Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης
Εργαστηριο Ορασης Υπολογιστων, Επικοινωνιας Λογου και Επεξεργασιας Σηματων
Ζωγραφου 157 73, Αθηνα

Πολυτροπική Κατανόηση Βίντεο με Τεχνικές Ασθενώς Επιβλεπόμενης Μάθησης

Διπλωματικη Εργασια

του

Γιώργου Μπουρίτσα

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εργαστηριο Ορασης Υπολογιστων, Επικοινωνιας Λογου και Επεξεργασιας Σηματων
Αθήνα, Ιούλιος 2017

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων
Ζωγράφου 157 73, Αθήνα

Πολυτροπική Κατανόηση Βίντεο με Τεχνικές Ασθενώς Επιβλεπόμενης Μάθησης

Διπλωματικη Εργασια

του

Γιώργου Μπουρίτσα

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18 Ιουλίου, 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Πέτρος Μαραγκός	Γεράσιμος Ποταμιάνος	Κωνσταντίνος Τζαφέστας
Καθηγητής	Αναπληρωτής Καθηγητής	Επίκουρος Καθηγητής
Ε.Μ.Π.	Πανεπιστήμιο Θεσσαλίας	ΕΜΠ

Αθήνα, Ιούλιος 2017

(Υπογραφή)

.....

Γιώργος Μπουρίτσας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων
Ζωγράφου 157 73, Αθήνα

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον καθηγητή κ. Πέτρο Μαραγκό για την ευκαιρία που μου έδωσε να εκπονήσω τη διπλωματική εργασία αυτή. Ακόμα, θέλω να τον ευχαριστήσω για τις συμβουλές, του όχι μόνο όσον αφορά την διπλωματική, αλλά και για τη μελλοντική μου επαγγελματική πορεία, καθώς και για την δυνατότητα που μου έδωσε να έρθω σε επαφή με την ερευνητική κοινότητα και να διευρύνω τους ορίζοντές μου.

Ακόμα, ευχαριστώ ιδιαίτερα τον αναπληρωτή καθηγητή κ. Josep Ramon Morros για την παραγωγική συνεργασία μας και για την δυνατότητα που μου έδωσε να κάνω τα πρώτα μου ερευνητικά βήματα στο πανεπιστήμιο UPC της Βαρκελώνης. Η εμπιστοσύνη που μου έδειξε και η ευκαιρία που μου παρείχε να συμμετάσχω σε ένα διαφορετικό ακαδημαϊκό περιβάλλον, έχουν ιδιαίτερη σημασία για εμένα.

Επίσης, θα ήθελα να εκφράσω τις ευχαριστίες μου στα μέλη του εργαστηρίου για τη βοήθεια που μου παρείχαν όλον αυτό τον καιρό. Ευχαριστώ ιδιαίτερα τον υποψήφιο διδάκτορα Πέτρο Κούτρα για την διαρκή και άμεση επικοινωνία που είχαμε, αλλά και για τις πάντα καίριας σημασίας συμβουλές του. Ακόμα, τους μεταδιδακτορικούς ερευνητές Νάνσυ Ζλατίντση και Βασσίλη Πιτσικάλη για τη βοήθεια να οργανωθεί όλη αυτή η δουλειά και να υπάρξει ένα ολοκληρωμένο και αξιόλογο αποτέλεσμα. Τέλος, το συμφοιτητή και φίλο Βαγγέλη Νικολουδάκη για την πραγματικά ευχάριστη συνεργασία που είχαμε.

Τελειώνοντας και επίσημα με αυτήν την εργασία αυτά τα 6 πολύ όμορφα φοιτητικά μου χρόνια, οφείλω να πω ένα μεγάλο ευχαριστώ στους φίλους μου για τις όλες τις αξέχαστες εμπειρίες που πήραμε, για τις ιδέες που ανταλλάξαμε, αλλά και για τη βοήθεια που μου προσέφεραν, ο καθένας με τον τρόπο του. Τέλος, προφανής είναι η ευγνωμοσύνη μου στους ανθρώπους της οικογένειάς μου που πάντα στηρίζουν τις αποφάσεις μου και δείχνουν κατανόηση.

Γιώργος Μπουρίτσας,
Ιούλιος 2017

Περίληψη

Στην παρούσα διπλωματική αντιμετωπίζουμε το πρόβλημα της αυτόματης κατανόησης βίντεο χρησιμοποιώντας κειμενικούς υπαινιγμούς ως μορφές ασθενούς επίβλεψης. Συγκεκριμένα, αν και υπάρχει μεγάλος όγκος βίντεο που συνοδεύονται από περιγραφικό κείμενο, δεν είναι πάντα εύκολο να αξιοποιηθεί η επίβλεψη που μας παρέχει, λόγω της χωροχρονικής ανακρίβειας των περιγραφών, αλλά και της δυσκολίας στην κατανόηση της σημασιολογίας τους.

Για κάθε κατηγορία οπτικών αντικειμένων υπό αναγνώριση, τα ερωτήματα που προκύπτουν είναι δύο: (i) Ποιο είναι το χωροχρονικό τμήμα του βίντεο στο οποίο αναφέρεται κάθε περιγραφή; (ii) Ποια είναι η ετικέτα που υπαινίσσεται κάθε περιγραφή; Απαντάμε στο πρώτο με Μάθηση Πολλαπλών Παραδειγμάτων και στο δεύτερο με Μάθηση Πιθανοτικών Ετικετών. Ακόμα, εισάγουμε την έννοια των Ασαφών Συνόλων Πολλαπλών Παραδειγμάτων για να μοντελοποιήσουμε τις διαφορετικές χρονικές επικαλύψεις μεταξύ των κειμενικών υπαινιγμών και των οπτικών αντικειμένων. Επίσης, εξερευνούμε τις δυνατότητες βελτίωσης της κατανόησης ενσωματώνοντας πληροφορία από άλλα υπό αναγνώριση οπτικά αντικείμενα και από τις προβλέψεις ενός προεκπαιδευμένου ταξινομητή. Τέλος, διατυπώνουμε μαθηματικά όλες αυτές τις μορφές ασθενούς επίβλεψης επεκτείνοντας έναν παλαιότερο φορμαλισμό διακριτικής ομαδοποίησης μέσω κυρτού προγραμματισμού.

Οι πτυχές του βίντεο που επιχειρούνται να κατανοηθούν είναι οι ανθρώπινοι χαρακτήρες και οι δράσεις που εκτελούν, αν και η μοντελοποίηση δεν περιορίζεται σε αυτές. Αφού εντοπιστούν τα αντικείμενα αυτά στο βίντεο, αναπαρίστανται μέσω χαρακτηριστικών βαθιάς μάθησης. Για να εξάγουμε τις ασθενείς ετικέτες από το κείμενο καθορίζουμε εκ των προτέρων ένα σταθερό σύνολο για κάθε μία από τις 2 περιπτώσεις και στη συνέχεια χρησιμοποιούμε ταίριασμα κανονικών εκφράσεων για τους χαρακτήρες και υπολογισμό σημασιολογικής ομοιότητας για τις δράσεις.

Αξιολογούμε τις μεθόδους μας, αφενός για την αναγνώριση προσώπου και για αφετέρου για την αναγνώριση δράσεων, σε ρεαλιστικά περιβάλλοντα και συγκεκριμένα σε 6 ταινίες της νεοεισαχθείσας στη διεθνή βιβλιογραφία βάσης COGNIMUSE, συνοδευόμενες από τα σενάρια και τους υπότιτλούς τους.

Λέξεις Κλειδιά

Αυτόματη Κατανόηση Βίντεο, Πολυτροπική Κατανόηση Γεγονότων, Ασθενώς Επιβλεπόμενη Μάθηση, Μάθηση Πολλαπλών Παραδειγμάτων, Ασαφή Σύνολα, Πιθανοτικές Ετικέτες, Διακριτική Ομαδοποίηση, Κυρτός Προγραμματισμός, Σημασιολογία Κειμένου, Σημασιολογική Ομοιότητα, Αναγνώριση Προσώπου, Αναγνώριση Δράσεων

Abstract

In this thesis we address the problem of automatic video understanding using textual cues as forms of weak supervision. Specifically, despite the fact that a huge amount of video data accompanied by a descriptive text are available, it is not always easy to exploit the supervision the text provides. The reason is the spatio-temporal imprecision of the descriptions, as well as the adversity to understand their semantics

The questions that are raised for each category of visual objects are the following: (i) To which spatio-temporal video region does each textual description refer? (ii) Which label is implied by each textual description? We address the former as a Multiple Instance Learning problem and the latter as a Probabilistic Label Learning one. We also introduce the concept of Fuzzy Multiple Instance Sets to model the variations in the temporal overlap between the textual cues and the visual objects. In addition, we explore the capabilities of improvement of the understanding procedure incorporating information created by the recognition of other categories of visual objects, as well as the prediction of a pre-trained classifier. All this forms of weak supervision are formulated using a discriminative clustering framework which is optimized with a convex relaxation.

The video content that we wish to retrieve comprises the human characters and the actions they perform. After detecting the objects in the video sequence, we represent them in a feature space using deep learning architectures. To extract the weak labels from the text we define the label set beforehand and then we apply either regular expression matching (concerning the characters) or semantic similarity calculation (concerning the actions).

We validate our methods, with respect to the characters and the actions, in the challenging and realistic setting of 6 movies of the newly introduced database COGNIMUSE, accompanied by their scripts and subtitles.

Keywords

Automatic Video Understanding, Multimodal Event Understanding, Weakly Supervised Learning, Multiple Instance Learning, Fuzzy Sets, Probabilistic Labels, Discriminative Clustering, Convex Programming, Text Semantics, Semantic Similarity, Face Recognition, Action Recognition

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Η Αυτόματη Κατανόηση Βίντεο

Στη σύγχρονη εποχή των μεγάλων δεδομένων και του διαδικτύου η τεχνολογία καλείται να ικανοποιήσει τις ολοένα και πιο σύνθετες ανάγκες των ανθρώπων για πρόσβαση σε πληροφορίες με γρήγορο και αυτοματοποιημένο τρόπο και ταυτόχρονα να διευκολύνει την ανθρώπινη καθημερινότητα. Μία από τους πιο πλούσιες, αλλά και χαώδης ταυτόχρονα σε όγκο πηγή γνώσης είναι η εικόνα και το βίντεο. Αν σκεφτεί κανείς το πλήθος των εικόνων και βίντεο που παράγονται καθημερινά μέσω του διαδικτύου (στο Facebook ανεβαίνουν καθημερινά 300 εκατομμύρια φωτογραφίες, ενώ στο Youtube κάθε λεπτό ανεβαίνουν 300 ώρες βίντεο) είναι εύκολο να αντιληφθούμε ότι η πληροφορία πλέον δεν είναι διαχειρίσιμη από έναν άνθρωπο. Οι σύγχρονες ανάγκες που προκύπτουν σχετικά με τις εικόνες και τα βίντεο εκτείνονται από τις πιο απλές και καθημερινές, όπως είναι η αναζήτηση μίας φωτογραφίας στο διαδίκτυο, μέχρι τις πιο σύνθετες, όπως είναι η σχεδίαση μίας υψηλού επιπέδου τεχνητής όρασης στα πρότυπα της ανθρώπινης. **Η Αυτόματη Κατανόηση Βίντεο (Automatic Video Understanding)** είναι ένα υπο-πεδίο της Μηχανικής Όρασης που δίνει λύσεις σε πολλά από αυτά τα προβλήματα. Συγκεκριμένα, η ανάλυση και η κατανόηση του περιεχομένου ενός βίντεο (στην βιβλιογραφία μπορεί να αναφέρεται και ως Video Content Analysis - VCA ή εντοπισμός γεγονότων - Event Detection) είναι η διαδικασία με την οποία από κάθε βίντεο εξάγονται υψηλού επιπέδου σημασιολογικές έννοιες που περιγράφουν το περιεχόμενό του, ενώ ταυτόχρονα εντοπίζονται χωροχρονικά τα οπτικά γεγονότα που καθορίζουν το περιεχόμενο αυτό. Ο τομέας αυτός περιέχει ένα μεγάλο πλήθος υποπροβλημάτων όπως είναι η αναγνώριση ανθρώπινων συμπεριφορών (π.χ χειρονομίες, δράσεις), αντικειμένων, προσώπων κ.ά, ο εντοπισμός των θεματικών ενοτήτων και η αναγνώριση των σκηνών που περιέχει το βίντεο, η ανάλυση συναισθήματος, η εξαγωγή των σημαντικών σημείων (salient) κ.ά. Επιλύοντας κάθε ένα από αυτά μπορεί να εξαχθεί μία περιγραφή του βίντεο, όμοια με αυτή που θα έκανε ένας άνθρωπος βλέποντας το. Χαρακτηριστικά παραδείγματα εφαρμογών της αυτόματης κατανόησης βίντεο είναι τα εξής:

- **Video Indexing and Retrieval**: Εντοπίζοντας τα γεγονότα μέσα σε ένα βίντεο, είναι δυνατόν να χωριστεί σε τμήματα τα οποία θα δεικτοδοτούνται από το περιεχόμενό

τους (Video indexing). Έτσι, καθίσταται δυνατό για ένα χρήστη να περιηγηθεί σε ένα βίντεο απλά αναζητώντας αυτά που είναι σημαντικά για αυτόν (Video Browsing). Γενικότερα, με αυτόν τον τρόπο καθίσταται δυνατή η ανάκτηση βίντεο (Video Retrieval) από βάσεις δεδομένων ή και από το σύνολο του διαδικτύου που περιέχουν γεγονότα που ανταποκρίνονται στο αίτημα ενός χρήστη. Αυτό σήμερα υλοποιείται κυρίως μέσα από ετικέτες (tags) που οι διαχειριστές των δεδομένων έχουν χειροκίνητα αποδώσει. Όμως, αυτό καλύπτει τις ανάγκες αναζητήσεων κυρίως σε εικόνες, ενώ στα βίντεο οι ετικέτες δεν συνοδεύουν κάθε κομμάτι του βίντεο ξεχωριστά αλλά το σύνολο του. Άλλωστε είναι ιδιαίτερα χρονοβόρα διαδικασία για έναν άνθρωπο να επισημειώσει χωροχρονικά ένα βίντεο με το σύνολο του περιεχομένου του. Έτσι, στο YouTube για παράδειγμα, ένας χρήστης ακόμα και να καταφέρει να εντοπίσει το βίντεο που τον ενδιαφέρει χρειάζεται να αναζητήσει μόνος του, σε όλη τη διάρκεια του, το κομμάτι του βίντεο που περιέχει την πληροφορία που ζήτησε. Γίνεται επομένως αντιληπτό ότι η αυτοματοποίηση και η επιτάχυνση τέτοιων διαδικασιών είναι αναγκαία.

- **Video Summarization** : Εδώ, και πάλι στα πλαίσια της σύγχρονης ανάγκης για επιτάχυνση διαδικασιών, εντάσσεται η εξαγωγή της περίληψης ενός βίντεο. Ως αντίστοιχο παράδειγμα μπορούμε να σκεφτούμε το trailer μίας ταινίας. Συχνά, ο θεατής προκειμένου να αποφασίσει αν θα παρακολουθήσει την ταινία ή όχι χρειάζεται να αποκτήσει μία γενική εικόνα για το περιεχόμενό της. Αυτό επιτυγχάνεται με τα trailer. Ακόμα, είναι συνηθισμένο οι χρήστες να θέλουν να δουν μόνο τα σημαντικά σημεία από ένα βίντεο προκειμένου να πάρουν την πληροφορία που χρειάζονται. Εκτός όμως από τις ταινίες, τα υπόλοιπα βίντεο δεν διαθέτουν τέτοιου είδους περίληψη. Έτσι, η αυτόματη κατανόηση ενός βίντεο είναι απαραίτητη προκειμένου να προσφέρονται στους χρήστες μόνο τα σημεία μεγάλης σημασίας.

- **Human - Machine Interaction**: Ίσως η πιο σημαντική από τις εφαρμογές είναι αυτή της αλληλεπίδρασης συστημάτων που διαθέτουν υπολογιστική όραση, με τον άνθρωπο. Συγκεκριμένα είναι απαραίτητο για ένα ρομπότ, ή γενικότερα για ένα σύστημα τεχνητής νοημοσύνης να αντιλαμβάνεται το περιβάλλον του και να εξάγει συμπεράσματα με βάση αυτό. Για παράδειγμα ένα σύστημα που παρέχει κάποια βοήθεια σε έναν άνθρωπο πρέπει να είναι σε θέση να αναγνωρίζει τις κινήσεις του.

Η κατανόηση του περιεχομένου του βίντεο, έχει εξεταστεί σε μεγάλο βαθμό χωρίς την παρουσία των άλλων 2 τροπικοτήτων που συνήθως το συνοδεύουν, δηλαδή του ήχου και του κειμένου. Ένα μεγάλο κομμάτι της έρευνας στην Όραση Υπολογιστών είναι η αναγνώριση διαφόρων στοιχείων μέσα σε ένα βίντεο χρησιμοποιώντας μεθόδους Επιβλεπόμενης Μάθησης που εκπαιδεύουν μοντέλα με τη βοήθεια άλλων βίντεο που είναι επισημειωμένα ως προς το οπτικό περιεχόμενο τους. Η επισημείωση όμως είναι δύσκολη και χρονοβόρα διαδικασία. Επίσης, η πλειοψηφία των βίντεο που συναντάμε σε ρεαλιστικά και όχι τεχνητά περιβάλλοντα περιέχει ένα μέρος της πληροφορίας της και στις άλλες τροπικότητες. Άλλωστε και στον ίδιο τον άνθρωπος οι αισθήσεις του λειτουργούν συνεργατικά προκειμένου να συνθέσουν την αντίληψη του. Έτσι, αξιοποιώντας τις άλλες τροπικότητες είτε υποβοηθητικά, είτε ισότιμα με

το βίντεο (όπου εντοπίζουμε τα γεγονότα και σε αυτές), μπορούμε να επιτύχουμε κατανόηση με τη χρήση λιγότερης πρότερης γνώσης και πιθανόν και πληρέστερη.

Στη συγκεκριμένη εργασία η πολυτροπική επεξεργασία γίνεται με τρόπο που η μία τροπικότητα βοηθά στην κατανόηση της άλλης και δεν είναι ισότιμες. Μέσω της βοήθειας από την δευτερεύουσα τροπικότητα, καταφέρνουμε πέρα από την κατανόηση της πρωτεύουσας, να συλλέξουμε δεδομένα σε μεγάλη κλίμακα. Με αυτά μπορούμε στη συνέχεια να εκπαιδεύσουμε συστήματα με σκοπό την εφαρμογή τους σε δεδομένα όπου δεν θα υπάρχει τέτοιου είδους βοήθεια. Στην εργασία αυτή η τροπικότητα που λειτουργεί υποβοηθητικά είναι το κείμενο, ενώ αυτή που υποβάλλεται σε διαδικασίες αναγνώρισης είναι το βίντεο. Το βασικό κίνητρο για αυτό είναι ότι ο πυρήνας της σημασιολογίας είναι η γλώσσα. Με την ίδια λογική που ο ανθρώπινος εγκέφαλος περιγράφει αυτά που βλέπει με λέξεις, το ίδιο πρέπει να μπορεί να κάνει και μια τεχνητή νοημοσύνη. Βέβαια, ο φορμαλισμός που σχεδιάσαμε δεν υπονοεί κάπου ποιες τροπικότητες συμμετέχουν, πράγμα που σημαίνει ότι μπορεί να χρησιμοποιηθεί για οποιοδήποτε ζεύγος τροπικοτήτων, αρκεί η μία να περιέχει πληροφορία που να περιγράφει σε ένα βαθμό την άλλη. Θεωρούμε ότι η πληροφορία παρέχεται υπό μορφή υπαινιγμών (cues), ή αλλιώς, αμφισημιών (ambiguities). Στην περίπτωση μας, το κείμενο περιγράφει μεν το βίντεο, αλλά χωρίς να ορίζονται επακριβώς τα χωροχρονικά όρια αυτού που περιγράφεται κάθε φορά, ενώ επίσης δεν είναι προφανές το τι είναι αυτό που περιγράφεται. Τέτοιου είδους δεδομένα υπάρχουν σε αφθονία, όπως είναι για παράδειγμα τα βίντεο στο YouTube ή και γενικότερα στο διαδίκτυο (ιστοσελίδες, social media, κ.ά) που στην τεράστια πλειοψηφία τους συνοδεύονται από γλωσσική περιγραφή, εκπομπές στην τηλεόραση όπου γίνεται χρήση λεζάντων (captions), βιντεοσκοπημένες συνεδρίες από τη Βουλή ή από δίκες που συνοδεύονται από τα πρακτικά τους και φυσικά κινηματογραφικές ταινίες ή τηλεοπτικές σειρές που συνοδεύονται από τους υπότιτλους και τα σενάρια τους (ή από λεπτομερείς περιγραφές για ανθρώπους με προβλήματα όρασης - Audio Described Movies). Η τελευταία κατηγορία είναι και αυτή που χρησιμοποιούμε στην παρούσα διπλωματική, λόγω της μεγαλής ποικιλίας των ταινιών σε γεγονότα και ταυτόχρονα λόγω της ρεαλιστικής φύσης τους.

1.2 Η Ασθενώς Επιβλεπόμενη Μάθηση

Η **Ασθενής Επίβλεψη - Weak Supervision** μπορεί να παρέχεται με πολλές διαφορετικές μορφές κοινό χαρακτηριστικό των οποίων είναι ότι δεν παρέχεται σε κάθε δεδομένο μία ετικέτα αλλά αυτή υπαινίσσεται (αναλυτικά στο κεφάλαιο 3) . Στόχος της Ασθενώς Επιβλεπόμενης Μάθησης είναι να βρεθεί για κάθε δεδομένο ακριβώς αυτή η υπαινισσόμενη ετικέτα και ταυτόχρονα να εκπαιδευτεί ένα σύστημα που να μπορεί να εφαρμοστεί σε δεδομένα όπου η ετικέτα τους είναι εντελώς άγνωστη. Η Μάθηση αυτή έχει αρχίσει να λαμβάνει ιδιαίτερη δημοφιλία τα τελευταία χρόνια. Ο πρώτος βασικός λόγος είναι ότι είναι ιδιαίτερα συνηθισμένο σε ρεαλιστικά προβλήματα μάθησης να παρέχονται επιπλέον πληροφορίες στα δεδομένα προς αναγνώριση (οι οποίες μπορούν να γίνουν αντιληπτές ως μορφές ασθενούς επίβλεψης). Ο δεύτερος είναι ότι μέσω αυτών των μεθόδων αντιμετωπίζεται σε μεγάλο βαθμό ένα από τα βασικά μειονεκτήματα των συστημάτων μάθησης, το κόστος της συλλογής επισημειωμένων

δεδομένων. Συγκεκριμένα, η μεγάλη πλειοψηφία των μοντέλων που χρησιμοποιούνται, προκειμένου να εκπαιδευτούν έχουν ανάγκη από το λεγόμενο σύνολο εκπαίδευσης (training set). Αυτό αποτελείται από ζεύγη δεδομένων - ετικετών, όπου η ετικέτα εκφράζει την κατηγορία στην οποία ανήκει το δεδομένο. Η προσπάθεια μείωσης ή ακόμα και εξάλειψης της ύπαρξης αυτού του συνόλου και η αντικατάσταση του από λιγότερο πληροφοριακά αλλά ευκολότερο να αποκτηθούν δεδομένα είναι ένα από τα βασικά κίνητρα της εν λόγω διπλωματικής εργασίας.

Ένα χαρακτηριστικό παράδειγμα ασθενούς επίβλεψης είναι το πρόβλημα όπου ένα δεδομένο μπορεί να συνοδεύεται από περισσότερες από μία ετικέτες εκ των οποίων μόνο μία είναι σωστή ([10]), ή από μία κατανομή πιθανότητας πάνω στις ετικέτες που εκφράζει την πιθανότητα κάθε ετικέτας να είναι η σωστή ([28]). Ακόμα, οι ετικέτες μπορεί να αποδίδονται σε υποσύνολα των δεδομένων και όχι σε μεμονωμένα δείγματα, υπονοώντας ότι σε ένα τουλάχιστον δείγμα του κάθε υποσυνόλου πρέπει να αποδοθεί η αντίστοιχη ετικέτα ([16]). Άλλη μία περίπτωση είναι να έχει εκπαιδευτεί ήδη ένα σύστημα πρόβλεψης και να εφαρμόζεται σε δεδομένα που συνοδεύονται από υπαινισσόμενες ετικέτες ([34]).

Η Ασθενώς Επιβλεπόμενη Μάθηση έχει ήδη αρχίσει να χρησιμοποιείται εκτενώς στην Όραση Υπολογιστών. Για παράδειγμα, έχουν προταθεί μέθοδοι εντοπισμού αντικειμένων οι οποίες έχουν εκπαιδευτεί από εικόνες επισημειωμένες μόνο με την ετικέτα αλλά όχι την ακριβή θέση του αντικειμένου ([8]), ή μέθοδοι αναγνώρισης και εντοπισμού δράσεων από αποσπάσματα βίντεο που τις περιέχουν χωρίς να είναι γνωστά τα χρονικά τους όρια ([63]).

Όπως είναι φανερό, το πρόβλημα της εν λόγω διπλωματικής κατέχει τα γνωρίσματα της Ασθενούς Επίβλεψης, καθώς αυτή παρέχεται σε ποικίλες μορφές στο συνοδευτικό κείμενο. Άλλωστε, υπάρχουν ήδη κάποιες ερευνητικές εργασίες που χρησιμοποιούν τέτοιες μοντελοποιήσεις όπως είναι τα [4, 5, 10] κ.ά. Έτσι, και εμείς αντιμετωπίζουμε το πρόβλημα της πολυτροπικής κατανόησης βίντεο υπό αυτό το πρίσμα. Τέλος, η προσοχή μας εστιάζεται κυρίως στον τρόπο με τον οποίο θα αξιοποιηθεί η βοηθητική πληροφορία και όχι στη φύση της και για αυτό το λόγο δίνουμε μεγαλύτερη βαρύτητα στις διάφορες μορφές μάθησης και όχι στους εναλλακτικούς τρόπους κατανόησης του βίντεο.

1.3 Περιγραφή του Προβλήματος

Στο πρόβλημα της εργασίας μας έχουμε στη διάθεση μας ένα βίντεο και ένα συνοδευτικό κείμενο. Κάθε τμήμα του κειμένου διαθέτει χρονικά όρια που δείχνουν ποιο τμήμα του βίντεο περιγράφει. Τα τμήματα αυτά είναι αρκετά μεγάλα, έτσι ώστε να μην υπάρχει μοναδική γλωσσική περιγραφή για κάθε οπτικό αντικείμενο. Ακόμα, οι γλωσσικές περιγραφές αυτές είναι ρεαλιστικές (δεν έχουν προκύψει για παράδειγμα από χειροκίνητη επισημείωση ή μεταδεδομένα). Αυτό σημαίνει ότι οι ετικέτες των υπό αναγνώριση αντικειμένων υπαινίσσονται από τις περιγραφές και δεν αναφέρονται ρητά. Στόχος είναι να εντοπιστούν χωροχρονικά και να αναγνωριστούν συγκεκριμένα αντικείμενα μέσα στο βίντεο αξιοποιώντας όσο το δυνατόν καλύτερα τη γλωσσική πληροφορία.

Εδώ, εξετάζουμε δύο υποκατηγορίες του προβλήματος της Αυτόματης Κατανόησης Βίντεο: Την αναγνώριση Χαρακτήρων - Προσώπων και την αναγνώριση Οπτικών Γεγονότων -

Ανθρώπινων Δράσεων. Τα δύο αυτά στοιχεία συνιστούν ένα πολύ μεγάλο κομμάτι του περιεχομένου του βίντεο και αποτελούν σημεία μεγάλου ενδιαφέροντος. Για παράδειγμα, σε ένα περιβάλλον ανάκτησης βίντεο είναι πολύ συχνή η αναζήτηση με βάση το όνομα ενός ανθρώπου. Ακόμα, σε ένα περιβάλλον αλληλεπίδρασης ανθρώπου - μηχανής η τεχνητή νοημοσύνη χρειάζεται να παρακολουθεί τις δράσεις που εκτελεί ο άνθρωπος προκειμένου για παράδειγμα να του παρέχει βοήθεια, να τον κατευθύνει ή και να δεχθεί οδηγίες από αυτόν. Βέβαια, η μοντελοποίηση δεν περιορίζεται σε αυτά. Για την ακρίβεια όπως προαναφέραμε, δεν περιορίζεται καν στο πρόβλημα της μάθησης βίντεο από κείμενο, καθώς οι 2 τροπικότητες γίνονται αντιληπτές σαν 2 παράλληλες ροές αντικειμένων (πληροφορίας) - βλέπε σχήμα 4.4 - με την ίδια σημασιολογία (δηλαδή περιγράφουν τα ίδια πράγματα) όπου η φύση τους δεν παίζει κάποιο ρόλο.

1.4 Στόχοι και Συνεισφορές της Διπλωματικής Εργασίας

Βασικός στόχος της εργασίας είναι να εξεταστεί η δυνατότητα κατανόησης οπτικών αντικειμένων χρησιμοποιώντας μόνο ασθενή επίβλεψη από το συνοδευτικό κείμενο και καμία πρότερη γνώση. Παράλληλα με αυτόν, και ορμώμενοι από την ενδιαφέρουσα φύση του προβλήματος, ερευνούμε διαφορετικά σενάρια μάθησης που μπορούν να εφαρμοστούν σε ποικίλα - εντελώς διαφορετικά - γενικότερα πλαίσια.

Οι συνεισφορές της εργασίας μπορούν να συνοψιστούν στα εξής:

- Παρουσιάζουμε ένα νέο τρόπο εξαγωγής των ασθενών ετικετών (weak labels) από το κείμενο βασισμένο στη σημασιολογική ομοιότητα των λέξεων. Οι μέχρι τώρα μέθοδοι ήταν αρκετά περιοριστικές και βασισμένες στην εκάστοτε εφαρμογή ή/και απαιτούσαν την εκπαίδευση ενός ταξινομητή, άρα και τη συλλογή δεδομένων, ο οποίος ταξινομούσε τα μέρη του κειμένου σε ετικέτες κλάσεων. Η δικιά μας μέθοδος είναι γενική και μπορεί να εφαρμοστεί σε οποιοδήποτε πρόβλημα αναγνώρισης, αρκεί φυσικά οι γλωσσικές περιγραφές των αντικειμένων να περιέχονται στο κείμενο, ενώ δεν απαιτεί την εκπαίδευση κάποιου επιπλέον συστήματος, άρα δεν χρειάζεται κανένα είδος επίβλεψης.

- Επεκτείνουμε την μοντελοποίηση του [4] σε νέα σενάρια ασθενούς επίβλεψης. Συγκεκριμένα, στο [4] γίνεται μία θεώρηση της επίβλεψης ως Μάθηση Πολλαπλών Παραδειγμάτων (Multiple Instance Learning) ενώ στο [10] ως Μάθηση Υποψήφιων Ετικετών (Candidate Labels). Εμείς, εισάγουμε ένα κυρτό φορμαλισμό (convex formulation) Μάθησης Πολλαπλών Παραδειγμάτων με Πιθανοτικές Ετικέτες (Multiple Instance Probabilistic Learning) καθώς και μία επέκταση των Απλών Συνόλων Πολλαπλών Παραδειγμάτων σε Ασαφή Σύνολα. Σύμφωνα με όσα γνωρίζουμε, δεν έχουν υπάρξει αντίστοιχες μοντελοποιήσεις ούτε για το εν λόγω πρόβλημα ούτε για κάποιο παρόμοιο.

- Ενσωματώνουμε την υπαινισσόμενη πληροφορία από τις επαναλήψεις κάθε συνόλου παραδειγμάτων με την ίδια ετικέτα ρυθμίζοντας την αβεβαιότητα της.

- Προτείνουμε ένα γενικότερο τρόπο ροής πληροφορίας μεταξύ των διαφόρων ειδών υπό αναγνώριση αντικειμένων (εδώ πρόσωπα και δράσεις) από αυτόν του [4]. Συγκεκριμένα, στο [4], περιγράφεται μόνο η περίπτωση όπου η αντιστοίχηση μεταξύ δύο διαφορετικών ειδών αντικειμένων είναι '1-1' (πρόσωπο και σώμα ενός ανθρώπου). Εμείς αίρουμε αυτόν τον περιορισμό και επιτρέπουμε τη ροή πληροφορίας ακόμα και όταν δεν ξέρουμε την ακριβή αντιστοίχηση αλλά μπορούμε να κάνουμε μόνο μία εκτίμηση.

- Προτείνουμε μία μέθοδο σύμπτυξης ενός ασθενώς επιβλεπόμενου ταξινομήτη με έναν προ-εκπαιδευμένο προκειμένου να εισάγουμε πρότερη γνώση στο μοντέλο.

- Ενσωματώνουμε στην τροπικότητα του βίντεο αναπαραστάσεις Βαθιάς Μηχανικής Μάθησης.

- Εφαρμόζουμε τους αλγορίθμους αυτούς στη βάση δεδομένων COGNIMUSE, όπως αυτή παρουσιάζεται στο [70].

Ακόμα, προτείνουμε μία μέθοδο πολυτροπικής κατανόησης βίντεο και κειμένου, όπου οι 2 τροπικότητες είναι ισότιμες και μαθαίνουν η μία από την άλλη. Η μέθοδος αυτή δεν έχει υλοποιηθεί λόγω του περιορισμένου χρόνου εκπόνησης της εργασίας αλλά παρουσιάζεται εδώ (στο κεφάλαιο 4) προκειμένου να υλοποιηθεί σε μελλοντική εργασία.

1.5 Διάρθρωση του Τόμου

Στο **κεφάλαιο 2** παρουσιάζουμε τις σχετικές εργασίες που έχουν δημοσιευθεί στη διεθνή βιβλιογραφία από τη σκοπιά της Κατανόησης Βίντεο από Κείμενο.

Στο **κεφάλαιο 3** κάνουμε μία σύντομη ανασκόπηση των μεθόδων μάθησης παραθέτοντας παραδείγματα από εφαρμογές στην Όραση Υπολογιστών, προκειμένου να αναδείξουμε τις διαφορές, τα πλεονεκτήματα και τα μειονεκτήματα, καθώς και τις αναγκαιότητες κάθε μίας από αυτές. Ιδιαίτερη έμφαση δίνεται στις υποκατηγορίες της Ασθενούς Επίβλεψης.

Στο **κεφάλαιο 4** περιγράφουμε το φορμαλισμό του προβλήματος που υιοθετήσαμε από παλαιότερες εργασίες και ιδιαίτερα από το [4], καθώς και τους τρόπους με τον οποίο τον επεκτείναμε και τον γενικεύσαμε.

Στο **κεφάλαιο 5** αναλύουμε τα ολοκληρωμένα συστήματα με τα οποία εντοπίζονται χωροχρονικά και αναπαρίστανται τα οπτικά αντικείμενα περιγράφοντας εν συντομία το κάθε υποσύστημα. Τα συστήματα αυτά είναι τυποποιημένα και εδώ περιοριζόμαστε στην προσαρμογή τους στα δικά μας δεδομένα.

Στο **κεφάλαιο 6** περιγράφουμε τον τρόπο με τον οποίο αντλήσαμε την πληροφορία από το κείμενο. Συγκεκριμένα, δείχνουμε τους πιθανούς τρόπους με τους οποίους μπορεί κανείς να επιλέξει το σύνολο των ετικετών, σύμφωνα και με παλαιότερες εργασίες, και αυτούς με τους οποίους μπορεί να κάνει την εξαγωγή τους από το κείμενο. Ακόμα, δείχνουμε στο εν λόγω πρόβλημα τον αλγόριθμο που αξιοποιήθηκε για την απόδοση χρονικών σημάνσεων (timestamps) στο κείμενο. Τέλος, παρουσιάζουμε τις δικές μας μεθόδους εξόρυξης πληροφορίας.

Να σημειώσουμε εδώ ότι η μέθοδος που ακολουθήθηκε για την αναγνώριση δράσεων είναι νέα στη διεθνή βιβλιογραφία.

Στο **κεφάλαιο 7** παρατίθενται τα πειράματα που εκτελέσαμε εφαρμόζοντας τους αλγορίθμους του κεφαλαίου 4. Εξηγούμε τις επιλογές κάποιων παραμέτρων και γίνεται σύγκριση μεταξύ των μεθόδων που εισάγαμε στη διπλωματική αυτή, καθώς και με μία υλοποίηση παρόμοια με αυτή του [4].

Στο **κεφάλαιο 8** τέλος, αναφέρουμε τις βασικές συνεισφορές της εργασίας και προτείνουμε κατευθύνσεις για μελλοντική έρευνα.

Στο παράρτημα Α΄ παραθέτουμε κάποιες βασικές γνώσεις κυρτού προγραμματισμού που χρειάζονται για να κατανοηθεί ο φορμαλισμός του κεφαλαίου 4.

1.6 Συμβολισμός

Πριν αναφερθούμε σε οποιαδήποτε μαθηματική σχέση κρίνουμε σκόπιμο να εξηγήσουμε τους βασικούς συμβολισμούς που θα χρησιμοποιηθούν στα περισσότερα κεφάλαια του τόμου. Εξηγούμε αρχικά την γενική μεθοδολογία με την οποία αναπαριστούμε τα διάφορα μεγέθη. Στη συνέχεια παρουσιάζουμε σε έναν συγκεντρωτικό πίνακα (1.1) τη σημασία των πιο βασικών εξ'αυτών, έτσι ώστε ο αναγνώστης να μπορεί να ανατρέξει ανά πάσα στιγμή εδώ για να λάβει εξηγήσεις για κάποιο σύμβολο.

Έτσι, αναπαριστούμε:

- βαθμωτά μεγέθη με μικρά γράμματα συνηθισμένης έντασης ,π.χ $x \in \mathbb{R}$
- διανύσματα με μικρά γράμματα τονισμένης (bold) γραμματοσειράς, π.χ $\boldsymbol{x} \in \mathbb{R}^n$
- πίνακες με κεφαλαία γράμματα τονισμένης γραμματοσειράς π.χ $\boldsymbol{A} \in \mathbb{R}^{m \times n}$
- σύνολα με κεφαλαία καλλιγραφική γραμματοσειρά π.χ $\mathcal{X}$,εκτός από κάποιες εξαιρέσεις όταν μιλάμε για ευρέως γνωστά σύνολο ,όπως το $\mathbb{R}$
- ακολουθίες m αντικειμένων με αποσιωπητικά π.χ $\boldsymbol{x_1 ... x_m}$
- συναρτήσεις με τα μικρά γράμματα f, h, g κ.ά και όταν γράφουμε $f : \mathcal{X} \to \mathcal{Y}$ εννοούμε ότι η f έχει σαν σύνολο ορισμού το $\boldsymbol{dom} f = \mathcal{X}$ και σύνολο άφιξης το $\mathcal{Y}$

Όταν θέλουμε να αναφερθούμε στο i-οστό στοιχείο ενός διανύσματος $\boldsymbol{x}$ θα γράφουμε x_i, ενώ για το στοιχείο i, j ενός πίνακα $\boldsymbol{A}$ θα γράφουμε a_{ij}. Η αρίθμηση των δεικτών ξεκινάει πάντα από το 1.

Το εσωτερικό γινόμενο δύο διανυσμάτων $\boldsymbol{x}, \boldsymbol{y}$ γράφεται $< \boldsymbol{x}, \boldsymbol{y} >$ ή απλούστερα $\boldsymbol{x} \cdot \boldsymbol{y}$ και ισούται με $\sum_{i=1}^{d} x_i y_i$ όπου d η διάσταση του χώρου. Ακόμα, συμβολίζουμε τις ℓ_p νόρμες ενός διανύσματος $\boldsymbol{x}$ ως $\|\boldsymbol{x}\|_p$ με $\|\boldsymbol{x}\|_p = (\sum_{i=1}^{d} |x_i|^p)^{1/p}$

Για να συμβολίσουμε την ελάχιστη τιμή του συνόλου $\{f(\boldsymbol{x}) : \boldsymbol{x} \in C\}$ γράφουμε $\min_{\boldsymbol{x} \in C} f(\boldsymbol{x})$. Επίσης, όταν η τιμή αυτή δεν εφικτή, συμβολίζουμε το infimum του συνόλου με $\inf_{\boldsymbol{x} \in C} f(\boldsymbol{x})$. Όταν, η τιμή είναι εφικτή, συμβολίζουμε το όρισμα με το οποίο επιτυγχάνεται η ελάχιστη τιμή

με $argmin_{\boldsymbol{x}\in C} f(\boldsymbol{x})$. Όμοια για max και sup. Για να αναφερθούμε στην παράγωγο-κλίση (gradient) της f χρησιμοποιούμε τον εξής συμβολισμό: $\nabla f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_1}(\boldsymbol{x}) \quad \frac{\partial f}{\partial x_2}(\boldsymbol{x}) \quad \cdots \quad \frac{\partial f}{\partial x_n}(\boldsymbol{x})\right]^T$. Για να αναφερθούμε στον εσσιανό (hessian) πίνακα της f γράφουμε:

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Ακόμα χρησιμοποιούμε συμβολισμούς που αφορούν πιθανοτικά μεγέθη. Συγκεκριμένα, αναφερόμαστε με το σύνολο $\mathcal{D}$ σε μία κατανομή πιθανότητας πάνω σε ένα σύνολο $\mathcal{Z}$ και όταν μια τυχαία μεταβλητή z ακολουθεί την συγκεκριμένη κατανομή γράφουμε $z \sim \mathcal{D}$. Η πιθανότητα πραγματοποίησης ενός ενδεχομένου A γράφεται $\mathbb{P}[A]$, ενώ αν το ενδεχόμενο ορίζεται από μία συνάρτηση $f : \mathcal{Z} \to \{true, false\}$, δηλαδή $A = \{z : f(z) = true\}$ τότε η πιθανότητα πραγματοποίησής του γράφεται $\mathbb{P}_{z\sim\mathcal{D}}[f(z)]$ ή $D(\{z : f(z) = true\})$ Τέλος, αν μία τυχαία μεταβλητή ορίζεται ως $f : \mathcal{Z} \to \mathbb{R}$,τότε η αναμενόμενη τιμή της συμβολίζεται ως $\mathbb{E}_{z\sim\mathcal{D}}[f(z)]$.

Πίνακας 1.1: Πίνακας Συμβολισμών

Σύμβολο	Επεξήγηση
$\mathbb{R}$	Η ευθεία των πραγματικών αριθμών $(-\infty, +\infty)$
$\mathbb{R}^d$	Το σύνολο των d-διάστατων διανυσμάτων όπου κάθε συντεταγμένη ορίζεται στο $\mathbb{R}$
$\mathbb{R}_+$	Το σύνολο των μη-αρνητικών πραγματικών αριθμών
$\mathbb{N}$	Το σύνολο των φυσικών αριθμών $\{1, 2...\}$
$\mathbb{Z}$	Το σύνολο των ακεραίων αριθμών $\{\cdots, -2, -1, 0, 1, 2 \cdots\}$
$\mathbb{1}$[λογική έκφραση]	Δείκτρια συνάρτηση: ισούται με 1 μόνο όταν η λογική έκφραση είναι αληθής, αλλιώς 0
$\mathbf{0}_{n\times m}$	πίνακας $n \times m$ διαστάσεων με κάθε στοιχείο του να είναι ίσο με το 0
$\mathbf{1}_{n\times m}$	πίνακας $n \times m$ διαστάσεων με κάθε στοιχείο του να είναι ίσο με το 1
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$	διανύσματα στήλες
x_i, y_i, z_i	η i-οστή συντεταγμένη των διανυσμάτων
$\boldsymbol{x}, \boldsymbol{y}$ ή $< \boldsymbol{x}, \boldsymbol{y} >$	εσωτερικό γινόμενο 2 διανυσμάτων
$\|\boldsymbol{x}\|_p$	ℓ_p νόρμα ενός διανύσματος
$\boldsymbol{A} \in \mathbb{R}^{m\times n}$	πίνακας $m \times n$ διαστάσεων
$\boldsymbol{A}^T$	ανάστροφος πίνακας
$\boldsymbol{A}^{-1}$	αντίστροφος πίνακας
$\boldsymbol{A}^{\succeq}0$	θετικά ημιορισμένος πίνακας. Όπου η ανίσωση εφαρμόζεται στοιχείο προς στοιχείο,αυτό θα αναφέρεται επιτόπου
$\mathcal{X}$	σύνολο αντικειμένων
$\boldsymbol{x_1 ... x_m}$	ακολουθία αντικειμένων
$\boldsymbol{dom} f$	το σύνολο ορισμού της συνάρτησης f
$\mathcal{D}$	κατανομή πιθανότητας πάνω σε κάποιο σύνολο $\mathcal{Z}$
$\mathcal{D}(A)$	πιθανότητα πραγματοποίησης του ενδεχομένου $A \subseteq \mathcal{Z}$
$z \sim \mathcal{D}$	δειγματοληψία μιας μεταβλητής z από μία κατανομή $\mathcal{D}$
$\mathcal{D}^m$	κατανομή πιθανότητας πάνω στο $\mathcal{Z}^m$ όταν κάθε μεταβλητή z_i δειγματοληπτείται i.i.d
$\mathbb{P}$	πιθανότητα
$\mathbb{E}$	αναμενόμενη τιμή
$\min_{x\in C} f(x)$	η ελάχιστη τιμή του συνόλου $\{f(x) : x \in C\}$
$argmin_{x\in C} f(x)$	το όρισμα με το οποίο επιτυγχάνεται η ελάχιστη τιμή
$\nabla f(\boldsymbol{x})$	η παράγωγος-κλίση της συνάρτησης f στο σημείο $\boldsymbol{x}$
$\nabla^2 f(\boldsymbol{x})$	ο εσσιανός πίνακας της f

Κεφάλαιο 2

Σχετική Έρευνα

Στο παρακάτω κεφάλαιο κάνουμε μία σύντομη ανασκόπηση των εργασιών που σχετίζονται με την εν λόγω διπλωματική. Η κατανόηση βίντεο με τη βοήθεια κειμένου έχει προσεγγιστεί κατά κύριο λόγο με τρεις διαφορετικούς τρόπους: Ως Ασθενής Επίβλεψη (που είναι και η μέθοδος που ακολουθούμε εδώ), ως Αυτόματη Εξαγωγή Περιγραφών (Automatic Video Captioning) και ως Ευθυγράμμιση του Βίντεο με το Κείμενο (Alignment). Η τρίτη προσέγγιση μπορεί να θεωρηθεί υποπερίπτωση της δεύτερης. Παρακάτω παρουσιάζουμε τους 3 αυτούς άξονες, ενώ στο τέλος αναφέρουμε τη σχετική βιβλιογραφία για κάποια επιμέρους ζητήματα της διπλωματικής.

2.1 Το Κείμενο ως Ασθενής Επίβλεψη σε Προβλήματα Μάθησης σε Βίντεο

Σε αυτές τις προσεγγίσεις οι συγγραφείς χρησιμοποιούν συνοδευτικά στο βίντεο κείμενα προκειμένου να εξάγουν ασθενείς ετικέτες. Οι ετικέτες αυτές αξιοποιούνται στη συνέχεια από μοντέλα μάθησης προκειμένου να αναγνωριστούν τα περιεχόμενα του βίντεο. Εδώ τα κείμενα που κυριαρχούν στη βιβλιογραφία είναι τα σενάρια ταινιών καθώς όπως θα δούμε και στο κεφάλαιο 6 περιέχουν πολύ πλούσια πληροφορία και είναι ταυτόχρονα εύκολο να γίνει μία αρχική ευθυγράμμιση τους με το αντίστοιχο βίντεο, δηλαδή να δοθούν στο κείμενο χρονικές σημάνσεις.

Αναγνώριση Προσώπου: Στο πρόβλημα αυτό πριν εξερευνηθούν οι δυνατότητές του σε βίντεο είχε γίνει σημαντική δουλεία στον τομέα της Κατανόησης Εικόνας. Συγκεκριμένα, στο [2] εξετάζεται η αναγνώριση προσώπων σε μία μεγάλη συλλογή φωτογραφιών με λεζάντες από εφημερίδες. Το πρόβλημα είναι ότι σε μία φωτογραφία μπορεί να εμφανίζονται παραπάνω από ένα πρόσωπα και αντίστοιχα οι λεζάντες μπορεί να περιέχουν παραπάνω από ένα ονόματα. Οι συγγραφείς εκτελούν αρχικά μία μέθοδο Linear Discriminant Analysis χρησιμοποιώντας μόνο τα πρόσωπα που συνοδεύονται από μία και μόνο λεζάντα και έτσι ρίχνουν τη διαστασιμότητα του χώρου με τρόπο που γίνεται μία αρχική διάκριση μεταξύ των κλάσεων. Στη συνέχεια στο νέο χώρο εκτελείται μία τροποποιημένη μορφή του k-means η οποία σέβεται τις πιθανές ετικέτες κάθε προσώπου προκειμένου να ομαδοποιηθούν τα διάφορα πρόσωπα.

Στο τέλος 'κλαδεύονται' μικρά clusters και συγχωνεύονται τα ιδιαίτερα όμοια μεταξύ τους. Η εργασία αυτή επεκτάθηκε στο [3] μέσω ενός πιθανοτικού μοντέλου.

Στο πρόβλημα της Κατανόησης Βίντεο που μας ενδιαφέρει κυρίως, μία από τις πρώτες εργασίες είναι αυτή του [19] όπου γίνεται προσπάθεια ταξινόμησης των προσώπων στα επεισόδια μίας τηλεοπτικής σειράς, χρησιμοποιώντας τα ονόματα των ομιλητών που παρέχουν τα αντίστοιχα σενάρια. Η αναγνώριση γίνεται με απλές μεθόδους όρασης υπολογιστών. Συγκεκριμένα, ανιχνεύοντας την κίνηση των χειλιών οι συγγραφείς έβγαζαν συμπέρασμα για το πρόσωπο του χαρακτήρα που αναφέρεται ως ομιλητής στο σενάριο. Έτσι, χρησιμοποιώντας τις περιπτώσεις όπου μόνο ένα πρόσωπο έχει έντονη κίνηση χειλιών και συνοδεύεται από ένα μόνο όνομα ομιλητή, καθώς και τις περιπτώσεις όπου ένα όνομα αντιστοιχίζεται σε ένα και μόνο πρόσωπο (είτε κινεί τα χείλια είτε όχι), έφτιαχναν μία συλλογή από tracks προσώπου που η ταυτότητα τους είναι σωστή με μεγάλη πιθανότητα. Τα tracks αυτά τα ονόμαζαν exemplars και χρησιμοποιούνταν για την ταξινόμηση των υπολοίπων tracks με ένα απλό πιθανοτικό μοντέλο

Η εργασία αυτή επεκτάθηκε στο [53]. Οι διαφορές εστιάζονταν κυρίως στον τρόπο του εντοπισμού των προσώπων (προστέθηκε και ένας ανιχνευτής profile προσώπων) και της αναπαράστασης, καθώς και στον τρόπο μάθησης των μη exemplar tracks (με τη χρήση ενός SVM πολλαπλών πυρήνων όπου το βάρος του καθενός υπολογίζεται με μια τεχνική γνωστή ως Multiple Kernel Learning - MKL). Παρ΄ όλα αυτά η πληροφορία του κειμένου εισάγεται με τον ίδιο τρόπο με πριν. Οι 2 αυτές εργασίες μπορούμε να πούμε ότι αντιμετωπίζουν το πρόβλημα σαν Ημι-επιβλεπόμενη Μάθηση-Semi-Supervised Learning καθώς θεωρούν ότι μόνο ένα ποσοστό των δεδομένων είναι επισημειωμένο. Δεν μοντελοποιούν όμως το θόρυβο και επίσης τα μοντέλα μάθησης τους εκφυλίζονται σε πλήρως επιβλεπόμενα όπου τα δεδομένα εκπαίδευσης είναι τα exemplar tracks.

Στο [11] και στην επέκταση του ([10]), η μάθηση γίνεται με πιο προηγμένο τρόπο. Συγκεκριμένα, μοντελοποιούν το πρόβλημα ως Μάθηση με Υποψήφιες Ετικέτες (υποκατηγορία των Πιθανοτικών) ή αλλιώς Αμφίσημες Ετικέτες (Ambiguous Labels), όπως τις αποκαλούν οι ίδιοι. Εδώ, θεωρούν ότι σε κάθε χαρακτήρα αποδίδονται παραπάνω από μία ετικέτες, εκ των οποίων μόνο μία είναι σωστή. Προτείνουν μία κυρτή συνάρτηση κόστους που ενσωματώνει αυτή την πληροφορία, την οποία και ελαχιστοποιούν για όλα τα δείγματα του βίντεο, συμπεριλαμβανομένων και αυτών που στα [19, 53] οι συγγραφείς θα χαρακτήριζαν ως exemplars. Να σημειώσουμε εδώ ότι παρότι τα βασικά πειράματα έγιναν για την αναγνώριση προσώπων σε τηλεοπτικές σειρές, η γενικότητα του φορμαλισμού των συγγραφέων τους έδωσε τη δυνατότητα να εκτελέσουν πειράματα και για το πρόβλημα της αναγνώρισης προσώπων σε εικόνες και της αναγνώρισης ομιλητή, πάντα με τη χρήση ασθενούς επίβλεψης από κείμενο.

Στο [4] που είναι και το βασικό baseline μας, το πρόβλημα μοντελοποιείται ως Μάθηση Πολλαπλών Παραδειγμάτων και εφαρμόζεται σε κινηματογραφικές ταινίες. Συγκεκριμένα, για κάθε όνομα που λαμβάνουμε από το κείμενο συνθέτουμε σύνολα πολλαπλών παραδειγμάτων από τα πρόσωπα που περιέχονται στα χρονικά όρια του ονόματος αυτού. Αυτό σημαίνει ότι τουλάχιστον ένα από τα πρόσωπα πρέπει να πάρει την ετικέτα του συνόλου. Η αντιμετώπιση του προβλήματος γίνεται μέσω μιας κυρτής χαλάρωσης (convex relaxation) μίας συνδυαστι-

κής συνάρτησης κόστους προσθέτοντας περιορισμούς που κωδικοποιούν την πληροφορία των πολλαπλών παραδειγμάτων. Η λύση αυτή είναι ιδιαίτερα κομψή και μοντελοποιεί καλύτερα τις διάφορες πτυχές του προβλήματος από αυτήν του [11]. Αυτό γιατί στο [11] έχει γίνει η υπόθεση ότι ανάμεσα στις υποψήφιες ετικέτες βρίσκεται και η σωστή. Αυτό δεν είναι πάντα αλήθεια σε ρεαλιστικά περιβάλλοντα όπως οι ταινίες. Αντίθετα, στο [4] μοντελοποιείται επαρκώς ο θόρυβος μέσα από slack variables. Επίσης, δεδομένης της υπόθεσης ότι σε κάθε σύνολο παραδειγμάτων αρκεί ένα να πάρει την ετικέτα του συνόλου, είναι επιτρεπτό κάποιο δεδομένο να μην πάρει καμία από τις ετικέτες των συνόλων στα οποία ανήκει, πράγμα που δεν θα μπορούσε να γίνει στο [9].

Στο [47] η ασθενής επίβλεψη επεκτείνεται και σε αντωνυμίες, ουσιαστικά και άλλες γλωσσικές περιγραφές που υποννοούν το όνομα ενός χαρακτήρα (συναναφορές - coreferences). Εδώ, παρουσιάζεται ένα κοινό μοντέλο για τα γλωσσικά και τα οπτικά χαρακτηριστικά που πετυχαίνει να αναγνωρίζει τις τα πρόσωπα στο βίντεο και ταυτόχρονα να επιλύει τις συναναφορές στο κείμενο.

Άλλες εργασίες είναι το [12], όπου ως μορφή επίβλεψης χρησιμοποιούνται μόνο οι υπότιτλοι και όχι το σενάριο και αξιοποιούνται αρκετοί επιπλέον υπαινιγμοί από την οπτική και την ακουστική τροπικότητα και το [55], όπου γίνεται μία προσέγγιση με γράφους. Τέλος, στο [43] επεκτείνονται οι μέθοδοι προκειμένου να γίνεται αναγνώριση και των χαρακτήρων παρασκηνίου, ενσωματώνονται αναπαραστάσεις βαθιάς μάθησης και η μοντελοποίηση γίνεται με την αρχικό ορισμό της Μάθησης Πολλαπλών Παραδειγμάτων (όπως αυτός δόθηκε στο [16]), όπου κατασκευάζονται θετικά και αρνητικά σύνολα (bags) που περιέχουν ένα τουλάχιστον δεδομένο με την αντίστοιχη ετικέτα ή κανένα αντίστοιχα. Η μέθοδος αυτή φαίνεται να επιτυγχάνει το state-of-the-art στο πρόβλημα της αναγνώρισης προσώπων σε ταινίες.

