότητας, κάτι που τους καθιστά ακατάλληλους για τις εφαρμογές σε πραγματικό χρόνο. Ωστόσο, κάθε κόμβος στο γράφο μπορεί να μοντελοποιηθεί με ένα σύνολο χαρακτηριστικών, το οποίο είναι χρήσιμο για την αντιμετώπιση ζητημάτων όπως τα περίπλοκα παρασκήνια [63] και τις παραλλαγές στο μέγεθος ή στο σχήμα της χειρονομίας.

5.5 Προκλήσεις και Μελλοντικές Κατευθύνσεις

Η αναγνώριση της φάσης της χειρονομίας αποτελεί μια σημαντική πρόκληση στην Αναγνώριση Χειρονομιών του Χεριού (HGR). Η παρουσία απρόβλεπτων και ασαφών αντικειμένων που δεν σχετίζονται με την χειρονομία του χεριού ή την κίνησή της, καθιστά το έργο δύσκολο. Η ικανότητα απόρριψης άγνωστων κλάσεων είναι ένα από τα σημαντικά απαιτούμενα για έναν αυτόματο σύστημα αναγνώρισης χειρονομιών. Το μοντέλο

με όριο που εισήχθη από τον Lee και Kim [10] είναι χρήσιμο για αυτό το σκοπό. Ο αλγόριθμος προτεινόμενος από τον Kim και άλλους [17] για συνεχή τμηματοποίηση της εικόνας στο χέρι και την αναγνώριση των χειρονομιών του χεριού, χρησιμοποιεί μια συνεχή εκτίμηση μέσω πιθανοτήτων για χειρονομίες και μη-χειρονομίες, ώστε να βρει τα σημεία έναρξης/λήξης της χειρονομίας. Ο Kang και άλλοι [131] πρότειναν ένα σχέδιο αναγνώρισης βασισμένο στην εντοπισμό χειρονομιών, φιλτράροντας εκτός τις κινήσεις που δεν έχουν συσχέτιση με την χειρονομία. Πρόσφατα, ο Yin και άλλοι [132] χρησιμοποίησαν ένα αθροιστικό HMM για τον εντοπισμό χειρονομιών σε συνεχή ροή δεδομένων, πετυχαίνοντας ενθαρρυντικά πειραματικά αποτελέσματα.

Οι μεταβατικές κινήσεις μεταξύ γειτονικών χειρονομιών αποτελούν άλλο ένα σχετικό ζήτημα στην αυτόματη αναγνώριση συνεχών χειρονομιών, ειδικά σε εφαρμογές όπως η αναγνώριση της νοηματικής γλώσσας. Ο Yang και άλλοι [133] ασχολήθηκαν με το ζήτημα της διαχείρισης της κίνησης επένθεσης με μια προσέγγιση βασισμένη στο δυναμικό προγραμματισμό. Ο Li και άλλοι [134] πρότειναν και σύγκριναν τρεις μεθόδους βασισμένες σε ένα μοντέλο χειρονομίας για τον εντοπισμό του τελικού σημείου της χειρονομίας. Οι μέθοδοι που εξετάστηκαν είναι η πολύ-κλιμακωτή αναζήτηση (multi-scale search), η δυναμική παραμόρφωση χρόνου (dynamic time wraping), και ο δυναμικός προγραμματισμός (dynamic programming). Η μέθοδος βασισμένη στο δυναμικό προγραμματισμό υπερέχει των άλλων δύο. Μια εμφωλευμένη, δομημένη μέθοδος βασισμένη στο δυναμικό προγραμματισμό προτείνεται από τον Sarkar και άλλους [135] για την αντιμετώπιση της αβεβαιότητας των ορίων στις προτάσεις των νοηματικών προτάσεων.

Η αντιστοίχιση μιας ακολουθίας εικόνων με ένα μοντέλο αποτελεί ένα κεντρικό ζήτημα στην Αναγνώριση Χειρονομιών του Χεριού (HGR). Ο Yang και άλλοι [136] πρότειναν έναν αλγόριθμο ελαχιστοποίησης για την αντιστοίχιση ομάδων από εικόνες, αρχικά με στατιστικά (HMM) καθώς και μη στατιστικά (βασισμένα σε δείγματα) μοντέλα. Ο αλγόριθμος δεν απαιτούσε ούτε τέλεια τμηματοποίηση της σκηνής ούτε την παρακολούθηση χαρακτηριστικών διαμέσου των πλαισίων.

Η πρόσφατη τάση της μάθησης με μία μόνο παρατήρηση (One-shot learning) [90, 97, 100, 101] στην αναγνώριση χειρονομιών είναι ελπιδοφόρα. Η μάθηση με μία μόνο παρατήρηση περιλαμβάνει την εκμάθηση μιας χειρονομίας παρατηρώντας μόνο ένα παράδειγμα αυτής της χειρονομίας, παρόμοια με την εκμάθηση που κάνει και ο άνθρωπος. Έχει δημιουργήσει την πρόκληση για να αντιμετωπιστεί η εξαγωγή διακριτικών χαρακτηριστικών καθώς και τον σχεδιασμό ταξινομητών χρησιμοποιώντας μόνο ένα παράδειγμα

εκπαίδευσης ανά κλάση. Επίσης, η μάθηση με μία μόνο παρατήρηση παρέχει κρίσιμη σύγκριση μεταξύ των αλγορίθμων αναγνώρισης χειρονομιών.

