



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΑΥΤΟΜΑΤΗ ΠΕΡΙΓΡΑΦΗ ΤΟΠΙΟΥ ΜΕ ΗΧΗΤΙΚΗ ΑΠΟΔΟΣΗ ΓΙΑ ΧΡΗΣΤΕΣ ΜΕ ΠΡΟΒΛΗΜΑΤΑ ΟΡΑΣΗΣ

ΓΑΝΤΖΟΥΛΑ ΣΤΑΥΡΟΥΛΑ, ΚΑΡΑΔΗΜΗΤΡΙΟΥ ΣΟΦΙΑ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος
Αναπληρωτής Καθηγητής

Λαμία Σεπτέμβριος 2026

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία έχει ως στόχο την ανάπτυξη ενός συστήματος αυτόματης περιγραφής του περιβάλλοντος με ηχητική απόδοση για την υποβοήθηση ατόμων με προβλήματα όρασης. Οι δυσκολίες στην αντίληψη του χώρου και στην αναγνώριση αντικειμένων μπορούν να επηρεάσουν την καθημερινότητα και την αυτονομία των ατόμων αυτών, ενώ η εξέλιξη της τεχνητής νοημοσύνης και της υπολογιστικής όρασης προσφέρει νέες δυνατότητες για την ανάπτυξη σχετικών εφαρμογών υποβοήθησης. Στο πλαίσιο της εργασίας αναπτύχθηκε ένα σύστημα το οποίο λαμβάνει εικόνες από κάμερα, αναγνωρίζει αντικείμενα και στοιχεία του περιβάλλοντος και δημιουργεί μία περιγραφή της σκηνής. Για την υλοποίησή του αξιοποιούνται μοντέλα βαθιάς μάθησης για την ανίχνευση αντικειμένων και την αυτόματη περιγραφή εικόνων, και συγκεκριμένα τα YOLOv8, YOLO-World και BLIP. Οι πληροφορίες που προκύπτουν συνδυάζονται για τη δημιουργία της τελικής περιγραφής, η οποία στη συνέχεια μετατρέπεται σε ομιλία μέσω τεχνολογίας Text-to-Speech. Παράλληλα, αναπτύχθηκε γραφική διεπαφή μέσω της οποίας προβάλλεται η εικόνα της κάμερας και πραγματοποιείται η ανάλυση της σκηνής. Η λειτουργία του συστήματος εξετάστηκε σε διαφορετικές περιπτώσεις, από απλές σκηνές έως πιο σύνθετα περιβάλλοντα, με σκοπό να αξιολογηθούν η αναγνώριση και περιγραφή των στοιχείων της εικόνας, η παροχή βασικών χωρικών πληροφοριών, καθώς και περιπτώσεις λανθασμένων ανιχνεύσεων. Μέσα από τις δοκιμές αναδείχθηκαν τόσο οι δυνατότητες του συστήματος όσο και περιορισμοί που σχετίζονται με την αναγνώριση ορισμένων αντικειμένων και την αξιοπιστία των παραγόμενων πληροφοριών. Τέλος, παρουσιάζονται δυνατότητες μελλοντικής βελτίωσης και επέκτασης του συστήματος, με στόχο την ανάπτυξη μιας περισσότερο αξιόπιστης και πρακτικής λύσης για την υποβοήθηση ατόμων με προβλήματα όρασης.

ABSTRACT

This thesis aims to develop an automatic scene description system with audio output for assisting people with visual impairments. Difficulties in perceiving the surrounding environment and recognizing objects can affect the daily lives and independence of visually impaired people, while advances in artificial intelligence and computer vision provide new possibilities for the development of assistive applications. As part of this thesis, a system was developed that captures images from a camera, recognizes objects and other elements of the environment, and generates a description of the scene. Deep learning models for object detection and automatic image captioning are utilized, specifically YOLOv8, YOLO-World, and BLIP. The information produced by these models is combined to generate the final description, which is subsequently converted into speech using Text-to-Speech technology. A graphical user interface was also developed to display the camera feed and perform scene analysis. The system was tested in different cases, ranging from simple scenes to more complex environments, in order to evaluate the recognition and description of image elements, the provision of basic spatial information, as well as cases of incorrect detections. The tests highlighted both the capabilities of the developed system and limitations related to the recognition of certain objects and the reliability of the generated information. Finally, possible future improvements and extensions of the system are presented, aiming toward the development of a more reliable and practical solution for assisting people with visual impairments.

Πίνακας περιεχομένων

ΚΕΦΑΛΑΙΟ 1 Εισαγωγή και Κίνητρο	8
ΚΕΦΑΛΑΙΟ 2 Ανάλυση Βιβλιογραφίας	10
2.1 Προβλήματα όρασης και ανάγκες υποβοήθησης	10
2.2 Υφιστάμενα συστήματα υποβοήθησης ατόμων με προβλήματα όρασης	11
2.3 Σύγχρονες ερευνητικές προσεγγίσεις στην αναγνώριση αντικειμένων	12
ΚΕΦΑΛΑΙΟ 3 Προηγμένες Τεχνικές Μηχανικής Μάθησης	16
3.1 Deep Learning	17
3.2 Συνελκτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs)	18
3.3 YOLO	18
3.4 YOLO-World / Open-Vocabulary Detection	20
3.5 BLIP	20
3.6 Μοντέλα Όρασης-Γλώσσας (Vision-Language Models - VLMs)	21
3.7 Text-to-Speech (TTS)	22
ΚΕΦΑΛΑΙΟ 4 Συστήματα Υποβοήθησης Ανθρώπων με Προβλήματα Όρασης	23
4.1 Σχεδιαστικές απαιτήσεις συστημάτων οπτικής υποβοήθησης	23
4.2 Επιλογή και ιεράρχηση της πληροφορίας	24
4.3 Προσβάσιμη Απόδοση της Περιγραφής και Αλληλεπίδραση με τον Χρήστη	24
4.4 Αξιοπιστία της Παραγόμενης Πληροφορίας	25
ΚΕΦΑΛΑΙΟ 5 Προτεινόμενο Σύστημα	27
5.1 Επισκόπηση Συστήματος	27
5.2 Ροή Λειτουργίας	27
5.3 Περιγραφή των επιμέρους υποσυστημάτων	29
5.3.1 Λήψη Εικόνας	29
5.3.2 Ανίχνευση Αντικειμένων	30
5.3.3 Δημιουργία περιγραφής της σκηνής	31
5.3.4 Υποσύστημα σύνθεσης ομιλίας	32
5.4 Πλεονεκτήματα και περιορισμοί του συστήματος	33
ΚΕΦΑΛΑΙΟ 6 Μελέτες Περίπτωσης	35
6.1 Μεθοδολογία αξιολόγησης	35
6.2 Περίπτωση απλής σκηνής	35
6.3 Περίπτωση σύνθετης σκηνής	37
6.4 Αξιολόγηση χωρικής πληροφορίας	39
6.5 Περιπτώσεις λανθασμένων ανιχνεύσεων	40
6.6 Επίδραση των συνθηκών λήψης	41
6.7 Συνολική αξιολόγηση	42

ΚΕΦΑΛΑΙΟ 7 Συμπεράσματα και Μελλοντικές Προεκτάσεις	43
7.1 Συμπεράσματα	43
7.2 Μελλοντικές Βελτιώσεις	44
ΚΕΦΑΛΑΙΟ 8 Επίλογος	49
Βιβλιογραφία	50

ΚΕΦΑΛΑΙΟ 1 Εισαγωγή και Κίνητρο

Η όραση αποτελεί τον θεμελιώδη πυλώνα της ανθρώπινης αντίληψης, διαδραματίζοντας κρίσιμο ρόλο σε κάθε πτυχή και στάδιο της ζωής μας. Αποτελεί μία από τις σημαντικότερες αισθήσεις για την αντίληψη και τη κατανόηση του περιβάλλοντος. Μέσω της οπτικής πληροφορίας, ο άνθρωπος αναγνωρίζει αντικείμενα, πρόσωπα, χώρους και πιθανές δυσκολίες που υπάρχουν γύρω του. Η απώλεια ή η σοβαρή μείωση αυτής της αίσθησης δημιουργεί σημαντικές προκλήσεις στην αυτόματη διαβίωση και τη προσβασιμότητα. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας και τις μελέτες που έγιναν τον Φεβρουάριο του 2026, τουλάχιστον 2.2 δισεκατομμύρια άνθρωποι παγκοσμίως αντιμετωπίζουν κάποιο πρόβλημα στην όραση τους. Η ανάπτυξη των τεχνολογιών, της Τεχνητής Νοημοσύνης και της Υπολογιστικής Όρασης ξεκλειδώνει νέες προοπτικές για την αντιμετώπιση αυτών των κοινωνικών και λειτουργικών προκλήσεων. Έχουν δημιουργήσει νέες δυνατότητες για τη διευκόλυνση των ατόμων αυτών, αφού πλέον μπορούν να επεξεργάζονται εικόνες, να αναγνωρίζουν αντικείμενα και να παράγουν φυσικές γλωσσικές περιγραφές.

Στο πλαίσιο αυτό, η παρούσα πτυχιακή εργασία επικεντρώνεται πάνω σε αυτό το θέμα και αναλύει ένα σύστημα αυτόματης περιγραφής τοπίου με ηχητική απόδοση για χρήστες με προβλήματα όρασης. Το σύστημα λαμβάνει στιγμιότυπα από ζωντανή κάμερα, συνδυάζει τεχνικές ανίχνευσης αντικειμένων, ερμηνείας της σκηνής και παραγωγής φυσικής γλωσσικής περιγραφής.

Βασικό κίνητρο της εργασίας είναι η αξιοποίηση της τεχνητής νοημοσύνης για τη βελτίωση της πρόσβασης στην οπτική πληροφορία και η ουσιαστική αξιοποίηση των νευρωνικών δικτύων για τη γεφύρωση του χάσματος μεταξύ του οπτικού κόσμου και των ατόμων με οπτική αναπηρία. Ένα άτομο με περιορισμένη έως καθόλου όραση έχει ανάγκη να γνωρίζει πληροφορίες σχετικά με το χώρο που βρίσκεται, τι υπάρχει μπροστά του, τι αντικείμενα μπορεί να του φανούν χρήσιμα ή να τον εμποδίζουν και γενικά ποια είναι η γενική κατάσταση μιας σκηνής.

Μία από τις κρίσιμες παραδοχές της παρούσας μελέτης είναι ότι η απομονωμένη ανίχνευση αντικειμένων δεν επαρκεί για την κάλυψη των σύνθετων αναγκών αντίληψης ενός ατόμου. Η απομονωμένη πληροφορία ότι σε μια εικόνα υπάρχουν απλώς ένας άνθρωπος και μία καρέκλα, είναι συχνά λιγότερο χρήσιμη από μια ολιστική περιγραφή που λαμβάνει υπόψη τη μεταξύ τους χωρική θέση και το γενικό περιεχόμενο της σκηνής, που μπορεί να προσφέρει πραγματική γνωστική αξία.

Επομένως, ο πρωταρχικός στόχος δεν περιορίζεται απλώς στην ανίχνευση αντικειμένων, αλλά επεκτείνεται στη σύνθεση συνοπτικής αλλά δομημένης και απόλυτα κατανοητής περιγραφής. Η περιγραφή αυτή έχει ως σκοπό να ερμηνεύσει το οπτικό ερέθισμα με όσο το δυνατό πιο φυσικό τρόπο, ώστε να μπορεί να αποδοθεί άμεσα και σε πραγματικό χρόνο προς τον χρήστη

μέσω τεχνολογιών εκφώνησης και ομιλίας, διευκολύνοντας έτσι την άμεση πρόσληψη της πληροφορίας.

Η συνεχής εξέλιξη των συστημάτων υπολογιστικής όρασης έχει οδηγήσει στην ανάπτυξη πληθώρας εφαρμογών που υποστηρίζουν άτομα με προβλήματα όρασης, είτε μέσω αναγνώρισης αντικειμένων είτε μέσω παραγωγής περιγραφών εικόνων. Παρά την πρόοδο αυτή, μεγάλο μέρος της υπάρχουσας βιβλιογραφίας αντιμετωπίζει τις επιμέρους λειτουργίες ως ανεξάρτητα προβλήματα, εστιάζοντας είτε στην ανίχνευση αντικειμένων είτε στην παραγωγή γενικών περιγραφών μίας εικόνας. Στη πράξη όμως, ένας χρήστης που κινείται σε πραγματικό περιβάλλον χρειάζεται μία περισσότερη ενοποιημένη μορφή πληροφόρησης, η οποία να συνδυάζει την αναγνώριση των σημαντικότερων στοιχείων της σκηνής με τη μεταξύ τους σχέση και να αποδίδεται άμεσα μέσω ομιλίας.

Η παρούσα εργασία προσεγγίζει αυτή την ανάγκη μέσα από ένα συνδυασμό τεχνικών που λειτουργούν συμπληρωματικά και όχι απομονωμένα. Αντί να παρουσιάζεται στον χρήστη μία απλή λίστα αντικειμένων που εντοπίστηκαν στην εικόνα, το σύστημα συνθέτει μία ενιαία περιγραφή της σκηνής, αξιοποιώντας πληροφορίες για τη παρουσία, τη σχετική θέση και τη σημασία των αντικειμένων μέσα στο περιβάλλον. Με αυτό το τρόπο η πληροφορία οργανώνεται όπως ένας άνθρωπος θα περιέγραφε προφορικά ένα χώρο, γεγονός που καθιστά την ηχητική απόδοση περισσότερο κατανοητή και χρηστική.

Η συνεισφορά της εργασίας δεν έγκειται στην ανάπτυξη ενός νέου αλγορίθμου μηχανικής μάθησης, αλλά στην ολοκληρωμένη αξιοποίηση υπαρχόντων σύγχρονων μοντέλων σε μία ενιαία διαδικασία ανάλυσης, περιγραφής και εκφώνησης της σκηνής. Η προσέγγιση αυτή δίνει έμφαση στη πρακτική χρησιμότητα του συστήματος σε πραγματικές συνθήκες χρήσης, όπου η ταχύτητα απόκρισης, η σαφήνεια της πληροφορίας και η δυνατότητα άμεσης κατανόησης της σκηνής αποτελούν κρίσιμους παράγοντες για την υποστήριξη του χρήστη.

ΚΕΦΑΛΑΙΟ 2 Ανάλυση Βιβλιογραφίας

2.1 Προβλήματα όρασης και ανάγκες υποβοήθησης

Η αυτόνομη διακίνηση και η καθημερινή αλληλεπίδραση των ατόμων με προβλήματα όρασης με το περιβάλλον βασίζονται παραδοσιακά σε φυσικά βοηθήματα, όπως το λευκό συνήθως σπαστό μαστούνι ή τους ειδικά εκπαιδευμένους σκύλους οδηγούς για αυτές τις περιστάσεις. Τα μέσα αυτά έχουν προσφέρει ανεκτίμητη υπηρεσία ώστε να ανεξαρτητοποιηθούν τα άτομα με οπτική αναπηρία, παρέχοντας φυσική προστασία εντοπίζοντας εμπόδια στο επίπεδο του εδάφους και καθοδηγώντας τους προς μία κατεύθυνση. Ωστόσο, παρά τη διαχρονική τους σημαντικότητα, οι συμβατικές αυτές λύσεις παρουσιάζουν δομικούς περιορισμούς. Η πληροφορία που μεταφέρουν στον χρήστη είναι αμιγώς μηχανική ή τακτική, ενώ στερείται εννοιολογικού και ποιοτικού περιεχομένου, καθώς και από μία πιο σαφή απάντηση για το τι έχουν να αντιμετωπίσουν.

Ένα λευκό μαστούνι είναι το σημαντικότερο τεχνικό βοήθημα στην προσπάθεια ενός ατόμου με πρόβλημα όρασης να κινηθεί με άνεση, ανεξαρτησία και ασφάλεια. Μπορεί εύκολα να προειδοποιήσει τον χρήστη για την ύπαρξη ενός φυσικού εμποδίου μπροστά του, όμως αδυνατεί να προσδιορίσει την ταυτότητα του αντικειμένου, όπως για παράδειγμα αν πρόκειται για ένα σταθμευμένο ποδήλατο, ένα κάδο απορριμμάτων, ή ένα τραπέζι. Αντίστοιχα, ένας σκύλος-οδηγός εκπαιδεύεται για να καθοδηγεί τον κάτοχό του με ασφάλεια γύρω από εμπόδια. Παρόλο που μπορούν να εκπαιδευτούν να κρατάνε τον άνθρωπο ασφαλή, αδυνατούν να αντιληφθούν και να ερμηνεύσουν το χρώμα από το φανάρι στο δρόμο, να μεταφέρουν τη πληροφορία μίας πινακίδας οδικής σήμανσης, ή να διαβάσουν την ετικέτα ενός προϊόντος στο ράφι ενός καταστήματος.

Ωστόσο, η επικρατούσα πρακτική βασίζεται ακόμη σε μεγάλο βαθμό στην ανθρώπινη υποστήριξη εξ αποστάσεων. Αν και η ανθρώπινη υποστήριξη είναι αποτελεσματική, ενέχει σημαντικά μειονεκτήματα, καθώς εισάγει καθυστερήσεις, απαιτεί συνεχή σύνδεση στο διαδίκτυο, περιορίζει την ιδιωτικότητα του χρήστη και περιορίζει την προσωπική αυτονομία. Η ανάγκη αυτή ανέδειξε την αναγκαιότητα για αυτόνομα, έξυπνα συστήματα που εκτελούν τη διαδικασία περιγραφής τοπίου αυτόματα, σε πραγματικό χρόνο και υπό πραγματικές, μεταβαλλόμενες συνθήκες χρήσης.

Η κάλυψη αυτού του πληροφοριακού κενού καθίσταται πλέον εφικτή, χάρη στη ραγδαία σύγκλιση τριών βασικών τεχνολογικών πυλώνων: της Τεχνητής Νοημοσύνης, της Υπολογιστικής Όρασης (Computer Vision) και της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing). Η πραγματική δύναμη δεν βρίσκεται σε καθέναν από αυτούς τους τομείς ξεχωριστά, αλλά στη σύγκλισή τους. Όταν η Υπολογιστική Όραση συνδυάζεται με την Επεξεργασία Φυσικής Γλώσσας υπό την ομπρέλα της Τεχνητής Νοημοσύνης, το σύστημα δεν βλέπει απλώς εικόνες. Μπορεί να τις περιγράψει, να απαντήσει σε ερωτήσεις γι' αυτές και να καταλάβει το πλαίσιο. Μέσω προηγμένων αρχιτεκτονικών βαθιάς μάθησης, όπως τα συνελκτικά νευρωνικά δίκτυα και τα μοντέλα Vision Transformers, τα σύγχρονα συστήματα

υπολογιστικής όρασης μπορούν όχι μόνο να ανιχνεύουν και να ταξινομούν αντικείμενα σε πραγματικό χρόνο, αλλά και να αντιλαμβάνονται τη χωρική τους διάταξη στο χώρο. Έτσι περνάμε από την απλή επεξεργασία δεδομένων στην ουσιαστική κατανόηση του φυσικού κόσμου.

Τα μοντέλα επεξεργασίας φυσικής γλώσσας και αυτόματης παραγωγής περιγραφών αναλαμβάνουν να μετατρέψουν τις ακατέργαστες οπτικές αναγνωρίσεις σε μια συνεκτική, συντακτικά ορθή και φυσική λεκτική περιγραφή της σκηνής, συνδέοντας τα αντικείμενα με τις μεταξύ τους αλληλεπιδράσεις. Αποτελούν κρίσιμο συνδετικό κρίκο που μετατρέπει την υπολογιστική ανάλυση σε προσβάσιμη ανθρώπινη εμπειρία. Οι τεχνολογίες σύνθεσης ομιλίας (Text-to-Speech) μετατρέπουν ακαριαία το παραγόμενο κείμενο σε φυσικό προφορικό λόγο, επιτρέποντας στον χρήστη να ακούει το περιβάλλον του χωρίς αισθητές καθυστερήσεις.

Η επιτακτική ανάγκη για την ανάπτυξη προηγμένων αυτόματων συστημάτων οπτικής υποβοήθησης αποδεικνύεται καθημερινά από τις ίδιες τις εμπειρίες και τις προκλήσεις που αντιμετωπίζουν τα άτομα με προβλήματα όρασης. Στη σύγχρονη ψηφιακή εποχή, η ευρεία εξάπλωση των έξυπνων κινητών τηλεφώνων, των φορητών καμερών και των ενσωματωμένων αισθητήρων ώθησε πολλούς χρήστες στο να χρησιμοποιούν τη λήψη φωτογραφιών ή βίντεο ως μέσο «αποκωδικοποίησης» του φυσικού και αστικού περιβάλλοντος που τους περιβάλλει.