Αναγνώριση Δράσης: Μία από τις πρώτες προσπάθειες Κατανόησης Βίντεο από κείμενο στο πρόβλημα της αναγνώρισης δράσης έγινε στο [32] όπου τα βίντεο είναι και πάλι κινηματογραφικές ταινίες. Εδώ, προκειμένου να εξαχθούν ετικέτες από το κείμενο εκπαιδεύεται ένας ταξινομητής με βάση μία Bag Of Words αναπαράσταση των γλωσσικών περιγραφών των δράσεων, ο οποίος εφαρμόζεται στις περιγραφές σκηνών των ταινιών δίνοντας ως έξοδο την ύπαρξη ή μη μίας ετικέτας. Για να γίνει αυτό, βέβαια, απαιτείται η συλλογή ενός αρκετά μεγάλου πλήθους διαφορετικών μεταξύ τους περιγραφών κάθε δράσης. Αφού γίνει η ταξινόμηση, τα χρονικά όρια της οπτικής δράσης καθορίζονται από τα χρονικά όρια της σκηνής στην οποία εντοπίστηκε η γλωσσική περιγραφή της. Τέλος, εκπαιδεύεται ένας ταξινομητής (οπτικών) δράσεων χρησιμοποιώντας τα δεδομένα που συνελέγησαν με αυτόματο τρόπο. Γίνεται επομένως αντιληπτό ότι εδώ δεν υπάρχει μοντελοποίηση μέσω ασθενούς επίβλεψης και η κατανόηση είναι ανακριβής ως προς τον εντοπισμό.

Ως συνέχεια αυτής της εργασίας παρουσιάστηκε το [37]. Εδώ χρησιμοποιείται γνώση από τα συμφραζόμενα μίας δράσης και συγκεκριμένα από τη σκηνή στην οποία εκτελείται. Συλλέγοντας με παρόμοιο τρόπο τα δεδομένα του κειμένου για τις δράσεις και τις σκηνές εκπαιδεύονται ταξινομητές οπτικών δράσεων και σκηνών οι οποίοι για να ταξινομήσουν κάθε δεδομένο λαμβάνουν υπόψη τους τα συμφραζόμενα. Ο λόγος που οι 2 αυτές εργασίες δεν εστιάζουν στην ακριβή κατανόηση των δράσεων μέσα στα βίντεο είναι γιατί το κυριότερο κίνητρο

τους είναι η συλλογή δεδομένων για την εκπαίδευση ενός ταξινομητή και όχι η κατανόηση. Έτσι, παρόλο που πετυχαίνουν καλά αποτελέσματα, χρειάζονται μεγάλο πλήθος ταινιών προκειμένου να εκπαιδεύσουν το σύστημα τους, έτσι ώστε να αντισταθμιστεί η ανακρίβεια του εντοπισμού.

Προκειμένου να λυθεί το πρόβλημα του ανακριβούς χρονικού εντοπισμού, στο [17] περιγράφεται μία προσέγγιση όπου η εκπαίδευση του ταξινομητή γίνεται από κοινού με την προσαρμογή των χρονικών ορίων. Για να το πετύχουν αυτό οι συγγραφείς εφαρμόζουν ένα χρονικό παράθυρο μέσα στα αρχικά χρονικά όρια κάθε δράσης και αναζητούν τη θέση εκείνη του παραθύρου που αυξάνει όσο περισσότερο γίνεται τη διακρισιμότητα μεταξύ των κλάσεων. Ο φορμαλισμός είναι ένα discriminative clustering που βελτιστοποιεί τις παραμέτρους του μοντέλου ταξινόμησης και τη θέση του παραθύρου εναλλάξ, πάντα με στόχο τη μέγιστη διακρισιμότητα. Η μέθοδος αυτή μπορεί να γίνει αντιληπτή ως παραλλαγή της Μάθησης Πολλαπλών Παραδειγμάτων, θεωρώντας ότι από το σύνολο των παραθύρων μόνο ένα (και όχι τουλάχιστον ένα) πρέπει να πάρει την αντίστοιχη ετικέτα.

Στο [4] δίνεται μία προσέγγιση αναγνώρισης δράσεων ίδια με αυτή των προσώπων (με Μάθηση Πολλαπλών Παραδειγμάτων). Ο εντοπισμός τους γίνεται με έναν αρκετά απλό τρόπο, απλά επεκτείνοντας το bounding box του αντίστοιχου ανθρώπινου προσώπου, ενώ η εξαγωγή των ετικετών γίνεται με τη χρήση του εργαλείου SEMAFOR που αξιοποιεί τη σημασιολογία των λέξεων (περισσότερα στα κεφάλαια 6, 7). Οι μέθοδοι αυτοί είναι αρκετά περιοριστικές και δεν επιτυγχάνουν αρκετά καλή κατανόηση. Υιοθετήθηκαν όμως από τους συγγραφείς προκειμένου να αναδειχθεί το πλεονέκτημα της από κοινού αναγνώρισης ατόμων και δράσεων σε σύγκριση με την απλή αναγνώριση δράσεων και δεν δόθηκε ιδιαίτερα βαρύτητα στην πληρότητα της κατανόησης.

Στο [5] αναγνωρίζονται δράσεις σε ένα βίντεο, όταν η επίβλεψη για αυτές δίνεται μέσα από μία ακολουθία τους. Μπορεί να γίνει αντιληπτό και ως πρόβλημα ευθυγράμμισης όπου πρέπει κάθε ετικέτα να αποδοθεί στο αντίστοιχο απόσπασμα. Εδώ, δεν υπάρχει ουσιαστική συνεισφορά του κειμένου, καθώς οι ετικέτες γίνονται αντιληπτές με εντελώς συμβολικό τρόπο. Παρ΄ όλα αυτά η μάθηση παραμένει ασθενής καθώς δεν γνωρίζουμε τα ακριβή όρια κάθε δράσης.

Τέλος, για λόγους πληρότητας, να σημειώσουμε ότι μία ακόμα προσέγγιση από κοινού μάθησης ατόμων και δράσεων παρουσιάζεται στο [46] όπου η μοντελοποίηση γίνεται με Conditional Random Fields και μοιάζει με το φορμαλισμό της Μάθησης Πολλαπλών Παραδειγμάτων με Πολλαπλές Ετικέτες. Οι συγγραφείς αντιλαμβάνονται το βίντεο ως κάτι ενιαίο στο εσωτερικό του οποίου μπορεί να συμβαίνουν παραπάνω από μία δράσεις-γεγονότα. Άρα πρέπει να του αποδοθούν παραπάνω από μία ετικέτες. Εδώ τα βίντεο είναι μικρής διάρκειας και συνοδεύονται από μία γλωσσική σύνοψη των συμβάντων τους.

2.2 Automatic Video Captioning

Τα προβλήματα αυτά μοιάζουν πολύ με τα προβλήματα Machine Translation, όπου μία πρόταση μεταφράζεται από μία γλώσσα σε μία άλλη. Όμοια και εδώ, ένα βίντεο μεταφράζεται στην αντίστοιχη γλωσσική αναπαράσταση. Στο πεδίο των εικόνων μία από τις πρώτες εργα-

σίες ήταν η [18], όπου τμήματα εικόνων αντιστοιχίζονται σε λέξεις με βάση τις συνοδευτικές λεζάντες. Αφού διασπαστεί η εικόνα σε περιοχές εκπαιδεύεται ένα μοντέλο με το όνομα IBM 2 που εκφράζεται μέσα από δύο κατανομές πιθανότητας: την πιθανότητα να πάρουμε μία περιοχή εικόνας δεδομένης μίας λέξης και την πιθανότητα να πάρουμε μία λέξη δεδομένης της εικόνας. Η εκπαίδευση του γίνεται με έναν κλασικό αλγόριθμο Expectation Maximization.

Στον τομέα της κατανόησης βίντεο αναφέρουμε για λόγους πληρότητας δύο χαρακτηριστικά παραδείγματα ([49, 42]), όπου μαθαίνονται αναπαραστάσεις που μεταφράζουν βίντεο σε προτάσεις και το αντίστροφο. Δεν θα επεκταθούμε παραπάνω γιατί αυτός ο τομέας ξεπερνά τα όρια της διπλωματικής.

2.3 Ευθυγράμμιση του Βίντεο με το Κείμενο

Όπως είδαμε στις προηγούμενες ενότητες η χρήση του κειμένου για την κατανόηση ενός βίντεο απαιτεί την, έστω και ανακριβή, ευθυγράμμιση των 2 τροπικοτήτων. Ακόμα, αποτελεί από μόνη της προσέγγιση για την κατανόηση του βίντεο, καθώς όσο πιο ακριβής είναι χρονικά, τόσο πιο ακριβείς είναι σημασιολογικά οι περιγραφές κάθε τμήματος του βίντεο που η ίδια παρέχει. Η πιο απλή προσέγγιση δίνεται στο [19] όπου το σενάριο μίας ταινίας ευθυγραμμίζεται με τους υπότιτλους της. Είναι αρκετά περιοριστική όμως γιατί τα 2 αυτά κείμενα υπάρχουν μόνο σε ταινίες και σειρές και οι υπότιτλοι είναι σχετικά αραιοί, με αποτέλεσμα να μην παίρνουμε ακριβείς περιγραφές.

Στο [50], γίνεται μία προσπάθεια ευθυγράμμισης ταινιών με το σενάριο τους, χωρίς να παρέχονται υπότιτλοι. Ακόμα, στο [54] επιχειρείται η ευθυγράμμιση βιβλίων με τις μεταφορές τους στο σινεμά. Τέλος, στο [6] περιγράφεται μία γενικότερη μορφή ευθυγράμμισης , με εφαρμογή σε βίντεο μαγειρικής συνοδευόμενα από γλωσσικές περιγραφές, αναζητώντας δύο γραμμικές αναπαραστάσεις που μετατρέπουν τη μία τροπικότητα στην άλλη. Ούτε εδώ θα επεκταθούμε περαιτέρω καθώς η ευθυγράμμιση είναι εκτός ορίων της διπλωματικής.

2.4 Λοιπή Σχετική Έρευνα

Όσον αφορά τις υποκατηγορίες και τις μεθόδους της ασθενούς επίβλεψης, υπάρχει ένας μεγάλος όγκος σχετικής έρευνας. Γι' αυτό, στο κεφάλαιο 3 περιγράφεται εν συντομία. Εκεί εξηγούμε διάφορα προβλήματα που αναφέραμε και εδώ καθώς και άλλες χρήσιμες έννοιες της Μηχανικής Μάθησης. Ακόμα, ο κορμός των αλγορίθμων μάθησης που προτείνουμε είναι ο αλγόριθμος DIFFRAC ([1]) ο οποίος, παρότι εισήχθη ως αλγόριθμος clustering, λόγω της ευελιξίας του έχει χρησιμοποιηθεί για να μοντελοποιήσει ποικίλα σενάρια μάθησης. Για αυτό το λόγο τον περιγράφουμε εκτενώς στο κεφάλαιο 4.

Ακόμα, όσον αφορά τις αναπαραστάσεις του βίντεο και του κειμένου που αποτέλεσαν βασικούς άξονες της διπλωματικής, η σχετική έρευνα, αλλά και η έρευνα που αξιοποιήθηκε άμεσα εδώ, περιγράφεται επί τόπου στα κεφάλαια 5 και 6 αντίστοιχα.

Μέρος Ι

Θεωρητικό Υπόβαθρο-Μαθηματικός Φορμαλισμός του Προβλήματος

Κεφάλαιο 3

Μηχανική Μάθηση

3.1 Εισαγωγή

Στο παρόν κεφάλαιο γίνεται μία σύντομη ανασκόπηση του τομέα της Μηχανικής Μάθησης. Προκειμένου να την περιγράψουμε παραθέτουμε τον ιδιαίτερα δημοφιλή ορισμό που δόθηκε το 1959, από τον Arthur Samuel και μιλάει για αυτήν ως ¨Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί¨. Συγκεκριμένα, οι μέθοδοι που αναπτύχθηκαν στα πλαίσια αυτού του πεδίου δεν δίνουν αυστηρά καθορισμένες οδηγίες σε ένα πρόγραμμα προκειμένου να εκτελέσει μία εργασία, αλλά περιγράφουν τη διαδικασία με την οποία το πρόγραμμα θα καταλήξει από μόνο του στον καθορισμό των οδηγιών. Το μόνο που χρειάζεται η μηχανή για να μάθει είναι ένα σύνολο δεδομένων τα οποία αντιλαμβάνεται ως πρότυπα. Από αυτά μαθαίνει να εξάγει συμπεράσματα και να εκτελεί προβλέψεις σε νέα άγνωστα δεδομένα (όπως π.χ η κατηγορία στην οποία ένα δεδομένο ανήκει) ή να λαμβάνει αποφάσεις. Έτσι, δεν χρειάζεται πλέον η μελέτη των ιδιοτήτων των δεδομένων από τον άνθρωπο, διαδικασία συνήθως δύσκολη ή ακόμα και αδύνατη, καθώς αυτή γίνεται αυτόματα από τον υπολογιστή ο οποίος στην ουσία αυτοεκπαιδεύεται. Για παράδειγμα, σε ένα πρόβλημα εντοπισμού προσώπου θα έπρεπε ο προγραμματιστής να δώσει ως είσοδο στον υπολογιστή ένα τεράστιο σύνολο από κανόνες που περιγράφουν τα στοιχεία που συνθέτουν ένα ανθρώπινο πρόσωπο (όπως την ύπαρξη και τις πιθανές θέσεις των ματιών, της μύτης κ.ά). Η εξαγωγή αυτού του συνόλου κανόνων είναι ανέφικτη αν αναλογιστεί κανείς την ποικιλομορφία των ανθρώπινων προσώπων. Αντίθετα, ένας αλγόριθμος μηχανικής μάθησης δεχόμενος ως είσοδο ένα μεγάλο και αντιπροσωπευτικό δείγμα της ποικιλομορφίας αυτής μπορεί από μόνος του να εξάγει τους απαραίτητους κανόνες. Άλλωστε, η διαδικασία αυτή είναι σε πλήρη αντιστοιχία με την λειτουργία του ανθρώπινου εγκεφάλου, ο οποίος δεν έχει κάποια εγγενή γνώση αλλά την αποκτά με την εμπειρία ερμηνεύοντας αυτά που βλέπει γύρω του. Η μηχανική μάθηση είναι ένας τομέας με μεγάλη εξέλιξη τα τελευταία 50 χρόνια και έχει δώσει λύσεις σε πολλά σύγχρονα προβλήματα και ώθηση σε άλλους τομείς όπως είναι η όραση υπολογιστών, η ρομποτική , η επεξεργασία φυσικής γλώσσας, η εξόρυξη πληροφορίας κ.ά. Δεδομένου του ότι η εν λόγω διπλωματική μελετά, μέσα από μία εφαρμογή, τις δυνατότητες ενός ιδιαίτερου σεναρίου μάθησης, κρίνουμε χρήσιμη την περιήγηση στα διαφορετικά σενάρια που έχουν εφαρμοστεί

προκειμένου να αναδειχτούν οι ομοιότητες και οι διαφορές τους, αλλά και η αναγκαιότητα ύπαρξης του καθενός. Ακόμα, αναφέρονται έννοιες οι οποίες θα χρησιμεύσουν στην πορεία προκειμένου να περιγραφούν οι διάφοροι αλγόριθμοι.

3.2 Φορμαλισμός ενός προβλήματος μάθησης

Ένας γενικός τρόπος να οριστεί μαθηματικά ένα πρόβλημα μηχανικής μάθησης παρουσιάζεται στο [51]. **Γενικά, σε ένα τέτοιο πρόβλημα στόχος είναι για κάθε δεδομένο x που παρουσιάζεται στον αλγόριθμο, να εξάγεται ένα συμπέρασμα το οποίο αναπαριστούμε με μία ετικέτα** y. Συγκεκριμένα, κάθε πρόβλημα μάθησης αποτελείται από τα εξής στοιχεία:

1. **Η είσοδος του μηχανισμού μάθησης**

   - **Σύνολο Δεδομένων-Domain Set**: Το σύνολο $\mathcal{X}$ στο οποίο ανήκουν τα αντικείμενα που θέλουμε να μάθουμε. Κάθε αντικείμενο αναπαρίσταται στο χώρο $\mathcal{X}$ με ένα διάνυσμα χαρακτηριστικών $\boldsymbol{x}$ και ονομάζεται και δείγμα (instance ή sample).
   - **Σύνολο Ετικετών-Label Set**: Το σύνολο $\mathcal{Y}$ στο οποίο ανήκουν οι τιμές των ετικετών που μπορούν να αποδοθούν σε κάθε δείγμα. Τα στοιχεία του συνόλου μπορεί να είναι διακριτές τιμές (όπως σε προβλήματα ταξινόμησης, π.χ $\mathcal{Y} = \{0, 1..Y\}$, όπου κάθε τιμή αντιπροσωπεύει μία διαφορετική κατηγορία) ,ή συνεχή διαστήματα (όπως σε προβλήματα παλινδρόμησης, π.χ $\mathcal{Y} = \mathbb{R}$)
   - **Σύνολο Εκπαίδευσης-Training Data**: Το σύνολο $\mathcal{S}$ στο οποίο ανήκουν οι τούπλες δειγμάτων-ετικετών στις οποίες έχει πρόσβαση ο αλγόριθμος εκμάθησης. Από το σύνολο αυτό ορίζεται η φύση της επίβλεψης. Οι διάφορες υποκατηγορίες με βάση αυτό το κριτήριο θα περιγραφούν παρακάτω.

2. **Η έξοδος του μηχανισμού μάθησης**: Η έξοδος $h : \mathcal{X} \to \mathcal{Y}$ αποτελεί μία συνάρτηση πρόβλεψης της ετικέτας ενός δεδομένου από την τιμή του διανύσματος χαρακτηριστικών του (**Ταξινομητής-Classifier ή Predictor**). Συνήθως, η συνάρτηση h επιλέγεται από μία κλάση συναρτήσεων ή αλλιώς μοντέλο $\mathcal{H}$. Η επιλογή γίνεται από την εκτέλεση ενός αλγορίθμου βελτιστοποίησης A. Η έξοδος αυτή του αλγορίθμου μπορεί να εκφραστεί και ως $A(\mathcal{S})$.

3. **Μοντέλο-κατανομή δεδομένων**: Η κατανομή πιθανότητας $\mathcal{D}$ πάνω στο $\mathcal{X}$ από την οποία παράγονται τα $\boldsymbol{x}$. Συγκεκριμένα, θεωρούμε ότι τα δεδομένα παράγονται δειγματοληπτώντας την $\mathcal{D}$ και στη συνέχεια τους αποδίδεται μία ετικέτα από μία συνάρτηση $f : \mathcal{X} \to \mathcal{Y}$, την οποία και θέλουμε να προσεγγίσουμε.

4. **Μετρική/Συνάρτηση σφάλματος**: Είναι ένας αριθμός που ποσοτικοποιεί την επιτυχία του αλγορίθμου. Συγκεκριμένα, ένας γενικός τρόπος να οριστεί η μετρική αυτή είναι το λεγόμενο πραγματικό σφάλμα (True Risk). Το μέγεθος αυτό είναι ίσο με

την πιθανότητα η έξοδος του αλγορίθμου να προβλέψει λάθος ετικέτα και δίνεται από τον τύπο:

$$L_{\mathcal{D},f}(h) \stackrel{def}{=} \mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})] = D(\{\boldsymbol{x} : h(\boldsymbol{x}) \neq f(\boldsymbol{x})\}) \tag{3.1}$$

Στόχος κάθε μηχανισμού μάθησης είναι να ελαχιστοποιήσει την πιθανότητα αυτή. Δεδομένης όμως της άγνοιας των $\mathcal{D}, f$ το σφάλμα αυτό δεν μπορεί να υπολογιστεί, επομένως χρησιμοποιούμε άλλα μεγέθη ως στόχους ελαχιστοποίησης προκειμένου να προσεγγίσουμε την ζητούμενη συνάρτηση f, τα οποία υπολογίζονται πάνω στα γνωστά δεδομένα.

5. **Μετρική Αξιολόγησης**: Συνήθως, σε ένα πρόβλημα μάθησης είναι αναγκαίο η έξοδος του αλγορίθμου να μπορεί να συμπεριφερθεί καλά σε δεδομένα άγνωστα στον αλγόριθμο κατά τη διάρκεια της μάθησης. Αυτή η ιδιότητα ονομάζεται **γενίκευση** και η έλλειψη της **υπερπροσαρμογή (overfitting)**. Η υπερπροσαρμογή είναι ένας πολύ συνηθισμένος κίνδυνος στη μηχανική μάθηση και συμβαίνει συνήθως όταν το μοντέλο που έχουμε επιλέξει είναι αρκετά πολύπλοκο με αποτέλεσμα να προβλέπει υπερβολικά καλά τα δεδομένα εκπαίδευσης αλλά μόνο αυτά. Έτσι, είναι αναγκαίο να ποσοτικοποιήσουμε την ικανότητα του συστήματος να γενικεύει. Αυτό μπορεί να γίνει είτε με τη χρήση της ίδιας συνάρτησης σφάλματος που χρησιμοποιήθηκε στα δεδομένα εκπαίδευσης, είτε με τη χρήση άλλων μετρικών αξιολόγησης όπως είναι οι μετρικές **accuracy, precision-recall, F1-measure** και άλλες. Όπου χρησιμοποιηθεί η εκάστοτε μετρική στον τόμο θα αναφέρεται και ο ορισμός της.

Ελαχιστοποίηση Εμπειρικού Σφάλματος-Συναρτήσεις Σφάλματος

Όπως προαναφέραμε, προκειμένου να προσεγγίσουμε το πραγματικό σφάλμα χρησιμοποιούμε μετρικές οι οποίες δεν απαιτούν την γνώση ούτε της κρυφής κατανομής $\mathcal{D}$ ούτε της κρυφής συνάρτησης απόδοσης ετικετών f. Αυτές οι μετρικές $\ell : \mathcal{H} \times \mathcal{S} \to \mathbb{R}_+$ ονομάζονται συναρτήσεις κόστους ή σφάλματος(**loss functions**). Με βάση αυτές μπορούμε να ορίσουμε γενικότερα το πραγματικό σφάλμα ως:

$$L_{\mathcal{D}}(h) \stackrel{def}{=} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\ell(h, \boldsymbol{x})] \tag{3.2}$$

Ακόμα, μπορούμε να ορίσουμε μία ακόμα ποσότητα που ονομάζεται **εμπειρικό ρίσκο-empirical risk** και αποτελεί μία εκτίμηση του πραγματικού, υπολογισμένο πάνω στα γνωστά δεδομένα. Δηλαδή, αν $\boldsymbol{x_1}, \boldsymbol{x_2}, ..\boldsymbol{x_m}$ τα δείγματα που βρίσκονται στο $\mathcal{S}$ τότε:

$$L_{\mathcal{S}}(h) \stackrel{def}{=} \frac{1}{m}\sum_{i=1}^{m} \ell(h, \boldsymbol{x}_i) \tag{3.3}$$

Το εμπειρικό ρίσκο είναι και η ποσότητα που αναλαμβάνει συνήθως να ελαχιστοποιήσει ο αλγόριθμος μάθησης και για αυτό το λόγο μπορούμε πλέον να τον ονομάζουμε **Αλγόριθμο Ελαχιστοποίησης Εμπειρικού Σφάλματος(Empirical Risk Minimization-ERM)**.

Ακόμα, αξίζει εδώ να σημειώσουμε ότι μία από τις βασικές υποθέσεις στο machine learning είναι ότι τα δείγματα $\boldsymbol{x_i}$ λαμβάνονται από την ίδια κατανομή $\mathcal{D}$ και ανεξάρτητα το ένα από το άλλο (**i.i.d-indipendently and identically distributed**). Έτσι, προκύπτει με προφανή τρόπο ότι:

$$L_{\mathcal{D}}(h) = \mathbb{E}_{\{\boldsymbol{x_1},\boldsymbol{x_2},\ldots,\boldsymbol{x_m}\}\sim\mathcal{D}^m}[L_{\mathcal{S}}(h)] \tag{3.4}$$

3.3 Οι βασικοί τύποι προβλημάτων της Μηχανικής Μάθησης

Προκειμένου κανείς να κατηγοριοποιήσει τις διάφορες μεθόδους και τα διάφορα προβλήματα μάθησης είναι σύνηθες να το κάνει με βάση τις διαφοροποιήσεις τους σε ένα από τα παραπάνω στοιχεία του φορμαλισμού. Για παράδειγμα, όσον αφορά το σύνολο ετικετών $\mathcal{Y}$ μπορούν να προκύψουν τα προβλήματα ταξινόμησης όπου προσπαθούμε να χωρίσουμε τα δεδομένα μας σε διάφορες κατηγορίες και τα προβλήματα παλινδρόμησης, όπου για κάθε δεδομένο ορίζεται μια τιμή από ένα συνεχές σύνολο μέσω μία συνάρτησης, την οποία και θέλουμε να προσεγγίσουμε. Ακόμα, όσον αφορά το μοντέλο/κλάση συναρτήσεων h μπορούν να προκύψουν μεγάλες οικογένειες ταξινομητών όπως οι γραμμικοί, οι τετραγωνικοί και άλλοι. Ένα από τα πιο σημαντικά κριτήρια ταξινόμησης των μεθόδων μηχανικής μάθησης είναι η φύση της επίβλεψης που παρέχεται στο μηχανισμό μάθησης. Συγκεκριμένα, διαχωρίζουμε τα προβλήματα με βάση τα περιεχόμενα του συνόλου $\mathcal{S}$, δηλαδή της πρότερης (a priori) γνώσης που χρησιμοποιεί ο αλγόριθμος προκειμένου να εξάγει την συνάρτηση h. Παρακάτω, κάνουμε μια μικρή ανάλυση των διαφόρων τύπων μάθησης, σύμφωνα και με το [9], με έμφαση στον τύπο της ασθενούς επίβλεψης, με τον οποίο και θα ασχοληθούμε στην παρούσα διπλωματική.

Επιβλεπόμενη Μάθηση-Supervised Learing

Η επιβλεπόμενη μάθηση είναι η πιο συχνή τεχνική επίλυσης προβλημάτων αναγνώρισης προτύπων καθώς προσφέρει στο σύστημα την περισσότερη a priori γνώση. Συγκεκριμένα, εδώ το σύνολο εκπαίδευσης είναι $\mathcal{S} = \cup_{i=1}^{m}(\boldsymbol{x_i}, y_i)$,όπου $\boldsymbol{x_i}$ τα δείγματα στα οποία έχει πρόσβαση ο αλγόριθμος και y_i οι ετικέτες που αντιστοιχούν στο κάθε ένα. Διαισθητικά αντιλαμβανόμαστε ότι με χρήση αυτής της γνώσης μπορούμε να προβλέψουμε αρκετά καλά τη συνάρτηση h αλλά και την κατανομή $\mathcal{D}$ αν έχουμε ένα αρκετά αντιπροσωπευτικό σύνολο δειγμάτων. Αυτή η διαίσθηση θεμελιώθηκε και θεωρητικά από τον κλάδο του Computational Learning όπως διαβάζουμε και πάλι στο [51]. Συγκεκριμένα,(χωρίς να επεκταθούμε πολύ) αξίζει να αναφέρουμε ότι ο Valiant στο [60] με τη θεωρία του probably approximately correct learning (PAC learning) και οι Vapnik–Chervonenkis στο [61] με τη θεωρία του **VC dimension** εξήγησαν με αλγοριθμικό τρόπο πότε ένα σύνολο $\mathcal{S}$ είναι αντιπροσωπευτικό και πότε μια κλάση συναρτήσεων $\mathcal{H}$ μπορεί με μεγάλη πιθανότητα να πετύχει μικρό σφάλμα γενίκευσης. Στα πλαίσια αυτής της θεωρίας αποδείχτηκε ότι προκειμένου ο αλγόριθμος να μπορέσει να μάθει ¨αρκετά καλά' (**probably approximately corect**)τη συνάρτηση f,η κλάση συναρτήσεων $\mathcal{H}$ δεν μπορεί να περιέχει όλες τις πιθανές συναρτήσεις απόδοσης ετικετών πάνω σε

ένα σύνολο $\mathcal{X}$. Αντίθετα, είναι συγκεκριμένες οι κλάσεις συναρτήσεων[1] που μπορούν εν γένει να επιτύχουν το στόχο της μάθησης και η πολυπλοκότητα της επιτυχίας εξαρτάται σε μεγάλο βαθμό από τις ιδιότητες της $\mathcal{H}$. Αυτό, μας οδηγεί στο να κατανοήσουμε τη σημασία ενός από τα βασικά κριτήρια επιλογής ενός αλγορίθμου μάθησης, δηλαδή της επιλογής του **μοντέλου μάθησης**. Το άλλο κριτήριο είναι η η συνάρτηση κόστους. Αναφέρουμε εδώ ενδεικτικά μερικές μεθόδους:

1. **Generative-παραγωγικές μέθοδοι**. Πρόκειται για μεθόδους οι οποίες προσπαθούν να προβλέψουν την κατανομή $\mathcal{D}$ και με βάση αυτή να βγάλουν συμπεράσματα για τη δεσμευμένη ,από την τιμή του δείγματος, πιθανότητα της επιλογής της κάθε ετικέτας. Συνήθως σε αυτά τα μοντέλα η συνάρτηση κόστους που μεγιστοποιείται/ελαχιστοποιείται είναι ή +/- **Εκτιμήτρια Μέγιστης Πιθανοφάνειας(maximum likelihood estimation-MLE)**, δηλαδή η πιθανότητα ένα μοντέλο να παράξει τα δείγματα μας, δεδομένων των παραμέτρων του. Συνήθως το μέγεθος που χρησιμοποιείται είναι ο λογάριθμος της πιθανοφάνειας, δηλαδή:

$$L(\boldsymbol{\theta}\,;\,\boldsymbol{x_1},\ldots,\boldsymbol{x_n}) = f(\boldsymbol{x_1},\boldsymbol{x_2},\ldots,\boldsymbol{x_n}\mid\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x_i}\mid\boldsymbol{\theta}) \quad \text{θεωρώντας i.i.d} \tag{3.5}$$

$$\hat{\ell} = \frac{1}{n}\ln L(\boldsymbol{\theta}\,;\,\boldsymbol{x_1},\ldots,\boldsymbol{x_n}) = \sum_{i=1}^{n}\ln f(\boldsymbol{x_i}\mid\boldsymbol{\theta}) \tag{3.6}$$

$$\{\hat{\boldsymbol{\theta}}_{\text{mle}}\} \subseteq \{\arg\max_{\theta\in\Theta}\,\hat{\ell}(\boldsymbol{\theta}\,;\,\boldsymbol{x_1},\ldots,\boldsymbol{x_n})\} \tag{3.7}$$

   Ενδεικτικά, αναφέρουμε τους ταξινομητές **Naive Bayes** και **Hidden Markov Models-Κρυφά μαρκοβιανά μοντέλα** που χρησιμοποιούν παραγωγικές μεθόδους για την εκπαίδευση τους.

2. **Discriminative-διακριτικές μέθοδοι**. Πρόκειται για μεθόδους οι οποίες προσπαθούν να βγάλουν απευθείας συμπέρασμα για την συνάρτηση h, χωρίς να εκτιμήσουν την κατανομή $\mathcal{D}$. Για την ακρίβεια, οι μέθοδοι αυτές εκτιμούν την δεσμευμένη πιθανότητα $\mathbb{P}[y|\boldsymbol{x}]$ αδιαφορώντας για την από κοινού πιθανότητα $\mathbb{P}[y,\boldsymbol{x}]$. Πολλές φορές οι αλγόριθμοι αυτοί ελαχιστοποιούν συναρτήσεις κόστους οι οποίες δεν έχουν άμεση πιθανοτική ερμηνεία, αλλά στοχεύουν στην βέλτιστη πρόβλεψη από το μοντέλο,των ετικετών των ήδη γνωστών δεδομένων.

   Χαρακτηριστικά παραδείγματα διακριτικών ταξινομητών είναι οι **Μηχανές Διανυσματικής Υποστήριξης-SVMs**, όπου η κλάση συναρτήσεων $\mathcal{H}$ από την οποία επιλέγεται ο βέλτιστος ταξινομητής είναι η κλάση των υπερεπιπέδων, και η **Γραμμική Παλινδρόμηση**, όπου η κλάση συναρτήσεων $\mathcal{H}$ που επιλέγεται προκειμένου να προσεγγιστεί η συνάρτηση f είναι μία αφινική συνάρτηση (βλ. παράρτημα Α′) των συντεταγμένων του διανύσματος $\boldsymbol{x}$. Ακόμα, άλλες περιπτώσεις είναι τα **Δένδρα Απόφασης** που χωρίζουν τον χώρο σε υποπεριοχές με βάση τις τιμές των διανυσμάτων

[1] αυτές που έχουν πεπερασμένο VC dimension.

εκπαίδευσης και αντιστοιχίζουν κάθε υποπεριοχή σε μία τιμή ετικέτας, και τα **Νευρωνικά Δίκτυα** που αποτελούν σύνθετα δίκτυα τεχνητών νευρώνων και μπορούν θεωρητικά να προσεγγίσουν μία οποιαδήποτε συνάρτηση.

Μη Επιβλεπόμενη Μάθηση-Unsupervised Learing

Η τεχνική αυτή στοχεύει στην εξαγωγή συμπερασμάτων για τη δομή ενός συνόλου δεδομένων χωρίς να έχουμε καμία πληροφορία για τη δομή αυτή. Συγκεκριμένα, στόχος είναι η εύρεση των τιμών κάποιων κρυφών μεταβλητών(latent variables) όπως είναι οι άγνωστες ετικέτες των δεδομένων, η εκτίμηση μίας άγνωστης κατανομής πιθανότητας κ.ά. Εδώ το σύνολο εκπαίδευσης είναι $\mathcal{S} = \cup_{i=1}^{m}(\boldsymbol{x_i})$. Είναι εύκολο να καταλάβει κανείς ότι συνήθως αυτά τα προβλήματα δεν έχουν μοναδική σωστή λύση και η κρυφή δομή στην οποία θα καταλήξει ο αλγόριθμος εξαρτάται σε μεγάλο βαθμό από τις υποθέσεις που κάνουμε. Παρακάτω παραθέτουμε μερικά υποπροβλήματα αυτής της κατηγορίας:

1. **Ανάλυση Συνιστωσών(Component Analyis)**- Μείωση Διάστασης(Dimensionality Reduction: Σε αυτά τα προβλήματα θεωρούμε ότι παρατηρούμε μία ακολουθία διανυσμάτων $\boldsymbol{x_1, x_2, ...x_m}$ τα οποία έχουν προκύψει από τον συνδυασμό μίας άλλης(μικρότερης συνήθως ακολουθίας $\boldsymbol{s_1, s_2, ...s_m}$ με την εφαρμογή του εξής μετασχηματισμού: $\boldsymbol{X} = \boldsymbol{A} \cdot \boldsymbol{S}$ ($\boldsymbol{X}$,$\boldsymbol{S}$ οι πίνακες όπου κάθε γραμμή είναι ένα διάνυσμα των ακολουθιών). Στόχος είναι να αντιστρέψουμε το πρόβλημα δηλαδή να προσεγγίσουμε τα $\boldsymbol{s_i}$ μέσα από την εύρεση ενός πίνακα $\boldsymbol{B}$ τέτοιου ώστε $\hat{\boldsymbol{S}} \simeq \boldsymbol{S} = \boldsymbol{B} \cdot \boldsymbol{X}$. Εδώ γίνεται προφανές ότι αναλόγως με τις παραδοχές που κάνουμε προκύπτουν διαφορετικές λύσεις. Για παράδειγμα στην **ανάλυση κύριων συνιστωσών(PCA)**, υποθέτουμε ότι τα διανύσματα $\boldsymbol{s_i}$ είναι τα γραμμικώς ανεξάρτητα διανύσματα που εξηγούν όσο το δυνατόν καλύτερα τη διασπορά των δεδομένων. Στην **ανάλυση ανεξάρτητων συνιστωσών (ICA)**, υποθέτουμε ότι οι συνιστώσες έχουν την μέγιστη στατιστική ανεξαρτησία. Άλλες μέθοδοι είναι η **μη αρνητική παραγοντοποίηση πίνακα(non-negative matrix factorization)** και η **ανάλυση πίνακα σε ιδιάζουσες τιμές(singular value decomposition)**.

2. **Συσταδοποίηση-clustering**: Εδώ τα προβλήματα στοχεύουν στην εύρεση ομάδων δεδομένων, δηλαδή στην απόδοση ετικετών σε αυτά.

   - Ένας βασικός τρόπος επίλυσης τέτοιων προβλημάτων είναι η μοντελοποίηση όπου θεωρούμε ότι τα δεδομένα μας προέρχονται από μία κοινή κατανομή (**distribution-based**). Ένα παράδειγμα είναι η υπόθεση μίξης (συνήθως γκαουσιανών) κατανομών πιθανότητας, όπου κάθε μίξη είναι μια ομάδα. Εδώ, θεωρούμε ως κρυφές μεταβλητές τις ετικέτες που δείχνουν την μίξη στην οποία ανήκει κάθε δεδομένο, ενώ ταυτόχρονα άγνωστες είναι και οι παράμετροι των γκαουσιανών. Τέτοια προβλήματα λύνονται με τον δημοφιλή αλγόριθμο **Expectation Maximization**. Ο αλγόριθμος αυτός είναι μία επαναληπτική μέθοδος επίλυσης της εκτίμησης μέγιστης πιθανοφάνειας όταν κάποια μεγέθη είναι άγνωστα. Συγκεκριμένα, αν κάνουμε

κάποιες παραδοχές για ένα μοντέλο με παραμέτρους $\boldsymbol{\theta}$ τότε η πιθανοφάνεια είναι:

$$L(\boldsymbol{\theta}; \boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{3.8}$$

όπου $\boldsymbol{X}$ ο πίνακας με τα δεδομένα, Z ο πίνακας με τις κρυφές μεταβλητές. Δεδομένης όμως της πολυπλοκότητας που προκύπτει λόγω του μεγάλου χώρου στον οποίο βρίσκονται οι πιθανοί πίνακες Z και οι πιθανές παράμετροι $\boldsymbol{\theta}$, ο EM αλγόριθμος εκτελεί επαναληπτικά 2 βήματα προκειμένου να βρει ένα (τοπικό) μέγιστο της συνάρτησης:

Βήμα αναμενόμενης τιμής(E step):

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}} [\log L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z})] \tag{3.9}$$

Βήμα μεγιστοποίησης (M step):

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \tag{3.10}$$

- Ένας ακόμα τρόπος ομαδοποίησης είναι το **hierarchical clustering-ιεραρχική συσταδοποίηση**. Εδώ οι ομάδες δημιουργούνται με βάση ένα μέγεθος ομοιότητας (**connectivity based**). Δηλαδή, τα δεδομένα ιεραρχικά ομαδοποιούνται σε όλο και μικρότερες (ή ολο και μεγαλύτερες-αναλόγως με το πως ξεκινά ο αλγόριθμος) ομάδες με τρόπο τέτοιο ώστε τα περιεχόμενα κάθε ομάδας να μοιάζουν μεταξύ τους και ταυτόχρονα να διαφέρουν από τα περιεχόμενα των άλλων ομάδων.
- Τέλος, άλλες δημοφιλείς μέθοδοι είναι αυτές που κάνουν την υπόθεση ότι κάθε ομάδα μπορεί να αναπαρασταθεί από ένα κεντρικό διάνυσμα-αντιπρόσωπο (**centroid-based**) όπως ο k-means.

Ασθενώς Επιβλεπόμενη Μάθηση-Weakly Supervised Learing

Η ασθενώς επιβλεπόμενη μάθηση είναι ένα από τα πιο σύγχρονα πεδία ενδιαφέροντος των ερευνητών της μηχανικής μάθησης. Ο ορισμός της δεν είναι κοινώς αποδεκτός καθώς πολλοί χρησιμοποιούν την έννοια weak supervision για να περιγράψουν εντελώς διαφορετικές μεθόδους. Γενικά, το κοινό όλων αυτών των μεθόδων είναι ότι οι αλγόριθμοι μάθησης έχουν πρόσβαση σε ένα είδος **μερικώς επισημειωμένων δεδομένων**. Ο στόχος τους είναι να ερευνηθούν οι δυνατότητες της αποφυγής της κοστοβόρας και χρονοβόρας διαδικασία της συλλογής μεγάλων ποσοτήτων πλήρως επισημειωμένων δεδομένων. Κάποιοι θεωρούν την ημιεπιβλεπόμενη μάθηση ως υποσύνολο αυτής της κατηγορίας. Θα ακολουθήσουμε και εμείς αυτήν την ταξινόμηση προκειμένου να εξηγήσουμε με όσον το δυνατόν πιο γενικό τρόπο την έννοια της ασθενούς επίβλεψης. Παρακάτω παραθέτουμε τα διάφορα προβλήματα που εμπίπτουν σε αυτή την κατηγορία ακολουθώντας την έρευνα του [25] στην οποία γίνεται ένας πολύ κομψός διαχωρισμός των περιπτώσεων όπου η επισημείωση είναι ελλιπής συγκρίνοντας πάντα με την πλήρη περίπτωση της επιβλεπόμενης μάθησης.

Τα κριτήρια που καθιστούν μία επίβλεψη πλήρη είναι 3. Συγκεκριμένα, πλήρη μάθηση έχουμε: (i) όταν το σύνολο $\mathcal{S}$ περιέχει τούπλες όπου μόνο ένα δείγμα αντιστοιχίζεται σε μόνο μία ετικέτα, (ii) όταν όλα τα δείγματα του $\mathcal{S}$ έχουν ετικέτα και (iii) όταν οι ετικέτες των νέων δειγμάτων, που δεν έχει δει ο ταξινομητής στο στάδιο της μάθησης, είναι άγνωστες, δηλαδή όταν στο στάδιο της πρόβλεψης δεν έχουμε καθόλου επίβλεψη. Όταν ένα από τα 3 κριτήρια καταργείται, τότε λέμε ότι έχουμε ασθενή επίβλεψη. Για τα πρώτα δύο η ασθενής επίβλεψη βρίσκεται στο στάδιο της μάθησης , ενώ για το τρίτο βρίσκεται στο στάδιο της πρόβλεψης.

1. **Σχέση δειγμάτων-ετικετών**. Με αυτό το κριτήριο ταξινόμησης διακρίνουμε 4 υποκατηγορίες. Συγκεκριμένα, στις τούπλες του συνόλου $\mathcal{S}$ μπορούμε να αντικαταστήσουμε είτε τα μοναδικά δείγματα (single instance-SI) με ομάδες δειγμάτων (multiple instance-MI), είτε τις μοναδικές ετικέτες(single label-SL) με ομάδες ετικετών (multiple label-ML) . Έτσι προκύπτουν τα προβλήματα SISL (που είναι το ίδιο με αυτά που αναφέραμε στην ενότητα της πλήρως επιβλεπόμενης μάθησης), SIML όπου κάθε δείγμα αντιστοιχίζεται σε παραπάνω από μία κλάσεις, MISL όπου ένα σύνολο δειγμάτων (το οποίο θα αποκαλούμε και bag) αντιστοιχίζεται σε μία κλάση και το MIML όπου ένα σύνολο δειγμάτων αντιστοιχίζεται σε ένα σύνολο ετικετών. Να σημειώσουμε εδώ ότι όταν η σχέση δειγμάτων-ετικετών δεν είναι SISL αυτό δεν σημαίνει απαραίτητα ότι η επίβλεψη είναι ασθενής. Αυτό γιατί, μπορεί σε ένα πρόβλημα να εκπαιδεύσουμε έναν ταξινομητή ο οποίος να μαθαίνει μη SISL σχέσεις. Για παράδειγμα στο [59] περιγράφονται διάφορα προβλήματα και μέθοδοι SIML προβλημάτων με πλήρη επίβλεψη όπως π.χ το [7] όπου μία σκηνη σε μία εικόνα μπορεί να ανήκει σε παραπάνω από μια κατηγορίες(π.χ ηλιοβασίλεμα και παραλία ταυτόχρονα). Ακόμα, στο [68] περιγράφεται ένα πρόβλημα MIML όπου ο ταξινομητής στοχεύει στην εξαγωγή πολλών ετικετών από πολλά αντικείμενα σε μία εικόνα. Αντίθετα, η επίβλεψη είναι ασθενής όταν στόχος του ταξινομητή είναι να εξάγει απλούστερες σχέσεις (π.χ SISL) από αυτές που περιγράφει το σύνολο $\mathcal{S}$ (π.χ MISL). Τέτοιες περιπτώσεις θα δούμε παρακάτω.

2. **Επίβλεψη στη φάση της εκπαίδευσης**. Αυτό το κριτήριο ταξινόμησης είναι ο κύριος τρόπος διαχωρισμού των διαφόρων μεθόδων ασθενούς επίβλεψης. Συγκεκριμένα, αναλόγως με τη σχέση που περιγράφει το σύνολο $\mathcal{S}$ και τη σχέση που θέλουμε να μάθει ο ταξινομητής μέσω της συνάρτησης h μπορεί να προκύψει και μία διαφορετική μέθοδος επίβλεψης. Θα απαριθμήσουμε εδώ μόνο μερικές παραπέμποντας τον αναγνώστη και πάλι στο [25] για μία πιο συγκεντρωτική παράθεση τους.

    - **Ημι-επιβλεπόμενη Μάθηση-Semi-Supervised Learing**. Το πρόβλημα αυτό δημιουργήθηκε όταν ερευνητές προβλημάτων επιβλεπόμενης μάθησης προσπάθησαν να περιορίσουν το κόστος σε χρόνο και ανθρώπινο δυναμικό της επισημείωσης μεγάλων συνόλων δεδομένων. Συγκεκριμένα, εδώ η ασθενής επίβλεψη προκύπτει γιατί το σύνολο $\mathcal{S}$ είναι $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 = \{\cup_{i=1}^{m_1}(\boldsymbol{x_i}, y_i)\} \cup \{\cup_{i=m_1+1}^{m}(\boldsymbol{x_i})\}$, δηλαδή μόνο ένα μέρος του συνόλου έχει ετικέτα. Μία μέθοδος που έχει προταθεί είναι η αυτό-εκπαίδευσή (self-training). Η γενική ιδέα(πάνω στην οποία εφαρμόστηκαν πολλές παραλλαγές) είναι ότι πρώτα εκπαιδεύεται ένας ταξινομητής

h πάνω στο σύνολο $\mathcal{S}_1$ όπου η γνώση είναι πλήρης, στη συνέχεια προβλέπεται η ετικέτα ενός δείγματος $\boldsymbol{x_j}$ από το άγνωστο κομμάτι του συνόλου $\mathcal{S}_2$ και τέλος η τούπλα $(\boldsymbol{x_j}, h(\boldsymbol{x_j}))$ προστίθεται στο $\mathcal{S}_1$ και η διαδικασία επαναλαμβάνεται. Ακόμα, παραγωγικά μοντέλα (generative) έχουν χρησιμοποιηθεί με μεγάλη επιτυχία σε τέτοια προβλήματα όπου ,όπως και στα πλήρως επιβλεπόμενα, γίνεται μία εικασία για την κατανομή πιθανότητας των δεδομένων με τη διαφορά ότι υπάρχουν κρυφές μεταβλητές οι οποίες είναι οι ετικέτες του συνόλου $\mathcal{S}_2$. Όπως αναφέραμε και στην προηγούμενη υποενότητα, ένας ιδιαίτερα αποτελεσματικός αλγόριθμος για την βελτιστοποίηση τέτοιων μοντέλων είναι ο Expectation Maximization. Τέλος, έχουν προταθεί γραφοθεωρητικές μέθοδοι, επεκτάσεις των SVMs κ.ά.

- **Μάθηση με πιθανοτικές ετικέτες-Probabilistic Label Learning**. Πρόκειται για μια κατηγορία προβλημάτων που δεν έχουν μελετηθεί επαρκώς μέχρις στιγμής, ενώ οι 2 βασικές ερευνητικές εργασίες που ασχολήθηκαν με αυτήν ([28, 10]) ασχολούνται με το SISL framework. Συγκεκριμένα, το [28] αποτελείται από ένα σύνολο $\mathcal{S} = \cup_{i=1}^{m}(\boldsymbol{x_i}, f_i(y))$, όπου $f_i(y) = \mathbb{P}[y_i = y]$ η κατανομή πιθανότητας της ετικέτας y_i πάνω στο σύνολο $\mathcal{Y}$. Δηλαδή, σε ένα δείγμα δεν αντιστοιχίζεται μία μοναδική ετικέτα y_i , αλλά μας δίνεται για κάθε ετικέτα η πιθανότητά της να είναι η πραγματική. Στο [10] για κάθε δεδομένο μας δίνεται ένα σύνολο $\mathcal{Y}_i \subseteq \mathcal{Y}$ που περιέχει κάποιες υποψήφιες ετικέτες (**candidate labels**) στις οποίες περιέχεται και η πραγματική. Να σημειώσουμε ότι η 2η περίπτωση αποτελεί υποπερίπτωση της 1ης καθώς μπορούμε να αποδώσουμε σε κάθε υποψήφια ετικέτα ίση πιθανότητα $\frac{1}{|\mathcal{Y}_i|}$ και μηδενική πιθανότητα σε αυτές που δεν θέτουν υποψηφιότητα. **Η συγκεκριμένη κατηγορία μας ενδιαφέρει ιδιαίτερα, καθώς στην εν λόγω διπλωματική συναντάμε το πρόβλημα των πιθανοτικών ετικετών αλλά σε πρόβλημα MISL, επεκτείνοντας την έρευνα σε αυτόν τον τομέα**.
- **Μάθηση Πολλαπλών παραδειγμάτων-Multiple Instance Learing**. Μελετήθηκε πρώτη φορά το 1997 στο [16] με στόχο να προβλεφθεί η δραστηριότητα ενός φαρμάκου. Συγκεκριμένα, οι επιστήμονες γνώριζαν πότε μία ουσία είναι κατάλληλη για να λειτουργήσει ως φάρμακο και πότε όχι, αλλά δεν μπορούσαν να καταλάβουν ποια από τις εναλλακτικές μοριακές δομές (σχήματα) που μπορούσε να έχει η ουσία ήταν υπεύθυνη κάθε φορά για την καλή επίδραση του φαρμάκου. Οπότε θεώρησαν κάθε κατάλληλη ουσία σαν ομάδα δειγμάτων(multiple instance) με θετική ετικέτα, όπου αυτό σήμαινε ότι τουλάχιστον μία από τις πιθανές εναλλακτικές δομές θα είχε στην πράξη την θετική ετικέτα. Στις ακατάλληλες ουσίες αποδιδόταν η αρνητική ετικέτα, πράγμα που σημαίνει ότι όλες οι εναλλακτικές δομές είχαν στην πράξη την ετικέτα αυτή. Αυτή η εργασία στην ουσία άνοιξε το πεδίο της ασθενούς μάθησης έξω από τα στενά όρια της ημι-επιβλεπόμενης. Άλλες εφαρμογές είναι ο εντοπισμός αντικειμένων σε εικόνες τοποθετώντας θετική ετικέτα σε αυτές που περιέχουν το αντικείμενο χωρίς να καθορίζεται αυστηρά η θέση στην οποία βρίσκεται ([36, 67]).

- **Μάθηση πολλαπλών παραδειγμάτων με υποψήφια διανύσματα ετικετών-Candidate Labeling Vectors Learning**. Στο [34] για κάθε **ομάδα** δεδομένων μας παρέχεται ένα σύνολο από υποψήφια διανύσματα ετικετών (**candidate labeling vectors**). Δηλαδή, για n δείγματα έχουμε m διανύσματα ετικετών n διαστάσεων, όπου η συντεταγμένη i κάθε διανύσματος εκφράζει την ετικέτα που το υποψήφιο διάνυσμα θα αποδώσει στο δείγμα i, αν τελικά το διάνυσμα αυτό επιλεγεί. Μια υποπερίπτωση αυτού είναι τα ποσοστά ετικετών (**label proportions**), όπου για κάθε ομάδα ξέρουμε σε πόσα δείγματα πρέπει να αποδοθεί κάθε ετικέτα, και οι κοινοί περιορισμοί (**mutual label constraints**), όπου γνωρίζουμε μια σχέση που περιγράφει τις ετικέτες της ομάδας (π.χ ¨όλα έχουν την ίδια ετικέτα αλλά δεν ξέρουμε ποια είναι αυτή¨-μπορεί να συμβεί άμα η ομάδα δειγμάτων είναι τα frames μιας παρακολούθησης προσώπου, ¨όλα έχουν διαφορετικές ετικέτες¨-μπορεί να συμβεί άμα η ομάδα δειγμάτων είναι μία ομάδα προσώπων σε μία φωτογραφία).
- **Άλλες κατηγορίες**. Για λόγους πληρότητας αναφέρουμε εν συντομία και κάποιες άλλες κατηγορίες όπως είναι οι **θορυβώδεις ετικέτες(noisy labels)** - σε κάθε δείγμα ανατίθεται μία ετικέτα που υποφέρει από ένα μέτρο αβεβαιότητας, και η **ημιτελής επίβλεψη(incomplete supervision)** - σε κάθε δείγμα πρέπει να αποδοθούν πολλαπλές ετικέτες (ML) αλλά οι ετικέτες των δεδομένων εκπαίδευσης είναι ελλιπείς.

3. **Επίβλεψη στη φάση της πρόβλεψης**. Σύμφωνα με αυτό το τελευταίο κριτήριο, μπορούμε να πούμε ότι ένα πρόβλημα έχει ασθενή επίβλεψη όταν τα άγνωστα στο στάδιο της εκπαίδευσης δεδομένα, παρέχονται στον ταξινομητή συνοδευόμενα από κάποιες πληροφορίες που διευκολύνουν την ταξινόμηση τους. Για παράδειγμα στο [34] χρησιμοποιείται η τεχνική των υποψήφιων διανυσμάτων ετικετών στο στάδιο της εκπαίδευσης, αλλά και στο στάδιο της πρόβλεψης. Συγκεκριμένα, ο ταξινομητής εκπαιδεύεται πάνω σε πρόσωπα από φωτογραφίες χρησιμοποιώντας ως ασθενή επίβλεψη υποψήφια διανύσματα όπως αυτά συμπεραίνονται από τα ονόματα που παρέχονται στις λεζάντες. Στο στάδιο της πρόβλεψης, τροφοδοτείται και πάλι με συνδυασμούς φωτογραφιών-λεζάντων και διαλέγει από τα υποψήφια διανύσματα ετικετών αυτό με τη μεγαλύτερη εμπιστοσύνη. Τέτοιες μορφές ασθενούς επίβλεψης, παρ΄ όλο που μπορούν να αυξήσουν κατά πολύ την επιτυχία ενός αλγορίθμου, δεν είναι πάντα εύκολο να βρεθούν και άλλωστε η ασθενής επίβλεψη έχει νόημα μόνο όταν βοηθάει στον περιορισμό της συμμετοχής του ανθρώπινου παράγοντα και του κόστους της εύρεσης κατάλληλα επισημειωμένων δεδομένων. Για αυτό το λόγο δεν υπάρχει μεγάλη γκάμα εργασιών πάνω σε αυτόν τον τομέα στη βιβλιογραφία.

Αξίζει εδώ να σημειώσουμε ότι συνδυάζοντας τις διάφορες υποκατηγορίες που προκύπτουν από τα 3 κριτήρια μπορούν να ερευνηθούν πολλά νέα ενδιαφέροντα προβλήματα ή να βελτιωθεί η επίλυση παλαιότερων, στα οποία μέχρις στιγμής δεν είχε ληφθεί υπ΄ όψιν η ασθενής επίβλεψη. Χαρακτηριστικά μπορούμε να αναφέρουμε τα προβλήματα αναγνώρισης της όρασης

υπολογιστών, όπου πολύ συχνά βίντεο και εικόνες περιγράφονται από μικρά ή μεγάλα κομμάτια κειμένου. Έτσι, ερευνώντας θεωρητικά και πρακτικά τα περιθώρια που μας δίνουν οι μέθοδοι αυτές, δημιουργούνται οι προϋποθέσεις αξιοποίησης πολύ μεγαλύτερου όγκου δεδομένων ως ¨πρώτης ύλης' για την εξέλιξη της τεχνητής νοημοσύνης.