Οι χειρονομίες που χρησιμοποιούνται στα υπάρχοντα συστήματα αναγνώρισης χειρονομιών περιορίζονται σε ένα προσεκτικά επιλεγμένο λεξιλόγιο συμβολικών χειρονομιών (εμβλημάτων και επεξηγήσεων), τα οποία χρησιμοποιούνται κυρίως για την εκτέλεση εντολών. Η αναγνώριση χειρονομιών από ρυθμιστές, εκφράσεις αισθημάτων και προσαρμογείς (Κεφάλαιο 1) είναι απαραίτητη για τη φυσική αλληλεπίδραση μεταξύ ανθρώπων και μηχανών. Χρειάζονται αλγόριθμοι με καλύτερες δυνατότητες αμεταβλησίας, οι οποίοι να έχουν την δυνατότητα να αναγνωρίζουν ένα ευρύ αριθμό κλάσεων χωρίς εκτεταμένη εκπαίδευση, για να αναπτυχθούν μηχανές με την ικανότητα να κατανοούν καλύτερα τις προθέσεις και τα μοτίβα κίνησης των ανθρώπων.

Το αποτέλεσμα των ενσωματωμένων αλληλεπιδράσεων μέσω χειρονομιών στον εμπλουτισμό της οπτικής επεξεργασίας είναι ο τομέας που έχει μελετηθεί λιγότερο. Για παράδειγμα, η εξερεύνηση του πώς ένα κυματίζον χέρι προσελκύει την ανθρώπινη προσοχή θα είναι χρήσιμη για την ανάπτυξη του μηχανισμού της προσοχής ενός διαδραστικού ρομπότ. Μια άλλη μελλοντική κατεύθυνση έρευνας είναι η εξερεύνηση του εγκεφάλου για την ανάπτυξη υπολογιστικών μοντέλων τα οποία θα μιμούνται την διαδικασία αναγνώρισης των μοτίβων μιας ανθρώπινης χειρονομίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] K. Hogan, R. Stubbs, Can't get Through 8 Barriers to Communication, Pelican Publishing Company, Grenta, LA, 2003.

[2] D.K. Spencer, M.M. Sarah, R. Sabrina, Gesture gives a hand to language and learning: perspectives from cognitive neuroscience, developmental psychology and education, Lang. Linguist. Compass 2 (4) (2008) 569–588.

[3] A. Kendon, Current issues in the study of gesture, in: The Biological Foundation of Gestures: Motor and Semiotic Aspects, Psychology Press, 1986, pp. 23–47.

[4] L.L. Barker, L.A. Malandro, A.B. Deborah, Nonverbal Communication, 2nd ed., Addison-Wesley, MA, 1989.

[5] A. Kendon, Gesture and speech: how they interact, in: John M. Wiemann, Randall P. Harrison (Eds.), Nonverbal Interaction, Sage Publications, Beverly Hills, 1983.

[6] S. Berman, H. Stern, Sensors for gesture recognition systems, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 42 (3) (2012) 277–290.

[7] S. Mitra, T. Acharya, Gesture recognition : a survey, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 37 (3) (2007) 311–324.

[8] A. Erol, G. Bebis, M. Nicolescu, R.D. Boyle, X. Twombly, Vision-based hand pose estimation: a review, Comput. Vis. Image Underst. 108 (2007) 52–73.

[9] S.C.W. Ong, S. Ranganath, Automatic sign language analysis: a survey and the future beyond lexical meaning, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 873–891.

[10] K.H. Lee, J.H. Kim, An HMM based threshold model approach for gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell. 21 (10) (1999) 961–973.

[11] H.S. Yoon, J. Soh, Y.J. Bae, H.S. Yang, Hand gesture recognition using combined features of location, angle, and velocity, Pattern Recognit. 34 (2001) 1491–1501.

[12] A. Ramamoorthy, N. Vaswani, S. Chaudhury, S. Banerjee, Recognition of dynamic hand gestures, Pattern Recognit. 36 (2003) 2069–2081.

[13] A. Just, S. Marcel, A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition, Comput. Vis. Image Underst. 113 (4) (2009) 532–543.