2.2 Υφιστάμενα συστήματα υποβοήθησης ατόμων με προβλήματα όρασης

Τα τελευταία χρόνια έχουν αναπτυχθεί πλήθος ερευνητικών εργασιών που αξιοποιούν τεχνικές ανίχνευσης αντικειμένων για την υποστήριξη ατόμων με προβλήματα όρασης. Μέσω εξειδικευμένων εφαρμογών, όπως το Be My Eyes, ο χρήστης πραγματοποιεί βιντεοκλήσεις με εθελοντές ή οικεία πρόσωπα που μιλάνε την ίδια γλώσσα και βρίσκονται στην ίδια χρονική ζώνη, προκειμένου να λάβει προφορικές περιγραφές για το τι βρίσκεται μπροστά του, από την αναγνώριση μιας πινακίδας ή ενός εμποδίου στον δρόμο, μέχρι την περιγραφή ενός τοπίου. Η εφαρμογή αυτή είναι δωρεάν για τους χρήστες και τους εθελοντές και βασίζεται στη Τεχνητή Νοημοσύνη (Be My AI) όπου ο χρήστης τραβάει μια φωτογραφία (π.χ. το εσωτερικό ενός ψυγείου, ένα χώρο, ένα τοπίο ή ένα ρούχο) και το σύστημα παράγει αυτόματα μια λεπτομερή, γραπτή και ηχητική περιγραφή. Η εφαρμογή αυτή υποστηρίζεται σε έξυπνα κινητά, σε συγκεκριμένες εκδόσεις για υπολογιστές Desktop (Windows & macOS), σε ειδική έκδοση για επαγγελματικά περιβάλλοντα, καθώς επίσης εφαρμόζεται πλήρως σε έξυπνα γυαλιά, επιτρέποντας στους χρήστες να λαμβάνουν φωνητικές περιγραφές hands-free χωρίς να κρατούν το κινητό.

Η εφαρμογή αυτή βασίζεται εξ ολοκλήρου στο Cloud AI και η εικόνα αποστέλλεται στους διακομιστές της Open AI όπου γίνεται η ανάλυση και η παραγωγή του αντίστοιχου κειμένου, κάτι που χρειάζεται αδιάκοπη και ταχεία σύνδεση στο διαδίκτυο. Επιπλέον, οι περιγραφές που δίνει η εργασία μας εστιάζει στην αντίληψη του χωρικού τοπίου και δίνει προτεραιότητα σε δομημένες, σύντομες και ουσιαστικές ηχητικές προτάσεις.

Παρόμοια νοοτροπία ακολουθεί και η εφαρμογή Envision, μια δωρεάν εφαρμογή με τεχνολογία αιχμής OCR που διατυπώνει καθημερινές οπτικές πληροφορίες στην ομιλία και απαντά σε ερωτήσεις σχετικά με κείμενα ή σκηνές. Χρησιμοποιεί την κάμερα του smartphone για να εκφωνήσει γραπτές πληροφορίες και να περιγράψει περιβάλλοντα και αντικείμενα. Βασίζεται στην ανάγνωση κειμένου, μετατρέποντας ένα κείμενο σε ομιλία αφού το σαρώσει, περιγράφει μια σκηνή είτε τραβώντας μια φωτογραφία εκείνη τη στιγμή είτε εισάγοντας μια εικόνα στο σύστημα, καθώς επίσης απαντάει σε πιθανές ερωτήσεις του χρήστη που αφορούν τη φωτογραφία που χρησιμοποίησε νωρίτερα. Κάτι επίσης σημαντικό για την Envision είναι πως μπορεί να βρει άτομα καθώς επίσης και να τα αναγνωρίζει εφόσον τα διδάξεις στην εφαρμογή.

Τα Be My Eyes και Envision App αποτελούν δύο από τις πιο διαδεδομένες εφαρμογές υποβοήθησης ατόμων με προβλήματα όρασης. Και οι δύο αξιοποιούν τεχνολογίες τεχνητής νοημοσύνης για την περιγραφή του περιβάλλοντος, ωστόσο ακολουθούν διαφορετική φιλοσοφία λειτουργίας από το προτεινόμενο σύστημα της παρούσας εργασίας. Εκεί που η εργασία μας ξεχωρίζει, είναι στη σχεδίαση της αρχιτεκτονικής και στον τρόπο που συνδυάζει διαφορετικά μοντέλα για να παράγει μία τελική περιγραφή. Δεν επιχειρεί να αντικαταστήσει τις υπάρχουσες εμπορικές εφαρμογές. Στόχος της είναι να διερευνήσει μία διαφορετική αρχιτεκτονική, όπου η περιγραφή του περιβάλλοντος δεν παράγεται από ένα μόνο μοντέλο, αλλά από τον συνδυασμό εξειδικευμένων μοντέλων αναγνώρισης αντικειμένων και image captioning, τα οποία συνθέτουν μία πιο δομημένη τελική περιγραφή.

Μία από τις σημαντικότερες εφαρμογές στον χώρο της ψηφιακής προσβασιμότητας είναι το Seeing AI, το οποίο αναπτύχθηκε από τη Microsoft με σκοπό να βοηθήσει άτομα με προβλήματα όρασης να αντιλαμβάνονται καλύτερα το περιβάλλον τους μέσω της κάμερας ενός έξυπνου κινητού τηλεφώνου. Η εφαρμογή αξιοποιεί τεχνικές υπολογιστικής όρασης, cloud και τεχνητής νοημοσύνης για την αναγνώριση αντικειμένων, προσώπων, προϊόντων, νομισμάτων, χειρόγραφου και τυπωμένου κειμένου, καθώς και για την παραγωγή σύντομων περιγραφών ολόκληρων σκηνών.

Ιδιαίτερο χαρακτηριστικό του Seeing AI είναι ότι οργανώνει τις λειτουργίες του σε διαφορετικά «κανάλια», επιτρέποντας στον χρήστη να επιλέγει αν επιθυμεί ανάγνωση κειμένου, αναγνώριση προϊόντων μέσω barcode, περιγραφή περιβάλλοντος ή αναγνώριση ατόμων. Με τον τρόπο αυτό προσαρμόζεται σε διαφορετικές καθημερινές ανάγκες χωρίς να απαιτείται εξειδικευμένος εξοπλισμός πέρα από το κινητό τηλέφωνο.

Παρότι η εφαρμογή προσφέρει υψηλό επίπεδο λειτουργικότητας, οι περιγραφές των σκηνών παραμένουν σχετικά σύντομες και δεν αποδίδουν πάντοτε λεπτομερώς τις χωρικές σχέσεις μεταξύ των αντικειμένων.

2.3 Σύγχρονες ερευνητικές προσεγγίσεις στην αναγνώριση αντικειμένων

Οι πρόσφατες ερευνητικές εργασίες δείχνουν ότι οι εικόνες που λαμβάνονται από άτομα με προβλήματα όρασης παρουσιάζουν ιδιαίτερες προκλήσεις. Πολλές φορές οι φωτογραφίες που

τραβούν με την κάμερά τους μπορεί να μην εστιάζει σωστά στο αντικείμενο ή στο εμπόδιο που έχουν μπροστά τους, ή η εικόνα να είναι χαμηλής ποιότητας, με αποτέλεσμα η εφαρμογή να οδηγεί σε εσφαλμένες ανιχνεύσεις (false positives). Όπως επίσης, θα μπορούσε η εικόνα να παρουσιάζει θόλωση λόγω κίνησης ή να μην εμφανίζονται επαρκώς τα όρια του αντικειμένου που θέλει να αποτυπώσει.

Πάνω σε αυτά τα προβλήματα βασίζεται το VizWiz Salient Object, ένα εξειδικευμένο σύνολο δεδομένων (dataset) και ένα αντίστοιχο υπολογιστικό πρόβλημα στο χώρο της Όρασης Υπολογιστών ειδικά σχεδιασμένο για να εξυπηρετεί τέτοιου είδους ανάγκες. Προέρχεται από το οικοσύστημα VizWiz, από το Πανεπιστήμιο του Τέξας, και εστιάζει στον εντοπισμό του σημαντικού και κυρίαρχου αντικειμένου στη φωτογραφία, όταν αυτή δεν είναι η ιδανική αλλά η πραγματική. Εστιάζει δηλαδή στην απομόνωση του αντικειμένου που τραβάει τη προσοχή στην εικόνα. Το dataset του βασίζεται πλήρως σε φωτογραφίες που έχουν τραβηχτεί από τα ίδια άτομα με προβλήματα όρασης στην καθημερινότητά τους.

Στην εικόνα [1] παρακάτω παρουσιάζονται μερικά παραδείγματα από πραγματικές φωτογραφίες που τράβηξαν οι χρήστες της εφαρμογής VizWiz, καθώς και οι περιγραφές που κατάφερε να αποδώσει η εφαρμογή.

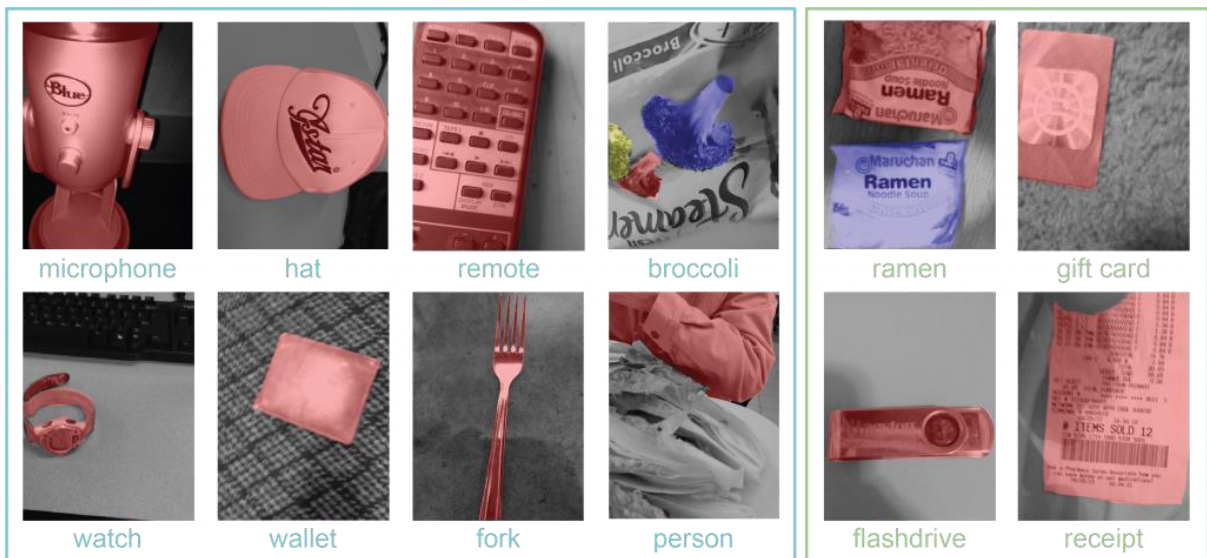


[1 - Image Quality Assessment – VizWiz]

Όπως φαίνεται και από τα περιγράμματα, χρησιμοποιείται πράσινο για τις «καθαρές» εικόνες, μωβ για τις θολές και κόκκινες για εκείνες που δεν ήταν δυνατή κάποια αναγνώριση.

Στις δύο εικόνες πιο κάτω εμφανίζονται διάφορα αντικείμενα που κατέγραψαν με το κινητό τους άτομα με προβλήματα στην όρασή τους. Είναι ένα αντιπροσωπευτικό παράδειγμα από το πρόβλημα Few-Shot Object Localization / Detection του VizWiz. Δείχνει πώς ένα νευρωνικό δίκτυο προσπαθεί να εντοπίσει και να διαχωρίσει το κύριο αντικείμενο (Segmentation Mask) από το φόντο σε πραγματικές λήψεις χρηστών με προβλήματα όρασης. Οι φωτογραφίες αυτές χρησιμοποιήθηκαν από το VizWiz για τις ανάγκες της έρευνας και λειτούργησαν σαν παραδείγματα για εκπαίδευση του.

Η πρώτη εικόνα [2] δείχνει καθαρά πώς το σύστημα απομονώνει το αντικείμενο με ροζ και μπλε επισήμανση, αγνοώντας το υπόλοιπο θορυβώδες φόντο. Οι κατηγορίες στις οποίες έχουν γίνει επισημάνσεις περιλαμβάνουν τόσο αυτές που είναι κοινές με προηγούμενες εργασίες (πράσινο πλαίσιο αριστερά) όσο και αυτές που είναι μοναδικές για το σύνολο δεδομένων.



[2 - Few-Shot Object Localization – VizWiz]

Στη δεύτερη εικόνα [3] τα αντικείμενα είναι εν μέρη κρυμμένα και δεν είναι ευδιάκριτα, μερικά εκ των οποίων επειδή δεν κατάφερε να τα «πιάσει» η κάμερα. Η εφαρμογή προσπαθεί να αναγνωρίσει αντικείμενα ακόμα και όταν δεν βρίσκονται στο επίκεντρο της φωτογραφίας ή βρίσκονται σε μεγάλη απόσταση από τον χρήστη.



[3 - Few-Shot Object Localization – VizWiz]

Μία ακόμη σημαντική συμβολή στη βιβλιογραφία αποτελεί το VizWiz Few Shot, το οποίο παρουσιάστηκε στο συνέδριο ECCV 2022 και εξετάζει μία ιδιαίτερα δύσκολη πρόκληση της υπολογιστικής όρασης, την αναγνώριση και τον εντοπισμό αντικειμένων όταν υπάρχουν ελάχιστα διαθέσιμα παραδείγματα εκπαίδευσης. Στα περισσότερα σύγχρονα συστήματα αναγνώρισης αντικειμένων, όπως τα μοντέλα που εκπαιδεύονται σε μεγάλα σύνολα δεδομένων (π.χ. COCO), απαιτούνται εκατοντάδες ή ακόμη και χιλιάδες εικόνες για κάθε κατηγορία αντικειμένου, ώστε το μοντέλο να μάθει να αναγνωρίζει σωστά τα χαρακτηριστικά του. Ωστόσο, στην πραγματική ζωή δεν υπάρχουν πάντοτε τόσα διαθέσιμα δεδομένα. Αυτό είναι ιδιαίτερα εμφανές στις εικόνες που τραβούν άτομα με προβλήματα όρασης, όπου πολλά αντικείμενα εμφανίζονται σπάνια, βρίσκονται σε ασυνήθιστες γωνίες λήψης ή είναι μερικώς ορατά. Η έννοια του few-shot learning αναφέρεται ακριβώς σε αυτήν την περίπτωση: το μοντέλο καλείται να αναγνωρίσει νέα αντικείμενα έχοντας εκπαιδευτεί με ελάχιστα παραδείγματα, συχνά μόλις μία, δύο ή λίγες εικόνες ανά κατηγορία.

Τα συμπεράσματα της έρευνας VizWiz-FewShot παρουσιάζουν ιδιαίτερο ενδιαφέρον για την παρούσα πτυχιακή εργασία, καθώς αναδεικνύουν τις δυσκολίες που αντιμετωπίζουν τα συστήματα αναγνώρισης αντικειμένων όταν λειτουργούν σε πραγματικές συνθήκες χρήσης. Το προτεινόμενο σύστημα της εργασίας αξιοποιεί μοντέλο YOLO για την ανίχνευση αντικειμένων σε πραγματικό χρόνο, στη συνέχεια χρησιμοποιεί το μοντέλο BLIP για τη δημιουργία περιγραφής της σκηνής και τέλος μετατρέπει την πληροφορία σε ομιλία μέσω τεχνολογίας Text-to-Speech. Επομένως, η ακρίβεια της αρχικής ανίχνευσης αντικειμένων αποτελεί καθοριστικό παράγοντα για την ποιότητα της τελικής ηχητικής περιγραφής. Οι προκλήσεις που επισημαίνει η βιβλιογραφία, όπως οι θολές εικόνες, οι ασυνήθιστες γωνίες λήψης, καθώς και τα πολύ μικρά ή υπερβολικά μεγάλα αντικείμενα, αποτελούν περιπτώσεις που μπορούν να επηρεάσουν άμεσα την απόδοση του προτεινόμενου συστήματος. Για τον λόγο αυτό, τα ευρήματα του VizWiz-FewShot προσφέρουν σημαντική θεωρητική βάση για την αξιολόγηση των περιορισμών της παρούσας υλοποίησης και αναδεικνύουν την ανάγκη ανάπτυξης πιο ανθεκτικών συστημάτων υποβοήθησης για άτομα με προβλήματα όρασης.

ΚΕΦΑΛΑΙΟ 3 Προηγμένες Τεχνικές Μηχανικής Μάθησης

Η ραγδαία ανάπτυξη της τεχνητής νοημοσύνης (Artificial Intelligence) και ειδικότερα της μηχανικής μάθησης (Machine Learning) έχει οδηγήσει σε σημαντική πρόοδο στην αυτόματη επεξεργασία και κατανόηση της οπτικής πληροφορίας. Η εξέλιξη των τεχνικών βαθιάς μάθησης (Deep Learning) έχει επιτρέψει την ανάπτυξη συστημάτων ικανών να αναγνωρίζουν και να εντοπίζουν αντικείμενα, να αναλύουν σύνθετες οπτικές σκηνές και να εξάγουν πληροφορίες υψηλότερου σημασιολογικού επιπέδου από εικόνες. Παράλληλα, η ανάπτυξη μοντέλων που συνδυάζουν οπτική πληροφορία και φυσική γλώσσα έχει συμβάλει σημαντικά στην εξέλιξη εφαρμογών όπως η αυτόματη περιγραφή εικόνων (Image Captioning) και η ανίχνευση αντικειμένων βάσει λεκτικών εννοιών.

Οι δυνατότητες που προσφέρουν τα μοντέλα αυτά παρουσιάζουν ιδιαίτερο ενδιαφέρον για την ανάπτυξη υποστηρικτικών τεχνολογιών (Assistive Technologies), καθώς επιτρέπουν την επεξεργασία οπτικής πληροφορίας και την απόδοσή της σε μορφή περισσότερο προσβάσιμη στον τελικό χρήστη. Στην περίπτωση ατόμων με προβλήματα όρασης, ένα τέτοιο σύστημα μπορεί να επεξεργάζεται εικόνες από το περιβάλλον, να αναγνωρίζει σημαντικά αντικείμενα και στοιχεία της σκηνής και να μετατρέπει τα αποτελέσματα της οπτικής ανάλυσης σε λεκτική περιγραφή, η οποία στη συνέχεια μπορεί να αποδοθεί μέσω συνθετικής ομιλίας. Με τον τρόπο αυτό, τεχνικές Computer Vision και Natural Language Processing μπορούν να λειτουργήσουν συμπληρωματικά για τη δημιουργία ενός συστήματος που δεν περιορίζεται στην απλή αναγνώριση αντικειμένων, αλλά επιχειρεί να παρέχει μία περισσότερο ολοκληρωμένη και κατανοητή περιγραφή του οπτικού περιβάλλοντος.

Στο πλαίσιο της παρούσας εργασίας, η ολοκληρωμένη ανάλυση και περιγραφή μιας οπτικής σκηνής προϋποθέτει τον συνδυασμό διαφορετικών τεχνικών, οι οποίες επιτελούν διακριτούς αλλά συμπληρωματικούς ρόλους. Η ανίχνευση αντικειμένων (Object Detection) αξιοποιείται για την αναγνώριση και τον χωρικό εντοπισμό αντικειμένων, ενώ οι τεχνικές Image Captioning συμβάλλουν στην παραγωγή μιας συνολικότερης περιγραφής της σκηνής σε φυσική γλώσσα. Παράλληλα, οι τεχνικές Open-Vocabulary Object Detection επιτρέπουν μεγαλύτερη ευελιξία ως προς τις κατηγορίες που μπορεί να αναγνωρίζει το σύστημα, ενώ η τεχνολογία Text-to-Speech καθιστά δυνατή τη μετατροπή της τελικής λεκτικής περιγραφής σε ηχητική πληροφορία για τον χρήστη.

Η ανάλυση που ακολουθεί επικεντρώνεται στις τεχνικές που σχετίζονται άμεσα με τη σχεδίαση και την υλοποίηση του προτεινόμενου συστήματος. Αρχικά γίνεται αναφορά στις βασικές αρχές του Deep Learning και των συνελκτικών νευρωνικών δικτύων (Convolutional Neural Networks - CNNs), τα οποία αποτελούν βασικό υπόβαθρο για την επεξεργασία οπτικών δεδομένων. Στη συνέχεια εξετάζεται η ανίχνευση Αντικειμένων (Object Detection), με έμφαση στην οικογένεια μοντέλων YOLO και στο YOLO-World στο πλαίσιο του Open-Vocabulary Object Detection. Εξετάζονται επίσης το BLIP και γενικότερα τα Μοντέλα Όρασης-Γλώσσας (Vision-Language Models - VLMs), τα οποία συνδέουν την οπτική πληροφορία με τη φυσική γλώσσα. Τέλος, γίνεται αναφορά στην τεχνολογία Text-to-Speech, μέσω της οποίας η παραγόμενη λεκτική περιγραφή αποδίδεται σε μορφή ομιλίας. Παράλληλα, γίνεται αναφορά

στη χρήση pre-trained μοντέλων, τα οποία αξιοποιούνται στην εφαρμογή χωρίς να πραγματοποιείται εκ νέου εκπαίδευσή τους.

3.1 Deep Learning

Η βαθιά μάθηση (Deep Learning) αποτελεί εξειδικευμένο υποσύνολο της μηχανικής μάθησης και βασίζεται στη χρήση τεχνητών νευρωνικών δικτύων με πολλαπλά επίπεδα επεξεργασίας. Σε αντίθεση με παραδοσιακές προσεγγίσεις, στις οποίες σημαντικό μέρος των χαρακτηριστικών των δεδομένων καθορίζεται χειροκίνητα, τα μοντέλα Deep Learning έχουν τη δυνατότητα να μαθαίνουν αυτόματα κατάλληλες αναπαραστάσεις απευθείας από τα δεδομένα. Τα διαδοχικά επίπεδα ενός βαθιού νευρωνικού δικτύου επιτρέπουν την αναπαράσταση της πληροφορίας σε διαφορετικά επίπεδα αφαίρεσης, από απλά χαρακτηριστικά έως περισσότερο σύνθετες έννοιες.