Άλλες μορφές Μάθησης

Δεδομένων των μεγάλων απαιτήσεων που προκύπτουν από την ταχεία εξέλιξη της τεχνολογίας προέκυψαν νέα μοντέλα στον κλάδο της μηχανικής μάθησης, τα οποία δεν μπορούσαν να διαχειριστούν οι ήδη υπάρχουσες μέθοδοι. Συγκεκριμένα, ο όγκος των δεδομένων που καλούνται οι ερευνητές σήμερα να διαχειριστούν γίνεται όλο και μεγαλύτερος, ενώ πολλές φορές δεν είναι διαθέσιμος στην πλήρη του μορφή αλλά παρουσιάζεται στα συστήματα μάθησης σταδιακά. Για παράδειγμα, εταιρείες κολοσσοί στον κλάδο του διαδικτύου όπως η Google καλούνται καθημερινά να επεξεργάζονται τεράστιο όγκο δεδομένων και από αυτόν να εξάγουν χρήσιμα συμπεράσματα και να ανταποκρίνονται σε μικρό χρόνο σε αιτήματα των πελατών τους, όπως είναι η αναζήτηση ενός θέματος, ή η λήψη οδηγιών πλοήγησης από ένα γεωγραφικό σημείο σε ένα άλλο. Ακόμα, στους κλάδους της ρομποτικής και των ηλεκτρονικών παιχνιδιών είναι συνήθως απολύτως απαραίτητη η αλληλεπίδραση του συστήματος με το περιβάλλον του και σπάνια υπάρχει πλήρης γνώση αυτού κατά τη διάρκεια του σχεδιασμού και της εκπαίδευσης των αλγορίθμων. Επίσης, η αλληλεπίδραση άλλων επιστημών με την επιστήμη των υπολογιστών έχει βοηθήσει πολλές φορές στην εξέλιξη της τεχνολογίας. Για παράδειγμα, η ψυχολογία έχει δώσει πολλές φορές έμπνευση στους ερευνητές προκειμένου να μιμηθούν τον τρόπο που ο άνθρωπος μαθαίνει. Ακόμα, οι νευροεπιστήμες και η φυσική έδωσαν τη δυνατότητα της δημιουργίας πληθώρας υπολογιστικών μοντέλων,όπως είναι τα νευρωνικά δίκτυα,που είναι βασισμένα στη λειτουργία του νευρικού συστήματος και του ανθρώπινου εγκεφάλου(από τις νευροεπιστήμες) ή οι μηχανές Boltzmann και η προσομοιωμένη ανόπτηση (από τη φυσική). Τέλος, οι εξελίξεις στον κλάδο του υλικού των υπολογιστών(hardware) έχουν κάνει πλέον δυνατή την εφαρμογή στην πράξη μοντέλων με τεράστιο υπολογιστικό κόστος τα οποία δεν θα μπορούσαν παλαιότερα παρά να μείνουν στη θεωρία. Παρακάτω αναφέρουμε μερικές μόνο από τις ολοένα και εξελισσόμενες μορφές μάθησης που ξεφεύγουν από τα όρια αυτών που προαναφέραμε:

- **Ενισχυτική μάθηση-Reinforcement learning**. Σε αυτήν την οικογένεια τεχνικών το σύστημα μαθαίνει αλληλεπιδρώντας με το περιβάλλον του. Συγκεκριμένα, έχει τις ρίζες της στη μέθοδο μάθησης με επιβράβευση και τιμωρία των ανθρώπινων και γενικά των έμβιων όντων. Έτσι και το υπολογιστικό μοντέλο, στοχεύει στην μεγιστοποίηση μίας συνάρτησης του σήματος ανταμοιβής που του στέλνει το περιβάλλον του. Το σύστημα χωρίς εξωτερική καθοδήγηση, ανακαλύπτει σιγά σιγά το βέλτιστο τρόπο συμπεριφοράς μαθαίνοντας να μην υποπίπτει στα ίδια λάθη. Βρίσκει εφαρμογές σε έλεγχο ρομπότ καθώς και ηλεκτρονικά παιχνίδια.

- **Ενεργός Μάθηση-Active Learning**. Αποτελεί μία επέκταση της ημι-επιβλεπόμενης μάθησης. Συγκεκριμένα, ο αλγόριθμος έχει στη διάθεση του ετικέτες μόνο για ένα

μέρος των δεδομένων του και σταδιακά επιλέγει δεδομένα που είναι όσο γίνεται πιο αντιπροσωπευτικά και ζητάει από κάποιον επισημειωτή να του παρέχει ετικέτες για αυτά. Ταυτόχρονα, σε κάθε βήμα του αλγορίθμου μαθαίνει να ταξινομεί τα δεδομένα όλο και καλύτερα. Έχει μεγάλη απήχηση τα τελευταία χρόνια, καθώς επιταχύνει τη διαδικασία της επισημείωσης και βοηθάει στην επιλογή αντιπροσωπευτικών δεδομένων.

- **Online Machine Learning**. Εδώ τα δεδομένα παρουσιάζονται στο σύστημα μάθησης ακολουθιακά, με αποτέλεσμα ο αλγόριθμος να πρέπει να ανανεώνει τη συνάρτηση πρόβλεψης κάθε φορά που του παρέχεται ένα νέο δεδομένο. Είναι απαραίτητο όταν δεν είναι εφικτό υπολογιστικά να βρεθεί λύση στο πρόβλημα παρέχοντας όλο το σύνολο δεδομένων ή όταν τα δεδομένα από τη φύση τους παρέχονται ακολουθιακά και για κάθε ένα από αυτά η απόφαση πρέπει να λαμβάνεται τη στιγμή της εμφάνισης του (π.χ πρόβλεψη τιμής μετοχής ή καταχώρηση μηνύματος σε ανεπιθύμητη αλληλογραφία)

- **Βαθιά μηχανική μάθηση-Deep Learning**. Τέλος, παρουσιάζουμε εδώ συνοπτικά τον κλάδο που έχει συγκεντρώσει το μεγαλύτερο ενδιαφέρον των ερευνητών τα τελευταία χρόνια λόγω των εντυπωσιακών αποτελεσμάτων που έχει επιδείξει σε τεράστια γκάμα προβλημάτων. Η απαρχή της εξέλιξης του ήταν η εξέλιξη της τεχνολογίας του hardware που κατέστησε εφικτή την εκπαίδευση των μοντέλων σε λογικά χρονικά πλαίσια (της τάξης των ωρών ή ημερών). Αυτό που διαχωρίζει επί της αρχής όλα τα μοντέλα βαθιάς μάθησης, δηλαδή τα βαθιά νευρωνικά δίκτυα, από τα απλά (ρηχά) νευρωνικά δίκτυα είναι ο μεγάλος αριθμός επιπέδων νευρώνων που χρησιμοποιούν. Έτσι, παρ΄ όλο που η λειτουργία τους δεν έχει εξηγηθεί πλήρως, οι ερευνητές θεωρούν ότι πετυχαίνουν τη μάθηση πολλαπλών επιπέδων αναπαραστάσεων, δομημένων ιεραρχικά. Δηλαδή, όσο κινούμαστε σε βάθος μίας αρχιτεκτονικής ενός τέτοιου δικτύου οι αναπαραστάσεις γίνονται όλο και πιο ¨αφηρημένες' ,δηλαδή περιγράφουν όλο και πιο σύνθετα αντικείμενα, και η μάθηση τους γίνεται από τις αναπαραστάσεις του αμέσως χαμηλότερου επιπέδου. Έτσι, προσομοιώνεται και η λειτουργία του ανθρώπινου εγκεφάλου, όπου θεωρείται ότι η πληροφορία, αφού εισαχθεί από τα αισθητηριακά όργανα, επεξεργάζεται ιεραρχικά σε διαφορετικούς υποδοχείς της, μέχρι να οδηγηθεί στο κέντρο λήψης αποφάσεων. Για παράδειγμα, στην αναγνώριση ενός αντικειμένου, οι πρώτες αναπαραστάσεις που μαθαίνονται μπορεί να είναι απλές ακμές, στη συνέχεια να μαθαίνεται η υφή, κατόπιν μικρά μοτίβα και σχήματα, μετά τα βασικά μέρη που συνθέτουν ένα αντικείμενο και στο τέλος το ίδιο το αντικείμενο.

  Στον κλάδο αυτό έχουν χρησιμοποιηθεί πολλών διαφορετικών ειδών αρχιτεκτονικές, όπως είναι τα **Αναδρομικά Δίκτυα(Recurrent Neural Networks-RNNs**,τα **Συνελικτικά Δίκτυα(Convolutional Neural Networks-CNNs)**, **οι μηχανές Boltzmann(Restricted Boltzmann Machines)** και **οι αυτο-κωδικοποιητές (Autoencoders)**. Συνήθως, τα στοιχεία που συναποτελούν τα δίκτυα (και άρα και οι τελικές συναρτήσεις που υλοποιούν) είναι μη γραμμικά, και η ελαχιστοποίηση της συνάρτησης κόστους είναι μία μη κυρτή βελτιστοποίηση. Έχουν χρησιμοποιηθεί σε πληθώρα προβλημάτων όπως είναι η αναγνώριση προτύπων σε προβλήματα όρασης υπολογιστών,

επεξεργασίας ήχου και επεξεργασίας φυσικής γλώσσας, σε προβλήματα παλινδρόμησης, σε αποθορυβοποίηση, σε εξαγωγή χαρακτηριστικών, στην ενισχυτική μάθηση κ.ά.

Στην παρούσα διπλωματική θα χρησιμοποιήσουμε βαθιά μηχανική μάθηση μέσω προεκπαιδευμένων συνελικτικών δικτύων σε μεγάλες βάσεις δεδομένων προκειμένου να εξάγουμε υψηλού επιπέδου χαρακτηριστικά από τα δείγματα μας. Η εξαγωγή χαρακτηριστικών στην όραση υπολογιστών πλέον γίνεται σχεδόν καθολικά με deep networks καθώς έχουν αποδείξει στην πράξη την ανωτερότητα τους σε σχέση με τα παλιά hand-crafted features. Περισσότερες πληροφορίες για τα συνελικτικά δίκτυα που χρησιμοποιήσαμε στην εργασία αυτή θα αναφέρουμε στο κεφάλαιο 6.

Κεφάλαιο 4

Μαθηματικός Φορμαλισμός του Προβλήματος - Προϋπάρχουσες και Νέες Μέθοδοι

4.1 Εισαγωγή

Στο παρακάτω κεφάλαιο παρουσιάζουμε τη μαθηματική μοντελοποίηση του προβλήματος, όπως αυτή περιγράφεται σε προηγούμενες ερευνητικές εργασίες, καθώς και την προσαρμογή της και τις επεκτάσεις της στο πρόβλημα που προσεγγίζει η παρούσα διπλωματική. Το ενδιαφέρον μας επικεντρώνεται σε προβλήματα αναγνώρισης στην όραση υπολογιστών και συγκεκριμένα σε βίντεο, όπου η επίβλεψη γίνεται με ασθενή τρόπο. Ερευνούμε τις δυνατότητες επίβλεψης που μπορεί να μας παράσχει ένα συνοδευτικό στο βίντεο κείμενο. Συγκεκριμένα, σπάνια οι 2 'τροπικότητες' (βίντεο και κείμενο) είναι απόλυτα χρονικά ευθυγραμμισμένες και ακόμα σπανιότερα χωρικά. Αυτό σημαίνει ότι δεν μπορούμε να συνθέσουμε ένα πλήρως επιβλεπόμενο σύνολο $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$, καθώς δεν υπάρχει '1-1' αντιστοιχία κάθε αντικειμένου που υπάρχει σε ένα βίντεο (από το οποίο εξάγεται το $\boldsymbol{x}_i$) με ένα αντίστοιχο αντικείμενο στο κείμενο (από το οποίο εξάγεται το y_i) . Γίνεται επομένως εύκολα αντιληπτό ότι πρέπει αφενός να αντιμετωπιστεί το πρόβλημα εξαγωγής των $\boldsymbol{x}_i$ και y_i και αφετέρου να σχεδιαστεί ένας αλγόριθμος μάθησης που να διαχειρίζεται την επίβλεψη με κάποιον από τους τρόπους που αναφέρθηκαν στο κεφάλαιο 3. Στο παρακάτω κεφάλαιο παρουσιάζεται ο μαθηματικός φορμαλισμός των προτεινόμενων μοντέλων μάθησης, χωρίς να δοθεί έμφαση στη μεθοδολογία εξαγωγής των στοιχείων που συναποτελούν το σύνολο $\mathcal{S}$, η οποία θα παρουσιαστεί εκτενώς στα κεφάλαια 5 και 6. Τέλος, να σημειώσουμε ότι ο συμβολισμός που θα χρησιμοποιηθεί εδώ είναι τέτοιος ώστε να συμφωνεί με την υπάρχουσα βιβλιογραφία και ίσως σε κάποια σημεία να ξεφεύγει από τα αυστηρά όρια που ορίσαμε στο κεφάλαιο 1.

4.2 Παρουσίαση του προβλήματος

Τα αντικείμενα που θα μας απασχολήσουν στην εν λόγω εργασία είναι τα ανθρώπινα πρόσωπα καθώς και οι δράσεις που εκτελούν. Αναζητούμε δηλαδή απαντήσεις στα ερωτήματα «Ποιος εμφανίζεται;» και «Τι δράση κάνει;» σε μία ακολουθία εικόνων και συγκεκριμένα σε μία ταινία. Σκοπός είναι να κάνουμε μία πλήρη καταγραφή του βίντεο απαντώντας σε αυτά τα ερωτήματα με τη θέση στο χώρο και στο χρόνο όλων των αντικειμένων. Θεωρούμε ότι δεν υπάρχει πρότερη γνώση για τα υπό αναγνώριση αντικείμενα και η μόνη πηγή πληροφοριών για τις ετικέτες τους παρέχεται από το συνοδευτικό κείμενο το οποίο στην περίπτωσή μας είναι το σενάριο. Το πρόβλημα μπορεί να χωριστεί σε τρία υποπροβλήματα: Αυτό του εντοπισμού στο χώρο του βίντεο και της εξαγωγής χαρακτηριστικών, αυτό της εξαγωγής πληροφορίας από το κείμενο και αυτό της μάθησης.

Χωρίς να επεκταθούμε εδώ αναφέρουμε επιγραμματικά αυτά τα προβλήματα για να δώσουμε μία καλύτερη διαίσθηση στον αναγνώστη.

Εντοπισμός και εξαγωγή οπτικών χαρακτηριστικών

Η ανάλυση του προβλήματος αυτού θα γίνει με λεπτομέρεια στο κεφάλαιο 5.

Στο κομμάτι του βίντεο πρέπει αρχικά να απομονωθούν τα υποψήφια προς αναγνώριση δείγματα από την περιττή πληροφορία. Στόχος αυτού του βήματος είναι να εντοπίσουμε χωροχρονικά κάθε οπτικό αντικείμενο v_i ($\mathcal{V}$ το σύνολο τους) δηλαδή να βρούμε τη θέση του ως $\{\mathcal{D}_k^i\}_{k=k_1}^{k_2}$,με $\mathcal{D}_k^i$ το σύνολο των σημείων του αντικείμενου πάνω στον χώρο που ορίζει η εικόνα και k η χρονική στιγμή, καθώς και τα χρονικά όρια k_1^i, k_2^i. Στη συνέχεια, για κάθε οπτικό αντικείμενο μπορεί να εξαχθεί ένα διάνυσμα χαρακτηριστικών $\boldsymbol{x}_i$.

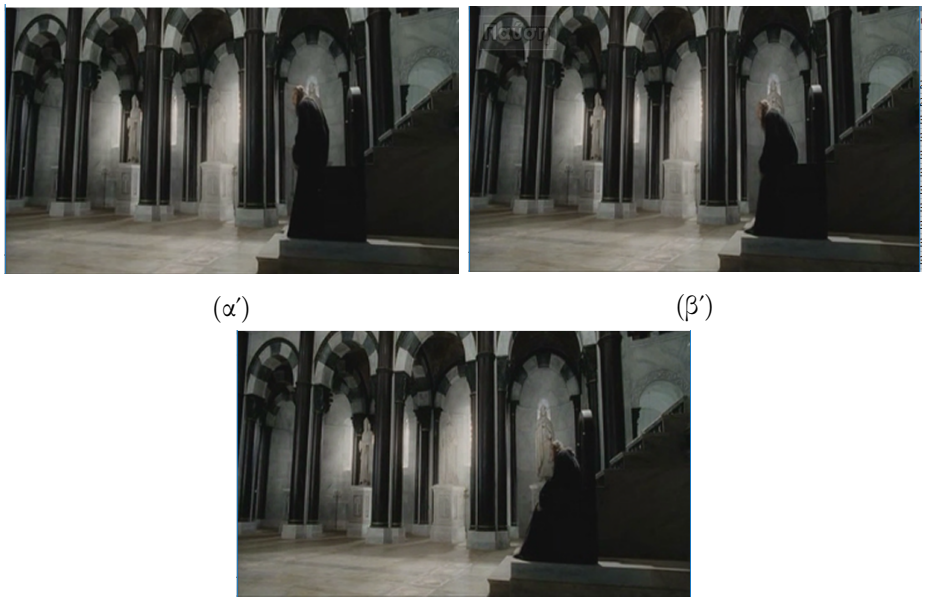
Σχήμα 4.1: Στιγμιότυπο ενός προσώπου v_i όπως αυτό έχει απομονωθεί από το στάδιο του εντοπισμού μία χρονική στιγμή k. Η πραγματική του ετικέτα είναι το όνομα Gandalf. Το σύνολο $\mathcal{D}_k^i$ ορίζεται από το παράθυρο με το κόκκινο χρώμα

Εξαγωγή πληροφορίας από το κείμενο

Η ανάλυση του προβλήματος αυτού θα γίνει με λεπτομέρεια στο κεφάλαιο 6.

Στόχος αυτού του βήματος είναι να δώσουμε στον αλγόριθμο μάθησης όσο γίνεται περισσότερη πληροφορία για τα υπό αναγνώριση αντικείμενα προκειμένου να βρει τις κρυφές τους ετικέτες. Ιδανικά δηλαδή θέλουμε για κάθε v_i να βρούμε την αντίστοιχη ετικέτα y_i που το περιγράφει. Στην πραγματικότητα όμως είναι πολύ δύσκολο να βρούμε με αυστηρό τρόπο αφενός την τιμή του y_i και αφετέρου τη θέση χωρικά και χρονικά του αντικειμένου που περιγράφει. **Αναλόγως με τη μεθοδολογία εξόρυξης της πληροφορίας προκύπτουν και οι αμφισημίες που πρέπει να επιλύσουμε στο στάδιο της μάθησης**. Αυτό που μπορούμε να πετύχουμε είναι να απομονώσουμε κάποιες φράσεις / γλωσσικά αντικείμενα w_j ($\mathcal{W}$ το σύνολο τους) που περιγράφουν ένα η περισσότερα αντικείμενα v_i καθώς και κάποια χαλαρά χρονικά όρια l_1^j, l_2^j. Η θέση στην εικόνα αυτού που περιγράφει το w_j δεν μπορεί να καθοριστεί επαρκώς [1] ενώ, όπως θα δούμε αναλυτικά στο κεφάλαιο 6, δεν είναι πάντα εύκολο να βρεθεί απευθείας η ετικέτα που κρύβεται μέσα στο w_j. Συγκεκριμένα, θεωρούμε ότι από κάθε w_j μπορούμε να συμπεράνουμε μόνο μία κατανομή πιθανότητας $f_j(y) = \mathbb{P}[y_j = y]$ πάνω στο σύνολο ετικετών $\mathcal{Y}$. Να σημειώσουμε εδώ ότι για να γίνει αυτό χρειαζόμαστε ένα σύνο-

[1] Κάποιες φορές μπορεί να περιγράφεται ή να υπονοείται η θέση του αντικείμενου (π.χ "Το μπουκάλι βρίσκεται πάνω στο τραπέζι"- δίνεται πληροφορία για τη θέση του μπουκαλιού μέσω της σχετικής του θέσης με το τραπέζι) αλλά είναι σπάνιο και δύσκολο να απομονωθεί τέτοια πληροφορία οπότε δεν λαμβάνεται υπ' όψη στην συγκεκριμένη εργασία.



(α′) (β′)

(γ′)

Σχήμα 4.2: (α′,β′,γ′) :Στιγμιότυπα της δράσης v_i σε 3 χρονικές στιγμές k_1, k_2, k_3 όπως αυτή έχει απομονωθεί από το στάδιο του εντοπισμού. Η πραγματική ετικέτα είναι η φράση sitting down. Το σύνολο $\mathcal{D}_k^i$ εδώ επιλέγεται να είναι ολόκληρη η εικόνα τη στιγμή k. Εναλλακτικά, μπορεί να επιλεχθεί το σύνολο που ορίζεται από το σώμα του ανθρώπου που εκτελεί τη δράση.

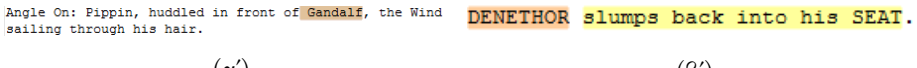
λο ετικετών $\mathcal{Y}$ που να έχει μία ¨1-1' αντιστοίχηση μία ετικέτας με μία γλωσσική περιγραφή δηλαδή να αποτελείται από τούπλες $\{y, t\}$ όπου y μία ακέραια τιμή από το 1 μέχρι το $|\mathcal{Y}|$ που δεικτοδοτεί τις ετικέτες και t μία λέξη, μία φράση ή πρόταση. Η κατανομή πιθανότητας προκύπτει ελέγχοντας την σημασιολογική ομοιότητα του w_j με κάθε t, δηλαδή εξετάζοντας κατά πόσο τα w_j, t σημαίνουν το ίδιο πράγμα. Όσον αφορά τα χρονικά όρια l_1^j, l_2^j αυτά στη συγκεκριμένη εργασία εξασφαλίζονται χρησιμοποιώντας ένα ακόμα συνοδευτικό κείμενο, αυτό των **υποτίτλων**. Η μεθοδολογία αυτή αποτελεί πάγια τακτική στην βιβλιογραφία και παρουσιάστηκε πρώτη φορά στο [19]. Συγκεκριμένα, ευθυγραμμίζει τα 2 συνοδευτικά κείμενα και κάθε πρόταση του σεναρίου κληρονομεί τα χρονικά όρια της πρότασης των υποτίτλων με την οποία έχει ευθυγραμμιστεί.

Μάθηση

Οι αμφισημίες που πρέπει να επιλύσουμε προκειμένου να κατηγοριοποιήσουμε τα δεδομένα είναι δύο:

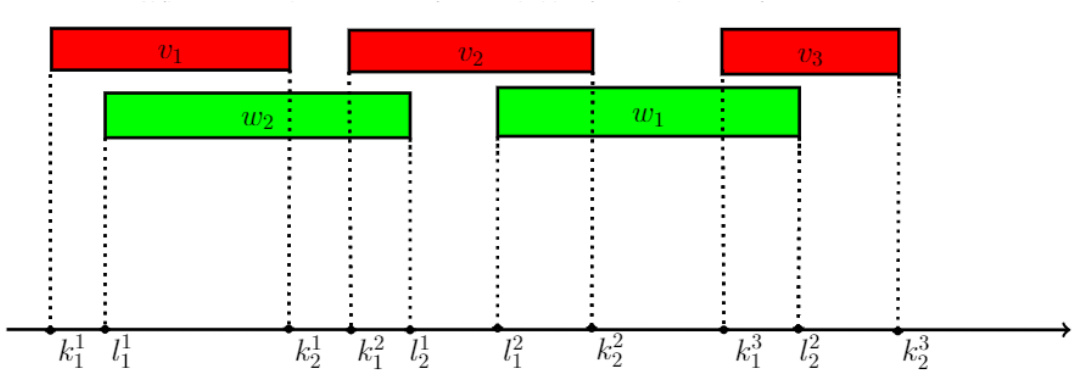
- Ποιο είναι το w_j που περιγράφει κάθε v_i;
- Ποιο είναι το y_j που υπονοείται από το w_j;

Για να απαντήσουμε στο πρώτο ερώτημα μπορούμε να μοντελοποιήσουμε την ασθενή επίβλεψη ως **μάθηση πολλαπλών παραδειγμάτων**. Στο σχήμα 4.4 δείχνουμε ένα α-



(α′) (β′)

Σχήμα 4.3: (α′): Το τμήμα του κειμένου που περιέχει τη λέξη w_j που περιγράφει το αντικείμενο-πρόσωπο v_i του σχήματος 4.1. Το γλωσσικό αντικείμενο είναι η λέξη Gandalf. Εδώ η ομοιότητα με την πραγματική ετικέτα Gandalf είναι προφανώς 1 και 0 για τις υπόλοιπες. (β′): Το τμήμα του κειμένου που περιέχει τη λέξη w_j που περιγράφει το αντικείμενο-δράση v_i του σχήματος 4.2. Το γλωσσικό αντικείμενο είναι η φράση slumps back into his SEAT. Ενδεικτικά αναφέρουμε ότι η ομοιότητα του w_j με την πραγματική ετικέτα sitting down είναι 0.65 ενώ με την ετικέτα sitting up είναι 0.27.



Σχήμα 4.4: Στιγμιότυπο από την ευθυγράμμιση των 2 τροπικοτήτων (βίντεο και κειμένου)

πόσπασμα των ροών του βίντεο και του κειμένου. Μέσα από αυτό μπορούν να γίνουν πιο εύκολα κατανοητά τα παρακάτω. Συγκεκριμένα:

1. Μπορούμε να θεωρήσουμε ότι σε κάθε w_j αντιστοιχίζονται όλα τα v_i που εμπίπτουν στην χρονική διάρκεια του. Δηλαδή κατασκευάζουμε ένα σύνολο $\mathcal{X}_j = \{\boldsymbol{x_i} | v_i \in \mathcal{V}, overlap(v_i, w_j) \neq 0\}$ για κάθε w_j και τελικά το σύνολο εκπαίδευσης είναι $\mathcal{S} = \bigcup_{w_j \in \mathcal{W}} \{(\mathcal{X}_j, w_j)\}$.

2. Εναλλακτικά, μπορούμε να επεκτείνουμε το φορμαλισμό της μάθησης πολλαπλών παραδειγμάτων δίνοντας την δυνατότητα στα σύνολα δειγμάτων (bags) να είναι ασαφή σύνολα (fuzzy), προκειμένου να ενσωματώσουμε την πληροφορία του μεγέθους της επικάλυψης μεταξύ των v_i, w_j. Σε αυτήν την περίπτωση έχουμε $\mathcal{X}_j = \{(\boldsymbol{x_i}, \mu(\boldsymbol{x_i})) | v_i \in \mathcal{V}, \mu(\boldsymbol{x_i}) = g(overlap(v_i, w_j))\}$, δηλαδή ορίζουμε κάθε $\mathcal{X}_j$ να περιέχει όλα τα $\boldsymbol{x_i}$ συνοδευόμενα από μία συνάρτηση συμμετοχής στο σύνολο, η οποία ορίζεται από μία αύξουσα συνάρτηση g της επικάλυψης του αντίστοιχου v_i με το w_j. Να σημειώσουμε ότι η επικάλυψη $overlap(v_i, w_j)$ πρέπει να εισάγεται στην συνάρτηση g ως ποσοστό της χρονικής διάρκειας του υπό αναγνώριση αντικειμένου v_i. Έτσι, μπορούμε να δώσουμε ένα ακόμα χαρακτηριστικό στη g. Συγκεκριμένα, πρέπει $g(0) = 0$ και $g(1) = 1$ έτσι ώστε αν ένα v_i δεν επικαλύπτεται με ένα w_j τότε η συμμετοχή του στο σύνολο να είναι μηδενι-

κή, ενώ αν ένα v_i περιέχεται πλήρως στο w_j τότε η συμμετοχή του να είναι η μέγιστη (1). Η περίπτωση των ασαφών συνόλων ανάγεται σε αυτήν των απλών αν ορίσουμε την $g = \mathbb{1}_{[overlap(v_i,w_j)\neq 0]}$.

Για να απαντήσουμε στο δεύτερο ερώτημα μπορούμε να μοντελοποιήσουμε την ασθενή επίβλεψη και πάλι με 2 τρόπους:

- Ως **μάθηση με ντετερμινιστικές ετικέτες**. Συγκεκριμένα, μπορούμε να επιλέξουμε για κάθε w_j την ετικέτα y με τη μεγαλύτερη νοηματική ομοιότητα και να κατασκευάσουμε ένα σύνολο εκπαίδευσης $\mathcal{S} = \bigcup_{w_j \in \mathcal{W}} \{(\mathcal{X}_j, y_j)\}$ με $y_j = argmax_y f_j(y)$.
- Ως **μάθηση με πιθανοτικές ετικέτες**. Συγκεκριμένα, σε κάθε σύνολο $\mathcal{X}_j$ αποδίδουμε την κατανομή πιθανότητας f που προαναφέραμε και σχηματίζουμε το τελικό σύνολο $\mathcal{S} = \bigcup_{w_j \in \mathcal{W}} \{(\mathcal{X}_j, f_j(y))\}$, όπου $f_j(y) = \mathbb{P}[y_j = y]$ και y_j η ετικέτα που υπονοείται από το w_j. Έτσι, το σενάριο μάθησης που χρησιμοποιούμε είναι ένα είδος μάθησης **πολλαπλών παραδειγμάτων με πιθανοτικές ετικέτες**.

Να σημειώσουμε εδώ ότι μία εναλλακτική προσέγγιση είναι η θεώρηση κάθε παραδείγματος ξεχωριστά αποδίδοντας του μία πιθανοτική ετικέτα η οποία θα προέκυπτε αφενός από την επικάλυψη του με κάθε w_j και αφετέρου από τις κατανομές f_j. Προτιμήθηκε όμως η ιδέα των πολλαπλών παραδειγμάτων προκειμένου να ενσωματώσουμε την πληροφορία ότι κάθε w_j θα αναφέρεται σίγουρα (ντετερμινιστικά) σε τουλάχιστον ένα από τα v_i με τα οποία ευθυγραμμίζεται. Η αποτελεσματικότητα της εναλλακτικής μεθόδου θα εξεταστεί σε μελλοντική εργασία.

Προκειμένου οι αμφισημίες αυτές να επιλυθούν χρειάζεται να αναπτυχθεί ένας αλγόριθμος ο οποίος να λαμβάνει υπ΄ όψη του αφενός την κρυφή δομή των δεδομένων και αφετέρου την πληροφορία που μας παρέχει το σύνολο $\mathcal{S}$. Παρ΄ όλο που τα σενάρια μάθησής μας καθορίστηκαν από τον τρόπο που εξαγάγαμε την πληροφορία, οι αλγόριθμοι έχουν καθολική εφαρμογή, καθώς χρειάζεται να γνωρίζουν μόνο την τελική μορφή του συνόλου εκπαίδευσης $\mathcal{S}$ και τίποτα από τα βήματα που οδηγούν στην κατασκευή του. Έτσι, αν κανείς διαθέτει ένα διακριτό σύνολο ετικετών $\mathcal{Y}$ και μοντελοποιήσει τη μάθηση του με οποιονδήποτε από τους τρόπους που προαναφέρθηκαν, τότε, ανεξάρτητα από τη φύση των διανυσμάτων χαρακτηριστικών ή των ετικετών, μπορεί να εφαρμόσει την μεθοδολογία μάθησης που παρουσιάζουμε στις επόμενες ενότητες.

4.3 Προϋπάρχοντες Αλγόριθμοι μάθησης

4.3.1 DIFFRAC

Παρακάτω περιγράφουμε το μοντέλο που εισήγαγαν οι Bach και Harchaoui στο [1] με το όνομα DIFFRAC-Discriminative and Flexible Framework for Clustering. Πρόκειται για μια μέθοδο ομαδοποίησης(clustering) βασισμένη σε μια γραμμική διαμέριση των ομάδων-κλάσεων. Επίσης, είναι ένας αλγόριθμος ανάκτησης κρυφών μεταβλητών μέσα από την επίλυση ενός κυρτού προγράμματος. Αυτό, τον καθιστά ιδιαίτερα ελκυστικό σε σχέση με την πλειοψηφία τον

αλγορίθμων αυτής της κατηγορίας, οι οποίοι επιλύονται με την μη κυρτή μέθοδο Expectation Maximization. Για αυτό το λόγο, χρησιμοποιούμε το DIFFRAC ως βάση για να σχεδιάσουμε τους αλγορίθμους μας.

Έστω το πρόβλημα ομαδοποίησης (clustering) N δεδομένων σε P κλάσεις και το σύνολο $\mathcal{P}$ των κλάσεων (δηλαδή $P = |\mathcal{P}|$ - χρησιμοποιείται αντί του $\mathcal{Y}$ για λόγους συμφωνίας με άλλες εργασίες). Κάθε δεδομένο $\boldsymbol{x_i}$ με $i \in \mathcal{N} = \{1, 2...N\}$ ανήκει στον χώρο $\mathbb{R}^{1\times D}$. Ορίζουμε τον πίνακα $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ με γραμμές τα $\boldsymbol{x_i}$. Κάθε δείγμα $\boldsymbol{x_i}$ ανήκει σε μία κλάση του $\mathcal{P}$. Ορίζουμε τη μεταβλητή $\boldsymbol{z_i}$ με $i \in \mathcal{N}$ η οποία ανήκει στο χώρο $\{0,1\}^{1\times P}$ με $\boldsymbol{z_i} \cdot \mathbf{1}_P = 1$, δηλαδή μία δυαδική μεταβλητή διάστασης P που παίρνει την τιμή 1 μόνο σε μία θέση p αν και μόνο αν το δεδομένο ανήκει στην κλάση p. Όμοια με τα $\boldsymbol{x_i}$, ορίζουμε την μεταβλητή $\boldsymbol{Z} \in \mathbb{R}^{N\times P}$ με γραμμές τα $\boldsymbol{z_i}$ καθώς και το σύνολο των πινάκων δεικτριών $\mathcal{Z}_{N,P} = \{Z \in \{0,1\}^{N\times P} | Z \cdot \mathbf{1}_P = \mathbf{1}_N\}$. Η μεταβλητή αυτή είναι κρυφή (latent variable) καθώς δεν έχουμε πρόσβαση στο ground truth των δεδομένων.

Ο σκοπός είναι να ανακτήσουμε τις τιμές της κρυφής μεταβλητής και ταυτόχρονα να εκπαιδεύσουμε έναν ταξινομητή $h : \mathbb{R}^D \to \mathcal{Z}_{1,P}$, ο οποίος θα δέχεται σαν είσοδο ένα διάνυσμα χαρακτηριστικών-δεδομένο διάστασης D και θα επιστρέφει το διάνυσμα δείκτρια της κλάσης που ανήκει το δεδομένο. Μπορούμε να επιλέξουμε τον ταξινομητή h ως εξής:

$$
\begin{aligned}
f(\boldsymbol{x}) &= (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..f_P(\boldsymbol{x})) \\
h(\boldsymbol{x}) &= (h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), ..h_P(\boldsymbol{x})) \\
h_j(\boldsymbol{x}) &= \begin{cases} 1 & j = \arg\max\limits_i f_i(\boldsymbol{x}) \\ 0 & \text{αλλιώς} \end{cases} .
\end{aligned}
$$

Όπου $f : \mathbb{R}^D \to \mathbb{R}^P$. Αν οι κλάσεις των δεδομένων ήταν γνωστές τότε με την αντικατάσταση της h από την f θα μπορούσαμε να μετατρέψουμε το πρόβλημα ταξινόμησης σε πρόβλημα παλινδρόμησης (regression). Έτσι επιλέγοντας κατάλληλη συνάρτηση κόστους $\ell : \mathcal{Z}_{1,P} \times \mathbb{R}^D \to \mathbb{R}$ η συνάρτηση f θα μπορούσε να βρεθεί μέσω ενός αλγορίθμου ERM-empirical risk minimization ως εξής:

$$\min_f \frac{1}{N} \sum_{i\in\mathcal{N}} \ell(\boldsymbol{z_i}, f(\boldsymbol{x_i}))$$

Ή καλύτερα, προσθέτοντας έναν όρο κανονικοποίησης για να αποφύγουμε τυχόν υπερπροσαρμογή (overffiting) του ταξινομητή στα δεδομένα, η f μπορεί να βρεθεί μέσα από έναν αλγόριθμο RLM-regularized loss minimization και ειδικότερα, από τη λύση ενός προβλήματος παλινδρόμησης κορυφογραμμής (ridge regression):

$$\min_f \frac{1}{N} \sum_{i\in\mathcal{N}} \ell(\boldsymbol{z_i}, f(\boldsymbol{x_i})) + \lambda\Omega(f)$$

Όπου $\lambda \geq 0$ η παράμετρος κανονικοποίησης, $\Omega : \mathcal{F} \to \mathbb{R}$ και $\mathcal{F}$ η κλάση συναρτήσεων από τις οποίες προέρχεται η f.

Επεκτείνοντας την τελευταία εξίσωση στο δικό μας πρόβλημα των άγνωστων κλάσεων, η αντικειμενική συνάρτηση ελαχιστοποιείται και ως προς τα $\boldsymbol{z_i}$, δηλαδή:

$$\min_{Z,f} \frac{1}{N} \sum_{i \in \mathcal{N}} \ell(\boldsymbol{z_i}, f(\boldsymbol{x_i})) + \lambda \Omega(f)$$

Αν θεωρήσουμε ότι τα δεδομένα έχουν προβληθεί σε ένα χώρο όπου οι κλάσεις είναι γραμμικά διαχωρίσιμες (οι επιφάνειες διαχωρισμού για κάθε ζεύγος κλάσεων είναι υπερεπίπεδα) τότε η συνάρτηση f μπορεί να πάρει την εξής μορφή:

$$f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{w} + \boldsymbol{b}, \boldsymbol{w} \in \mathbb{R}^{D \times P}, \boldsymbol{b} \in \mathbb{R}^{1 \times P} \tag{4.1}$$

Τέλος, αν ορίσουμε την συνάρτηση ℓ ως το τετραγωνικό σφάλμα και τον όρο κανονικοποίησης ως την $\mathcal{L}_2$ νόρμα των w ,όπως συνηθίζεται σε προβλήματα ridge regression τότε το πρόβλημα που λύνουμε παίρνει την μορφή:

$$\min_{Z,\boldsymbol{w},\boldsymbol{b}} \frac{1}{2N} \|\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{w} - \boldsymbol{1}_N \boldsymbol{b}\|_F^2 + \frac{\lambda}{2} Tr(\boldsymbol{w}^T \boldsymbol{w}) \tag{4.2}$$

Κρατώντας σταθερό το $\boldsymbol{Z}$ μπορούμε να βρούμε την ελάχιστη τιμή της συνάρτησης πάνω στα $\boldsymbol{w}, \boldsymbol{b}$ σε κλειστή μορφή. Η εύρεση των συντελεστών με αυτόν τον τρόπο είναι ευρέως γνωστή ιδιαίτερα στη Στατιστική και για αυτό δεν θα επεκταθούμε περαιτέρω. Συνήθως οι τύποι που συναντάμε στην διεθνή βιβλιογραφία αφορούν γραμμική παλινδρόμηση για πίνακες δείκτριες, δεν επεκτείνονται δηλαδή συνήθως στην παλινδρόμηση κορυφογραμμής ([22]), αλλά είναι πολύ εύκολο να βρούμε τα σημεία ελαχίστου απλά βρίσκοντας που μηδενίζεται η κλίση, δηλαδή το gradient της συνάρτησης κόστους. Οι βέλτιστες τιμές των 2 μεγεθών είναι : $\boldsymbol{w}^* = (\boldsymbol{X}^T \boldsymbol{\Pi}_N \boldsymbol{X} + N\lambda \boldsymbol{I}_D)^{-1} \boldsymbol{X}^T \boldsymbol{\Pi}_N \boldsymbol{Z}$ και $\boldsymbol{b}^* = \frac{1}{N} \boldsymbol{1}_N^T (\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{w}^*)$, όπου $\boldsymbol{I}_N$ ο μοναδιαίος πίνακας διάστασης N και $\boldsymbol{\Pi}_N = \boldsymbol{I}_N - \frac{1}{N} \boldsymbol{1}_N \boldsymbol{1}_N^T$ ο πίνακας centering[2]. Έτσι, αντικαθιστώντας με τις τιμές αυτές στήν εξίσωση (4.2) παίρνουμε:

$$\min_{Z} \frac{1}{2} (\boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{A}(\boldsymbol{X}, \lambda)) \quad \text{με} \quad \boldsymbol{A}(\boldsymbol{X}, \lambda) = \frac{1}{N} \boldsymbol{\Pi}_N (\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{\Pi}_N \boldsymbol{X} + N\lambda \boldsymbol{I}_D)^{-1} \boldsymbol{X}^T) \boldsymbol{\Pi}_N \tag{4.3}$$

Για να εκφραστεί ο πίνακας $\boldsymbol{A}(\boldsymbol{X}, \lambda)$ αρκεί ο Gram πίνακας $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^T$ ως εξής

$$\boldsymbol{A}(\boldsymbol{K}, \lambda) = \lambda \boldsymbol{\Pi}_N (\tilde{\boldsymbol{K}} + N\lambda \boldsymbol{I}_N)^{-1} \boldsymbol{\Pi}_N \qquad \text{με} \qquad \tilde{\boldsymbol{K}} = \boldsymbol{\Pi}_N \boldsymbol{K} \boldsymbol{\Pi}_N \tag{4.4}$$

Οι αποδείξεις για τα (4.3), (4.4) προκύπτουν από διαδοχικές πράξεις μεταξύ πινάκων και εφαρμογές ιδιοτήτων οι οποίες δεν παρουσιάζουν ιδιαίτερο ενδιαφέρον, οπότε παραλείπονται.

Με τη χρήση του λήμματος (4.4) μπορούμε να χρησιμοποιήσουμε οποιονδήποτε θετικά-ημιορισμένο πυρήνα στην θέση του $\boldsymbol{X}\boldsymbol{X}^T$. Δεδομένου του ότι ο πίνακας Gram εκφράζει τα

[2] Ο πίνακας αυτός εξασφαλίζει ότι η παλινδρόμηση δεν επηρεάζεται από την απόσταση των $\boldsymbol{x_i}$ από την αρχή των αξόνων. Υπάρχει στη λύση λόγω του όρου πόλωσης $\boldsymbol{b}$

εσωτερικά γινόμενα όλων των πιθανών συνδυασμών των διανυσμάτων του συνόλου εκπαίδευσης, τότε μπορούμε πρώτα να προβάλλουμε τα διανύσματα σε κάποιον άλλο χώρο και στην συνέχεια να υπολογίσουμε τον πίνακα Gram. Αυτό όμως, όπως είναι γνωστό, δεν είναι τίποτα άλλο από την εφαρμογή μίας συνάρτησης πυρήνα σε όλους τους πιθανούς συνδυασμούς των διανυσμάτων. Το πόρισμα αυτό καθιστά το DIFFRAC ακόμα πιο ελκυστικό καθώς μπορούμε πολύ εύκολα να αξιοποιήσουμε τις δυνατότητες που μας παρέχουν οι πολυμελετημένες μέθοδοι πυρήνων, όπως για παράδειγμα είναι η προβολή των διανυσμάτων σε χώρους υψηλότερων διαστάσεων όπου εκεί είναι πιο πιθανό οι κλάσεις να είναι γραμμικά διαχωρίσιμες.

Το πρόβλημα αυτό πάνω στο χώρο $\mathcal{Z}_{N,P}$ είναι NP-hard και για αυτό το λόγο γίνεται relax πάνω στον συνεχή χώρο $\hat{\mathcal{Z}}_{N,P} = \{Z \in [0,1]^{N\times P} | Z \cdot \mathbf{1}_P = \mathbf{1}_N\}$. Οι Bach και Harchaoui κάνουν μία μικρή ανάλυση για την αποδοτικότητα του προσεγγιστικού αλγορίθμου και παραπέμπουμε τον αναγνώστη στο στο [1] για περισσότερες λεπτομέρειες.

Τώρα, στο συνεχή χώρο κάθε όρος της παραπάνω συνάρτησης (4.2) είναι κυρτή συνάρτηση άρα και το άθροισμα τους είναι επίσης κυρτό.

Απόδειξη. Για τον πρώτο όρο έχουμε ότι: Η συνάρτηση $f(\boldsymbol{A}) = \|\boldsymbol{A}\|_F^2$ είναι ίση με $\|vec(\boldsymbol{A})\|_2^2 = vec(\boldsymbol{A})^T vec(\boldsymbol{A}) = vec(\boldsymbol{A})^T \boldsymbol{I} vec(\boldsymbol{A})$, όπου $vec(\cdot)$ η ανάπτυξη του πίνακα σε μορφή διανύσματος. Αυτή είναι μία τετραγωνική μορφή και αφού ο πίνακας $\boldsymbol{I}$ είναι θετικά ημιορισμένος τότε είναι κυρτή. Άρα, αφού η συνάρτηση $g(\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{b}) = \boldsymbol{Z} - \boldsymbol{X}\boldsymbol{w} - \mathbf{1}_N \boldsymbol{b}$ είναι αφινική ως προς τα $\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{b}$, τότε η σύνθεση $f(g(\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{b}))$ είναι κυρτή (βλέπε κυρτές συναρτήσεις στο παράρτημα Α΄). Για τον δεύτερο όρο ισχύει το ίδιο με τη διαφορά ότι η συνάρτηση $g(\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{b})$ είναι ίση με $\boldsymbol{w}$ άρα είναι και πάλι αφινική ως προς $\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{b}$. □

Άρα και η (4.3) είναι κυρτή εφόσον είναι μία ειδική περίπτωση της (4.2) και επομένως καλούμαστε να λύσουμε ένα πρόβλημα κυρτού προγραμματισμού.

4.3.2 Επέκταση του DIFFRAC σε προβλήματα μάθησης πολλαπλών παραδειγμάτων

Στο [4] αξιοποιείται με ιδιαίτερα κομψό τρόπο ο DIFFRAC προκειμένου να επιλυθεί το ίδιο πρόβλημα με το δικό μας. Συγκεκριμένα, επεκτείνεται το μοντέλο από την ακραία περίπτωση του clustering, όπου δεν έχουμε καμία πληροφορία για τις ετικέτες των δεδομένων, στην περίπτωση των πολλαπλών παραδειγμάτων, όπου αποδίδεται μία ετικέτα σε κάθε υποομάδα δεδομένων. Όπως είδαμε στην ενότητα 4.2, είναι δυνατή η εφαρμογή ενός τέτοιου σεναρίου μάθησης στο εν λόγω πρόβλημα. Στο [4] κατασκευάζεται ένα bag για κάθε γλωσσικό αντικείμενο w_j λαμβάνοντας όλα τα οπτικά αντικείμενα v_i για τα οποία ισχύει ότι $overlap(w_j, v_i) \neq 0$. Ακόμα, οι συγγραφείς θεωρούν ότι σε κάθε bag αποδίδεται μοναδική ετικέτα, ενώ για να το πετύχουν αυτό ακολουθούν μία μεθοδολογία αρκετά περιοριστική που δύσκολα μπορεί να γενικευτεί (βλέπε κεφάλαιο 6). Οι επεκτάσεις τους περιγράφονται παρακάτω.

Constraints

Ορίζεται το σύνολο των bags $\mathcal{X}_J$ με στοιχεία τα σύνολα $\mathcal{X}_j$. Ακόμα, για κάθε $\mathcal{X}_j$ ορίζουμε το σύνολο $\mathcal{N}_j = \{i | \boldsymbol{x_i} \in \mathcal{X}_j\} \subseteq \mathcal{N}$. Τα bags δεικτοδοτούνται από το $\mathcal{J} = \{0, 1..J\}$, όπου J το πλήθος τους. Να σημειώσουμε εδώ ότι κάποια από τα $\mathcal{N}_j$ μπορούν να ταυτίζονται. Σε κάθε $\mathcal{N}_j$ αποδίδεται μία ετικέτα $p_j \in \mathcal{P}$. Αυτό σημαίνει ότι το p_j πρέπει να αποδοθεί σε τουλάχιστον ένα από τα στοιχεία του $\mathcal{X}_j$. Στην ειδική περίπτωση του [4] κάθε $\mathcal{N}_j$ περιέχει τους δείκτες i των v_i που έχουν μη μηδενική επικάλυψη με το w_j, ενώ p_j η ετικέτα που προκύπτει απευθείας από το w_j . Μεταφράζοντας αυτήν την πληροφορία στη μαθηματική της έκφραση μπορούμε να πούμε ότι

$$\forall j \in \mathcal{J},\ p = p_j,\ \sum_{i \in \mathcal{N}_j} z_{ip} \geq 1 \tag{4.5}$$

Ακόμα, ο χώρος $\hat{\mathcal{Z}}_{N,P}$ μπορεί να γραφτεί ως

$$\forall i \in \mathcal{N},\ \forall p \in \mathcal{P},\ z_{ip} \geq 0 \tag{4.6}$$

$$Z \cdot \mathbf{1}_P = \mathbf{1}_N \tag{4.7}$$

Επίσης, προκειμένου να μοντελοποιήσουμε το θόρυβο στις ετικέτες χρησιμοποιούμε τα γνωστά ,σε προβλήματα machine learning και convex optimization, slack variables :

$$\forall j \in \mathcal{J},\ p = p_j,\ \sum_{i \in \mathcal{N}_j} z_{ip} \geq 1 - \xi_j \tag{4.8}$$

$$\forall j \in \mathcal{J},\ \xi_j \geq 0 \tag{4.9}$$

και η αντικειμενική συνάρτηση γίνεται:

$$\min_{Z, \boldsymbol{\xi}} Tr(\boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{A}(\boldsymbol{X}, \lambda)) + \kappa \boldsymbol{\xi}^T \boldsymbol{\xi} \tag{4.10}$$

Με τα slack variables γίνεται πιο χαλαρή η ικανοποίηση των ανισοτήτων, δηλαδή επιτρέπεται σε κάποιες περιπτώσεις να μην πάρει κανένα από τα δεδομένα ενός bag την ετικέτα που του αποδίδεται. Έτσι, μοντελοποιούνται τα λάθη που μπορεί να γίνουν κατά την απόδοση των ετικετών. Τέτοια μπορεί να υπάρξουν είτε από ανεπαρκή ευθυγράμμιση των 2 συνοδευτικών κειμένων, είτε από λάθη στην εξόρυξη πληροφορίας από το κείμενο.

Ακόμα, όπως αναφέρεται στο [4] παρατηρήθηκε εμπειρικά ότι αντικαθιστώντας τη σταθερά 1 στον περιορισμό (4.8) με μία μεγαλύτερη σταθερά α, ο αλγόριθμος αποκτά μεγαλύτερη σθεναρότητα και οδηγείται σε υψηλότερα ποσοστά αναγνώρισης. Δηλαδή ο νέος περιορισμός είναι:

$$\forall j \in \mathcal{J},\ p = p_j,\ \sum_{i \in \mathcal{N}_j} z_{ip} \geq a - \xi_j \tag{4.11}$$

Αυτό, μπορεί να ερμηνευθεί από το γεγονός ότι όταν σε ένα bag συμμετέχει μεγάλος αριθμός δειγμάτων, τότε για μικρές τιμές του α είναι εύκολο να ικανοποιηθεί ο περιορισμός με

αποτέλεσμα να μην διοχετεύει επαρκώς την πληροφορία του στον αλγόριθμο βελτιστοποίησης. Από εδώ και στο εξής οι περιορισμοί που θα αναφέρουμε αντικαθιστούν τη σταθερά 1 με α.

Τελικά το πρόγραμμα κυρτού προγραμματισμού που πρέπει να λυθεί έχει σαν αντικειμενική συνάρτηση την (4.10) και σαν περιορισμούς τις (4.6), (4.7),(4.9),(4.11). Ο όρος που προστίθεται στη συνάρτηση είναι κυρτός και η απόδειξη είναι ίδια με αυτήν που κάναμε στην προηγούμενη ενότητα. Τα constraints είναι προφανώς αφινικά ως προς όλες τις μεταβλητές άρα το εφικτό σύνολο (feasible set) είναι επίσης κυρτό (βλέπε κυρτά σύνολα στο παράρτημα Α΄).

Rounding

Αφού βρούμε το ολικό ελάχιστο του προβλήματος $\hat{\boldsymbol{Z}}^*$ στον χώρο $\hat{\mathcal{Z}}_{N,P}$ πρέπει να επιστρέψουμε στον διακριτό χώρο $\mathcal{Z}_{N,P}$ για να αποφασίσουμε σε ποια κλάση ανήκει το κάθε δεδομένο. Έτσι βρίσκουμε μία προσέγγιση της βέλτιστης λύσης του αρχικού συνδυαστικού προβλήματος. Η διαδικασία του rounding που προτείνεται στο [4] είναι $\boldsymbol{Z}^* = \arg\min_{\boldsymbol{Z}\in\mathcal{Z}} \|\hat{\boldsymbol{Z}}^* - \boldsymbol{Z}\|_F^2$. Αυτό ισοδυναμεί με:

$$\arg\min_{\boldsymbol{Z}\in\mathcal{Z}} \|\hat{\boldsymbol{Z}}^* - \boldsymbol{Z}\|_F^2 =$$

$$\arg\min_{\boldsymbol{Z}\in\mathcal{Z}} Tr((\hat{\boldsymbol{Z}}^* - \boldsymbol{Z})(\hat{\boldsymbol{Z}}^* - \boldsymbol{Z}^T)) =$$

$$\arg\min_{\boldsymbol{Z}\in\mathcal{Z}} Tr(\hat{\boldsymbol{Z}}^*\hat{\boldsymbol{Z}}^{*T} + \boldsymbol{Z}\boldsymbol{Z}^T - \hat{\boldsymbol{Z}}^*\boldsymbol{Z}^T - \boldsymbol{Z}\hat{\boldsymbol{Z}}^{*T}) =$$

$$\arg\min_{\boldsymbol{Z}\in\mathcal{Z}} Tr(\hat{\boldsymbol{Z}}^*\hat{\boldsymbol{Z}}^{*T}) + Tr(\boldsymbol{Z}\boldsymbol{Z}^T) - Tr(\hat{\boldsymbol{Z}}^*\boldsymbol{Z}^T) - Tr(\boldsymbol{Z}\hat{\boldsymbol{Z}}^{*T}) = \quad \text{(ισχύει ότι } Tr(\hat{\boldsymbol{Z}}^*\hat{\boldsymbol{Z}}^{*T}) \text{ σταθερό)}$$

$$\arg\min_{\boldsymbol{Z}\in\mathcal{Z}} Tr(\boldsymbol{Z}\boldsymbol{Z}^T) - 2Tr(\hat{\boldsymbol{Z}}^*\boldsymbol{Z}^T) = \quad \text{(λόγω των ιδιοτήτων του πίνακα δείκτρια } Tr(\boldsymbol{Z}\boldsymbol{Z}^T) = N)$$

$$\arg\max_{\boldsymbol{Z}\in\mathcal{Z}} Tr(\hat{\boldsymbol{Z}}^*\boldsymbol{Z}^T)$$

Αυτό είναι ισοδύναμο με την επιλογή της μέγιστης τιμής ανά γραμμή του πίνακα $\hat{\boldsymbol{Z}}^*$. Οι θέσεις των μέγιστων αυτών τιμών είναι και οι θέσεις του $\boldsymbol{Z}$ που πρέπει να πάρουν την τιμή 1.

4.4 Νέοι Αλγόριθμοι Μάθησης - Επιπλέον Μορφές Επίβλεψης

4.4.1 Επέκταση σε προβλήματα μάθησης πολλαπλών παραδειγμάτων με πιθανοτικές ετικέτες

Όπως είδαμε παραπάνω οι ετικέτες που εξάγονται από το κείμενο πολλές φορές παρουσιάζουν αβεβαιότητα και δεν μπορεί να αποδοθεί μοναδική ετικέτα σε κάθε bag. Συγκεκριμένα, η ιδέα για αυτήν την περίπτωση προέκυψε στο πρόβλημα της ταξινόμησης δράσεων, καθώς για την εξόρυξη της πληροφορίας υπολογίζουμε τις τιμές ομοιότητας της γλωσσικής περιγραφής κάθε ετικέτας με διάφορες προτάσεις-γλωσσικά αντικείμενα w_j του κειμένου.