[14] F.S. Chen, C.M. Fu, C.L. Huang, Hand gesture recognition using a real-time tracking method and hidden Markov models, Image Vis. Comput. 21 (2003) 745–758.
[15] S. Marcel, O. Bernier, J.E. Viallet, D. Collobert, Hand gesture recognition using input/output hidden Markov models, in: Proceedings of IEEE Automatic Face and Gesture Recognition, FG, 2000, pp. 456–461.
[16] N. Liu, B.C. Lovell, P.J. Kootsookos, Evaluation of HMM training algorithms for letter hand gesture recognition, in: Proceedings of the 3rd IEEE International Sym- posium on Signal Processing and Information Technology, Darmstadt, Germany, 2003, pp. 648–651.
[17] D. Kim, J. Song, D. Kim, Simultaneous gesture segmentation and recognition based on forward spotting accumulative HMMS, Pattern Recognit. 40 (11) (2007) 3012–3026.
[18] W.H.A. Wang, C.L. Tung, Dynamic hand gesture recognition using hierarchical dynamic Bayesian networks through low-level image processing, in: Proceed- ings of the 7th International Conference on Machine Learning and Cybernetics, Kunming, China, 2008, pp. 3247–3253.
[19] C.L. Huang, M.S. Wu, S.H. Jeng, Gesture recognition using the multi-PDM method and hidden Markov model, Image Vis. Comput. 18 (11) (2000) 865–879.
[20] J. Beh, D.K. Han, R. Durasiwami, H. Ko, Hidden Markov model on a unit hypersphere space for gesture trajectory recognition, Pattern Recognit. Lett. 36 (2014a) 144–153.
[21] J. Beh, D. Han, H. Ko, Rule-based trajectory segmentation for modeling hand motion trajectory, Pattern Recognit. 47 (4) (2014b) 1586–1601.
[22] J.H. Lee, T. Delbruck, et al., Real-time gesture interface based on event-driven processing from stereo silicon retinas, IEEE Trans. Neural Netw. Learning Syst. 25 (12) (2014) 2250–2263.
[23] S. Theodorakis, V. Pitsikalis, P. Maragos, Dynamic-static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition, Image Vis. Comput. 32 (8) (2014) 533–549.
[24] H.I. Suk, B.K. Sin, S.W. Lee, Hand gesture recognition based on dynamic Bayesian network framework, Pattern Recognit. 43 (9) (2010) 3059–3072.

[25] Q. Chen, N.D. Georganas, E.M. Petriu, Hand gesture recognition using Haar-like features and a stochastic context-free grammar, IEEE Trans. Instrum. Meas. 57 (8) (2008) 1562–1571.
[26] G. Caridakis, K. Karpouzis, A. Drosopoulos, S. Kollias, SOMM: self-organizing Markov map for gesture recognition, Pattern Recognit. Lett. 31 (1) (2010) 52–59.
[27] M. Abid, E. Petriu, E. Amjadian, Dynamic sign language recognition for smart home interactive application using stochastic linear formal grammar, IEEE Trans. Instrum. Meas. 64 (3) (2014) 596–605.
[28] W.W. Kong, S. Ranganath, Towards subject independent continuous sign language recognition: a segment and merge approach, Pattern Recognit. 47 (3) (2014) 1294–1308.
[29] M. Yeasin, S. Chaudhuri, Visual understanding of dynamic hand gestures, Pattern Recognit. 33 (11) (2000) 1805–1817.
[30] P. Hong, M. Turk, T.S. Huang, Gesture modeling and recognition using finite state machines, in: Proceedings of IEEE Automatic Face and Gesture Recognition, FG, 2000, pp. 410–415.
[31] J. Davis, M. Shah, Recognizing hand gestures, in: Proceedings of the European Conference on Computer Vision, 1994, pp. 331–340.
[32] M.H. Yang, N. Ahuja, M. Tabb, Extraction of 2D motion trajectories and its application to hand gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell 24 (8) (2002) 1061–1074.
[33] M.H. Yang, N. Ahuja, Extraction and classification of visual motion patterns for hand gesture recognition, in: Proceedings of IEEE Computer Vision and Pattern Recognition, CVPR, Santa Barbara, CA, USA, 1998, pp. 892–897.
[34] C.W. Ng, S. Ranganath, Real-time gesture recognition system and application, Image Vis. Comput. 20 (2002) 993–1007.
[35] X.H. Shen, G. Hua, L. Williams, Y. Wu, Dynamic hand gesture recognition: an exemplar-based approach from motion divergence fields, Image Vis. Comput. 30 (3) (2012) 227–235.
[36] J. Cheng, C. Xie, W. Bian, D.C. Tao, Feature fusion for 3d hand gesture recognition by learning a shared hidden space, Pattern Recognit. Lett. 33 (4) (2012) 476–484.