Η διαδικασία μάθησης πραγματοποιείται μέσω της προσαρμογής των εσωτερικών παραμέτρων του δικτύου. Κατά την εκπαίδευση, η έξοδος του μοντέλου συγκρίνεται με την επιθυμητή έξοδο μέσω μιας συνάρτησης απώλειας (loss function). Στη συνέχεια, μέσω του αλγορίθμου backpropagation, υπολογίζεται ο τρόπος με τον οποίο πρέπει να μεταβληθούν τα βάρη (weights) του δικτύου ώστε να μειωθεί το σφάλμα. Η δυνατότητα αυτή έχει συμβάλει σημαντικά στη βελτίωση εργασιών όπως η αναγνώριση εικόνων, η ανίχνευση αντικειμένων, η επεξεργασία ομιλίας και η επεξεργασία φυσικής γλώσσας.

Ιδιαίτερη σημασία για την υπολογιστική όραση έχει η δυνατότητα αυτόματης εξαγωγής χαρακτηριστικών (automatic feature extraction). Μία εικόνα αποτελείται από μεγάλο αριθμό pixels, τα οποία από μόνα τους δεν παρέχουν άμεσα σημασιολογική πληροφορία σχετικά με το περιεχόμενο της σκηνής. Μέσω των διαδοχικών επιπέδων ενός νευρωνικού δικτύου μπορούν να εξαχθούν σταδιακά περισσότερο σύνθετες αναπαραστάσεις, επιτρέποντας τελικά την αναγνώριση αντικειμένων και οπτικών προτύπων.

Η βαθιά μάθηση συνδέεται άμεσα με την παρούσα εργασία, καθώς τόσο η ανίχνευση αντικειμένων όσο και η παραγωγή περιγραφών εικόνων βασίζονται σε βαθιά νευρωνικά δίκτυα. Οι εικόνες που λαμβάνονται από την κάμερα προέρχονται από πραγματικό περιβάλλον και τα αντικείμενα μπορεί να εμφανίζονται σε διαφορετικές θέσεις, μεγέθη, γωνίες θέασης και συνθήκες φωτισμού. Η χρήση Deep Learning επιτρέπει την αξιοποίηση μοντέλων που έχουν μάθει σύνθετες οπτικές αναπαραστάσεις και μπορούν να χρησιμοποιηθούν για την ανάλυση τέτοιων σκηνών.

Στην παρούσα υλοποίηση δεν πραγματοποιείται εκπαίδευση των βαθιών νευρωνικών δικτύων από την αρχή. Αντίθετα, αξιοποιούνται pre-trained μοντέλα, τα οποία εφαρμόζονται στις εικόνες που λαμβάνονται από την κάμερα για την εξαγωγή των αντίστοιχων αποτελεσμάτων. Με αυτόν τον τρόπο, η υλοποίηση επικεντρώνεται κυρίως στον συνδυασμό των διαφορετικών

μοντέλων και στην επεξεργασία των αποτελεσμάτων τους, ώστε να προκύπτει η τελική περιγραφή της οπτικής σκηνής.

3.2 Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs)

Τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNNs) αποτελούν μία από τις βασικές αρχιτεκτονικές βαθιάς μάθησης για την επεξεργασία και ανάλυση εικόνων. Η χρήση τους έχει συνδεθεί με σημαντική πρόοδο σε διάφορες εργασίες υπολογιστικής όρασης, μεταξύ των οποίων η ταξινόμηση εικόνων (Image Classification), η ανίχνευση αντικειμένων (Object Detection) και η σημασιολογική κατάτμηση εικόνων (Semantic Segmentation). Παρά την ανάπτυξη νεότερων αρχιτεκτονικών, όπως οι Vision Transformers, τα συνελικτικά δίκτυα εξακολουθούν να χρησιμοποιούνται ευρέως και σύγχρονες αρχιτεκτονικές τους παρουσιάζουν υψηλή απόδοση σε εργασίες αναγνώρισης και ανίχνευσης αντικειμένων.

Βασικό χαρακτηριστικό των CNNs είναι η χρήση συνελικτικών επιπέδων (convolutional layers) στα οποία πραγματοποιείται η πράξη της συνέλιξης. Στα επίπεδα αυτά, φίλτρα ή πυρήνες (filters/kernels) εφαρμόζονται σε τοπικές περιοχές των δεδομένων εισόδου με σκοπό την εξαγωγή χαρακτηριστικών. Από τη διαδικασία παράγονται χάρτες χαρακτηριστικών (feature maps), οι οποίοι περιέχουν πληροφορία σχετικά με τα χαρακτηριστικά που ανιχνεύονται στις διαφορετικές περιοχές της εικόνας.

Η επεξεργασία της εικόνας πραγματοποιείται σταδιακά μέσα από διαδοχικά επίπεδα του δικτύου. Στα πρώτα επίπεδα συνήθως εντοπίζονται σχετικά απλά χαρακτηριστικά, όπως ακμές και βασικά οπτικά μοτίβα, ενώ στα βαθύτερα επίπεδα οι πληροφορίες αυτές συνδυάζονται για τη δημιουργία πιο σύνθετων αναπαραστάσεων του οπτικού περιεχομένου. Με αυτόν τον τρόπο, το δίκτυο μπορεί να μαθαίνει κατά την εκπαίδευσή του χαρακτηριστικά που είναι χρήσιμα για την εκάστοτε εργασία.

Στην παρούσα εργασία, τα συνελικτικά νευρωνικά Δίκτυα σχετίζονται κυρίως με το στάδιο της ανίχνευσης αντικειμένων. Για την επεξεργασία των εικόνων που λαμβάνονται από την κάμερα χρησιμοποιείται, μεταξύ άλλων, ένα pre-trained μοντέλο YOLOv8, το οποίο αξιοποιεί συνελικτικές δομές για την εξαγωγή και επεξεργασία οπτικών χαρακτηριστικών. Μέσω της ανίχνευσης αντικειμένων προσδιορίζονται οι κατηγορίες και οι θέσεις των αντικειμένων στην εικόνα. Οι πληροφορίες αυτές αξιοποιούνται στη συνέχεια κατά τη δημιουργία της τελικής περιγραφής της σκηνής.

3.3 YOLO

Η οικογένεια μοντέλων YOLO (You Only Look Once) αποτελεί μία από τις πλέον διαδεδομένες προσεγγίσεις για την ανίχνευση αντικειμένων, ιδιαίτερα σε εφαρμογές όπου απαιτείται γρήγορη επεξεργασία εικόνων ή βίντεο. Η αρχική προσέγγιση YOLO παρουσιάστηκε από τους Redmon et al. το 2016 και αντιμετώπισε το Object Detection ως ένα ενιαίο πρόβλημα πρόβλεψης. Αντί να εφαρμόζονται διαδοχικά ανεξάρτητα στάδια για τον εντοπισμό πιθανών περιοχών ενδιαφέροντος και την ταξινόμησή τους, ένα ενιαίο νευρωνικό δίκτυο επεξεργάζεται την εικόνα και προβλέπει τις θέσεις των αντικειμένων μαζί με τις αντίστοιχες κατηγορίες τους.

Η βασική λειτουργία ενός μοντέλου YOLO περιλαμβάνει την εξαγωγή οπτικών χαρακτηριστικών από την εικόνα και την αξιοποίησή τους για την πρόβλεψη των αντικειμένων που εμφανίζονται σε αυτή. Για κάθε ανίχνευση παράγεται ένα πλαίσιο οριοθέτησης (bounding box), το οποίο προσδιορίζει τη θέση του αντικειμένου, καθώς και μία εκτίμηση της κατηγορίας στην οποία ανήκει. Παράλληλα, η έξοδος συνοδεύεται από τιμή εμπιστοσύνης (confidence score), η οποία εκφράζει τη βεβαιότητα του μοντέλου για τη συγκεκριμένη ανίχνευση. Η ταυτόχρονη εκτίμηση θέσης και κατηγορίας επιτρέπει στο YOLO να πραγματοποιεί Object Detection με μία ενιαία διαδικασία επεξεργασίας της εικόνας.

Η αρχιτεκτονική των σύγχρονων μοντέλων YOLO μπορεί γενικά να διακριθεί σε τρία βασικά τμήματα: το backbone, το neck και το detection head. Το backbone είναι υπεύθυνο για την εξαγωγή χαρακτηριστικών από την αρχική εικόνα. Στη συνέχεια, το neck συνδυάζει πληροφορία από διαφορετικές κλίμακες χαρακτηριστικών, ώστε το δίκτυο να μπορεί να επεξεργάζεται αντικείμενα διαφορετικού μεγέθους. Τέλος, το detection head χρησιμοποιεί τις παραγόμενες αναπαραστάσεις για την πρόβλεψη των bounding boxes και των αντίστοιχων κατηγοριών. Η συγκεκριμένη δομή έχει εξελιχθεί σημαντικά στις διαδοχικές εκδόσεις της οικογένειας YOLO, με στόχο τη βελτίωση της ακρίβειας και της υπολογιστικής αποδοτικότητάς.

Στην παρούσα εργασία χρησιμοποιείται το YOLOv8x, δηλαδή η έκδοση x της οικογένειας YOLOv8, μέσω της βιβλιοθήκης Ultralytics. Το YOLOv8x χρησιμοποιεί συνελκτικές δομές για την εξαγωγή χαρακτηριστικών και διαθέτει αρχιτεκτονική τύπου backbone-neck-head. Σε σχέση με προηγούμενες εκδόσεις της οικογένειας, η αρχιτεκτονική του περιλαμβάνει, μεταξύ άλλων, τα δομικά στοιχεία C2f και έναν anchor-free detection head, στον οποίο η πρόβλεψη των αντικειμένων δεν βασίζεται σε προκαθορισμένα anchor-boxes. Η επιλογή του YOLOv8x στην παρούσα εφαρμογή εξυπηρετεί την ανάγκη εντοπισμού πολλαπλών αντικειμένων μέσα στις εικόνες που λαμβάνονται από την κάμερα, παρέχοντας τόσο την κατηγορία τους όσο και τη θέση τους στο οπτικό πεδίο.

Το pre-trained μοντέλο YOLOv8x εφαρμόζεται στις εικόνες που λαμβάνονται από την κάμερα και τα αποτελέσματα της ανίχνευσης αξιοποιούνται στη συνέχεια για τη δημιουργία της τελικής περιγραφής της σκηνής. Για τον περιορισμό λιγότερο αξιόπιστων ή μη χρήσιμων αποτελεσμάτων, εφαρμόζεται φιλτράρισμα των ανιχνεύσεων με βάση τις επιλεγμένες κατηγορίες και τις τιμές εμπιστοσύνης. Οι επιμέρους παράμετροι και ο ακριβής τρόπος επεξεργασίας των αποτελεσμάτων παρουσιάζονται αναλυτικότερα στο Κεφάλαιο 5.

Παρότι το YOLOv8x επιτρέπει την αποτελεσματική ανίχνευση αντικειμένων, οι κατηγορίες που μπορεί να αναγνωρίσει καθορίζονται από το σύνολο κατηγοριών πάνω στο οποίο έχει εκπαιδευτεί. Για την επέκταση της διαδικασίας ανίχνευσης σε πρόσθετες κατηγορίες αντικειμένων, χρησιμοποιείται παράλληλα το YOLO-World, το οποίο υποστηρίζει OpenVocabulary Object Detection και επιτρέπει τον καθορισμό των επιθυμητών κατηγοριών μέσω κειμένου. Η λειτουργία και ο ρόλος του μοντέλου αυτού εξετάζονται στην επόμενη ενότητα.

3.4 YOLO-World / Open-Vocabulary Detection

Τα συμβατικά μοντέλα ανίχνευσης αντικειμένων λειτουργούν με ένα προκαθορισμένο σύνολο κατηγοριών, το οποίο καθορίζεται κατά την εκπαίδευσή τους. Ο περιορισμός αυτός μπορεί να μειώσει την ευελιξία τους σε εφαρμογές πραγματικού περιβάλλοντος, όπου ενδέχεται να εμφανίζονται αντικείμενα που δεν περιλαμβάνονται στις αρχικές κατηγορίες εκπαίδευσης. Η προσέγγιση Open-Vocabulary Object Detection επιχειρεί να αντιμετωπίσει αυτόν τον περιορισμό, συνδέοντας την οπτική πληροφορία με λεκτικές έννοιες και επιτρέποντας την ανίχνευση ενός ευρύτερου συνόλου κατηγοριών.

Το YOLO-World παρουσιάστηκε από τους Cheng et al. Το 2024 ως προσέγγιση OpenVocabulary Object Detection, συνδυάζοντας την ανίχνευση αντικειμένων της οικογένειας YOLO με πληροφορία φυσικής γλώσσας. Η αρχιτεκτονική του περιλαμβάνει το Reparameterizable Vision-Language Path Aggregation Network (RepVL-PAN), μέσω του οποίου συνδυάζονται οπτικά και γλωσσικά χαρακτηριστικά. Κατά την εκπαίδευσή του χρησιμοποιείται επίσης region-text contrastive loss, με σκοπό τη συσχέτιση περιοχών της εικόνας με τις αντίστοιχες λεκτικές έννοιες.

Βασικό χαρακτηριστικό του YOLO-World είναι η δυνατότητα καθορισμού κατηγοριών αντικειμένου μέσω κειμένου. Έτσι η διαδικασία ανίχνευσης μπορεί να προσαρμόζεται στις ανάγκες της εκάστοτε εφαρμογής, χωρίς να απαιτείται εκ νέου εκπαίδευση του μοντέλου για κάθε διαφορετικό σύνολο κατηγοριών. Στην παρούσα εργασία χρησιμοποιείται το pre-trained μοντέλο YOLOv8x-Worldv2, παράλληλα με το YOLOv8x. Ο ρόλος του είναι να διευρύνει τις κατηγορίες αντικειμένων που μπορούν να αναζητηθούν στις εικόνες που λαμβάνονται από την κάμερα. Τα αποτελέσματα των δύο μοντέλων αξιοποιούνται συνδυαστικά για τη δημιουργία μιας πληρέστερης περιγραφής της οπτικής σκηνής, ενώ οι επιλεγμένες κατηγορίες και ο τρόπος συνδυασμού των αποτελεσμάτων παρουσιάζονται αναλυτικότερα στο Κεφάλαιο 5.

3.5 BLIP

Η αυτόματη περιγραφή εικόνων (Image Captioning) αποτελεί εργασία που συνδυάζει την επεξεργασία οπτικής πληροφορίας με την παραγωγή φυσικής γλώσσας. Σε αντίθεση με την ανίχνευση αντικειμένων, η οποία παρέχει πληροφορίες σχετικά με τις κατηγορίες και τις θέσεις επιμέρους αντικειμένων, το Image Captioning αποσκοπεί στην παραγωγή μιας λεκτικής

περιγραφής του συνολικού περιεχομένου μιας εικόνας. Με αυτόν τον τρόπο μπορεί να αποδώσει πληροφορίες για τη σκηνή και τις σχέσεις μεταξύ των στοιχείων που εμφανίζονται σε αυτή.

Το BLIP (Bootstrapping Language-Image Pre-Training) παρουσιάστηκε από τους Li et al. Το 2022 ως ένα Vision-Language Pre-Training πλαίσιο σχεδιασμένο ώστε να υποστηρίζει τόσο εργασίες κατανόησης όσο και παραγωγής περιεχομένου. Η αρχιτεκτονική του βασίζεται στο Multimodal Mixture of Encoder-Decoder (MED), το οποίο επιτρέπει την επεξεργασία και σύνδεση οπτικής και γλωσσικής πληροφορίας. Παράλληλα, οι συγγραφείς πρότειναν την μέθοδο Captioning and Filtering (CapFilt), κατά την οποία παράγονται νέες λεζάντες για εικόνες και φιλτράρονται ζεύγη εικόνας-κειμένου με θόρυβο, με στόχο τη βελτίωση των δεδομένων που χρησιμοποιούνται κατά το pre-training.

Μία από τις εφαρμογές του BLIP είναι το Image Captioning, όπου μία εικόνα δίνεται ως είσοδος και το μοντέλο παράγει μία περιγραφή της σε φυσική γλώσσα. Σε αντίθεση με ένα μοντέλο Object Detection, η έξοδος δεν αποτελείται από ξεχωριστές κατηγορίες και πλαίσια οριοθέτησης, αλλά από μία πρόταση που επιχειρεί να αποδώσει το περιεχόμενο της εικόνας.

Στην παρούσα εργασία χρησιμοποιείται το pre-trained μοντέλο Salesforce/blip-imagecaptioning-base για την παραγωγή περιγραφής των εικόνων που λαμβάνονται από την κάμερα. Η περιγραφή που παράγεται από το BLIP χρησιμοποιείται συμπληρωματικά με τα αποτελέσματα της ανίχνευσης αντικειμένων από τα YOLOv8x και YOLO-World. Με αυτόν τον τρόπο, η πληροφορία για τα αντικείμενα και τις θέσεις τους συνδυάζεται με μία συνολικότερη περιγραφή του περιεχομένου της εικόνας, συμβάλλοντας στη δημιουργία της τελικής λεκτικής περιγραφής της σκηνής.

3.6 Μοντέλα Όρασης-Γλώσσας (Vision-Language Models - VLMs)

Τα Μοντέλα Όρασης-Γλώσσας αποτελούν κατηγορία μοντέλων που συνδυάζουν την επεξεργασία οπτικής πληροφορίας με τη φυσική γλώσσα. Σε αντίθεση με μοντέλα που επεξεργάζονται αποκλειστικά εικόνες ή κείμενο, τα VLMs συνδέουν τις δύο μορφές πληροφορίας, ώστε οπτικά χαρακτηριστικά να μπορούν να συσχετιστούν με λέξεις και λεκτικές έννοιες. Η προσέγγιση αυτή χρησιμοποιείται σε εργασίες όπως η αντιστοίχιση εικόνας και κειμένου, η αυτόματη περιγραφή εικόνων και η ανίχνευση αντικειμένων με βάση λεκτικές κατηγορίες.

Η λειτουργία των VLMs βασίζεται στην εκμάθηση σχέσεων μεταξύ οπτικών και γλωσσικών αναπαραστάσεων. Ένα χαρακτηριστικό παράδειγμα είναι το CLIP, το οποίο εκπαιδεύτηκε σε μεγάλο αριθμό ζευγών εικόνας και κειμένου, μαθαίνοντας να συσχετίζει το οπτικό περιεχόμενο μιας εικόνας με αντίστοιχες λεκτικές περιγραφές. Με αυτόν τον τρόπο, η πληροφορία που περιέχεται σε μία εικόνα μπορεί να συσχετιστεί με έννοιες που εκφράζονται μέσω φυσικής γλώσσας.

Στην παρούσα εργασία, η σύνδεση οπτικής πληροφορίας και γλώσσας αξιοποιείται με διαφορετικό τρόπο από το YOLO-World και το BLIP. Το YOLO-World χρησιμοποιεί γλωσσική πληροφορία για την ανίχνευση αντικειμένων, ενώ το BLIP χρησιμοποιείται για την παραγωγή περιγραφής της εικόνας σε φυσική γλώσσα. Τα δύο μοντέλα επιτελούν διαφορετικούς αλλά συμπληρωματικούς ρόλους, καθώς τα αποτελέσματά τους αξιοποιούνται για τη δημιουργία της τελικής λεκτικής περιγραφής της οπτικής σκηνής.

3.7 Text-to-Speech (TTS)

Η τεχνολογία Text-to-Speech, γνωστή και ως σύνθεση ομιλίας (speech synthesis), αφορά τη μετατροπή γραπτού κειμένου σε συνθετική ομιλία. Ένα σύστημα TTS δέχεται ως είσοδο κείμενο και παράγει ως έξοδο το αντίστοιχο ηχητικό σήμα, με στόχο η παραγόμενη ομιλία να είναι κατανοητή και όσον το δυνατό πιο φυσική. Η εξέλιξη των τεχνικών βαθιάς μάθησης έχει συμβάλλει σημαντικά στη βελτίωση της ποιότητας των σύγχρονων συστημάτων σύνθεσης ομιλίας.

Σε ένα σύστημα TTS, το κείμενο υφίσταται αρχικά επεξεργασία ώστε να προετοιμαστεί για τη σύνθεση της ομιλίας. Στα σύγχρονα νευρωνικά συστήματα, η επεξεργασμένη γλωσσική πληροφορία χρησιμοποιείται για την παραγωγή μιας ακουστικής αναπαράστασης, η οποία στη συνέχεια μετατρέπεται σε ηχητικό σήμα μέσω ενός vocoder. Η ποιότητα ενός συστήματος TTS σχετίζεται, μεταξύ άλλων, με την καταληπτότητα και τη φυσικότητα της παραγόμενης ομιλίας.

Η σύνθεση ομιλίας είναι ιδιαίτερα χρήσιμη σε εφαρμογές προσβασιμότητας, καθώς επιτρέπει την απόδοση πληροφορίας χωρίς να απαιτείται η οπτική ανάγνωση κειμένου. Στην περίπτωση ενός συστήματος υποβοήθησης ατόμων με προβλήματα όρασης, η λεκτική πληροφορία που προκύπτει από την ανάλυση μιας εικόνας μπορεί να μετατραπεί σε ομιλία, ώστε να μεταφερθεί στον χρήστη μέσω ακουστικής εξόδου.