Έτσι, για κάθε πρόταση η πληροφορία που παίρνουμε είναι ένα διάνυσμα $\boldsymbol{similarity}_{jp}$ διάστασης P που τα στοιχεία του εκφράζουν ομοιότητες. Για να ενσωματώσουμε την πληροφορία της αβεβαιότητας αυτής, μοντελοποιούμε τη μάθηση σαν σενάριο πιθανοτικών ετικετών (probabilistic labels). Μετατρέπουμε δηλαδή τις ομοιότητες σε μια κατανομή πιθανότητας διαιρώντας κάθε στοιχείο του $\boldsymbol{similarity}_j$ με το άθροισμα των στοιχείων του: $\mathbb{P}[p_j = p] = \frac{similarity_{jp}}{\sum_{k=1}^{P} similarity_{jk}}$,όπου p ο δείκτης που αντιπροσωπεύει κάθε ετικέτα. Ορίζουμε έναν πίνακα $\boldsymbol{S} \in \mathcal{S}_{J,P} = \{\boldsymbol{S} \in [0,1]^{J \times P} \boldsymbol{S} \cdot \mathbf{1}_P = \mathbf{1}_J\}$ με γραμμές $\boldsymbol{s_j} = \mathbb{P}[p_j = p]_{p=1}^{P}$. Έτσι, αντικαθιστούμε κάθε p_j της προηγούμενης μεθόδου από τις κατανομές πιθανότητας $\boldsymbol{s_j}$. Η έννοια της πιθανοτικής ετικέτας ,όπως αναφέρεται και στο [28], εκφράζει ουσιαστικά για κάθε δείγμα (ή για κάθε bag όπως εδώ) μία αρχική εκτίμηση για το ποια ετικέτα πρέπει να του αποδοθεί και με τι ποσό βεβαιότητας (confidence). Ειδική περίπτωση είναι και η πλήρως επιβλεπόμενη μάθηση όπου για κάθε δείγμα εκπαίδευσης η αρχική εκτίμηση δεν ενέχει καθόλου αβεβαιότητα καθώς δίνει πιθανότητα/βεβαιότητα ίση με 1 στην πραγματική ετικέτα και 0 στις υπόλοιπες. Εμείς αυτό που επιδιώκουμε είναι να βρούμε μία τελική εκτίμηση της βεβαιότητας αυτής. Η πληροφορία αυτή επιστρέφεται στη μεταβλητή $\hat{\boldsymbol{Z}}$ ως confidence scores ή στην $\boldsymbol{Z}$ ως πίνακας ταξινόμησης των δεδομένων. Στην πλήρως επιβλεπόμενη μάθηση η τελική εκτίμηση για τα δεδομένα εκπαίδευσης προφανώς δεν διαφέρει από την αρχική.

Για να εισάγουμε τις πιθανοτικές ετικέτες στο πρόβλημα βελτιστοποίησης χρησιμοποιήσαμε μία ad hoc μέθοδο, η οποία δεν επιδέχεται κάποια προφανή πιθανοτική ερμηνεία. Συγκεκριμένα, αν για μία πρόταση w_j ορίσουμε ως $\mathcal{P}_j$ το σύνολο των ετικετών για τις οποίες η πιθανότητα είναι μη μηδενική τότε μπορούμε να κατασκευάσουμε $P_j = |\mathcal{P}_j|$ περιορισμούς της μορφής (4.11). Η ιδέα είναι ότι εφόσον δεν μπορούμε να γνωρίζουμε ποια είναι η πραγματική ετικέτα που περιγράφει κάθε σύνολο δειγμάτων τότε οποιοσδήποτε από τους P_j διαφορετικούς περιορισμούς που μπορεί να προκύψουν, είναι δυνατόν να είναι σωστός. Γνωρίζουμε, όμως, ότι ο καθένας έχει διαφορετική πιθανότητα να είναι αυτός που πράγματι πρέπει να εφαρμοστεί. Έτσι, μπορούμε να να τους χαλαρώσουμε αντιστρόφως ανάλογα με την πιθανότητα τους να είναι σωστοί. Δηλαδή, όσο πιο μεγάλη πιθανότητα έχει μία ετικέτα να αποδοθεί σε ένα bag, τόσο λιγότερο πρέπει να χαλαρώσει ο αντίστοιχος περιορισμός. Έτσι, τώρα κάθε ξ_j αντικαθίσταται από τα ξ_{jp} με $p \in \mathcal{P}_j$ και το σύνολο περιορισμών της (4.11) γίνεται:

$$\forall j \in \mathcal{J},\ \forall p \in \mathcal{P}_j,\ \sum_{i \in \mathcal{N}_j} z_{ip} \geq a - \xi_{jp} \tag{4.12}$$

Η διαφοροποίηση των P_j περιορισμών μεταξύ τους γίνεται μέσω της αντικειμενικής συνάρτησης ως εξής:

$$\min_{Z,\boldsymbol{\xi}} Tr(\boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{A}(\boldsymbol{X}, \lambda)) + \kappa \sum_{j \in \mathcal{J}} \sum_{p \in \mathcal{P}_j} \rho_{jp} \xi_{jp}^2 \tag{4.13}$$

όπου ρ_{jp} κατάλληλα επιλεγμένα βάρη. Εύκολα αντιλαμβάνεται κανείς ότι όσο μεγαλύτερο είναι το ρ_{jp}, τόσο μικρότερο πρέπει να είναι το ξ_{jp}, πάντα σεβόμενο τους περιορισμούς βέβαια, προκειμένου να φτάσουμε στο ολικό ελάχιστο. Άρα έτσι, πετυχαίνουμε διαφορετική χαλάρωση μεταξύ των περιορισμών, τους δίνουμε δηλαδή διαφορετική βαρύτητα. Ταυτόχρονα, όμως, είναι

εφικτό ένας περιορισμός με μικρό βάρος να αποκτήσει τελικά μικρή μεταβλητή χαλάρωσης αν η ομαδοποίηση των δεδομένων το επιβάλλει. Με αυτόν τον τρόπο δίνεται η δυνατότητα να αποδοθεί σε ένα bag μία ετικέτα η οποία δεν έχει αρχικά μεγάλη πιθανότητα απόδοσης. Για να επιλέξουμε τα βάρη ρ_{jp}, πειραματιστήκαμε με κάποιες συναρτήσεις εξαγωγής τους. Τελικά, καταλήξαμε στην απλή επιλογή $\forall p \in \mathcal{P}_j : \rho_{jp} = s_{jp}$, δηλαδή να είναι ίσα με τις αντίστοιχες πιθανότητες. Πιθανόν βέβαια άλλες συναρτήσεις να αποδίδουν καλύτερα.

Σημειώνουμε εδώ, ότι θέτουμε αρχικά ένα κατώφλι (*similarity threshold*) στις τιμές των διανυσμάτων ομοιοτήτων, έτσι ώστε όταν μία ομοιότητα δεν το ξεπερνά να τίθεται ίση με το 0. Ο λόγος είναι ότι ένα μεγάλο πλήθος των w_j που εξάγουμε δεν εκφράζουν κάποια από τις ετικέτες και η εισαγωγή τους στη διαδικασία βελτιστοποίησης θα εισάγει μόνο θόρυβο. Το κατώφλι πρέπει να τεθεί με προσοχή έτσι ώστε στη γενική περίπτωση να μην μηδενίζει ωφέλιμες τιμές ομοιοτήτων.

Αναφέρουμε, τέλος, ότι μία ίσως πιο γενική και καλά θεμελιωμένη μέθοδος αντιμετώπισης του προβλήματος θα μπορούσε να γίνει όπως στο [28]. Εκεί οι συγγραφείς, όπως έχουμε προαναφέρει, αντιμετωπίζουν το πρόβλημα των πιθανοτικών ετικετών για single instance προβλήματα. Εκπαιδεύουν ένα μοντέλο και ταυτόχρονα καταλήγουν σε τελικές εκτιμήσεις βεβαιότητας για τις κλάσεις των δεδομένων προσπαθώντας να πετύχουν ένα trade-off μεταξύ της απόστασης της τελικής εκτίμησης από την εκτιμώμενη από το μοντέλο a posteriori πιθανότητα και της απόστασης από την αρχική εκτίμηση. Ο αλγόριθμος βελτιστοποίησης όμως είναι expectation maximization και άρα μη κυρτός. Για αυτόν τον λόγο αποφύγαμε αυτήν τη μέθοδο. Η επέκταση όμως του [28] σε multiple instance προβλήματα δεν παύει να είναι ένα ενδιαφέρον πρόβλημα για μελλοντική μελέτη που ενδεχομένως να μοντελοποιεί καλύτερα τις διάφορες πτυχές αυτού του σεναρίου μάθησης.

4.4.2 Επέκταση για fuzzy σύνολα πολλαπλών παραδειγμάτων

Όπως ακριβώς και στην προηγούμενη ενότητα, προκειμένου να ενσωματώσουμε την πληροφορία της συνάρτησης συμμετοχής ενός δεδομένου σε ένα σύνολο, διαλέγουμε έναν ad hoc τρόπο. Συγκεκριμένα, εδώ πρέπει να υπάρξει μεροληψία ως προς τα δεδομένα που συμμετέχουν σε ένα bag j, καθώς όσο μεγαλύτερη είναι η συνάρτηση συμμετοχής ενός δεδομένου τόσο πιο πιθανό είναι το δεδομένου αυτό να χαρακτηρίζεται από την ιδιότητα του bag, δηλαδή να περιγράφεται από το γλωσσικό αντικείμενο w_j. Ορίζουμε εδώ τον πίνακα των βαθμών συμμετοχής $\boldsymbol{M} \in \mathcal{M}_{J,N} = \{\boldsymbol{M} \in [0,1]^{J \times N}\}$ με στοιχεία $\mu_{ji} = g(overlap(w_j, v_i))$ (βλ. ενότητα 4.2). Ο περιορισμός που θέτουμε προκειμένου να δωθεί διαφορετική βαρύτητα μεταξύ των δεδομένων κάθε συνόλου είναι:

$$\forall j \in \mathcal{J}, \forall p \in \mathcal{P}_j \sum_{i \in \mathcal{N}} z_{ip} \mu_{ji} \geq \alpha - \xi_{jp} \tag{4.14}$$

Να σημειώσουμε εδώ ότι η επιλογή της συνάρτησης g έχει μεγάλη σημασία καθώς αν τα μ_{ji} παίρνουν γενικά πολύ μικρές τιμές, ο αλγόριθμος θα προκαλέσει μεγάλη αύξηση στις τιμές των z_{ip} προκειμένου να ικανοποιηθούν οι περιορισμοί και η αντικειμενική συνάρτηση δεν θα

παίζει σημαντικό ρόλο. Από την άλλη, αν τα μ_{ji} παίρνουν μεγάλες τιμές ακόμα και για μικρές τιμές του $overlap(w_j, v_i)$ τότε οδηγούμαστε σε εξάλειψη της μεροληψίας και άρα εκφυλίζεται το πρόβλημα σε αυτό των απλών συνόλων πολλαπλών παραδειγμάτων. Ακόμα, μπορούμε να εισάγουμε μία υπερπαράμετρο που να επηρεάζει την χρονική επικάλυψη και άρα και τη συμμετοχή κάθε v_i σε ένα σύνολο παραδειγμάτων. Στόχος είναι να περιορίσουμε όσο γίνεται τα λάθη στα χρονικά όρια των w_j όπως αυτά προκύπτουν από το βήμα της ευθυγράμμισης των 2 κείμενων που προαναφέραμε. Συγκεκριμένα, μπορούμε να επεκτείνουμε τα όρια κάθε w_j κατά ένα συγκεκριμένο χρονικό διάστημα 2ϵ, δηλαδή να θέσουμε $l\acute{}_j^1 = l_j^1 - \epsilon$, $l\acute{}_j^2 = l_j^2 + \epsilon$. Έτσι, τα μ_{ji} καθορίζονται από την παράμετρο ϵ και την επιλογή της συνάρτησης συμμετοχής.

4.5 Άλλες επεκτάσεις - Εισαγωγή επιπλέον πληροφορίας

4.5.1 Υπαινισσόμενη Πληροφορία από τις Επαναλήψεις των bags

Δεν είναι δύσκολο να παρατηρήσει κανείς, με μία προσεκτικότερη ματιά στους περιορισμούς, ότι κάποιοι από αυτούς μπορεί να επαναλαμβάνονται. Οι επαναλήψεις αυτές μπορεί να προκύψουν αν για δύο διαφορετικές προτάσεις w_j τα οπτικά δείγματα συμμετέχουν με τον ίδιο τρόπο στα bags που προκύπτουν. Τότε για κάθε ετικέτα που έχει μη μηδενική πιθανότητα απόδοσης και στα 2 bags, θα προκύπτει ο ίδιος περιορισμός, πιθανόν με διαφορετική βαρύτητα. Για παράδειγμα, σε μία περιγραφή μίας σκηνής μπορεί να εμφανίζεται πολλές φορές το όνομα ενός χαρακτήρα. Συνήθως, ταυτόχρονα, ο χαρακτήρας αυτός έχει πράγματι έντονη παρουσία σε εκείνο το χρονικό σημείο στο βίντεο. Επομένως, θεωρούμε σημαντική τη διατήρηση της πληροφορίας της επανάληψης των περιορισμών, προκειμένου να δίνεται μεγαλύτερη βαρύτητα σε αυτούς που παρουσιάζουν μεγάλο πλήθος επαναλήψεων. Άλλωστε, όσο πιο πολλές φορές αναφέρεται κάτι στο κείμενο τόσο μικρότερη είναι η πιθανότητα να εισάγει θόρυβο στο σύστημα αναγνώρισης. Έτσι, εισάγουμε στον αλγόριθμο ένα ακόμα είδος μεροληψίας που αφορά τις επαναλήψεις των bags.

Φυσικά, αν ένας περιορισμός επαναληφθεί αυτούσιος δεν αλλάζει το πρόβλημα. Όμως, για κάθε περιορισμό υπάρχει διαφορετική μεταβλητή χαλάρωσης ξ. Επομένως, ο αλγόριθμος συγκλίνει σε διαφορετική λύση. Είναι εύκολο να δει κανείς ότι οι διαφορετικές αυτές μεταβλητές στο ολικό ελάχιστο θα πάρουν την ίδια τιμή (άλλωστε δεν έχουν λόγο να είναι διαφορετικές). Άρα, αυτό ισοδυναμεί με το να κρατήσουμε μόνο έναν περιορισμό και να αντικαταστήσουμε όλες τις διαφορετικές μεταβλητές ξ που του αντιστοιχούν με μία κοινή. Τελικά, αν από τους $\{j_{1p}, j_{2p}, \cdots, j_{\tau p}\}$ δείκτες όπου ο περιορισμός είναι κοινός κρατήσουμε μόνο τον j_{1p} και αντικαταστήσουμε τα $\xi_{j_{1p}}, \xi_{j_{2p}}, \cdots, \xi_{j_{\tau p}}$ με $\xi\acute{}_{j_{1p}}$ στον όρο $\kappa \sum_{j \in \mathcal{J}} \sum_{p \in \mathcal{P}_j} \rho_{jp} \xi_{jp}^2$ της (4.13), τότε φαίνεται καθαρά ότι η τιμή βάρους $\rho\acute{}_{j_{1p}}$ που πρέπει να αποδθεί στην μεταβλητή $\xi\acute{}_{j_{1p}}$ δίνεται από τον τύπο: $\rho\acute{}_{j_{1p}} = \rho_{j1p} + \rho_{j2p} + \cdots + \rho_{\tau p}$.

4.5.2 Ενσωμάτωση γνώσης από την αναγνώριση προσώπου

Μία επιπλέον πηγή πληροφορίας για την αναγνώριση των αντικειμένων σε μία εικόνα ή ένα βίντεο είναι τα ήδη αναγνωρισμένα αντικείμενα τα οποία γνωρίζουμε ότι συσχετίζονται με αυτό που αναζητούμε. Συγκεκριμένα, στο εν λόγω πρόβλημα, αν γνωρίζουμε ότι κάποιος χαρακτήρας εκτελεί μία δράση και έχουμε εντοπίσει τον χαρακτήρα σε κάποια στιγμιότυπα, τότε περιορίζουμε σημαντικά το σύνολο των υποψήφιων tracks για την αναγνώριση της δράσης πάνω στα συγκεκριμένα στιγμιότυπα. Φυσικά ισχύει και το αντίστροφο, δηλαδή αν έχουμε αναγνωρίσει τη δράση σε κάποια στιγμιότυπα, μπορούμε να περιορίσουμε τα υποψήφια tracks για την αναγνώριση του χαρακτήρα. Βέβαια, όπως έχουμε προαναφέρει είναι εύκολο από τη μία να συσχετίσεις την ετικέτα ενός χαρακτήρα με ένα κομμάτι του κειμένου, αλλά αρκετά πιο δύσκολο να κάνεις το ίδιο με μία δράση. Επομένως, αξιοποιούμε την ροή πληροφορίας μόνο στην κατεύθυνση χαρακτήρας $\rightarrow$ δράση. **Συγκεκριμένα, μπορούμε να συσχετίσουμε έναν χαρακτήρα με μία πρόταση w_j αν ο χαρακτήρας αυτός είναι συντακτικά το υποκείμενο της πρότασης**. Έτσι, σε κάθε bag δράσεων j μπορούμε να αντιστοιχήσουμε έναν χαρακτήρα p. [3]

Στο [4] περιγράφεται η απλή εκδοχή του προβλήματος, όπου κάθε υποψήφιο track για αναγνώριση προσώπου έχει '1-1' αντιστοιχία με ένα υποψήφιο track για αναγνώριση δράσης. Εμείς εδώ εξετάζουμε τη γενικότερη περίπτωση όπου δεν γνωρίζουμε την αντιστοιχία αυτή αλλά μπορούμε μόνο να την εκτιμήσουμε και πάλι μέσα από την χρονική τους επικάλυψη.[4] Ορίζουμε τον πίνακα $\boldsymbol{B} \in \mathcal{B}_{N_A,N_P+1}$, που περιέχει μία μετρική αντιστοίχησης (π.χ την χρονική επικάλυψη) για όλα τα ζεύγη των N_A υποψηφίων tracks για αναγνώριση δράσεων με τα N_P υποψήφια tracks για αναγνώριση προσώπου. Ακόμα, υπολογίζουμε μία μετρική που δείχνει κατά πόσο ένα track δράσης δεν αντιστοιχίζεται με κανένα από αυτά του προσώπου. Αυτά τα tracks δεν πρέπει με κάποιο τρόπο να επιβαρυνθούν γιατί δεν έχουμε καμία πληροφορία για τη κλάση του χαρακτήρα με τον οποίο συσχετίζονται - αυτόν δηλαδή που εκτελεί τη δράση. Ο υπολογισμός αυτής της μετρικής μπορεί να γίνει αθροίζοντας τα χρονικά διαστήματα στα οποία το track δράσης δεν επικαλύπτεται από κανένα track προσώπου. Συμβολίζουμε με $\mathcal{N}_A$ το σύνολο των πρώτων και με $\mathcal{N}_P$ το σύνολο των δεύτερων. Επίσης, με $\mathcal{A}$ συμβολίζουμε τώρα τις κλάσεις των δράσεων για να τις διαχωρίσουμε από τις αντίστοιχες των χαρακτήρων που συμβολίζουμε με $\mathcal{P}$. Ακόμα με $\boldsymbol{T}\in\mathbb{R}^{N_P+1\times P+1}$ (εξηγούμε τις διαστάσεις παρακάτω) συμβολίζουμε τη μεταβλητή που εκφράζει την κλάση κάθε track προσώπου. Με (j, p_j) συμβολίζουμε το ζεύγος που προκύπτει από το bag j και από την κλάση του υποκειμένου της πρότασης p_j. Τελικά, ό νέος περιορισμός είναι ο εξής:

$$\forall j \in \mathcal{J}, \forall a \in \mathcal{A}_j \sum_{i \in \mathcal{N}_A} z_{ia}\mu_{ji} \sum_{k \in \mathcal{N}_P} \left(\frac{b_{ik}t_{kp_j}}{\sum_{k \in \mathcal{N}_P} b_{ik}}\right) \geq \alpha - \xi_{ja} \tag{4.15}$$

[3] συμπεριλαμβάνεται και ο "κενός' χαρακτήρας, όταν δηλαδή η δράση δεν έχει υποκείμενο που να ανήκει στο σύνολο των χαρακτήρων.

[4] Φυσικά αυτός ο τρόπος είναι αρκετά απλός, καθώς μπορεί να γίνει πιο αποτελεσματικά, για παράδειγμα μέσω της χωρικής επικάλυψης ή μέσα από την εύρεση των τμημάτων του ανθρώπινου κορμού και άρα την αντιστοίχηση καθέ προσώπου με το υπόλοιπο σώμα που εκτελεί τη δράση

Οι μεταβλητές t_{kp} μπορούν να έχουν προκύψει είτε ως confidence scores, είτε κατόπιν στρογγυλοποίησης (rounding) ως δυαδικές μεταβλητές. Στην απλή περίπτωση της '1-1' αντιστοίχησης, ο πινακας $\boldsymbol{B}$ είναι μοναδιαίος και επομένως κάθε δείγμα i του bag j συμμετέχει στο άθροισμα ανάλογα με την τιμή t_{ip} του αντίστοιχου track του προσώπου. Όταν η αντιστοίχηση δεν είναι '1-1' σταθμίζουμε τις πιθανότητες t_{kp_j} ανάλογα με την ποσότητα αντιστοίχησης των 2 tracks.

Ο πίνακας $\boldsymbol{T}$ είναι επέκταση του αρχικού πίνακα που προκύπτει ως έξοδος του συστήματος αναγνώρισης προσώπου. Συγκεκριμένα, $\boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_{init} & \mathbf{1}_{N_P \times 1} \\ \mathbf{1}_{1 \times P} & 1 \end{bmatrix}$ Η προσθήκη της τελευταίας γραμμής εξασφαλίζει ότι ένα track που δεν αντιστοιχίζεται σε κάποιο πρόσωπο δεν θα επιβαρυνθεί. Η προσθήκη της τελευταίας στήλης εξασφαλίζει ότι όταν μία πρόταση δεν έχει υποκείμενο από το σύνολο των χαρακτήρων, τότε δεν θα υπάρξει μεροληψία για κανένα από τα tracks που ανήκουν στο bag της πρότασης (σε όλα θα δοθεί βαρύτητα ίση με 1).

4.5.3 Ενσωμάτωση πρότερης γνώσης - Προεκπαιδευμένος Ταξινομητής

Αν με κάποιο τρόπο έχουμε πρόσβαση σε πρότερη γνώση για τις ετικέτες ενός υποσυνόλου των δεδομένων τότε είναι ιδιαίτερα χρήσιμο η πληροφορία αυτή να ενσωματωθεί προκειμένου η επίβλεψη να γίνει ισχυρότερη και ο αλγόριθμος να διευκολυνθεί στην ανάκτηση των κρυφών μεταβλητών. Για παράδειγμα στο [5] αναφέρεται το σενάριο μάθησης όπου κάποια δεδομένα είναι πλήρως επισημειωμένα, δηλαδή γνωρίζουμε τις ετικέτες τους. Γενικά η πρότερη γνώση σε προβλήματα βελτιστοποίησης μπορεί να εισαχθεί είτε με την προσθήκη παραπάνω περιορισμών που να την περιγράφουν, είτε στην αντικειμενική συνάρτηση μέσω ενός ακόμα όρου. Στην παρούσα διπλωματική εξετάζεται η περίπτωση όπου δεν έχουμε γνώση των πραγματικών ετικετών των δεδομένων αλλά μπορούμε να χρησιμοποιήσουμε έναν προεκπαιδευμένο ταξινομητή ο οποίος να μας δώσει μία πρόβλεψη για αυτές. Η πρόβλεψη αυτή δίνεται είτε μέσω των τελικών εκτιμήσεων των ετικετών, είτε μέσα από τα confidence scores τους . Μπορούμε να συνδυάσουμε τους 2 ταξινομητές (προεκπαιδευμένος και ασθενώς επιβλεπόμενος) είτε μέσα από κάποια τεχνική ensemble, είτε διοχετεύοντας την πληροφορία που μας παρέχει ο προεκπαιδευμένος στην είσοδο του ασθενώς επιβλεπόμενου. Συγκεκριμένα, προσθέτουμε στην συνάρτηση (4.13) τον όρο $\frac{u}{2N}\|\boldsymbol{Z} - \boldsymbol{Z}_{prior}\|_F^2$, όπου u μία παράμετρος που ρυθμίζει το βάρος που θα δοθεί στον όρο αυτό και $\boldsymbol{Z}_{prior}$ ο πίνακας που περιέχει τις προβλέψεις του προεκπαιδευμένου ταξινομητή. Δηλαδή:

$$\min_{Z,\boldsymbol{\xi}} Tr(\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{A}(\boldsymbol{X},\lambda)) + \kappa \sum_{j\in\mathcal{J}} \sum_{p\in\mathcal{P}_j} \rho_{jp}\xi_{jp}^2 + \frac{u}{2N}\|\boldsymbol{Z} - \boldsymbol{Z}_{prior}\|_F^2 \tag{4.16}$$

Ο όρος αυτός είναι κυρτός για τους ίδιους λόγους που αναφέραμε στην ενότητα 4.3.1. Έτσι, με κατάλληλη επιλογή του u, ο αλγόριθμος προσπαθεί από την μία να ομαδοποιήσει τα δεδομένα σεβόμενος τους περιορισμούς που του θέτει η ασθενής επίβλεψη, ενώ από την άλλη προσπαθεί να αποκλίνει όσο λιγότερο γίνεται από τις προβλέψεις που του δίνονται. Στο

τέλος, όπως θα δούμε στο κεφάλαιο 7 καταφέρνει να συνδυάσει τους δύο ταξινομητές βελτιώνοντας τα ποσοστά επιτυχίας και των 2. Να σημειώσουμε εδώ ότι αυτό το συμπέρασμα είναι αρκετά σημαντικό καθώς στο [5] χρησιμοποιούνται πλήρως επισημειωμένα δεδομένα (παρόμοια με αυτά που πρέπει να αναγνωριστούν) τα οποία δεν είναι πάντα εύκολο να βρεθούν. Αντίθετα, ο προεκπαιδευμένος ταξινομητής που αναφέρουμε εμείς μπορεί να έχει εκπαιδευτεί σε διαφορετικής φύσης δεδομένα τα οποία δεν είναι απαραίτητο να γνωρίζουμε.

4.6 Από κοινού μάθηση οπτικών και γλωσσικών δεδομένων

Η παρακάτω υποενότητα δεν έχει υλοποιηθεί αλλά η μοντελοποίηση της έχει γίνει και αποτελεί ενδιαφέρουσα ιδέα για πολυτροπικό εντοπισμό αντικειμένων/συμβάντων για μελλοντική εργασία.

Παρατηρώντας ότι η απόδοση των ετικετών στα μέρη του κειμένου αποτελεί και αυτή ένα πρόβλημα ταξινόμησης, θεωρούμε πιθανό να μπορούν να λυθούν τα 2 προβλήματα από κοινού. Δηλαδή, να γίνει αντιληπτό το Κείμενο ως ισότιμη τροπικότητα με το Βίντεο και να γίνει Κατανόηση και των 2. Συγκεκριμένα, αν αναπαραστήσουμε τις προτάσεις σε ένα διανυσματικό χώρο τότε μπορούμε να εργαστούμε παρόμοια με τα προηγούμενα. Συγκεκριμένα, κάθε πρόταση w_j μπορεί να αναπαρασταθεί σε ένα χώρο $\mathbb{R}^{1\times D_2}$ μέσω ενός διανύσματος $\boldsymbol{y_j}$ [5]. Ορίζουμε τον πίνακα $Y\in\mathbb{R}^{J\times D_2}$ με γραμμές τα $\boldsymbol{y_j}$. Ορίζουμε τη μεταβλητή $\boldsymbol{q_j}$ η οποία ανήκει στο χώρο $\{0,1\}^{1\times P}$ με $\boldsymbol{q_j}\cdot\mathbf{1}_P = 1$, η οποία υποδεικνύει την κλάση στην οποία ανήκει το w_j. Όμοια ορίζουμε την μεταβλητή $\boldsymbol{Q}\in\mathbb{R}^{J\times P}$ με γραμμές τα $\boldsymbol{q_j}$ καθώς και το σύνολο των πινάκων δεικτριών $\mathcal{Q}_{J,P} = \{\boldsymbol{Q}\in\{0,1\}^{J\times P} | Q\cdot\mathbf{1}_P = \mathbf{1}_J\}$.

Τώρα εφαρμόζοντας το ίδιο relaxation με πριν μπορούμε να ελαχιστοποιήσουμε την:

$$\min_{\boldsymbol{Q},\xi_{text}} \frac{1}{2}(\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}_{text}(\boldsymbol{Y},\lambda_{text})) + \kappa_{text}\boldsymbol{\xi}_{text}^T\boldsymbol{\xi}_{text} \tag{4.17}$$

$$\text{με } \boldsymbol{A}_{text}(\boldsymbol{Y},\lambda_{text}) = \frac{1}{J}\boldsymbol{\Pi}_J(\boldsymbol{I}_J - \boldsymbol{Y}(\boldsymbol{Y}^T\boldsymbol{\Pi}_J\boldsymbol{Y} + J\lambda_{text}I_{D_2})^{-1}\boldsymbol{Y}^T)\boldsymbol{\Pi}_J \tag{4.18}$$

Ταυτόχρονα μπορούμε να κατασκευάσουμε κατάλληλα constraints αξιοποιώντας γνώση που παρέχεται μέσα από το βίντεο, ακολουθώντας ακριβώς την ίδια λογική με όσα έχουμε προαναφέρει. Αντιστρέφουμε δηλαδή τη ροή της πληροφορίας η οποία τώρα ακολουθεί την κατεύθυνση βίντεο→ κείμενο. Ο πίνακας $\boldsymbol{S}$ που ορίστηκε στο 4.4.1 είναι μια πιθανή λύση του προβλήματος αυτού. Έτσι, αρχικοποιώντας τον $\boldsymbol{Q}$ στον $\boldsymbol{S}$ μπορούμε αρχικά να βελτιστοποιήσουμε την (4.13). Στη συνέχεια αξιοποιώντας τις τιμές της μεταβλητής $\hat{\boldsymbol{Q}}$ [6] ως πιθανοτικές ετικέτες ή τις τιμές της $\boldsymbol{Q}$ ως ντετερμινιστικές, καθώς και τις επικαλύψεις μεταξύ οπτικών και γλωσσικών δεδομένων, μπορούμε να κατασκευάσουμε όμοια τους περιορισμούς της (4.17) και να βελτιστοποιήσουμε και αυτήν. Η διαδικασία μπορεί να επαναληφθεί μέχρι τη σύγκλιση.

[5] Χρησιμοποιούμε το σύμβολο y που είχαμε χρησιμοποιήσει για τις ετικέτες σε προηγούμενο κεφάλαιο αλλά τα 2 μεγέθη δεν πρέπει να συγχέονται

[6] πρόκειται για το relaxation της $\boldsymbol{Q}$ στο συνεχή χώρο

Το πρόβλημα τώρα προφανώς δεν είναι convex και άρα έχει μεγάλη σημασία η αρχικοποίηση. Σε κάθε μία συνιστώσα του (κείμενο και βίντεο), όμως, είναι κυρτό, και επομένως μπορούμε να ευελπιστούμε σε καλή σύγκλιση. Παρόμοια μέθοδος ακολουθείται στο [4] όπου οι 2 συνιστώσες είναι δράσεις και πρόσωπα σε ένα βίντεο και στο [47], όπου η μία συνιστώσα είναι συναναφορές (coreferences) ονομάτων σε ένα κείμενο και η άλλη πρόσωπα σε ένα βίντεο.

Μέρος II

Υλοποίηση και Πειραματική Αξιόλογηση των Μεθόδων

Κεφάλαιο 5

Εντοπισμός και Εξαγωγή Οπτικών Χαρακτηριστικών

5.1 Εισαγωγή

Στο παρόν κεφάλαιο αναλύεται η μεθοδολογία που ακολουθήσαμε προκειμένου να εντοπίσουμε χωροχρονικά τα υποψηφία προς αναγνώριση αντικείμενα στο χώρο του βίντεο, καθώς και οι τρόποι με τους οποίους εξάγουμε τις περιγραφές τους. Τα προβλήματα του εντοπισμού (spatio-temporal localization) και του σχεδιασμού αναπαραστάσεων βίντεο σε d-διάστατους χώρους είναι από μόνα τους σύνθετα προβλήματα της όρασης υπολογιστών και έχουν προταθεί πολλοί τρόποι επίλυσης τους. Για παράδειγμα, όπως διαβάζουμε στα [52, 66], έχουν προταθεί μέθοδοι που βασίζονται στην εύρεση συγκεκριμένων χαρακτηριστικών (π.χ χρώμα, σχήμα, υφή κ.ά), στο ταίριασμα με προϋπολογισμένα πρότυπα (templates) (σταθερά ή παραμορφώσιμα), στην δυαδική ταξινόμηση μεταξύ παρασκηνίου και προσκηνίου ή ακόμα και στην ανιχνευόμενη κίνηση των αντικειμένων. Όσον αφορά την αναπαράσταση του βίντεο η έρευνα έχει κινηθεί γύρω από 2 βασικές τεχνικές, τη 'χειροκίνητη' σχεδίαση περιγραφητών (hand-crafted descriptors), η οποία βασίζεται σε παρατηρήσεις της κρυμμένης δομής του αντικειμένου που θέλουμε να περιγράψουμε και στην αποτύπωση της με τέτοιο τρόπο ώστε να είναι όσο πιο γενική γίνεται, και στην αυτόματη μάθηση χαρακτηριστικών (feature learning), όπου αλγόριθμοι μηχανικής μάθησης με ή χωρίς επίβλεψη εντοπίζουν μόνοι τους την κρυμμένη δομή και μας δίνουν τις ζητούμενες αναπαραστάσεις. Στα δύο αυτά προβλήματα, τα τελευταία χρόνια έχει αρχίσει να επικρατεί κατά κράτος η βαθιά μηχανική μάθηση είτε με τη χρήση επιβλεπόμενων δικτύων όπως είναι τα CNNs (χρησιμοποιούνται σε όλο σχεδόν το φάσμα των προβλημάτων), είτε με τη χρήση μη επιβλεπόμενων συστημάτων, όπως είναι οι autoencoders (χρησιμοποιούνται για την κωδικοποίηση αντικειμένων σε υψηλού επιπέδου αναπαραστάσεις). Στην εργασία αυτή δεν επεκταθήκαμε στο σχεδιασμό αλγορίθμων για τα συστήματα αυτά, αλλά χρησιμοποιήσαμε τυποποιημένες τεχνικές (off-the-shelf) ιδιαίτερα διαδεδομένες και αποδεκτές, προσαρμόζοντας τες στις ανάγκες της εφαρμογής μας. Να σημειώσουμε τέλος, ότι διαδικασίες του εντοπισμού, της αναπαράστασης και της αναγνώρισης γίνονται σε διαφορετικά στάδια, πρακτική η οποία είναι και η πιο συνηθισμένη, αν και τελευταία έχουν προταθεί μέθοδοι

για την από κοινού εκτέλεση τους (π.χ για εντοπισμό και αναγνώριση δράσεων [26, 63]).

5.2 Η βάση δεδομένων COGNIMUSE

Οι αλγόριθμοι που θα παρουσιαστούν στην παρούσα διπλωματική αξιολογούνται πάνω στην βάση δεδομένων COGNIMUSE η οποία έχει κατασκευαστεί από το εργαστήριο CVSP του ΕΜΠ στα πλαίσια του ερευνητικού προγράμματος COGNIMUSE και παρουσιάζεται στο [70]. Παρακάτω κάνουμε μία σύντομη περιγραφή της βάσης αυτής και εξηγούμε τις προκλήσεις που πρέπει να αντιμετωπιστούν. Η COGNIMUSE αποτελείται από αποσπάσματα οκτώ ταινιών, στα οποία έχει πραγματοποιηθεί, από τα μέλη του εργαστηρίου, εντοπισμός και επισημείωση (annotation) των ακουστικών και οπτικών γεγονότων καθώς και επισημείωση της κατάτμησης του βίντεο σε λήψεις (shots) και σκηνές (scenes). Επισημείωση ανθρωπίνων προσώπων δεν προϋπήρχε. Οι ταινίες είναι οι εξής:

- *Beautiful Mind (BMI)*: Βιογραφία, Δράμα – 2001
- *Chicago (CHI)*: Μιούζικαλ – 2002
- *Crash (CRA)*: Δράμα – 2004
- *The Departed (DEP)*: Δράμα, Περιπέτεια, Θρίλερ – 2006
- *Finding Nemo (FNE)*: Περιπέτεια, Κινούμενα Σχέδια – 2003
- *Gladiator (GLA)*: Δράμα – 2000
- *Gone with the Wind (GWW)*: Αισθηματική, Δράμα – 1939
- *Lord of the Rings: The Return of the King (LOR)*: Φαντασίας, Περιπέτεια – 2003

Απο εδώ και πέρα οι ταινίες θα αναφέρονται με τις συντομογραφίες τους. Ο ρυθμός δειγματοληψίας των ταινιών εκτός του Gone with the Wind είναι περίπου 25 καρέ ανά δευτερόλεπτο (24.9997500025) ενώ για το Gone with the Wind ο ρυθμός είναι περίπου 24 καρέ ανά δευτερόλεπτο (23.9760237365033). Σημειώνουμε ότι στην διπλωματική εργασία αυτή δεν λαμβάνεται υπόψη η ταινία κινουμένων σχεδίων *Finding Nemo* καθώς για προφανείς λόγους δεν είναι κατάλληλη για αναγνώριση ρεαλιστικών οπτικών αντικειμένων, όπως είναι οι δράσεις και τα πρόσωπα. Ακόμα, δεν λαμβάνεται υπόψη η ταινία *Chicago* καθώς το σενάριο που τη συνοδεύει προστατεύεται από copyrights και δεν είναι δημόσια διαθέσιμο. Στον Πίνακα 5.1 αναγράφεται η διάρκεια σε λεπτά και καρέ των 6 ταινιών που θα μας απασχολήσουν στη συνέχεια.

Για την επισημείωση των οπτικών γεγονότων των ταινιών (visual event annotation) χρησιμοποιήθηκαν κλάσεις ανθρώπινων δράσεων παρόμοιες με αυτές άλλων αντίστοιχων βάσεων δεδομένων, όπως η HMDB51 και η Hollywood2. Οι κλάσεις αυτές είναι 69 και διαχωρίζονται σε 6 ευρύτερες κατηγορίες:

1. Γενικές Κινήσεις Προσώπου: *smile, cry, laugh, chew, talk, other*

	Duration (min)	Total Frames
BMI	31:17	46937
CRA	26:37	39926
DEP	30:28	45707
GLA	30:02	45062
GWW	104:10	449568
LOR	37:33	56339

Πίνακας 5.1: Διάρκεια των 6 ταινιών της COGNIMUSE που χρησιμοποιούμε, σε λεπτά και καρέ

2. Κινήσεις Προσώπου με Αλληλεπίδραση με Αντικείμενα: *eat, drink, smoke, other*

3. Γενικές Κινήσεις Σώματος: *cartwheel, clap hands, climb, climb stairs, dance, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, running, sitting down, sitting up, somersault, standing up, turn, walk, other*

4. Χειρονομίες: *pantomime, point at something, wave hands, other*

5. Κινήσεις Σώματος με Αλληλεπίδραση με Αντικείμενα: *answering phone, brush hair, catch, draw sword, dribble, driving car, getting out of the car, golf, hit something, kick ball, open car door, open door, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw, other*

6. Κινήσεις Σώματος με Αλληλεπίδραση με Άλλους Ανθρώπους: *fencing, fighting, grab hand, hugging, kick someone, kissing, punch, shake hands, sword fight, threaten person, other*

Για τις ανάγκες της παρούσας Διπλωματικής Εργασίας, οι κλάσεις των ανθρώπινων δράσεων υπό αναγνώριση περιορίστηκαν αρκετά προκειμένου να αποφευχθούν φαινόμενα μεγάλης ανισοκατανομής στο πλήθος των δεδομένων κάθε κλάσης, μερικής ή πλήρους ανυπαρξίας δεδομένων σε κάποιες κλάσεις κ.ά. Σε πρώτη φάση οι κλάσεις που κρατήθηκαν είναι 44 και είναι οι εξής: *smile,cry,laugh,chew,smoke,eat,drink,sitting down,sitting up,standing up,running,clap hands,climb,climb stairs,fall on the floor,jump,turn,dance,walk,wave hands,point at something,answering phone,driving car,getting out of the car,open car door,open door,catch,draw sword,hit something,pick,pour,ride horse,shoot bow,shoot gun,throw,fighting,hugging,kissing,grab hand,threaten person,kick someone,punch,shake hands,sword fight*. Οι κλάσεις αυτές μπορεί να μην εμφανίζονται με την ίδια συχνότητα σε όλες τις ταινίες (πράγμα προφανές καθώς μερικές εξαρτώνται σε μεγάλο βαθμό από το γενικότερο θέμα το οποίο πραγματεύεται η ταινία), αλλά για κάθε ταινία υπάρχει ένα υποσύνολο των 44 αυτών κλάσεων που τα περιεχόμενα του εμφανίζονται επαρκή αριθμό φορών. Σε δεύτερη φάση, αν κανείς χρειάζεται να κρατήσει ένα σύνολο κλάσεων που κάθε μία από αυτές να εμφανίζεται ικανοποιητικό αριθμό φορών σε

όλες τις ταινίες, μπορούν να επιλεχθούν 20 κλάσεις οι οποίες είναι οι εξής: *climb stairs, cry, dance, fall on the floor, grab hand, hugging, laugh, open door, pick, point at something, ride horse, running, sitting down, sitting up, smile, standing up, throw, turn, walk, wave hands*. Περισσότερα, για την επιλογή των δεδομένων μπορεί να διαβάσει κανείς στη διπλωματική [56] όπου πέρα από την κατασκευή μίας περιορισμένης βάσης δεδομένων κατάλληλης για αναγνώριση ανθρώπινων δράσεων, παρουσιάζονται μέθοδοι ταξινόμησης των δεδομένων με πλήρως επιβλεπόμενα βαθιά συνελικτικά δίκτυα.

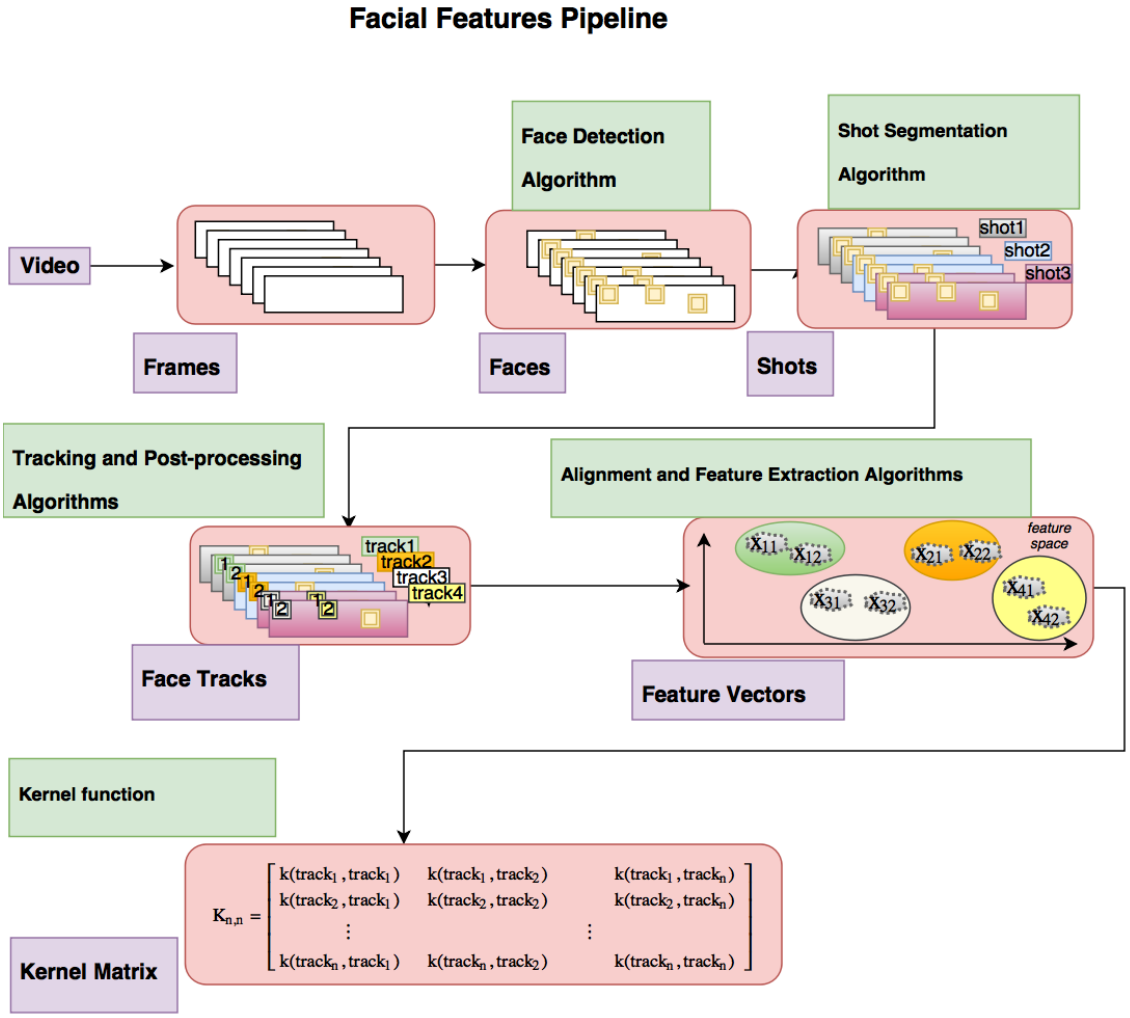
5.3 Αναπαράσταση Ανθρώπινων Προσώπων

Το ολοκληρωμένο σύστημα εντοπισμού και αναπαράστασης προσώπου είναι βασισμένο πάνω στο pipeline που περιγράφεται στο [4] το οποίο αποτελεί επέκταση του αντίστοιχου pipeline των [19, 53]. Συγκεκριμένα, όπως φαίνεται από το σχήμα 5.1 αποτελείται από τα εξής στάδια:

1. Εντοπισμός Προσώπου
2. Κατάτμηση Βίντεο σε Λήψεις
3. Παρακολούθηση Προσώπου
4. Ευθυγράμμιση Προσώπου
5. Εξαγωγή Χαρακτηριστικών
6. Υπολογισμός πυρήνων - kernels

5.3.1 Εντοπισμός Προσώπου

Ο εντοπισμός προσώπου υλοποιείται σύμφωνα με το [69]. Ο εν λόγω αλγόριθμος εντοπίζει ένα πρόσωπο χρησιμοποιώντας ταίριασμα με παραμορφώσιμα πρότυπα, τα γνωστά Deformable Part Models. Τα μοντέλα αυτά ([13, 21, 20]) αναπαριστούν τα αντικείμενα μέσω ενός αριθμού από κομμάτια (parts) των οποίων η θέση δεν είναι σταθερή, πράγμα που κάνει το μοντέλο παραμορφώσιμο. Η εύρεση της θέσης ενός αντικειμένου γίνεται μέσα από το βέλτιστο ταίριασμα των χαρακτηριστικών κάθε υποψηφίου κομματιού με το αντίστοιχο εκπαιδευμένο πρότυπο και την ταυτόχρονη ελαχιστοποίηση της μετατόπισης των σχετικών θέσεων των κομματιών από τις αναμενόμενες (από την διαδικασία της εκπαίδευσης) τιμές τους. Στη μεθοδολογία που ακολουθήσαμε εδώ, χρησιμοποιείται μία μίξη από DPMs, όπου κάθε όρος μίξης αντιπροσωπεύει μία διαφορετική γωνία θέασης του προσώπου (viewpoint). Έτσι, το ταίριασμα δεν γίνεται μόνο ως προς τις θέσεις των κομματιών, αλλά και ως προς τη μίξη που το μεγιστοποιεί. Αυτό έχει σαν αποτέλεσμα, να είναι δυνατός ο από κοινού εντοπισμός της πόζας του προσώπου και των σημείων αναφοράς του (facial landmarks). Τα κομμάτια κάθε DPM είναι συνδεδεμένα μεταξύ τους με βάση μια δενδρική δομή, στην οποία οι κόμβοι είναι τα κομμάτια και οι ακμές είναι οι συνδέσεις τους που καθορίζουν τις παραμορφώσεις. Το σύνολο των κόμβων-landmarks που



Σχήμα 5.1: Διάγραμμα του συστήματος εντοπισμού και αναπαράστασης προσώπου

αποτελούν ένα πρόσωπο ονομάζεται V και αποτελείται από σημεία πάνω στη μύτη, τα μάτια το στόμα κ.τ.λ. Κάθε δέντρο ορίζεται ως (V_m, E_m) όπου το m είναι ο δείκτης της μίξης, V_m το σύνολο των κόμβων του με $V_m \subseteq V$ - δεν φαίνονται όλα τα σημεία ενδιαφέροντος από κάποιες γωνίες θέασης - και E_m το σύνολο των ακμών (i, j) που συνδέουν τους κόμβους. Η δομή του δέντρου m, δηλαδή οι ακμές του E_m, έχει εκπαιδευτεί από πλήρως επισημειωμένα πρόσωπα (θέσεις των landmarks και viewpoints). Επίσης, για κάθε μίξη m και κόμβο i έχει εκπαιδευτεί ένα διάνυσμα πρότυπο $\boldsymbol{w_i^m}$ και για κάθε ακμή (i, j) ένα διάνυσμα $\boldsymbol{deform_{ij}^m}$ που εκφράζει το πρότυπο της αναμενόμενης παραμόρφωσης. Περισσότερα για την εκπαίδευση των μοντέλων μίξης στο [69]. Για να βρούμε τα πρόσωπα σε μία εικόνα υπολογίζουμε ένα σκορ για κάθε σύνολο θέσεων landmarks L και κάθε γωνία θέασης m. Έτσι, αν $\boldsymbol{l_i} = (x_i, y_i)$ η θέση

του σημείου $i \in V$ και $L = \{\boldsymbol{l_i} : i \in V\}$ και I η εικόνα, μεγιστοποιούμε την εξής συνάρτηση:

$$S(I, L, m) = App_m(I, L) + Shape_m(L) + \alpha^m \tag{5.1}$$

Όπου ο όρος $App_m(I, L)$ υποδεικνύει για κάθε σύνολο L το ταίριασμά του με τα προεκπαιδευμένα πρότυπα, δηλαδή την εμφάνιση ή όχι των χαρακτηριστικών ενός προσώπου. Συγκεκριμένα, αν $\boldsymbol{\phi(I, l_i)}$ το διάνυσμα χαρακτηριστικών γύρω από τη θέση $\boldsymbol{l_i}$ στην εικόνα I τότε:

$$App_m(I, L) = \sum_{i \in \mathcal{V}_m} \boldsymbol{w_i^m} \cdot \boldsymbol{\phi}(I, \boldsymbol{l_i}) \tag{5.2}$$

Ο όρος $Shape_m(L)$ εκφράζει την επιτρεπτή ελαστική παραμόρφωση μεταξύ των κομματιών κάθε μίξης και ορίζεται ως εξής:

$$Shape_m(L) = \sum_{(i,j) \in E_m} a_{i,j}^m dx_{ij}^2 + b_{ij}^m dx_{ij} + c_{ij}^m dy_{ij}^2 + d_{ij}^m dy_{ij} \tag{5.3}$$

Όπου $dx_{ij} = x_i - x_j$ και $dy_{ij} = y_i - y_j$ και $\boldsymbol{deform_{ij}^m} = (a_{ij}^m, b_{ij}^m, c_{ij}^m, d_{ij}^m)$.

Τέλος ο όρος a^m εκφράζει την a priori πιθανότητα της κάθε μίξης να εμφανιστεί.

Για τις ανάγκες της διπλωματικής εργασίας, χρησιμοποιούμε τα προεκπαιδευμένα μοντέλα όπως αυτά παρέχονται από τους συγγραφείς. Προκειμένου να εντοπιστούν παραπάνω από ένα πρόσωπα σε μία εικόνα, λαμβάνεται κάθε pixel της εικόνας ως ρίζα του δέντρου και στη συνέχεια υπολογίζεται το δέντρο που μεγιστοποιεί την (5.1) δεδομένης της θέσης της ρίζας του, μέσω δυναμικού προγραμματισμού. Στη συνέχεια, διατηρούνται μόνο τα σύνολα σημείων L που το σκορ τους ξεπερνά κάποιο κατώφλι (**-0.7** στην περίπτωση μας). Στο τέλος, εφαρμόζεται non-maximum supression προκειμένου από διαφορετικά detections του ίδιου προσώπου να διατηρηθεί μόνο αυτό με το υψηλότερο σκορ. Στα εναπομείναντα πρόσωπα σχεδιάζεται το bounding box έτσι ώστε να περιέχει όλα τα landmarks. Η ίδια διαδικασία επαναλμβάνεται για όλες τις γωνίες θέασης.

5.3.2 Κατάτμηση Βίντεο σε Λήψεις

Η κατάτμηση της ταινίας σε λήψεις γίνεται σύμφωνα με το [19]. Η διαδικασία είναι πολύ απλή και γίνεται με μεθόδους low level επεξεργασίας βίντεο. Συγκεκριμένα, για κάθε καρέ του βίντεο υπολογίζεται το χρωματικό ιστόγραμμά του (συνένωση των ιστογραμμάτων των 3 χρωματικών καναλιών) και στη συνέχεια υπολογίζεται η L_1 νόρμα της διαφοράς του από το αντίστοιχο ιστόγραμμα του επόμενου καρέ. Έτσι, δημιουργείται ένα χρονικό σήμα που δείχνει τις μεταβολές του χρωματικού ιστογράμματος κατά την εξέλιξη του βίντεο. Τώρα, ανιχνεύοντας τις απότομες μεταβολές του σήματος αυτού (spikes) μπορούμε να βρούμε τις μεταβάσεις των λήψεων (shot transitions). Η μέθοδος αυτή εντοπίζει άψογα τα λεγόμενα hard cuts, δηλαδή τις απότομες μεταβάσεις, αλλά αποτυγχάνει σε ορισμένα soft cuts, δηλαδή όταν η μετάβαση γίνεται με σταδιακό τρόπο, πράγμα συνηθισμένο στον κινηματογράφο καθώς υπάρχουν ποικίλες σκηνοθετικές τεχνικές (βλέπε https://en.wikipedia.org/wiki/Shot_transition_detection). Μία πιο αναλυτική και αποτελεσματική τεχνική κατάτμησης σε

λήψεις παρουσιάζεται στο [38]. Παρ΄ όλα αυτά, τα λάθη του μέρους αυτού του συστήματος δεν είναι σημαντικά καθώς, (i) όταν δεν εντοπίζεται μία μετάβαση τότε συνήθως το σύστημα παρακολούθησης προσώπου σταματά την παρακολούθηση έτσι και αλλιώς από μόνο του, ενώ (ii) όταν εντοπίζεται μη υπαρκτή μετάβαση το μόνο πρόβλημα είναι η διάσπαση ορισμένων tracks, τα οποία αναμένουμε να κατηγοριοποιηθούν με ευκολία στην ίδια κλάση από το σύστημα μάθησης λόγω της μεγάλης ομοιότητας τους.