[37] K.S. Patwardhan, S.D. Roy, Hand gesture modelling and recognition involving changing shapes and trajectories, using a predictive eigentracker, Pattern Recog- nit. Lett. 28 (2007) 329–334.
[38] K. Daniel, M. John, M. Charles, A person independent system for recognition of hand postures used in sign language, Pattern Recognit. Lett. 31 (2010) 1359– 1368.
[39] M.C. Shin, L.V. Tsap, D.B. Goldgof, Gesture recognition using Bezier curves for visualization navigation from registered 3-d data, Pattern Recognit. 37 (5) (2004) 1011–1024.
[40] T. Kuremoto, Y. Kinoshita, L. Feng, S. Watanabe, K. Kobayashi, M. Obayashi, A gesture recognition system with retina-v1 model and one-pass dynamic program- ming, Neurocomputing 116 (2012) 291–300.
[41] A. Corradini, Dynamic time warping for off-line recognition of a small gesture vocabulary, in: Proceedings of IEEE International Workshop on Computer Vision, ICCVW, 2001, pp. 82–89.
[42] J.F. Lichtenauer, E.A. Hendriks, M.J.T. Reinders, Sign language recognition by combining statistical DTW and independent classification, IEEE Trans. Pattern Anal. Mach. Intell. 30 (11) (2008) 2040–2046.
[43] K. Tohyama, K. Fukushima, Neural network model for extracting optic flow, Neural Netw. 18 (5–6) (2005) 549–556.
[44] Cao, C., Zhang, Y., Wu, Y., Lu, H., & Cheng, J. (n.d.). "Egocentric Gesture Recognition Using Recurrent 3D Convolutional Neural Networks with Spatiotemporal Transformer Modules."
[45] Dānishgāh-i Yazd, & Institute of Electrical and Electronics Engineers. (n.d.). "ICEE 2019 : 27th Iranian Conference on Electrical Engineering : 30 April-2 May 2019, Yazd University, Yazd, Iran."
[46] Dhingra, N., & Kunz, A. (2019). "Res3ATN-Deep 3D Residual Attention Network for Hand Gesture Recognition in Videos." Proceedings - 2019 International Conference on 3D Vision, 3DV 2019, 491–501. https://doi.org/10.1109/3DV.2019.00061
[47] Duan, J., Zhou, S., Wan, J., Guo, X., & Li, S. Z. (2016). "Multi-Modality Fusion based on Consensus-Voting and 3D Convolution for Isolated Gesture Recognition." http://arxiv.org/abs/1611.06689

[48] Elboushaki, A., Hannane, R., Afdel, K., & Koutti, L. (2020). "MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences." Expert Systems with Applications, 139. https://doi.org/10.1016/j.eswa.2019.112829
[49] International Association for Pattern Recognition, & Mexican Association for Computer Vision, R. and N. C. (n.d.). "2016 23rd International Conference on Pattern Recognition (ICPR) : 4-8 Dec. 2016."
[50] Kurmanji, M., & Ghaderi, F. (2020). "Hand Gesture Recognition from RGB-D Data using 2D and 3D Convolutional Neural Networks: a comparative study." Journal of AI and Data Mining, 8(2), 177–188. https://doi.org/10.22044/JADM.2019.7903.1929
[51] Li, Y., Miao, Q., Tian, K., Fan, Y., Xu, X., Li, R., & Song, J. (2018). "Large-Scale Gesture Recognition with a Fusion of RGB-D Data Based on Saliency Theory and C3D Model." IEEE Transactions on Circuits and Systems for Video Technology, 28(10), 2956–2964. https://doi.org/10.1109/TCSVT.2017.2749509
[52] Li, Y., Miao, Q., Tian, K., Fan, Y., Xu, X., Ma, Z., & Song, J. (2019). "Large-scale gesture recognition with a fusion of RGB-D data based on optical flow and the C3D model." Pattern Recognition Letters, 119, 187–194. https://doi.org/10.1016/j.patrec.2017.12.003
[53] Lu, Z., Qin, S., Li, X., Li, L., & Zhang, D. (2019). "One-shot learning hand gesture recognition based on modified 3d convolutional neural networks." Machine Vision and Applications, 30(7–8), 1157–1180. https://doi.org/10.1007/s00138-019-01043-7
[54] Miao, Q., Li, Y., Ouyang, W., Ma, Z., Xu, X., Shi, W., & Cao, X. (n.d.). "Multimodal Gesture Recognition Based on the ResC3D Network."
[55] Molchanov Xiaodong Yang Shalini Gupta Kihwan Kim Stephen Tyree Jan Kautz NVIDIA, P. (n.d.). "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks." https://research.nvidia.com/publication/
[56] Oudah, M., Al-Naji, A., & Chahl, J. (2020). "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques." In Journal of Imaging (Vol. 6, Issue 8). MDPI. https://doi.org/10.3390/JIMAGING6080073
[57] Sarma, D., Kavyasree, V., & Bhuyan, M. K. (2020). "Two-stream Fusion Model for Dynamic Hand Gesture Recognition using 3D-CNN and 2D-CNN Optical Flow guided Motion Template." http://arxiv.org/abs/2007.08847

