ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ, ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΗΣΗΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΤΙΚΩΝ, ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΚΟΙΝΩΝΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

DEPARTMENT OF ARCHIVAL, LIBRARY AND INFORMATION STUDIES
SCHOOL OF MANAGEMENT, ECONOMICS AND SOCIAL SCIENCES

Πτυχιακή Εργασία

Πολυμεσική Ανάκτηση Πληροφοριών: Εικόνα και Ήχος

ΦΑΡΑΟΣ ΔΙΟΝΥΣΙΟΣ (ΑΜ: 17027)

ΝΤΟΝΙ ΑΝΤΖΕΛΟ (ΑΜ: 17040)

Επιβλέπων: **Ιωάννης Τριανταφύλλου**

Αθήνα, Σεπτέμβριος 2022

Επιτροπή Εξέτασης

1. Ονοματεπώνυμο

Ioannis Triantafyllou

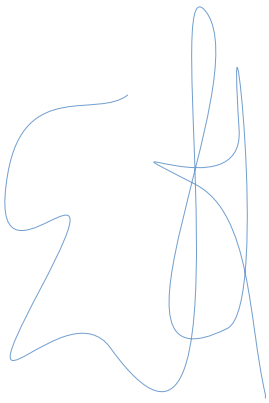
Digitally signed by Ioannis Triantafyllou
Date: 2022.09.30 09:26:17 +03'00'

2. Ονοματεπώνυμο

Markos Dendrinos

Digitally signed by Markos Dendrinos
Date: 2022.09.30 10:19:51 +03'00'

3. Ονοματεπώνυμο



Sarantos Kapidakis
2022.09.30
07:39:44 Z

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

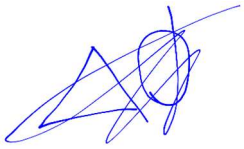
Ο κάτωθι υπογεγραμμένος ΦΑΡΑΟΣ ΔΙΟΝΥΣΙΟΣ, με αριθμό μητρώου 59917027 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Διοικητικών, Οικονομικών και Κοινωνικών Επιστημών του Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο Δηλών

ΦΑΡΑΟΣ ΔΙΟΝΥΣΙΟΣ



ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

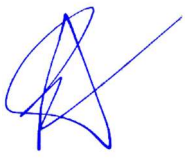
Ο κάτωθι υπογεγραμμένος NDONI ANXHELO, με αριθμό μητρώου 59917040 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Διοικητικών, Οικονομικών και Κοινωνικών Επιστημών του Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο Δηλών

NDONI ANXHELO



Ευχαριστίες – Αφιερώσεις

Ημερομηνία

Συγγραφέας

Περίληψη στα ελληνικά

Η εργασία ασχολείται με την ανάγκη του ανθρώπου για ανάκτηση πολυμεσικού υλικού. Έμφαση έχει δοθεί στην εικόνα και τον ήχο. Παρουσιάζονται οι παράγοντες που οδήγησαν στην άνοδο του πεδίου αυτού και τον καίριο συνδυασμό του με το machine learning και τα νευρωνικά δίκτυα. Αναφέρονται οι σημαντικότερες τεχνικές ανάκτησης, εξαγωγής και αντιστοίχισης δεδομένων, από τις αρχές του 20ου αιώνα μέχρι σήμερα. Αναφορά γίνεται στις βασικές εφαρμογές ανάκτησης, που έχουν γίνει μέρος της καθημερινότητας των ανθρώπων. Αξιολογούνται με βάση την ταχύτητα, την ακρίβεια και την πρακτικότητα οι πιο προηγμένες μηχανές αναζήτησης με σκοπό να συγκριθούν. Τέλος, συλλέγονται οι σκέψεις για το παρόν και το μέλλον της ανάκτησης πληροφοριών και τα κριτήρια που μπορούν να φέρουν νέους χρήστες, ακόμα και εάν κύριο μέλημα παραμένει η ικανοποίηση και η πρόσβαση του χρήστη στην πληροφορία.

Λέξεις Κλειδιά: ανάκτηση, μηχανή αναζήτησης, εικόνα, ήχος, νευρωνικά δίκτυα.

Περίληψη στα αγγλικά

The thesis deals with the human need to retrieve multimedia content. The emphasis has been placed on image and sound. The factors that led to the rise of this field and its key combination with machine learning and neural networks are presented. The most important techniques for data retrieval, extraction and matching from the early 20th century to the present day are mentioned. Reference is made to the key retrieval applications that have become part of people's everyday lives. The most advanced search engines are evaluated and compared with each other on the basis of speed, accuracy and practicality. Finally, thoughts on the present and future of information retrieval are collected along with the criteria that can bring in new users, even if the main concern remains user satisfaction and access to information.

Keywords: retrieval, search engine, image, sound, neural networks

Πίνακας περιεχομένων

Κεφάλαιο 1. Εισαγωγή

Πλαίσιο, σκοπός και στόχοι της πτυχιακής εργασίας

Την σημερινή εποχή, τα πολυμέσα επεκτείνονται με ταχύτατους ρυθμούς. Η ανάκτηση τους κρίνεται αναγκαία από την υπέρμετρη χρήση ηλεκτρονικών συσκευών. Άλλοτε για προσωπικούς ή ακόμα και επαγγελματικούς λόγους, με απώτερο σκοπό την ικανοποίηση κάποιου αιτήματος. Εξαιτίας αυτού η έρευνα ,που ακολουθεί, εστιάζει στη μελέτη μηχανών αναζήτησης σχετικές με την εικόνα και τον ήχο. Μία ιστορική αναδρομή μας ωθεί στην ευκολότερη κατανόηση του ζητήματος. Τέλος, αξιολογούνται μέθοδοι και εφαρμογές για τον εντοπισμό πλεονεκτημάτων και μειονεκτημάτων, αλλά και πιθανών βελτιώσεων.

Στόχοι της εργασίας είναι:

- Να οριστούν τα πολυμέσα και οι μηχανές αναζήτησης.
- Να παρουσιαστεί η ιστορική αναδρομή των μέσων ανάκτησης.
- Να αναλυθούν και να αξιολογηθούν οι σύγχρονες και μη μέθοδοι ανάκτησης.
- Να παρουσιαστούν οι διαθέσιμες πολυμεσικές μηχανές αναζήτησης.

Μεθοδολογία

Η μεθοδολογία που ακολουθεί είναι η εξής:

-Βιβλιογραφική επισκόπηση σε έντυπη βιβλιογραφία και στο διαδίκτυο και πιο συγκεκριμένα σε επιστημονικές δημοσιεύσεις σε βιβλιογραφικές βάσεις δεδομένων (Scopus, Google scholar, κλπ).

- Ταξινόμηση με βάση την απήχηση της πηγής/μεθοδολογίας, σε ήχο ή/και εικόνα,

στο επιστημονικό κοινό και περιγραφή-σύγκριση των πιο αποτελεσματικών και γνωστών μηχανών αναζήτησης ή/και εφαρμογών.

Περιορισμοί

Η ανάκτηση ηλεκτρονικών δεδομένων αφορά πολλές κατηγορίες υλικού. Στην εργασία αυτή θα μιλήσουμε μόνο για την διαδικασία, τις μεθόδους και τις εφαρμογές της ανάκτησης για την εικόνα και τον ήχο.

Ορισμοί

Πολυμέσο (Multimedia): Εικόνα, ήχος, βίντεο και κείμενο.

Μηχανή αναζήτησης (Search engine): Λογισμικό που χρησιμοποιείται για την αναζήτηση δεδομένων.

Ανάκτηση (Retrieval): Δυνατότητα επαναφοράς αποθηκευμένου υλικού στη μνήμη υπολογιστή.

Πολυμεσική ανάκτηση (Multimedia retrieval): Ανάκτηση που στηρίζεται στα πολυμέσα όπως εικόνα και ήχος.

Τεχνητή νοημοσύνη (Artificial intelligence): Επιστημονικό πεδίο που εστιάζει στην νοημοσύνη των μηχανών.

Machine learning: Η διαδικασία αυτόματης εκπαίδευσης του υπολογιστή.

Deep learning: Μέθοδος machine learning που λειτουργεί πάνω στα νευρωνικά δίκτυα**.**

Νευρωνικά δίκτυα (Neural Networks): Σύνολο τεχνιτών νευρώνων βασισμένοι στον ανθρώπινο εγκέφαλο**.**

Οργάνωση Κεφαλαίων

Η δομή της εργασίας έχει ως εξής. Στο πρώτο κεφάλαιο περιγράφεται το θέμα, ο σκοπός και η δομή της εργασίας, μαζί με τους βασικούς ορισμούς του θέματος. Στο δεύτερο κεφάλαιο αναφέρεται το επιστημονικό πεδίο της ανάκτησης πληροφοριών, οι λόγοι ανάπτυξης των τελευταίων χρόνων και η σημασία της στην καθημερινότητα. Επίσης, παρουσιάζονται οι βασικές έννοιες και τεχνολογίες των πεδίων που λειτουργούν πάνω στην ανάπτυξη της ανάκτησης του ήχου και της εικόνας. Το τρίτο κεφάλαιο αφορά αποκλειστικά την ανάκτηση εικόνας. Γίνεται ιστορική αναδρομή και αναφέρονται οι πιο σημαντικές έννοιες. Παρουσιάζονται οι σύγχρονοι μέθοδοι ανάκτησης και η σημασία της κάθε μίας. Μετά, αναφέρονται οι μηχανές αναζήτησης που χρησιμοποιούν την τεχνολογία της ανάκτησης εικόνας και περιγράφονται οι δυνατότητες και οι αδυναμίες της κάθε μίας. Στο τέταρτο και πέμπτο κεφάλαιο η δομή του κεφαλαίου είναι ίδια με το προηγούμενο, αλλά σχετίζεται με την ανάκτηση του ήχου. Ο ήχος χωρίζεται σε ανάκτηση μουσικής και αναγνώριση της ομιλίας (φωνητικές εντολές). Τέλος, στο έκτο κεφάλαιο γίνεται η αξιολόγηση των πιο αναγνωρισμένων πολυμεσικών μηχανών αναζήτησης καθώς και η ανακεφαλαίωση όσων συλλέξαμε.

Κεφάλαιο 2. Η ανάκτηση πληροφοριών τον 21ο αιώνα

Ανάκτηση Πληροφοριών

Η ανάκτηση πληροφοριών (information retrieval) ορίζεται ως "η εύρεση υλικού σε μεγάλες συλλογές που ικανοποιεί τις πληροφοριακές ανάγκες του χρηστή" (Manning, Raghavan, & Schütze, 2008). Το υλικό είναι αποθηκευμένο ηλεκτρονικά με την μορφή εικόνας, ήχου και κειμένου. Ο χρήστης ανακτά την πληροφορία δίνοντας δεδομένα. Η δραστηριότητα αυτή δεν αφορά μόνο την επιστημονική κοινότητα αλλά απασχολεί και τον καθημερινό άνθρωπο (“Η ανάκτηση πληροφοριών είναι πολύ σημαντική λειτουργία και ενδιαφέρει όλους τους χρήστες (Zaidi, 2019)”). Οτιδήποτε υπάρχει σε ένα υπολογιστικό σύστημα μπορεί να εντοπιστεί. Η ανάκτηση πληροφοριών ως επιστημονικό πεδίο, εστιάζει στην επικοινωνία μεταξύ ανθρώπου και συστήματος στοχεύοντας στην εξυπηρέτηση του χρήστη.

Το κάθε είδος υλικού φέρει και διαφορετική πορεία μέσα στο χρόνο. Πιο συγκεκριμένα, η ανάκτηση κειμένου προηγήθηκε της εικόνας και του ήχου. Αν και πρόκειται για πεδίο που έχει ιστορία δεκαετιών, δεν παρουσίασε παρόμοια πρόοδο με άλλα υπολογιστικά πεδία. Οι εφαρμογές και η ερευνά γύρω του ήταν ελάχιστες. Χρειάστηκαν δεκαετίες για να φτάσουμε στην σημερινή κατάσταση.

Η διαρκής ανάπτυξη της υπολογιστικής δύναμης: Από την δημιουργία του πρώτου υπολογιστή, παρουσιάζεται τεράστια πρόοδος. Τα εξαρτήματα του αποκτούν ταχύτητα και νέες δυνατότητες. Σημαντική η συμβολή του "νόμου του Moore". Ο Gordon Moore το 1965, είχε αναφέρει ότι κάθε δυο χρόνια ο αριθμός των τρανζίστορ (transistor) που θα χωρούν σε ένα συγκεκριμένο χώρο θα διπλασιαστεί και μαζί με αυτό θα μειωθεί η τιμή των υπολογιστών στο μισό. Η πρόβλεψη ήταν πετυχημένη, και ώθησε τις εταιρείες σε περεταίρω χρηματοδότηση και έρευνα (Chandra, 2018).