5.3.3 Παρακολούθηση Προσώπου

Η παρακολούθηση κάθε προσώπου που έχει ανιχνευθεί γίνεται μέσα στα όρια των λήψεων που βρέθηκαν από το προηγούμενο βήμα. Έτσι, μειώνεται η πολυπλοκότητα της παρακολούθησης καθώς περιορίζεται ο αριθμός των ήδη ανιχνευμένων προσώπων που προσπαθούμε να συνδέσουμε. Ακόμα περιορίζονται τα λάθη της παρακολούθησης, όπως είναι η σύνδεση διαφορετικών προσώπων μεταξύ τους[1]. Το σύστημα παρακολούθησης που χρησιμοποιείται παρουσιάστηκε στα [19, 53] και αποτελείται από το γνωστό KLT feature tracker ο οποίος επιλέγει έναν αριθμό σημείων ενδιαφέροντος και παρακολουθεί τις μετακινήσεις τους από frame σε frame. Όταν, ένα σημείο ενδιαφέροντος αλλάξει πολύ σταματά η παρακολούθηση του και επιλέγεται άλλο στη θέση του. Τα σημεία ενδιαφέροντος επιλέγονται έτσι ώστε να είναι πυκνά στο εσωτερικό των bounding boxes των προσώπων και πιο αραιά στο εξωτερικό τους. Η παρακολούθηση γίνεται 2 φορές. Στην πρώτη ξεκινά από το πρώτο frame (forward) και στην δεύτερη από το τελευταίο (backward). Τα σημεία ενδιαφέροντος τώρα είναι η συνένωση αυτών που προκύπτουν από το forward tracking και αυτών που προκύπτουν από το backward. Προκειμένου τώρα να γίνει η σύνδεση των προσώπων υπολογίζεται μία μετρική ομοιότητας για κάθε ζεύγος προσώπων A , B ως το πηλίκο του πλήθους των σημείων ενδιαφέροντος που εντοπίζονται στο εσωτερικό και του A και του B, προς το πλήθος των σημείων ενδιαφέροντος που εντοπίζονται είτε μόνο στο A είτε μόνο στο B. Αν το πηλίκο είναι κάτω από ένα κατώφλι (**0.5**) τότε η ομοιότητα τίθεται ίση με 0. Τελικά, εκτελείται ένα απλό agglomerative clustering με χρήση αυτής της ομοιότητας πάνω σε όλα τα πρόσωπα και τα clusters που προκύπτουν καθορίζουν τα tracks.

Κατόπιν, ακολουθεί μία μετα - επεξεργασία των tracks η οποία αποτελείται από (i) την αφαίρεση tracks μικρής διάρκειας (κάτω από **8** frames), καθώς θεωρούνται λανθασμένοι εντοπισμοί, (ii) παρεμβολή bounding boxes μεταξύ μη διαδοχικών frames που περιέχουν εντοπισμένα πρόσωπα που ανήκουν στο ίδιο track (η απόσταση τους δεν πρέπει να ξεπερνά έναν αριθμό frames που τίθεται ίσος με **6**, προκειμένου να αποφευχθεί η παρεμβολή σε περιπτώσεις που το πρόσωπο κρύβεται στα ενδιάμεσα frames ή δεν φαίνεται καθαρά), (iii) 'κλάδεμα' επικαλυπτόμενων tracks που η μέση επικάλυψη των bounding boxes τους ξεπερνά ένα κατώφλι (**0.3**), (iv) εξομάλυνση των θέσεων των bounding boxes κάθε track προκειμένου να αποφευχθούν περιπτώσεις όπου το bounding box δεν περικλείει σωστά το πρόσωπο και τέλος (v) 'κλάδεμα' tracks των οποίων το μέσο σκορ εντοπισμού προσώπου δεν ξεπερνά κάποιο

[1] αυτό μπορεί να συμβεί αν σε 2 διαδοχικά frames η κάμερα κινηθεί απότομα και ταυτόχρονα εντοπιστούν πρόσωπα διαφορετικού χαρακτήρα στην ίδια περίπου θέση της εικόνας. Τότε, αν δεν ανιχνευθεί η μετάβαση της λήψης ο tracker μπορεί να μην καταφέρει να διαχωρίσει τα 2 πρόσωπα

κατώφλι (**-0.6**). Το τελευταίο βήμα, δεν υπήρχε στις προηγούμενες εργασίες και αποτελεί δικιά μας προσθήκη καθώς παρατηρήσαμε ότι στα δεδομένα μας, και γενικά στα περιβάλλοντα κινηματογραφικών ταινιών, εμφανιζόταν ένα αξιοσημείωτο πλήθος λάθος εντοπισμών, ή εντοπισμών χαρακτήρων παρασκηνίου. Αυτοί θέτουν μία παραπάνω δυσκολία στη διαδικασία της μάθησης και έτσι έπρεπε να απορριφθούν. Σημειώνουμε ότι με το κατώφλι αυτό, το πλήθος των σωστών εντοπισμών που απορρίπτεται είναι αμελητέο.

5.3.4 Ευθυγράμμιση Προσώπου

Το επόμενο βήμα του αλγορίθμου είναι η ευθυγράμμιση κάθε εντοπισμένου προσώπου (face alignment) έτσι ώστε οι θέσεις των σημείων ενδιαφέροντος να είναι περίπου ίδιες σε όλα. Η χρησιμότητα αυτού του βήματος έγκειται στον περιορισμό των παρενεργειών που μπορούν να προκαλέσουν οι διαφορετικές κλίμακες και πόζες στη διαδικασία της αναγνώρισης. Να σημειώσουμε εδώ ότι με την εξέλιξη της βαθιάς μάθησης οι παρενέργειες αυτές σχεδόν έχουν εξαλειφθεί. Παρ΄ όλα αυτά οι συγγραφείς του [44] που εισήγαγαν το δίκτυο αναγνώρισης προσώπου που χρησιμοποιούμε και εμείς συστήνουν την ευθυγράμμιση των προσώπων για ακόμα καλύτερα αποτελέσματα. Η διαδικασία που ακολουθείται είναι αυτή που προτείνεται στα [53, 19]. Συγκεκριμένα, στο εσωτερικό κάθε bounding box γίνεται εντοπισμός ενός μικρού πλήθους βασικών σημείων ενδιαφέροντος (μάτια, μύτη, στόμα), διαφορετικών για frontal και profile πρόσωπα. Στη συνέχεια υπολογίζεται μέσω ελαχίστων τετραγώνων ένας αφινικός μετασχηματισμός που μετατρέπει τις θέσεις των σημείων αυτών έτσι ώστε να συμπίπτουν όσο το δυνατόν καλύτερα με τις αντίστοιχες θέσεις ενός προεκπαιδευμένου προτύπου (ένα για frontal και ένα για profile - χρησιμοποιήθηκε το ίδιο με αυτό του [4] καθώς δεν έχει βαρύνουσα σημασία). Τέλος, ο μετασχηματισμός αυτός εφαρμόζεται σε ολόκληρο το πρόσωπο προκειμένου να το φέρει στην ευθυγραμμισμένη θέση.

5.3.5 Εξαγωγή Χαρακτηριστικών

Στο σημείο αυτό εξάγουμε 2 διαφορετικές αναπαραστάσεις οι οποίες θα συγκριθούν στη διαδικασία της ταξινόμησης. Η πρώτη αναπαράσταση υλοποιείται στο baseline του [4] και είναι βασισμένη σε hand-crafted χαρακτηριστικά, ενώ η δεύτερη είναι νέα προσθήκη στο σύστημα και αποσκοπεί στην ανάδειξη των χαρακτηριστικών βαθιάς μάθησης (deep feature learning) ως state-of-the-art για το εν λόγω πρόβλημα.

- Η πρώτη αναπαράσταση προκύπτει από την εφαρμογή του γνωστού περιγραφητή SIFT ([33]). Εδώ επανυπολογίζονται τα landmarks του ευθυγραμμισμένου προσώπου και στη συνέχεια ο περιγραφητής εφαρμόζεται σε κάθε ένα από αυτά (12 για frontal πρόσωπα και 7 για profile) σε 2 κλίμακες. Ο τελικός (παγκόσμιος-global) περιγραφητής προκύπτει ως συνένωση των επιμέρους τοπικών (local) και έχει 128*2*12=3072 διαστάσεις για frontal και 128*2*7=1792 διαστάσεις για profile πρόσωπα.

- Η δεύτερη αναπαράσταση προκύπτει από την εφαρμογή του περιγραφητή VGG-face. Λαμβάνεται από την έξοδο του τελευταίου πλήρως συνδεδεμένου επιπέδου της αρχιτε-

κτονικής VGG-Very-Deep-16 CNN που περιγράφεται στο [44] και έχει διάσταση 4096. Η αναπαράσταση υφίσταται κανονικοποίηση L_2 νόρμας, όπως ακριβώς προτείνεται στο [44] ιδιαίτερα για προβλήματα ταυτοποίησης προσώπου. Αυτά είναι αρκετά κοντινά στο δικό μας καθώς και αυτό στηρίζεται στις ομοιότητες μεταξύ προσώπων. Ως είσοδος στο δίκτυο δίνεται το κομμάτι της εικόνας που περικλείεται από το bounding box αφού μετασχηματιστεί προκειμένου να έχει διαστάσεις 224×224 και αφαιρεθεί από αυτό η μέση εικόνα προσώπου που έχει υπολογιστεί από το σύνολο εκπαίδευσης του δικτύου.

5.3.6 Υπολογισμός πυρήνων - kernels

Όπως είδαμε στο κεφάλαιο 4, από την εξίσωση (4.4) προκύπτει ότι η διαδικασία μάθησης που ακολουθούμε δεν χρειάζεται να γνωρίζει τις τιμές των διανυσμάτων χαρακτηριστικών αλλά αρκούν οι τιμές του πίνακα Gram, ο οποίος προκύπτει από την εφαρμογή μίας συνάρτησης πυρήνα $k(\boldsymbol{x_i}, \boldsymbol{x_j})$. Το πρόβλημα της αναγνώρισης προσώπου μπορεί να αντιμετωπιστεί είτε θεωρώντας ως ξεχωριστή οντότητα προς αναγνώριση κάθε εντοπισμένο πρόσωπο, είτε κάθε track. Η δεύτερη μέθοδος πλεονεκτεί καθώς γνωρίζουμε ότι σε ένα track όλα τα πρόσωπα ανήκουν στον ίδιο χαρακτήρα και άρα μπορούμε να πάρουμε περισσότερη πληροφορία για τις διακυμάνσεις και παραλλαγές που εμφανίζονται στο εσωτερικό κάθε κλάσης. Για αυτό το λόγο αναπαριστούμε κάθε πρόσωπο με το σύνολο των διανυσμάτων χαρακτηριστικών κάθε track και εφαρμόζουμε μία συνάρτηση πυρήνα η οποία παίρνει σαν είσοδο ζεύγη tracks, δηλαδή $k(track_i, track_j)$. Ο πυρήνας που επιλέγεται ονομάζεται min-min kernel και παρουσιάζεται στο [53]. Συγκεκριμένα, χρησιμοποιώντας τον συμβολισμό των συγγραφέων λέμε ότι, αν $\mathcal{F}_i$ το σύνολο που περιέχει όλα τα διανύσματα χαρακτηριστικών του track i και $\mathcal{F}_j$ το σύνολο που περιέχει όλα τα διανύσματα χαρακτηριστικών του track j τότε η min-min απόσταση του $\mathcal{F}_i$ από το $\mathcal{F}_j$ ορίζεται ως:

$$d(\mathcal{F}_i, \mathcal{F}_j) = \min_{\mathbf{F_k} \in \mathcal{F}_i} \min_{\mathbf{F_l} \in \mathcal{F}_j} \|\mathbf{F_k} - \mathbf{F_l}\| \tag{5.4}$$

και ο min-min πυρήνας, αν μπορεί να εκφραστεί ως φθίνουσα συνάρτηση της απόστασης, ως:

$$K(i, j) = k(d(\mathcal{F}_i, \mathcal{F}_j)) \tag{5.5}$$

Να σημειώσουμε εδώ ότι ο πίνακας αυτός δεν είναι σίγουρο ότι θα είναι θετικά ημιορισμένος, για αυτό τον μετατρέπουμε πάντα σε τέτοιο μηδενίζοντας τις αρνητικές ιδιοτιμές του. Με βάση τους ορισμούς αυτούς κατασκευάζουμε τους εξής πυρήνες:

- **SIFT38**: Υπολογίζουμε τους min-min rbf kernels όπως αυτοί περιγράφονται στο [53] για κάθε σημείο ενδιαφέροντος και κάθε κλίμακα ξεχωριστά (συνολικά 12*2+7*2=38 χαρακτηριστικά). Δηλαδή, για κάθε χαρακτηριστικό f ο πυρήνας δίνεται από τον τύπο:

$$K_f(i, j) = exp(-\gamma_f d(\mathcal{F}_i^f, \mathcal{F}_j^f)^2) \tag{5.6}$$

Όπου γ_f η παράμετρος που καθορίζει τη διασπορά της γκαουσσιανής και $d(\mathcal{F}_i^f, \mathcal{F}_j^f)$ η min-min απόσταση πάνω στα 2 σύνολα περιγραφητών $\mathcal{F}_i^f, \mathcal{F}_j^f$, όπως αυτή ορίστηκε προηγουμένως. Ο τελικός πυρήνας υπολογίζεται αθροίζοντας με βάρη τους επιμέρους:

$$K(i,j) = \sum_f b_f K_f(i,j) \tag{5.7}$$

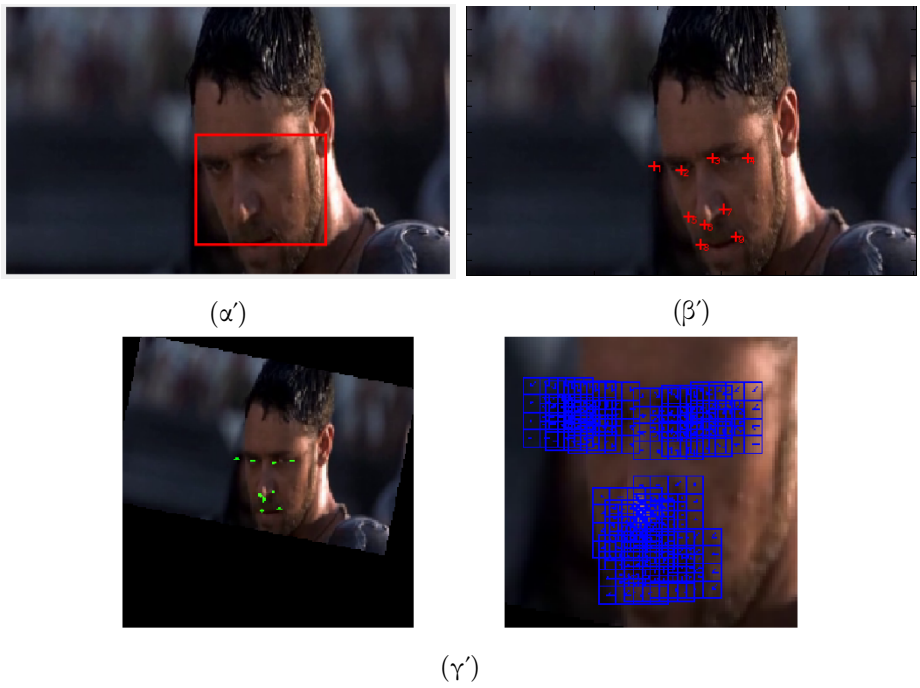
Όπου b_f τα βάρη του αθροίσματος. Αν κάποιο από τα σύνολα $\mathcal{F}_i^f$ είναι κενό τότε ο πυρήνας τίθεται ίσος με 0. Τα βάρη ορίζονται να είναι ομοιόμορφα και ίσα με 1 δια το πλήθος των χαρακτηριστικών για τα οποία και τα 2 σύνολα $\mathcal{F}_i^f$, $\mathcal{F}_j^f$ είναι μη κενά. Αν το πλήθος αυτό είναι 0, τότε και ο συνολικός πυρήνας είναι 0

- **VGG2**: Υπολογίζουμε τους min-min rbf kernels ξεχωριστά για frontal και profile πρόσωπα. Δηλαδή, θεωρούμε 2 χαρακτηριστικά, 1 για frontal και 1 για profile. Ο τελικός πυρήνας υπολογίζεται και πάλι με άθροιση με βάρη (για την ακρίβεια ως μέσος όρος) όπως πριν.

- **VGG1**: Υπολογίζουμε τους min-min rbf kernels για όλους τους πιθανούς συνδυασμούς frames μεταξύ του ζεύγους tracks είτε αυτά είναι profile είτε είναι frontal. Αυτό σημαίνει ότι τα σύνολα $\mathcal{F}_i^f$, $\mathcal{F}_j^f$ είναι πάντα μη κενά και υπάρχει 1 μόνο χαρακτηριστικό, άρα ο τελικός πυρήνας προκύπτει απευθείας. Αυτή η μορφή πυρήνα ταιριάζει καλύτερα με την global αναπαράσταση του περιγραφητή VGG, αλλά δεδομένου ότι συγκρίνει frontal με profile πρόσωπα, πιθανόν να δίνει κάποιες φορές μεγαλύτερες τιμές πυρήνα από τις αναμενόμενες για πρόσωπα διαφορετικής ταυτότητας. Δεδομένης όμως της εκπαίδευσης της βαθιάς αρχιτεκτονικής με πρόσωπα πολλών διαφορετικών γονιών θέασης αναμένουμε να μην υπάρχει μεγάλη διαφορά μεταξύ του VGG1 και του VGG2.

5.4 Αναπαράσταση Ανθρώπινων Δράσεων

Το σύστημα εντοπισμού και αναπαράστασης ανθρώπινων δράσεων είναι αρκετά πιο απλοϊκό από αυτό του προσώπου καθώς δεν ελήφθη συγκεκριμένη μέριμνα για την χωροχρονική ανίχνευσή τους, αλλά χρησιμοποιήθηκαν τα ήδη διαθέσιμα αποσπάσματα που επισημειώθηκαν για τις ανάγκες της βάσης COGNIMUSE. Ο λόγος που έγινε αυτό είναι η μεγάλη δυσκολία του προβλήματος του χωροχρονικού εντοπισμού, καθώς δεν υπάρχουν ακόμα τόσο εξελιγμένες και γενικές μέθοδοι όπως αυτές του εντοπισμού προσώπου. Το βασικό μειονέκτημα των περισσότερων είναι η αδυναμία τους να γενικεύσουν τις δυνατότητες εντοπισμού σε κλάσεις οι οποίες ήταν άγνωστες κατά τη διάρκεια της εκπαίδευσης. Αυτό μπορεί να αντιμετωπιστεί επανεκπαιδεύοντας κάποιο σύστημα εντοπισμού με παραδείγματα δράσεων που ανήκουν στις κλάσεις που θέλουμε να αναγνωρίσουμε. Δεδομένης όμως της περαιτέρω αβεβαιότητας που θα εισαγόταν στο σύστημα μάθησης που προτείνουμε, αποφύγαμε αυτή τη μεθοδολογία προκειμένου να αξιολογήσουμε τους αλγορίθμους μας αυτούς καθαυτούς. Στόχος είναι, όμως, σε



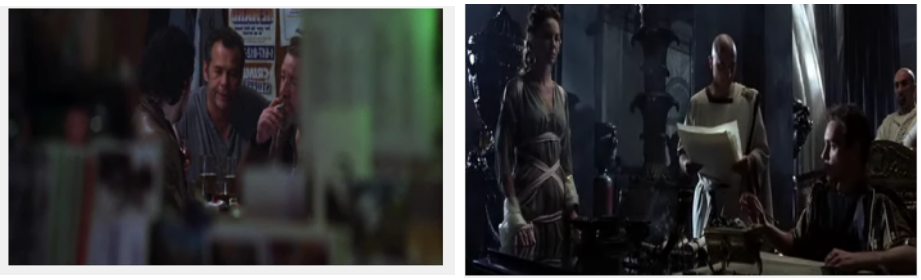
(α′) (β′)

(γ′)

Σχήμα 5.2: (α′): Παράδειγμα του bounding box ενός εντοπισμένου προσώπου. (β′): Υπολογισμός facial landmarks και (γ′) ευθυγράμμιση προσώπου - επανυπολογισμός facial landmarks και εξαγωγή SIFT descriptor γύρω από το καθένα.

μελλοντική έρευνα να ελεγχθεί η δυνατότητα μίας πλήρους κατανόησης του βίντεο χωρίς την ανθρώπινη παρέμβαση που ενέχει η χρήση των επισημειωμένων αποσπασμάτων. Άλλωστε, τα αποσπάσματα αυτά περιλαμβάνουν και περιπτώσεις δράσεων που δεν συνεισφέρουν ιδιαίτερα στην κατανόηση (δεν έχουν μεγάλη σημασία - saliency) και είναι ταυτόχρονα δύσκολες να αναγνωριστούν καθώς δεν περιγράφονται στο συνοδευτικό κείμενο (βλέπε σχήμα 5.3). Έτσι, ένα σύστημα αυτόματου εντοπισμού, παρόλο που δεν μπορεί να προσφέρει μία πλήρη καταγραφή, εκτιμάμε ότι θα οδηγήσει σε υψηλότερα ποσοστά αναγνώρισης και ταυτόχρονα σε ανάκτηση δράσεων πιο πλούσιων σε πληροφορία. Μία πιθανή λύση για αυτό το ζήτημα είναι ο συνδυασμός ενός συστήματος εντοπισμού του ανθρώπινου κορμού και ενός συστήματος ανίχνευσης έντονης κίνησης. Περισσότερα για τις μεθόδους εντοπισμού δράσεων μπορεί να βρει κανείς στα [45, 29, 23], ενώ κάποιες σύγχρονες μέθοδοι περιγράφονται στα [27, 65, 64, 63] κ.ά.

Η αναπαράσταση των δράσεων έγινε με τη χρήση του συνελικτικού δικτύου C3D ([58]) και σύμφωνα με αυτά που προτείνουν οι εμπνευστές του. Δηλαδή, κάθε απόσπασμα προς αναγνώριση χωρίζεται σε τμήματα αποτελούμενα από 16 καρέ (συνήθως επικαλυπτόμενα με επικάλυψη 50%), τα οποία δίνονται ως είσοδος στο δίκτυο. Η αναπαράσταση του καθενός λαμβάνεται ως η έξοδος του πλήρως συνδεδεμένου επιπέδου fc6. Η τελική αναπαράσταση λαμβάνεται ως ο μέσος όρος των επιμέρους, ακολουθούμενος από L_2 κανονικοποίηση. Τα χαρακτηριστικά που παρέχει το C3D θεωρούνται generic και μπορούν να χρησιμοποιηθούν σε ποικίλα προβλήματα αναγνώρισης σε βίντεο.



(α′) (β′)

Σχήμα 5.3: (α′,β′,γ′) Στιγμιότυπα αποσπασμάτων για τις δράσεις smoke (α′) και wave hands (β′) τα οποία έχουν μικρή σημασία, είναι δύσκολο να αναγνωριστούν (ειδικά στη δεύτερη περίπτωση ακόμα και ένας άνθρωπος δεν θα παρατηρούσε την δράση) και προφανώς δεν αναφέρονται στο σενάριο.

Τέλος για κάθε ζεύγος $\boldsymbol{x}, \boldsymbol{y}$ διανυσμάτων χαρακτηριστικών υπολογίστηκαν 2 είδη πυρήνων:

- $\boldsymbol{\chi^2}$**C3D**: Ο πυρήνας χ^2 που δίνεται από τον τύπο :

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\gamma \sum_i \frac{(x[i] - y[i])^2}{x[i] + y[i]}\right) \tag{5.8}$$

  Όπου $x[i]$ η τιμή του $\boldsymbol{x}$ στη διάσταση i.

- **linC3D**: Ο γραμμικός πυρήνας που δίνεται από τον τύπο:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y} \tag{5.9}$$

  Υπενθυμίζουμε εδώ ότι τα διανύσματα είναι ήδη κανονικοποιημένα, οπότε δεν χρειάζεται να διαιρέσουμε την τιμή του πυρήνα με το γινόμενο των μέτρων των 2 διανυσμάτων.

Κεφάλαιο 6

Εξόρυξη πληροφορίας από το κείμενο - Απόδοση ασθενών ετικετών

6.1 Εισαγωγή

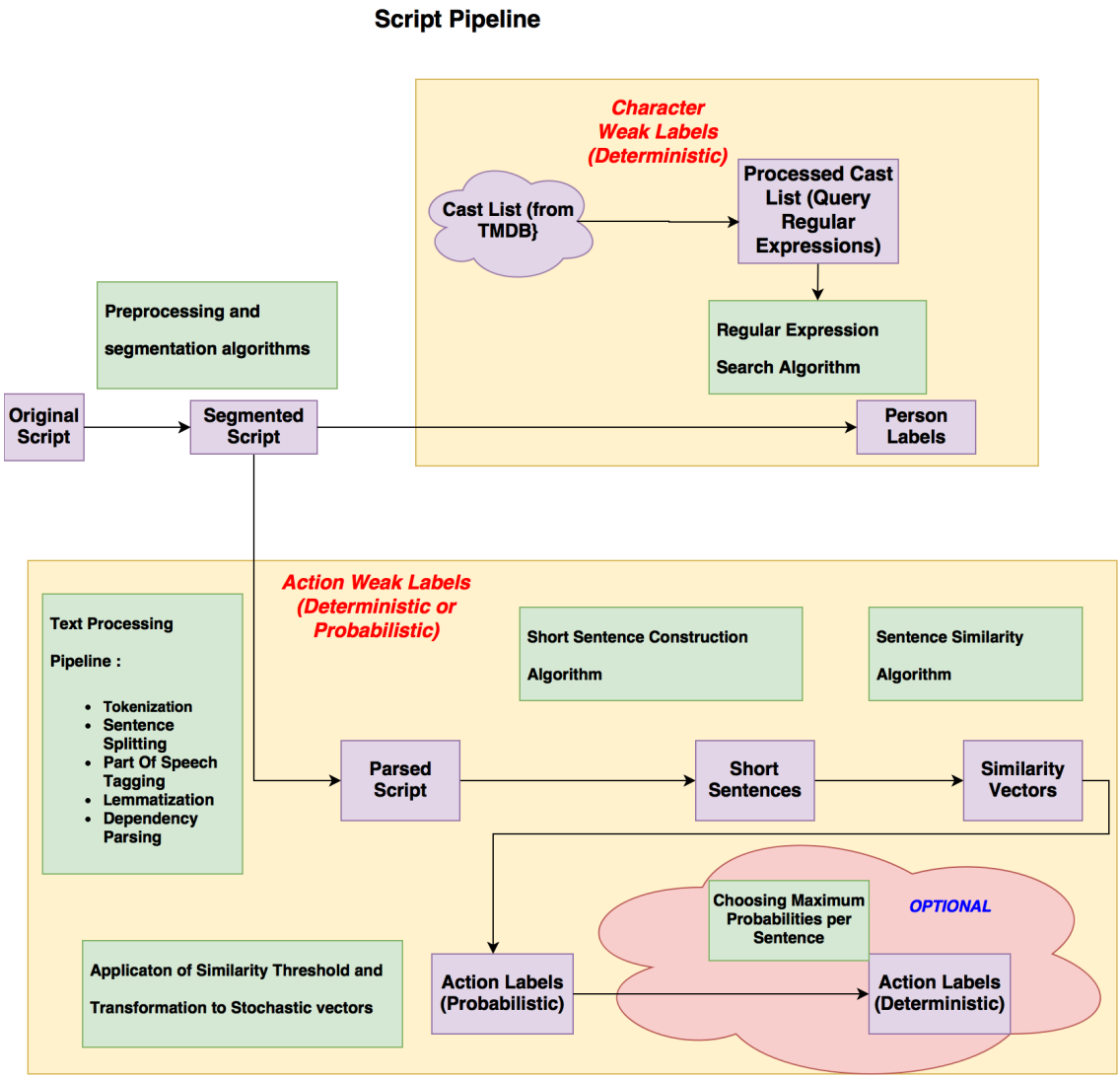
Στο παρακάτω κεφάλαιο παρουσιάζονται οι μέθοδοι επεξεργασίας κειμένου που εφαρμόσαμε προκειμένου να αναλύσουμε σημασιολογικά τα σενάρια των ταινιών και να προσφέρουμε στον αλγόριθμο τις ασθενείς ετικέτες (weak labels) που χρειάζεται προκειμένου να υλοποιηθούν τα διάφορα σενάρια μάθησης. Στο σχήμα 6.1 αναπαριστούμε κάθε στάδιο της επεξεργασίας που οδηγεί στον καθορισμό των τελικών ετικετών. Τα προβλήματα που καλούμαστε να αντιμετωπίσουμε εδώ είναι η **επιλογή του συνόλου ετικετών** και η **εύρεσή τους μέσα στο κείμενο**. Η τελική μορφή της επίβλεψης του αλγορίθμου μάθησης λαμβάνεται με την απόδοση χρονικών ορίων στις ετικέτες που εξήχθησαν, με τη χρήση ενός αλγορίθμου ευθυγράμμισης του σεναρίου με τους υπότιτλους.

6.2 Προεπεξεργασία και Κατάτμηση Σεναρίου

Το πρώτο στάδιο, όπως φαίνεται και από το σχήμα 6.1, είναι η μετατροπή του κειμένου σε μία μορφή κατάλληλη για επεξεργασία. Για αυτό το λόγο αφαιρούνται κάποια περιττά σημεία στίξης, κενά, νούμερα όπως αρίθμηση σελίδων κ.ά. Στη συνέχεια εκμεταλλευόμενοι τη δομή που έχουν σχεδόν όλα τα σενάρια ταινιών [1] χωρίζουμε το κάθε κείμενο στα εξής 4 τμήματα:

- Περιγραφή Σκηνής - SCENE
- Όνομα ομιλητή - SPEAKER
- Μονόλογος - MONOLOGUE

[1] Σε περιπτώσεις που η δομή αυτή δεν υπήρχε εξ΄ αρχής ήταν εύκολο να την κατασκευάσουμε εφαρμόζοντας αντικαταστάσεις μέσω κανονικών εκφράσεων



Σχήμα 6.1: Διάγραμμα του συστήματος εντοπισμού και αναπαράστασης προσώπου

- Περιγραφή γεγονότων - DESCRIPTION

Η διαδικασία της κατάτμησης είναι ιδιαίτερα εύκολη ακριβώς λόγω της γενικευμένης αυτής δομής και στηρίζεται σε απλές παρατηρήσεις. Συγκεκριμένα, όπως φαίνεται στο σχήμα 6.2 παρατηρούμε ότι οι περιγραφές σκηνών γράφονται με κεφαλαία γράμματα και χωρίς εσοχή, τα ονόματα των ομιλητών γράφονται επίσης με κεφαλαία γράμματα και με μία μικρή εσοχή, οι μονόλογοι είτε με κεφαλαία είτε με μικρά και με μεγαλύτερη εσοχή, ενώ οι περιγραφές και πάλι με κεφαλαία ή μικρά και χωρίς εσοχή. Τα χαρακτηριστικά αυτά ανιχνεύονται με προφανή

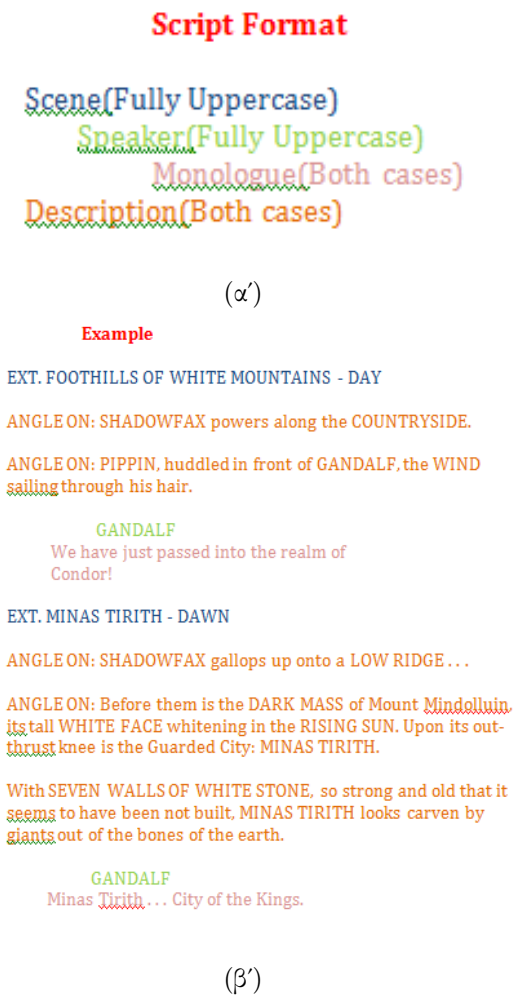
τρόπο και έτσι με μεγάλη ευκολία μπορούμε να χωρίσουμε το κείμενο σε μέρη που προσφέρουν πλούσια πληροφορία το καθένα. Για παράδειγμα, οι κλάσεις των πρωταγωνιστών μπορούν να αναζητηθούν στα κομμάτια του κειμένου που επισημειώνονται ως SPEAKER, ή οι σκηνές (για ένα παρόμοιο πρόβλημα αναγνώρισης σκηνών σε ταινίες) σε αυτά που επισημειώνονται ως SCENE. Εμείς χρησιμοποιούμε αυτή τη δομή με 2 τρόπους. Πρώτον, για την ευθυγράμμιση του σεναρίου με τους υπότιτλους χρησιμοποιούμε μόνο τα κομμάτια που επισημειώνονται ως MONOLOGUE. Δεύτερον, εξάγουμε τις ετικέτες των ανθρώπινων προσώπων - χαρακτήρων μόνο από τα τμήματα SPEAKER και DESCRIPTION και τις ετικέτες των δράσεων μόνο από τα τμήματα DESCRIPTION. Ο λόγος είναι ότι τα υπόλοιπα τμήματα μπορεί να δώσουν αποπροσανατολιστική πληροφορία, που να μην εμφανίζεται στο βίντεο εκείνη την ώρα. Για παράδειγμα σε ένα διάλογο οι συνομιλητές συχνά περιγράφουν κάτι άσχετο που δεν λαμβάνει χώρα στην ίδια τη σκηνή.

Τέλος, για να διευκολύνουμε το ταίριασμα κανονικών εκφράσεων και τους αλγορίθμους επεξεργασίας φυσικής γλώσσας που θα περιγραφούν παρακάτω, διορθώσαμε τα κεφαλαία γράμματα σε όλο το κείμενο, μετατρέποντας τις λέξεις που περιείχαν μόνο κεφαλαία σε λέξεις που κεφαλαίο είναι μόνο το αρχικό τους γράμμα.

6.3 Απόδοση Ετικετών για το Πρόβλημα της Αναγνώρισης Προσώπου

6.3.1 Σύνολο Ετικετών για το Πρόβλημα της Αναγνώρισης Προσώπου

Όπως προαναφέραμε, η καταγραφή των χαρακτήρων του βίντεο απαιτεί αρχικά να βρούμε το σύνολο αυτών που εμφανίζονται σε αυτό. Εδώ το συνοδευτικό κείμενο είναι μία πλούσια πηγή πληροφορίας και είναι αρκετό προκειμένου να καθοριστεί το σύνολο των ετικετών των χαρακτήρων που έχουν πρωταγωνιστικό ρόλο στην ταινία (και άρα έχουν μεγάλη σημασία στην αυτόματη κατανόηση της). Για παράδειγμα, στα [9, 53, 19] το σύνολο των χαρακτήρων καθορίζεται από τα μοναδικά (διαφορετικά μεταξύ τους) ονόματα (λέξεις ή φράσεις) που αναφέρονται στα τμήματα του σεναρίου που έχουν επισημειωθεί ως SPEAKER. Αυτός ο τρόπος έχει το πλεονέκτημα ότι δεν χρειάζεται κάποια ανθρώπινη παρέμβαση, αλλά δεν λαμβάνει υπόψιν ότι συχνά κάποιοι χαρακτήρες αναφέρονται με περισσότερα από ένα ονόματα σε μία ταινία και στο σενάριο της (όνομα και επίθετο ή κάποιο ψευδώνυμο) και επίσης αγνοεί χαρακτήρες που δεν μιλούν (αυτό είναι λιγότερο προβληματικό, καθώς αυτοί συνήθως δεν έχουν πρωταγωνιστικό ρόλο). Μία διαφορετική μέθοδος είναι η χρήση ενός συστήματος Name Entity Recognition, το οποίο εντοπίζει στο κείμενο λέξεις ή φράσεις που έχουν την έννοια κάποιου ονόματος ανθρώπου, περιοχής, οργανισμού κ.τ.λ. Το σύνολο ετικετών μπορεί κατόπιν να επιλεγεί μέσω των μοναδικών φράσεων από αυτές που έχουν εντοπιστεί. Με αυτόν τον τρόπο διευρύνεται μεν το σύνολο ετικετών και σε μη ομιλούντες χαρακτήρες, αλλά και πάλι δεν είναι εύκολο να συμπεράνουμε τους διαφορετικούς τρόπους αναφοράς ενός ατόμου. Ακόμα, δοκιμάζοντας αυτήν την τεχνική παρατηρήσαμε ότι τα αποτελέσματα περιείχαν πολλά False



Σχήμα 6.2: Δομή σεναρίου και παράδειγμα

Positives (φράσεις που δεν ήταν στην πραγματικότητα ονόματα) και False Negatives (αποτυχία επισημείωσης κάποιων φράσεων) και για αυτό το λόγο απορρίφθηκε. Τελικά, επιλέξαμε την εύρεση του πλήρους συνόλου ετικετών όμοια με το [12] μέσω του online API που παρέχει η ιστοσελίδα **TMDB** (`https://www.themoviedb.org/`). Από αυτό αποκτούμε τη λίστα με τα πλήρη ονόματα των χαρακτήρων η οποία ορίζει και το σύνολο ετικετών. Γενικότερα, λόγω της δημοφιλίας των ταινιών, υπάρχουν βάσεις δεδομένων που προσφέρουν μεγάλο όγκο πληροφοριών για αυτές. Αυτό τις καθιστά ιδανικές για τέτοια προβλήματα αναγνώρισης και για το σχεδιασμό αλγορίθμων (πολυτροπικής και μη) κατανόησης βίντεο. Φυσικά, αυτό ενέχει μία έμμεση ανθρώπινη παρέμβαση, η οποία, παρότι εδώ δεν έχει μεγάλη σημασία, σε άλλες κατηγορίες βίντεο με συνοδευτικό κείμενο (όπως αυτά στο YouTube) δεν θα είναι διαθέσιμη. Έτσι, σημειώνουμε την ανάγκη ύπαρξης κάποιας εναλλακτικής και πιο γενικής μεθοδολογίας.

6.3.2 Εντοπισμός Ετικετών για το Πρόβλημα της Αναγνώρισης Προσώπου

Η λίστα των πρωταγωνιστών (cast list) που αποκτούμε χρειάζεται μία προεπεξεργασία, προκειμένου να απομονωθεί η ουσιαστική και σημαντική πληροφορία. Για αυτό το λόγο αφαιρούνται χαρακτήρες που το όνομα τους δεν εμφανίζεται στους συντελεστές της ταινίας (uncredited), χαρακτήρες που ακούγεται μόνο η φωνή τους (Voice) κ.ά. Στη συνέχεια, κάθε όνομα χωρίζεται στις επιμέρους λέξεις του, κάποιες από τις οποίες αφαιρούνται (π.χ όταν βρίσκονται μετά από προθέσεις κ.ά). Δεν θα επεκταθούμε παραπάνω στην επεξεργασία της λίστας των πρωταγωνιστών καθώς είναι απλοί κανόνες που συγκεντρώθηκαν παρατηρώντας τα δεδομένα μας (data driven). Κάποια παραδείγματα επεξεργασίας της λίστας δείχνουμε στον πίνακα 6.1. Τελικά, οι λέξεις που απομένουν είναι το πολύ 3 ανά χαρακτήρα (4 σε σπάνιες περιπτώσεις) και αποτελούν το όνομα, το επίθετο και πιθανόν κάποιο μεσαίο όνομα ή ψευδώνυμο. Από κάθε μία κατασκευάζεται μία κανονική έκφραση έτσι ώστε να περιέχει κάποιες πιθανές εναλλακτικές χρήσεις κεφαλαίων και μικρών γραμμάτων. Τελικά, χρησιμοποιώντας ταίριασμα κανονικών εκφράσεων - Regular Expression Matching αποδίδουμε σε κάθε λέξη που ταιριάζεται, την ετικέτα του αντίστοιχου χαρακτήρα.

Σημειώνουμε ότι οι 2 πρώτες τεχνικές εύρεσης του συνόλου ετικετών που αναφέραμε (SPEAKER names και NER) μπορούν να φανούν χρήσιμες στην περίπτωση που το σενάριο της ταινίας διαφέρει από αυτήν που τελικά διανέμεται στους κινηματογράφους, όπως συμβαίνει στην ταινία Beautiful Mind - BMI της βάσης μας. Εκεί, κάποια ονόματα έχουν αλλαχθεί στην τελική έκδοση της ταινίας και άρα η λίστα πρωταγωνιστών διαφέρει από το σενάριο. Είναι προφανές ότι αυτό είναι ιδιαίτερα σπάνιο και άρα η πιθανότητα αποτυχίας της προτεινόμενης μεθόδου είναι μικρή.

Τέλος να πούμε ότι μία πιθανή επέκταση της μεθόδου είναι η εφαρμογή κάποιου αλγορίθμου Επίλυσης Συναναφοράς - Coreference Resolution προκειμένου να λάβουμε πιο πλούσια πληροφορία για τις εμφανίσεις κάθε χαρακτήρα. Η πληροφορία αυτή δεν μπορεί να εντοπιστεί με καμία από τις προηγούμενες μεθόδους καθώς έχει τη μορφή αντωνυμιών ή ουσιαστικών που αντικαθιστούν συχνά το κύριο όνομα. Μία προσέγγιση στο πρόβλημα παρουσιάζεται στο [47].

6.4 Απόδοση Ετικετών για το Πρόβλημα της Αναγνώρισης Δράσης

6.4.1 Σύνολο Ετικετών για το Πρόβλημα της Αναγνώρισης Δράσεων

Το πρόβλημα της κατανόησης της σημασιολογίας του βίντεο ως προς τις ανθρώπινες δράσεις που περιέχει είναι αρκετά πιο πολύπλοκο από αυτό των ανθρώπινων προσώπων. Η δυσκολία εντοπίζεται αφενός στην εύρεση του συνόλου των διαφορετικών δράσεων που εμφανίζονται και αφετέρου στον εντοπισμό τους. Παρουσία του συνοδευτικού κειμένου, η δυσκολία περιορίζεται καθώς ανάγεται σε ένα πρόβλημα αυτόματης κατανόησης κειμένου όπου εκεί η

Table 6.1: Πίνακας παραδειγμάτων επεξεργασίας της λίστας χαρακτήρων

Example	Output (Query Regular Expresion)	Rule
Girl **in** Bar	Removed	**in** token
Billy **'s** Aunt:	Removed	**'s / s'** token (possesives)
Reporter **(uncredited)**	Removed	**(uncredited)** token
Treebeard **(Voice)**	Removed	**(Voice)** token
John Nash	• John • (J|j)ohn (N|n)ash • Nash	none
Gandalf **the** White	Gandalf	**the** token : remove the rest that follows
Théoden, King **of** Rohan	• Theoden • King	**of** token : remove the rest that follows
Younger Priest	• Younger • Priest • (Y|y)ounger (P|p)riest	none

έρευνα είναι πιο ώριμη. Αναφορικά με τον καθορισμό του συνόλου των ετικετών (των πιθανών δράσεων δηλαδή), έχουν προταθεί διάφορες μέθοδοι στην υπάρχουσα βιβλιογραφία η ειδοποιός διαφορά των οποίων είναι κατά βάση ο βαθμός της ανθρώπινης παρέμβασης. Συγκεκριμένα, έχουν προταθεί τεχνικές πλήρως αυτοματοποιημένες όπου δεν καθορίζονται διακριτές κλάσεις αλλά η σημασιολογική αναπαράσταση κάθε φράσης του κειμένου. Για παράδειγμα στα [48, 49, 42] εκπαιδεύεται ένα σύστημα μηχανικής μετάφρασης (machine translation) από τη μία τροπικότητα στην άλλη, δηλαδή εξαγωγής προτάσεων που περιγράφουν ένα βίντεο και αντίστροφα, ή στο [6] υλοποιείται μία ευθυγράμμιση των 2 τροπικοτήτων. Αυτές πετυχαίνουν υψηλού επιπέδου κατανόηση η οποία, δεδομένου του ότι εκφράζεται με φυσική γλώσσα, είναι απευθείας αντιληπτή από τον άνθρωπο. Χρειάζονται όμως μεγάλο πλήθος δεδομένων για να εκπαιδευτούν ή/και πολύ καλή αρχική χρονική ευθυγράμμιση των 2 τροπικοτήτων. Κανένα

από τα 2 αυτά χαρακτηριστικά δεν παρουσιάζεται στο πρόβλημα αυτόματης κατανόησης μιας ταινίας. Επίσης, με εξαίρεση το [42], δεν επιλύουν το πρόβλημα της μάθησης των δράσεων αλλά περισσότερο αυτού του συμπερασμού γλωσσικών περιγραφών. Το [42] είναι πιο κοντά στη δική μας εργασία αλλά η εκπαίδευση του χρειάζεται πολύ μεγάλο πλήθος βίντεο (από το YouTube) συνοδευόμενο από λίγες προτάσεις για το καθένα. Οι εργασίες που εστιάζουν περισσότερο στο πρόβλημα της αναγνώρισης δράσεων ([17, 37, 31, 4, 32]) επιλέγουν από πριν ένα σύνολο ετικετών τις οποίες και εντοπίζουν μέσα στο κείμενο. Έτσι, δεδομένης της προϋπάρχουσας επισημείωσης των ταινιών και προκειμένου να δούμε το πρόβλημα υπό το πρίσμα της αναγνώρισης δράσεων καθορίσαμε και εμείς a-priori των σύνολο ετικετών μας βάσει όσων αναφέρθηκαν στο κεφάλαιο 5 και πειραματιστήκαμε με διάφορα υποσύνολα του (βλέπε κεφάλαιο 7). Ως εναλλακτική πρόταση εστιασμένη στο πρόβλημα της αναγνώρισης δράσεων αναφέρουμε την ομαδοποίηση (clustering) ρημάτων με βάση τα νοήματα τους (word senses) όπως αυτά παρέχονται από κάποιο λεξικό όπως το wordnet [39, 40] και το verbnet [30]. Η πρόταση αυτή θα ερευνηθεί σε μελλοντική εργασία.

6.4.2 Εντοπισμός Ετικετών για το Πρόβλημα της Αναγνώρισης Δράσεων

Θεωρώντας ως δεδομένο ένα σύνολο ετικετών προκύπτει το εξής ερώτημα. *Σε ποια σημεία του κειμένου περιγράφεται η κάθε δράση;*. Η απάντηση σε αυτό δεν είναι όσο προφανής είναι η στο πρόβλημα της αναγνώρισης χαρακτήρων, καθώς η έννοια μίας ανθρώπινης δράσης κρύβει πιο σύνθετη σημασιολογία. Στα [32, 17, 37] εκπαιδεύεται ένας γραμμικός ταξινομητής από ένα πλήθος γλωσσικών παραδειγμάτων (με αναπαράσταση bag of words) που υπονοούν κάθε κλάση. Στη συνέχεια ο ταξινομητής αυτός εφαρμόζεται σε κάθε τμήμα DESCRIPTION του κειμένου επισημειώνοντας το με την ύπαρξη ή μη της κάθε κλάσης. Η τεχνική αυτή ενέχει αφενός το μειονέκτημα του μεγάλου βαθμού της ανθρώπινης παρέμβασης προκειμένου να συγκεντρωθεί το σύνολο εκπαίδευσης και αφετέρου την αδυναμία να συμπεριληφθούν πιο σύνθετες σημασιολογίες. Στο [4], που είναι και το βασικό baseline μας, προκειμένου να εντοπιστούν οι ετικέτες χρησιμοποιήθηκε το σύστημα SEMAFOR ([14]). Αυτό κατηγοριοποιεί τα διάφορα μέρη του κειμένου, με βάση τη σημασιολογία τους, στα λεγόμενα Frames. Οι συγγραφείς του [4] ανέλυσαν το κείμενο στα Frames του και επέλεξαν τα 2 πιο συχνά εμφανιζόμενα (συγκεκριμένα το ChangePosture και το SelfMotion). Στη συνέχεια, επέλεξαν το ρήμα με τις περισσότερες εμφανίσεις σε κάθε frame (sitting down και walking) προκειμένου να αποδώσουν την ετικέτα. Με τη διαδικασία αυτή καθόρισαν αυτόματα το σύνολο ετικετών και από κοινού με τον εντοπισμού τους, αλλά γίνεται αντιληπτό ότι είναι αρκετά περιοριστική και δεν μπορεί να γενικευθεί. Παρ΄ όλο που περιέχει μικρή ανθρώπινη παρέμβαση δεν μπορεί, όπως και οι προηγούμενες να ενσωματώσει σύνθετη σημασιολογία. Ο βασικός λόγος είναι ότι μία δράση μπορεί να αναφέρεται με πολλούς διαφορετικούς τρόπους, ακόμα και με ρήματα με εντελώς διαφορετική ρίζα.

Για αυτό το λόγο στην παρούσα διπλωματική εφαρμόσαμε μία διαφορετική προσέγγιση εξόρυξης τέτοιου είδους πληροφορίας από το κείμενο. Συγκεκριμένα, παρατηρώντας τους

τρόπους με τους οποίους περιγράφονται διάφορες δράσεις στα σενάρια ταινιών είδαμε ότι οι σεναριογράφοι χρησιμοποιούν ποικιλία στο λόγο τους για να αποδώσουν ίδια νοήματα. Έτσι, θεωρούμε ότι οι ετικέτες που κρύβονται μέσα στο κείμενο μπορούν να μοντελοποιηθούν καλύτερα υπολογίζοντας ακριβώς αυτήν την ομοιότητα μεταξύ των νοημάτων. Για να το πετύχουμε αυτό, υπολογίζουμε τη σημασιολογική ομοιότητα προτάσεων του κειμένου με τις φράσεις που συνθέτουν το σύνολο των ετικετών μας. Το ολοκληρωμένο σύστημα εξαγωγής των ετικετών παρουσιάζεται στο σχήμα 6.1 και αποτελείται από τα εξής στάδια:

- Επεξεργασία Κειμένου με χρήση τεχνικών Επεξεργασίας φυσικής Γλώσσας (NLP). Υλοποιείται με τη βοήθεια της εργαλειοθήκης CoreNLP [35]. Στόχος είναι η συντακτική ανάλυση των προτάσεων του κειμένου προκειμένου να συνθέσουμε μικρότερες προτάσεις που θα δοθούν σαν είσοδοι στο σύστημα υπολογισμού σημασιολογικής ομοιότητας. Τα υποστάδια είναι τα εξής:

  1. Διάσπαση σε λέξεις-Tokenization
  2. Διάσπαση σε προτάσεις-Sentence Splitting
  3. Επισημείωση Μέρους του Λόγου-Part Of Speech Tagging (POS Tagging)
  4. Λημματοποίηση-Lemmatization
  5. Εξαγωγή Εξαρτήσεων-Dependency Parsing

- Σύνθεση υποψήφιων φράσεων

  Προκειμένου να συνθέσουμε τις υποψήφιες φράσεις (δηλαδή αυτές που πιθανόν να εκφράζουν κάποια δράση) βρίσκουμε όλα τα ρήματα του κειμένου, εκτός από τα βοηθητικά (be, can κ.τ.λ) και αυτά που βρίσκονται σε αρνητική μορφή, με τη βοήθεια του POS Tagging. Στη συνέχεια χρησιμοποιώντας την έξοδο του Dependency Parser εντοπίζουμε σε κάθε συντακτικό υποδέντρο του οποίου η ρίζα είναι ένα από τα ρήματα, κάποιους όρους φύλα. Αυτοί οι όροι είναι οι εξής[2]:

  1. **dobj**: Το άμεσο αντικείμενο του ρήματος
  2. **compound:prt**: Η πρόθεση ενός phrasal verb
  3. **advmod**: Επίρρημα που τροποποιεί το νόημα του ρήματος
  4. **nmod**: Ουσιαστικά που λειτουργούν σαν επιρρήματα τροποποιώντας το νόημα του ρήματος

  Παραδείγματα:

  1. Denethor slumps (**verb**) back (**advmod**) into (**preprosition of nmod**) his Seat (**nmod**)

[2]Χρησιμοποιούμε την ορολογία των Univesal Dependencies που εισήχθη από το πανεπιστήμιο του Stanford και δίνεται ως έξοδος από το CoreNLP toolbox. Περισσότερα για αυτήν αναφέρονται στο [15] και στο documentation των Univesal Dependencies (http://universaldependencies.org/en/dep/index.html)

2. Gandalf hurries (**verb**) off (**compound:prt**)
3. Gandalf lowers (**verb**) his voice (**dobj**)

- Υπολογισμός Ομοιοτήτων

  Το τελευταίο στάδιο είναι η εφαρμογή ενός αλγορίθμου υπολογισμού ομοιότητας φράσεων / προτάσεων. Εδώ υπολογίζουμε για κάθε υποψήφια φράση την ομοιότητά της με κάθε μία από τις γλωσσικές εκφράσεις των ετικετών των δράσεων. Έτσι αποκτούμε ένα σύνολο διανυσμάτων ομοιοτήτων με διάσταση ίση με το πλήθος των κλάσεων.

  Το πρόβλημα της σημασιολογικής ομοιότητας έχει συγκεντρώσει ιδιαίτερο ενδιαφέρον στην κοινότητα της επεξεργασία φυσικής γλώσσας. Εμείς για τις ανάγκες τις διπλωματικής εργασίας αξιοποιούμε ένα off - the - shelf σύστημα που παρουσιάστηκε στο [24] και έλαβε υψηλή κατάταξη στο task STS του συνεδρίου SEM 2013. Ο λόγος που το επιλέξαμε είναι η εύκολη ενσωμάτωση του σε ένα συνολικό σύστημα μέσω του διαδικτυακού API που παρέχουν οι δημιουργοί του και η υβριδική υλοποίησή του. Συγκεκριμένα, για τον υπολογισμό ομοιότητας λέξεων συνδυάζει μία μέθοδο Latent Semantic Analysis εμπλουτισμένη με πληροφορίες από το λεξικό του Wordnet ([39]). Αρχικά, σύμφωνα με τις γνωστές πρακτικές, υπολογίζεται ένας πίνακας συναναφορών λέξεων μέσα από ένα μεγάλο όγκο κειμένων, από τον οποίο εξάγονται οι διανυσματικές αναπαραστάσεις (κατόπιν εφαρμογής truncated SVD). Η πρωταρχική τιμή ομοιότητας υπολογίζεται με την εφαρμογή της μετρικής cosine similarity πάνω σε ένα ζεύγος διανυσματικών αναπαραστάσεων. Οι αναπαραστάσεις αυτές, όμως, συνήθως υποφέρουν από χαμηλές τιμές ομοιοτήτων μεταξύ λέξεων που είναι συνώνυμα αλλά έχουν ταυτόχρονα πολλά διαφορετικά νοήματα. Ο λόγος είναι ότι αυτές οι λέξεις αναφέρονται με πολλά διαφορετικά συμφραζόμενα και άρα οι συναναφορές τους δεν είναι αρκετά πληροφοριακές. Για αυτό το λόγο, η τιμή ομοιότητας για δύο λέξεις ενισχύεται για κάποιες περιπτώσεις, όπως είναι η συμμετοχή τους στο ίδιο σύνολο συνωνύμων του Wordnet. Προκειμένου τώρα να υπολογιστεί η ομοιότητα για δύο φράσεις, υπολογίζεται ένας βαθμός σημασιολογικής ευθυγράμμισης τους, βασισμένος στους επιμέρους υπολογισμούς ομοιοτήτων των λέξεων των δύο φράσεων. Περισσότερες λεπτομέρειες για το σύστημα στο [24].

  Δεν θα επεκταθούμε παραπάνω καθώς το συγκεκριμένο πεδίο ξεπερνά τα όρια της παρούσας διπλωματικής. Αναφέρουμε μόνο ότι μία εναλλακτική λύση που έχει αρχίσει να επικρατεί τα τελευταία χρόνια είναι τα λεγόμενα word embeddings όπως τα μοντέλα word2vec. Ειδικά για την μελλοντική υλοποίηση της από κοινού μάθησης γλωσσικών και οπτικών αντικειμένων (4.17), είναι απαραίτητη μία διανυσματική (συνεχής) αναπαράσταση των προτάσεων όπως αυτή που παρέχεται από την LSA ή τα word embeddings.

- Τελικές Ετικέτες

  Προκειμένου να πάρουμε τις τελικές ετικέτες που χρειάζεται ο αλγόριθμος μάθησης, εφαρμόζουμε αρχικά ένα κατώφλι ομοιότητας (*similarity threshold*) το οποίο μηδενίζει όσα στοιχεία των διανυσμάτων δεν το ξεπερνούν. Με αυτόν τον τρόπο, μπορούμε να απορρίψουμε τις θορυβώδεις προτάσεις που δεν κρύβουν καμία από τις ετικέτες μας.