[58] Singh, D. K. (2021). "3D-CNN based Dynamic Gesture Recognition for Indian Sign Language Modeling." Procedia CIRP, 189, 76–83.
https://doi.org/10.1016/j.procs.2021.05.071
[59] Ur Rehman, M., Ahmed, F., Khan, M. A., Tariq, U., Alfouzan, F. A., Alzahrani, N. M., & Ahmad, J. (2022). "Dynamic hand gesture recognition using 3D-CNN and LSTM networks." Computers, Materials and Continua, 70(3), 4675–4690.
https://doi.org/10.32604/cmc.2022.019586
[60] Yu, B., Luo, Z., Wu, H., & Li, S. (2020). "Hand gesture recognition based on attentive feature fusion." Concurrency and Computation: Practice and Experience, 32(22).
https://doi.org/10.1002/cpe.5910
[61] Zhang, E., Xue, B., Cao, F., Duan, J., Lin, G., & Lei, Y. (2019). "Fusion of 2D CNN and 3D densenet for dynamic gesture recognition." Electronics (Switzerland), 8(12).
https://doi.org/10.3390/electronics8121511
[62] Zhang, Y., Wang, C., Zheng, Y., Zhao, J., Li, Y., & Xie, X. (2019). "Short-Term Temporal Convolutional Networks for Dynamic Hand Gesture Recognition."
http://arxiv.org/abs/2001.05833
[63] Zholshiyeva, L. Z., Zhukabayeva, T. K., Turaev, S., Berdiyeva, M. A., & Jambulova, D. T. (2021, October 11). "Hand Gesture Recognition Methods and Applications: A Literature Survey." ACM International Conference Proceeding Series.
https://doi.org/10.1145/3492547.3492578
[64] Zhu, G., Zhang, L., Shen, P., Song, J., Shah, S. A. A., & Bennamoun, M. (2019). "Continuous gesture segmentation and recognition using 3dcnn and convolutional lstm." IEEE Transactions on Multimedia, 21(4), 1011–1021.
https://doi.org/10.1109/TMM.2018.2869278
[65] Z. Zhang, Microsoft kinect sensor and its effect, IEEE MultiMed. 19 (2) (2012) 04–10.
[66] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: Proceedings of IEEE Computer Vision and Pattern Recognition, CVPR, Col- orado Springs, 2011.
[67] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with microsoft Kinect sensor: a review, IEEE Trans. Cybern. 43 (5) (2013) 1318–1334.

[68] R. Munoz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, A. Carmona-Poyato, Depth silhouettes for gesture recognition, Pattern Recognit. Lett. 29 (3) (2008) 319–329.
[69] F. Dominio, M. Donadeo, P. Zanuttigh, Combining multiple depth-based descriptors for hand gesture recognition, Pattern Recognit. Lett. 50 (2014) 101–111.
[70] M.R. Malgireddy, I. Inwogu, V. Govindaraju, A temporal Bayesian model for classifying, detecting and localizing activities in video sequences, in: Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, CVPRW, 2012, pp. 43–48.
[71] J. Sung, C. Ponce, B. Selman, A. Saxena, Human activity detection from RGBD images, in: Proceedings of AAAI workshop on Pattern, Activity and Intent Recogni- tion, PAIR, 2011.
[72] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from RGBD images, in: Proceedings of IEEE International Conference on Robotics and Automation, ICRA, 2012.
[73] P.K. Pisharady, M. Saerbeck, Kinect based body posture detection and recognition system, in: Proceedings of International Conference on Graphic and Image Processing (ICGIP), 2012.
[74] M.B. Holte, T. Moeslund, P. Fihl, Fusion of range and intensity information for view invariant gesture recognition, in: Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, CVPRW, 2008, pp. 1–7.
[75] L. Gallo, A.P. Placitell, M. Ciampi, Controller-free exploration of medical image data: experiencing the Kinect, in: Proceedings of International Symposium on Computer-Based Medical Systems, CBMS, 2011.
[76] W. Di, Z. Fan, S. Ling, One shot learning gesture recognition from RGBD images, in: Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, CVPRW, 2012.
[77] Y.M. Lui, A least squares regression framework on manifolds and its applica- tion to gesture recognition, in: Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, CVPRW, 2012, pp. 13–18.
[78] K. Lai, J. Konrad, P. Ishwar, A gesture-driven computer interface using Kinect, in: Proceedings of IEEE Southwest Symposium on Image Analysis and Interpreta- tion, SSIAI, 2012, pp. 185–188.

[79] D. Frolova, H. Stern, S. Berman, Most probable longest common subsequence for recognition of gesture character input, IEEE Trans. Cybern. 43 (3) (2013) 871– 880.
[80] H. Jiang, B.S. Duerstock, J.P. Wachs, A machine vision-based gestural interface for people with upper extremity physical impairments, IEEE Trans. Syst. Man Cybernet. Syst. 44 (5) (2014) 630–641.
[81] P.K. Pisharady, M. Saerbeck, A robust gesture detection and recognition algorithm for domestic robot interactions, in: Proceedings of International Conference on Control, Automation, Robotics and Vision, ICARCV, Singapore, 2014.
[82] R. Krishnan, S. Sarkar, Conditional distance based matching for one-shot gesture recognition, Pattern Recognit. 48 (4) (2015) 1302–1314.
[83] C. Zhang, T. Yingli, Histogram of 3d facets: a depth descriptor for human ac- tion and hand gesture recognition, Comput. Vis. Image Underst. (2015) in press, doi:10.1016/j.cviu.2015.05.010.
[84] Y.M. Lui, Human gesture recognition on product manifolds, J. Mach. Learn. Res. 13 (2012) 3297–3321.
[85] U. Mahbub, H. Imtiaz, T. Roy, M. Rahman, M. Ahad, A template matching approach of one-shot-learning gesture recognition, Pattern Recognit. Lett. 34 (2013) 1780–1788.
[86] J. Wan, Q. Ruan, W. Li, S. Deng, One-shot learning gesture recognition from RGB- D data using bag of features, J. Mach. Learn. Res. 14 (2013) 2549–2582.
[87] P.K. Pisharady, M. Saerbeck, Robust gesture detection and recognition using dynamic time warping and multi-class probability estimates, in: Proceedings of IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vi- sion Processing, CIMSIVP, 2013.
[88] H. Cheng, Z. Dai, Z. Liu, Image-to-class dynamic time warping for 3d hand gesture recognition, in: Proceedings of IEEE International Conference on Multime- dia and Expo, ICME, 2013.
[89] H. Cheng, J. Luo, X. Chen, A windowed dynamic time warping approach for 3d continuous hand gesture recognition, in: Proceedings of IEEE International Con- ference on Multimedia and Expo, ICME, 2014.
[90] E. Ohn-Bar, M. Trivedi, Hand gesture recognition in real time for automotive interfaces: a multimodal vision-based approach and evaluations, IEEE Trans. Intell. Transp. Syst. 15 (6) (2014) 2368–2377.