Σήμερα κάθε μέρος του υπολογιστή και των κινητών ηλεκτρονικών συσκευών, όπως είναι ο επεξεργαστής (CPU), τα γραφικά (GPU) και η μνήμη (RAM), υπολογίζουν εκατομμύρια πράξεις ταυτόχρονα, με τον αριθμό να αυξάνεται συνεχώς. Από την στιγμή που έγινε υπαρκτή η αποθήκευση και άλλων πολυμέσων πέραν του κειμένου, η ανάκτηση έγινε περίπλοκη. Η υπολογιστική δύναμη που απαιτείται από τα σύγχρονα μοντέλα, είναι μεγάλη. Πριν από μερικά χρόνια δεν ήταν καν δυνατή η ανάπτυξη ενός τέτοιου συστήματος, λόγω της πολυπλοκότητας του.

Προσβασιμότητα: Οι πρώτοι υπολογιστές ήταν πολύ μεγάλοι και υπερβολικά ακριβοί για τον μέσο άνθρωπο. Με την πάροδο του χρόνου, το μέγεθος μίκρυνε και η τιμή μειώθηκε. Η ιδέα για έναν προσιτό προσωπικό υπολογιστή ξεκίνησε την δεκαετία του 1950, όμως η πραγματική του άφιξη έγινε την δεκαετία του 1970. Κάθε χρόνο, όσο το κόστος μειωνόταν, οι χρήστες ηλεκτρονικών υπολογιστών αυξάνονται ραγδαία. Σημαντική αύξηση έγινε επίσης με την δημιουργία των κινητών ηλεκτρονικών συσκευών, όπως το λάπτοπ, το τάμπλετ και το τηλέφωνο.

Ανακάλυψη του διαδικτύου: Η κυκλοφορία του Παγκόσμιου Ιστού το 1992, βοήθησε σημαντικά στην ανάπτυξη μεθόδων ανάκτησης. Η διαρκής πρόσβαση στο διαδίκτυο έγινε απαραίτητη, καθώς δισεκατομμύρια χρήστες έχουν την δυνατότητα να δημιουργήσουν υλικό και να το αποθηκεύσουν. Στην ανάκτηση πληροφοριών στο διαδίκτυο, βοήθησαν πολύ οι μηχανές αναζήτησης. Σήμερα, οποιοσδήποτε κάτοχος ηλεκτρονικής συσκευής δύναται να ανεβάσει στο διαδίκτυο μία εικόνα, ένα κομμάτι ήχου και βίντεο (Computer history museum, 2021).

Όλα τα παραπάνω οδήγησαν στην σημερινή έννοια που ονομάζουμε ανάκτηση. Χάρη στην επιστημονική κοινότητα που αφιέρωσε δεκαετίες έρευνας στο αντικείμενο, δημιούργησε νέες τεχνολογίες και τεχνικές, προσιτές για όλους.

Οι τεχνολογίες ανάκτησης

Τα πρώτα συστήματα ανάκτησης ξεκίνησαν από την αναζήτηση κειμένου. Το πρώτο βήμα έγινε με την εύρεση του ονόματος του αρχείου, τις λέξεις-κλειδιά μέσα στο ηλεκτρονικό έγγραφο και τα εργαλεία που περιορίζουν την αναζήτηση με την χρήση πεδίων (αλφαβητική σειρά, χρονολογία, μέγεθος και τύπος αρχείου). Εδώ και μερικές δεκαετίες, έχει επικεντρωθεί το ενδιαφέρον στο πολυμεσικό υλικό.

Για αυτό, έχουν δημιουργηθεί νέα επιστημονικά πεδία που ασχολούνται σε ένα βαθμό με την ανάκτηση του υλικού αυτού. Η αρχή της εφαρμογής της ανάκτησης του ήχου και της εικόνας έγινε με τον πιο προσιτό και εφαρμόσιμο τρόπο μέσα από ένα Text-based περιβάλλον το οποίο έθεσε και τις βάσεις για τις προοπτικές της ανάκτησης του μέλλοντος. Τον 21ο αιώνα, μια σειρά από νέες τεχνολογίες έχουν κάνει την έρευνα ανάκτησης πιο προσιτή από ποτέ.

Machine learning

Το machine learning είναι ένας κλάδος της τεχνητής νοημοσύνης (artificial intelligence) και ένα σημαντικό κομμάτι του αναπτυσσόμενου τομέα της επιστήμης των δεδομένων (data science).

Η λειτουργία του machine learning στηρίζεται στο γεγονός ότι μια μηχανή, με τη χρήση πολύπλοκων στατιστικών μεθόδων, μεγάλων ποσοτήτων δεδομένων και αλγορίθμων για την ανάλυση των δεδομένων, εκπαιδεύεται στο να κάνει ταξινομήσεις ή προβλέψεις χωρίς να χρειαστεί προγραμματισμό άμεσα από τον άνθρωπο, επιτρέποντας στο μέλλον να μάθει η μηχανή πώς να εκτελεί μια εργασία αυτόματα (Education, I. (2020).

Είναι μια σχετικά καινούρια τεχνολογία που αναπτύσσεται και βελτιώνεται συνεχώς. Καθώς εξελίσσεται, προσπαθεί να κάνει τη ζωή μας ευκολότερη. Η συγκεκριμένη τεχνολογία ήδη εφαρμόζεται, αφού υπάρχουν αυτοματοποιημένα συστήματα για την μεταφορά εμπορίου, εικονικοί προσωπικοί βοηθοί όπως Alexa, Siri και Google Assistant και συστήματα ασφαλείας που εντοπίζουν τον κάθε κίνδυνο και δρουν αντίστοιχα. Οι εφαρμογές του machine learning εντοπίζονται πέρα από την καθημερινότητα μας και στον επιχειρηματικό τομέα.

Computer vision

Ένας από τους καλύτερους τομείς εφαρμογής του machine learning για πολλά χρόνια ήταν το computer vision.

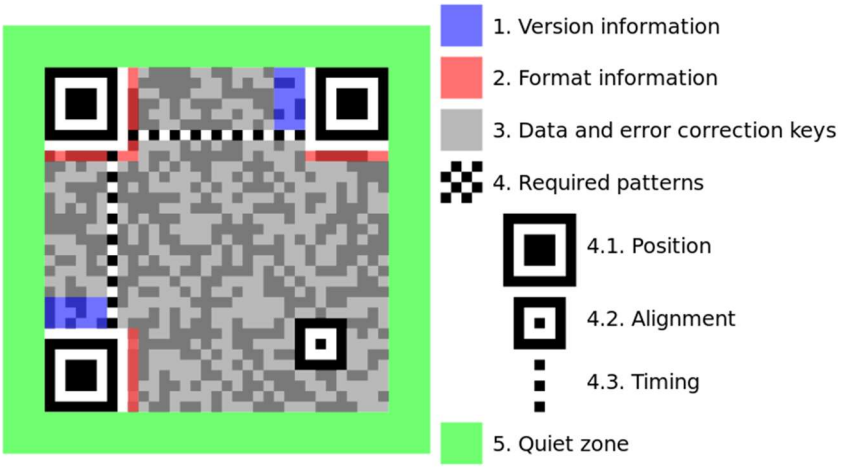
Το computer vision ανήκει κυρίως στον τομέα της μηχανικής αλλά απευθύνεται σε διάφορους τομείς εφαρμογής για την επίλυση κρίσιμων προβλημάτων της πραγματικής ζωής. Σκοπός του είναι να φτιάξει μηχανές οι οποίες είναι ικανές να συλλέγουν, επεξεργάζονται και να αναλύουν τις πληροφορίες της εικόνας και οποιαδήποτε άλλης οπτικής πληροφορίας, όπως η ανθρώπινη όραση (Prince, 2012).

Εδώ και 6 δεκαετίες, πολλοί ερευνητές προσπάθησαν να βρουν τρόπους για να πετύχουν αυτόν τον σκοπό και την δημιουργία του τέλειου μηχανήματος computer vision, χωρίς όμως αυτό ακόμα να είναι δυνατό.

Πρόκειται για ένα εξαιρετικά δύσκολο έργο, καθώς το computer vision είναι ένα αρκετά πολύπλοκο πεδίο και η χρήση αλγορίθμων βασισμένων στην ανθρώπινη βιολογική όραση αποτελεί μια τεράστια πρόκληση. Πιο συγκεκριμένα διότι η οπτική πληροφορία ερμηνεύεται πιο δύσκολα. Ένας αλγόριθμος computer vision μπορεί να χρονοτριβεί αρκετά για τον εντοπισμό των όριων ενός αντικειμένου.

Το machine learning βοήθησε σημαντικά το computer vision ως προς την αναγνώριση και τον εντοπισμό δεδομένων πάνω στην εικόνα και ύστερα στην δημιουργία πληροφοριών από αυτές, προσφέροντας αποτελεσματικές μεθόδους ανάκτησης, επεξεργασίας εικόνας (image processing) και εστίασης αντικειμένων (object focus).

Τα πλεονεκτήματα που προκύπτουν έχουν εφαρμοστεί σε χιλιάδες οργανισμούς. Αυτή την στιγμή, εντοπίζονται σε βιομηχανίες ενέργειας, υπηρεσίες κοινής ωφέλειας, κατασκευαστικές εταιρίες ακόμη και στην αυτοκινητοβιομηχανία.



Εικόνα 1. Περιγραφή ενός QR Code. Το κάθε σημείο της εικόνας είναι μία σειρά από δεδομένα που κάνουν το QR μοναδικό.

https://commons.wikimedia.org/w/index.php?curid=25534216

Με την πρόοδο της τεχνολογίας, η αγορά μεγαλώνει. Πέρα από τους οργανισμούς και τις επιχειρήσεις, η τεχνολογία του computer vision είναι πλέον διαθέσιμη σε τεράστιο αριθμό ανθρώπων. Η πιο γνωστή συνήθεια είναι η σάρωση barcode και QR Codes.

Μπορούμε επίσης να εντοπίζουμε την χρησιμότητα της τεχνολογίας του computer vision μέσα από πολλές ηλεκτρονικές συσκευές, όπως κινητά, κάμερες, κονσόλες παιχνιδιών αλλά και καθιερωμένες ηλεκτρονικές εργασίες όπως object detection, face detection, hand writing recognition, content-based image retrieval, κλπ.

Εικόνα 2. Computer Vision. https://soft-cluster.com/wp-content/uploads/2020/11/Computer-Vision.png

Εφαρμογές του Computer Vision και Image Retrieval

- Image classification και Computer Vision

Σε συνδυασμό με την γνώση που διαδίδεται από το computer vision, το image classification είναι από τις βασικές λειτουργίες ενός συστήματος ανάκτησης. Ταξινομεί αυτό που αναγνωρίζει μέσα σε μια εικόνα και προβλέπει με ακρίβεια οτιδήποτε ανήκει σε μια συγκεκριμένη κατηγορία (π.χ. "Αυτή είναι εικόνα μιας γάτας").

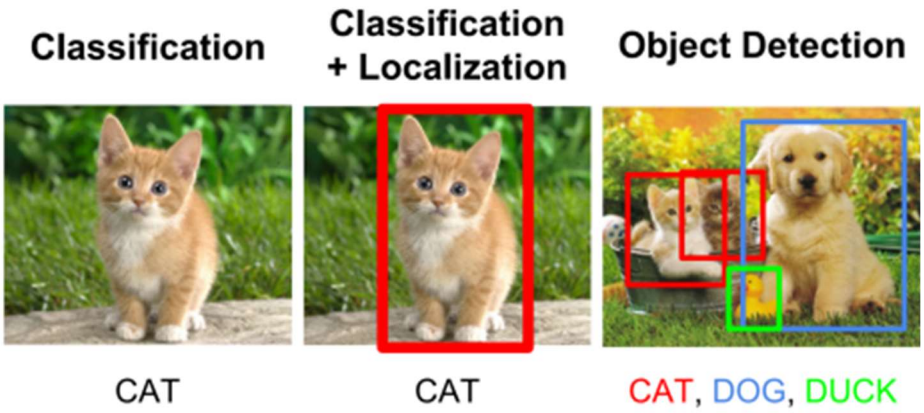
- Object detection και Computer Vision

Το object detection ενσωματώνει τις τεχνολογίες του computer vision και τη διαδικασία του image classification. Παράλληλα με τις προσεγγίσεις των machine learning και deep learning τεχνολογιών, προσδιορίζει μια συγκεκριμένη κατηγορία εικόνας, ανιχνεύει το περιεχόμενο της και τέλος την κατατάσσει με την αντίστοιχη κατηγόρια αποτελεσμάτων. Εφαρμόζεται σε μεγάλο βαθμό κατά την αναγνώριση προσώπων.