Στη συνέχεια ακολουθεί μια κανονικοποίηση των διανυσμάτων προκειμένου να γίνουν στοχαστικά, να εκφράζουν δηλαδή μία κατανομή πιθανότητας. Η κανονικοποίηση αυτή γίνεται διαιρώντας με το άθροισμα των στοιχείων κάθε διανύσματος. Αν η μάθηση γίνει με *πιθανοτικές ετικέτες*, τότε τα διανύσματα που προκύπτουν είναι αυτά που δίνονται ως είσοδοι στον αλγόριθμο μάθησης. Να σημειώσουμε εδώ ότι σε αυτήν την περίπτωση έχει ακόμα μεγαλύτερη σημασία το κατώφλι ομοιότητας καθώς καθιστά πιο εύκολη την αποσαφήνιση της πραγματικής ετικέτας που κρύβεται μέσα στην πρόταση. Από την άλλη, αν η μάθηση γίνει με *ντετερμινιστικές ετικέτες* επιλέγουμε την κρυμμένη ετικέτα ως αυτή με την μέγιστη τιμή πιθανότητας.

6.5 Ευθυγράμμιση σεναρίου - υποτίτλων

Η ευθυγράμμιση των 2 κειμένων (για την ακρίβεια των υποτίτλων με τα τμήματα MONOLOGUE του σεναρίου) γίνεται με έναν αλγόριθμο Dynamic Time Warping όπως αυτός περιγράφεται στο [19]. Συγκεκριμένα, αφού μετατραπούν όλα τα κεφαλαία γράμματα σε μικρά και αφαιρεθούν τα σημεία στίξης προκύπτουν 2 ακολουθίες λέξεων. Στόχος είναι να βρεθεί ένα μονοπάτι που να αντιστοιχίζει τις λέξεις της μίας ακολουθίας με τις λέξεις της άλλης (όχι '1-1' φυσικά) ξεκινώντας από τις πρώτες 2 και καταλήγοντας στις 2 τελευταίες και κάνοντας όσο πιο λίγα βήματα γίνεται όταν συναντά λέξεις με μεγάλη διαφορά (μεγάλο κόστος βήματος). Ακόμα, το μονοπάτι πρέπει στο τέλος να δίνει σαν έξοδο 2 αύξουσες ακολουθίες λέξεων. Τα διαδοχικά στοιχεία αυτών θα πρέπει να είναι είτε ίδια, είτε διαδοχικά και στις αρχικές ακολουθίες προκειμένου να μην αγνοείται κάποια λέξη. Το βέλτιστο μονοπάτι βρίσκεται μέσω δυναμικού προγραμματισμού. Τελικά, ευθυγραμμίζοντας τους μονολόγους του σεναρίου με τους υπότιτλους αποδίδονται στους πρώτους τα ζητούμενα χρονικά όρια. Τα ίδια όρια τίθενται στα αντίστοιχα τμήματα SPEAKER. Τα τμήματα DESCRIPTION λαμβάνουν τα χρονικά όρια που ορίζονται από το πέρας του αμέσως προηγούμενου διαλόγου και την αρχή του αμέσως επόμενου. Τέλος, τα τμήματα SCENE λαμβάνουν τα χρονικά όρια που ορίζονται από την αρχή του αμέσως επόμενου διαλόγου και το πέρας του τελευταίου πριν την εμφάνιση ενός επόμενου τμήματος SCENE. Περισσότερο λεπτομερή περιγραφή των βημάτων και της υλοποίησης του αλγορίθμου μπορεί να διαβάσει κανείς στο [57].

Κεφάλαιο 7

Αξιολόγηση των αλγορίθων μάθησης

7.1 Εισαγωγή

Στο παρακάτω κεφάλαιο παρουσιάζονται τα αποτελέσματα των αλγορίθμων μάθησης που περιγράφηκαν στο κεφάλαιο 4. Συγκεκριμένα, η αξιολόγηση γίνεται στις έξι από τις οκτώ ταινίες της βάσης COGNIMUSE όπως αυτές αναφέρονται στο κεφάλαιο 4.

Το **σύνολο ετικετών** για το πρόβλημα της αναγνώρισης χαρακτήρων καθορίζεται για κάθε ταινία ξεχωριστά με τη βοήθεια της λίστας των χαρακτήρων που συλλέγουμε από τη διαδικτυακή βάση TMDB (βλέπε κεφάλαιο 6). Ο **εντοπισμός των γλωσσικών αντικειμένων** w_j και η **απόδοση ετικετών** στα bags γίνεται μέσω της απλής μεθόδου των κανονικών εκφράσεων που περιγράφηκε στο κεφάλαιο 6. Το **τελικό σύνολο ετικετών** που δίνεται ως είσοδος στο σύστημα μάθησης προκύπτει ως υποσύνολο του αρχικού με βάση τις ετικέτες που βρέθηκαν μέσα στο κείμενο και μόνο αυτές. **Επισημείωση** για τις πραγματικές ετικέτες των προσώπων κάθε χρονική στιγμή δεν υπάρχει. Για αυτό το λόγο στα πλαίσια της παρούσας διπλωματικής και προκειμένου να αξιολογήσουμε τις μεθόδους μας, επισημειώθηκαν οι ταινίες LOR και GLA. Η επισημείωση, όμως, έγινε καθαρά για τα tracks που παίρνουμε από τη διαδικασία εντοπισμού προσώπου οπότε δεν μπορεί να θεωρηθεί πλήρης. Αυτό γιατί τα αυστηρά χρονικά και χωρικά όρια δεν ορίζονται χειροκίνητα, αλλά από τον αυτόματο εντοπισμό, ενώ επίσης κάποια αποσπάσματα σημαντικών προσώπων δεν εντοπίζονται καν. Βέβαια, παρ' όλο που η επισημείωση μπορεί να μην είναι επαρκής για άλλες εργασίες πάνω στη βάση αυτή, είναι αρκετή για την αξιολόγηση των αλγορίθμων μάθησης της παρούσας διπλωματικής καθώς μας παρέχει ακριβείς ετικέτες για κάθε δεδομένο $\boldsymbol{x_i}$ που συναντούν και οι ίδιοι οι αλγόριθμοι σε αυτές τις 2 ταινίες. Τα **διανύσματα χαρακτηριστικών** $\boldsymbol{x_i}$ για κάθε πρόσωπο προκύπτουν από τη διαδικασία εντοπισμού που περιγράφεται στο κεφάλαιο 5 και στη συνέχεια από την εφαρμογή των περιγραφητών SIFT και VGG σε κάθε ένα από αυτά. Οι **πυρήνες** που χρησιμοποιεί ο αλγόριθμος μάθησης και άρα μας ενδιαφέρουν τελικά είναι οι 3 που αναφέρθηκαν στο κεφάλαιο 5 και θα αναφερόμαστε σε αυτούς με τις συντομογραφίες τους: SIFT38, VGG2 ,VGG1. Η **ευθυγράμμιση κειμένου - βίντεο**, δηλαδή

ο υπολογισμός των χρονικών ορίων κάθε φράσης w_j, γίνεται ευθυγραμμίζοντας το σενάριο με τους διαλόγους με τη μέθοδο Dynamic Time Warping που αναφέρουμε στο κεφάλαιο 6.

Το **σύνολο ετικετών** για το πρόβλημα της αναγνώρισης ανθρώπινων δράσεων καθορίζεται από τις 44 κατηγορίες που αναφέραμε στο κεφάλαιο 5. Τα πειράματα γίνονται ξεχωριστά για κάθε ταινία και επομένως μας ενδιαφέρουν οι κλάσεις που εμφανίζονται συχνά σε κάθε ταινία και όχι σε όλη τη βάση. Το **τελικό σύνολο ετικετών** για κάθε πείραμα προκύπτει ως υποσύνολο του αρχικού επιλέγοντας κάθε φορά διαφορετικό αριθμό κλάσεων. Συγκεκριμένα, εκτελούμε τα πειράματα για τις 2, 4, 6, 8 και 10 πολυπληθέστερες κλάσεις τις εκάστοτε ταινίας. Για πλήθος κλάσεων μεγαλύτερο από 10 παρατηρήθηκε ότι τα αποτελέσματα είχαν μεγάλη τυχαιότητα και μικρές τιμές, οπότε δεν παρουσιάζουν ενδιαφέρον. Ο **εντοπισμός των γλωσσικών αντικειμένων** w_j και η **απόδοση ετικετών** στα bags γίνεται μέσω της μεθόδου της σημασιολογικής ομοιότητας που περιγράφηκε στο κεφάλαιο 6. Εδώ υπάρχει πλήρης χρονική **επισημείωση** για τις πραγματικές ετικέτες κάθε χρονική στιγμή. Δεν υπάρχει όμως χωρική επισημείωση για την ακριβή θέση του ανθρώπου που εκτελεί τη δράση κάθε φορά. Ο εντοπισμός των ανθρώπινων δράσεων όπως αναφέραμε στο κεφάλαιο 5 επίσης γίνεται μόνο χρονικά αξιοποιώντας ακριβώς αυτά τα όρια που μας δίνει η επισημείωση. Τα **διανύσματα χαρακτηριστικών** $\boldsymbol{x_i}$ για κάθε δράση προκύπτουν από την εφαρμογή του περιγραφητή C3D στα αποσπάσματα που παίρνουμε μετά τον εντοπισμό, ενώ οι **πυρήνες** που εφαρμόζονται σε αυτά είναι ο γραμμικός (linear) και ο χ^2. Θα αναφερόμαστε σε αυτούς με τις συντομογραφίες τους: linC3D, χ^2C3D. Η **ευθυγράμμιση κειμένου - βίντεο** γίνεται με την ίδια μέθοδο που προαναφέραμε.

Σημειώνουμε ότι και στα 2 προβλήματα η μάθηση γίνεται μόνο για τα δεδομένα προσκηνίου (foreground), δηλαδή σε αυτά που η πραγματική τους ετικέτα ανήκει στο σύνολο ετικετών. Για την αναγνώριση προσώπου, δεδομένα παρασκηνίου (background) προκύπτουν από λάθη στον εντοπισμό, ή από τον εντοπισμό χαρακτήρων που δεν αναφέρονται στο κείμενο. Για την αναγνώριση δράσης, δεδομένα παρασκηνίου προκύπτουν όταν περιορίζουμε το σύνολο ετικετών μας σε λιγότερες κλάσεις από αυτές που εμφανίζονται συνολικά (όπως αναφέρθηκε στην παραπάνω παράγραφο). Τα δεδομένα αυτά δεν λαμβάνονται υπ΄ όψη, καθώς δεν έχει ληφθεί ειδική μέριμνα για την αναγνώρισή τους στην παρούσα διπλωματική (όπως γίνεται π.χ στο [4]). Το πρόβλημα επομένως αντιμετωπίζεται σαν πρόβλημα αναγνώρισης κλειστού συνόλου (closed set recognition). Σημειώνουμε, όμως, ότι εκτελώντας τις διαδικασίες βελτιστοποίησης σε όλα τα δεδομένα, πολλές φορές τα αποτελέσματα δεν παρουσίαζαν μεγάλη διαφορά με αυτά που παίρνουμε εκτελώντας τη βελτιστοποίηση μόνο στα δεδομένα προσκηνίου. Φυσικά, οι περιορισμοί και η αντικειμενική συνάρτηση του προβλήματος αλλάζουν, οπότε πολλές φορές τα δεδομένα παρασκηνίου μπορεί να επηρεάσουν πολύ το τελικό αποτέλεσμα. Έτσι, προκειμένου να μην αλλοιώνονται τα συμπεράσματα μας από αυτόν τον παράγοντα, τον παραλείψαμε εντελώς.

Η σύγκριση των αλγορίθμων γίνεται με βάση τη γνωστή και απλή μετρική accuracy ή αλλιώς ποσοστό επιτυχίας, η οποία εκφράζει το ποσοστό των σωστά ταξινομημένων δεδομένων προς το συνολικό πλήθος των δεδομένων. Δηλαδή, αν m το πλήθος των δεδομένων $\boldsymbol{x_i}$, τότε $accuracy = \frac{|\{\boldsymbol{x_i}|\text{true label}(\boldsymbol{x_i})=\text{predicted label}(\boldsymbol{x_i})\}|}{m}$, όπου στο σύστημα μας όπως αναφέραμε στο

κεφάλαιο 4, predicted label($\boldsymbol{x_i}$) = arg $\max_p$ z_{ip} και προφανώς true label($\boldsymbol{x_i}$) = y_i.

Το βασικό σύστημα (baseline) με το οποίο θα συγκριθούν οι μέθοδοι που προτείνουμε είναι αυτό που παρουσιάζεται στο [4]. Το σύστημα μας αποτελεί επέκταση αυτού λαμβάνοντας υπόψιν κάποια χαρακτηριστικά του προβλήματος που δεν είχαν μοντελοποιηθεί επαρκώς. Η σύγκριση δεν μπορεί να γίνει πάνω στο ίδιο σύνολο δεδομένων που οι συγγραφείς είχαν χρησιμοποιήσει, καθώς τα δεδομένα που έχουν διανείμει δημόσια είναι στην τελική τους μορφή και έτσι δεν μπορούμε να εφαρμόσουμε όλες τις διαφοροποιήσεις που προτείνουμε.

7.2 Αναγνώριση Προσώπου

Για το πρόβλημα της αναγνώρισης προσώπου, όπως εξηγήσαμε στο κεφάλαιο 6 δεν υπάρχει ζήτημα πιθανοτικών ετικετών, καθώς κάθε γλωσσικό αντικείμενο w_j που απομονώνουμε έχει μοναδική αντιστοίχηση με μία ετικέτα. Κάθε φορά που εντοπίζουμε ένα όνομα στο κείμενο που ταυτίζεται με κάποιο από το σύνολο ετικετών μας, είμαστε σίγουροι για την ετικέτα που του αποδίδεται. Επίσης, δεν υπάρχει πρότερη γνώση για τα πρόσωπα που θέλουμε να αναγνωρίσουμε. Επομένως, οι επεκτάσεις που αξιολογούμε εδώ αφορούν:

- Τη μοντελοποίηση των bags ως ασαφή σύνολα. Καθορίζεται από τη συνάρτηση συμμετοχής g και την παράμετρο επέκτασης των χρονικών ορίων του bag, ϵ,
- Τη μοντελοποίηση των επαναλήψεων των bags. Καθορίζεται από την αλλαγή των βαρών κάθε μεταβλητής χαλάρωσης στην αντικειμενική συνάρτηση. Για κάθε επαναλαμβανόμενο περιορισμό, το τελικό βάρος προκύπτει ως άθροισμα των αρχικών βαρών των μεταβλητών χαλάρωσης των επιμέρους επαναλήψεων του (εδώ τα αρχικά βάρη είναι ίσα με 1),
- Την ενσωμάτωση βαθιών νευρωνικών δικτύων στο σύστημα αναγνώρισης χρησιμοποιώντας έναν από τους πυρήνες VGG2 ,VGG1.

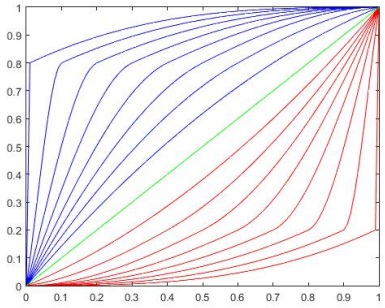
Αρχικά παρουσιάζουμε δύο πρωταρχικά πειράματα που εκτελέσαμε προκειμένου να επιλέξουμε κάποιες υπερπαραμέτρους και να εξετάσουμε σε πρώτο βαθμό ποιοτικά την αποτελεσματικότητα των μεθόδων. Στη συνέχεια τις συγκρίνουμε και ποσοτικά για να αποφανθούμε για την υπεροχή της κάθε μίας.

Ασαφή Σύνολα Πολλαπλών Παραδειγμάτων

Η μοντελοποίηση με ασαφή σύνολα προκύπτει από εφαρμογή του περιορισμού (4.14) ο οποίος αντικαθιστά τον (4.11). Η απόδοση της μεθόδου εξαρτάται σε μεγάλο βαθμό από την επιλογή της συνάρτησης συμμετοχής και από την υπερπαράμετρο ϵ. Οι συναρτήσεις συμμετοχής g που συγκρίναμε ανήκουν σε δύο διαφορετικές οικογένειες οι οποίες υπακούν στις ιδιότητες που περιγράψαμε στην ενότητα 4.2, δηλαδή g αύξουσα, $g(0) = 0$, $g(1) = 1$.

- Τμηματική Κυβική Παρεμβολή Hermite (Piecewise Cubic Hermite Interpolating Polynomial-pchip). Επιλέγουμε ένα ενδιάμεσο σημείο στο διάστημα $[0, 1]$ και του αποδίδουμε μία

τιμή επίσης στο $[0, 1]$. Παρεμβάλλοντας μία συνάρτηση από τα 3 δεδομένα σημεία με τη μέθοδο αυτή παίρνουμε τη συνάρτηση g. Τα ζεύγη ενδιάμεσων σημείων-τιμών $(x_0, g(x_0))$ είναι αυτά που παράγουν την οικογένεια. Επιλέγουμε μία πολύ εύχρηστη ειδική κατηγορία της παρεμβολής αυτής, η οποία εξηγείται στο [41] και χαρακτηρίζεται από τη διατήρηση του σχήματος που έχουν τα σημεία που παρεμβάλλουμε. Δηλαδή, διατηρείται η μονοτονία αλλά και άλλες ιδιότητες που μπορεί να έχουν τα σημεία αυτά. Επιλέγουμε ένα υποσύνολο της οικογένειας αυτής, το οποίο φαίνεται στο σχήμα 7.1, θέτοντας το ενδιάμεσο σημείο να ανήκει στο διάστημα $(0, 0.8]$ και την τιμή του να είναι ίση με 0.8 ή θέτοντας το ενδιάμεσο σημείο να ανήκει στο διάστημα $[0.2, 1)$ και την τιμή του να είναι ίση με 0.2. Η οικογένεια περιέχει συναρτήσεις με κοίλη γραφική παράσταση



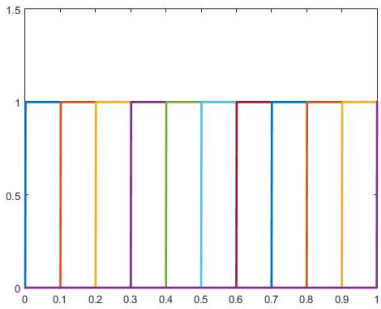
Σχήμα 7.1: Η οικογένεια συναρτήσεων που παράγεται από την pchip

(αυτές με το μπλε χρώμα - $g(x_0) = 0.8$) οι οποίες αναγκάζουν δείγματα με έστω και μικρή επικάλυψη να συμμετέχουν με μεγάλη τιμή στο σύνολο, συναρτήσεις με κυρτή γραφική παράσταση (αυτές με το κόκκινο χρώμα - $g(x_0) = 0.2$) οι οποίες κάνουν μόνο τα δείγματα με μεγάλη επικάλυψη να συμμετέχουν με μεγάλη τιμή στο σύνολο, καθώς και την γραμμική συνάρτηση όπου κάνει τα δείγματα να συμμετέχουν στο σύνολο με ίδια τιμή με το ποσοστό επικάλυψής τους.

- Βηματικές Συναρτήσεις. Ορίζονται ως:

$$g(x) = \begin{cases} 1 & x > t \\ 0 & \text{αλλιώς} \end{cases}.$$

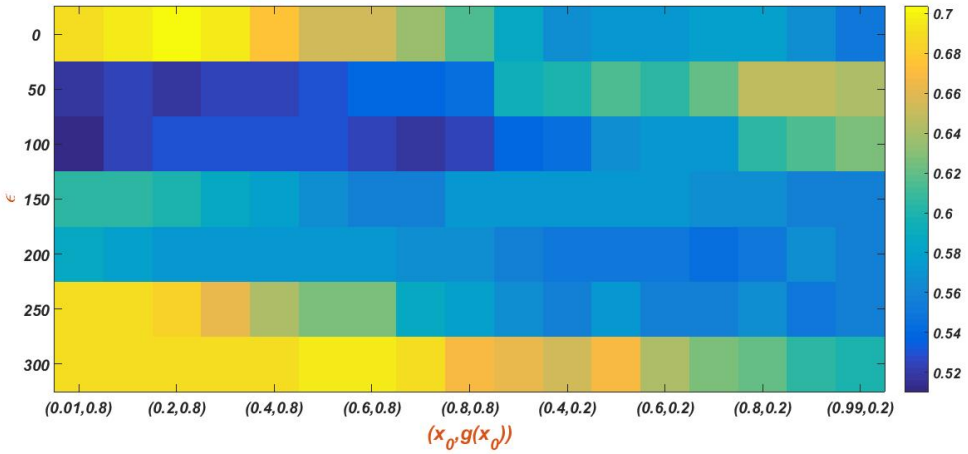
Να σημειώσουμε ότι αυτές οι συναρτήσεις ουσιαστικά ορίζουν απλά σύνολα αλλά γενικεύουν σε ένα βαθμό το baseline καθώς απαιτούν η επικάλυψη να είναι μεγαλύτερη από το κατώφλι t και όχι απλά μεγαλύτερη του μηδενός.



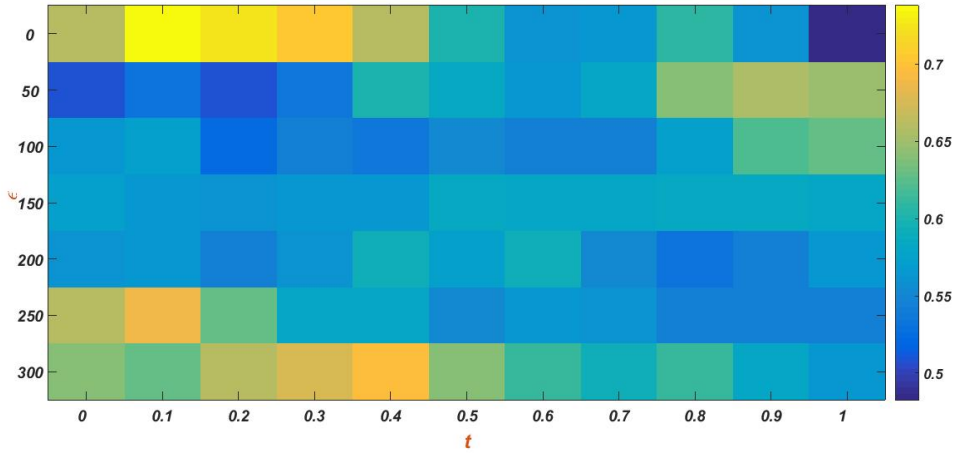
Σχήμα 7.2: Η οικογένεια των βηματικών συναρτήσεων

Για να επιλέξουμε τις παραμέτρους ϵ, $(x_0, g(x_0))$ για την pchip και t για την βηματική, δημιουργήσαμε μερικά πρωταρχικά πειράματα για τις 2 επισημειωμένες ταινίες. Συγκεκριμένα, για κάθε έναν από τους 3 πυρήνες εκτελέσαμε τη διαδικασία βελτιστοποίησης για διάφορες τιμές των παραμέτρων αυτών.[1]. Παρακάτω παρουσιάζουμε τα αποτελέσματα μόνο για τον πυρήνα VGG1 καθώς ακολουθείται πάνω κάτω το ίδιο μοτίβο και για τους άλλους πυρήνες:

[1]Τα πειράματα έγιναν για ίσα βάρη στις μεταβλητές χαλάρωσης και για τις τιμές των υπερπαραμέτρων $\alpha = 2.5, \kappa = 20, \lambda = 10^{-4}$ που θα χρησιμοποιηθούν και στα τελικά πειράματα

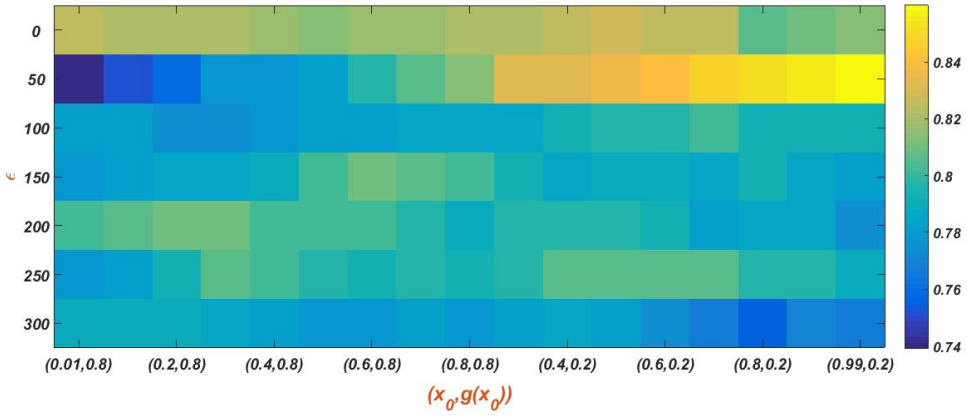


(α′)

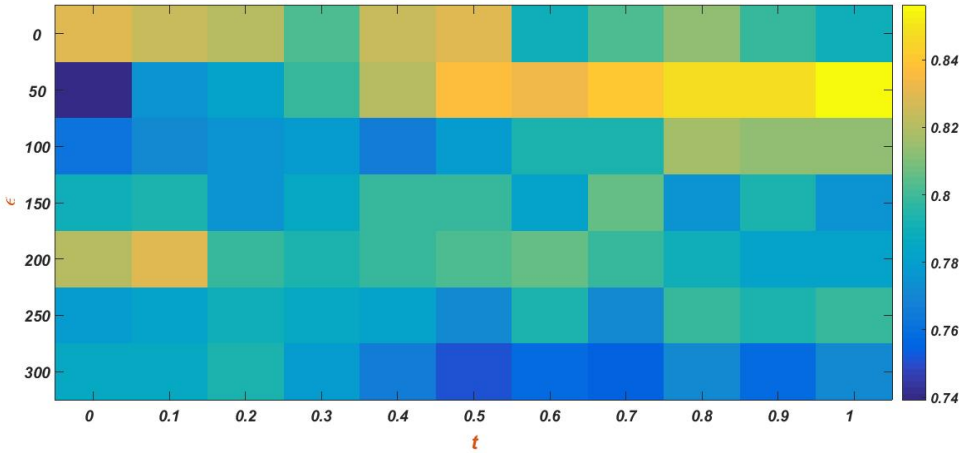


(β′)

Σχήμα 7.3: Στο σχήμα (α′) παρουσιάζεται το ποσοστό επιτυχίας για την ταινία GLA για την οικογένεια συναρτήσεων συμμετοχής που προκύπτει από την παρεμβολή pchip ως συνάρτηση του ζεύγους $(x_0, g(x_0))$ και της χρονικής επέκτασης ϵ μετρημένης σε καρέ (frames). Στο σχήμα (β′) παρουσιάζεται το αντίστοιχο ποσοστό επιτυχίας για την οικογένεια συναρτήσεων συμμετοχής που προκύπτει από τις βηματικές συναρτήσεις ως συνάρτηση του κατωφλίου t και της χρονικής επέκτασης ϵ.



(α′)



(β′)

Σχήμα 7.4: Όμοια με το σχήμα 7.3 για την ταινία LOR

Παρατηρούμε ότι ως προς τη χρονική επέκταση τα αποτελέσματα σε γενικές γραμμές επιδεινώνονται όσο μεγαλώνει το ϵ. Αυτό μπορεί να εξηγηθεί αν σκεφτεί κανείς ότι στο πρόβλημα της αναγνώρισης προσώπου οι ετικέτες που παρέχονται από το κείμενο είναι πολύ πυκνές. Επίσης, η μεγάλη πλειοψηφία τους προέρχεται από τμήματα του κειμένου που αντιστοιχούν σε μονολόγους (βλέπε 6). Άρα λόγω του ότι όταν κάποιος μιλάει σε μία ταινία, συνήθως εμφανίζεται και το πρόσωπο του, οι ετικέτες αυτές είναι συνήθως σωστές. Επομένως, δεν υπάρχει λόγος να αυξήσουμε τα χρονικά όρια γιατί έτσι μάλλον θα εισάγουμε περισσότερο θόρυβο παρά θα αντιμετωπίσουμε τον ήδη υπάρχον. Η μόνη περίπτωση που βλέπουμε αύξηση των αποτελεσμάτων είναι για $\epsilon = 50$ στην ταινία LOR και όταν ταυτόχρονα περιορίζουμε τα δείγματα που συμμετέχουν σε κάθε σύνολο (στην δεξιά πλευρά των διαγραμμάτων). Στην ταινία GLA στα ίδια σημεία των διαγραμμάτων παρατηρούμε ικανοποιητικά υψηλές τιμές των ποσοστών επιτυχίας, αν και όχι καλύτερες από αυτές που παρουσιάζονται στην αριστερή πλευρά του διαγράμματος για $\epsilon = 0$. Είναι εύκολο να ερμηνεύσουμε αυτό το φαινόμενο αν σκεφτούμε ότι επεκτείνοντας τα όρια κάθε γλωσσικού αντικειμένου μεγαλώνουμε τις επικαλύψεις του με τα

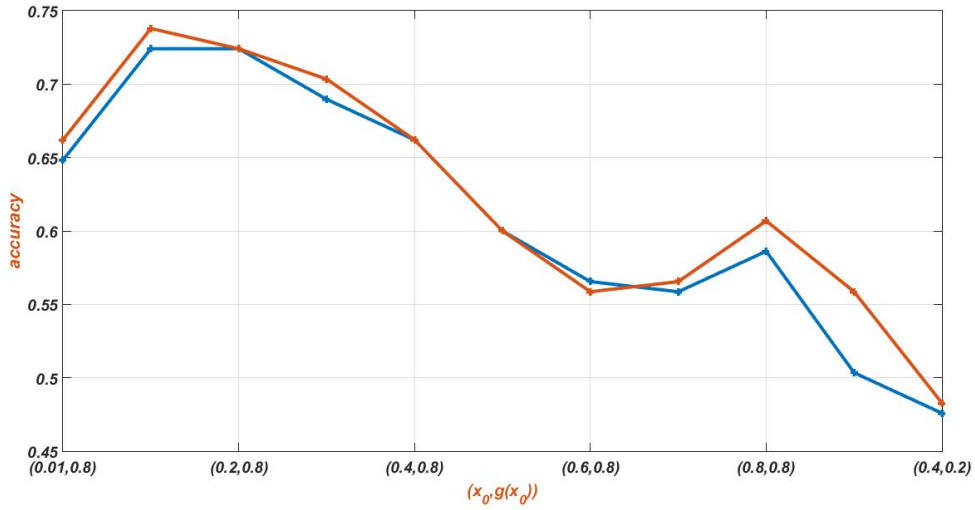
οπτικά. Στη δεξιά πλευρά των διαγραμμάτων η συμμετοχή ενός δείγματος σε ένα σύνολο έχει αξιοσημείωτη τιμή μόνο όταν η επικάλυψη του είναι αρκετά μεγάλη. Αντίθετα, στην αριστερή πλευρά η συμμετοχή είναι αξιοσημείωτη ακόμα και για μικρές επικαλύψεις. Άρα, είναι περίπου ισοδύναμο να επιλέξει κανείς συναρτήσεις συμμετοχής από την αριστερή πλευρά του διαγράμματος με επέκταση ϵ και συναρτήσεις συμμετοχής από την δεξιά πλευρά με επέκταση $\acute{\epsilon} > \epsilon$. Η ποσότητα κατά την οποία πρέπει να αυξηθεί η επέκταση δεν είναι προφανής και εξαρτάται από την πυκνότητα των ετικετών και τη χρονική διάρκεια των οπτικών αντικειμένων, αλλά διαισθητικά μπορούμε να αντιληφθούμε ότι υπάρχει κάποια σχέση μεταξύ των 2 παραμέτρων.

Η επιλογή των $(x_0, g(x_0))$, t θα γίνει για **μηδενική επέκταση**, εφόσον κρίνουμε ότι για το πρόβλημα της αναγνώρισης προσώπου είναι καταλληλότερη. Όπως φαίνεται από το διάγραμμα 7.4α′ και κυρίως από το 7.3α′ που αφορά την ταινία GLA, παρατηρούμε υψηλότερα ποσοστά επιτυχίας στην αριστερή πλευρά τους, δηλαδή για την υποοικογένεια των κοίλων συναρτήσεων που παράγει η pchip. Αυτό είναι λογικό γιατί στο συγκεκριμένο πρόβλημα με την μεγάλη πυκνότητα ετικετών και το μηδενικό θόρυβο στην αντιστοίχηση γλωσσικό αντικείμενο-ετικέτα, είναι αρκετά πιθανό η ετικέτα του γλωσσικού αντικειμένου να πρέπει να αποδοθεί και στο οπτικό, ακόμα και αν δεν έχουν πολύ μεγάλη επικάλυψη. Άλλωστε, οι διαφορές στα ποσοστά επιτυχίας είναι μικρές μεταξύ διαφορετικών κοίλων συναρτήσεων, πράγμα που σημαίνει ότι το ποσό συμμετοχής ενός δείγματος σε ένα σύνολο δεν επηρεάζει έντονα, αρκεί να μην είναι κοντά στο 0. Παρόμοια, συμπεράσματα μπορούμε να βγάλουμε και για τις βηματικές συναρτήσεις. Τελικά, επιλέγουμε $\boldsymbol{(x_0, g(x_0)) = (0.2, 0.8)}$ και $\boldsymbol{t = 0.1}$ που πετυχαίνουν τα υψηλότερα ποσοστά στο GLA και ικανοποιητικά υψηλά στο LOR.

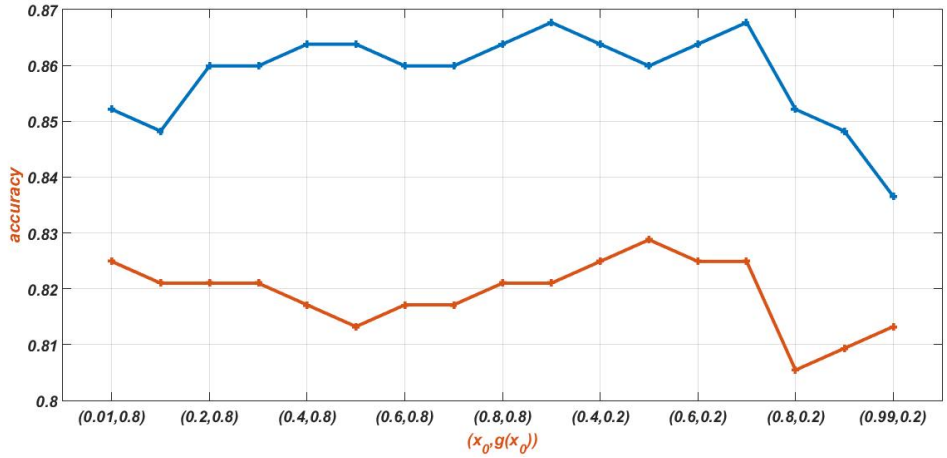
Τελικά βάρη των Μεταβλητών Χαλάρωσης - Επαναλήψεις των bags

Εδώ εκτελούμε ένα ακόμα πρωταρχικό πείραμα για να ελέγξουμε αν είναι βάσιμος ο ισχυρισμός μας ότι οι μεταβλητές χαλάρωσης των περιορισμών που επαναλαμβάνονται πρέπει να παίρνουν μεγαλύτερο βάρος στην αντικειμενική συνάρτηση. Ελαχιστοποιώντας την αντικειμενική συνάρτηση 4.13 με τιμές βαρών όπως αυτές που περιγράφονται στο κεφάλαιο 4 καταφέρνουμε να βελτιώσουμε την απόδοση του συστήματος αναγνώρισης[2]. Στα σχήματα 7.5 7.6 απεικονίζεται το accuracy για την οικογένεια συναρτήσεων που παράγονται από την pchip(παρόμοια είναι τα αποτελέσματα για βηματικές συναρτήσεις), για τον πυρήνα VGG1 και για τις τιμές των παραμέτρων που επιλέξαμε προηγουμένως, δηλαδή $\epsilon = 0$, $(x_0, g(x_0)) = (0.2, 0.8)$. Στις 2 μεθόδους που συγκρίνονται επιλέγονται μόνο οι ξεχωριστοί περιορισμοί και στη συνέχεια τους αποδίδονται είτε ίσα βάρη, είτε βάρη ίσα με το άθροισμα των αρχικών βαρών των επιμέρους επαναλήψεων.

[2]εδώ τα αρχικά βάρη είναι 1 άρα για κάθε επαναλαμβανόμενο περιορισμό το τελικό βάρος θα είναι ίσο με το πλήθος των επαναλήψεων του



Σχήμα 7.5: GLA: Ποσοστά επιτυχίας συναρτήσει του $(x_0, g(x_0)) = (0.2, 0.8)$ για βάρη ίσα με τα αρχικά (εδώ είναι 1 εφόσον δεν υπάρχουν πιθανοτικές ετικέτες) -κόκκινο χρώμα- και για βάρη ίσα με το άθροισμα των βαρών των επιμέρους επαναλήψεων κάθε περιορισμού -μπλε χρώμα.



Σχήμα 7.6: Όμοια με το σχήμα 7.5 για την ταινία LOR.

Στο σχήμα 7.6 φαίνεται καθαρά η σημασία της επιβολής διαφορετικών βαρών σε κάθε μεταβλητή χαλάρωσης. Τα αποτελέσματα επιβεβαιώνουν την αρχική μας διαίσθηση ότι όταν ένα γλωσσικό αντικείμενο εμφανίζεται πολλές φορές στο ίδιο σημείο του κειμένου τότε αυξάνεται η πιθανότητα να βρεθεί το αντίστοιχο οπτικό αντικείμενο στο βίντεο. Στο σχήμα 7.5 φαίνεται να μην υπάρχει ουσιαστική διαφορά είτε επιβάλλουμε βάρη είτε όχι. Συμπερασματικά, δεδομένου του ότι στην μία ταινία τα αποτελέσματα βελτιώνονται αισθητά και στην άλλη δεν επιδεινώνονται, θεωρούμε ότι η χρήση των βαρών είναι μάλλον ωφέλιμη.

Συγκρίσεις Μεθόδων

Παρακάτω συγκρίνουμε τις διαφορετικές μεθόδους μεταξύ τους. Στο **baseline** του [4] αντιμετωπίζεται το πρόβλημα ως μάθηση πολλαπλών παραδειγμάτων όπου τα σύνολα των παραδειγμάτων (bags) είναι απλά σύνολα, δηλαδή η συμμετοχή ενός δείγματος σε ένα σύνολο είναι δυαδική (είτε ανήκει είτε δεν ανήκει). Για να ανήκει ένα δεδομένο σε ένα bag αρκεί η επικάλυψη να είναι διάφορη του μηδενός. Ο πυρήνας που χρησιμοποιείται είναι ο SIFT38. Η μέθοδος αυτή υλοποιείται εκτελώντας τη βελτιστοποίηση που ορίζεται από την αντικειμενική συνάρτηση της σχέσης (4.10) και τους περιορισμούς (4.6), (4.7),(4.9),(4.11). Οι επεκτάσεις είναι η αλλαγή του τρόπου συμμετοχής των δεδομένων στα bags, είτε αυτός ο τρόπος κατασκευάζει απλά σύνολα (step), είτε ασαφή (pchip), η προσθήκη βαρών στις μεταβλητές χαλάρωσης (weights), και η αντικατάσταση των hand-crafted χαρακτηριστικών (SIFT38) με χαρακτηριστικά βαθιάς μάθησης (VGG2,VGG1).Στον πίνακα 7.1 παρουσιάζονται τα αποτελέσματα για τις 2 επισημειωμένες ταινίες. Όλα τα πειράματα εκτελέστηκαν για τις τιμές των υπερπαραμέτρων $\alpha = 2.5, \kappa = 20, \lambda = 10^{-4}$.

Τα δεδομένα μας σε αυτό το πρόβλημα έχουν τα εξής ποσοτικά χαρακτηριστικά: Για την ταινία **Lord of the Rings: The Return of the King (LOR)** το αρχικό σύνολο ετικετών αποτελείται από 28 διαφορετικές κατηγορίες ανθρώπινων προσώπων, ενώ το τελικό από 14. Τα γλωσσικά αντικείμενα που εντοπίζουμε είναι 427. Τα συνολικά δεδομένα είναι 295 tracks που περιέχουν 19,283 εντοπισμένα πρόσωπα, ενώ τα δεδομένα προσκηνίου είναι 257 tracks. Για την ταινία **Gladiator (GLA)** το αρχικό σύνολο ετικετών αποτελείται από 38 διαφορετικές κατηγορίες, ενώ το τελικό από 12 και τα γλωσσικά αντικείμενα είναι 194. Τα συνολικά δεδομένα είναι 185 tracks που περιέχουν 13,593 εντοπισμένα πρόσωπα, ενώ τα δεδομένα προσκηνίου είναι 145 tracks.

Όπως προαναφέραμε η προσθήκη των βαρών στις μεταβλητές χαλάρωσης δεν φαίνεται να βελτιώνει πάντα τα ποσοστά. Βλέπουμε, όμως, ότι για τους πυρήνες του VGG στην ταινία LOR, η βελτίωση είναι αισθητή. Αυτό δείχνει τη σημασία που έχει η ομαδοποίηση στον αλγόριθμο. Συγκεκριμένα, μπορούμε να σκεφτούμε ένα παράδειγμα όπου σε μία μεταβλητή χαλάρωσης δίνεται μεγάλο βάρος λόγω πολλών επαναλήψεων του περιορισμού, αλλά στην πραγματικότητα ο περιορισμός αυτός είναι θορυβώδης. Τότε, θα πρέπει να υποδειχτεί από τον όρο της ομαδοποίησης ότι αυτός ο περιορισμός είναι θορυβώδης, δηλαδή θα πρέπει η θέση στο χώρο των διανυσμάτων που συμμετέχουν σε αυτόν να μην ταιριάζει με την τοποθέτηση τους στην κλάση που υποδεικνύει η ετικέτα. Για να συμβεί αυτό όμως θα πρέπει τα διανύσματα να περιγράφουν πολύ καλά τα δείγματα και να κάνουν τις κλάσεις γραμμικά διαχωρίσιμες. Επομένως, αν δεν έχουμε αρκετά ποιοτικά διανύσματα, τότε η προσθήκη των βαρών μπορεί να οδηγήσει σε αντίθετα αποτελέσματα λόγω της ενίσχυσης θορυβωδών περιορισμών. Αυτό το φαινόμενο λογικά εμφανίζεται στην ταινία GLA και τα αποτελέσματα δεν βελτιώνονται.

Αναφορικά με τις συναρτήσεις συμμετοχής, παρατηρούμε ότι στο LOR η pchip έχει καλύτερα αποτελέσματα από τη step, ειδικά όταν συνδυάζεται με βάρη, ενώ στο GLA συμβαίνει το ακριβώς αντίθετο. Η εξήγηση που μπορούμε να δώσουμε είναι παρόμοια με αυτό που αναφέραμε στην προηγούμενη παράγραφο. Αν θεωρήσουμε δεδομένο ότι στην ταινία GLA,

	LOR	GLA
SIFT (Baseline)	72.37%	48.97%
SIFT + weights	69.65%	47.59%
SIFT+step	73.93%	51.72%
SIFT+pchip	71.21%	49.66%
SIFT+step+weights	72.37%	51.72%
SIFT+pchip+weights	73.54%	48.97%
VGG2	83.66%	64.83%
VGG2+weights	85.60%	61.38%
VGG2+step	82.88%	68.97%
VGG2+pchip	82.10%	64.14%
VGG2+step+weights	85.21%	64.83%
VGG2+pchip+weights	**86.38%**	63.45%
VGG1	82.88%	66.21%
VGG1+weights	84.82%	64.83%
VGG1+step	82.49%	**73.79**
VGG1+pchip	82.10%	70.34%
VGG1+step+weights	85.60%	72.41%
VGG1+pchip+weights	85.99%	69.66%

Πίνακας 7.1: Συγκεντρωτικά αποτελέσματα συγκρίσεων για το πρόβλημα της αναγνώρισης προσώπου στις ταινίες LOR και GLA

οι περιορισμοί έχουν περισσότερο θόρυβο, τότε ίσως είναι καλύτερη η θεώρηση των απλών συνόλων καθώς απορρίπτει από τους περιορισμούς πιθανόν ακατάλληλα δεδομένα και μειώνει την πολυπλοκότητα της βελτιστοποίησης. Αντίθετα, όταν ο θόρυβος είναι μικρός, η συμμετοχή στα bags όλων των δειγμάτων ανάλογα με την επικάλυψη τους φαίνεται να βοηθάει το σύστημα. Για παράδειγμα, μπορεί ο όρος ομαδοποίησης να υποδεικνύει με μεγάλη σιγουριά για ένα δείγμα ακόμα και με μικρή συμμετοχή στο σύνολο, ότι πρέπει να πάρει την ετικέτα του συνόλου. Σε αυτήν την περίπτωση η χρήση της step θα απέρριπτε εντελώς το δείγμα από το σύνολο και ο περιορισμός δεν θα το αφορούσε.

Τέλος όσον αφορά τα διανύσματα χαρακτηριστικών, είναι έκδηλη η υπεροχή του VGG, πράγμα αναμενόμενο καθώς είναι πλέον ευρέως γνωστό ότι η βαθιά μάθηση ξεπερνά κατά πολύ τις παλιές τεχνικές σε πολλά προβλήματα της όρασης υπολογιστών. Όπως φαίνεται και από τον πίνακα 7.1 η φύση και η ποιότητα των διανύσματα χαρακτηριστικών επηρεάζουν σε μεγαλύτερο βαθμό τα αποτελέσματα από ότι οι άλλες επεκτάσεις. Ακόμα, συγκρίνοντας τους δύο πυρήνες VGG2,VGG1, παρατηρούμε ότι ο πρώτος συμπεριφέρεται καλύτερα στο GLA, ενώ ο δεύτερος στο LOR. Δεν μπορούμε να είμαστε σίγουροι για το ποιος πλεονεκτεί, αλλά από ότι φαίνεται δεν υπάρχει μεγάλη διαφορά μεταξύ τους. Άλλωστε, το VGG είναι εκπαιδευμένο και για frontal και για profile πρόσωπα και μπορεί να ταυτοποιεί πρόσωπα με

διαφορετικές γωνίες θέασης. Οπότε, η λογική υπόθεση είναι ότι δεν θα υπάρχει μεγάλη διαφορά μεταξύ των 2 πυρήνων.

Συμπερασματικά:

Για την αναγνώριση προσώπου, προτείνουμε ανεπιφύλακτα τη χρήση χαρακτηριστικών από βαθιά νευρωνικά δίκτυα. Ακόμα, προτείνουμε τη χρήση μίας από τις 2 συναρτήσεις συμμετοχής, προκειμένου σε πρώτη φάση να τοποθετηθεί διαφορετικό κατώφλι στην επικάλυψη γλωσσικών και οπτικών αντικειμένων. Σε δεύτερη φάση, η επιλογή μεταξύ pchip και step (ή και άλλων που δεν έχουν μελετηθεί εδώ), καθώς και η προσθήκη ή όχι των βαρών επαναλήψεων στις μεταβλητές χαλάρωσης, πρέπει να γίνει ανάλογα με μία εκτίμηση του πόσο θορυβώδεις είναι οι περιορισμοί. Αυτό, μπορεί να γίνει για παράδειγμα μέσα από την αξιολόγηση της ευθυγράμμισης του dynamic time warping. Για του λόγου το αληθές, η ταινία LOR έχει μεγαλύτερο alignment score από την ταινία GLA(0.922 και 0.878) και άρα καλύτερη ευθυγράμμιση μεταξύ βίντεο και κειμένου.

7.3 Αναγνώριση Δράσεων

Στο πρόβλημα της αναγνώρισης δράσεων μπορούν να εφαρμοστούν όλες οι επεκτάσεις που προτείναμε στο κεφάλαιο 4. Συγκεκριμένα, αξιολογούμε:

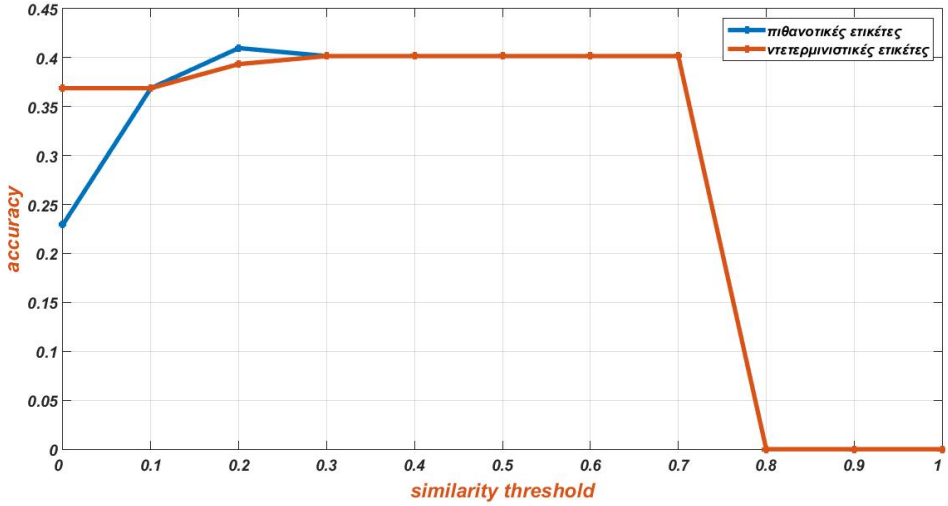
- Τη χρήση της σημασιολογικής ομοιότητας των γλωσσικών αντικειμένων για την εξαγωγή των ετικετών. Επηρεάζεται από το κατώφλι *similarity threshold* που απορρίπτει τιμές ομοιότητας κάτω από αυτό,
- Την μοντελοποίηση των πιθανοτικών ετικετών, συγκριτικά με τις αντίστοιχες ντετερμινιστικές. Καθορίζεται από την προσθήκη αρχικών βαρών στις μεταβλητές χαλάρωσης με βάση την πιθανότητα κάθε ετικέτας (εκφυλίζονται στο 1 στην περίπτωση των ντετερμινιστικών ετικετών),
- Τη μοντελοποίηση των bags ως ασαφή σύνολα. Καθορίζεται όπως προαναφέραμε,
- Τη μοντελοποίηση των επαναλήψεων των bags. Καθορίζεται όπως προαναφέραμε,
- Την αξιοποίηση της πρότερης γνώσης που προκύπτει από το σύστημα αναγνώρισης προσώπου,
- Την αξιοποίηση της πρότερης γνώσης που προκύπτει από τις προβλέψεις ενός προεκπαιδευμένου συστήματος αναγνώρισης δράσεων.

Δεδομένου του μεγάλου πλήθους συνδυασμών που παράγονται από τις παραμέτρους κάθε επέκτασης δεν εκτελέσαμε μία πλήρη αναζήτηση σε όλο το πλέγμα που παράγουν, αλλά τις βελτιστοποιήσαμε διαδοχικά. Δηλαδή, όπως και πριν εκτελέσαμε κάποια πρωταρχικά πειράματα για να επιλέξουμε κάποιες παραμέτρους και για να εξεταστούν ποιοτικά οι μέθοδοι και στη συνέχεια συγκρίνονται όλες οι επεκτάσεις και οι συνδυασμοί τους ποσοτικά. Τα πρωταρχικά πειράματα έγιναν στις ίδιες ταινίες με αυτές που χρησιμοποιήθηκαν στο σύστημα αναγνώρισης προσώπου, ενώ τα τελικά για όλες τις ταινίες της βάσης.

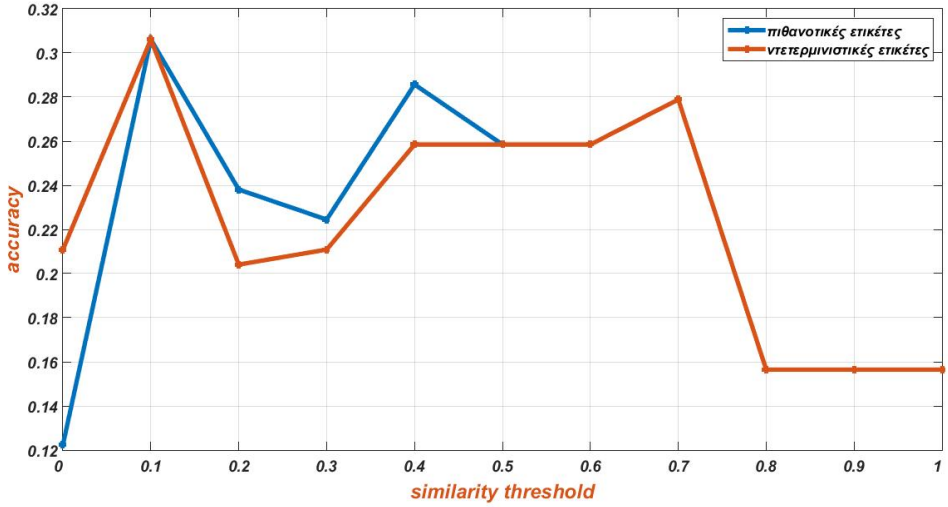
Χρήση Σημασιολογικής Ομοιότητας

Η χρήση της σημασιολογικής ομοιότητας για την εξαγωγή των ετικετών περιγράφεται στο κεφάλαιο 6. Είναι μία διαφορετική προσέγγιση από αυτές που έχουν προταθεί σε παρόμοια προβλήματα ([4, 32]). Αφού υπολογίσουμε την ομοιότητα κάθε γλωσσικού αντικειμένου w_j με τις γλωσσικές αναπαραστάσεις των ετικετών μας, απορρίπτουμε όσες τιμές ομοιότητες βρίσκονται κάτω από το κατώφλι *similarity threshold* (δηλαδή τις θέτουμε ίσες με το 0). Η επιλογή του είναι αρκετά σημαντική καθώς πρέπει από τη μία να απορρίπτει φράσεις που δεν υπονοούν κάποια ανθρώπινη δράση, άρα είναι θόρυβος για το σύστημα μας, αλλά από την άλλη να μην περιορίζεται μόνο στις απόλυτα όμοιες φράσεις, οι οποίες είναι σπάνιες σε αυτά τα κείμενα. Οι μέθοδοι που υλοποιούνται στα [4, 32] αμφότερες περιορίζονται σε φράσεις με μεγάλη σημασιολογική ομοιότητα με κάποια από τις ετικέτες. Καθώς δεν υπάρχει υλοποίηση των μεθόδων αυτών για να γίνει ευθεία σύγκριση, ποιοτικά μπορούμε να δούμε τη σημασία της δικιάς μας παρατηρώντας τα ποσοστά επιτυχίας σαν συνάρτηση του κατωφλίου ομοιότητας. Στα σχήματα 7.7,7.8 παρουσιάζουμε ενδεικτικά πειράματα για 4 και 6 κλάσεις για τις ταινίες GLA, LOR [3]. Με μπλε χρώμα αναπαρίσταται το ποσοστό επιτυχίας για τη μέθοδο των πιθανοτικών ετικετών, ενώ με κόκκινο για αυτή των ντετερμινιστικών.

[3] Τα πειράματα έγιναν για συνάρτηση συμμετοχής step με μηδενικό κατώφλι t, $\epsilon = 0$, πυρήνα χ^2C3D $\alpha = 2$, $\kappa = 1$, $\lambda = 10^{-6}$. Ακόμα, τα τελικά βάρη καθορίζονται από τις επαναλήψεις των bags.