Στην παρούσα εργασία, το TTS αποτελεί το τελικό στάδιο της επεξεργασίας. Μετά τον συνδυασμό των αποτελεσμάτων ανίχνευσης αντικειμένων και της αυτόματης περιγραφής εικόνας, η τελική λεκτική περιγραφή μετατρέπεται σε ομιλία μέσω της βιβλιοθήκης pyttsx3. Η συγκεκριμένη βιβλιοθήκη λειτουργεί τοπικά, αξιοποιώντας μηχανές σύνθεσης ομιλίας που είναι διαθέσιμες στο λειτουργικό σύστημα, χωρίς να απαιτεί σύνδεση στο διαδίκτυο για την μετατροπή κειμένου σε ομιλία.

ΚΕΦΑΛΑΙΟ 4 Συστήματα Υποβοήθησης Ανθρώπων με Προβλήματα Όρασης

Η αξιοποίηση της υπολογιστικής όρασης σε συστήματα υποβοήθησης ατόμων με προβλήματα όρασης έχει διευρύνει τις δυνατότητες πρόσβασης σε πληροφορίες του φυσικού περιβάλλοντος. Σύγχρονες λύσεις μπορούν να υποστηρίξουν πολλές διαφορετικές λειτουργίες, όπως αναγνώριση αντικειμένων, περιγραφή σκηνών, ανάγνωση κειμένου και υποστήριξη της πλοήγησης. Ωστόσο, η αποτελεσματικότητα ενός συστήματος υποβοήθησης δεν εξαρτάται αποκλειστικά από την ικανότητά του να επεξεργάζεται σωστά μία εικόνα. Εξίσου σημαντικός είναι ο τρόπος με τον οποίο επιλέγεται η πληροφορία, οργανώνεται και τελικά παρουσιάζεται στον χρήστη.

Για τον λόγο αυτό, για τον σχεδιασμό τέτοιων συστημάτων απαιτείται να λαμβάνονται υπόψη τόσο οι δυνατότητες της χρησιμοποιούμενης τεχνολογίας όσο και οι συνθήκες στις οποίες πρόκειται να χρησιμοποιηθεί. Πρόσφατες μελέτες τονίζουν τη σημασία της σύνδεσης των τεχνικών χαρακτηριστικών ενός συστήματος με τις πραγματικές ανάγκες των χρηστών, ιδιαίτερα σε εφαρμογές που έχουν σκοπό να παρέχουν πληροφορίες για το περιβάλλον και το περιεχόμενο μιας σκηνής.

4.1 Σχεδιαστικές απαιτήσεις συστημάτων οπτικής υποβοήθησης

Η σχεδίαση ενός συστήματος οπτικής υποβοήθησης οφείλει να ξεκινά από τον τρόπο με τον οποίο το σύστημα πρόκειται να χρησιμοποιηθεί και όχι μόνο από τις τεχνικές δυνατότητες των επιμέρους μοντέλων. Μία εφαρμογή που λειτουργεί ικανοποιητικά σε ελεγχόμενες συνθήκες δεν είναι απαραίτητο ότι θα είναι εξίσου λειτουργική σε πραγματικά περιβάλλοντα. Για τον λόγο αυτό, παράγοντες όπως η ευκολία αλληλεπίδρασης, ο χρόνος απόκρισης και η καταλληλότητα της παρεχόμενης πληροφορίας αποτελούν σημαντικές παραμέτρους κατά τον σχεδιασμό.

Ιδιαίτερη σημασία έχει η προσέγγιση με επίκεντρο τον χρήστη (user-centered approach). Κατά την ανάπτυξη συστημάτων κατανόησης σκηνής είναι σημαντικό να εξετάζεται όχι μόνο το τεχνικό πρόβλημα που επιλύεται, αλλά και ο τρόπος με τον οποίο η αντίστοιχη λειτουργία ανταποκρίνεται στις ανάγκες του χρήστη. Ένα σύστημα περιγραφής περιβάλλοντος, για παράδειγμα, χρειάζεται να παρέχει πληροφορίες που μπορούν να γίνουν άμεσα αντιληπτές, χωρίς να απαιτεί σύνθετη αλληλεπίδραση για την απόκτησή τους.

Η ταχύτητα απόκρισης αποτελεί επίσης σημαντικό χαρακτηριστικό, ιδιαίτερα όταν η πληροφορία αφορά το άμεσο περιβάλλον του χρήστη. Μεγάλη καθυστέρηση μεταξύ της λήψης μιας εικόνας και της απόδοσης της αντίστοιχης πληροφορίας μπορεί να μειώσει τη χρησιμότητα της περιγραφής, καθώς στο μεταξύ η σκηνή μπορεί να έχει μεταβληθεί. Ως αποτέλεσμα η υπολογιστική αποδοτικότητα και η κατάλληλη επιλογή των λειτουργιών που

εκτελούνται σε κάθε ανάλυση αποκτούν μεγάλη σημασία σε ένα σύστημα που προορίζεται για καθημερινή χρήση.

4.2 Επιλογή και ιεράρχηση της πληροφορίας

Η δυνατότητα ενός συστήματος να εξάγει μεγάλο αριθμό πληροφοριών από μία εικόνα δεν σημαίνει ότι όλες οι πληροφορίες έχουν την ίδια χρησιμότητα για τον τελικό χρήστη. Μία σύνθετη σκηνή μπορεί να περιλαμβάνει πολλά αντικείμενα και επιμέρους στοιχεία, όμως η διαδοχική αναφορά όλων αυτών μπορεί να οδηγήσει σε μία εκτενή περιγραφή, στην οποία η σημαντικότερη πληροφορία δεν γίνεται εύκολα αντιληπτή. Επομένως, η επιλογή του περιεχομένου αποτελεί ξεχωριστό ζήτημα από την ίδια την αναγνώριση των στοιχείων μιας εικόνας.

Η σημασία του πλαισίου χρήσης έχει επισημανθεί σε μελέτες σχετικά με την προσβασιμότητα των περιγραφών εικόνων. Η αξιολόγηση μιας περιγραφής επηρεάζεται από το κατά πόσο οι πληροφορίες που περιλαμβάνει είναι σχετικές με το πλαίσιο στο οποίο εμφανίζεται η εικόνα. Επομένως, μία περιγραφή δεν είναι απαραίτητα περισσότερο χρήσιμη επειδή περιέχει περισσότερες λεπτομέρειες. Εκτός από την ορθότητα της περιγραφής του οπτικού περιεχομένου, σημαντική είναι και η συνάφεια των πληροφοριών που επιλέγονται.

Σε ένα σύστημα περιγραφής του περιβάλλοντος για άτομα με προβλήματα όρασης, διαφορετικά είδη πληροφορίας μπορούν να συμβάλουν στην κατανόηση μιας σκηνής. Η αναγνώριση σημαντικών αντικειμένων παρέχει πληροφορίες για τα στοιχεία που υπάρχουν στην εικόνα, ενώ η σχετική θέση τους βοηθά στον προσδιορισμό της διάταξής τους. Παράλληλα, μία γενικότερη περιγραφή μπορεί να αποδώσει το συνολικό περιεχόμενο της σκηνής. Ο συνδυασμός αυτών των πληροφοριών επιτρέπει τη δημιουργία μιας πιο συνεκτικής περιγραφής σε σχέση με την απλή παράθεση των αντικειμένων που έχουν αναγνωριστεί.

Η λογική αυτή ακολουθείται και στο σύστημα που αναπτύχθηκε στην παρούσα εργασία. Τα αποτελέσματα της ανίχνευσης αντικειμένων και της αυτόματης περιγραφής εικόνας συνδυάζονται πριν από τη δημιουργία της τελικής λεκτικής περιγραφής, με στόχο να αποφεύγονται περιττές επαναλήψεις και να οργανώνεται καλύτερα η πληροφορία που μεταδίδεται στον χρήστη.

4.3 Προσβάσιμη Απόδοση της Περιγραφής και Αλληλεπίδραση με τον Χρήστη

Μετά την επιλογή της κατάλληλης πληροφορίας, ιδιαίτερη σημασία έχει ο τρόπος με τον οποίο αυτή παρουσιάζεται στον χρήστη. Στα συστήματα περιγραφής σκηνής, η ακουστική απόδοση επιτρέπει την πρόσβαση στο αποτέλεσμα χωρίς να απαιτείται οπτική αλληλεπίδραση με την εφαρμογή.

Η λεκτική περιγραφή χρειάζεται να είναι σαφής και να αποδίδει την απαιτούμενη πληροφορία χωρίς περιττές επαναλήψεις. Η υπερβολικά αναλυτική παρουσίαση μπορεί να αυξήσει τον χρόνο που απαιτείται μέχρι ο χρήστης να ακούσει την πληροφορία που τον ενδιαφέρει. Αντίθετα, μία υπερβολικά περιορισμένη περιγραφή ενδέχεται να παραλείπει στοιχεία που είναι χρήσιμα για την κατανόηση της σκηνής. Συνεπώς, απαιτείται να διατηρείται ισορροπία μεταξύ πληρότητας και συντομίας, η οποία εξαρτάται και από το πλαίσιο χρήσης.

Η αλληλεπίδραση με το σύστημα πρέπει επίσης να παραμένει όσο το δυνατόν απλή. Σε εφαρμογές όπου ο χρήστης ζητά πληροφορίες για μία σκηνή, η διαδικασία καταγραφής της σκηνής, ενεργοποίησης της ανάλυσης και λήψης της περιγραφής δεν θα πρέπει να απαιτεί μεγάλο αριθμό ενεργειών. Παράλληλα, η συνεχής επανάληψη της ίδιας ή παρόμοιας ηχητικής πληροφορίας μπορεί να δυσχεράνει τη χρήση αντί να την υποστηρίζει. Ο σχεδιασμός της αλληλεπίδρασης συνδέεται επομένως με το πότε ενεργοποιείται η ανάλυση και με το πώς παρουσιάζεται το αποτέλεσμα της.

Η σημασία αυτών των παραμέτρων φαίνεται και σε πρόσφατες μελέτες πραγματικής χρήσης εφαρμογών περιγραφής σκηνής. Οι Gonzalez et al. μελέτησαν για δύο εβδομάδες τη χρήση μιας εφαρμογής αυτόματης περιγραφής από 16 άτομα με τύφλωση ή μειωμένη όραση και διαπίστωσαν ότι οι χρήστες αξιοποιούσαν τις περιγραφές για διαφορετικούς καθημερινούς σκοπούς. Ταυτόχρονα, οι αξιολογήσεις τους ανέδειξαν περιθώρια βελτίωσης τόσο ως προς την ικανοποίηση από τις περιγραφές όσο και ως προς την εμπιστοσύνη σε αυτές. Τα αποτελέσματα αυτά υπογραμμίζουν ότι η τεχνική δυνατότητα παραγωγής μιας περιγραφής δεν συνεπάγεται αυτομάτως και αποτελεσματική εμπειρία χρήσης.

4.4 Αξιοπιστία της Παραγόμενης Πληροφορίας

Η αξιοπιστία αποτελεί ιδιαίτερα σημαντική παράμετρο σε ένα σύστημα που μετατρέπει αυτόματα τα αποτελέσματα οπτικής ανάλυσης σε πληροφορία για τον χρήστη. Τα μοντέλα υπολογιστικής όρασης λειτουργούν πιθανοτικά και οι προβλέψεις τους δεν είναι πάντοτε ορθές. Μπορεί, για παράδειγμα, να ανιχνευθεί ένα αντικείμενο που δεν υπάρχει στην εικόνα (false positive) ή να μην αναγνωριστεί ένα αντικείμενο που υπάρχει (false negative). Αντίστοιχα, ένα σύστημα παραγωγής περιγραφών μπορεί να παραλείψει στοιχεία της σκηνής ή να παράγει πληροφορία που δεν ανταποκρίνεται πλήρως στο οπτικό περιεχόμενο.

Η σημασία αυτών των σφαλμάτων είναι κρίσιμη σε ένα σύστημα οπτικής υποβοήθησης. Σε μία συμβατική εφαρμογή ανίχνευσης αντικειμένων, μία εσφαλμένη πρόβλεψη αποτελεί κυρίως σφάλμα στην έξοδο του μοντέλου. Όταν όμως το αποτέλεσμα μετατρέπεται αυτόματα σε προφορική πληροφορία, το ίδιο σφάλμα μπορεί να παρουσιαστεί στον χρήστη ως μέρος της περιγραφής του πραγματικού περιβάλλοντος. Για τον λόγο αυτό, η αξιολόγηση ενός τέτοιου συστήματος δεν πρέπει να περιορίζεται μόνο στο αν παράγεται τελικά μία περιγραφή, αλλά να εξετάζει και κατά πόσο το περιεχόμενό της αντιστοιχεί στην πραγματική σκηνή.

Για τον περιορισμό των λανθασμένων αποτελεσμάτων μπορούν να εφαρμοστούν μηχανισμοί ελέγχου πριν από τη δημιουργία της τελικής περιγραφής, όπως η χρήση κατωφλίων

εμπιστοσύνης και το φιλτράρισμα των ανιχνεύσεων. Οι μηχανισμοί αυτοί δεν εξαλείφουν την πιθανότητα σφάλματος, μπορούν όμως να περιορίσουν την ένταξη λιγότερο αξιόπιστων αποτελεσμάτων στην πληροφορία που τελικά μεταδίδεται στον χρήστη.

ΚΕΦΑΛΑΙΟ 5 Προτεινόμενο Σύστημα

5.1 Επισκόπηση Συστήματος

Το προτεινόμενο σύστημα προτείνει ένα ολοκληρωμένο σύστημα υποβοήθησης ατόμων με προβλήματα όρασης. Δέχεται εικόνες που λαμβάνονται από ζωντανή κάμερα στον υπολογιστή, τις αναλύει και παρέχει μια περιγραφή με βάση τα αποτελέσματά της, την οποία έπειτα μετατρέπει από οπτική πληροφορία σε ηχητική περιγραφή σε πραγματικό χρόνο. Στόχος του συστήματος δεν είναι μόνο η αναγνώριση των αντικειμένων που εμφανίζονται στο οπτικό πεδίο του χρήστη, αλλά η δημιουργία μιας περισσότερο ολοκληρωμένης περιγραφής της σκηνής, ώστε ο χρήστης να αποκτά καλύτερη αντίληψη του περιβάλλοντός του.

Η λειτουργία του συστήματος βασίζεται στον συνδυασμό τεχνικών υπολογιστικής όρασης, μοντέλων όρασης-γλώσσας και σύνθεσης ομιλίας, με τη βασική φιλοσοφία του να στηρίζεται στον συνδυασμό πολλαπλών εξειδικευμένων μοντέλων τεχνητής νοημοσύνης, καθένα από τα οποία επιτελεί διαφορετικό ρόλο. Η ανίχνευση αντικειμένων πραγματοποιείται μέσω των μοντέλων YOLOv8x και YOLO-World, η συνολική κατανόηση της εικόνας υποστηρίζεται από το μοντέλο BLIP, ενώ η τελική περιγραφή μετατρέπεται σε ομιλία μέσω τεχνολογίας Text-to-Speech.

Σε αντίθεση με προσεγγίσεις που βασίζονται αποκλειστικά σε ένα ενιαίο μοντέλο περιγραφής εικόνων, το προτεινόμενο σύστημα αξιοποιεί τη συμπληρωματική λειτουργία διαφορετικών τεχνολογιών. Με τον τρόπο αυτό επιδιώκεται μεγαλύτερη ακρίβεια στην αναγνώριση αντικειμένων, καλύτερη αξιοποίηση της χωρικής πληροφορίας και παραγωγή φυσικότερης περιγραφής.

Η εφαρμογή έχει υλοποιηθεί στην γλώσσα προγραμματισμού Python, χρησιμοποιώντας τη βιβλιοθήκη OpenCV για την επεξεργασία εικόνας και τη λήψη δεδομένων από την κάμερα, τη βιβλιοθήκη Ultralytics για τα μοντέλα YOLO, το μοντέλο Salesforce BLIP για Image Captioning και την βιβλιοθήκη pyttsx3 για την τοπική σύνθεση ομιλίας.

5.2 Ροή Λειτουργίας

Η αρχιτεκτονική του προτεινόμενου συστήματος οργανώνεται σε πέντε διαδοχικά στάδια επεξεργασίας, μέσω των οποίων η αρχική εικόνα μετατρέπεται σε τελική ηχητική πληροφορία.

Η διαδικασία ξεκινά με τη συνεχή λήψη εικόνων από την κάμερα του υπολογιστή. Κάθε εικόνα αναλύεται παράλληλα από δύο διαφορετικά μοντέλα ανίχνευσης αντικειμένων. Τα αποτελέσματα των δύο μοντέλων συγχωνεύονται, φιλτράρονται και αξιοποιούνται μαζί με την περιγραφή που δημιουργεί το BLIP. Τέλος, η παραγόμενη περιγραφή μετατρέπεται σε ομιλία, ώστε ο χρήστης να λαμβάνει άμεσα την πληροφορία.

Η επιλογή αυτής της πολυεπίπεδης αρχιτεκτονικής επιτρέπει στο σύστημα να αξιοποιεί τα πλεονεκτήματα κάθε μοντέλου ξεχωριστά, ενώ παράλληλα περιορίζει ορισμένους από τους περιορισμούς που θα παρουσίαζε η αποκλειστική χρήση ενός μόνο αλγορίθμου.

Η εργασία είναι ένα σύνολο από αρχεία που περιέχουν κώδικα σε python. Συγκεκριμένα, αποτελείται από:

- Το `main.py` αποτελεί το κεντρικό αρχείο εκτέλεσης της εφαρμογής και συντονίζει τη λειτουργία των επιμέρους τμημάτων του συστήματος. Είναι υπεύθυνο για τη λειτουργία της κάμερας, την ενεργοποίηση της διαδικασίας ανάλυσης, την ενημέρωση της γραφικής διεπαφής και την ηχητική απόδοση της τελικής περιγραφής.

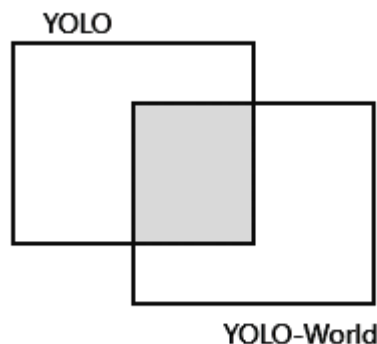
Οι βιβλιοθήκες που χρησιμοποιεί είναι οι `os`, `threading`, `datetime`, `cv2`, `tkinter`, `PIL`, `combined_detector` και `config`. Πιο συγκεκριμένα για το `config`, χρησιμοποιεί τις `camera_INDEX`, `USE_BLIP_CAPTION`, `SAVE_CAPTURED_FRAMES` και `CAPTURED_FRAMES_FOLDER`. Μέσα στη `main` σχηματίζεται όλο το GUI και δημιουργούνται όλες οι συναρτήσεις.

- Το `config.py`, το οποίο κρατάει όλες τις ρυθμίσεις, ποιό YOLO και YOLO-World χρησιμοποιείται, το γενικό confidence threshold και το IoU threshold.

Εκεί κρατάμε κυρίως αντικείμενα χρήσιμα για τον προσανατολισμό, την αποφυγή εμποδίων, την ενημέρωση για το περιβάλλον και γενικότερα τη βασική καθημερινή αλληλεπίδραση. Μπορούμε σε αυτό το σημείο να απαιτήσουμε μεγαλύτερη βεβαιότητα για κατηγορίες που παρουσιάζουν συχνά λανθασμένες ανιχνεύσεις.

- Το `combined_detector.py`, όπου συνδυάζει YOLO και YOLO-World.

Υπολογίζει το Intersection over Union (IoU) για να καθορίσει εάν και τα δύο, YOLO και YOLO-World, αναφέρονται στο ίδιο αντικείμενο και προχωράει αναλόγως.



[4 – YOLO – YOLO-World]

Τρέχει τα δύο YOLO, ενώνει τα αποτελέσματα, κανονικοποιεί τα ονόματα, βρίσκει εάν υπάρχουν διπλότυπα για να τα διαγράψει και κρατάει το μεγαλύτερο confidence. Η εικόνα [4] ([ChatGPT: Chat, Work, Create & Code with AI](#)) είναι AI Generated και δημιουργήθηκε για να εξηγήσει πώς λειτουργούν μαζί YOLO. Έστω ότι κάθε τετράγωνο αποτελεί τα YOLO και YOLO-World, τα οποία εφάπτονται και δημιουργούν ένα κοινό σχήμα με γκρι χρώμα. Αυτό το σχήμα δείχνει πως συνεργάζονται και στην εργασία. Το καθένα YOLO παράγει δικά του αποτελέσματα και στο τέλος βρίσκουν τα κοινά τους και τα συνδυάζουν.

- Το `scene_description.py`, το οποίο παράγει τη περιγραφή της εικόνας.

Δίνει προτεραιότητα σε αντικείμενα που μπορεί να είναι περισσότερο χρήσιμα για τον χρήστη και εκτιμά το πόσο μεγάλα εμφανίζονται μέσα στην εικόνα.