- Image Processing

Η επεξεργασία εικόνας (image processing) αποτελεί υποσύνολο του computer vision και είναι ο τομέας της βελτίωσης των εικόνων, εισάγοντας πολλούς παραμέτρους και χαρακτηριστικά των εικόνων. Επικεντρώνεται κυρίως στην επεξεργασία των ακατέργαστων εικόνων (raw images).

Η λειτουργία του είναι η εξής: Εισάγεται μια εικόνα, εκτελούνται οι κατάλληλες μετατροπές σε αυτή (όξυνση (sharpening), αποθορυβοποίηση (denoising), λείανση (smoothing), περικοπή (cut), κλπ.) και στο τέλος εξάγεται η νέα επεξεργασμένη εικόνα.



Εικόνα 3. Παράδειγμα Image Classification και Object Classification.
https://appsilondatascience.com/assets/uploads/2018/08/types.png

- Computer Vision και CBIR

Το CBIR χρησιμοποιεί το computer vision για να περιηγηθεί, αναζητήσει και ανακτήσει εικόνες από μία μεγάλη βάση δεδομένων, χρησιμοποιώντας το περιεχόμενο ως βασικό στοιχείο αντί για μεταδεδομένα ή τις ετικέτες (tags) της εικόνας. Αυτή η εργασία επιτυγχάνει τον αυτόματο σχολιασμό εικόνας, με στόχο να αντικαταστήσει την χειροκίνητη διαδικασία της επισήμανσης ετικετών.

Deep Learning

Εάν και αποτελεί μονάχα ένα μέρος του machine learning, είναι το ανερχόμενο και με μεγαλύτερη προοπτική εργαλείο του. Οι ερευνητές προβλέπουν ότι θα είναι αυτό που θα καταφέρει να ανεβάσει το επίπεδο της ποιότητας ανάκτησης για τα επόμενα χρόνια (Chandra, 2018).

Η ανάγκη του ανθρώπου για μηχανές ικανές να εφαρμόζουν καθημερινές λειτουργίες, ξεκινάει από τα αρχαία χρόνια. Η ανάπτυξη της τεχνολογίας και οι μαθηματικές εξισώσεις, μαζί με πολύπλοκους αλγορίθμους έχουν βοηθήσει στην εφαρμογή αυτής της ανάγκης.

Το μεγαλύτερο ζήτημα που προέκυψε από την δημιουργία της τεχνητής νοημοσύνης, είναι ως προς την αναπαράσταση που χρειάζεται η μηχανή ώστε να είναι ικανή να καταλάβει τις λειτουργείς που απαιτούνται. Για παράδειγμα, ένας άνθρωπος μπορεί εύκολα να αναγνωρίσει έναν αριθμό, έναν χαρακτήρα, την γλώσσα κτλ.

Για αυτό και η δημιουργία του deep learning στηρίχθηκε πάνω στην ιδέα ότι με την χρήση εξειδικευμένων αλγορίθμων, μπορούμε να μιμηθούμε τον ανθρώπινο εγκέφαλο ώστε να λύνουμε τα προβλήματα γρηγορότερα και πιο αποτελεσματικά.

Ο ανθρώπινος εγκέφαλος είναι ένας δυνατός υπολογιστής που μπορεί να λύνει προβλήματα σε σχετικά μικρό χρονικό διάστημα που ένας υπολογιστής θα χρειαζόταν χρόνια για να τα λύσει. Η αρχή της ανάπτυξης ενός τέτοιου συστήματος ξεκίνησε το 1943, όταν ο νευρολόγος Warren MuCulloch και ο θεωρητικός της λογικής, Walter Pitts σχεδίασαν το πρώτο υπολογιστικό μοντέλο ενός νευρώνα. Η λογική πίσω από αυτό το σύστημα είναι η χρήση boolean εντολών για τον υπολογισμό του αποτελέσματος.

Πολλές ανακαλύψεις έχουν κριθεί σημαντικές στην ανάπτυξη του deep learning, όπως οι αποκωδικοποιητές των Rumelhart, Hinton και Williams το 1986, και το LeNet από τον LeCun το 1990. Η πιο σημαντική από αυτές θεωρείται η δημιουργία του Deep Belief Network το 2006 από τον Geoffrey Hinton. Πρόκειται για ένα μοντέλο αναπαράστασης που αποτελείται από πολλαπλές στρώσεις, οι οποίες με μη επιβλεπόμενη εκπαίδευση (τεχνική machine learning στην οποία οι χρήστες δεν χρειάζεται να επιβλέπουν το μοντέλο), ανακτούν την πληροφορία (Chandra, 2018).

Ο πιο πετυχημένος τύπος νευρωνικού δικτύου είναι το Convolutional neural network (CNN) και οι πιο συνηθισμένες προσεγγίσεις για τις διάφορες εργασίες ανάκτησης εικόνας και ήχου οπού χρησιμοποιείται το deep learning βασίζονται σε convolutional neural networks. Πρόκειται για ένα είδος deep neural network που περιέχει ένα σύνολο επιπέδων ανάμεσα στην είσοδο και έξοδο των δεδομένων του. Ένα τεχνητό νευρωνικό δίκτυο είναι ένα υπολογιστικό σύστημα εμπνευσμένο από το βιολογικό νευρωνικό δίκτυο του εγκεφάλου. Ένας από τους λόγους της επιτυχίας του είναι η δυνατότητα να εκπαιδεύεται μέσα από μία μεγάλη συλλογή δεδομένων και να εφαρμόζει σε σύντομο χρονικό διάστημα τις εργασίες που του ανατίθενται. Μερικά από τα πιο γνωστά CNN είναι το VGGNet, GoogLeNet και AlexNet. Αξίζει να επισημανθεί ότι το 2012, μέσα από το CNN μοντέλο AlexNet, μειώθηκε σημαντικά το ποσοστό σφάλματος στην αναγνώριση των εικόνων.

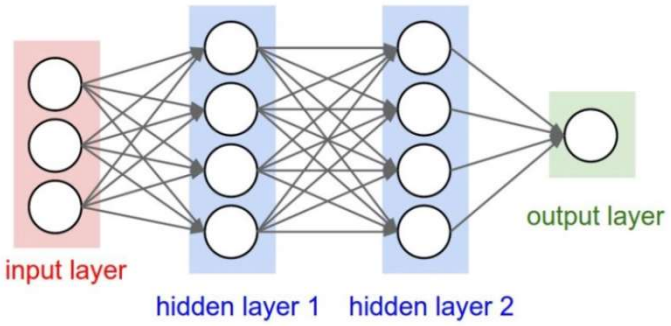
Δομή deep learning

Σε ένα τυπικό νευρωνικό δίκτυο υπάρχουν τρεις τύποι επιπέδων:

- **Επίπεδο εισόδου (Input layer):** Τα αρχικά δεδομένα για το νευρωνικό δίκτυο. Αυτό το επίπεδο δέχεται τα δεδομένα και τα μεταβιβάζει στο υπόλοιπο δίκτυο.
- **Κρυφά επίπεδα (Hidden layers):** Το ενδιάμεσο επίπεδο μεταξύ του επιπέδου εισόδου και του επιπέδου εξόδου και το μέρος όπου γίνεται όλη η υπολογιστική

διαδικασία. Είναι υπεύθυνα για τις εξαιρετικές επιδόσεις και την πολυπλοκότητα των νευρωνικών δικτύων. Εκτελούν πολλαπλές λειτουργίες ταυτόχρονα, όπως η μετατροπή των δεδομένων, κ.λπ.

- **Επίπεδο εξόδου (Output layer)**: Παράγει και περιέχει το αποτέλεσμα ή την έξοδο του προβλήματος των δεδομένων εισόδου.



Εικόνα 4. Τα τρία επίπεδα του deep learning μοντέλου.

https://www.i2tutorials.com/what-are-different-layers-in-neural-networks/

Τέτοια δίκτυα CNN έχουν αναπτυχθεί από πολλούς οργανισμούς με ανάμεικτα αποτελέσματα. Ένα πολύ διαδεδομένο δίκτυο είναι το GoogLeNet. Αποτελείται από 22 επίπεδα και οι ερευνητές κατά την διάρκεια ανάπτυξης του, ανακάλυψαν ότι όσο περισσότερα επίπεδα υπάρχουν στο νευρωνικό δίκτυο, τόσο καλύτερη θα είναι η απόδοσή τους. Το deep learning μεγαλώνει και γίνεται πιο χρήσιμο όσο αναπτύσσεται η υπολογιστική δύναμη και όσο η πληροφορία πάνω σε αυτό αυξάνεται.

Στην ανάκτηση της εικόνας προσφέρει την δυνατότητα συλλογής υψηλού και χαμηλού επιπέδου χαρακτηριστικών και τον συνδυασμό της πληροφορίας αυτής με σκοπό την επιτυχή ανάκτηση. Ο συνδυασμός των νευρωνικών μοντέλων, στατιστικών μοντέλων και η πληθώρα των εξειδικευμένων αλγορίθμων για την εξαγωγή των χαρακτηριστικών, έχει κάνει το deep learning ένα αρκετά αποτελεσματικό μοντέλο.

Από την δημιουργία του, έχει καταφέρει να λύσει πολύπλοκες εφαρμογές με μεγάλη επιτυχία. Τα πιο σημαντικά προβλήματα που λύνει το deep learning στον τομέα της ανάκτησης εικόνας είναι (Voulodimos, 2019):

Αναγνώριση αντικειμένων (Object recognition).

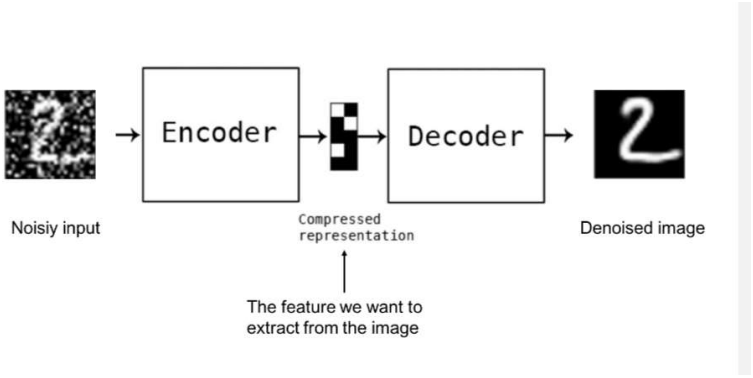
Human pose estimation: Αναγνωρίζει τους ανθρώπους και την στάση τους.

Semantic segmentation: Αναγνωρίζει όμοιες οντότητες και τις κατηγοριοποιεί.

Παράδειγμα επίδειξης Νευρικού δικτύου deep learning.

Το denoising autoencoder ή αυτόματος κωδικοποιητής αποθορυβοποίησης είναι ένα είδος νευρωνικού δικτύου που αφαιρεί τον "θόρυβο" από τις εικόνες. Μέσα από μία σειρά από δοκιμαστικές εικόνες, το νευρωνικό δίκτυο μαθαίνει τα πιο σημαντικά και ιδιαίτερα χαρακτηριστικά τους.

Έπειτα, εξάγει χαρακτηριστικά από παρομοίου τύπου εικόνες και δημιουργήσει το επιθυμητό αποτέλεσμα για την αποθορυβοποίηση οποιασδήποτε εικόνας (Voulodimos, 2019).



Εικόνα 5. Η δομή ενός denoising autoencoder.

https://camo.githubusercontent.com/64eb8f839ccb2cd4150f31bada17b47cd86a712ef440a04a04b9df4a6cfc6ab1/687474706733a2f2f63646e2d696d616765732d312e6d656469756d2e636f6d2f6d61782f313830302f312a47305634647a34524b544b47706562656f53574230412e706e67

Μέσα από ζήτηση που προκύπτει για την ανάπτυξη των συγχρόνων συστημάτων image retrieval, το deep learning συνθέτει ένα μεγάλο κομμάτι ανάπτυξης για την εύρεση και εξαγωγή των χαρακτηριστικών εικόνας. Όμως η επιρροή του δεν σταματά εκεί καθώς συνεχίζει να αναδύεται στον τομέα της ανάκτησης της μουσικής πληροφορίας (Music Information Retrieval (MIR) και της ανάκτησης ήχου (Sound Retrieval) (Choi, Fazekas, Cho & Sandler, 2018).