[91] M.G. Jacob, J.P. Wachs, Context-based hand gesture recognition for the operating room, Pattern Recognit. Lett. 36 (2014) 196–203.
[92] O. Mendels, H. Stern, S. Berman, User identification for home entertainment based on free-air hand motion signatures, IEEE Trans. Syst. Man Cybern.: Syst. 44 (11) (2014) 1461–1473.
[93] S.-Z. Li, B. Yu, W. Wu, S.-Z. Su, R.-R. Ji, Feature learning based on SAE-PCA network for human gesture recognition in RGBD images, Neurocomputing 151 (2015) 565–573.
[94] Y. Ming, Hand fine-motion recognition based on 3d mesh mosift feature descriptor, Neurocomputing 151 (2015) 574–582.
[95] R. Schramm, C.R. Jung, E.R. Miranda, Dynamic time warping for music conducting gestures evaluation, IEEE Trans. Multimed. 17 (2) (2015) 243–255.
[96] F.A. Kondori, S. Yousefi, J.-P. Kouma, L. Liu, H. Li, Direct hand pose estima- tion for immersive gestural interaction, Pattern Recognit. Lett. (2015) in press, doi:10.1016/j.patrec.2015.03.013.
[97] Z. Ren, J. Yuan, J. Meng, Z. Zhang, Robust part-based hand gesture recognition using Kinect sensor, IEEE Trans. Multimed. 15 (5) (2013) 1110–1120.
[98] R. Zhou, M. Jingjing, Y. Junsong, Depth camera based hand gesture recognition and its applications in human computer interaction, in: Proceedings of Interna- tional Conference on Information, Communications and Signal Processing, ICICS, 2011.
[99] Y. Li, Hand gesture recognition using kinect, in: Proceedings of the 3rd International Conference on Software Engineering and Service Science, ICSESS, 2012.
[100] P. Doliotis, V. Athitsos, D. Kosmopoulos, S. Perantonis, Hand shape and 3d pose estimation using depth data from a single cluttered frame, Adv. Vis. Comput. 7431 (2012) 148–158.
[101] F. Kirac, Y.E. Kara, L. Akarun, Hierarchically constrained 3d hand pose estimation using regression forests from single frame depth data, Pattern Recognit. Lett. 50 (2014) 91–100.
[102] Y. Yao, Y. Fu, Contour model based hand-gesture recognition using Kinect sensor, IEEE Trans. Circuits Syst. Video Technol. 24 (11) (2014) 1935–1944.
[103] F. Kirac, Y.E. Kara, L. Akarun, Hierarchically constrained 3d hand pose estimation using regression forests from single frame depth data, Pattern Recognit. Lett. 50 (2014) 91–100.

[104] C. Wang, Z. Liu, S.-C. Chan, Superpixel-based hand gesture recognition with Kinect depth camera, IEEE Trans. Multimed. 17 (1) (2015) 29–39.
[105] A. Bobick, J. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001).
[106] H. Liang, J. Yuan, D. Thalmann, Parsing the hand in depth images, IEEE Trans. Multimed. 16 (5) (2014) 1241–1253.
[107] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 509–522.
[108] F. Erden, A.E. Cetin, Hand gesture based remote control system using infrared sensors and a camera, IEEE Trans. Consum. Electron. 60 (4) (2014) 675–680.
[109] F. Simon, M.M. Helena, K. Pushmeet, N. Sebastian, Instructing people for train- ing gestural interactive systems, in: Proceedings of International Conference on Human Factors in Computing Systems, CHI, ACM, 2012, pp. 1737–1746.
[110] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, H. Escalante, Chalearn gesture challenge: design and first results, in: Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, CVPRW, 2012, pp. 1–6.
[111] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athistos, H.J. Escalante, Multi-modal gesture recognition challenge 2013: dataset and results, in: Proceedings of the 15th ACM International Conference on Multimodal Interac- tion, ICMI, Sydney, Australia, 2013.
[112] M. Chen, G. AlRegib, B.H. Juang, 6dmg: a new 6d motion gesture database, in: Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, CVPRW, 2011.
[113] A. Just, O. Bernier, S. Marcel, HMM and IOHMM for the recognition of mono- and bi-manual 3d hand gestures, in: Proceedings of British Machine Vision Confer- ence, BMVC, 2004.
[114] S. Yale, D. David, D. Randall, Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, FG, Santa Barbara, CA, 2011, pp. 500–506.
[115] S. Marcel, Hand posture recognition in a body-face centered space, in: Proceedings of the Conference on Human Factors in Computer Systems, CHI, 1999.