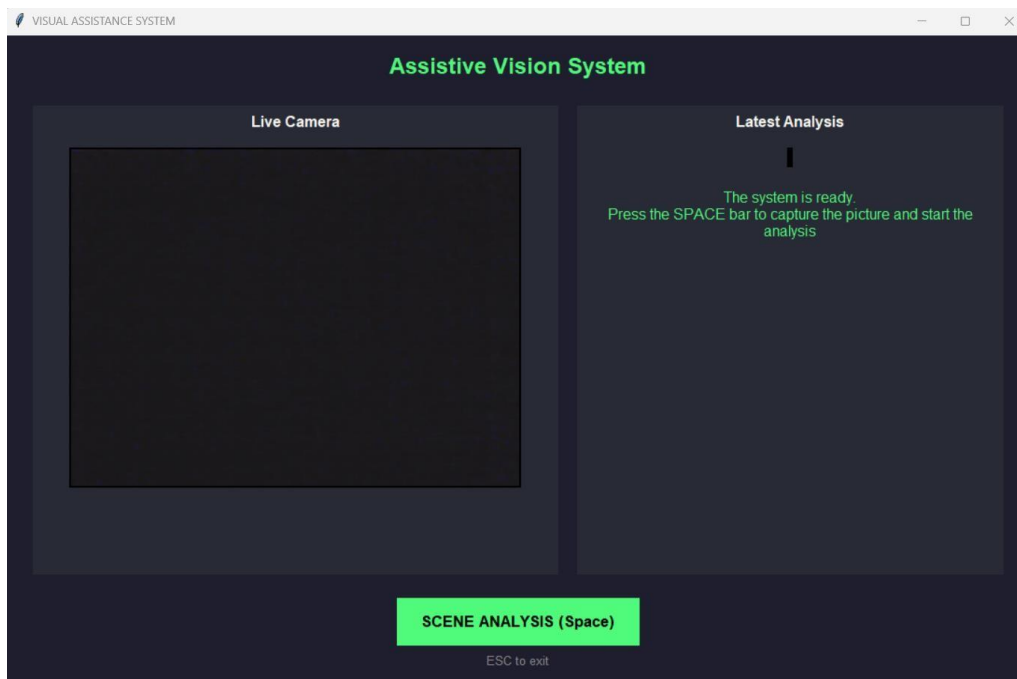
- Το `image_caption.py`, για να λαμβάνει τη περιγραφή από το BLIP.
- Το `tts_local.py`, για να μετατρέπει το κείμενο σε ομιλία.
- Το `utils.py` περιλαμβάνει βοηθητικές συναρτήσεις για την οπτική απεικόνιση των αποτελεσμάτων της ανίχνευσης. Χρησιμοποιείται για την εμφάνιση των bounding boxes, της κατηγορίας του αντικειμένου, του βαθμού εμπιστοσύνης και της πηγής της αντίστοιχης ανίχνευσης πάνω στην εικόνα.
- Σχεδιάζει τα bounding boxes, τα labels, το confidence και τη πηγή του μοντέλου.

5.3 Περιγραφή των επιμέρους υποσυστημάτων

5.3.1 Λήψη Εικόνας

Η διαδικασία ξεκινά με το άνοιγμα του GUI περιβάλλοντος της εργασίας. Η παρακάτω εικόνα [5] δείχνει ακριβώς τι εμφανίζεται στην οθόνη του χρήστη όταν ξεκινήσει να τρέχει το πρόγραμμα, το περιβάλλον που ανοίγει και οι περιγραφές για το πώς να το χρησιμοποιήσει ο χρήστης. Η επιλογή των χρωμάτων που χρησιμοποιούνται στο γραφικό περιβάλλον της εφαρμογής δεν έγινε τυχαία αλλά βασίστηκε σε κριτήρια ευκρίνειας και λειτουργικότητας. Στόχος ήταν οι πληροφορίες που εμφανίζονται στην οθόνη να διακρίνονται άμεσα από το περιεχόμενο της εικόνας, διευκολύνοντας την παρακολούθηση της λειτουργίας του συστήματος. Στο αριστερό πλαίσιο ανοίγει αυτόματα η ζωντανή κάμερα του υπολογιστή, στο οποίο εμφανίζεται συνεχώς η εικόνα που λαμβάνεται από την κάμερα. Στη δεξιά μεριά αρχικά εμφανίζεται ένα μήνυμα στον χρήστη, ότι το σύστημα είναι έτοιμο και μπορεί να ξεκινήσει την ανάλυση πατώντας το Space bar ή το κουμπί που του εμφανίζεται στο κάτω μέρος της οθόνης του. Ο χρήστης μπορεί να τραβήξει όσα στιγμιότυπα επιθυμεί, καθώς η διαδικασία δεν τερματίζεται μετά τη λήψη της πρώτης φωτογραφίας. Για κάθε στιγμιότυπο που θα τραβήξει,

θα λάβει και από μία αντίστοιχη περιγραφή. Ο χρήστης μπορεί να κλείσει το παράθυρο και να τερματίσει όλες τις ενέργειες πατώντας το Escape στο πληκτρολόγιο ή το X στην οθόνη του.



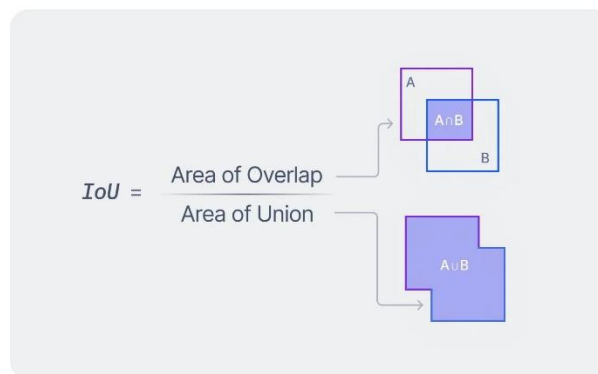
[5 - Γραφικό περιβάλλον της εφαρμογής κατά τη λειτουργία της]

5.3.2 Ανίχνευση Αντικειμένων

Η αναγνώριση αντικειμένων πραγματοποιείται μέσω δύο διαφορετικών μοντέλων που λειτουργούν συμπληρωματικά. Το πρώτο μοντέλο είναι το YOLOv8x, το οποίο χρησιμοποιείται ως βασικός μηχανισμός ανίχνευσης αντικειμένων. Έχει σχεδιαστεί για γρήγορη και ακριβή ανάλυση εικόνων και βίντεο. Μπορεί να αναγνωρίσει αντικείμενα, να προβλέψει τη θέση τους, να τα εντοπίσει με πλαίσια οριοθέτησης και να εκτιμήσει ανθρώπινες κινήσεις και καταστάσεις. Το μοντέλο αυτό επιστρέφει τη κατηγορία του αντικειμένου, τη θέση του στην εικόνα και τον βαθμό εμπιστοσύνης, ενώ τα αποτελέσματα χαμηλής αξιοπιστίας απορρίπτονται μέσω κατάλληλου κατωφλίου εμπιστοσύνης, ώστε να μειωθούν οι λανθασμένες ανιχνεύσεις.

Παράλληλα με το YOLO χρησιμοποιείται και το YOLO-World, μια προηγμένη προσέγγιση σε πραγματικό χρόνο που βασίζεται στο YOLOv8x. Αυτή η καινοτομία επιτρέπει την αναζήτηση αντικειμένων βάσει λεκτικά καθορισμένων κατηγοριών, πέρα από ένα αυστηρά προκαθορισμένο σύνολο κλάσεων. Μειώνοντας σημαντικά τις υπολογιστικές απαιτήσεις διατηρώντας παράλληλα την ανταγωνιστική απόδοση, το YOLO-World αναδεικνύεται ως ένα ευέλικτο εργαλείο για πολλές εφαρμογές που βασίζονται στην όραση, και αξιοποιώντας την υπολογιστική ταχύτητα των CNN, το προσφέρει μια γρήγορη λύση ανίχνευσης.

Το YOLO-World δεν ήρθε για να αντικαταστήσει το YOLOv8x, αλλά για να συμπληρώσει τα σημεία που δεν καλύπτονται επαρκώς από τις προκαθορισμένες κατηγορίες του. Η παράλληλη χρήση των δύο μοντέλων αυξάνει την ευελιξία του συστήματος και οδηγεί σε πλουσιότερη αναπαράσταση της σκηνής. Για τον συνδυασμό των αποτελεσμάτων των δύο μοντέλων χρησιμοποιείται η μετρική Intersection over Union (IoU), η οποία εκφράζει τον βαθμό επικάλυψης μεταξύ δύο πλαισίων οριοθέτησης, όπως φαίνεται στην εικόνα [6]. Όταν δύο ανιχνεύσεις της ίδιας κατηγορίας παρουσιάζουν επαρκή επικάλυψη, θεωρείται ότι είναι πιθανό να αντιστοιχούν στο ίδιο αντικείμενο. Στην περίπτωση αυτή διατηρείται η ανίχνευση με τον υψηλότερο βαθμό εμπιστοσύνης, περιορίζοντας έτσι την εμφάνιση διπλότυπων αντικειμένων στην τελική περιγραφή.

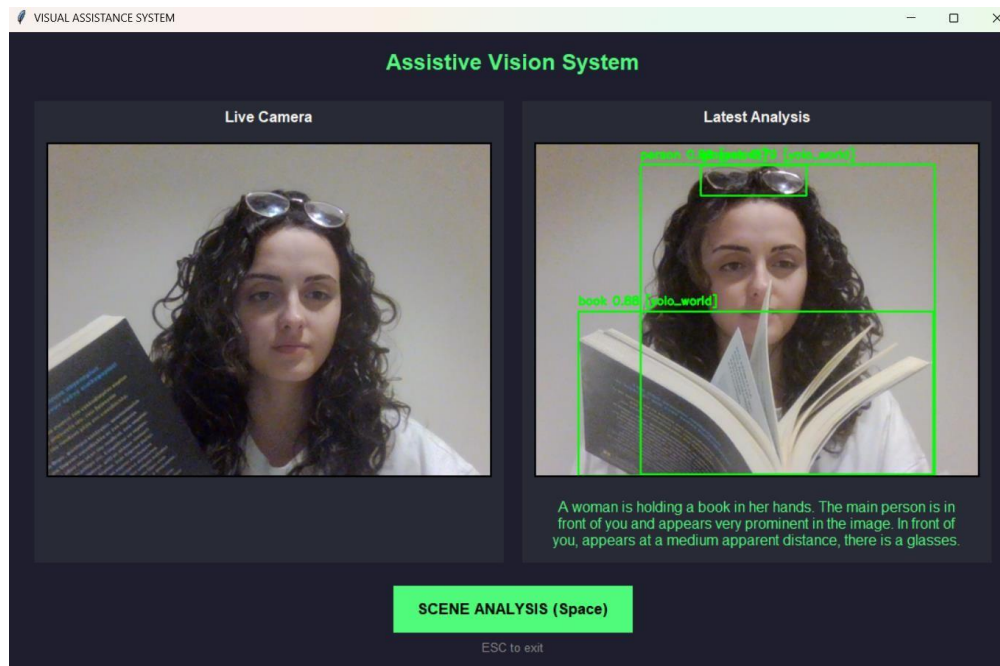


[6 - Intersection over Union (IoU): Definition, Calculation, Code]

Το σύστημα δεν περιορίζεται στο να αναφέρει ότι υπάρχει ένα αντικείμενο, αλλά προσπαθεί να ενημερώσει τον χρήστη για τη σχετική θέση του μέσα στο οπτικό πεδίο.

5.3.3 Δημιουργία περιγραφής της σκηνής

Η αναγνώριση αντικειμένων από μόνη της δεν επαρκεί για την περιγραφή του συνολικού περιεχομένου μιας σκηνής. Για τον λόγο αυτό, στο σύστημα χρησιμοποιείται το μοντέλο BLIP, το οποίο λαμβάνει ως είσοδο το στιγμιότυπο της κάμερας και παράγει μία περιγραφή του περιεχομένου της εικόνας σε φυσική γλώσσα. Η λειτουργία του BLIP πραγματοποιείται ανεξάρτητα από την ανίχνευση αντικειμένων των YOLOv8x και YOLO-World. Στη συνέχεια, η περιγραφή που παράγεται από το BLIP συνδυάζεται με τα αποτελέσματα της ανίχνευσης αντικειμένων, ώστε να δημιουργηθεί η τελική λεκτική περιγραφή της σκηνής. Με αυτόν τον τρόπο αξιοποιούνται τόσο οι πληροφορίες για τα αντικείμενα και τις σχετικές θέσεις τους όσο και η γενικότερη περιγραφή του περιεχομένου της εικόνας.



[7 - Ζωντανή λήψη εικόνας από την κάμερα]

Η φωτογραφία [7] είναι πραγματική λήψη στιγμιότυπου για τον έλεγχο λειτουργίας του συστήματος. Στη δεξιά πλευρά του GUI, βλέπουμε το στιγμιότυπο που τράβηξε ο χρήστης, καθώς και τη περιγραφή για το στιγμιότυπο αυτό. Στο συγκεκριμένο παράδειγμα, το σύστημα αναγνωρίζει πως υπάρχει μια γυναίκα στην εικόνα, η οποία κρατάει ένα βιβλίο και φοράει γυαλιά. Η εργασία αναγνωρίζει με επιτυχία όλα αυτά τα αντικείμενα, τα πρόσωπα και τις θέσεις τους σε σχέση με τη κάμερα. Το BLIP καταφέρνει να δώσει μια ολοκληρωμένη περιγραφή με όσα περισσότερα στοιχεία μπορεί να βρει.

5.3.4 Υποσύστημα σύνθεσης ομιλίας

Το τελικό στάδιο λειτουργίας του προτεινόμενου συστήματος είναι η μετατροπή της παραγόμενης περιγραφής της σκηνής σε ηχητική πληροφορία, ώστε ο χρήστης να μπορεί να ενημερώνεται άμεσα για το περιεχόμενο του περιβάλλοντός του χωρίς να απαιτείται οπτική επαφή με την οθόνη. Η διαδικασία αυτή υλοποιείται μέσω της τεχνολογίας Text-to-Speech TTS, η οποία μετατρέπει αυτόματα το παραγόμενο κείμενο σε ανθρώπινη ομιλία.

Στην παρούσα εργασία χρησιμοποιήθηκε η βιβλιοθήκη pyttsx3, ένα module της γλώσσας προγραμματισμού Python που υποστηρίζει τοπική σύνθεση ομιλίας. Η επιλογή της συγκεκριμένης βιβλιοθήκης βασίστηκε κυρίως στο γεγονός ότι λειτουργεί εκτός σύνδεσης. Το χαρακτηριστικό αυτό είναι ιδιαίτερα σημαντικό για εφαρμογές υποβοήθησης ατόμων με προβλήματα όρασης, καθώς μειώνει την εξάρτηση από σύνδεση στο διαδίκτυο και επιτρέπει την αξιοποίηση της εφαρμογής ακόμη και σε περιοχές με περιορισμένη ή μηδενική πρόσβαση σε δίκτυο.

Ένα από τα σημαντικότερα πλεονεκτήματα της pyttsx3 είναι ότι παρέχει πλήρη έλεγχο των βασικών χαρακτηριστικών της συνθετικής ομιλίας. Συγκεκριμένα, προσφέρει μικρή καθυστέρηση και επιτρέπει την άμεση εκφώνηση των αποτελεσμάτων του BLIP. Επιπλέον, με την pyttsx3 έχουμε τον πλήρη έλεγχο για τη ταχύτητα, την ένταση και τον τύπο της φωνής.

Η επιλογή τοπικής σύνθεσης ομιλίας συνάδει με τη συνολική φιλοσοφία της εφαρμογής, η οποία επιδιώκει να μειώσει την εξάρτηση από εξωτερικές υπηρεσίες και να προσφέρει μία περισσότερο αυτόνομη εμπειρία χρήσης

5.4 Πλεονεκτήματα και περιορισμοί του συστήματος

Η ανάπτυξη του προτεινόμενου συστήματος επέτρεψε την αξιοποίηση σύγχρονων τεχνολογιών τεχνητής νοημοσύνης για τη δημιουργία της εφαρμογής. Η αρχιτεκτονική που υλοποιήθηκε βασίζεται στον συνδυασμό διαφορετικών μοντέλων, τα οποία συνεργάζονται μεταξύ τους για να καλύψουν διαφορετικές πτυχές της κατανόησης μίας σκηνής. Αυτό που τη διαφοροποιεί είναι το ότι αξιοποιεί τα πλεονεκτήματα κάθε τεχνολογίας ξεχωριστά.

Τα μοντέλα YOLOv8x και YOLO-World αποτελούν από τα σημαντικότερα πλεονεκτήματα του συστήματος για την πολυεπίπεδη ανάλυση της εικόνας. Στη συνέχεια, το μοντέλο BLIP δεν περιορίζεται στην απλή απαρίθμηση των αντικειμένων, αλλά παράγει μία συνολικότερη, πιο ολοκληρωμένη και σωστή περιγραφή της σκηνής σε φυσική γλώσσα. Ο συνδυασμός αυτών των μοντέλων επιτρέπει στο σύστημα να αξιοποιεί τόσο τη χωρική πληροφορία των αντικειμένων, όσο και τη συνολική σημασιολογική κατανόηση της εικόνας, δημιουργώντας περιγραφές περισσότερο χρήσιμες για τον τελικό χρήστη.

Ένα δεύτερο σημαντικό πλεονέκτημα αφορά την αξιοποίηση της χωρικής πληροφορίας. Η εφαρμογή δεν αναφέρει απλώς ότι ένα αντικείμενο υπάρχει μέσα στην εικόνα, αλλά προσπαθεί να προσδιορίσει τη σχετική θέση του ως προς τον χρήστη. Επιπλέον, πραγματοποιείται εκτίμηση της σχετικής απόστασης των αντικειμένων με βάση το μέγεθος του πλαισίου αντίληψης, ενώ σε ορισμένες περιπτώσεις επιχειρεί να αναγνωρίσει απλές χωρικές σχέσεις μεταξύ αντικειμένων, όπως τη συσχέτιση ενός κινητού τηλεφώνου με ένα άτομο που πιθανώς το κρατάει. Η πληροφορία αυτή είναι ιδιαίτερα σημαντική για εφαρμογές υποβοήθησης, καθώς αυξάνει την πρακτική αξία της παραγόμενης περιγραφής.

Ένα ακόμα πλεονέκτημα αποτελεί η επιλογή χρήσης προεκπαιδευμένων μοντέλων, η αξιοποίηση των οποίων επέτρεψε την ανάπτυξη ενός λειτουργικού πρωτοτύπου χωρίς να απαιτηθεί η ιδιαίτερα απαιτητική διαδικασία εκπαίδευσης από την αρχή, η οποία θα απαιτούσε σημαντικούς υπολογιστικούς πόρους και μεγάλα σύνολα δεδομένων. Με τον τρόπο αυτό η εργασία επικεντρώθηκε κυρίως στον σχεδιασμό της αρχιτεκτονικής, στον συνδυασμό των μοντέλων και στη βελτίωση της τελικής εμπειρίας χρήσης.

Ένα ακόμη σημαντικό χαρακτηριστικό του συστήματος είναι η τοπική σύνθεση ομιλίας. Σε αντίθεση με υπηρεσίες που απαιτούν συνεχή σύνδεση στο διαδίκτυο, η συγκεκριμένη λύση

λειτουργεί εξ ολοκλήρου τοπικά. Αυτό ενισχύει την αυτονομία του συστήματος, κάτι που είναι ιδιαίτερα σημαντικό σε εφαρμογές προσβασιμότητας.

Παρά τα παραπάνω πλεονεκτήματα, το προτεινόμενο σύστημα παρουσιάζει και ορισμένους περιορισμούς που πρέπει να ληφθούν υπόψη κατά την αξιολόγηση των αποτελεσμάτων.

Ο σημαντικότερος περιορισμός αφορά την ποιότητα της εικόνας. Η απόδοση τόσο των μοντέλων YOLO όσο και του BLIP εξαρτάται άμεσα από την ποιότητα του frame που λαμβάνεται από την κάμερα. Εικόνες με χαμηλό φωτισμό, θόλωση λόγω κίνησης ή μερικώς κρυμμένα αντικείμενα μπορούν να μειώσουν σημαντικά την ακρίβεια της ανίχνευσης. Ο συγκεκριμένος περιορισμός συμφωνεί και με τη σύγχρονη βιβλιογραφία, ιδιαίτερα με τις εργασίες του VizWiz, όπως αναφέρθηκε και πιο πάνω.

Ένας δεύτερος περιορισμός σχετίζεται με την εμφάνιση λανθασμένων ανιχνεύσεων, false positives. Εξακολουθούν να υπάρχουν περιπτώσεις στις οποίες το σύστημα αναγνωρίζει αντικείμενα που στην πραγματικότητα δεν υπάρχουν στη σκηνή ή αποδίδει λανθασμένη κατηγορία σε κάποιο αντικείμενο. Εφόσον η τελική περιγραφή βασίζεται στις αρχικές αναγνωρίσεις, ένα σφάλμα στο στάδιο του Object Detection μπορεί να οδηγήσει σε ανακριβή ηχητική περιγραφή.

Παράλληλα, το σύστημα χρησιμοποιεί απλουστευμένη εκτίμηση χωρικής θέσης. Η θέση ενός αντικειμένου προσδιορίζεται με βάση το κέντρο του bounding box, το πράσινο περίγραμμα γύρω από κάθε αναγνωρίσιμο για τα YOLO αντικείμενα, και η εκτίμηση της απόστασης βασίζεται στο σχετικό μέγεθος που καταλαμβάνει μέσα στην εικόνα. Η προσέγγιση αυτή είναι υπολογιστικά αποδοτική, ωστόσο δεν αντιστοιχεί σε πραγματική μέτρηση απόστασης όπως θα μπορούσε να προσφέρει ένας αισθητήρας βάθους. Επομένως, οι πληροφορίες εγγύτητας θεωρούνται ενδεικτικές.