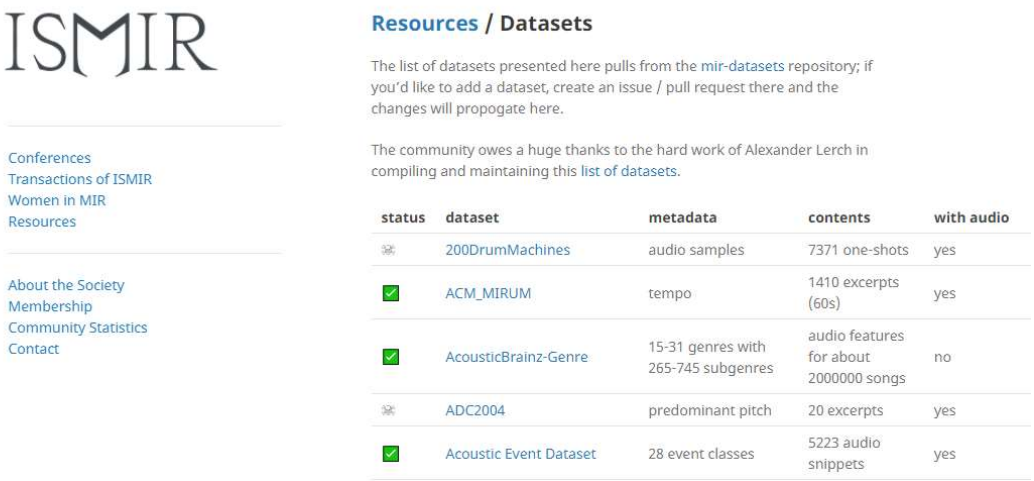
Τα CNN αποτελούν και σε αυτήν την περίπτωση ένα πρακτικό πρότυπο για την ανάκτηση του ήχου μέσω του deep learning. Χρησιμοποιείται για την εκτέλεση εργασιών όπως την κατάταξη του είδους της μουσικής (music classification), την εκτίμηση της ομοιότητας του ήχου (similarity estimation), την ανίχνευση και τον εντοπισμό του σημείου έναρξης του ήχου (onset detection). Με την υποστήριξη αρκετών νευρωνικών δικτύων όπως τα Recurrent Neural Networks, για εργασίες ανίχνευσης της έναρξης ηχητικού κομματιού (onset detection), ανίχνευση ρυθμών (beats) ή διάφορων ηχητικών γεγονότων (audio-event

detection) και το Siamese Networks ή Twin Neural Network, για την εκμάθηση αναπαραστάσεων και συναρτήσεων απόστασης ήχου οπού είναι αρκετά χρήσιμα και ειδικά για να βρεθούν ομοιότητες μουσικών κομματιών (similarity estimation) (Voulodimos, 2019).

Η αρχή του deep learning στον τομέα του MIR, έγινε το 2010 όταν επιστημονικά άρθρα ανέφεραν την ανάκτηση της μουσικής και τα πλεονεκτήματα που θα είχε αν το εκμεταλλεύονταν. Από τότε πολλοί επιστήμονες πιστεύουν ότι είναι το μέλλον της ανάκτησης της μουσικής πληροφορίας. Όπως και στην εικόνα, έτσι και στον ήχο, τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στρώματα (layers) για την εξαγωγή των δεδομένων. Στην περίπτωση της μουσικής, τα δεδομένα που συλλέγονται είναι τα ηχητικά κύματα (sound waves). Μέσα από datasets, το νευρωνικό δίκτυο μπορεί να εκπαιδευτεί ώστε να εφαρμόζει αυτόματα τις κατάλληλες λειτουργίες (Voulodimos, 2019).

Dataset

Για την εκπαίδευση των νευρωνικών δικτύων, υπάρχουν μεγάλες συλλογές από δεδομένα. Ονομάζονται datasets και παράγονται από κυβερνητικές υπηρεσίες ή μη κερδοσκοπικούς οργανισμούς (Shorten & Khoshgoftaar, 2019). Συνήθως μπορεί ο καθένας να τα αποκτήσει χωρίς κάποια χρέωση και να γίνει η λήψη τους από το διαδίκτυο. Ένας τέτοιος οργανισμός είναι το International Society for Music Information Retrieval (ISMIR) και έχει ως βασικό στόχο να προωθήσει την πρόσβαση, την οργάνωση και την κατανόηση της μουσικής πληροφορίας.



Εικόνα 6. Μέσα από την κεντρική ιστοσελίδα του ISMIR, οι χρήστες μπορούν να περιηγηθούν και να και να επιλέξουν το dataset ποου τους ταιριάζει. https://www.ismir.net/

Προσφέρει έναν κατάλογο από δωρεάν datasets όπως :

- **AcousticBrainz Genre Dataset:** είναι μια συλλογή από διαφορετικές πηγές μεταδεδομένων ήχου οπού επιτρέπει στους ερευνητές να διερευνήσουν πώς τα ίδια μουσικά κομμάτια σχολιάζονται από διαφορετικές κοινότητες ακολουθώντας τις δικές τους ταξινομήσεις ειδών μουσικής.
- **Aligned Scores and Performances (ASAP) dataset:** στοχεύει σε συγκεκριμένες εργασίες ανάκτησης όπως το beat/downbeat tracking, signature estimation, time signature estimation.
- **AudioSet:** περιέχει 632 ηχητικά στιγμιότυπα και μια συλλογή 2.084.320 ηχητικών αποσπασμάτων διάρκειας 10 δευτερολέπτων που προέρχονται από βίντεο στο YouTube.

Αυτά και πολλά άλλα βρίσκονται σε ενεργή και ανενεργή κατάσταση.

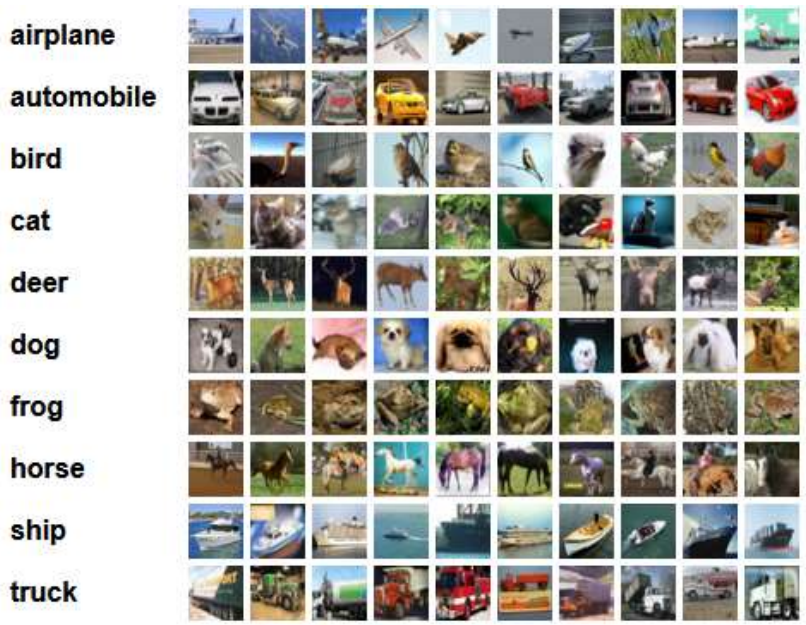
Στις εικόνες, τα dataset βοηθούν στην αξιολόγηση των μοντέλων και στην βελτίωση των δυνατοτήτων τους. Στα περισσότερα, η συλλογή μπορεί να περιλαμβάνει μερικές χιλιάδες έως και εκατομμύρια εικόνες. Αυτές κατηγοριοποιούνται με βάση το τι παρουσιάζουν (κτήρια, ψάρια, αεροπλάνα κλπ.). Τα μοντέλα ανάκτησης εξάγουν τα χαρακτηριστικά (χρώματα, γωνίες) των εικόνων και ύστερα επαληθεύεται αν η εξαγωγή των χαρακτηριστικών αυτών ήταν η σωστή (Shorten & Khoshgoftaar, 2019).

Ως προς την αξιολόγηση, τα dataset μπορούν να γίνουν μέτρο σύγκρισης μεταξύ μοντέλων ανάκτησης και να βρεθούν τα πιο δυνατά και αδύναμα σημεία του κάθε μοντέλου. Κάθε χρόνο, διεξάγονται διαγωνισμοί που κατατάσσουν τα μοντέλα ανάκτησης για αξιολόγηση και τους δίνεται ένας τελικός βαθμός. Έτσι μπορούν οι ερευνητές να συγκρίνουν την πρόοδο της τεχνολογίας και της ανάκτησης ανά τα χρόνια. Ένας τέτοιος διαγωνισμός, στην περίπτωση των datasets εικόνας είναι το ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

Το ILSVRC βασίζεται στο dataset του Imagenet και από το 2010 έως το 2017 οργάνωνε διαγωνισμούς που αξιολογούν τους αλγόριθμους που εξειδικεύονται στην αναγνώριση των αντικειμένων και στην ταξινόμηση των εικόνων. Επιστήμονες από όλο τον κόσμο είχαν την δυνατότητα να δοκιμάσουν τον αλγόριθμό τους μέσα από δεκάδες χιλιάδες δοκιμαστικές εικόνες. Κάθε χρόνο, η ακρίβεια αυξήθηκε και τα μοντέλα ήταν πολύ ανταγωνιστικά μεταξύ τους, κυρίως από την χρήση των νευρωνικών δικτύων και μετά. Το dataset του Imagenet είναι ελεύθερο στην χρήση, διαθέσιμο στην ιστοσελίδα τους.

Τα πιο γνωστά dataset για την αναγνώριση εικόνων είναι:

- **ImageNet**: μια συνεχής ερευνητική προσπάθεια που παρέχει σε ερευνητές δεδομένα εικόνας για την εκπαίδευση μοντέλων αναγνώρισης αντικειμένων (object recognition models) μεγάλης κλίμακας. ("About ImageNet", 2022).
- **SUN09:** αποτελείται από 12.000 εικόνες με περισσότερες από 200 κατηγορίες αντικειμένων.
- **Open Images:** εμπεριέχει πάνω από 9.000.000 εικόνες διαφορετικών κατηγοριών με πλούσιο σχολιασμό.
- **LabelMe:** μεγάλη συλλογή εικόνων που χρησιμοποιείται κατά βάση για τον εντοπισμό και την αναγνώριση αντικειμένων.
- **CIFAR-10:** αποτελείται από 60.000 έγχρωμες εικόνες, μεγέθους 32x32 χωρισμένες σε 10 κλάσεις, με 6.000 εικόνες ανά κλάση. Υπάρχουν 50000 εικόνες εκπαίδευσης (training images) και 10.000 δοκιμαστικές εικόνες.



Εικόνα 7. Οι 10 κατηγορίες από το CIFAR-10 (Alex Krizhevsky, 2009).

Η δομή ενός μοντέρνου συστήματος ανάκτησης μέσα από την λειτουργία του Computer Vision

Για την επιτυχή λειτουργία των συστημάτων ανάκτησης, απαιτείται τεράστιος όγκος πληροφορίας. Ένα τέτοιο σύστημα δουλεύει και εκπαιδεύεται μέσα από datasets, με σκοπό να μαθαίνει να αναγνωρίζει τις ιδιαιτερότητες κάθε είδους υλικού. Με νέες τεχνολογίες όπως Deep Learning και CNN να προσφέρουν άμεσα και κάνουν τα συστήματα αυτά πιο επιδέξια από ποτέ. Παρατηρείται ότι όσο εξελίσσονται οι νέες τεχνικές ανάκτησης, τόσο μεγαλώνει το computer vision.

Τα συστήματα που χρησιμοποιούν Computer vision εξοικειώνονται με πιθανά δείγματα εικόνων και αναπτύσσονται μέσα από τα datasets. Το Deep learning, χρησιμοποίει αλγορίθμους που επιτρέπουν σε ένα υπολογιστικό σύστημα να εκπαιδεύεται αυτόνομα για το περιεχόμενο της οπτικής πληροφορίας. Έτσι όταν έχει αποκτήσει αρκετή πληροφορία, όπως προαναφέραμε, το υπολογιστικό σύστημα μαθαίνει να ξεχωρίζει μόνο του τις διαφορετικές εικόνες. Εδώ είναι που παίρνουν μέρος τα CNN. Βοηθούν ένα σύστημα machine learning ή deep learning να αναλύσει τις εικόνες σε pixels, και να δοθούν ετικέτες (tags). Οι ετικέτες αυτές χρησιμοποιούνται για να γίνουν οι συσπειρώσεις (convolutions), οι οποίες μέσα από αυτές, το σύστημα καταφέρνει να βγάλει ένα αποτέλεσμα για το τι εμφανίζεται. Η μοναδική συνδρομή των ανθρώπων στην διαδικασία αυτή, είναι να ελέγχουν την ακρίβεια των προβλέψεων του συστήματος μέσα από μια σειρά από επαναλήψεις, μέχρι να επιτευχθεί η επιθυμητή πρόβλεψη.