[116] S. Ruffieux, D. Lalanne, E. Mugellini, Chairgest: a challenge for multimodal mid-air gesture recognition for close HCI, in: Proceedings of 15th ACM on Interna- tional Conference on Multimodal Interaction, ICMI, 2013.
[117] L. Liu, L. Shao, Learning discriminative representations from RGB-D video data, in: Proceedings of International Joint Conference on Artificial Intelligence, IJCAI, 2013.
[118] J. Zhuolin, L.S. Davis, Recognizing actions by shape-motion prototype trees, in: Proceedings of IEEE International Conference on Computer Vision, ICCV, 2009, pp. 444–451.
[119] T.-K. Kim, S.-F. Wong, R. Cipolla, Tensor canonical correlation analysis for action classification, in: Proceedings of IEEE Computer Vision and Pattern Recognition, CVPR, 2007, pp. 1–8.
[120] Q. Chen, S. Xiao, W. Yichen, T. Xiaoou, S. Jian, Realtime and robust hand tracking from depth, in: Proceedings of IEEE Computer Vision and Pattern Recognition, CVPR, 2014.
[121] C. Neidle, A. Thangali, S. Sclaroff, Challenges in development of the american sign language lexicon video dataset corpus, in: Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC, 2012.
[122] R. Kehl, L. Van Gool, Real-time pointing gesture recognition for an immersive environment, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, FG, Seoul, Korea, 2004, pp. 577–582.
[123] N. Jojic, B. Brumitt, B. Meyers, S. Harris, T. Huang, Detection and estimation of pointing gestures in dense disparity maps, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, FG, Grenoble, France, 2000, pp. 1000–1007.
[124] C.B. Park, S.W. Lee, Real-time 3d pointing gesture recognition for mobile robots with cascade HMM and particle filter, Image Vis. Comput. 29 (1) (2011) 51–63.
[125] C.B. Park, M.C. Roh, S.W. Lee, Real-time 3D pointing gesture recognition in mobile space, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, FG, 2008.
[126] K. Nickel, R. Stiefelhagen, Visual recognition of pointing gestures for human– robot interaction, Image Vis. Comput. 25 (12) (2007) 1875–1884.

[127] J.L. Raheja, A. Chaudhary, S. Maheshwari, Hand gesture pointing location detection, Optik 125 (3) (2014) 993–996.
[128] M. Pateraki, H. Baltzakis, P. Trahanias, Visual estimation of pointed targets for robot guidance via fusion of face pose and hand orientation, Comput. Vis. Image Underst. 120 (2014) 1–13.
[129] M. Pateraki, H. Baltzakis, P. Trahanias, Visual estimation of pointed targets for robot guidance via fusion of face pose and hand orientation, Proceedings of IEEE International Conference on Computer Vision Workshops, ICCVW 2011, 2011.
[130] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 411–426.
[131] H. Kang, W.L. Chang, K.C. Jung, Recognition-based gesture spotting in video games, Pattern Recognit. Lett. 25 (15) (2004) 1701–1714.
[132] Y. Yin, R. Davis, Gesture spotting and recognition using salience detection and concatenated hidden Markov models, in: Proceedings of 15th ACM on Interna- tional conference on multimodal interaction, ICMI, 2013, pp. 489–494.
[133] R.D. Yang, S. Sarkar, B. Loeding, Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dy- namic programming, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2010) 462– 477.
[134] L. Hong, G. Michael, Model-based segmentation and recognition of dynamic gestures in continuous video streams, Pattern Recognit. 44 (8) (2011).
[135] S. Sarkar, B. Loeding, R. Yang, S. Nayak, A. Parashar, Segmentation-robust representations, matching, and modeling for sign language, in: Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, CVPRW, 2011, pp. 13–19.
[136] R.D. Yang, S. Sarkar, Coupled grouping and matching for sign and gesture recognition, Comput. Vis. Image Underst. 113 (6) (2009) 663–681.
[137] A. Betancourt, P. Morerio, E. Barakova, L. Marcenaro, M. Rauterberg, C. Regazzoni, A dynamic approach and a new dataset for hand-detection in first person vision, In International Conference on Computer Analysis of Images and Patterns (2015).
[138] 2020 25th International Conference on Pattern Recognition (ICPR). (n.d.). IEEE.