Παρά τους παραπάνω περιορισμούς, τα αποτελέσματα της παρούσας εργασίας δείχνουν ότι ο συνδυασμός τεχνολογιών Υπολογιστικής Όρασης, μοντέλων Όρασης-Γλώσσας και τοπικής σύνθεσης ομιλίας αποτελεί μία ιδιαίτερα αποτελεσματική προσέγγιση για την ανάπτυξη εφαρμογών υποβοήθησης ατόμων με προβλήματα όρασης. Η αρχιτεκτονική που υλοποιήθηκε αποδεικνύει ότι η συνεργασία εξειδικευμένων μοντέλων μπορεί να παράγει περισσότερο χρήσιμες και φυσικές περιγραφές σε σχέση με την αποκλειστική χρήση ενός μόνο μοντέλου.

ΚΕΦΑΛΑΙΟ 6 Μελέτες Περίπτωσης

Η αξιολόγηση του προτεινόμενου συστήματος πραγματοποιείται μέσω μελετών περίπτωσης σε διαφορετικές οπτικές σκηνές, με σκοπό να εξεταστεί η συμπεριφορά του κατά την πρακτική χρήση. Στο κεφάλαιο αυτό εξετάζεται το αποτέλεσμα της συνολικής λειτουργίας της, από τη λήψη του στιγμιότυπου μέχρι την παραγωγή της τελικής λεκτικής και ηχητικής περιγραφής.

Οι δοκιμές επικεντρώνονται κυρίως στην ορθότητα των ανιχνεύσεων, στην ικανότητα του συστήματος να συνδυάζει τις πληροφορίες των διαφορετικών μοντέλων και στην ποιότητα της τελικής περιγραφής. Παράλληλα, εξετάζονται περιπτώσεις στις οποίες εμφανίζονται λανθασμένες ανιχνεύσεις ή παραλείψεις, καθώς η αξιοπιστία της πληροφορίας αποτελεί ιδιαίτερα σημαντικό παράγοντα για ένα σύστημα που απευθύνεται σε άτομα με προβλήματα όρασης.

6.1 Μεθοδολογία αξιολόγησης

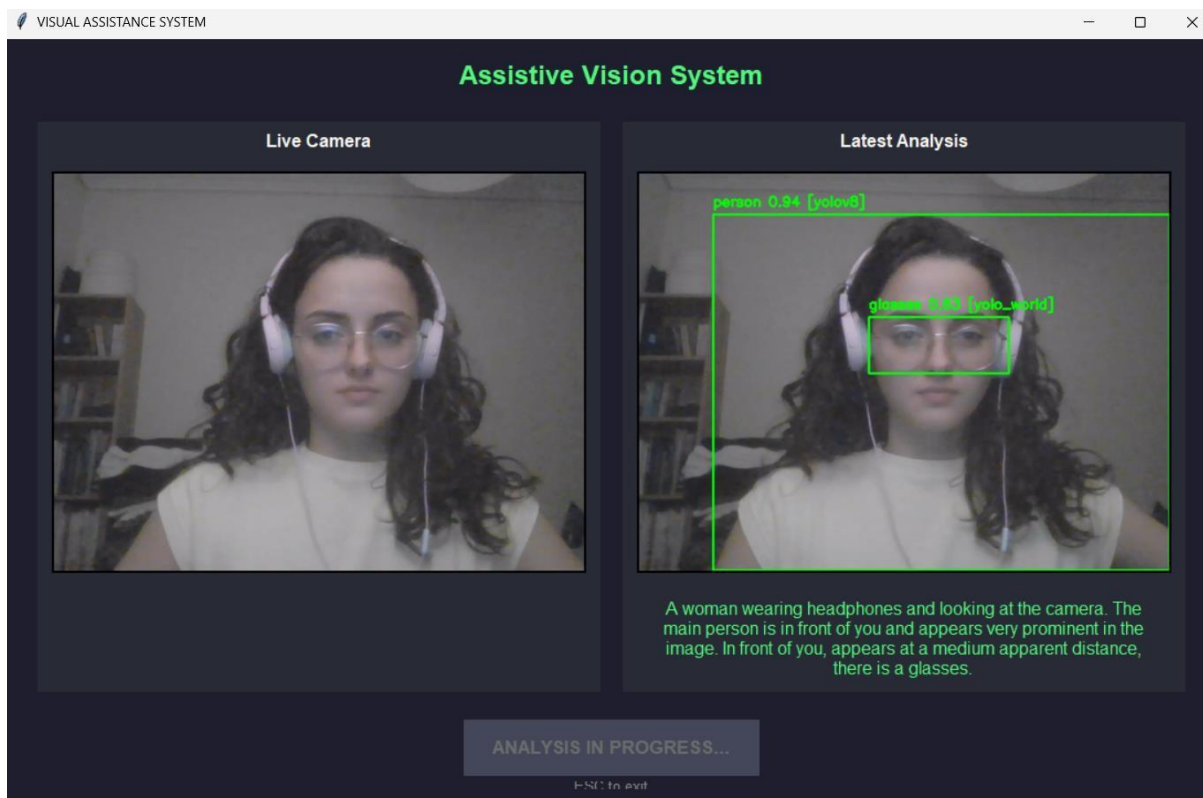
Για την αξιολόγηση χρησιμοποιούνται στιγμιότυπα που λαμβάνονται από την κάμερα του υπολογιστή μέσω της γραφικής διεπαφής της εφαρμογής. Μετά την ενεργοποίηση της ανάλυσης, το ίδιο στιγμιότυπο επεξεργάζεται από τα επιμέρους υποσυστήματα και στη συνέχεια εμφανίζεται στη διεπαφή μαζί με τα αντικείμενα που ανιχνεύθηκαν και την τελική περιγραφή. Η περιγραφή μετατρέπεται επίσης σε ομιλία μέσω του υποσυστήματος Text-toSpeech.

Η αξιολόγηση πραγματοποιείται σε διαφορετικές κατηγορίες σκηνών, ώστε να εξεταστεί η λειτουργία του συστήματος υπό διαφορετικές συνθήκες. Περιλαμβάνονται απλές σκηνές με περιορισμένο αριθμό αντικειμένων, περισσότερο σύνθετες σκηνές, περιπτώσεις στις οποίες τα αντικείμενα εμφανίζονται σε διαφορετικές θέσεις ή μεγέθη μέσα στην εικόνα, καθώς και περιπτώσεις στις οποίες παρουσιάζονται λανθασμένες ή ελλιπείς ανιχνεύσεις.

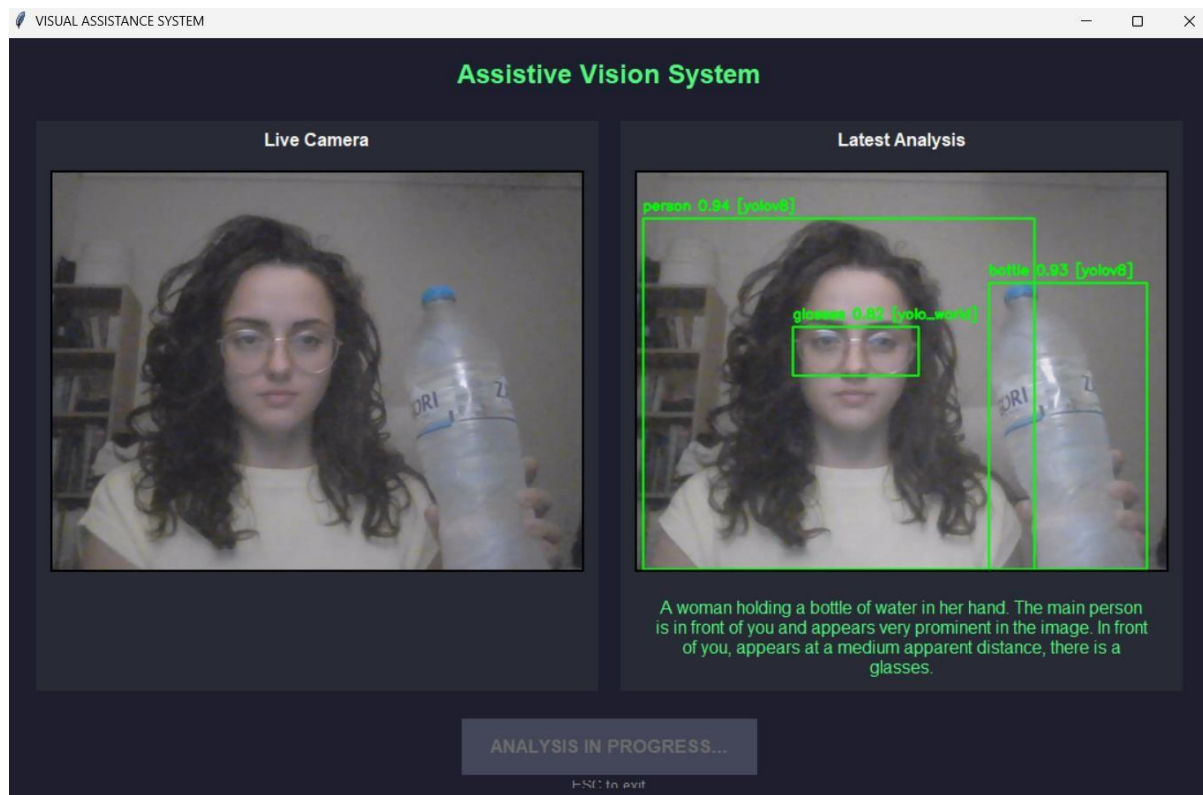
Για κάθε περίπτωση εξετάζονται τα αντικείμενα που υπάρχουν πραγματικά στη σκηνή, οι ανιχνεύσεις που πραγματοποιούνται από το σύστημα και η τελική περιγραφή που παράγεται. Με τον τρόπο αυτό μπορεί να αξιολογηθεί όχι μόνο η λειτουργία κάθε επιμέρους μοντέλου, αλλά κυρίως το τελικό αποτέλεσμα που λαμβάνει ο χρήστης.

6.2 Περίπτωση απλής σκηνής

Η πρώτη μελέτη περίπτωσης αφορά μία απλή σκηνή, στην οποία το κύριο στοιχείο της εικόνας είναι ένα άτομο μπροστά από την κάμερα. Σκοπός της δοκιμής είναι να εξεταστεί η βασική λειτουργία του συστήματος σε μία εικόνα με περιορισμένο αριθμό αντικειμένων, καθώς και η μεταβολή της παραγόμενης περιγραφής όταν προστίθεται ένα ακόμη αντικείμενο στη σκηνή.



[8 - Screenshot της πρώτης δοκιμής απλής σκηνής]



[9 - Screenshot της πρώτης δοκιμής απλής σκηνής]

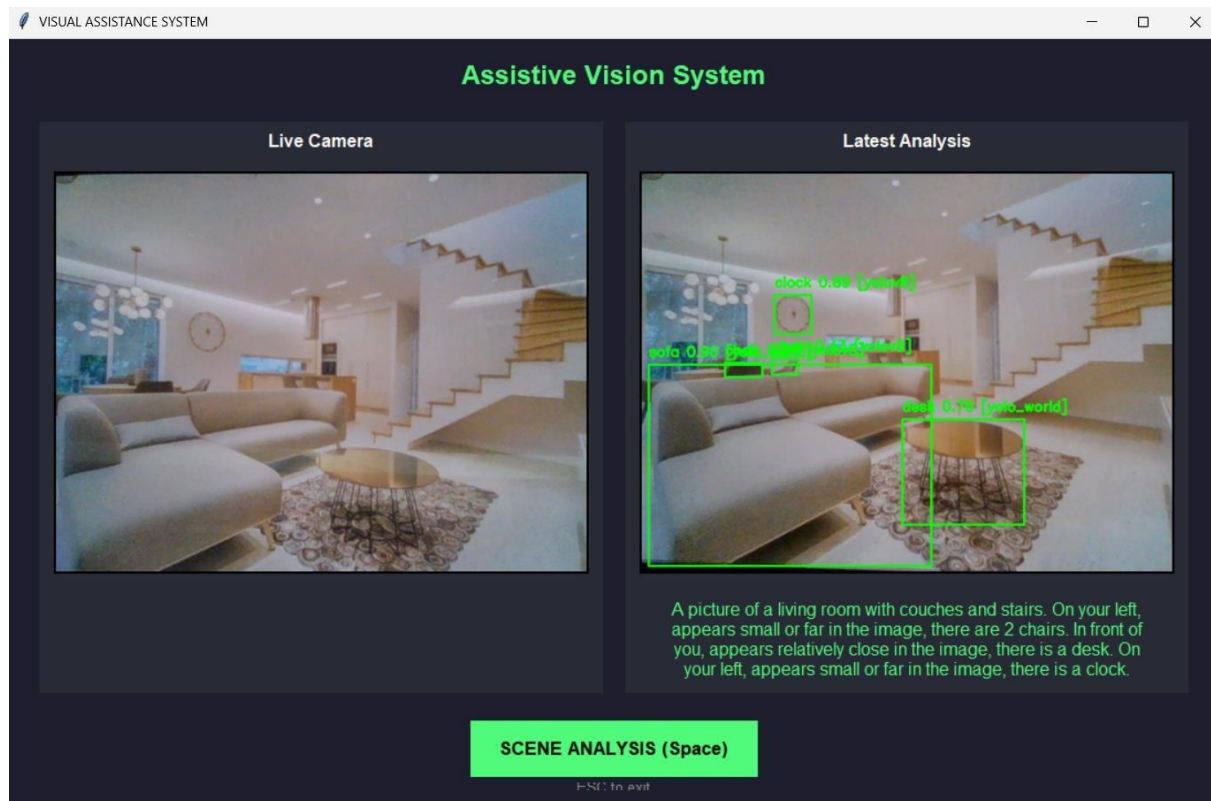
Στην πρώτη εικόνα [8], το YOLOv8 αναγνωρίζει σωστά την παρουσία του ατόμου με βαθμό εμπιστοσύνης 0,94. Παράλληλα, το YOLO-World εντοπίζει τα γυαλιά με βαθμό εμπιστοσύνης 0,63. Η γενική περιγραφή της εικόνας αναφέρει μία γυναίκα που φορά ακουστικά και κοιτάζει προς την κάμερα, ενώ στην τελική περιγραφή προστίθενται πληροφορίες από την ανίχνευση αντικειμένων σχετικά με την παρουσία του ατόμου και των γυαλιών. Το παράδειγμα δείχνει τον συμπληρωματικό ρόλο των διαφορετικών μοντέλων, καθώς τα ακουστικά εμφανίζονται στη γενική περιγραφή της εικόνας χωρίς να έχουν εντοπιστεί ως ξεχωριστό αντικείμενο από τους ανιχνευτές.

Στη δεύτερη δοκιμή [9] προστίθεται ένα μπουκάλι νερού, ενώ τα υπόλοιπα βασικά στοιχεία της σκηνής παραμένουν παρόμοια. Το άτομο αναγνωρίζεται και πάλι από το YOLOv8 με βαθμό εμπιστοσύνης 0,94, ενώ το YOLO-World εντοπίζει τα γυαλιά με βαθμό εμπιστοσύνης 0,62 και το μπουκάλι με 0,83. Η γενική περιγραφή αποδίδει τη σκηνή ως μία γυναίκα που κρατά ένα μπουκάλι νερού, ενώ η τελική έξοδος περιλαμβάνει επίσης τις πληροφορίες για το άτομο, τα γυαλιά και τη σχετική θέση τους.

Η σύγκριση των δύο περιπτώσεων δείχνει ότι το σύστημα προσαρμόζει την περιγραφή του όταν μεταβάλλεται το περιεχόμενο της εικόνας. Η προσθήκη του μπουκαλιού αναγνωρίζεται και ενσωματώνεται στην περιγραφή, χωρίς να χάνεται η πληροφορία για τα υπόλοιπα στοιχεία της σκηνής. Παράλληλα, η δεύτερη περίπτωση αναδεικνύει τον τρόπο με τον οποίο η γενική περιγραφή της εικόνας μπορεί να παρέχει πληροφορία για τη σχέση μεταξύ των στοιχείων της σκηνής, καθώς αναφέρεται ότι η γυναίκα κρατά το μπουκάλι, αντί να παρουσιάζονται αποκλειστικά ως δύο ανεξάρτητα αντικείμενα.

6.3 Περίπτωση σύνθετης σκηνής

Η δεύτερη μελέτη περίπτωσης αφορά μία περισσότερο σύνθετη σκηνή εσωτερικού χώρου, στην οποία εμφανίζονται ταυτόχρονα διαφορετικά αντικείμενα και στοιχεία του περιβάλλοντος. Σκοπός της συγκεκριμένης δοκιμής είναι να εξεταστεί η συμπεριφορά του συστήματος όταν η εικόνα περιέχει μεγαλύτερο αριθμό οπτικών πληροφοριών και απαιτείται ο συνδυασμός των αποτελεσμάτων της ανίχνευσης αντικειμένων με τη γενικότερη περιγραφή της σκηνής.



[10 - Ανάλυση σύνθετης σκηνής εσωτερικού χώρου]

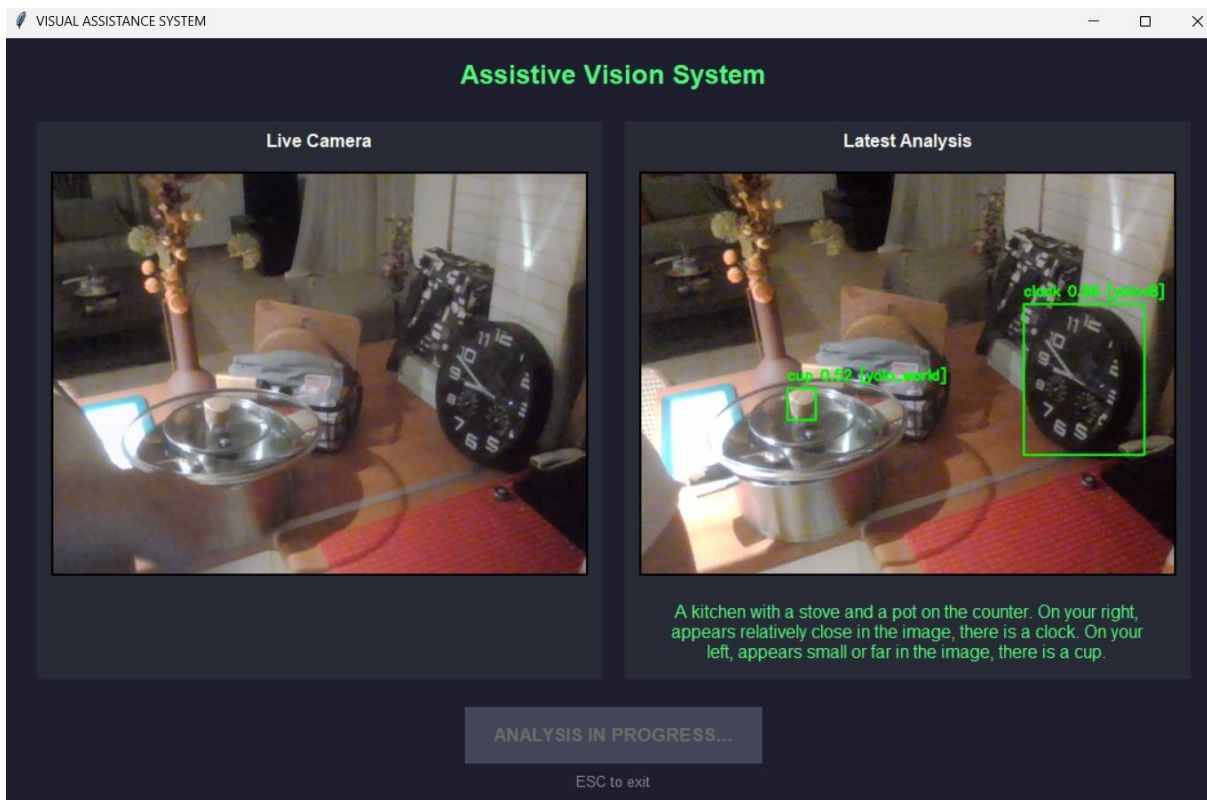
Στο συγκεκριμένο στιγμιότυπο [10] απεικονίζεται ένας εσωτερικός χώρος με καναπέδες, τραπέζακι, ρολόι και σκάλα. Το σύστημα αναγνωρίζει έναν καναπέ με βαθμό εμπιστοσύνης 0,90 και ένα ρολόι με 0,89, ενώ εμφανίζονται επίσης ανιχνεύσεις για καρέκλες και ένα τραπέζι που ταξινομείται ως γραφείο (desk). Παράλληλα, η περιγραφή της συνολικής εικόνας αποδίδει τη σκηνή ως καθιστικό με καναπέδες και σκάλα. Η αναφορά της σκάλας παρουσιάζει ιδιαίτερο ενδιαφέρον, καθώς αποτελεί στοιχείο του γενικότερου περιεχομένου της εικόνας χωρίς να εμφανίζεται ως ξεχωριστή ανίχνευση με πλαίσιο οριοθέτησης.

Η τελική περιγραφή προκύπτει από τον συνδυασμό των δύο μορφών πληροφορίας και περιλαμβάνει τόσο το γενικό περιεχόμενο της σκηνής όσο και πληροφορίες για επιμέρους αντικείμενα και τη σχετική θέση τους στο οπτικό πεδίο. Για ορισμένα αντικείμενα παρέχεται επίσης ένδειξη σχετικά με το πόσο μεγάλα εμφανίζονται στην εικόνα, η οποία χρησιμοποιείται ως προσεγγιστική πληροφορία φαινομενικής εγγύτητας και όχι ως πραγματική μέτρηση απόστασης.