Όπως ένας άνθρωπος που βλέπει μία εικόνα από μακρινή απόσταση, έτσι και ένα CNN εντοπίζει πρώτα τις άκρες ενός αντικειμένου ή του περιβάλλοντος. Σε συνδυασμό με απλά σχήματα, συμπληρώνει τις πληροφορίες καθώς εκτελεί επαναλήψεις των προβλέψεών του.

Image Mining

Το Image Mining ή αλλιώς εξόρυξη εικόνας είναι συνώνυμη με την έννοια της εξόρυξης δεδομένων (data mining) και έχει να κάνει με την απόσπαση της γνώσης, των σχέσεων των δεδομένων ή άλλων μοτίβων που μπορεί να είναι εμφανή στο περιεχόμενο της εικόνας και δεν είναι αποθηκευμένα στην βάση δεδομένων.

Η νέα εποχή της προηγμένης τεχνολογίας και της υψηλής αποθηκευτικής δυνατότητας, φέρνει την ανάπτυξη μίας βάσης δεδομένων εικόνων, που διευκολύνει την διαδικασία του image mining. Βοηθάει ιδιαίτερα τα συστήματα ανάκτησης εικόνας στην προσπάθεια ανάπτυξης και βελτίωσης τους.

Χρησιμοποιεί μεθόδους από computer vision, image processing, image retrieval, machine learning και τεχνητής νοημοσύνης (artificial intelligence) (Meherban, 2016).

Ασχολείται κυρίως με την εύρεση άρρητης γνώσης και επικεντρώνεται στην εξαγωγή χαρακτηριστικών και μοτίβων από μεγάλες συλλογές εικόνων.

Μπορεί να γίνει χειροκίνητα, χωρίζοντας τα δεδομένα σε μικρότερα για την επίτευξη ενός συγκεκριμένου μοτίβου, με τη χρήση προγραμμάτων που συλλέγουν τα δεδομένα αυτόματα.

Λειτουργίες Image mining

- Πρώτα, οι εικόνες από μια βάση δεδομένων εικόνων επεξεργάζονται, βελτιώνοντας την ποιότητάς τους.
- Στην συνέχεια, υποβάλλονται σε διάφορες αλλαγές και περνάνε από την διαδικασία εξαγωγής χαρακτηριστικών (feature extraction) για την αναγνώριση και καταγραφή των σημαντικών χαρακτηριστικών τους.
- Μέσα από τις τεχνικές data mining ολοκληρώνεται η εξαγωγή τους.

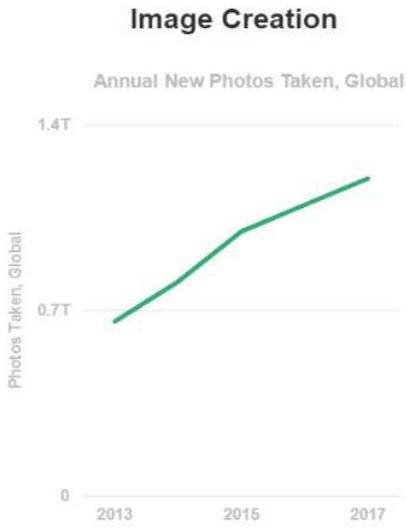
Τα μοτίβα αυτά αξιολογούνται και ερμηνεύονται για την απόκτηση της τελικής γνώσης και μπορεί πλέον να ξεκινήσει την χρήση της στις διάφορες εφαρμογές. Υποστηρίζει ένα ευρύ πεδίο εφαρμογών και επαγγελμάτων, όπως την ιατρική, την γεωργία, την βιομηχανία, την διαστημική έρευνα και τον εκπαιδευτικό τομέα.

Είναι ένας πρόσφατος και πολλά υποσχόμενος τομέας για την εξόρυξη γνώσης από εικόνες, ωστόσο βρίσκεται ακόμη σε ένα πρώιμο στάδιο της ερευνάς του. Για αυτό, πρέπει να γίνουν περεταίρω μελέτες για τη συνεχή ανάπτυξη. Μόνο έτσι θα βελτιωθούν οι τεχνικές όπως η επεξεργασία εικόνας (image processing), η εξαγωγή των χαρακτηριστικών (feature extraction), η κατάτμηση εικόνας (image segmentation) και ο εντοπισμός αντικειμένων (object recognition).

Κεφάλαιο 3. Ανάκτηση εικόνας

Η διάδοση της ανάκτησης εικόνας

Από την τελευταία δεκαετία υπάρχει ραγδαία αύξηση των ενεργών χρηστών του Παγκοσμίου Ιστού και κυρίως των μέσων κοινωνικής δικτύωσης, επηρεάζοντας ταυτόχρονα την παραγωγή του πολυμεσικού υλικού. Ο συνολικός αριθμός των εικόνων που υπάρχουν διαθέσιμες στο διαδίκτυο δεν γίνεται να καταμετρηθεί, καθώς ο αριθμός αυτός σταδιακά αυξάνεται. Το Bond internet report το 2019 εκτίμησε ότι μόνο για το 2017 δημιουργήθηκαν πάνω από 1,2 δισεκατομμύρια εικόνες στο διαδίκτυο, ο διπλάσιος αριθμός σε σύγκριση με το 2013.



Εικόνα 8. Η ετήσια αύξηση της δημιουργίας φωτογραφιών στο διαδίκτυο (Bondcap, 2019).

Η περιγραφή των χαρακτηριστικών της εικόνας σε μία συλλογή μπορεί να γίνει από τον χρήστη σε πολύ μικρό χρονικό διάστημα. Ακόμα και σε μία ηλεκτρονική συλλογή, όπως μίας βιβλιοθήκης ή ενός αρχείου, με τις σωστές εργασίες, η συλλογή μπορεί να γίνει εύκολα και γρήγορα ανακτήσιμη.

Το σύνολο των εικόνων που υπάρχουν στο διαδίκτυο δεν γίνεται να ανακτηθεί με πεδία δημιουργημένα από τον άνθρωπο, αφού ο χρόνος και τα χρήματα που θα χρειαστούν είναι απλησίαστα.

Η εύρεση συγκεκριμένου ηλεκτρονικού υλικού είναι κάτι που απασχολεί όχι μόνο έναν απλό χρήστη αλλά διευρύνεται σε όλους τους επαγγελματικούς κλάδους (προγραμματιστές, επιστήμονες, ιατρούς, καθηγητές κλπ.).

Ο γενικός όρος που έχει προκύψει για την ανάκτηση εικόνας είναι "η διαδικασία της αναζήτησης, περιήγησης και τελικώς την ανάκτησης της εικόνας από μία βάση δεδομένων." Εκτός αυτού, τα συστήματα έχουν αναπτυχθεί ώστε να έχουν την δυνατότητα να αντιστοιχούν, να παρουσιάζουν αποτελέσματα και να ανταπεξέλθουν στις ανάγκες του χρήστη συγκρίνοντας και αξιολογώντας χαρακτηριστικά, έτσι ώστε να βρει στις εικόνες ομοιότητες που σχετίζονται με το περιεχόμενό τους (Elmogy, 2015).

Σύστημα ανάκτησης εικόνας

Σήμερα η ανάκτηση εικόνων αποτελεί μια από τις πιο συχνές ασχολίες των χρηστών του διαδικτύου. Όμως η τεχνολογία αυτή υπάρχει δεκαετίες πριν γίνει γνωστή στο διαδίκτυο. Η ανάπτυξη ενός τέτοιου συστήματος ξεκινάει από την δεκαετία του 1980. Η υλοποίηση έγινε την δεκαετία του 1990 και από τότε χρησιμοποιείται σε τομείς της έρευνας και της διαφήμισης.

Τα περισσότερα έχουν σχεδιαστεί και αναπτυχθεί για ερευνητικούς σκοπούς σε σχολεία, ηλεκτρονικές βιβλιοθήκες, νοσοκομεία και διάφορα συστήματα πληροφόρησης. Μπορούν να χρησιμοποιηθούν για αναζήτηση εικόνας με την χρήση κειμένου, εικόνων ή και οποιασδήποτε άλλης μεθόδου αναζήτησης. Τα τελευταία χρόνια, τα συστήματα έχουν στραφεί στους απλούς χρήστες του διαδικτύου.

Για την εφαρμογή αυτών των συστημάτων, έχουν προκύψει διάφορα μοντέλα ανάκτησης που διαφοροποιούνται στον τρόπο λειτουργίας. Με την πάροδο του χρόνου, τα μοντέλα αυτά είτε γίνονται ξεπερασμένα, είτε αναπτύσσονται περεταίρω. Υπάρχουν μέχρι στιγμής 2 βασικά μοντέλα για την ανάκτηση των εικόνων (Elmogy, 2015):

- Text-based image retrieval (TBIR)
- Content-based image retrieval (CBIR)

Text-Based Image Retrieval Research

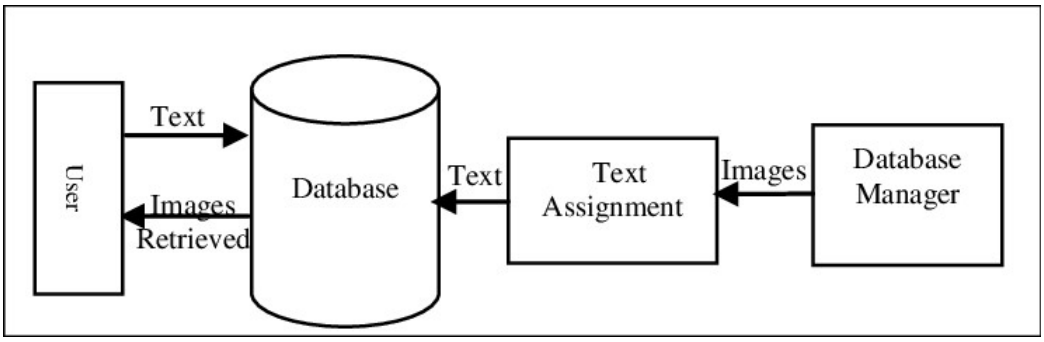
Δεκαετίες πριν γίνει η ψηφιοποίηση των εικόνων πραγματικότητα, η πρόσβαση τους δημόσια ήταν δυνατή μόνο με άδεια από τους διάφορους επιμελητές, αρχειονόμους και βιβλιοθηκονόμους.

Η αρχική εφαρμογή των συστημάτων ανάκτησης εικόνας άρχισε με το κείμενο ως το μοναδικό εργαλείο για την αναζήτηση, γνωστό και ως Text-Based Image Retrieval System.

Οι εικόνες όμως δεν παρέχουν την ίδια πληροφορία με το κείμενο, έτσι ήταν αναγκαίο να βρεθεί ένας νέος τρόπος περιγραφής της εικόνας. Η λύση ήταν να αναπτυχθούν έννοιες και σχήματα ταξινόμησης για την περιγραφή τους.

Τα πρώτα χνάρια του Text-Based Image Retrieval μπορούν να εντοπιστούν γύρω στο 1970. Η ανάκτηση των εικόνων γίνεται με την χρήση κειμένου και πιο συγκεκριμένα με όρους και λέξεις κλειδιά. Αυτά αξιοποιούνται μέσα από μία βάση δεδομένων, η οποία κάνει αναζήτηση τον όρο (μεταδεδομένα) (Elmogy, 2015).

Σε αυτήν την περίπτωση, το περιεχόμενο της εικόνας περιγράφεται από τον άνθρωπο. Ο πιο συνήθης τρόπος αναζήτησης είναι ο χρήστης να ψάξει τον όρο σε μία μηχανή αναζήτησης. Ο όρος αυτός μεταφέρεται στην βάση δεδομένων και αναζητεί οποιαδήποτε εικόνα που περιέχει τον όρο στον τίτλο ή στην περιγραφή της. Στην περίπτωση που ο όρος βρεθεί, τότε η μηχανή αναζήτησης εμφανίζει στον χρήστη τις σχετικές εικόνες που εντοπίστηκαν. Αυτός ο τρόπος ανάκτησης χρησιμοποιείται μέχρι και σήμερα και θα τον βρούμε σε δεκάδες μηχανές αναζήτησης.