[139] Abavisani, M., Vaezi, H. R., Microsoft, J., & Patel, V. M. (n.d.). "Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition with Multimodal Training."
[140] Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., & Mekhtiche, M. A. (2020). "Hand Gesture Recognition for Sign Language Using 3DCNN." IEEE Access, 8, 79491–79509. https://doi.org/10.1109/ACCESS.2020.2990434
[141] Alsulaiman, M., Faisal, M., Mekhtiche, M., Bencherif, M., Alrayes, T., Muhammad, G., Mathkour, H., Abdul, W., Alohali, Y., Alqahtani, M., Al-Habib, H., Alhalafi, H., Algabri, M., Al-hammadi, M., Altaheri, H., & Alfakih, T. (2023). "Facilitating the communication with deaf people: Building a largest Saudi sign language dataset." Journal of King Saud University - Computer and Information Sciences, 35(8). https://doi.org/10.1016/j.jksuci.2023.101642
[142] de Smedt, Q., Wannous, H., Vandeborre, J.-P., Guerry, J., le Saux, B., Filliat, D., de Smedt, Q., Wannous, H., Vandeborre, J.-P., Guerry, J., le Saux, B., & Filliat, D. (2017). "SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset." 1–6. https://doi.org/10.2312/3dor.20171049
[143] Devineau, G., Moutarde, F., Xi, W., & Yang, J. (2018). "Deep learning for hand gesture recognition on skeletal data." Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, 106–113. https://doi.org/10.1109/FG.2018.00025
[144] European Association for Signal Processing, & Institute of Electrical and Electronics Engineers. (n.d.). "IPTA 2017 : proceedings of the Seventh International Conference on Image Processing Theory, Tools and Applications : Montreal, Canada, November 28 - December 1."
[145] Goos, G., Bertino, E., Gao, W., Steffen, B., & Yung, M. (n.d.). "Lecture Notes in Computer Science 13532 Founding Editors Editorial Board Members." https://link.springer.com/bookseries/558
[146] Köpüklü, O., Gunduz, A., Kose, N., & Rigoll, G. (2019). "Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks." http://arxiv.org/abs/1901.10323
[147] Kopuklu, O., Gunduz, A., Kose, N., & Rigoll, G. (2020). "Online Dynamic Hand Gesture Recognition including Efficiency Analysis." IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(2), 85–97. https://doi.org/10.1109/TBIOM.2020.2968216

[148] Liu, C., Yang, Y., Liu, X., Fang, L., & Kang, W. (2021). "Dynamic-Hand-Gesture Authentication Dataset and Benchmark." IEEE Transactions on Information Forensics and Security, 16, 1550–1562. https://doi.org/10.1109/TIFS.2020.3036218
[149] Lu, W., Tong, Z., & Chu, J. (2016). "Dynamic hand gesture recognition with leap motion controller." IEEE Signal Processing Letters, 23(9), 1188–1192. https://doi.org/10.1109/LSP.2016.2590470
[150] Materzynska, J., Berger, G., & Memisevic, R. (n.d.). "The Jester Dataset: A Large-Scale Video Dataset of Human Gestures."
[151] Oudah, M., Al-Naji, A., & Chahl, J. (2020). "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques." In Journal of Imaging (Vol. 6, Issue 8). MDPI. https://doi.org/10.3390/JIMAGING6080073
[152] Tran, D. S., Ho, N. H., Yang, H. J., Baek, E. T., Kim, S. H., & Lee, G. (2020). "Real-time hand gesture spotting and recognition using RGB-D Camera and 3D convolutional neural network." Applied Sciences (Switzerland), 10(2). https://doi.org/10.3390/app10020722
[153] Ur Rehman, M., Ahmed, F., Khan, M. A., Tariq, U., Alfouzan, F. A., Alzahrani, N. M., & Ahmad, J. (2022). "Dynamic hand gesture recognition using 3D-CNN and LSTM networks." Computers, Materials and Continua, 70(3), 4675–4690. https://doi.org/10.32604/cmc.2022.019586
[154] Wan, J., Li, S. Z., Zhao, Y., Zhou, S., Guyon, I., & Escalera, S. (n.d.). "ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition." http://www.cbsr.ia.ac.cn/users/jwan/database/isogd.html
[155] Zhang, Y., Cao, C., Cheng, J., & Lu, H. (2018). "EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition." IEEE Transactions on Multimedia, 20(5), 1038–1050. https://doi.org/10.1109/TMM.2018.2808769

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Ο Δημήτριος Έξαρχος γεννήθηκε στο Μόναχο Γερμανίας, το 1996. Τον Σεπτέμβρη του 2013 ξεκίνησε την φοίτησή του στο τμήμα Πληροφορικής του Ιονίου Πανεπιστημίου και το Φεβρουάριο του 2020. Το Φεβρουάριο του 2022 ξεκίνησε το μεταπτυχιακό του στο τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής στο Πανεπιστήμιο των Ιωαννίνων. Το ερευνητικό του ενδιαφέρον επικεντρώνεται στην υπολογιστική ό-ραση και στην μηχανική μάθηση.