(α′)

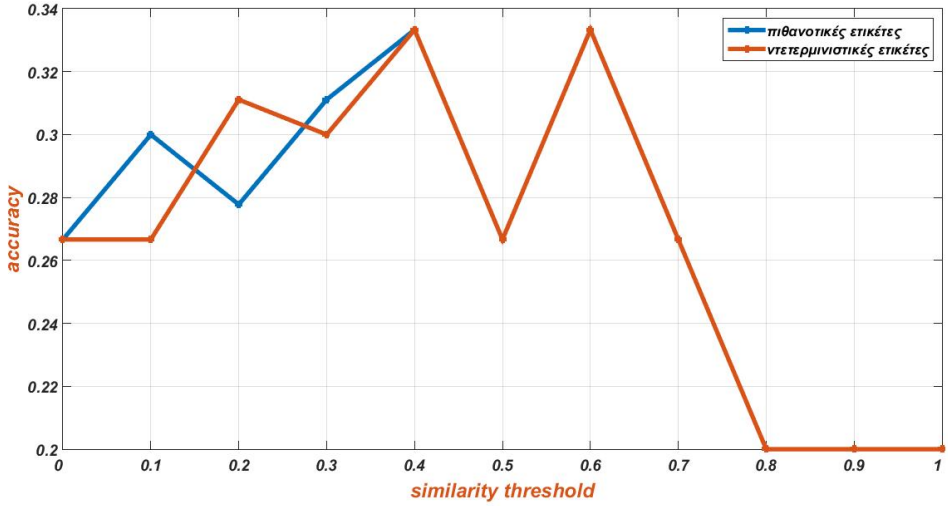


(β′)

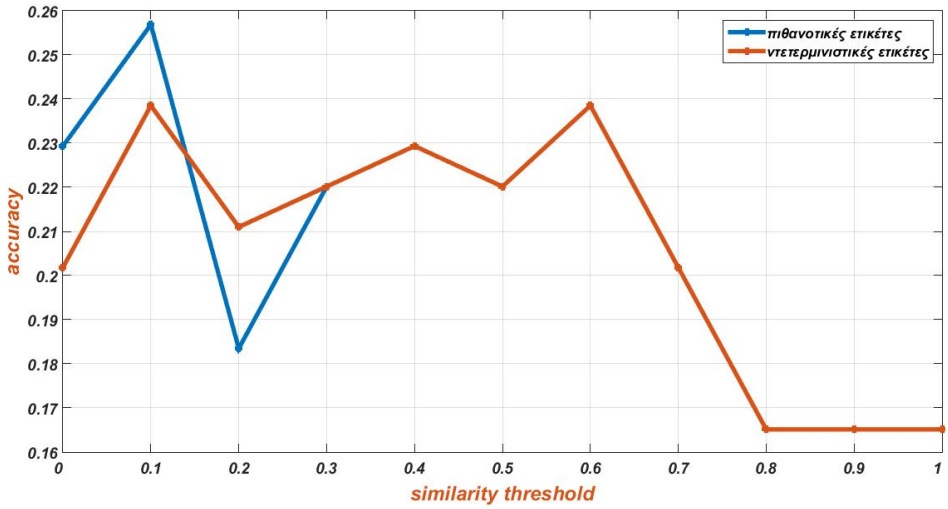
Σχήμα 7.7: Ποσοστά επιτυχίας για πιθανοτικές και ντετερμινιστικές ετικέτες συναρτήσει του *similarity threshold* για την ταινία GLA για 4 και 6 κλάσεις.

Παρατηρούμε, ότι για τιμές κατωφλίου μεγαλύτερες του 0.7 τα αποτελέσματα παρουσιάζουν αξιοσημείωτη επιδείνωση. Ειδικά στην περίπτωση των 4 κλάσεων τα αποτελέσματα για κάποιες τιμές του *similarity threshold* έχουν τεθεί καταχρηστικά ίσα με το 0[4] καθώς δεν υπάρχει κανένας περιορισμός, δηλαδή δεν υπάρχει καμία τιμή ομοιότητας πάνω από το κατώφλι. **Είναι χαρακτηριστικό ότι στο σενάριο της ταινίας LOR δεν υπάρχει καμία εμφάνιση της λέξης walk ή άλλων παραγώγων της, ενώ στο βίντεο είναι η 2η πιο πολυπληθής κλάση!**. Αντιλαμβανόμαστε έτσι τη σημασία της αξιοποίησης περι-

[4]Αυτό δεν ανταποκρίνεται στην πραγματικότητα καθώς η κυρτή βελτιστοποίηση μπορεί να εκτελεστεί και χωρίς περιορισμούς, αλλά όπως έχουν δείξει στο [1] και αλλού, συχνά τέτοια προβλήματα discriminative clustering εκφυλίζονται σε τετριμμένες λύσεις όπου ομαδοποιούν όλα τα δεδομένα σε ένα cluster.



(α′)



(β′)

Σχήμα 7.8: Όμοια με 7.7 για την ταινία LOR

γραφών με παρόμοια αλλά όχι ακριβώς ίδια σημασιολογία. Το πόσο παρόμοια πρέπει να είναι αυτή η σημασιολογία καθορίζεται από το *similarity threshold*. Η επιλογή του γίνεται εδώ εμπειρικά στην τιμή **0.4** καθώς όπως φαίνεται και από τα διαγράμματα παρουσιάζει σταθερά υψηλές τιμές αποτελεσμάτων.

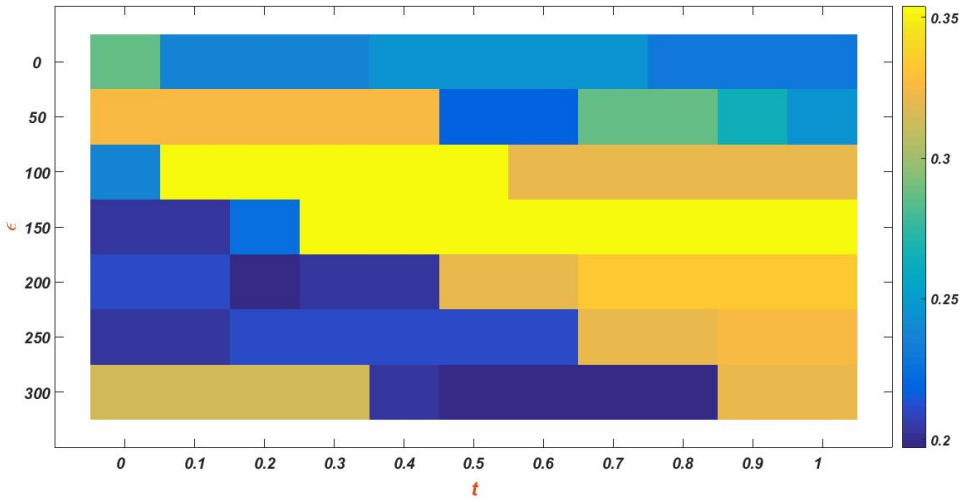
Όσον αφορά τη σύγκριση μεταξύ των μεθόδων των πιθανοτικών και των ντετερμινιστικών ετικετών, δεν είναι καθαρό αν κάποια από τις 2 υπερέχει και για αυτό το λόγο θα αξιολογηθούν ποσοτικά παρακάτω. Σημειώνουμε εδώ μόνο για υπενθύμιση ότι η μέθοδος των ντετερμινιστικών ετικετών εξάγει μία ετικέτα για κάθε φράση και συγκεκριμένα αυτήν με τη μέγιστη τιμή ομοιότητας. Αντίθετα, η μέθοδος των πιθανοτικών εξάγει μία κατανομή πιθανότητας πάνω στις ετικέτες, κανονικοποιώντας το διάνυσμα ομοιοτήτων, και στη συνέχεια παράγει τόσους

περιορισμούς όσες και οι ετικέτες που έχουν μη μηδενική τιμή πιθανότητας. Τα αρχικά βάρη των μεταβλητών χαλάρωσης στην αντικειμενική συνάρτηση τίθενται ίσα με τη τιμή πιθανότητας που αντιστοιχεί στην ετικέτα κάθε περιορισμού (βλέπε κεφάλαιο 4).

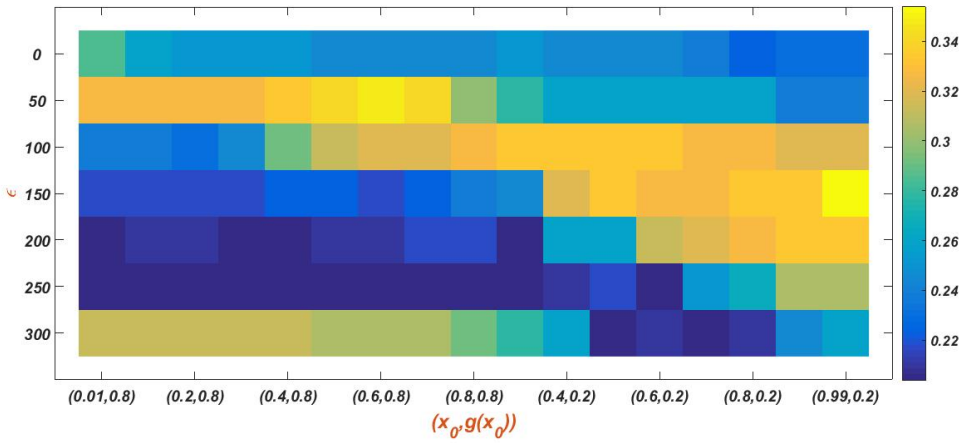
Ασαφή σύνολα πολλαπλών παραδειγμάτων

Η μοντελοποίηση εδώ γίνεται ακριβώς όπως περιγράφηκε στην ενότητα της αναγνώρισης προσώπου. Εκτελούμε και πάλι παρόμοια πειράματα προκειμένου να καθορίσουμε τις υπερπαραμέτρους. Ενδεικτικά, εδώ παραθέτουμε μέσω των σχημάτων 7.9, 7.10 τα ποσοστά επιτυχίας για τις ταινίες GLA και LOR για την περίπτωση των 6 κλάσεων και για κάθε μία συνάρτηση συμμετοχής[5].

[5]Τα πειράματα έγιναν για κατώφλι *similarity threshold* $= 0.4$, πιθανοτικές ετικέτες, επαναλαμβανόμενα bags, πυρήνα χ^2C3D, $\alpha = 2$, $\kappa = 1$, $\lambda = 10^{-6}$

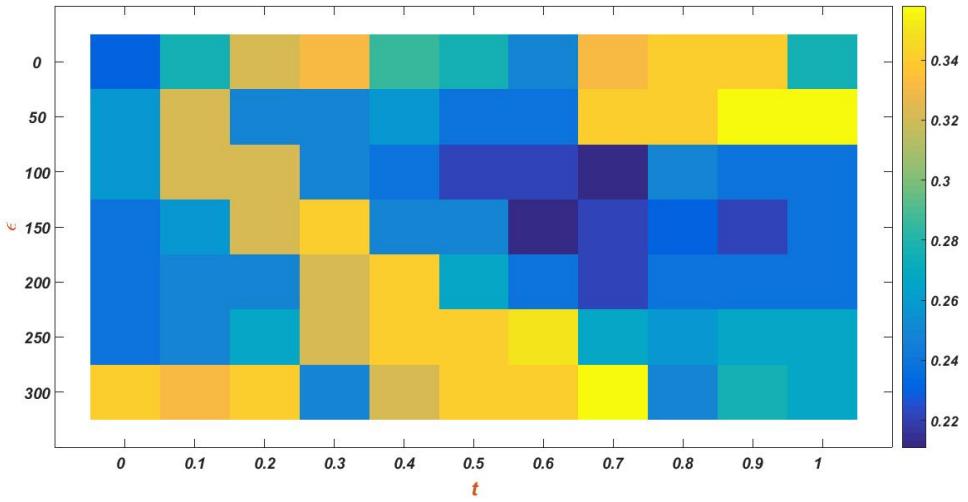


(α′)

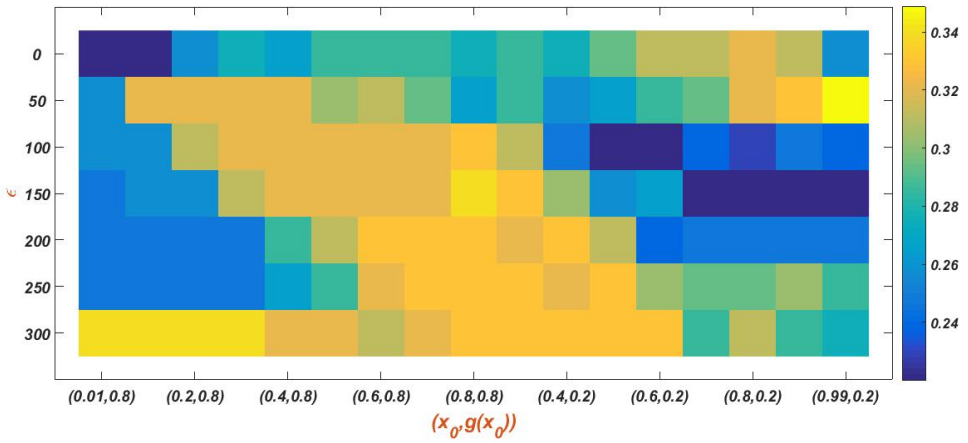


(β′)

Σχήμα 7.9: Ποσοστά επιτυχίας για την ταινία GLA για την οικογένεια συναρτήσεων συμμετοχής που προκύπτει από την παρεμβολή pchip - (α′) και για την οικογένεια των βηματικών συναρτήσεων - (β′)



(α′)



(β′)

Σχήμα 7.10: Όμοια με 7.9 για την ταινία LOR

Εδώ ,όπως αναμέναμε, η χρονική επέκταση ϵ είναι ιδιαίτερα ευεργετική. Συγκεκριμένα, οι ετικέτες που προέρχονται από το κείμενο είναι πολύ αραιές και λίγες, λόγω της εφαρμογής του κατωφλίου ομοιότητας. Μεγαλώνοντας τα χρονικά όρια κάθε γλωσσικού αντικειμένου, αυξάνουμε τα δείγματα που θα περιέχονται στο bag του. Με αυτόν τον τρόπο, παρ ΄όλο που κάνουμε τους περιορισμούς πιο σύνθετους, υποδεικνύουμε στον αλγόριθμο να εξερευνήσει ομοιότητες μεταξύ δεδομένων που δεν μπορούσε να εξερευνήσει πριν. Άρα, αν τα ίδια τα δείγματα έχουν υψηλή ικανότητα διαχωρισιμότητας, τότε από τους πιο γενικούς περιορισμούς ίσως προκύψουν καλύτερες ομαδοποιήσεις τους. Σε αντιπαραβολή με το πρόβλημα της αναγνώρισης χαρακτήρων, βλέπουμε ότι η χρονική επέκταση που χρειάζεται για τα 2 διαφορετικά προβλήματα κατά τη διάρκεια της βελτιστοποίησης είναι διαφορετική. Αυτό δείχνει ότι δεν αντισταθμίζει απαραίτητα τον θόρυβο που προκύπτει από την αρχική ευθυγράμμιση βίντεο -

κειμένου, αλλά μάλλον η συνεισφορά της είναι η κατασκευή γενικότερων bags, ακόμα και αν πολλά δείγματα από αυτά που θα περιέχονται σε αυτά πλέον είναι άχρηστα. Έτσι, με τεχνητό τρόπο δίνουμε βοήθεια στον αλγόριθμο έτσι ώστε να βρει καλύτερες λύσεις.

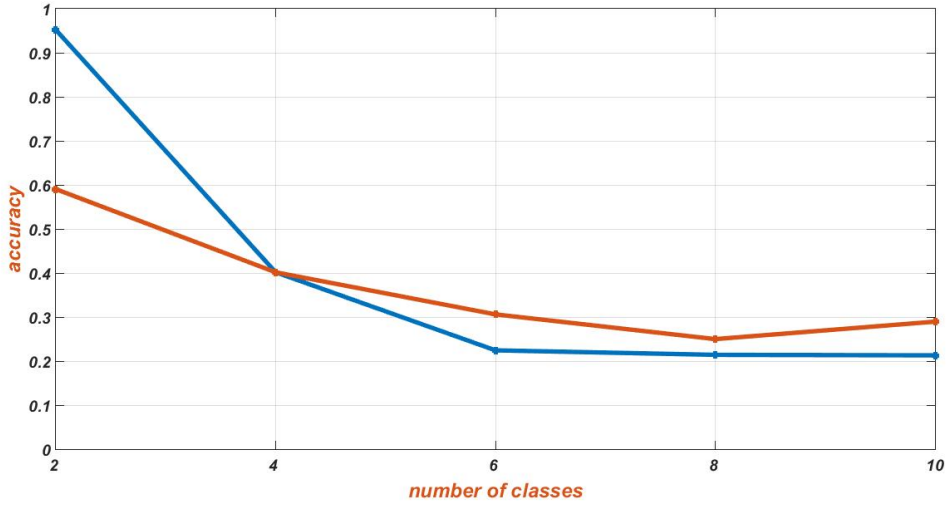
Ακόμα, να αναφέρουμε ότι εδώ είναι πολύ έντονο το φαινόμενο που αναφέραμε στην ενότητα της αναγνώρισης προσώπου. Ότι, δηλαδή, όσο αυξάνουμε το ϵ, πρέπει ταυτόχρονα να αυξάνεται και η αντίστοιχη παράμετρος - 'κατώφλι' της συνάρτησης συμμετοχής προκειμένου τα ποσοστά επιτυχίας να διατηρούνται σε υψηλά επίπεδα. Με αυτόν τον τρόπο περιορίζουμε σε ένα βαθμό την δημιουργία υπερβολικά μεγάλων bags και ταυτόχρονα αξιοποιούμε τη βοήθεια που προσφέρει η χρονική επέκταση και εξηγήσαμε στην προηγούμενη παράγραφο.

Να σημειώσουμε τέλος, ότι είναι δύσκολο να βρεθεί ένα αρκετά γενικό μοτίβο που να μας υποδεικνύει τη σωστή επιλογή των παραμέτρων, καθώς όπως παρατηρήσαμε και από αρκετά ακόμα πρωταρχικά πειράματα, υπάρχει μεγάλη εξάρτηση από ένα πλήθος παραγόντων που δεν μπορούν να προσδιοριστούν με σιγουριά και άρα τα αποτελέσματα εμφανίζουν μεγάλη διασπορά. Ως ένα ικανοποιητικό trade off επιλέξαμε τις τιμές $\boldsymbol{\epsilon = 150 frames}$, $\boldsymbol{(x_0, g(x_0)) = (0.5, 0.8)}$ για την pchip και $\boldsymbol{t = 0.3}$ για την step.

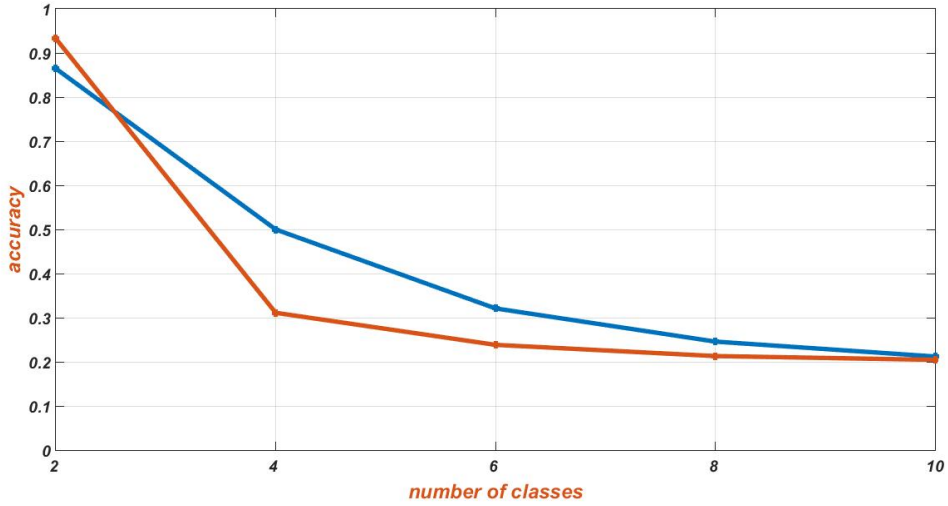
Τελικά βάρη των Μεταβλητών Χαλάρωσης - Επαναλήψεις των bags

Το τρίτο και τελευταίο ποιοτικό πείραμα που γίνεται αφορά την αλλαγή των βαρών των μεταβλητών χαλάρωσης με βάση τις επαναλήψεις κάθε bag. Το τελικό βάρος για κάθε περιορισμό τίθεται, όπως αναφέραμε στο κεφάλαιο 4 ίσο με το άθροισμα των αρχικών βαρών των μεταβλητών χαλάρωσης των επιμέρους επαναλήψεων του. Θέτοντας τις παραμέτρους στις τιμές που αναφέραμε προηγουμένως και για συνάρτηση συμμετοχής παραγόμενη από την pchip, δείχνουμε στα σχήματα 7.11α΄, 7.11β΄ τα ποσοστά επιτυχίας για 2, 4, 6, 8 και 10 κλάσεις για τις ταινίες GLA και LOR[6]. Οι 2 μέθοδοι που συγκρίνονται διαφοροποιούνται μόνο στην μοντελοποίηση των επαναλήψεων των bags. Δηλαδή, στην μία περίπτωση διατηρούμε μόνο τους διαφορετικούς περιορισμούς που έχουν ταυτόχρονα και διαφορετικά βάρη. Άρα για έναν επαναλαμβανόμενο περιορισμό με διαφορετικά βάρη στις επαναλήψεις του, το τελικό του βάρος θα προκύψει ως άθροισμα τους. Στην δεύτερη περίπτωση, κάθε ένα από αυτά τα αρχικά βάρη μπορεί να εμφανίζεται παραπάνω από μία φορές. Τελικά, θα συμμετέχει στο άθροισμα πολλαπλασιασμένο ακριβώς με αυτό τον αριθμό φορών. Παρατηρούμε, όπως ακριβώς και με την αναγνώριση προσώπου, ότι η προσθήκη αυτή βελτιώνει σταθερά τα αποτελέσματα στο LOR, ενώ στο GLA δεν φαίνεται να προσφέρει κάτι (με εξαίρεση την περίπτωση των 2 κλάσεων που είναι ιδιαίτερα απροσδόκητο αποτέλεσμα). Δεν θα επεκταθούμε εδώ παραπάνω καθώς η ανάλυση έγινε προηγουμένως. Θα διατηρήσουμε παντού την μεθοδολογία της αλλαγής των βαρών με βάση τις επαναλήψεις, καθώς όπως είπαμε, δεν αναμένουμε να έχει αρνητική συνεισφορά στα αποτελέσματα, αλλά, πιθανόν σε κάποιες περιπτώσεις, ουδέτερη.

[6]Τα πειράματα έγιναν για κατώφλι *similarity threshold* $= 0.4$, πιθανοτικές ετικέτες, συνάρτηση pchip,$(x_0, g(x_0)) = (0.5, 0.8)$, $\epsilon = 150$, πυρήνα χ^2C3D, $\alpha = 2$, $\kappa = 1$, $\lambda = 10^{-6}$



(α′)



(β′)

Σχήμα 7.11: Ποσοστά επιτυχίας για 2, 4, 6, 8 και 10 κλάσεις για βάρη ίσα με τα αρχικά (καθοριζόμενα από τις πιθανοτικές ετικέτες) -κόκκινο χρώμα- και για βάρη ίσα με το άθροισμα των αρχικών βαρών των μεταβλητών χαλάρωσης των επιμέρους επαναλήψεων του -μπλε χρώμα. Στο σχήμα (α′) απεικονίζεται το GLA και στο (β′) το LOR.

Συγκρίσεις Μεθόδων

Όπως και στην αναγνώριση προσώπου, το **baseline** του [4] αντιμετωπίζει το πρόβλημα ως μάθηση πολλαπλών παραδειγμάτων όπου για να ανήκει ένα δεδομένο σε ένα bag αρκεί η επικάλυψη να είναι διάφορη του μηδενός. Οι ετικέτες που αποδίδονται στα bags είναι ντετερμινιστικές. Η διαδικασία που προτείνουν είναι δύσκολο να γενικευθεί και είναι αρκετά περιοριστική καθώς επιλέγει τις υπό αναγνώριση δράσεις που η γλωσσική τους αναπαράσταση (λέξη/φράση) εμφανίζεται πολύ συχνά στο κείμενο. Είναι παρόμοιας λογικής με την επιλογή

φράσεων με μεγάλη τιμή ομοιότητας και για αυτό το λόγο θα την υλοποιήσουμε λαμβάνοντας ένα υψηλό κατώφλι ομοιότητας (κοντά στο 1). Στον αντίποδα υλοποιούμε το baseline λαμβάνοντας το κατώφλι ομοιότητας να είναι μηδενικό, προκειμένου να αναδείξουμε ακριβώς τη σημασία ύπαρξης του ενδιάμεσου κατωφλίου. Τα διανύσματα χαρακτηριστικών που χρησιμοποίησαν οι συγγραφείς είναι ιστογράμματα που προκύπτουν από μοντέλα bag of visual words πυκνών τροχιών ([62]) και ο πυρήνας είναι ο χ^2. Στην εν λόγω διπλωματική δεν επεκταθήκαμε ιδιαίτερα στο ζήτημα της επιλογής χαρακτηριστικών για το πρόβλημα της αναγνώρισης δράσεων όποτε επιλέξαμε για όλα μας τα πειράματα, καθώς και για την υλοποίηση του baseline να χρησιμοποιήσουμε τον ίδιο περιγραφητή ο οποίος είναι ο C3D. Η επιλογή έγινε λόγω της ευκολίας εξαγωγής των διανυσμάτων σε συνδυασμό με τα προσδοκώμενα αποτελέσματα, λόγω της βαθιάς αρχιτεκτονικής του C3D και της δυνατότητας του να εφαρμόζεται σε πολλά διαφορετικά προβλήματα όρασης.

Οι επεκτάσεις του baseline έγκεινται:

- στην προσθήκη του κατωφλίου ομοιότητας σε μία ενδιάμεση τιμή - 0.4 (τα απλά deterministic και probabilistic settings περιέχουν μόνο αυτήν την αλλαγή),
- στη μετάβαση από το απλό deterministic setting της 1 ετικέτας για κάθε bag (αυτής με τη μέγιστη ομοιότητα) και άρα ενός περιορισμού, στο probabilistic setting με την προσθήκη ενός περιορισμού για κάθε ετικέτα με μη μηδενική πιθανότητα με αρχικό βάρος μεταβλητής χαλάρωσης ίσο με την πιθανότητα,
- στην αλλαγή του τρόπου συμμετοχής των δεδομένων στα bags, είτε αυτός ο τρόπος κατασκευάζει απλά σύνολα (step), είτε ασαφή (pchip),
- και στην επιλογή μεταξύ των 2 πυρήνων χ^2 (χ^2C3D) και γραμμικού (linC3D).

Στους πίνακες 7.7, 7.8, 7.9, 7.10, 7.11, 7.12, συγκεντρώνονται τα τελικά αποτελέσματα για κάθε ταινία, ενώ στον 7.13 παρουσιάζεται ο μέσος όρος τους. Όλα τα πειράματα έγιναν για τιμές υπερπαραμέτρων $\alpha = 2$, $\kappa = 1$, $\lambda = 10^{-6}$. Τέλος, στον πίνακα 7.2 αναφέρουμε με φθίνουσα σειρά τις 10 πολυπληθέστερες κλάσεις κάθε ταινίας, πάνω στις οποίες πειραματιστήκαμε.

Γενικές παρατηρήσεις:

Αναφέρουμε εδώ τα γενικά συμπεράσματα που προκύπτουν από τις συγκρίσεις των πειραμάτων που παρουσιάζονται παρακάτω. Αρχικά, βλέπουμε ότι η τοποθέτηση του κατωφλίου similarity threshold σε μία ενδιάμεση θέση (*similarity threshold* $= 0.4$) βελτιώνει σημαντικά τα αποτελέσματα σε σύγκριση με την επιλογή μόνο των προφανώς σωστών ετικετών, δηλαδή αυτών που έχουν ομοιότητα 1, όπως περίπου γίνεται στο baseline του [4]. Αυτό γιατί μοντελοποιούμε καλύτερα το γεγονός ότι οι ετικέτες μας μπορεί να εκφράζονται στο κείμενο με διαφορετικούς τρόπους με παρόμοια σημασιολογία. Επίσης, βελτιώνει τα αποτελέσματα συγκριτικά με την επιλογή της ετικέτας με τη μέγιστη τιμή ομοιότητας, όποια και αν είναι αυτή, δηλαδή θέτοντας (*similarity threshold* $= 0$). Αυτό γιατί, αποφεύγουμε έτσι να συμπεριλάβουμε προτάσεις που δεν σημαίνουν κάτι από αυτά που προσπαθούμε να αναγνωρίσουμε και

Movie Name	**Action Classes**
BMI	*walk, turn, running, cry, smile, open door, driving car, sitting up, sitting down, getting out of the car*
CRA	*walk, turn, cry, threaten person, running, driving car, fall on the floor, pick, standing up, getting out of the car*
DEP	*smile, walk, answering phone, smoke, turn, laugh, chew, drink, eat, standing up*
GLA	*walk, ride horse, wave hands, point at something, pick, standing up, shoot bow*
GWW	*walk, turn, smile, wave hands, running, grab hand, climb stairs, pick, hugging, cry*
LOR	*ride horse, walk, running, standing up, climb, chew, eat, sitting down, smile, cry*

Πίνακας 7.2: Οι 10 πολυπληθέστερες κλάσεις δράσεων για κάθε ταινία σε φθίνουσα σειρά

άρα εισάγουν θορυβώδεις ετικέτες. Η μόνη εξαίρεση είναι η ταινία GWW, όπου το μεγάλο πλήθος των δειγμάτων, η δυσκολία ομαδοποίησης τους και η ύπαρξη μεγάλου θορύβου από τις ετικέτες εισάγουν μεγάλη τυχαιότητα στο πρόβλημα την οποία δεν μπορούμε να μοντελοποιήσουμε σωστά. Αναφέρουμε εδώ ότι όπως είναι λογικό το off-the-shelf σύστημα υπολογισμού ομοιότητας που έχουμε χρησιμοποιήσει ([24]) πολλές φορές υποεκτιμά τιμές ομοιοτήτων με αποτέλεσμα λανθασμένα να μην ξεπερνούν το κατώφλι. Παρ΄ όλα αυτά, όμως, δεδομένου του ότι δεν χρειάζεται (επαν)εκπαίδευση πάνω στο σύνολο φράσεων που μας ενδιαφέρουν όπως γίνεται στο [32] και ταυτόχρονα προσφέρει στο σύστημα αναγνώρισης περιορισμούς που δεν παρέχουν οι άλλες μέθοδοι, κρίνουμε ότι μπορεί να θέσει τις βάσεις για ένα πιο εξελιγμένο σύστημα εξαγωγής γλωσσικών ετικετών .

Ακόμα, βλέπουμε ότι η ταυτόχρονη επέκταση των χρονικών ορίων κάθε bag μέσω της παραμέτρου $\epsilon = 150$ και η απόρριψη δειγμάτων με μικρή επικάλυψη μέσω της επιλογής των παραμέτρων $t = 0.3$ ή $(x_0, g(x_0)) = (0.5, 0.8)$ πετυχαίνει τη συμμετοχή σε κάθε bag ενός λιγότερο θορυβώδους συνόλου δειγμάτων. Δηλαδή, παίρνουμε ένα καλό trade off μεταξύ των δειγμάτων που σωστά συμμετέχουν στο bag και αυτών που συμμετέχουν λανθασμένα. Αυτό φαίνεται ιδιαίτερα στην ταινία DEP, όπου οι τιμές του accuracy για $\epsilon = 0$, $t = 0$ είναι πολύ χαμηλές (χαρακτηριστικά για 2 κλάσεις linC3D+probabilistic:39.76% , ενώ για linC3D+probabilistic+step:96.39%, linC3D+probabilistic+pchip: 90.36%).

Όσον αφορά τις συναρτήσεις πυρήνων που εφαρμόσαμε, ο γραμμικός υπερέχει του χ^2 στις περισσότερες περιπτώσεις, έστω και με μικρές διαφορές. Αυτό είναι λογικό, καθώς ο χ^2, παρ΄ όλο που είναι συνηθισμένος στην αναγνώριση δράσεων, έχει μεγάλη καταλληλότητα σε περιπτώσεις υπολογισμού αποστάσεων μεταξύ ιστογραμμάτων. Άλλωστε, όπως διαβάζουμε

στο [58] το δίκτυο C3D πετυχαίνει σημαντικά ποσοστά επιτυχίας σε προβλήματα ταξινόμησης δράσεων όταν εισάγεται σε ένα SVM γραμμικού πυρήνα, επομένως το συμπέρασμα μας ταυτίζεται με αυτό των δημιουργών του. Να σημειώσουμε εδώ ότι τα καλύτερα αποτελέσματα που παρουσιάζονται στο [58] προκύπτουν μετά από συνδυασμό του C3D με ένα μοντέλο Bag of Visual Words πυκνών τροχιών (iDT) και κατόπιν εισαγωγής των διανυσμάτων σε γραμμικό SVM. Μπορούμε να χρησιμοποιήσουμε αυτό το είδος πυρήνα σε μελλοντική εργασία.

Οι πιθανοτικές ετικέτες, παρατηρήσαμε ότι πολύ συχνά οδηγούν σε έστω και μικρή βελτίωση των αποτελεσμάτων σε σύγκριση με τις αντίστοιχες ντετερμινιστικές. Η συνεισφορά τους γίνεται πιο έντονη όσο αυξάνεται το πλήθος των κλάσεων, όπου γίνεται όλο και λιγότερο ξεκάθαρο ποια ετικέτα κρύβεται μέσα σε κάθε πρόταση. Συγκεκριμένα, δεδομένου ότι πολλές δράσεις έχουν έτσι και αλλιώς σημασιολογική ομοιότητα, είναι πιθανό κάποιες προτάσεις να έχουν κοντινές τιμές ομοιότητας με παραπάνω από μία κλάσεις. Έτσι, όπως φαίνεται και από τα πειράματα μας, δεν αρκεί να κρατήσουμε μόνο την ετικέτα με τη μέγιστη τιμή ομοιότητας (ως ντετερμινιστική ετικέτα) καθώς έτσι χάνεται πολλή πληροφορία. Αντίθετα, όταν οι κλάσεις είναι λίγες είναι πιο εύκολο να διακρίνουμε την κρυμμένη ετικέτα κάθε πρότασης και άρα η μοντελοποίηση με τις πιθανοτικές ετικέτες κάνει το πρόβλημα πιο σύνθετο από ότι χρειάζεται καθώς προσθέτει επιπλέον - λανθασμένους - περιορισμούς στη διαδικασία της βελτιστοποίησης. Παρόμοιο συμπέρασμα μπορεί να βγάλει κανείς και για τη μοντελοποίηση με ασαφή σύνολα, καθώς όπως είδαμε η συνάρτηση που προκύπτει από την παρεμβολή pchip υπερέχει της βηματικής συνάρτησης όσο αυξάνουμε το πλήθος των κλάσεων, με εξαίρεση την ταινία GLA.

Επιμέρους παρατηρήσεις για την κάθε ταινία παρέχονται στους αντίστοιχους πίνακες.

Χρήση πρότερης γνώσης από τη διαδικασία αναγνώρισης προσώπου

Εφαρμόζουμε εδώ τη ενσωμάτωση αυτού του είδους πρότερης γνώσης όπως περιγράφεται στο κεφάλαιο 4 και ειδικότερα από το constraint της σχέσης (4.15). Για να αξιολογήσουμε τα αποτελέσματα της αξιοποίησης της πρότερης γνώσης από την αναγνώριση προσώπου, πρέπει αρχικά να ελέγξουμε κατά πόσο υπάρχει αντιστοιχία μεταξύ των tracks του προσώπου και των tracks των δράσεων. Αυτό γιατί, αν η συνολική αντιστοίχηση είναι μικρή, τότε η ροή πληροφορίας από τα πρόσωπα στις δράσεις δεν θα είναι ουσιώδης και δεν θα μπορεί να παρατηρηθεί αξιοσημείωτη βελτίωση. Συγκεκριμένα, ποσοτικοποιούμε τη συνολική αντιστοίχηση υπολογίζοντας το λόγο $\frac{sum_i^{NA} b_{i(N_P+1)}}{N_A}$, που δείχνει τι ποσοστό των tracks των δράσεων δεν αντιστοιχίζονται σε κάποιο track προσώπου. Παίρνοντας την τιμή $1 - \frac{sum_i^{NA} b_{i(N_P+1)}}{N_A}$, αποκτούμε μία διαίσθηση της συνολικής αντιστοίχησης. Αυτό δείχνει ο πίνακας 7.3. Συγκεκριμένα, για κάθε ταινία επιλέγουμε το σύνολο των tracks προσώπου (και αυτά που ανήκουν στο παρασκήνιο - άλλωστε δεν μπορούμε να ξέρουμε ποια είναι τα δεδομένα προσκηνίου σε όλες τις ταινίες ελλείψει επισημείωσης) και τα tracks δράσεων για 2 ,4 ,6 ,8 και 10 κλάσεις και ποσοτικοποιούμε τη συνολική αντιστοίχισή τους.

Παρατηρούμε ότι προφανώς η καταλληλότερη ταινία για αξιολόγηση της εν λόγω μεθόδου είναι η DEP, η οποία δεν έχει επισημειωμένα πρόσωπα. Για αυτό θα εκτελέσουμε τα πειράματα

number of classes	2	4	6	8	10	average
BMI	0.35	0.32	0.35	0.36	0.35	0.35
CRA	0.26	0.33	0.30	0.27	0.26	0.28
DEP	0.69	0.66	0.65	0.66	0.65	0.66
GLA	0.17	0.27	0.26	0.23	0.23	0.23
GWW	0.42	0.41	0.42	0.43	0.43	0.42
LOR	0.15	0.20	0.19	0.19	0.21	0.19

Πίνακας 7.3: Ποσοτικοποίηση της συνολικής αντιστοίχησης μεταξύ tracks προσώπων και tracks δράσεων

και στην ταινία GLA που έχει την υψηλότερη συνολική αντιστοίχηση από τις 2 επισημειωμένες. Με αυτήν μπορούμε να ελέγξουμε πως συμπεριφέρεται το σύστημα αν χρησιμοποιήσουμε τις πραγματικές ετικέτες των προσώπων (GT). Οι άλλες 2 μέθοδοι που συγκρίνουμε είναι η επιλογή του πίνακα των κλάσεων των προσώπων πριν ($\hat{\boldsymbol{Z}}$) και μετά το rounding ($\boldsymbol{Z}$)

Η διαδικασία της βελτιστοποίησης για το πρόσωπο γίνεται για συνάρτηση step με $t = 0.1$ για την ταινία GLA (όπως είδαμε παραπάνω πετυχαίνει τα καλύτερα αποτελέσματα στην αναγνώριση προσώπου του GLA) και για συνάρτηση pchip με $(x_0, g(x_0)) = (0.2, 0.8)$ για το DEP. Χρησιμοποιούμε βάρη επαναλήψεων και πυρήνα VGG1 και για τις 2. Για τις δράσεις χρησιμοποιούμε *similarity threshold* $= 0.4$, probabilistic labels και γραμμικό πυρήνα linC3D που είδαμε ότι γενικά πετυχαίνουν καλά αποτελέσματα. Ελέγχουμε και τις 2 συναρτήσεις συμμετοχής των δειγμάτων στα bags (step, pchip) για τις τιμές των παραμέτρων που έχουμε επιλέξει.

number of classes	2	4	6	8	10
linC3D+probabilistic+step	66.27%	40.16%	**33.33%**	**27.98%**	15.85%
linC3D+probabilistic+pchip	66.27%	40.16%	23.13%	17.86%	17.49%
linC3D+probabilistic+step+$\boldsymbol{Z}$	**67.47%**	40.98%	**33.33%**	24.40%	**31.69%**
linC3D+probabilistic+pchip+$\boldsymbol{Z}$	66.27%	40.16%	21.77%	14.29%	22.40%
linC3D+probabilistic+step+$\hat{\boldsymbol{Z}}$	66.27%	**43.44%**	**33.33%**	23.21%	25.68%
linC3D+probabilistic+pchip+$\hat{\boldsymbol{Z}}$	66.27%	**43.44%**	21.77%	14.88%	21.86%
linC3D+probabilistic+step+GT	66.27%	40.16%	33.33%	27.98%	15.30%
linC3D+probabilistic+pchip+GT	66.27%	40.16%	21.09%	16.07%	16.94%

Πίνακας 7.4: Συγκριτικός πίνακας πειραμάτων για τη χρήση πρότερης γνώσης από την αναγνώριση προσώπου για την ταινία GLA

Όπως είναι φανερό από τους πίνακες 7.4, 7.5, η μετάδοση γνώσης από το σύστημα αναγνώρισης προσώπου σε αυτό της αναγνώρισης δράσης, μπορεί πράγματι να βελτιώσει αισθητά τα αποτελέσματα. Μάλιστα, βλέπουμε ότι στο GLA, παρά τη μικρή συνολική αντιστοίχηση, υπάρχουν κάποιοι συνδυασμοί μεθόδων που βελτιώνουν τα αποτελέσματα. Ειδικότερα, για 4 και 10 κλάσεις, τα αποτελέσματα 43.44% και 31.69% είναι τα καλύτερα που παίρνουμε συγκρι-

number of classes	2	4	6	8	10
linC3D+probabilistic+step	**96.39%**	**74.17%**	43.75%	36.14%	31.87%
linC3D+probabilistic+pchip	90.36%	71.67%	45.14%	39.16%	35.16%
linC3D+probabilistic+step+Z	86.75%	69.17%	**56.94%**	50.00%	42.86%
linC3D+probabilistic+pchip+Z	90.36%	68.33%	**56.94%**	**50.60%**	43.41%
linC3D+probabilistic+step+$\hat{Z}$	72.29%	72.50%	53.47%	49.40%	42.86%
linC3D+probabilistic+pchip+$\hat{Z}$	71.08%	72.50%	53.47%	48.80%	**43.96%**

Πίνακας 7.5: Συγκριτικός πίνακας πειραμάτων για τη χρήση πρότερης γνώσης από την αναγνώριση προσώπου για την ταινία DEP

τικά με όλες τις διαφορετικές τεχνικές του συστήματος αναγνώρισης δράσης όταν λειτουργεί μόνο του (πίνακας 7.10). Το παράξενο είναι ότι οι προβλέψεις του συστήματος λειτουργούν καλύτερα από ότι οι πραγματικές ετικέτες, ενώ περιμέναμε αυτές να μας έδιναν ένα upper bound. Πιθανόν, ό λόγος που συμβαίνει αυτό είναι οι αντιστοιχίσεις μεταξύ των tracks, οι οποίες εκτός του ότι είναι μικρές, δεν είναι απαραίτητο ότι πάντα θα βοηθούν τον αλγόριθμο καθώς ο υπολογισμός τους είναι απλά χρονικός άρα οι τιμές τους δεν ανταποκρίνονται πάντα στην πραγματικότητα. Άρα, πιθανόν, στη συγκεκριμένη περίπτωση τα λάθη του συστήματος του προσώπου να λειτουργούν βοηθητικά αντισταθμίζοντας τα λάθη των αντιστοιχήσεων. Ακόμα πιο σημαντικά είναι τα αποτελέσματα για την ταινία DEP. Εδώ, βλέπουμε ότι παρότι δεν έχουμε καμία γνώση για την απόδοση του συστήματος του προσώπου, η βελτίωση που προσφέρει στα αποτελέσματα για 6, 8 και 10 κλάσεις είναι αισθητή. Συγκεκριμένα, τα αποτελέσματα 45.14%, 39.16%, 35.16 % ήταν μέχρι τώρα τα καλύτερα που είχαμε πετύχει και τώρα βλέπουμε ότι ξεπερνιόνται όλα κατά 10% περίπου (56.94%, 50.60%, 43.96%). Αυτό δείχνει ότι έχει πολύ μεγάλη σημασία να υπάρχει σχετικά μεγάλη αντιστοίχηση μεταξύ των tracks όπως συμβαίνει εδώ. Υποθέτουμε ότι η αναγνώριση προσώπου έχει και εδώ καλά αποτελέσματα. Να σημειώσουμε ότι από τους πίνακες αυτούς φαίνεται να υπερτερεί η επιλογή της integral λύσης $\boldsymbol{Z}$ του αλγορίθμου, χωρίς αυτό να έχει όμως τόσο μεγάλη σημασία

Χρήση πρότερης γνώσης από προεκπαιδευμένο ταξινομητή

Η τελευταία μοντελοποίηση που αξιολογούμε είναι αυτή που δίνεται από τη σχέση (4.16). Συγκεκριμένα, ο προεκπαιδευμένος ταξινομητής μας δίνει τα confidence scores για ένα ενδεικτικό πείραμα 5 κλάσεων (*sitting down, standing up, ride horse , open door, point at something*), και πάλι μόνο για τις δράσεις προσκηνίου που μας αφορούν. Για κάθε ταινία που εξετάζουμε, ο ταξινομητής αυτός είναι ένα γραμμικό SVM προεκπαιδευμένο σε όλες τις υπόλοιπες ταινίες της βάσης (τεχνική leave one out). Περισσότερα για αυτόν υπάρχουν στο [56]. Δεν θα μας ενδιαφέρει ιδιαίτερα η προεκπαίδευση αυτή καθώς θέλουμε να εξετάσουμε περισσότερο ποιοτικά την συμπεριφορά του συνδυασμού των συστημάτων.

Στον πίνακα 7.6 συγκρίνουμε τα αποτελέσματα και πάλι για τον καλύτερο συνδυασμό μεθόδων (probabilistic labels, linC3D, *similarity threshold* = 0.4) για τις 2 διαφορετικές συναρτήσεις συμμετοχής. Η παράμετρος u θυμίζουμε ότι ρυθμίζει το βάρος του όρου της

πρότερης γνώσης. Για $u = 0$ παίρνουμε τα ποσοστά επιτυχίας για τη βελτιστοποίηση μόνο μέσω του κειμένου, ενώ για u πολύ μεγάλο, τα ποσοστά για τη βελτιστοποίηση μόνο μέσω του προεκπαιδευμένου ταξινομητή (τα constraints που δεν μπορούν να ικανοποιηθούν παραβιάζονται καθώς αναγκαστικά οι μεταβλητές χαλάρωσης γίνονται πολύ μεγάλες, ενώ ο όρος του clustering στην αντικειμενική συνάρτηση είναι αμελητέος μπροστά στον prior όρο). Είναι φανερό, από τον πίνακα ότι η μέθοδος αυτή, μπορεί να ξεπεράσει τα ποσοστά επιτυχίας του προεκπαιδευμένου ταξινομητή για ενδιάμεσες τιμές του u. Φαίνεται, ότι για αυτές τις τιμές η διαδικασία της βελτιστοποίησης θέτει τη μεταβλητή που προσπαθούμε να ανακτήσουμε κοντά στις τιμές των confidence scores της πρότερης γνώσης και ταυτόχρονα σέβεται τα constraints που θέτει το κείμενο. Έτσι, παρ΄ όλο που ο όρος του clustering μάλλον δεν συνεισφέρει ιδιαίτερα στην τελική λύση, οι άλλοι 2 παράγοντες συνδυάζονται καλά καθώς περιέχουν διαφορετικό είδος πληροφορίας και μας οδηγούν σε καλύτερες λύσεις.

Συμπερασματικά:

Για την αναγνώριση δράσης, προτείνουμε τη χρήση ενός σύγχρονου συστήματος υπολογισμού σημασιολογικής ομοιότητας (ή κατηγοριοποίησης κειμένου - text classification) και εξαγωγή των αμφίσημων ετικετών με βάση αυτό. Επίσης, όπως και στο πρόβλημα της αναγνώρισης προσώπου έχει μεγάλη σημασία η συμμετοχή των δειγμάτων στα bags να γίνεται όταν η επικάλυψη τους ξεπερνά κάποιο κατώφλι, ενώ ειδικά για αυτό το πρόβλημα συνήθως είναι απαραίτητη μία μικρή επέκταση των χρονικών ορίων κάθε bag. Ακόμα είδαμε ότι οι μέθοδοι των ασαφών συνόλων (pchip) και των πιθανοτικών ετικετών μπορούν να αυξήσουν τα ποσοστά αναγνώρισης ιδιαίτερα όσο αυξάνεται το πλήθος των κλάσεων. Τέλος, η χρήση πρότερης γνώσης από ήδη αναγνωρισμένα αντικείμενα, καθώς και ο συνδυασμός της ασθενούς επίβλεψης με τις προβλέψεις ενός προεκπαιδευμένου συστήματος, εξασφαλίζουν ακόμα καλύτερη επίδοση.

u	**0**	**0.1**	**1**	**10**	**100**	**1000**	**10000**	**100000**
BMI:linC3D+ probabilistic+ step	44.00%	44.00%	44.00%	**60.00%**	56.00%	52.00%	60.00%	**60.00%**
BMI:linC3D+ probabilistic+ pchip	44.00%	44.00%	44.00%	**60.00%**	56.00%	52.00%	60.00%	**60.00%**
CRA:linC3D+ probabilistic+ step	36.84%	36.84%	26.32%	26.32%	42.11%	**52.63%**	**52.63%**	**52.63%**
CRA:linC3D+ probabilistic+ pchip	36.84%	36.84%	26.32%	26.32%	42.11%	**52.63%**	**52.63%**	**52.63%**
DEP:linC3D+ probabilistic+ step	29.17%	29.17%	29.17%	33.33%	33.33%	**50.00%**	45.83%	45.83%
DEP:linC3D+ probabilistic+ pchip	29.17%	29.1%	29.17%	33.33%	33.33%	**50.00%**	45.83%	45.83%
GLA:linC3D+ probabilistic+ step	15.28%	47.22%	58.33%	59.72%	**63.89%**	59.72%	61.11%	**61.11%**
GLA:linC3D+ probabilistic+ pchip	15.28%	47.22%	58.33%	59.72%	**63.89%**	59.72%	61.11%	**61.11%**
GWW:linC3D+ probabilistic+ step	20.93%	20.93%	45.35%	**60.47%**	55.81%	56.98%	50.00%	**50.00%**
GWW:linC3D+ probabilistic+ pchip	20.93%	22.09%	45.35%	**60.47%**	55.81%	56.98%	50.00%	**50.00%**
LOR:linC3D+ probabilistic+ step	62.71%	66.10%	66.10%	67.80%	66.10%	**71.19%**	62.71%	**61.02%**
LOR:linC3D+ probabilistic+ pchip	54.24%	66.10%	66.10%	67.80%	66.10%	**71.19%**	62.71%	**61.02%**

Πίνακας 7.6: Συγκριτικός πίνακας πειραμάτων για τη χρήση πρότερης γνώσης από προεκπαιδευμένο ταξινομητή.

number of classes	2	4	6	8	10
linC3D+similarity threshold ≈ 1 (Baseline)	28.70%	22.63%	19.50%	**16.95%**	**15.87%**
linC3D+similarity threshold = 0	78.70%	**54.74%**	**28.93%**	10.17%	6.35%
linC3D+deterministic	77.78%	41.61%	3.77%	5.65%	4.23%
linC3D+deterministic+step	68.52%	30.66%	15.09%	10.73%	8.99%
linC3D+determinstc+pchip	68.52%	27.01%	15.09%	12.43%	10.05%
linC3D+probabilistic	**79.63%**	51.82%	11.32%	5.65%	10.05%
linC3D+probabilistic+step	68.52%	38.69%	11.95%	6.21%	9.52%
linC3D+probabilistic+pchip	60.19%	33.58%	11.95%	10.73%	9.52%
χ^2C3D+similarity threshold ≈ 1 (Baseline)	28.70%	22.63%	19.50%	**23.73%**	**22.22%**
χ^2C3D+similarity threshold = 0	**77.78%**	41.61%	**30.19%**	19.77%	10.05%
χ^2C3D+deterministic	75.00%	39.42%	11.32%	6.78%	4.76%
χ^2C3D+deterministic+step	64.81%	35.77%	25.16%	18.64%	14.29%
χ^2C3D+deterministic+pchip	62.96%	33.58%	25.16%	16.95%	14.81%
χ^2C3D+probabilistic	75.00%	**51.82%**	21.38%	6.78%	10.05%
χ^2C3D+probabilistic+step	62.04%	38.69%	18.24%	14.12%	12.70%
χ^2C3D+probabilistic+pchip	59.26%	35.04%	16.35%	13.56%	11.64%

Πίνακας 7.7: Συγκεντρωτικά αποτελέσματα συγκρίσεων για το πρόβλημα της αναγνώρισης δράσεων στην ταινία BMI για 2, 4, 6, 8 και 10 κλάσεις. Στη συγκεκριμένη ταινία φαίνεται να υπάρχει μεγάλος θόρυβος από τη μία στον χρονικό προσδιορισμό των bags και από την άλλη στις προτάσεις από τις οποίες εξάγουμε τις ετικέτες. Δηλαδή, πολλά από αυτά που περιγράφονται στο κείμενο δεν εμφανίζονται στο βίντεο. Αυτό έχει σαν αποτέλεσμα να μην λειτουργεί επαρκώς ούτε η μέθοδος των ασαφών συνόλων, ούτε η μέθοδος των πιθανοτικών ετικετών, καθώς δεν μπορούν να διαχειριστούν με επιτυχία το πρόβλημα. Αντίθετα, βλέπουμε οτί οι απλές μέθοδοι λειτουργούν καλύτερα. Πράγματι, με μία γρήγορη επισκόπηση του σεναρίου είδαμε ότι υπάρχουν μεγάλες διαφορές μεταξύ βίντεο και κειμένου. Επομένως, η ταινία αυτή δεν προσφέρεται για να βγάλουμε ασφαλή συμπεράσματα.

number of classes	**2**	**4**	**6**	**8**	**10**
linC3D+similarity threshold ≈ 1 (Baseline)	33.33%	22.12%	18.11%	16.08%	14.65%
linC3D+similarity threshold = 0	53.62%	32.69%	19.69%	13.29%	9.55%
linC3D+deterministic	**59.42%**	**39.42%**	14.96%	13.29%	10.83%
linC3D+deterministic+step	55.07%	36.54%	21.26%	22.38%	10.19%
linC3D+determinstc+pchip	55.07%	36.54%	20.47%	22.38%	10.19%
linC3D+probabilistic	55.07%	36.54%	**27.56%**	16.08%	12.10%
linC3D+probabilistic+step	55.07%	33.65%	19.69%	**23.08%**	13.38%
linC3D+probabilistic+pchip	53.62%	34.62%	19.69%	**23.08%**	**16.56%**
χ^2C3D+similarity threshold ≈ 1 (Baseline)	33.33%	22.12%	18.11%	16.08%	14.65%
χ^2C3D+similarity threshold = 0	55.07%	35.58%	25.20%	11.89%	8.92%
χ^2C3D+deterministic	**59.42%**	**36.54%**	18.90%	14.69%	10.83%
χ^2C3D+deterministic+step	49.28%	**36.54%**	24.41%	19.58%	15.29%
χ^2C3D+deterministic+pchip	49.28%	**36.54%**	23.62%	20.28%	15.92%
χ^2C3D+probabilistic	**59.42%**	**36.54%**	23.62%	19.58%	15.92%
χ^2C3D+probabilistic+step	47.83%	34.62%	24.41%	**22.38%**	19.11%
χ^2C3D+probabilistic+pchip	47.83%	35.58%	**25.98%**	**22.38%**	**19.75%**

Πίνακας 7.8: Συγκεντρωτικά αποτελέσματα συγκρίσεων για το πρόβλημα της αναγνώρισης δράσεων στην ταινία CRA για 2, 4, 6, 8 και 10 κλάσεις. Εδώ παρατηρούμε ότι ο πυρήνας linC3D συμπεριφέρεται λίγο καλύτερα στις περισσότερες περιπτώσεις πετυχαίνοντας ποσοστά επιτυχίας 2-3% καλύτερα. Ακόμα, η μέθοδος των πιθανοτικών ετικετών συμπεριφέρεται καλύτερα από αυτή των ντετερμινιστικών ετικετών όσο αυξάνουμε το πλήθος των κλάσεων, ενώ ακόμα και για λίγες κλάσεις δεν υπάρχει σημαντική υπεροχή της 2ης μεθόδου. Όσον αφορά τις συναρτήσεις συμμετοχής, ισχύουν όσα αναφέραμε και στις άλλες ταινίες, ότι δηλαδή η pchip υπερέχει, ιδιαίτερα για περισσότερες κλάσεις.

number of classes	**2**	**4**	**6**	**8**	**10**
linC3D+similarity threshold ≈ 1 (Baseline)	60.24%	56.67%	22.92%	19.88%	13.19%
linC3D+similarity threshold = 0	39.76%	27.50%	26.39%	22.89%	12.64%
linC3D+deterministic	39.76%	27.50%	22.92%	21.08%	18.68%
linC3D+deterministic+step	**96.39%**	**74.17%**	38.89%	31.93%	32.97%
linC3D+determinstc+pchip	90.36%	71.67%	42.36%	37.35%	34.62%
linC3D+probabilistic	39.76%	27.50%	22.92%	21.08%	18.68%
linC3D+probabilistic+step	**96.39%**	**74.17%**	43.75%	36.14%	31.87%
linC3D+probabilistic+pchip	90.36%	71.67%	**45.14%**	**39.16%**	**35.16%**
χ^2C3D+similarity threshold ≈ 1 (Baseline)	60.24%	56.67%	30.56%	26.51%	17.03%
χ^2C3D+similarity threshold = 0	51.81%	42.50%	33.33%	**31.93%**	18.13%
χ^2C3D+deterministic	45.78%	47.50%	29.17%	26.51%	19.23%
χ^2C3D+deterministic+step	85.54%	65.00%	36.11%	28.92%	28.57%
χ^2C3D+deterministic+pchip	**89.16%**	**65.83%**	**40.28%**	30.72%	**30.22%**
χ^2C3D+probabilistic	45.78%	47.50%	29.86%	26.51%	19.78%
χ^2C3D+probabilistic+step	85.54%	65.00%	37.50%	28.92%	28.57%
χ^2C3D+probabilistic+pchip	**89.16%**	**65.83%**	37.50%	28.92%	28.57%