Εικόνα 9. Η δομή ενός συστήματος TBIR (Elmogy, 2015).

Το TBIR όμως παρουσιάζει κάποια σημαντικά μειονεκτήματα:

1. Τα σχετικά αποτελέσματα μειώνονται αν ο χρήστης κάνει ορθογραφικά λάθη ή γράψει με διαφορετικό τρόπο τον όρο αναζήτησης. Οι σχολιασμοί της εικόνας εξαρτώνται επίσης από την γλώσσα και την διάλεκτο που γίνεται η περιγραφή. Πιο χαρακτηριστικό παράδειγμα είναι η διαφορά των όρων από τα Αγγλικά του Ηνωμένου Βασιλείου με τα Αγγλικά της Αυστραλίας, των Ηνωμένων Πολιτειών κτλ.
2. Όλοι οι σχολιασμοί των εικόνων πρέπει να γίνουν από τον άνθρωπο, κάτι που είναι αρκετά χρονοβόρο, απαιτείται χρηματική αμοιβή για κάθε εγγραφή και σε μια τεράστια βάση δεδομένων αυτό θα ήταν μη πρακτικό.
3. Πολλές εικόνες χρειάζονται λεπτομερή περιγραφή για να είναι αποτελεσματική η ανάκτηση. Ένας πίνακας ζωγραφικής θα χρειαστεί μεγαλύτερη προσοχή στην

περιγραφή από μία εικόνα ενός προσώπου. Έτσι το άτομο που αποδίδει την περιγραφή μπορεί να μην είναι ικανό να περιγράψει απόλυτα μία εικόνα, με αποτέλεσμα να χαθεί σημαντικό κομμάτι της περιγραφής.

Content-Based Image Retrieval (CBIR)

Τα προβλήματα της ανάκτησης εικόνας που υπήρχαν με βάση το κείμενο και η σταδιακή αύξηση των χρηστών μετά την δημιουργία του διαδικτύου, έστρεψαν την έρευνα στην δημιουργία πιο γρήγορων και αποτελεσματικών μοντέλων. Για αυτό δημιουργήθηκε το Content-Based Image Retrieval. Είναι ένα μοντέλο για την ανάκτηση των κατάλληλων εικόνων μέσα από τα χαρακτηριστικά τους που μας επιτρέπουν την καλύτερη και με ακρίβεια ανάκτηση.

Το Content-Based Image Retrieval (γνωστό ως query by image content , content-based visual information retrieval ή reverse image search) είναι η σύγκριση και ανάκτηση εικόνας με βάση τα οπτικά στοιχεία της, δηλαδή δουλεύει μέσα από μια βάση δεδομένων που είναι διαθέσιμη στον χρήστη.

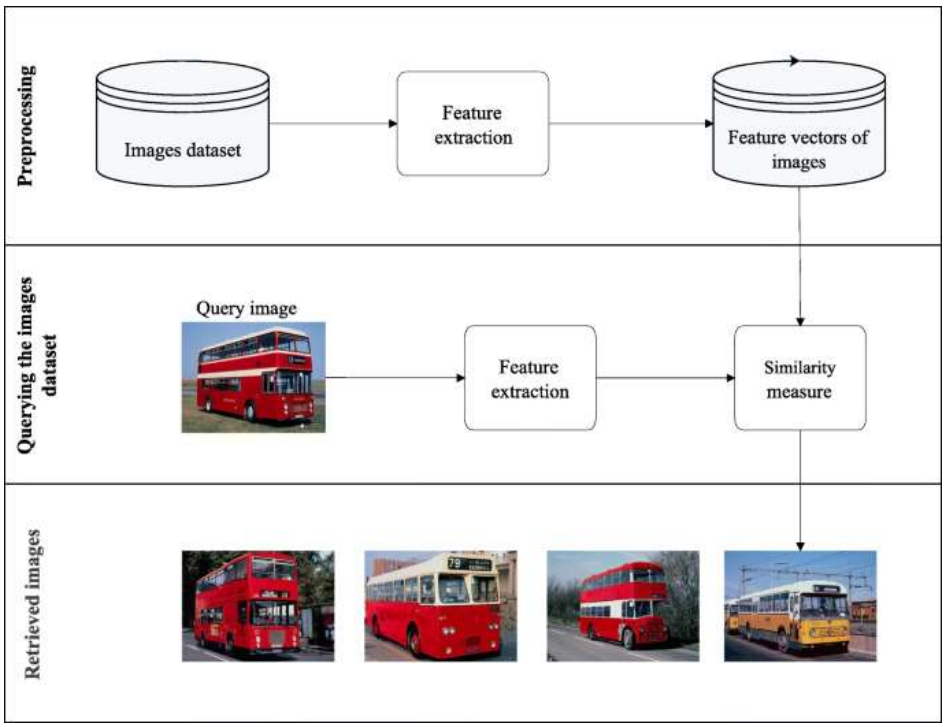
Πρώτα, ο χρήστης στέλνει το ερώτημα που ταιριάζει με τα οπτικά υποδείγματα του τύπου της εικόνας ή των χαρακτηριστικών της εικόνας προς αναζήτηση. Έπειτα, το σύστημα χρησιμοποιεί τις οπτικές ιδιότητες της εικόνας ως στοιχείο της αναζήτησης, (σε αντίθεση με το TBIR, που χρησιμοποιήσει μεταδομένα) και επιστρέφεται από την βάση η εικόνα που έχει την ίδια ή παρόμοια οπτική παρουσία (Elmogy, 2015).

Για να γίνει με επιτυχία η ανάκτηση των εικόνων, πρέπει τα χαρακτηριστικά αυτά να περιγραφούν με υψηλή ακρίβεια. Η μεγαλύτερη δυσκολία στην ανάπτυξη ενός συστήματος CBIR είναι η εύρεση των αλγορίθμων που ανακαλύπτουν τις ομοιότητες στις εικόνες και η δυνατότητα περιγραφής των χαρακτηριστικών μίας εικόνας. Πρέπει να τονιστεί ότι η απόδοση ενός συστήματος ανάκτησης εικόνων με βάση το περιεχόμενο περιορίζεται από τα χαρακτηριστικά που υιοθετούνται για την αναπαράσταση των εικόνων στη βάση δεδομένων.

Το ποσοστό της αποδοτικότητας της ανάκτησης στηρίζεται πλήρως στην επιλογή κατάλληλων χαρακτηριστικών εικόνας. Για αυτό είναι σημαντικό να επιλέγουμε τα κατάλληλα χαρακτηριστικά καθώς και τις πιο αποδοτικές τεχνικές για εξαγωγή χαρακτηριστικών.

Συνοπτικά το CBIR :

1. Εντοπίζει στην εικόνα το χρώμα, την υφή και τα σχήματα μέσα από ερμηνευτές που αναγνωρίζουν από τα pixels.
2. Υπολογίζουν τις αποστάσεις των χαρακτηριστικών της εικόνας.
3. Συγκρίνει την εικόνα του ερωτήματος με τις εικόνες που βρίσκονται στην βάση δεδομένων.
4. Επιστέφει το πλήθος των εικόνων που αντιστοιχούν περισσότερο με την αρχική εικόνα.



Εικόνα 10. Η δομή ενός συστήματος CBIR (Singh, 2019).

Σημασιολογική ανάκτηση

Για να μπορέσουν ανταπεξέλθουν στα ειδικά χαρακτηριστικά των οπτικών δεδομένων, εμφανίστηκαν οι μέθοδοι της ανάκτησης βάσεων περιεχομένου.

Σε μεγάλο βαθμό τα μοντέλα του CBIR αναζητούν για τα ειδικά χαρακτηριστικά που υπάρχουν στην εικόνα και τα αντιστοιχούν με άλλες εικόνες. Για την βελτίωση της ακρίβειας και της ανάκτησης των CBIR, οι ερευνητές επικεντρωθήκαν στην μείωση του σημασιολογικού κενού (Elmogy, 2015).

Η δομή ενός συστήματος που επικεντρώνεται στην σημασιολογική ανάκτηση, ξεκινά με τον χρήστη να στέλνει ένα ερώτημα στο σύστημα με την μορφή κειμένου ή εικόνας. Στην περίπτωση της εικόνας, το σύστημα εξάγει τα χαρακτηριστικά και γίνεται η μετατροπή τους σε σημασιολογικά χαρακτηριστικά από τον διερμηνέα του συστήματος. Στην περίπτωση του κειμένου, η μετάφραση γίνεται απευθείας. Ύστερα, το σύστημα αναζητά στην βάση δεδομένων για εικόνες που έχουν όμοια σημασιολογικά χαρακτηριστικά (έννοιες) και παρουσιάζονται στον χρήστη.

Η σημασιολογική ανάκτηση εστιάζει σε συγκεκριμένο είδος ερωτημάτων, όπως "βρες φωτογραφίες ενός κάστρου". Αυτού του είδους το ερώτημα είναι αρκετά περίπλοκο για τον υπολογιστή, καθώς ανάλογα με το είδους του κάστρου, θα διαφέρει το υλικό που έχει κατασκευαστεί άρα και στο χρώμα και στην υφή και η δομή του θα είναι διαφορετική ανάλογα με την εποχή και τον πολιτισμό του (Elmogy, 2015).

Συνεπώς, πολλά συστήματα CBIR χρησιμοποιούν χαρακτηριστικά χαμηλότερου επιπέδου, όπως η υφή, το χρώμα και το σχήμα. Αυτά τα χαρακτηριστικά χρησιμοποιούνται είτε σε συνδυασμό με διεπαφές που επιτρέπουν την ευκολότερη εισαγωγή των κριτηρίων είτε με βάσεις δεδομένων που έχουν ήδη εκπαιδευτεί για την αντιστοίχιση υψηλών χαρακτηριστικών (πρόσωπα, δακτυλικά αποτυπώματα κλπ.).

Αξιολόγηση συστημάτων

Η ανάκτηση εικόνων βρίσκεται σε ικανοποιητικό σημείο, όμως έχει ακόμα το περιθώριο για φτάσει στο 100% της απόδοσης. Ο υπολογισμός που επιφέρει σε κάθε δεδομένη φάση μπορεί να αποτελέσει ένα σημαντικό πυλώνα για να φτάσουμε στο μέγιστο επίπεδο.

Υπάρχουν πολλοί μέθοδοι για την αξιολόγηση της ανάκτησης. Η πιο συχνή είναι αυτή της ακρίβειας και ανάκλησης (precision and recall).

Η ακρίβεια (precision) (P) ορίζεται ως η αναλογία του αριθμού των σχετικών εικόνων που ανακτήθηκαν προς τον αριθμό των συνολικών εικόνων που ανακτήθηκαν.

Η ανάκληση ή (recall) (R) ορίζεται ως ο αριθμός των σχετικών εικόνων που ανακτήθηκαν σε σχέση με το συνολικό αριθμό των σχετικών εικόνων που είναι διαθέσιμες στη βάση δεδομένων.

Για την μέτρηση της ακρίβειας και της ανάκλησης, υπάρχει ο πίνακας αποτελεσμάτων (confusion matrix). Η κάθε τιμή που υπάρχει στον πίνακα είναι το αποτέλεσμα της προβλεπόμενης από το σύστημα απάντησης (predicted), και των πραγματικών τιμών (actual). Για να λειτουργήσει σωστά η αξιολόγηση, πρέπει οι πραγματικές τιμές να είναι γνωστές από τον άνθρωπο αλλά όχι από το σύστημα (Buckland & Gey, 1994).

Οι τιμές αυτές χωρίζονται σε 4 κατηγορίες:

True positive: Το σύστημα έχει προβλέψει σωστά ότι η πραγματική τιμή είναι θετική.

False positive: Το σύστημα λανθασμένα έχει προβλέψει ότι η πραγματική τιμή είναι θετική.

False negative: Το σύστημα λανθασμένα έχει προβλέψει ότι η πραγματική τιμή είναι αρνητική.

True negative: Το σύστημα έχει προβλέψει σωστά ότι η πραγματική τιμή είναι αρνητική.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Εικόνα 11. Confusion matrix.

Αφού γίνει η συλλογή αυτών των τιμών, μπορούμε να υπολογίσουμε την ακρίβεια και την ανάκληση. Οι τυπικοί ορισμοί αυτών των δύο μέτρων δίνονται από τις ακόλουθες εξισώσεις.

$$\text{ΑΝΑΚΛΗΣΗ} = \frac{\text{TRUE POSITIVES}}{\text{TRUE POSITIVES + FALSE NEGATIVES}} \quad \text{ΑΚΡΙΒΕΙΑ} = \frac{\text{TRUE POSITIVES}}{\text{TRUE POSITIVES + FALSE POSITIVES}}$$

Οι τιμές αυτές έχουν μέγιστη τιμή το 1 και έτσι μπορούμε να τις μετατρέψουμε σε ποσοστό. Όσο υψηλότερο είναι το ποσοστό, τόσο καλύτερη είναι η αξιολόγηση στην ανάκτηση του συστήματος.