Η συγκεκριμένη περίπτωση αναδεικνύει παράλληλα έναν περιορισμό της ανίχνευσης. Ορισμένα στοιχεία της σκηνής δεν αντιστοιχούν με σαφήνεια στις κατηγορίες που αποδίδει το σύστημα, όπως η ανίχνευση του τραπεζιού ως desk ενώ οι δύο ανιχνεύσεις chair εμφανίζουν σχετικά χαμηλότερους βαθμούς εμπιστοσύνης. Το παράδειγμα αυτό δείχνει ότι σε μία σύνθετη σκηνή η τελική περιγραφή μπορεί να περιλαμβάνει χρήσιμες πληροφορίες για το συνολικό περιβάλλον, χωρίς αυτό να σημαίνει ότι όλες οι επιμέρους ανιχνεύσεις είναι εξίσου ακριβείς.

6.4 Αξιολόγηση χωρικής πληροφορίας

Η συγκεκριμένη δοκιμή εξετάζει την ικανότητα του συστήματος να παρέχει πληροφορίες σχετικά με τη θέση των αντικειμένων μέσα στο οπτικό πεδίο. Εκτός από την αναγνώριση της κατηγορίας ενός αντικειμένου, η αναφορά της σχετικής θέσης του μπορεί να συμβάλει στην καλύτερη κατανόηση της διάταξης μιας σκηνής από τον χρήστη.



[11 - Αξιολόγηση της χωρικής πληροφορίας που παρέχει το σύστημα]

Στο συγκεκριμένο στιγμιότυπο [11] το σύστημα εντοπίζει ένα ρολόι (clock) με βαθμό εμπιστοσύνης 0,88 και ένα αντικείμενο που αναγνωρίζεται ως κύπελλο (cup) με βαθμό εμπιστοσύνης 0,52. Το ρολόι βρίσκεται στη δεξιά πλευρά της εικόνας και το σύστημα αποδίδει σωστά τη θέση του στην τελική περιγραφή με τη φράση “On your right”. Αντίστοιχα, το αντικείμενο που ταξινομεί ως κύπελλο βρίσκεται στην αριστερή πλευρά και περιγράφεται με τη φράση “On your left”. Στη συγκεκριμένη περίπτωση, επομένως, η χωρική πληροφορία που παράγεται με βάση τη θέση των πλαισίων οριοθέτησης συμφωνεί με την πραγματική διάταξη των αντικειμένων στην εικόνα.

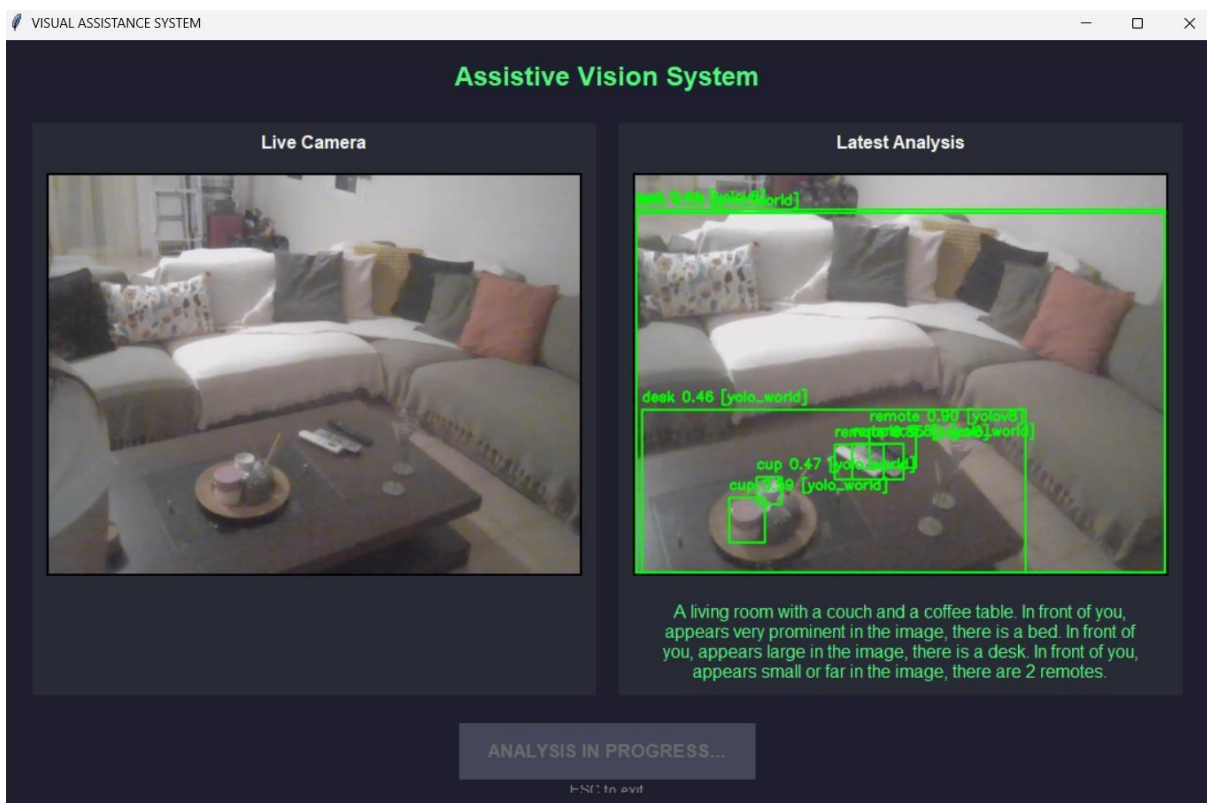
Εκτός από την κατεύθυνση, η τελική περιγραφή περιλαμβάνει και μία προσεγγιστική ένδειξη της φαινομενικής εγγύτητας των αντικειμένων. Το ρολόι χαρακτηρίζεται ως σχετικά κοντινό (“relatively close in the image”), ενώ το κύπελλο ως μικρό ή μακρινό (“small or far in the image”). Η εκτίμηση αυτή προκύπτει από το σχετικό μέγεθος που καταλαμβάνει το πλαίσιο

οριοθέτησης κάθε αντικειμένου μέσα στην εικόνα και δεν αποτελεί πραγματική μέτρηση της φυσικής απόστασής του από την κάμερα.

Η δοκιμή δείχνει ότι η πληροφορία των bounding boxes μπορεί να αξιοποιηθεί όχι μόνο για τον εντοπισμό των αντικειμένων, αλλά και για την παροχή βασικών στοιχείων σχετικά με τη θέση τους. Ωστόσο, η εκτίμηση της φαινομενικής εγγύτητας παρουσιάζει περιορισμούς, καθώς το μέγεθος ενός bounding box επηρεάζεται τόσο από την απόσταση όσο και από το πραγματικό μέγεθος του αντικειμένου. Για τον λόγο αυτό, οι χαρακτηρισμοί αυτοί χρησιμοποιούνται ως συμπληρωματικές ενδείξεις για τη διάταξη της σκηνής και όχι ως ακριβείς μετρήσεις απόστασης.

6.5 Περιπτώσεις λανθασμένων ανιχνεύσεων

Η αξιολόγηση του συστήματος είναι σημαντικό να περιλαμβάνει και περιπτώσεις στις οποίες οι ανιχνεύσεις δεν αντιστοιχούν πλήρως στο πραγματικό περιεχόμενο της εικόνας. Τέτοιου είδους σφάλματα παρουσιάζουν ιδιαίτερο ενδιαφέρον στην παρούσα εφαρμογή, καθώς οι πληροφορίες που προκύπτουν από την οπτική ανάλυση μπορούν να ενσωματωθούν στην τελική περιγραφή και στη συνέχεια να μετατραπούν σε ομιλία που θα λάβει ο τελικός χρήστης.



[12 - Παράδειγμα λανθασμένων ανιχνεύσεων σε σκηνή εσωτερικού χώρου]

Στο συγκεκριμένο στιγμιότυπο [12] απεικονίζεται ένα καθιστικό με έναν γωνιακό καναπέ και ένα τραπεζάκι στο προσκήνιο. Η γενική περιγραφή της εικόνας αποδίδει σωστά το βασικό περιεχόμενο της σκηνής ως “A living room with a couch and a coffee table”. Ωστόσο, στις επιμέρους ανιχνεύσεις εμφανίζονται ορισμένες λανθασμένες κατηγοριοποιήσεις. Ο καναπές αναγνωρίζεται ως bed, ενώ το τραπεζάκι ταξινομείται ως desk. Οι συγκεκριμένες ανιχνεύσεις μεταφέρονται και στην τελική περιγραφή, με αποτέλεσμα να γίνεται αναφορά σε ένα κρεβάτι και ένα γραφείο, παρόλο που τα αντικείμενα αυτά δεν υπάρχουν στη σκηνή.

Παράλληλα, το σύστημα εντοπίζει μικρότερα αντικείμενα επάνω στο τραπέζι, όπως δύο τηλεχειριστήρια (remote) και αντικείμενα που ταξινομούνται ως κύπελλα (cup). Η περίπτωση αυτή δείχνει ότι μέσα στην ίδια εικόνα μπορούν να συνυπάρχουν ορθές και λανθασμένες ανιχνεύσεις και ότι ένας σχετικά υψηλός βαθμός εμπιστοσύνης δεν εγγυάται από μόνος του ότι η αντίστοιχη κατηγοριοποίηση είναι σωστή.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η διαφορά μεταξύ της γενικής περιγραφής της εικόνας και των αποτελεσμάτων της ανίχνευσης αντικειμένων. Ενώ η πρώτη αναγνωρίζει σωστά το γενικό περιβάλλον ως καθιστικό με καναπέ και τραπεζάκι, ορισμένες από τις επιμέρους ανιχνεύσεις αποδίδουν διαφορετικές κατηγορίες στα ίδια στοιχεία. Κατά τον συνδυασμό των πληροφοριών, αυτές οι ασυμφωνίες μπορούν να οδηγήσουν σε αντιφατική τελική περιγραφή. Το παράδειγμα αναδεικνύει επομένως έναν περιορισμό του συστήματος: ο συνδυασμός διαφορετικών μοντέλων προσφέρει περισσότερες πληροφορίες για τη σκηνή, αλλά δεν εξασφαλίζει ότι οι πληροφορίες αυτές θα συμφωνούν πάντοτε μεταξύ τους.

Η συγκεκριμένη περίπτωση είναι ιδιαίτερα σημαντική για τον σκοπό της εφαρμογής, καθώς μία λανθασμένη ανίχνευση δεν παραμένει μόνο ως οπτικό σφάλμα στην οθόνη, αλλά μπορεί να μεταφερθεί στην ηχητική πληροφορία που λαμβάνει ο χρήστης. Τα κατώφλια εμπιστοσύνης και το φιλτράρισμα των αποτελεσμάτων μπορούν να περιορίσουν ορισμένες αναξιόπιστες ανιχνεύσεις, χωρίς όμως να εξαλείφουν πλήρως τέτοιου είδους σφάλματα. Για τον λόγο αυτό, η αντιμετώπιση των ασυμφωνιών μεταξύ των διαφορετικών πηγών πληροφορίας αποτελεί ένα σημείο που μπορεί να βελτιωθεί σε μελλοντική επέκταση του συστήματος.

6.6 Επίδραση των συνθηκών λήψης

Ένας ακόμη παράγοντας που μπορεί να επηρεάσει την απόδοση του συστήματος είναι η ποιότητα του στιγμιότυπου που λαμβάνεται από την κάμερα. Για την εξέταση του παράγοντα αυτού μπορούν να πραγματοποιηθούν δοκιμές με μεταβολές στον φωτισμό, στη θέση της κάμερας και στην ορατότητα των αντικειμένων.

Σε μία καθαρή εικόνα με επαρκή φωτισμό και ευδιάκριτα αντικείμενα αναμένεται τα μοντέλα να διαθέτουν περισσότερη οπτική πληροφορία για την πραγματοποίηση των προβλέψεών τους. Αντίθετα, η θόλωση, ο περιορισμένος φωτισμός ή η μερική απόκρυψη ενός αντικειμένου μπορούν να δυσκολέψουν την ανίχνευση και να επηρεάσουν την τελική περιγραφή.

6.7 Συνολική αξιολόγηση

Οι παραπάνω μελέτες περίπτωσης επιτρέπουν την εξέταση της συνολικής συμπεριφοράς του συστήματος σε διαφορετικές συνθήκες και όχι μόνο της λειτουργίας των επιμέρους μοντέλων. Η αξιολόγηση της τελικής περιγραφής είναι κρίσιμης σημασίας, καθώς αυτή αποτελεί την πληροφορία που τελικά μεταδίδεται στον χρήστη.

Ο συνδυασμός της ανίχνευσης αντικειμένων με την αυτόματη περιγραφή εικόνας επιτρέπει την αξιοποίηση διαφορετικών μορφών πληροφορίας για την ίδια σκηνή. Τα YOLOv8x και YOLO-World παρέχουν πληροφορίες σχετικά με τα αντικείμενα και τις θέσεις τους, ενώ το BLIP συμβάλλει με μία γενικότερη λεκτική περιγραφή του περιεχομένου της εικόνας. Η τελική έξοδος προκύπτει από την επεξεργασία και τον συνδυασμό αυτών των αποτελεσμάτων.

Οι δοκιμές μπορούν να αναδείξουν περιπτώσεις στις οποίες το σύστημα παρουσιάζει περιορισμούς, όπως λανθασμένες ή ελλιπείς ανιχνεύσεις και δυσκολίες σε λιγότερο ευνοϊκές συνθήκες λήψης. Η παρουσία τέτοιων περιπτώσεων είναι σημαντικό να λαμβάνεται υπόψη, ιδιαίτερα επειδή η εφαρμογή μετατρέπει αυτόματα την παραγόμενη πληροφορία σε ομιλία.

ΚΕΦΑΛΑΙΟ 7 Συμπεράσματα και Μελλοντικές Προεκτάσεις

7.1 Συμπεράσματα

Η παρούσα πτυχιακή εργασία είχε ως αντικείμενο τη σχεδίαση και την υλοποίηση ενός συστήματος αυτόματης περιγραφής τοπίου με ηχητική απόδοση, το οποίο μπορεί να λειτουργήσει ως υποστηρικτικό εργαλείο για άτομα με προβλήματα όρασης. Η βασική φιλοσοφία του συστήματος δεν περιορίστηκε στην απλή αναγνώριση αντικειμένων μέσα σε μία εικόνα, αλλά επικεντρώθηκε στη δημιουργία μίας περισσότερο ολοκληρωμένης και κατανοητής περιγραφής του περιβάλλοντος, ώστε η πληροφορία να αποκτά ουσιαστική αξία για τον τελικό χρήστη. Κατά τη διάρκεια της εργασίας έγινε σαφές ότι η αναγνώριση αντικειμένων από μόνη της δεν αρκεί για να περιγράψει επαρκώς μία πραγματική σκηνή. Η διαπίστωση αυτή αποτέλεσε βασικό σημείο σχεδίασης της εφαρμογής και δικαιολογεί την επιλογή του συνδυασμού διαφορετικών μοντέλων αντί της χρήσης ενός μόνο αλγορίθμου.

Η υλοποίηση βασίστηκε στη συνεργασία τεσσάρων διαφορετικών τεχνολογιών τεχνητής νοημοσύνης, καθεμία από τις οποίες αναλαμβάνει διαφορετικό ρόλο μέσα στη συνολική διαδικασία επεξεργασίας. Το YOLOv8x χρησιμοποιήθηκε για την ταχύτατη ανίχνευση αντικειμένων, το YOLO-World διεύρυνε το εύρος των κατηγοριών που μπορούν να αναγνωριστούν μέσω της προσέγγισης Open-Vocabulary Detection, το BLIP συνέβαλε στη δημιουργία μίας συνολικής περιγραφής της σκηνής και, τέλος, η βιβλιοθήκη pytsx3 μετέτρεψε το παραγόμενο κείμενο σε ομιλία χωρίς να απαιτείται σύνδεση στο διαδίκτυο. Η συγκεκριμένη αρχιτεκτονική επέτρεψε στο σύστημα να αξιοποιεί τα πλεονεκτήματα κάθε μοντέλου ξεχωριστά, δημιουργώντας μία περισσότερο ολοκληρωμένη τελική έξοδο.

Ένα ακόμη σημαντικό συμπέρασμα αφορά τη σημασία της τοπικής επεξεργασίας. Σε αντίθεση με εφαρμογές που βασίζονται αποκλειστικά σε cloud υπηρεσίες, η χρήση τοπικής σύνθεσης ομιλίας μειώνει την εξάρτηση από συνεχή σύνδεση στο διαδίκτυο και προσφέρει μεγαλύτερη αμεσότητα στην απόδοση της πληροφορίας. Για έναν χρήστη που κινείται σε εξωτερικό περιβάλλον, ακόμη και μία μικρή καθυστέρηση μπορεί να μειώσει τη χρησιμότητα της πληροφορίας, καθώς το περιβάλλον αλλάζει διαρκώς. Η δυνατότητα τοπικής εκφώνησης αποτελεί επομένως σημαντικό στοιχείο της προτεινόμενης προσέγγισης.

Παράλληλα, η εργασία επιβεβαίωσε αρκετές από τις προκλήσεις που έχουν ήδη αναδειχθεί στη διεθνή βιβλιογραφία. Οι εικόνες πραγματικών χρηστών δεν παρουσιάζουν πάντοτε ιδανικές συνθήκες λήψης. Χαμηλός φωτισμός, θόλωση λόγω κίνησης, μερικώς κρυμμένα αντικείμενα και ασυνήθιστες γωνίες λήψης μπορούν να επηρεάσουν σημαντικά την ακρίβεια των προβλέψεων. Οι δυσκολίες αυτές δεν αποτελούν αδυναμία μόνο της παρούσας εφαρμογής, αλλά χαρακτηρίζουν γενικότερα τα περισσότερα σύγχρονα συστήματα υπολογιστικής όρασης που λειτουργούν σε πραγματικές συνθήκες.

Τελικά, η εργασία καταδεικνύει ότι η ανάπτυξη συστημάτων προσβασιμότητας δεν αποτελεί αποκλειστικά τεχνικό ζήτημα. Η πραγματική πρόκληση βρίσκεται στη σωστή επιλογή της πληροφορίας που μεταδίδεται στον χρήστη, στην οργάνωση της περιγραφής και στη δημιουργία μίας εμπειρίας που να μπορεί να χρησιμοποιηθεί εύκολα στην καθημερινότητα.

7.2 Μελλοντικές Βελτιώσεις

Η εξέλιξη των μοντέλων κυρίως της τεχνητής νοημοσύνης δημιουργεί σημαντικές προοπτικές για τη βελτίωση της εφαρμογής. Η παρούσα υλοποίηση μπορεί να αποτελέσει τη βάση πάνω στην οποία θα αναπτυχθεί μία περισσότερο ολοκληρωμένη πλατφόρμα υποβοήθησης, ικανή να προσφέρει πλουσιότερη πληροφορία, μεγαλύτερη αξιοπιστία και καλύτερη εμπειρία χρήσης. Επιπλέον, ο συνδυασμός της εργασίας με κάποιο ήδη υπάρχον σύστημα ή εφαρμογή θα μπορούσε να φέρει ένα πιο ολοκληρωμένο αποτέλεσμα.

Μέχρι και σήμερα, η εφαρμογή μπορεί να αναγνωρίσει ότι ένα ποδήλατο βρίσκεται μπροστά στον χρήστη, χωρίς όμως να γνωρίζει πόσο κοντά βρίσκεται και αν κινείται ή είναι σταθμευμένο. Με τη χρήση μοντέλων όπως το ή το Depth Anything, θα μπορούσε να υπολογίζεται κατά προσέγγιση η απόσταση κάθε αντικειμένου από την κάμερα. Έτσι, η περιγραφή θα μπορούσε να μετατραπεί από «υπάρχει ποδήλατο μπροστά» σε «υπάρχει ποδήλατο περίπου δύο μέτρα μπροστά σου», επιτρέποντας στον χρήστη να αντιληφθεί καλύτερα τον πιθανό κίνδυνο και να προσαρμόσει την κίνησή του εγκαίρως.

Τα δύο μοντέλα MiDaS και Depth Anything [13] αποτελούν δύο από τα πιο δημοφιλή μοντέλα τεχνητής νοημοσύνης στον χώρο της Υπολογιστικής Όρασης για την εκτίμηση του βάθους, της απόστασης, κάθε σημείου σε μια σκηνή χρησιμοποιώντας μόνο μία απλή επίπεδη (2D) φωτογραφία χωρίς την ανάγκη για ειδικούς αισθητήρες. Το MiDaS (Mixed Dataset Solution) εκπαιδεύτηκε συνδυάζοντας πολλά διαφορετικά σύνολα δεδομένων με διαφορετικές κλίμακες και τύπους κάμερας και κατάφερε να παράγει σχετικό βάθος (relative depth) με μεγάλη ευστάθεια σε εντελώς άγνωστες εικόνες (zero-shot generalization), είτε πρόκειται για εσωτερικούς χώρους είτε για εξωτερικά τοπία.