Πίνακας 7.9: Συγκεντρωτικά αποτελέσματα συγκρίσεων για το πρόβλημα της αναγνώρισης δράσεων στην ταινία DEP για 2, 4, 6, 8 και 10 κλάσεις. Ο πυρήνας linC3D πετυχαίνει ποσοστά επιτυχίας 5-10% καλύτερα σχεδόν σε όλες τις περιπτώσεις. Μεταξύ των μεθόδων των πιθανοτικών και των ντετερμινιστικών ετικετών παρατηρούμε μικρή υπεροχή της μεθόδου των πιθανοτικών ετικετών ιδιαίτερα για τον γραμμικό πυρήνα της τάξης του 2%. Όσον αφορά τις συναρτήσεις συμμετοχής, παρατηρούμε και πάλι την υπεροχή της pchip ιδιαίτερα όσο αυξάνεται το πλήθος των κλάσεων.

number of classes	**2**	**4**	**6**	**8**	**10**
linC3D+similarity threshold ≈ 1 (Baseline)	0.00%	0.00%	14.29%	12.50%	11.48%
linC3D+similarity threshold = 0	59.04%	39.34%	28.57%	17.26%	12.57%
linC3D+deterministic	59.04%	40.16%	16.33%	13.10%	6.56%
linC3D+deterministic+step	59.04%	40.16%	25.17%	19.64%	16.39%
linC3D+determinstc+pchip	59.04%	40.16%	20.41%	17.86%	16.94%
linC3D+probabilistic	59.04%	40.16%	24.49%	16.07%	7.10%
linC3D+probabilistic+step	**66.27%**	40.16%	**33.33%**	**27.98%**	15.85%
linC3D+probabilistic+pchip	**66.27%**	40.16%	23.13%	17.86%	**17.49%**
χ^2C3D+similarity threshold ≈ 1(Baseline)	0.00%	0.00%	15.65%	13.69%	12.57%
χ^2C3D+similarity threshold = 0	60.24%	36.89%	21.09%	17.26%	15.85%
χ^2C3D+deterministic	59.04%	40.16%	25.85%	19.05%	**22.95%**
χ^2C3D+deterministic+step	63.86%	40.16%	22.45%	22.62%	18.58%
χ^2C3D+deterministic+pchip	63.86%	40.16%	20.41%	20.24%	21.31%
χ^2C3D+probabilistic	59.04%	40.16%	28.57%	16.67%	20.77%
χ^2C3D+probabilistic+step	**95.18%**	40.16%	**35.37%**	**25.00%**	19.67%
χ^2C3D+probabilistic+pchip	**95.18%**	40.16%	22.45%	21.43%	21.31%

Πίνακας 7.10: Συγκεντρωτικά αποτελέσματα συγκρίσεων για το πρόβλημα της αναγνώρισης δράσεων στην ταινία GLA για 2, 4, 6, 8 και 10 κλάσεις. Εδώ ο πυρήνας χ^2C3D υπερέχει σημαντικά στην περίπτωση των 2 κλάσεων όταν συνδυάζεται με πιθανοτικές ετικέτες (95.18% για τον χ^2C3D, 66.27% για τον linC3D). Στη συνέχεια για μεγαλύτερο πλήθος κλάσεων η επίδοση του δεν διαφέρει σημαντικά από αυτήν του linC3D. Μεταξύ των μεθόδων των πιθανοτικών και των ντετερμινιστικών ετικετών παρατηρούμε σταθερή υπεροχή των πιθανοτικών με ελάχιστες εξαιρέσεις. Όσον αφορά τις συναρτήσεις συμμετοχής, παρατηρούμε μία υπεροχή της step ,όπως ακριβώς και στο πρόβλημα αναγνώρισης προσώπου, που φτάνει μέχρι και 13% διαφορά (22.45% χ^2C3D+probabilistic+pchip - 35.37% χ^2C3D+probabilistic+step), η διαφορά αυτή όμως εξαλείφεται όσο αυξάνεται το πλήθος των κλάσεων.

number of classes	2	4	6	8	10
linC3D+similarity threshold ≈ 1 (Baseline)	39.34%	27.46%	22.99%	20.83%	7.51%
linC3D+similarity threshold=0	**58.03%**	**38.44%**	**24.33%**	**24.65%**	16.29%
linC3D+deterministic	47.87%	31.58%	19.16%	17.36%	**19.01%**
linC3D+deterministic+step	46.89%	30.43%	15.13%	15.80%	14.54%
linC3D+determinstc+pchip	48.85%	30.66%	15.71%	15.63%	16.13%
linC3D+probabilistic	42.95%	28.83%	17.24%	16.49%	17.89%
linC3D+probabilistic+step	43.28%	30.21%	16.67%	17.71%	15.65%
linC3D+probabilistic+pchip	43.61%	30.43%	17.43%	16.67%	17.57%
χ^2C3D+similarity threshold ≈ 1 (Baseline)	39.67%	21.97%	20.88%	19.27%	7.03%
χ^2C3D+similarity threshold = 0	**54.10%**	**34.10%**	**23.95%**	**22.05%**	12.94%
χ^2C3D+deterministic	44.92%	26.54%	19.73%	15.80%	15.34%
χ^2C3D+deterministic+step	51.48%	31.35%	19.35%	16.84%	15.50%
χ^2C3D+deterministic+pchip	51.80%	32.04%	18.58%	17.88%	**16.45%**
χ^2C3D+probabilistic	44.59%	26.32%	17.62%	14.58%	12.94%
χ^2C3D+probabilistic+step	47.21%	29.75%	22.41%	18.75%	15.34%
χ^2C3D+probabilistic+pchip	46.56%	29.29%	20.88%	19.27%	16.13%

Πίνακας 7.11: Συγκεντρωτικά αποτελέσματα συγκρίσεων για το πρόβλημα της αναγνώρισης δράσεων στην ταινία GWW για 2, 4, 6, 8 και 10 κλάσεις. Εδώ παρ΄ όλο που υπάρχει σχετικά καλή ευθυγράμμιση μεταξύ κειμένου και βίντεο και πλούσια πληροφορία στο σενάριο, δεν βελτιώνουμε τα αποτελέσματα με τις μεθόδους που προτείνουμε. Αντίθετα, λαμβάνοντας την ετικέτα με τη μέγιστη τιμή ομοιότητας, ανεξαρτήτως του πόσο μικρή είναι η τιμή της και σχηματίζοντας τα bags με όλα τα επικαλυπτόμενα tracks, πετυχαίνουμε τα μεγαλύτερα ποσοστά επιτυχίας. Βέβαια, η ταινία αυτή έχει κάποιες ιδιαιτερότητες. Συγκεκριμένα, έχει πολύ μεγαλύτερη διάρκεια από τις υπόλοιπες, συνεπώς έχει πολλά περισσότερα δείγματα προς αναγνώριση, αλλά και πολλές περισσότερες ετικέτες παρεχόμενες από το κείμενο. Ακόμα, η ανάλυση της είναι χαμηλή πράγμα που δυσχεραίνει την ποιότητα της περιγραφής από το C3D. Αυτά τα 2 καθιστούν την ομαδοποίηση ιδιαίτερα δύσκολη διαδικασία. Επομένως, πιθανόν χρησιμοποιώντας τις δικές μας μεθόδους, 'υπερμοντελοποιούμε' το πρόβλημα και αποτυγχάνουμε να επιλύσουμε τις αμφισημίες του. Όσον αφορά τους πυρήνες, παρατηρούμε κατά κύριο λόγο την υπεροχή του γραμμικού linC3D.

number of classes	**2**	**4**	**6**	**8**	**10**
linC3D+similarity threshold ≈ 1 (Baseline)	0.00%	20.00%	16.51%	14.75%	13.64%
linC3D+similarity threshold =0	47.46%	25.56%	18.35%	17.21%	15.91%
linC3D+deterministic	**93.22%**	42.22%	19.27%	13.93%	12.88%
linC3D+deterministic+step	74.58%	48.89%	26.61%	28.69%	23.48%
linC3D+determinstc+pchip	52.54%	50.00%	**30.28%**	30.33%	**31.06%**
linC3D+probabilistic	**93.22%**	42.22%	19.27%	13.93%	12.12%
linC3D+probabilistic+step	74.58%	**51.11%**	29.36%	22.13%	20.45%
linC3D+probabilistic+pchip	52.54%	50.00%	**30.28%**	**31.15%**	28.03%
χ^2C3D+similarity threshold ≈ 1 (Baseline)	0.00%	20.00%	16.51%	13.93%	12.88%
χ^2C3D+similarity threshold = 0	52.54%	26.67%	20.18%	19.67%	18.94%
χ^2C3D+deterministic	**93.22%**	33.33%	22.94%	17.21%	17.42%
χ^2C3D+deterministic+step	**93.22%**	42.22%	32.11%	31.15%	**28.79%**
χ^2C3D+deterministic+pchip	86.44%	46.67%	32.11%	**32.79%**	28.03%
χ^2C3D+probabilistic	**93.22%**	33.33%	22.94%	16.39%	15.91%
χ^2C3D+probabilistic+step	**93.22%**	42.22%	**33.94%**	24.59%	21.97%
χ^2C3D+probabilistic+pchip	86.44%	**50.00%**	32.11%	24.59%	21.21%

Πίνακας 7.12: Συγκεντρωτικά αποτελέσματα συγκρίσεων για το πρόβλημα της αναγνώρισης δράσεων στην ταινία LOR για 2, 4, 6, 8 και 10 κλάσεις. Εδώ παρατηρούμε ότι και οι 2 συναρτήσεις πυρήνων επιτυγχάνουν περίπου το ίδιο καλά αποτελέσματα ενώ επίσης μεταξύ πιθανοτικών και ντετερμινιστικών ετικετών δεν υπάρχει μεγάλη διαφοροποίηση (οι διαφορές είναι της τάξης του 1%-3%). Όσον αφορά τις συναρτήσεις συμμετοχής, παρατηρούμε ότι όσο αυξάνεται το πλήθος των κλάσεων, η συνάρτηση που προκύπτει από την pchip ξεπερνά την συνάρτηση step σε όλες τις περιπτώσεις.

number of classes	**2**	**4**	**6**	**8**	**10**
linC3D+similarity threshold=1 (Baseline)	26.94%	24.81%	19.05%	16.83%	12.72%
linC3D+similarity threshold=0	56.10%	36.38%	24.38%	17.58%	12.22%
linC3D+deterministic	62.85%	37.08%	16.07%	14.07%	12.03%
linC3D+deterministic+step	66.75%	43.47%	23.69%	21.53%	17.76%
linC3D+determinstc+pchip	62.40%	42.67%	24.05%	22.66%	19.83%
linC3D+probabilistic	61.61%	37.85%	20.47%	14.89%	12.99%
linC3D+probabilistic+step	**67.35%**	**44.66%**	**25.79%**	22.21%	17.79%
linC3D+probabilistic+pchip	61.10%	43.41%	24.60%	**23.11%**	**20.72%**
χ^2C3D+similarity threshold=1 (Baseline)	26.99%	23.90%	20.20%	18.87%	14.40%
χ^2C3D+similarity threshold=0	58.59%	36.22%	25.66%	20.43%	14.14%
χ^2C3D+deterministic	62.90%	37.25%	21.32%	16.67%	15.09%
χ^2C3D+deterministic+step	68.03%	41.84%	26.60%	22.96%	20.17%
χ^2C3D+deterministic+pchip	67.25%	42.47%	26.69%	**23.14%**	**21.13%**
χ^2C3D+probabilistic	62.84%	39.28%	24.00%	16.75%	15.90%
χ^2C3D+probabilistic+step	**71.84%**	41.74%	**28.65%**	22.29%	19.56%
χ^2C3D+probabilistic+pchip	70.74%	**42.65%**	25.88%	21.69%	19.77%

Πίνακας 7.13: Συγκεντρωτικά αποτελέσματα συγκρίσεων για το πρόβλημα της αναγνώρισης δράσεων κατά μέσο όρο για όλες τις ταινίες, για 2, 4, 6, 8 και 10 κλάσεις.Παρατηρούμε εδώ ότι ο χ^2C3D υπερισχύει πράγμα που οφείλεται κυρίως στα υψηλά ποσοστά συγκεκριμένων ταινιών. Επίσης, βλέπουμε τη μεγάλη σημασία της ενσωμάτωσης της πληροφορίας της χρονική επικάλυψης, είτε αυτή γίνεται μέσω της συνάρτησης step, είτε μέσω της pchip. Όσον αφορά τις πιθανοτικές ετικέτες, τα αποτελέσματα είναι καλύτερα με μικρές εξαιρέσεις.

Κεφάλαιο 8

Συμπεράσματα

8.1 Ανακεφαλαίωση - Συμβολή της Διπλωματικής Εργασίας

Στη διπλωματική εργασία αυτή αντιμετωπίσαμε το πρόβλημα της Αυτόματης Κατανόησης Βίντεο χρησιμοποιώντας τις δυνατότητες Ασθενούς Επίβλεψης που μας παρέχει ένα συνοδευτικό κείμενο. Εστιάσαμε την προσοχή μας σε δύο σημεία - κλειδιά του γενικότερου προβλήματος της Κατανόησης και συγκεκριμένα στον Εντοπισμό και την Αναγνώριση των Χαρακτήρων και των Δράσεων που εκτελούν.

Προτείναμε μία νέα και γενική μέθοδο εξαγωγής ετικετών, αυτήν της Σημασιολογικής Ομοιότητας. Με αυτήν είναι δυνατή η εξαγωγή ετικετών από το κείμενο για οποιοδήποτε πρόβλημα αναγνώρισης. Οι μέχρι τώρα μέθοδοι είτε είχαν περιορισμένη εφαρμογή, είτε απαιτούσαν μεγάλο βαθμό ανθρώπινης παρέμβασης προκειμένου να εκπαιδεύσουν ταξινομητές κειμένου οι οποίοι έβρισκαν τις ετικέτες. Αντίθετα, η δικιά μας μέθοδος είναι μη επιβλεπόμενη και χρησιμοποιεί μόνο ένα ήδη εκπαιδευμένο σύστημα σημασιολογικής ομοιότητας και ένα κατώφλι απόρριψης χαμηλών τιμών ομοιότητας. Το πρόβλημα της είναι ότι τις περισσότερες φορές, σε προβλήματα κατανόησης με σύνθετη σημασιολογία όπως η αναγνώριση δράσεων, δεν επιστρέφει κάποια ετικέτα με μεγάλη βεβαιότητα, όπως για παράδειγμα πετυχαίνουν οι μέθοδοι των ταξινομητών κειμένου. Για αυτό το λόγο όσο αυξάνονται οι κλάσεις οι αμφισημίες είναι όλο και πιο δύσκολο να επιλυθούν. Για να τις αντιμετωπίσουμε μοντελοποιήσαμε το πρόβλημα ως σενάριο Μάθησης Πιθανοτικών Ετικετών. Όπως είδαμε από το κεφάλαιο 7, κυρίως η εξαγωγή ετικετών μέσω σημασιολογικής ομοιότητας αλλά και το νέο σενάριο μάθησης, βελτίωσαν σχεδόν σε όλες τις περιπτώσεις τα αποτελέσματα σε σύγκριση με τη χρήση ετικετών με ίδια σημασιολογία, όπως είναι αυτές που μας επιστρέφουν οι τεχνικές των ταξινομητών κειμένου.

Ακόμα, εισαγάγαμε την έννοια των Ασαφών Συνόλων Πολλαπλών Παραδειγμάτων, ορμώμενοι από τις διαφορετικές χρονικές επικαλύψεις οπτικών και γλωσσικών αντικειμένων που προκύπτουν από την ευθυγράμμιση των 2 τροπικοτήτων. Οι μοντελοποιήσεις που προτείναμε για αυτό το φορμαλισμό καθώς και για το συνδυασμό των Πιθανοτικών Ετικετών με τα Σύνολα Πολλαπλών Παραδειγμάτων, αν και σχετικά απλές, είναι οι πρώτες που έχουν γίνει σύμφωνα με όσα ξέρουμε και βελτίωσαν την Κατανόηση του Βίντεο σε πολλές περιπτώσεις.

Επίσης, όπως είδαμε η χρήση επιπλέον πληροφορίας υπαινισσόμενης από τις επαναλήψεις πανομοιότυπων συνόλων με ίδιες ετικέτες, από την αυτόματη κατανόηση κάποιας άλλης πτυχής του βίντεο ή από έναν προεκπαιδευμένο ταξινομήτη συνεισέφερε στην καλύτερη κατανόηση. Από αυτό συμπεραίνουμε ότι όσο ισχυρότερους υπαινιγμούς συμπεριλαμβάνουμε, τόσο πιο αποτελεσματική γίνεται η μέθοδος.

Τέλος, δείξαμε την υπεροχή των αναπαραστάσεων βαθιάς μηχανικής μάθησης σε σύγκριση με τις παραδοσιακές (hand-crafted), ακόμα και χωρίς την επανεκπαίδευση τους.

8.2 Μελλοντική Έρευνα

Το μεγάλο πλήθος των προβλημάτων που αντιμετωπίσαμε κατά την εκπόνηση της διπλωματικής αυτής έχει δώσει το κίνητρο για μελλοντική έρευνα σε ποικίλες κατευθύνσεις.

Συγκεκριμένα, όσον αφορά τον εντοπισμό και την αναπαράσταση των οπτικών δεδομένων και την εξαγωγή των γλωσσικών ετικετών:

- **Για τα πρόσωπα:** Μπορούν να εφαρμοστούν πιο σύγχρονες τεχνικές εντοπισμού τους στο βίντεο (και παρακολούθησής τους) προκειμένου να πετύχουμε πληρέστερη καταγραφή (όπως αυτές που υλοποιούνται από βαθιά νευρωνικά δίκτυα). Ακόμα, εφόσον έχει φανεί από το [47] ότι η Επίλυση Συναναφορών είναι χρήσιμη, καλό θα ήταν να ενσωματωθεί στο συνολικό σύστημα.

- **Για τις Δράσεις:** Θεωρούμε απαραίτητη την αυτόματη ανίχνευση των δράσεων προκειμένου να μειώσουμε αφενός την ανθρώπινη παρέμβαση και αφετέρου να πάρουμε τα πιο salient αποσπάσματα. Επίσης, δεδομένου του ότι διαρκώς εισάγονται νέες βελτιωμένες αναπαραστάσεις δράσεων στη διεθνή βιβλιογραφία, το σύστημα μπορεί να επωφεληθεί από αυτές. Ακόμα, σχεδιάζουμε την εξαγωγή των ετικετών με πιο προηγμένους τρόπους, όπως είναι η ομαδοποίηση ρημάτων του κειμένου (μηδενίζει την ανθρώπινη παρέμβαση του καθορισμού του συνόλου ετικετών) και η χρήση σύγχρονων γλωσσικών αναπαραστάσεων, όπως το word2vec και πιο εξελιγμένων συστημάτων υπολογισμού σημασιολογικής ομοιότητας.

Επίσης, μία περισσότερη εξελιγμένη τεχνική συσχέτισης των ανθρώπων με τις δράσεις τους (για παράδειγμα εντοπίζοντας το ανθρώπινο σώμα και συσχετίζοντας το με το πρόσωπο του) θεωρούμε ότι θα βοηθήσει σε μεγάλο βαθμό την κατανόηση μέσα από τη ροή πληροφορίας από το ένα σύστημα αναγνώρισης στο άλλο (βλέπε εξίσωση (4.15)).

Ακόμα, μία πτυχή του προβλήματος που δεν εξετάστηκε καθόλου εδώ είναι η ύπαρξη των δεδομένων παρασκηνίου. Αν δεν ληφθεί μέριμνα για αυτά, η μάθηση των δεδομένων προσκηνίου μπορεί να επηρεαστεί καθώς ο ταξινομητής εκπαιδεύεται με περισσότερα λάθος δείγματα. Αντίθετα, λαμβάνοντας μέριμνα, αναμένουμε τα συνολικά αποτελέσματα να βελτιωθούν καθώς μέσα από τα δεδομένα παρασκηνίου προσφέρεται ένα μεγάλο πλήθος αρνητικών παραδειγμάτων. Για αυτό το λόγο θεωρούμε απαραίτητη στο μέλλον τη γενίκευση του προβλήματος ως Μάθηση Ανοιχτού Συνόλου - Open Set Recognition.

Επίσης, δεδομένης της πολυπλοκότητας των δεδομένων της βάσης COGNIMUSE είναι απαραίτητη η εφαρμογή των μοντελοποιήσεων που προτείναμε σε πιο απλά, πιθανώς ελεγχόμενα σύνολα δεδομένων όπου θα είναι πιο εύκολο να ερμηνευτεί η επίδοση κάθε μεθόδου. Άλλωστε, όπως έχουμε αναφέρει πολλές φορές, τα μοντέλα είναι γενικά και μπορούν να εφαρμοστούν σε οποιοδήποτε ζεύγος παράλληλων ροών πληροφορίας. Για αυτό το λόγο, θα ήταν ενδιαφέρον να εξεταστεί η αποδοτικότητα των μεθόδων και σε άλλες πτυχές της Αυτόματης Κατανόησης Βίντεο, όπως είναι η αναγνώριση σκηνών ή αντικειμένων ή και γενικότερα σε άλλες περιπτώσεις ζεύγους ροών πληροφορίας, όπως είναι η κατανόηση ακουστικών γεγονότων από κείμενο. Επίσης, δεδομένου του ότι φάνηκε ότι η χρήση επιπλέον υπαινισσόμενης πληροφορίας είναι ευεργετική, κρίνουμε σκόπιμο να εξεταστούν στο μέλλον και άλλοι υπαινιγμοί όπως αυτοί που προκύπτουν κατά την αναγνώριση πολλών ειδών αντικειμένων μαζί ή από πληροφορία που περιέχεται στο ακουστικό κανάλι.

Ακόμα, δεδομένου του ότι η μοντελοποίηση του φορμαλισμού έγινε με ad hoc τρόπους, όπως αναφέραμε στο κεφάλαιο 4, θα ήταν ενδιαφέρον να εξετάσουμε άλλους τρόπους μοντελοποίησης προερχόμενους για παράδειγμα από μία πιθανοτική ερμηνεία. Αυτή η μελέτη δεν είναι απαραίτητο να αφορά την Κατανόηση Βίντεο. Αντίθετα, χρειάζεται να είναι γενική έτσι ώστε να μοντελοποίησει με τρόπο περισσότερο κομψό τις νέες μορφές Μάθησης.

Τέλος, το σημαντικότερο ίσως κίνητρο για μελλοντική έρευνα είναι η υλοποίηση της από κοινού Κατανόησης Βίντεο και Κειμένου όπως περιγράφεται από την (4.17) και ίσως η επέκταση της ώστε να συμπεριλαμβάνει ακουστικά γεγονότα, καθώς και επιπλέον υπαινισσόμενες πληροφορίες. Συγκεκριμένα, αξιοποιώντας όλες τις τροπικότητες ισότιμα στοχεύουμε σε μία πολυτροπική κατανόηση η οποία θα είναι αρκετά πληρέστερη και πληροφοριακή για τη σημασιολογία του οπτικοακουστικού περιεχομένου ενός βίντεο.

Παράρτημα Α΄

Κυρτός Προγραμματισμός

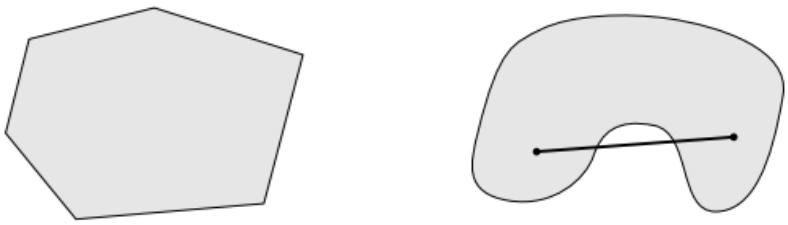
Α΄.1 Κυρτά Σύνολα και Κυρτές Συναρτήσεις

Κυρτά Σύνολα

Ορισμός 1.1. *Θα λέμε ότι ένα* ***σύνολο*** $\mathcal{C}$ ***είναι κυρτό*** *αν* $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ $\theta \in \mathbb{R}$ *με* $0 \leq \theta \leq 1$ *ισχύει ότι:*

$$\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y} \in \mathcal{C} \tag{Α΄.1}$$

Αυτό μπορεί να ερμηνευτεί λέγοντας ότι αν ενώσουμε οποιαδήποτε δύο σημεία ενός κυρτού συνόλου με ένα ευθύγραμμο τμήμα, τότε ολόκληρο το ευθύγραμμο τμήμα θα βρίσκεται και το ίδιο μέσα στο σύνολο, όπως φαίνεται και από το σχήμα Α΄.1. Ένα σημείο που ορίζεται από την εξίσωση Α΄.1 ονομάζεται κυρτός συνδυασμός των σημείων $\boldsymbol{x}, \boldsymbol{y}$.



(α΄) (β΄)

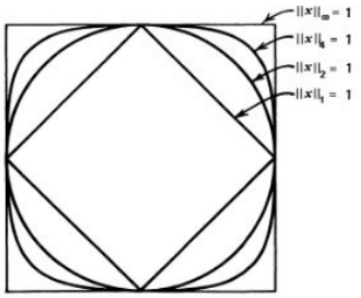
Σχήμα Α΄.1: Παράδειγμα κυρτού (α΄) και μη κυρτού (β΄) συνόλου

Κάποια χαρακτηριστικά παραδείγματα είναι τα εξής:

- Όλος ο χώρος $\mathbb{R}^n$
- Το σύνολα που ορίζονται από εξισώσεις $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^n | \ \|\boldsymbol{x}\|_p \leq r\}$ (**norm balls**) με $p > 1$. Συγκεκριμένα, αν $\|\boldsymbol{x}\| \leq r, \|\boldsymbol{y}\| \leq r$, $0 \leq \theta \leq 1$, τότε:

$$\|\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}\| \leq \|\theta \boldsymbol{x}\| + \|(1-\theta)\boldsymbol{y}\| \leq \theta \cdot r + (1-\theta) \cdot r = r$$

Για την 1η ανίσωση χρησιμοποιήσαμε την τριγωνική ανισότητα, ενώ για την δεύτερη το επιχείρημα ότι τα $\boldsymbol{x}, \boldsymbol{y}$ ανήκουν στο $\mathcal{C}$ και ότι $0 \leq \theta \leq 1$.



Σχήμα Α΄.2: Κάποιες ενδεικτικές μπάλες νορμών

- Οι αφινικοί υπόχωροι και τα πολύεδρα , δηλαδή τα σύνολα που ορίζονται από τις εξισώσεις $\mathcal{C} = \{\ \boldsymbol{x} \in \mathbb{R}^n |\ \boldsymbol{Ax} = \boldsymbol{b}\}$, $\mathcal{C} = \{\ \boldsymbol{x} \in \mathbb{R}^n |\ \boldsymbol{Ax} \leq \boldsymbol{b}\}$ αντίστοιχα. Σημειώνουμε ότι η εξίσωση του πολύεδρου σημαίνει ότι όλα τα στοιχεία του $\boldsymbol{Ax}$ είναι μικρότερα ή ίσα από τα αντίστοιχα στοιχεία του $\boldsymbol{b}$. Οι αποδείξεις είναι εξίσου απλές με αυτές των νορμών: Για $\boldsymbol{Ax} \leq \boldsymbol{b}$, $\boldsymbol{Ay} \leq \boldsymbol{b}$, $0 \leq \theta \leq 1$

$$\boldsymbol{A} \cdot (\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}) = \theta \boldsymbol{Ax} + (1-\theta)\boldsymbol{Ay} \leq \theta \boldsymbol{b} + (1-\theta)\boldsymbol{b} = \boldsymbol{b}.$$

- Τομή κυρτών συνόλων. Και πάλι η απόδειξη προκύπτει άμεσα από τους ορισμούς. Έστω $\mathcal{C}_1, \mathcal{C}_2, ...\mathcal{C}_k$ κυρτά σύνολα. Τότε η τομή τους ορίζεται ως εξής:

$$\bigcap_{i=1}^{k} \mathcal{C}_i = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} \in \mathcal{C}_i\ \forall i = 1, 2..., k\}$$

  Έστω, $\boldsymbol{x}, \boldsymbol{y} \in \bigcap_{i=1}^{k} \mathcal{C}_i$ και $0 \leq \theta \leq 1$. Άρα, $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}_i$, $\forall \mathcal{C}_i$ από τον ορισμό της τομής. Επομένως, ισχύει ότι: $\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y} \in \mathcal{C}_i$, $\forall \mathcal{C}_i$ αφού τα $\mathcal{C}_i$ είναι κυρτά σύνολα. Άρα, και πάλι από τον ορισμό της τομής $\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y} \in \bigcap_{i=1}^{k} \mathcal{C}_i$.

  Η ένωση όμως εν γένει δεν είναι κυρτό σύνολο, καθώς ο ορισμός της ένωσης μας εξασφαλίζει την συμμετοχή του $\boldsymbol{x}$ σε ένα τουλάχιστον $\mathcal{C}_i$ και του $\boldsymbol{y}$ σε ένα άλλο $\mathcal{C}_j$, χωρίς αυτό να σημαίνει απαραίτητα ότι $\mathcal{C}_j = \mathcal{C}_j$, άρα δεν μπορούμε να ισχυριστούμε ότι το $\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}$ θα ανήκει σε κάποια από τα 2 σύνολα.

- Το σύνολο $\mathbb{S}_+^n$ των θετικά ημιορισμένων πινάκων. Δηλαδή όλων των πινάκων $\boldsymbol{A}$ για τους οποίους ισχύει $\boldsymbol{A} = \boldsymbol{A}^T, \forall \boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x}^T \boldsymbol{Ax} \geq 0$ Έστω $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}_+^n$ και $0 \leq \theta \leq 1$. Τότε, $\forall \boldsymbol{x} \in \mathbb{R}^n$

$$\boldsymbol{x}^T(\theta \boldsymbol{A} + (1-\theta)\boldsymbol{B})\boldsymbol{x} = \theta \boldsymbol{x}^T \boldsymbol{Ax} + (1-\theta)\boldsymbol{x}^T \boldsymbol{Bx} \geq 0$$

Κυρτές Συναρτήσεις

Ορισμός 1.2. *Θα λέμε ότι μία* ***συνάρτηση*** $f : \mathbb{R}^n \to \mathbb{R}$ ***είναι κυρτή*** *όταν το πεδίο ορισμού της* $\boldsymbol{dom} f$ *είναι κυρτό σύνολο και ισχύει ότι* $\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{dom} f$, $\theta \in [0, 1]$*:*

$$f(\theta \boldsymbol{x} + (1 - \theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1 - \theta) f(\boldsymbol{y}) \tag{Α΄.2}$$

Διαισθητικά αυτό σημαίνει ότι αν ενώσουμε οποιαδήποτε δύο σημεία της γραφικής παράστασης της f με ένα ευθύγραμμο τμήμα, τότε ολόκληρο το ευθύγραμμο τμήμα θα βρίσκεται πάνω από το γράφημα της f, όπως φαίνεται και από το σχήμα Α΄.3.

Σχήμα Α΄.3: Παράδειγμα κυρτής συνάρτησης

Θα λέμε ακόμα ότι όταν η σχέση Α΄.2 ισχύει με αυστηρή ανίσωση για $\boldsymbol{x} \neq \boldsymbol{y}$, $\theta \in (0, 1)$ τότε η συνάρτηση είναι αυστηρά κυρτή. Επίσης, όταν η συνάρτηση $-f$ είναι κυρτή, τότε η συνάρτηση f είναι ***κοίλη***. Όμοια με πριν ορίζουμε την αυστηρά κοίλη συνάρτηση.

Παρακάτω παρουσιάζουμε τα δύο βασικά θεωρήματα που χρησιμοποιούνται για να αποδειχτεί ότι μία συνάρτηση είναι κυρτή

Θεώρημα 1.1. *Συνθήκη πρώτης τάξης για την κυρτότητα*

Έστω μία συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}$ *που είναι παραγωγίσιμη (δηλαδή υπάρχει το* $\nabla f(\boldsymbol{x})$ *σε κάθε σημείο του* $\boldsymbol{dom} f$*). Η* f *είναι κυρτή* ***αν και μόνο αν*** *το* $\boldsymbol{dom} f$ *είναι κυρτό σύνολο και* $\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{dom} f$ *ισχύει:*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) \tag{Α΄.3}$$

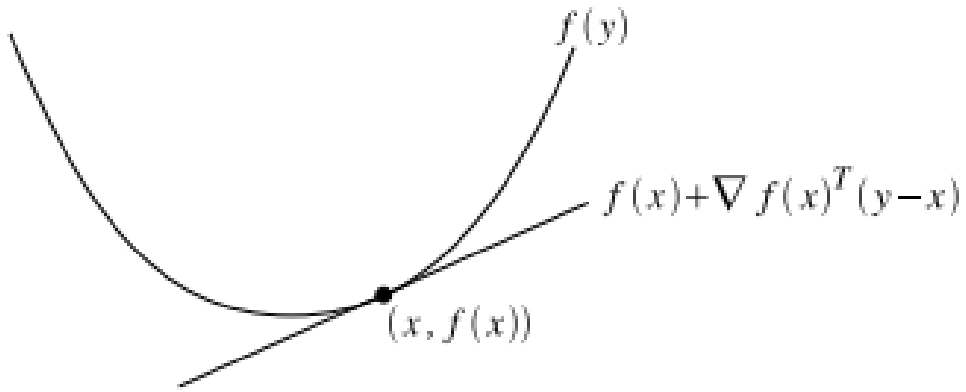
Η ερμηνεία του θεωρήματος αυτού είναι ότι η εφαπτομένη μιας κυρτής συνάρτησης σε ένα σημείο $(\boldsymbol{x}, f(\boldsymbol{x}))$ είναι πάντα κάτω από τη γραφική παράσταση της f (βλέπε σχήμα Α΄.4).

Θεώρημα 1.2. *Συνθήκη δεύτερης τάξης για την κυρτότητα*

Έστω μία συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}$ *που είναι διπλά παραγωγίσιμη (δηλαδή υπάρχει ο εσσιανός πίνακας* $\nabla^2 f(\boldsymbol{x})$ *σε κάθε σημείο του* $\boldsymbol{dom} f$*). Η* f *είναι κυρτή* ***αν και μόνο αν*** *το* $\boldsymbol{dom} f$ *είναι κυρτό σύνολο και* $\forall \boldsymbol{x} \in \boldsymbol{dom} f$ *ισχύει:*

$$\nabla^2 f(\boldsymbol{x}) \geq 0 \tag{Α΄.4}$$

Όπου η ανίσωση σημαίνει ότι ο πίνακας είναι θετικά ημιορισμένος



Σχήμα Α΄.4: Γεωμετρική ερμηνεία της συνθήκης πρώτης τάξης

Χρησιμοποιώντας τον ορισμό καθώς και τις συνθήκες 1ης και 2ης τάξης αποδεικνύουμε την κυρτότητα μερικών χρήσιμων συναρτήσεων. Για τα παρακάτω παραδείγματα θεωρούμε ότι το $\boldsymbol{dom}f$ είναι το $\mathbb{R}^n$ το οποίο όπως εξηγήσαμε παραπάνω είναι κυρτό σύνολο.

- Αφινικές συναρτήσεις. Έστω $f : \mathbb{R}^n \to \mathbb{R}$ με $f = \boldsymbol{w}^T\boldsymbol{x} + b$. Η f είναι διπλά παραγωγίσιμη με $\nabla^2 f(\boldsymbol{x}) = \mathbf{0}_{n\times n}$. Άρα, από συνθήκη δεύτερης τάξης η συνάρτηση είναι κυρτή και ταυτόχρονα κοίλη και μάλιστα αυτές οι συναρτήσεις είναι οι μόνες για τις οποίες γίνεται να ισχύουν και τα 2.

- Τετραγωνικές μορφές $f = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}^T\boldsymbol{x} + b$ με $\boldsymbol{A} \geq 0$. Και πάλι Η f είναι διπλά παραγωγίσιμη με $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}$ και άρα από συνθήκη δεύτερης τάξης είναι κυρτή.

- Όλες οι νόρμες. Έστω $f : \mathbb{R}^n \to \mathbb{R}$ μία νόρμα. Τότε:

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \overset{\substack{\text{Για κάθε νόρμα ισχύει}\\ \text{η τριγωνική ανισότητα}}}{\leq} f(\theta\boldsymbol{x}) + f((1-\theta)\boldsymbol{y}) \overset{\substack{\text{Για κάθε νόρμα ισχύει}\\ \text{η ομογένεια}}}{=}$$
$$= |\theta| f(\boldsymbol{x}) + |1-\theta| f(\boldsymbol{y}) \overset{\theta\in(0,1)}{=} \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y})$$

  Να σημειώσουμε εδώ ότι δεν μπορούμε να κάνουμε την απόδειξη μέσα από κάποια από τις 2 συνθήκες καθώς υπάρχουν νόρμες που δεν είναι παντού παραγωγίσιμες.

- Συγκεκριμένες πράξεις μεταξύ κυρτών συναρτήσεων όπως: άθροισμα κυρτών συναρτήσεων με μη αρνητικά βάρη ($f(\boldsymbol{x}) = \sum_{i=1}^{m} w_i f_i(\boldsymbol{x})$), η σύνθεση μιας κυρτής συνάρτησης με μία αφινική ($f(\boldsymbol{w}^T\boldsymbol{x} + b)$), η συνάρτηση που ισούται με το μέγιστο άλλων κυρτών συναρτήσεων στοιχείο προς στοιχείο ($f(\boldsymbol{x}) = max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_m(\boldsymbol{x})\}$) και άλλες. Αποδεικνύουμε εδώ ενδεικτικά την κυρτότητα του αθροίσματος συναρτήσεων

με μη αρνητικά βάρη:

$$\begin{aligned} f(\theta\boldsymbol{x}+(1-\theta)\boldsymbol{y}) &= \sum_{i=1}^{m} w_i f_i(\theta\boldsymbol{x}+(1-\theta)\boldsymbol{y}) \text{ (από κυρτότητα των } f_i \text{ και από μη αρνητικότητα των } w_i) \\ &\leq \sum_{i=1}^{m} w_i(\theta f_i(\boldsymbol{x})+(1-\theta)f_i(\boldsymbol{y})) \\ &= \theta\sum_{i=1}^{m} w_i f_i(\boldsymbol{x}) + (1-\theta)\sum_{i=1}^{m} w_i f_i(\boldsymbol{y}) \end{aligned}$$

Τέλος, αξίζει να αναφερθεί η έννοια των sublevel sets, που προκύπτουν από κυρτές συναρτήσεις. Συγκεκριμένα,αν $f:\mathbb{R}^n\to\mathbb{R}$ και $a\in\mathbb{R}$ το a-sublevel set ορίζεται ως:

$$\{\boldsymbol{x}\in\boldsymbol{dom}f : f(\boldsymbol{x})\leq a\}$$

Τα σύνολα αυτά είναι κυρτά και χρησιμοποιούνται συχνά ως περιορισμοί σε προβλήματα κυρτού προγραμματισμού. Η απόδειξη ότι είναι κυρτά έχει ως εξής: Αν $\boldsymbol{x}\in\boldsymbol{dom}f, \boldsymbol{y}\in\boldsymbol{dom}f$ με $f(\boldsymbol{x})\leq a, f(\boldsymbol{y})\leq a$

$$f(\theta\boldsymbol{x}+(1-\theta)\boldsymbol{y})\leq\theta f(\boldsymbol{x})+(1-\theta)f(\boldsymbol{y})\leq\theta a+(1-\theta)a=a$$

Άρα και το $\theta\boldsymbol{x}+(1-\theta)\boldsymbol{y}$ ανήκει στο a-sublevel set.

Α΄.2 Προβλήματα Κυρτής Βελτιστοποίησης

Ορισμός 1.3. *Γενικά ένα πρόβλημα κυρτής βελτιστοποίησης είναι ένα πρόβλημα της μορφής:*

$$\begin{aligned} &\textit{minimize } f(\boldsymbol{x}) \\ &\textit{subject to } \boldsymbol{x}\in\mathcal{C} \end{aligned} \tag{Α΄.5}$$

όταν η f είναι κυρτή συνάρτηση και το σύνολο $\mathcal{C}$ είναι επίσης κυρτό. Η ελαχιστοποίηση γίνεται ως προς τη μεταβλητή $\boldsymbol{x}$.

Συνήθως, τα προβλήματα εκφράζονται με τη μορφή:

$$\begin{aligned} &min_{\boldsymbol{x}}\ f(\boldsymbol{x}) \\ &\text{s.t. } g_i(\boldsymbol{x})\leq 0 \quad i=1,2,...,m \\ &\qquad h_i(\boldsymbol{x})=0 \quad i=1,2,...,p \end{aligned} \tag{Α΄.6}$$

Όπου f, g κυρτές συναρτήσεις (άρα το πρώτο σύνολο περιορισμών είναι κυρτό ως τομή sublevel sets) και h αφινικές συναρτήσεις (άρα οι περιορισμοί του δεύτερου συνόλου δημιουργούν αφινικούς υπόχωρους και άρα κυρτά σύνολα).

Ολικά βέλτιστα σε κυρτά προβλήματα

Το βασικό στοιχείο που κάνει τα προβλήματα κυρτού προγραμματισμού τόσο ελκυστικά είναι το γεγονός ότι **κάθε τοπικό ελάχιστο είναι ταυτόχρονα και ολικό**. Χάρις σε αυτήν την ιδιότητα, οι αλγόριθμοι εύρεσης του ολικού ελαχίστου δεν υποφέρουν από προβλήματα κακών αρχικοποιήσεων και δεν υπάρχουν κίνδυνοι σύγκλισης σε τοπικό ελάχιστο. Έτσι, με τον τερματισμό ενός τέτοιου αλγορίθμου (που λειτουργεί συνήθως επαναληπτικά) είμαστε σίγουροι ότι έχουμε βρει την βέλτιστη τιμή του προβλήματος.

Απόδειξη. Η απόδειξη θα γίνει με την εις άτοπο απαγωγή.

Έστω $\boldsymbol{x_0}$ τοπικό ελάχιστο, αλλά όχι ολικό. Τότε θα υπάρχει $\boldsymbol{y_0}$ στην εφικτή περιοχή και έξω από την γειτονιά όπου το $\boldsymbol{x_0}$ είναι ελάχιστο, τέτοιο ώστε $f(\boldsymbol{y_0}) < f(\boldsymbol{x_0})$. Για να βρίσκεται το $\boldsymbol{y_0}$ έξω από τη γειτονιά του $\boldsymbol{x_0}$ θα πρέπει να ισχύει $\|\boldsymbol{x_0} - \boldsymbol{y_0}\|_2 > R$. Από τον ορισμό της κυρτότητας συνάρτησης παίρνουμε:

$$f(\theta \boldsymbol{y_0} + (1-\theta)\boldsymbol{x_0}) \leq \theta f(\boldsymbol{y_0}) + (1-\theta) f(\boldsymbol{x_0}) \quad (\text{και επειδή έχουμε ότι } \theta \geq 0 \ , \ f(\boldsymbol{y_0}) < f(\boldsymbol{x_0}))$$
$$f(\theta \boldsymbol{y_0} + (1-\theta)\boldsymbol{x_0}) < \theta f(\boldsymbol{x_0}) + (1-\theta) f(\boldsymbol{x_0})$$
$$f(\theta \boldsymbol{y_0} + (1-\theta)\boldsymbol{x_0}) < f(\boldsymbol{x_0})$$

Τώρα με κατάλληλη επιλογή του θ μπορούμε να πάρουμε ένα $\boldsymbol{z} = \theta \boldsymbol{y_0} + (1-\theta)\boldsymbol{x_0}$ στο εσωτερικό της γειτονιάς του $\boldsymbol{x_0}$. Συγκεκριμένα, αν για παράδειγμα επιλέξουμε $\theta = \frac{R}{2\|\boldsymbol{x_0} - \boldsymbol{y_0}\|_2}$, τότε ,αφού $R > 0$ και $\|\boldsymbol{x_0} - \boldsymbol{y_0}\|_2 > R$,έχουμε ότι $\theta \in [0, 1]$, άρα η τιμή αυτή είναι αποδεκτή, ενώ ακόμα:

$$\begin{aligned} \|\boldsymbol{x_0} - \boldsymbol{z}\|_2 &= \|\boldsymbol{x_0} - (\frac{R}{2\|\boldsymbol{y_0} - \boldsymbol{y_0}\|_2}\boldsymbol{y_0} + (1 - \frac{R}{2\|\boldsymbol{x_0} - \boldsymbol{y_0}\|_2})\boldsymbol{x_0})\|_2 \\ &= \|\frac{R}{2\|\boldsymbol{x_0} - \boldsymbol{y_0}\|_2}(\boldsymbol{x_0} - \boldsymbol{y_0})\|_2 \\ &= R/2 < R \end{aligned}$$

Άρα βρήκαμε ένα $\boldsymbol{z}$ με $\|\boldsymbol{x_0} - \boldsymbol{z}\|_2 \leq R$ για το οποίο ισχύει ότι $f(\boldsymbol{z}) < f(\boldsymbol{x_0})$. **Άτοπο**!. Άρα, κάθε τοπικό ελάχιστο είναι και ολικό. □

Bibliography

[1] F. R. Bach and Z. Harchaoui. ≪Diffrac: a discriminative and flexible framework for clustering≫. In: *Advances in Neural Information Processing Systems*. 2008, pp. 49–56.

[2] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. ≪Names and faces in the news≫. In: *Computer Vision and Pattern Recognition, 2004. CVPR 2004. IEEE Conference on*. Vol. 2. IEEE. 2004, pp. II–II.

[3] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. ≪Who's in the picture≫. In: *Advances in Neural Information Processing Systems*. 2005, pp. 137–144.

[4] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. ≪Finding actors and actions in movies≫. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 2280–2287.

[5] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. ≪Weakly supervised action labeling in videos under ordering constraints≫. In: *European Conference on Computer Vision*. 2014, pp. 628–643.

[6] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid. ≪Weakly-supervised alignment of video with text≫. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4462–4470.

[7] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. ≪Learning multi-label scene classification≫. In: *Pattern recognition* 37.9 (2004), pp. 1757–1771.

[8] R. G. Cinbis, J. Verbeek, and C. Schmid. ≪Weakly supervised object localization with multi-fold multiple instance learning≫. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.1 (2017), pp. 189–203.

[9] T. Cour. ≪Weakly supervised learning from multiple modalities: Exploiting video, audio and text for video understanding≫. PhD thesis. University of Pennsylvania, 2009.

[10] T. Cour, B. Sapp, and B. Taskar. ≪Learning from partial labels≫. In: *Journal of Machine Learning Research* (2011).

[11] T. Cour, B. Sapp, C. Jordan, and B. Taskar. ≪Learning from ambiguously labeled images≫. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. 2009, pp. 919–926.

[12] T. Cour, B. Sapp, A. Nagle, and B. Taskar. ≪Talking pictures: Temporal grouping and dialog-supervised person recognition≫. In: *Computer Vision and Pattern Recognition, 2010. CVPR 2010. IEEE Conference on*. IEEE. 2010, pp. 1014–1021.

[13] N. Dalal and B. Triggs. ≪Histograms of oriented gradients for human detection≫. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Conference on*. Vol. 1. IEEE. 2005, pp. 886–893.

[14] D. Das, D. Chen, A. F. Martins, N. Schneider, and N. A. Smith. ≪Frame-semantic parsing≫. In: *Computational linguistics* 40.1 (2014), pp. 9–56.

[15] M.-C. De Marneffe and C. D. Manning. *Stanford typed dependencies manual*. Tech. rep. 2008.

[16] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. ≪Solving the multiple instance problem with axis-parallel rectangles≫. In: *Artificial Intelligence* 89.1 (1997), pp. 31–71.

[17] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. ≪Automatic annotation of human actions in video≫. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2009, pp. 1491–1498.

[18] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth. ≪Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary≫. In: *European Conference on Computer Vision*. Springer. 2002, pp. 97–112.

[19] M. Everingham, J. Sivic, and A. Zisserman. ≪"Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video≫. In: *Proceedings of the British Machine Vision Conference*. 2006.

[20] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. ≪Visual object detection with deformable part models≫. In: *Communications of the ACM* 56.9 (2013), pp. 97–105.

[21] P. F. Felzenszwalb and D. P. Huttenlocher. ≪Pictorial structures for object recognition≫. In: *International Journal of Computer Vision* 61.1 (2005), pp. 55–79.

[22] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, 2001.

[23] S. Gong and T. Xiang. *Visual analysis of behaviour: from pixels to semantics*. Springer Science & Business Media, 2011.

[24] L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese. ≪UMBC EBIQUITY-CORE: Semantic textual similarity systems≫. In: *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*. 2013.

[25] J. Hernández-González, I. Inza, and J. A. Lozano. ≪Weak supervision and other non-standard classification problems: a taxonomy≫. In: *Pattern Recognition Letters* 69 (2016), pp. 49–55.

[26] M. Hoai, Z.-Z. Lan, and F. De la Torre. ≪Joint segmentation and classification of human actions in video≫. In: *Computer Vision and Pattern Recognition, 2011. CVPR 2011. IEEE Conference on*. IEEE. 2011, pp. 3265–3272.

[27] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy, and C. G. Snoek. ≪Action localization with tubelets from motion≫. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2014, pp. 740–747.

[28] R. Jin and Z. Ghahramani. ≪Learning with multiple labels≫. In: *Advances in Neural Information Processing Systems*. 2003, pp. 921–928.

[29] S. Kang and R. P. Wildes. ≪Review of Action Recognition and Detection Methods≫. In: *CoRR* (2016).

[30] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. ≪A large-scale classification of English verbs≫. In: *Language Resources and Evaluation* 42.1 (2008), pp. 21–40.

[31] H. Kuehne, A. Richard, and J. Gall. ≪Weakly supervised learning of actions from transcripts≫. In: *Computer Vision and Image Understanding* (2017).

[32] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. ≪Learning realistic human actions from movies≫. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008, pp. 1–8.

[33] D. G. Lowe. ≪Object recognition from local scale-invariant features≫. In: *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2. 1999, pp. 1150–1157.

[34] J. Luo and F. Orabona. ≪Learning from candidate labeling sets≫. In: *Advances in Neural Information Processing Systems*. 2010, pp. 1504–1512.

[35] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. ≪The Stanford CoreNLP Natural Language Processing Toolkit≫. In: *Association for Computational Linguistics (ACL) System Demonstrations*. 2014, pp. 55–60.

[36] O. Maron and A. L. Ratan. ≪Multiple-Instance Learning for Natural Scene Classification.≫ In: *International Conference on Machine Learning*. 1998.

[37] M. Marszalek, I. Laptev, and C. Schmid. ≪Actions in context≫. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE. 2009, pp. 2929–2936.

[38] J. Mas and G. Fernandez. ≪Video shot boundary detection based on color histogram≫. In: *Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST* (2003).

[39] G. A. Miller. ≪WordNet: a lexical database for English≫. In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[40] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. «Introduction to WordNet: An on-line lexical database». In: *International Journal of Lexicography* 3.4 (1990), pp. 235–244.

[41] C. B. Moler. *Numerical computing with MATLAB*. SIAM, 2004.

[42] S. Naha and Y. Wang. «Beyond verbs: Understanding actions in videos with text». In: *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE. 2016, pp. 1833–1838.

[43] O. M. Parkhi, E. Rahtu, and A. Zisserman. «It's in the bag: Stronger supervision for automated face labelling». In: *ICCV Workshop: Describing and Understanding Video & The Large Scale Movie Description Challenge*. IEEE. 2015.

[44] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. «Deep Face Recognition.» In: *Proceedings of the British Machine Vision Conference*. 2015.

[45] R. Poppe. «A survey on vision-based human action recognition». In: *Image and Vision Computing* 28.6 (2010), pp. 976–990.

[46] V. Ramanathan, P. Liang, and L. Fei-Fei. «Video event understanding using natural language descriptions». In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 905–912.

[47] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. «Linking people in videos with "their" names using coreference resolution». In: *European Conference on Computer Vision*. Springer. 2014, pp. 95–110.

[48] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. «A dataset for movie description». In: *Computer Vision and Pattern Recognition, 2015. CVPR 2015. IEEE Conference on*. 2015, pp. 3202–3212.

[49] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. «Movie Description». In: *International Journal of Computer Vision* 123.1 (2017), pp. 94–120.

[50] P. Sankar, C. V. Jawahar, and A. Zisserman. «Subtitle-Free Movie to Script Alignment». In: *Proceedings of the British Machine Vision Conference*. 2009.

[51] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[52] S. Shantaiya, K. Verma, and K. Mehta. «A survey on approaches of object detection». In: *International Journal of Computer Applications* 65.18 (2013).

[53] J. Sivic, M. Everingham, and A. Zisserman. «"Who are you?"-Learning person specific classifiers from video». In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE. 2009, pp. 1145–1152.

[54] M. Tapaswi, M. Bauml, and R. Stiefelhagen. «Book2movie: Aligning video scenes with book chapters». In: *Computer Vision and Pattern Recognition, 2015. CVPR 2015. IEEE Conference on*. 2015, pp. 1827–1835.

[55] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. «"Knock! Knock! Who is it?" probabilistic person identification in TV-series». In: *Computer Vision and Pattern Recognition, 2012. CVPR 2012. IEEE Conference on*. IEEE. 2012, pp. 2658–2665.

[56] Ευάγγελος Α. Νικολουδάκης. «Αναγνώριση Ανθρώπινων Δράσεων και Χειρονομιών χρησιμοποιώντας Βαθιά Συνελικτικά Νευρωνικά Δίκτυα». Greek. MA thesis. Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, 2017.

[57] Ολίβια Σ. Καραθάνου. «Πολυτροπική Κατάτμηση Ταινιών σε Σκηνές». Greek. MA thesis. Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, 2015.

[58] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. «Learning spatiotemporal features with 3d convolutional networks». In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4489–4497.

[59] G. Tsoumakas and I. Katakis. «Multi-label classification: An overview». In: *International Journal of Data Warehousing and Mining* 3.3 (2006).

[60] L. G. Valiant. «A theory of the learnable». In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[61] V. N. Vapnik and A. Y. Chervonenkis. «On the uniform convergence of relative frequencies of events to their probabilities». In: *Measures of Complexity*. Springer, 2015, pp. 11–30.

[62] H. Wang and C. Schmid. «Action recognition with improved trajectories». In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 3551–3558.

[63] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. «UntrimmedNets for Weakly Supervised Action Recognition and Detection». In: *arXiv preprint arXiv:1703.03329* (2017).

[64] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. «Learning to track for spatio-temporal action localization». In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3164–3172.

[65] P. Weinzaepfel, X. Martin, and C. Schmid. «Human Action Localization with Sparse Spatial Supervision». working paper or preprint. 2017.

[66] S. Zafeiriou, C. Zhang, and Z. Zhang. «A survey on face detection in the wild: past, present and future». In: *Computer Vision and Image Understanding* 138 (2015), pp. 1–24.

[67] C. Zhang, J. C. Platt, and P. A. Viola. «Multiple instance boosting for object detection». In: *Advances in Neural Information Processing Systems*. 2006, pp. 1417–1424.

[68] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. ≪Multi-instance multi-label learning≫. In: *Artificial Intelligence* 176.1 (2012), pp. 2291–2320.

[69] X. Zhu and D. Ramanan. ≪Face detection, pose estimation, and landmark localization in the wild≫. In: *Computer Vision and Pattern Recognition, 2012. CVPR 2012. IEEE Conference on*. IEEE. 2012, pp. 2879–2886.

[70] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, and P. Maragos. ≪Multimodal Video Database Annotated with Saliency, Events, Semantics and Emotion with Application to Summarization≫. In: *EURASIP Journal on Image and Video Processings* (submitted).