Η υψηλή ακρίβεια σημαίνει ότι πολλές από τις εικόνες που επιστρέφονται από το ερώτημα είναι σχετικές, ενώ η **υψηλή ανάκληση** σημαίνει ότι το σύστημα κατάφερε να εμφανίσει τις περισσότερες σχετικές εικόνες που υπάρχουν διαθέσιμες (Buckland & Gey, 1994).

Η εφαρμογή της αξιολόγησης

Ας δούμε την αξιολόγηση στην πράξη.

Μία βάση αποτελείται από εικόνες με ζώα και εμείς θέλουμε να ανακτήσουμε όσες σχετίζονται και περιέχουν σκίουρους. Το σύστημα κάνει σάρωση όλες τις εικόνες. Αν το σύστημα προβλέψει ότι στην εικόνα υπάρχει ένας σκίουρος, τότε επιλέγεται η τιμή **true positive** για την συγκεκριμένη και **false negative** για οποιοδήποτε άλλη με διαφορετικό ζώο. **True negative** παίρνουν οι εικόνες οι οποίες το σύστημα επιτυχώς πρόβλεψε ότι δεν ταιριάζουν στο ερώτημα που θέσαμε και συνεπώς δεν είναι σκίουροι στις εικόνες αυτές. Αν το σύστημα κάνει λάθος και ανακτηθούν εικόνες που δεν ταιριάζουν στο ερώτημα, τότε αυτές παίρνουν την τιμή **false positive** και το σύστημα επομένως έχει ανακτήσει λάθος αποτελέσματα.

Δηλαδή:

Ας υποθέσουμε ότι σε μια βάση έχουμε εικόνες διάφορων άλλων ζωών, εκ των οποίων οι 40 από αυτές είναι με σκίουρους. Το σύστημα επιστρέφει τις 20, από τις οποίες οι 15 είναι σχετικές (true positive) και 5 έχουν λανθασμένα ανακτηθεί (false positive), χωρίς να καταφέρει να ανακτήσει τις υπόλοιπες 25 σχετικές εικόνες (false negative).

Τότε η ακρίβεια του είναι 15/(15+5)=15/20=0.75=75%.

Η ανάκληση του είναι 15/(15+25)=15/40=0.375=37.5%.

Η συγκεκριμένη μέθοδος αποτελεί αρκετά εξυπηρετική και βοηθά τα συστήματα ανάκτησης να αξιολογηθούν και να παρέχουν ένα μετρό, στο οποίο θα τα βοηθήσει στο να προσανατολιστούν ως προς το πόσο απέχουν για την ορθότερη και απολυτή ανάκτηση. Βέβαια, είναι μονάχα ένα μέρος του ευρύτερου πλαισίου που θα βοηθήσει στην βελτίωση των υπηρεσιών που παρέχουν τα σύστημα ανάκτησής.

Τα χαρακτηριστικά της εικόνας

Μία εικόνα αποτελείται από χαρακτηριστικά (features). Ένα χαρακτηριστικό ορίζεται ως "μία συγκεκριμένη οπτική ιδιότητα σε μία εικόνα. Τα χαρακτηριστικά αναφέρονται μερικές φορές ως περιγραφείς ή descriptors". Υπάρχουν δύο είδη descriptors ή χαρακτηριστικών, τα global και τα local (Elmogy, 2015).

- **global**

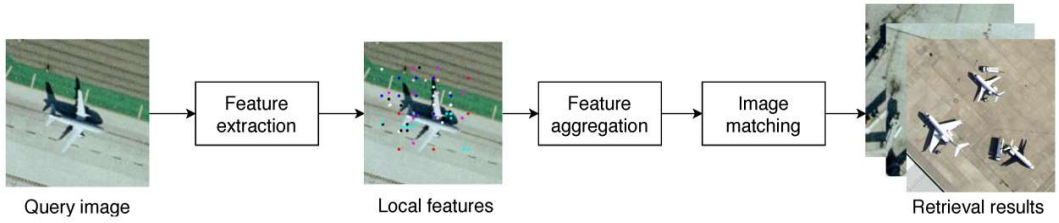
Περιγράφουν την εικόνα στο σύνολό της με σκοπό την γενίκευση του περιεχόμενου της. Παρέχουν πολύ γρήγορη ταχύτητα στην αναγνώριση των χαρακτηριστικών και των ομοιοτήτων μεταξύ άλλων. Χρησιμοποιούνται στην ανάκτηση εικόνων και γενικά για

εφαρμογές χαμηλού επιπέδου, όπως η ανίχνευση και η ταξινόμηση αντικειμένων (object detection and classification). Έχουν ένα σημαντικό μειονέκτημα, δεν μπορούν να περιγράψουν σωστά στην περίπτωση που η εικόνα είναι πολύπλοκη.

- **local**

Τα local περιγράφουν τμήματα μίας εικόνας ή αλλιώς τα βασικά σημεία της εικόνας , όπως μία έννοια που μπορεί να εντοπιστεί μέσα σε αύτη. Είναι πιο αποτελεσματικά στην ανάκτηση, αφού δεν σχηματίζουν μία περιγραφή για την εικόνα, αλλά πολλές, ανάλογα με τα σημαντικά σημεία που ξεχωρίζουν. Για αυτό τον λόγο χρησιμοποιούνται σε εφαρμογές υψηλότερου επιπέδου όπως η αναγνώριση αντικειμένων. Είναι πιο πολύπλοκα στον υπολογισμό και χρειάζονται περισσότερη υπολογιστική ισχύ.

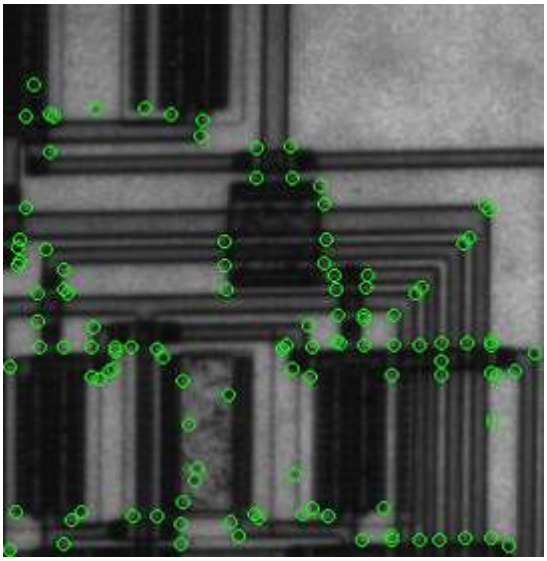
Η πιο αποτελεσματική προσέγγιση είναι ο συνδυασμός των δύο. Συμβάλουν έτσι στην βελτίωση της ακρίβειας αναγνώρισης, αλλά με την υπολογιστική επιβάρυνση του συστήματος.



Εικόνα 12. Ο ρόλος των local features (Imbriaco, Sebastian, Bondarev & de With, 2019).

Πέρα από τα local και global χαρακτηριστικά, υπάρχουν άλλες δύο κατηγορίες χαρακτηριστικών, εξίσου σημαντικά, τα low-level features και high-level features.

- **low-level features:** Χαρακτηριστικά χαμηλού επιπέδου είναι μικρές λεπτομέρειες της εικόνας, όπως γραμμές ή κουκκίδες.
- **high-level features:** Τα χαρακτηριστικά υψηλού επιπέδου χτίζονται πάνω στα χαρακτηριστικά χαμηλού επιπέδου για την ανίχνευση αντικειμένων και μεγαλύτερων σχημάτων στην εικόνα.



Εικόνα 13. Αποτέλεσμα συστήματος που εντοπίζει τις γωνίες.

https://www.mathworks.com/help/vision/ug/local-feature-detection-and-extraction.html

Εξαγωγή χαρακτηριστικών (feature extraction)

Για την επιτυχή διαδικασία ανάκτησης εικόνας, το πρώτο βήμα που χρειάζεται είναι η ανάλυση και εξαγωγή των κατάλληλων πληροφοριών.

Η ανάλυση εικόνας αφορά την ερευνά των δεδομένων εικόνας για μια συγκεκριμένη εφαρμογή. Συνήθως, τα ακατέργαστα δεδομένα των εικόνων αναλύονται με σκοπό να γίνουν κατανοητά και να χρησιμοποιηθούν για να εξάγουν τις επιθυμητές πληροφορίες.

Η εξαγωγή χαρακτηριστικών παίζει σημαντικό ρόλο στην επεξεργασία εικόνων και στην αναγνώριση μοτίβων (pattern recognition). Αποτελεί μία μορφή μείωσης των διαστάσεων της εικόνας, όταν η εισαγόμενη εικόνα είναι αρκετά μεγάλη για να επεξεργαστεί και να αφαιρεθεί η περιττή πληροφορία. Οι πληροφορίες αυτές μετατρέπονται σε διαιρούμενες σειρές χαρακτηριστικών. (Elmogy, 2015).

Σε γενικές γραμμές τα χαρακτηριστικά που εξάγονται είναι πρωτόγονα χαρακτηριστικά (primitive features) όπως χρώμα, υφή, σχήμα ή πληροφορίες σχετικές με το με το γενικό πλαίσιο και περιεχόμενο εικόνας. Το καθένα από αυτά κουβαλάει μεγάλο ποσοστό πληροφορίας που χρειάζεται για την εξαγωγή (feature extraction).

Χρώμα (Color)

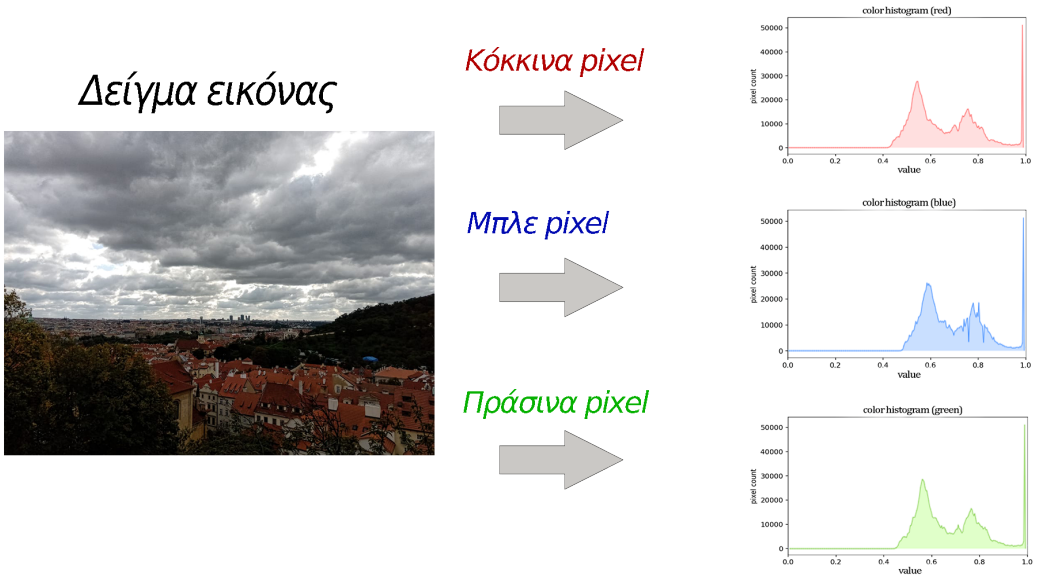
Ο συνηθέστερος τρόπος για την αντιστοίχιση εικόνων είναι το χρώμα. Αυτό οφείλεται στην ευκολία και ταχύτητα του υπολογισμού του. Το χρώμα αποτελεί ένα αρκετά εύληπτο χαρακτηριστικό και παίζει σημαντικό ρολό στην αντιστοιχία εικόνων. Για την περιγραφή του

χρώματος στην ηλεκτρονική εικόνα πρέπει να γίνει η μέτρηση της τιμής σε κάθε pixel. Το πιο διαδεδομένο πρότυπο για την περιγραφή των pixel, είναι το RGB. Για την εξαγωγή των δεδομένων σε μία εικόνα και την σύγκρισή της με άλλες εικόνες, υπάρχουν διαφορετικά μοντέλα, τα οποία αξιολογούνται με βάση την πολυπλοκότητα, την αξιοπιστία και την αποτελεσματικότητα (Meherban, 2016). Από τις περισσότερες τεχνικές που εντοπίζουν το χρώμα, οι παρακάτω είναι οι πιο διαδεδομένες και χρησιμοποιούνται στην πλειοψηφία των συστημάτων ανάκτησης εικόνας.