Το μοντέλο Depth Anything αναπτύχθηκε ως η επόμενη γενιά (Foundation Model) στο depth estimation. Βασίζεται στην αρχιτεκτονική Vision Transformers (DINOv2) και εκπαιδεύτηκε σε ένα τεράστιο όγκο δεδομένων και ξεπερνά αισθητά το MiDaS σε ακρίβεια, ειδικά στις λεπτομέρειες, όπως λεπτά αντικείμενα ή αντανakλάσεις. Στη συνέχεια έρχεται το Depth Anything V2, η νεότερη έκδοσή του προσφέρει ακόμα υψηλότερη ανάλυση, καλύτερη διαχείριση αντανakλαστικών επιφανειών και πολύ γρήγορη ταχύτητα εκτέλεσης.

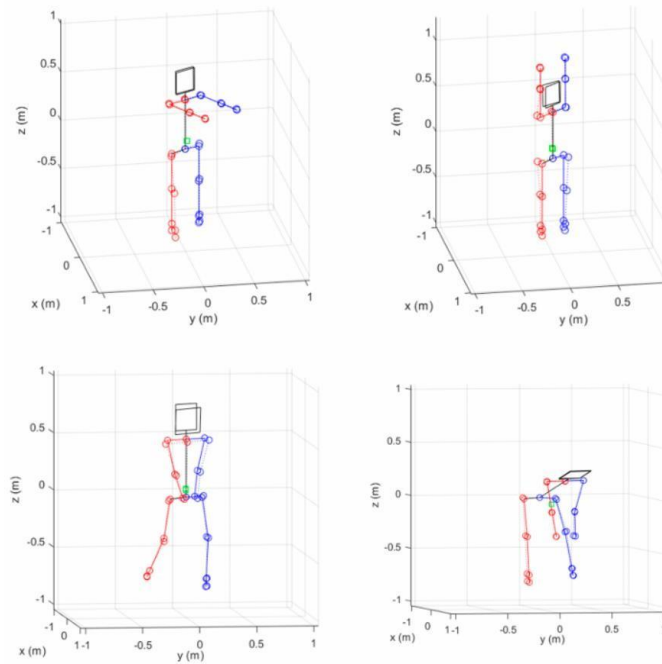


[13 - Depth Anything]

Η δυνατότητα αυτή θα ήταν ιδιαίτερα χρήσιμη σε καθημερινά σενάρια. Για παράδειγμα, κατά την είσοδο σε ένα κατάστημα, το σύστημα θα μπορούσε να ενημερώνει ότι «υπάρχει σκαλοπάτι σε απόσταση περίπου τριάντα εκατοστών», ενώ σε έναν εξωτερικό χώρο θα μπορούσε να αναφέρει ότι «υπάρχει κάδος απορριμμάτων περίπου ενάμιση μέτρο δεξιά». Η προσθήκη τέτοιων πληροφοριών θα μετέτρεπε την εφαρμογή από εργαλείο περιγραφής σε περισσότερο ενεργό σύστημα υποστήριξης της κίνησης.

Εξίσου σημαντική βελτίωση θεωρείται η ενσωμάτωση τεχνικών αναγνώρισης ανθρώπινων ενεργειών (Action Recognition). Στην παρούσα μορφή, η εφαρμογή μπορεί να αναγνωρίσει την παρουσία ενός ανθρώπου, αλλά δεν είναι πάντοτε σε θέση να κατανοήσει ή να αντιληφθεί πλήρως τι κάνει. Η αξιοποίηση τεχνολογιών όπως το MediaPipe Pose ή άλλων μοντέλων Pose Estimation θα επέτρεπε την αναγνώριση κινήσεων και στάσεων του σώματος. Για παράδειγμα, αντί να αναφέρει απλώς «υπάρχει άνθρωπος μπροστά», θα μπορούσε να ενημερώνει ότι «ένα άτομο κάθεται σε παγκάκι», «ένα παιδί τρέχει προς τα αριστερά» ή «κάποιος ανοίγει μία πόρτα». Τέτοιου είδους πληροφορίες θα προσέφεραν σημαντικά πλουσιότερη αντίληψη του περιβάλλοντος.

Το Pose Estimation πρόκειται για τη διαδικασία χαρτογράφησης των αρθρώσεων και των βασικών σημείων του σώματος (π.χ. ώμοι, αγκώνες, γόνατα, αστράγαλοι) σε συντεταγμένες (x, y) ή (x, y, z). Το MediaPipe Pose είναι το εξειδικευμένο framework της Google που εκτελεί αυτή τη διαδικασία σε πραγματικό χρόνο. Χρησιμοποιεί μια προσέγγιση δύο σταδίων για μέγιστη ταχύτητα και ακρίβεια, τον detector (ανιχνευτής) ο οποίος αρχικά εντοπίζει την παρουσία και την περιοχή του ανθρώπινου σώματος στην εικόνα, και το landmark model (μοντέλο σημείων), το οποίο εκτελεί την πρόβλεψη των 33 σημείων πάνω στην εντοπισμένη περιοχή. Η εικόνα [14] αποτελεί παραδείγματα ανθρώπινης κίνησης και των συντεταγμένων τους με τη χρήση του MediaPipe Pose.



[14 - applsci-13-02700.pdf]

Η παρούσα εργασία χρησιμοποιεί το BLIP, το οποίο αποτελεί αξιόπιστο μοντέλο παραγωγής περιγραφών εικόνων. Ωστόσο, η εξέλιξη του χώρου έχει οδηγήσει στην ανάπτυξη νέων μοντέλων, όπως το Florence-2, το LLaVA και το Qwen2.5-VL, τα οποία διαθέτουν σημαντικά αυξημένες δυνατότητες κατανόησης πολύπλοκων σκηνών. Το Florence-2, ένα εξαιρετικά ελαφρύ και εξειδικευμένο μοντέλο, το LLaVA, μοντέλο που άνοιξε τον δρόμο για τα ανοιχτά (Open-Source) πολυτροπικά μοντέλα παίρνοντας τα οπτικά χαρακτηριστικά μιας εικόνας και τα μεταφράζει στη γλώσσα που καταλαβαίνει το LLaMA, και Qwen2.5-VL, το οποίο εκπροσωπεί τη σύγχρονη γενιά (State-of-the-Art) των Vision-Language Models. Είναι ένα τεράστιο, εξαιρετικά ισχυρό μοντέλο που ανταγωνίζεται ευθέως εμπορικά μοντέλα όπως το GPT-4o και το Claude 3.5 Sonnet.

Είναι μοντέλα τεχνητής νοημοσύνης που συνδυάζουν την κατανόηση εικόνας και βίντεο με τη γλώσσα. Τα μοντέλα αυτά μπορούν να συνδυάζουν περισσότερες πληροφορίες ταυτόχρονα, να κατανοούν σχέσεις μεταξύ αντικειμένων και να παράγουν περισσότερο φυσικό λόγο. Για παράδειγμα, αντί για μία απλή περιγραφή όπως «υπάρχει τραπέζι και καρέκλα», ένα τέτοιο μοντέλο θα μπορούσε να δημιουργήσει την πρόταση «έναν άνθρωπος κάθεται δίπλα σε ξύλινο τραπέζι μέσα σε μία καφετέρια, ενώ στο βάθος υπάρχουν ακόμη δύο καθίσματα».

Η τεχνολογία TTS αποτελεί πλέον μέρος μιας σειράς εφαρμογών, από εικονικούς βοηθούς έως Siri και Alexa μέχρι τα εργαλεία προσβασιμότητας για άτομα με προβλήματα όρασης. Και παρόλο που έχει σημειωθεί τεράστια πρόοδος στην ομιλία των συστημάτων, χρειάζεται ακόμη πολύς χρόνος και δουλειά για να φτάσει η ομιλία να μοιάζει με μια ανθρώπινη, καθημερινή συζήτηση. Ιδιαίτερα όσον αφορά τη φυσικότητά της, την σύνθεση αλλά και την έκφρασή της, και όλα αυτά σε πραγματικό χρόνο.

Με την ανάπτυξη της μάθησης μεταφοράς και της αύξησης δεδομένων, τα μελλοντικά μοντέλα TTS πιθανότατα θα είναι σε θέση να παράγουν σύνθεση υψηλής ποιότητας για γλώσσες με ανεπαρκείς πόρους. Αυτό θα βοηθήσει στην επικοινωνία και τη διδασκαλία σε διαφορετικά γλωσσικά περιβάλλοντα, καθιστώντας έτσι την τεχνολογία πιο προσιτή.

Επιπλέον, αυτές οι εξελίξεις είναι πιθανό να οδηγήσουν σε μαζική προσαρμογή της τεχνολογίας TTS σε ένα ευρύ φάσμα βιομηχανιών που κυμαίνονται από την εκπαίδευση και την ψυχαγωγία έως τις υπηρεσίες που περιλαμβάνουν προσβασιμότητα, έτσι ώστε να υπάρχει μεγαλύτερη ενσωμάτωση της συνθετικής ομιλίας σε κοινές αλληλεπιδράσεις με τεχνολογίες.

Παράλληλα, ένα ενδιαφέρον κομμάτι που θα μπορούσε να αναπτυχθεί είναι η εξατομίκευση της εμπειρίας χρήσης. Διαφορετικοί χρήστες έχουν διαφορετικές ανάγκες. Κάποιος μπορεί να ενδιαφέρεται κυρίως για εμπόδια στο έδαφος, ενώ κάποιος άλλος μπορεί να επιθυμεί να αναγνωρίζει πινακίδες, πρόσωπα ή αντικείμενα μέσα σε ένα κατάστημα, ή κάποιος ο οποίος θέλει να το χρησιμοποιεί για να εντοπίζει ή να αναγνωρίζει αντικείμενα μέσα στο σπίτι του. Μία μελλοντική έκδοση θα μπορούσε να επιτρέπει τη δημιουργία διαφορετικών προφίλ λειτουργίας, ώστε το σύστημα να προσαρμόζει αυτόματα την πληροφορία που παρουσιάζει ανάλογα με τις προτιμήσεις του χρήστη.

Σημαντική επέκταση αποτελεί επίσης η υποστήριξη περισσότερων γλωσσών. Η δυνατότητα επιλογής τόσο της γλώσσας παραγωγής του κειμένου όσο και της γλώσσας σύνθεσης ομιλίας θα επέτρεπε τη χρήση της εφαρμογής από μεγαλύτερο αριθμό χρηστών, χωρίς να απαιτούνται αλλαγές στη βασική αρχιτεκτονική του συστήματος.

Σε μακροπρόθεσμο επίπεδο, η εφαρμογή μπορεί να εξελιχθεί από ένα σύστημα περιγραφής εικόνων σε μία ολοκληρωμένη πλατφόρμα προσωπικής υποβοήθησης για άτομα με προβλήματα όρασης. Η εξέλιξη αυτή δεν απαιτεί απαραίτητα την αντικατάσταση της υπάρχουσας αρχιτεκτονικής, αλλά μπορεί να πραγματοποιηθεί σταδιακά μέσω της ενσωμάτωσης νέων λειτουργιών και αισθητήρων.

Μία από τις σημαντικότερες προοπτικές αφορά τη μεταφορά της εφαρμογής από τον υπολογιστή σε φορητές συσκευές, όπως έξυπνα κινητά τηλέφωνα ή έξυπνα γυαλιά. Σε μία τέτοια περίπτωση, ο χρήστης δεν θα χρειάζεται να κρατά σταθερά μία κάμερα συνδεδεμένη με υπολογιστή, αλλά θα μπορεί να λαμβάνει πληροφορίες καθώς κινείται φυσιολογικά στον χώρο. Η εμπειρία αυτή θα είναι περισσότερο συμβατή με την καθημερινή χρήση και θα επιτρέπει συνεχή αλληλεπίδραση με το περιβάλλον.

Ένα ακόμη σημαντικό βήμα θα ήταν η σύνδεση της εφαρμογής με υπηρεσίες πλοήγησης και αισθητήρες θέσης. Η αξιοποίηση δεδομένων GPS, σε συνδυασμό με την αναγνώριση διαβάσεων πεζών, φωτεινών σηματοδοτών και πινακίδων οδικής κυκλοφορίας, θα μπορούσε να δημιουργήσει έναν ολοκληρωμένο ψηφιακό βοηθό μετακίνησης. Για παράδειγμα, ο χρήστης θα μπορούσε να λαμβάνει ειδοποιήσεις όπως «πλησιάζεις διάβαση πεζών», «το φανάρι είναι πράσινο» ή «υπάρχει ποδηλάτης που κινείται από τα δεξιά».

Τέλος, ιδιαίτερη σημασία έχει η πραγματική αξιολόγηση της εφαρμογής από άτομα με προβλήματα όρασης. Μέχρι σήμερα η παρούσα εργασία επικεντρώθηκε κυρίως στη σχεδίαση και στην τεχνική υλοποίηση του συστήματος. Το επόμενο ερευνητικό βήμα θα μπορούσε να περιλαμβάνει δοκιμές πεδίου με πραγματικούς χρήστες, αξιολόγηση της χρηστικότητας μέσω

ερωτηματολογίων, καθώς και καταγραφή του χρόνου απόκρισης, της κατανόησης των περιγραφών και της εμπιστοσύνης που αναπτύσσουν οι χρήστες προς το σύστημα. Τα αποτελέσματα μίας τέτοιας αξιολόγησης θα επέτρεπαν τη βελτίωση όχι μόνο των τεχνικών χαρακτηριστικών, αλλά κυρίως της πραγματικής εμπειρίας χρήσης.

ΚΕΦΑΛΑΙΟ 8 Επίλογος

Κατά τη διάρκεια της μελέτης έγινε εμφανές ότι η έννοια της προσβασιμότητας δεν αφορά αποκλειστικά την τεχνική ικανότητα ενός αλγορίθμου να αναγνωρίζει αντικείμενα, αλλά κυρίως την ποιότητα της πληροφορίας που τελικά φτάνει στον χρήστη. Ένα άτομο με περιορισμένη ή μηδενική όραση δεν χρειάζεται απλώς να γνωρίζει ότι υπάρχουν αντικείμενα γύρω του. Χρειάζεται να αποκτήσει μία συνολική εικόνα του χώρου, να αντιληφθεί ποια στοιχεία είναι σημαντικά για την κίνησή του και να λάβει την πληροφορία αυτή με τρόπο άμεσο, φυσικό και εύκολα κατανοητό. Η συγκεκριμένη παρατήρηση αποτέλεσε τον βασικό άξονα σχεδιασμού της εφαρμογής και επηρέασε όλες τις επιλογές που πραγματοποιήθηκαν κατά την ανάπτυξή της.

Παράλληλα, η υλοποίηση επιβεβαίωσε αρκετές από τις προκλήσεις που καταγράφονται στη διεθνή βιβλιογραφία σχετικά με τα συστήματα υπολογιστικής όρασης. Η ποιότητα της εικόνας αποδείχθηκε καθοριστικός παράγοντας για την επιτυχία της ανίχνευσης. Θολές εικόνες, χαμηλός φωτισμός, αντικείμενα που βρίσκονται σε μεγάλη απόσταση ή εμφανίζονται μόνο μερικώς μέσα στο κάδρο μπορούν να επηρεάσουν σημαντικά την ακρίβεια των αποτελεσμάτων. Επιπλέον, παρατηρήθηκε ότι ακόμη και σύγχρονα μοντέλα υψηλής απόδοσης είναι δυνατόν να παράγουν ψευδώς θετικές ανιχνεύσεις ή να παραλείπουν υπαρκτά αντικείμενα, γεγονός που επιβεβαιώνει ότι η αξιοπιστία εξακολουθεί να αποτελεί σημαντικό ερευνητικό ζήτημα.

Βιβλιογραφία

1. Kreiss, E., Bennett, C., Hooshmand, S., Zelikman, E., Morris, M. R., & Potts, C. (2022, October 27). *Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics*. arXiv.Org. <https://arxiv.org/abs/2205.10646>
2. Παπαδακάκης, Ν. (2022, September 18). *Υλοποίηση ενός ηχητικού ψηφιακού βοηθού βασισμένου σε τεχνολογία διαδικτύου των πραγμάτων για την υποστήριξη της αυτόνομης κίνησης ατόμων με προβλήματα όρασης*. Apothesis.Eap.Gr. <https://apothesis.eap.gr/archive/item/172485>
3. Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4). <https://doi.org/10.1007/s10462-024-10721-6>
4. Chinnaiyan, B., Balasubramanian, S., Jeyabalu, M., & Warriar, G. S. (2024). AI applications – computer vision and natural language processing. *Model Optimization Methods for Efficient and Edge AI*, 25–41. <https://doi.org/10.1002/9781394219230.ch2>
5. Khanam, F., Munmun, F. A., Ritu, N. A., Saha, A. K., & Mridha, M. F. (2022). Text to speech synthesis: A systematic review, deep learning based architecture and future research direction. *Journal of Advances in Information Technology*, 13(5). <https://doi.org/10.12720/jait.13.5.398-412>
6. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Thecvf.Com*, 779–788. https://openaccess.thecvf.com/content_cvpr_2016/html/Redmon_You_Only_Look_C_VPR_2016_paper.html
7. Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June 28). *BLIP: Bootstrapping languageimage pre-training for unified vision-language understanding and generation*. PMLR. Proceedings.Mlr.Press. <https://proceedings.mlr.press/v162/li22n.html>
8. Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv:2106.15561 [Cs, Eess]*. <https://arxiv.org/abs/2106.15561>
9. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/https://doi.org/10.1038/nature14539>
10. Valipoor, M., de Antonio, A., & Cabrera, J. (2024b). Analysis and design framework for the development of indoor scene understanding assistive solutions for the person with visual impairment/blindness. *Multimedia Systems*, 30(3). <https://doi.org/10.1007/s00530-024-01350-8>
11. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., & Shan, Y. (2024, February 22). *YOLO-World: Real-time open-vocabulary object detection*. arXiv.Org. <https://arxiv.org/abs/2401.17270>
12. dos Santos Caraíba, A. F., & Cardoso, A. A. (2025). Object distance estimation system with deep neural networks. *2025 Brazilian Conference on Robotics (CROS)*, 1–6. <https://doi.org/10.1109/cros66186.2025.11066153>
13. Gonzalez Penuela, R. E., Collins, J., Bennett, C., & Azenkot, S. (2024). Investigating use cases of AI-Powered scene description applications for blind and low vision people. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–21. <https://doi.org/10.1145/3613904.3642211>

14. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., Tiktok, Cuhk, & Zju. (n.d.). *Depth anything: Unleashing the power of large-scale unlabeled data*. Retrieved September 18, 2026, from <https://arxiv.org/pdf/2401.10891>
15. Chauhan, R., Dhyan, I., & Vaidya, H. (2023). A review on human pose estimation using mediapipe. *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, 1–6. <https://doi.org/10.1109/cisct57197.2023.10351240>
16. Kukana, P., Kaur, B., Singh, G., Kaur, A., & Kaur, D. (2025). Next-generation text-to-speech: Neural models, benchmarks and future prospects. *2025 2nd International Conference on Computational Intelligence and Computing Applications (ICCICA)*, 1–6. <https://doi.org/10.1109/iccica67008.2025.11337494>
17. Singh, A. K., Kumbhare, V. A., & Arthi, K. (2022). Real-time human pose detection and recognition using MediaPipe. *Advances in Intelligent Systems and Computing*, 145–154. https://doi.org/10.1007/978-981-16-7088-6_12
18. Singh, S. (2024, July 11). *Florence-2: Vision model for diverse AI applications*. Labellerr AI. <https://www.labellerr.com/blog/unveiling-florence-2-microsofts-cuttingedge-vision-model-for-diverse-ai-applications/>
19. Ultralytics. (2024, February 14). *YOLO-World model*. Ultralytics. Ultralytics Docs. <https://docs.ultralytics.com/models/yolo-world>
20. *What is YOLOv8: aN in-depth exploration of the internal features of the next generation object detector*. (2015). Arxiv.Org. <https://arxiv.org/html/2408.15857v1>
21. Reynolds, J., Kanth Nagesh, C., & Gurari, D. (n.d.). *Salient object detection for images taken by people with vision impairments*. Retrieved September 18, 2026, from https://openaccess.thecvf.com/content/WACV2024/papers/Reynolds_Salient_Object_Detection_for_Images_Taken_by_People_With_Vision_WACV_2024_paper.pdf
22. Louw, J. M., Verster, J., & Dickens, J. (2025). Practical implementation of depth anything V2 as a LiDAR alternative in robotics navigation. *MATEC Web of Conferences*, 417, 04003. <https://doi.org/10.1051/mateconf/202541704003>
23. Ultralytics. (2026, July 7). *YOLO architecture explained*. Ultralytics. Ultralytics Docs. <https://docs.ultralytics.com/guides/yolo-architecture>
24. *Salesforce/blip-image-captioning-base · hugging face*. (2024, October 30). Huggingface.Co. <https://huggingface.co/Salesforce/blip-image-captioning-base>
25. Gonzalez, R., Collins, J., Bennett, C. L., & Shiri Azenkot. (2024). *Understanding use cases for AI-Powered visual interpretation services*. Research.Google. <https://research.google/pubs/understanding-use-cases-for-ai-powered-visualinterpretation-services/>
26. LLaVA
27. Blindness and vision impairment
28. Λευκό Μπαστούνι: Τι συμβολίζει και ποια η ιστορία του;
29. ECVA | European Computer Vision Association
30. Few-Shot Object Localization – VizWiz
31. VizWiz – Algorithms to assist people who are blind
32. Practical implementation of Depth Anything V2 as a LiDAR alternative in robotics navigation | MATEC Web of Conferences