Χρωματικό Ιστόγραμμα (Color Histogram)

Το χρωματικό ιστόγραμμα, είναι η γραφική αναπαράσταση όλων των χρωμάτων που υπάρχουν σε μία εικόνα. Είναι ένα σύνολο από διαστήματα (intervals) που ονομάζονται bins. Κάθε bin δηλώνει την τιμή του pixel για ένα συγκεκριμένο χρώμα. Χρησιμεύει ως μια αναπαράσταση του χρωματικού περιεχομένου μιας εικόνας εάν το χρωματικό μοτίβο είναι μοναδικό σε σύγκριση με το υπόλοιπο σύνολο δεδομένων. Ο τρόπος που λειτουργεί είναι με την μέτρηση των τιμών των pixel σε ολόκληρη την εικόνα και έπειτα την αναπαράστασή τους στο ιστόγραμμα. Το ιστόγραμμα ξεκινάει την μέτρηση από την μικρότερη τιμή (0) στα αριστερά (μαύρο), μέχρι την μεγαλύτερη (1) στα δεξιά (άσπρο).

Χρησιμοποιείται σε μεγαλύτερο βαθμό από τα υπόλοιπα μοντέλα, γιατί είναι απλό στην κατανόηση και στην εφαρμογή, είναι ακριβές, γρήγορο και σταθερό στα αποτελέσματα του.



Εικόνα 14. Παράδειγμα από χρωματικό ιστόγραμμα.

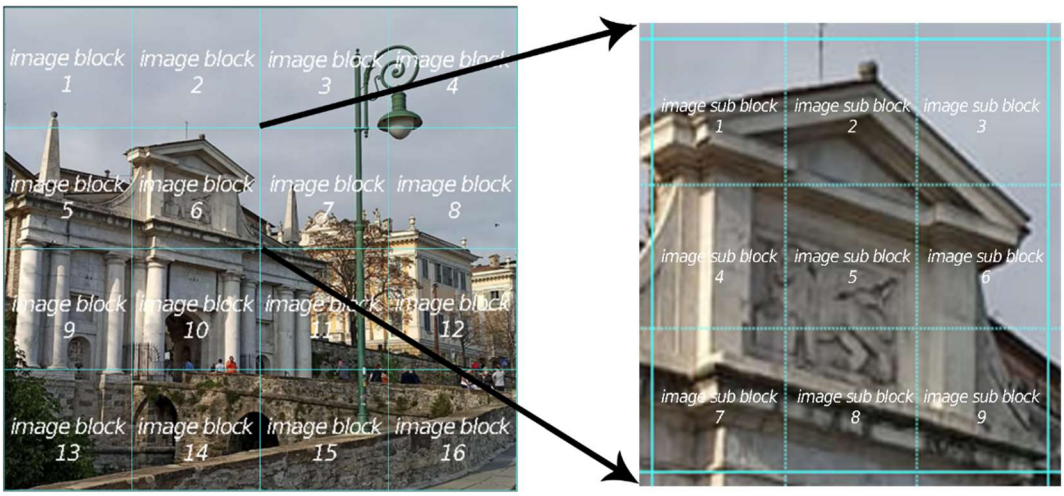
Color Moments

Είναι η μέτρηση των τιμών και η σύγκριση των εικόνων με μαθηματικούς τύπους. Αποτελείται από τρία μέρη (Mutlag, 2020):

- Μέσος όρος των τιμών
- Τυπική απόκλιση
- Ασυμμετρία

Ένα σημαντικό πλεονέκτημα τους είναι η χαμηλή πολυπλοκότητα, όμως αυτό έχει ως αποτέλεσμα η ακρίβεια των αποτελεσμάτων να μην είναι υψηλή.

Αν και οι παραπάνω τεχνικές είναι αρκετά ευκολονόητες για την μηχανή και τον άνθρωπο, δεν είναι αρκετές ακόμα για να πετύχουν την τέλεια ανάκτηση. Για αυτό προτείνεται η εικόνα να χωριστεί σε πλέγματα και να γίνει ο εντοπισμός και η αναγνώριση των χρωμάτων από κάθε πλέγμα ξεχωριστά. Πρόκειται για μια τεχνική που παρέχει αρκετή ακρίβεια ως προς τα αποτελέσματα, όμως υπάρχει μια ανησυχία σχετικά με την αξιοπιστία στον διαμοιρασμό της εικόνας σε πλέγματα (Zenggang, 2019).



Εικόνα 15. Παράδειγμα διαμοιρασμού εικόνας σε sub-blocks.

Υφή (Texture)

Η υφή αντιλαμβάνεται ως η περιγραφή της αίσθησης της αφής των αντικειμένων. Στον κόσμο των υπολογιστών όμως αποτελεί την εναλλαγή της φωτεινότητας των εικονοστοιχείων.

Είναι ουσιαστικά τα οπτικά μοτίβα που παρουσιάζουν μία ομοιομορφία εκτός του χρώματος και της έντασης. Ο σκοπός της ανάκτησης με την χρήση της υφής είναι η εύρεση ομοιοτήτων στις εικόνες με την πιθανότητα της αναγνώρισης αντικειμένων ή περιοχών μέσα από υποδεέστερα μοτίβα και χαρακτηριστικά όπως χρώμα, φωτεινότητα, μέγεθος, κλπ.

Περιέχει σημαντικές πληροφορίες σχετικά με τη βασική διάταξη της επιφάνειας (σύννεφα, φύλλα, τούβλα, ύφασμα).

Η αναγνώριση αυτή από τον άνθρωπο είναι εύκολη, όμως όχι για την ανάλυση δεδομένων.

Έχουν αναπτυχθεί διάφοροι μέθοδοι και νευρωνικά δίκτυα για την αναγνώριση της υφής των εικόνων και γενικότερα δεν μπορεί κάποιος να ισχυριστεί ότι υπάρχει μία γενική λύση σε αυτό το πρόβλημα. Κάθε νευρωνικό δίκτυο δημιουργείται ανάλογα με τις παραμέτρους που πρέπει να λάβει και το πρόβλημα το οποίο καλείται να λύσει. Η υφή θεωρείται ένα χαρακτηριστικό που μπορεί να εξαχθεί εύκολα από μία εικόνα χωρίς να απαιτεί τη δημιουργία μεγάλων και συνεπώς υπολογιστικά ασύμφορων νευρωνικών δικτύων.

Μέσα από την υφή έχουν προκύψει διάφορα μοντέλα αναγνώρισης.

Tamura

Είναι τεχνική που διακρίνει τα χαρακτηριστικά της υφής όπως η ανθρώπινη όραση.

Δίνει όμοια περιγραφή για όλους τους τύπους της υφής των εικόνων. Παρέχει έξι γνωρίσματα (Mutlag, 2020):

- Contrast (**αντίθεση χρώματος**)
- Coarseness (**Βάθος χρώματος**)
- Directionality (**Κατεύθυνση**)
- Line-Likeness (**γραμμή ομοιότητας**)
- Regularity (**Κανονικότητα**)
- Roughness (**Τραχύτητα**)

Είναι πολύ αποτελεσματικό στον εντοπισμό των χαρακτηριστικών της υφής και στην ανάκτηση όταν πρόκειται για υψηλού επιπέδου περιγραφή. Όμως λόγω της πολύπλοκης δομής του μοντέλου, σε χαμηλό επίπεδο περιγραφής και πιο απλών χρήσεων, το μοντέλο δεν παρουσιάζει την ίδια αποτελεσματικότητα.

Gray level co-occurrence matrix (GLCM)

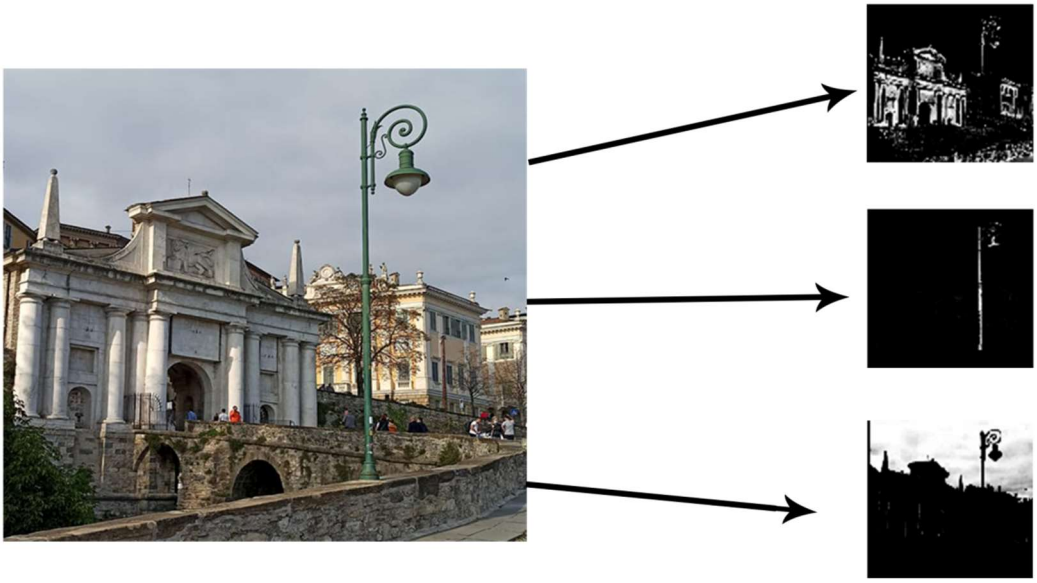
Αποτελεί μία από τις πρώτες μεθόδους εξόρυξης για την υφή. Χρησιμοποιεί ένα είδους ιστόγραμμα για τον υπολογισμό των γκρι τιμών που προκύπτουν σε μια συγκεκριμένη έκταση (offset) της εικόνας.

Οι συναρτήσεις του GLCM χαρακτηρίζουν την υφή την εικόνας, υπολογίζοντας πόσο συχνά εμφανίζονται σε μια εικόνα ζεύγη από pixels που έχουν συγκεκριμένες τιμές και βρίσκονται σε συγκεκριμένη έκταση τη εικόνας. (Zenggang, 2019).

Αποτελείται από 14 γνωρίσματα, με 5 από αυτά να είναι τα πιο διαδεδομένα. Ονομαστικά, τα γνωρίσματα αυτά είναι (Mutlag, 2020) :

- Entropy (**εντροπία**)
- Contrast (**αντίθεση**)
- Correlation **(συσχέτιση)**
- Energy (**ενέργεια και ομοιομορφία**)
- Homogeneit (**Ομοιογένεια**)

Λόγω της μικρής πιθανότητας λάθους, το μοντέλο αυτό έχει χρησιμοποιηθεί σε διάφορες εφαρμογές και τομείς, όπως στην ιατρική για τις ακτινογραφίες.



Εικόνα 16. Με την κλίμακα του γκρι, τα συστήματα μπορούν να εντοπίσουν σημεία που ξεχωρίζουν από τα άλλα.

Σχήμα (Shape)

Ορίζεται ως οι γεωμετρικές απεικονίσεις πάνω στην εικόνα ή η περιγραφή ενός αντικειμένου ανεξαρτήτως της θέσης, του προσανατολισμού και του μεγέθους του. Το σχήμα είναι η κύρια πηγή πληροφοριών που χρησιμοποιείται για αναγνώριση αντικειμένων (object recognition). Για να κριθεί χρήσιμο ένα σύστημα ανάκτησης, πρέπει να αναγνωρίσει τα αντικείμενα ακόμα και αν διαφέρουν στα χαρακτηριστικά τους.

Σημαντικό είναι να τονίσουμε πως πριν την περιγραφή των σχημάτων, απαραίτητος είναι ο διαχωρισμός της εικόνας σε μικρότερα μέρη (Chi, 2019). Είναι μία διαδικασία αρκετά απαιτητική και δύσκολη για να επιτευχθεί με απολυτή ακρίβεια. Για αυτό, η χρήση των χαρακτηριστικών του σχήματος έχει περιοριστεί σε ειδικές εφαρμογές που το αντικείμενο ή η περιοχή θα είναι άμεσα διαθέσιμες.

Αυτή την στιγμή, έχουν προκύψει 2 κατηγορίες εξαγωγής χαρακτηριστικών με βάση το σχήμα (Elmogy, 2015).

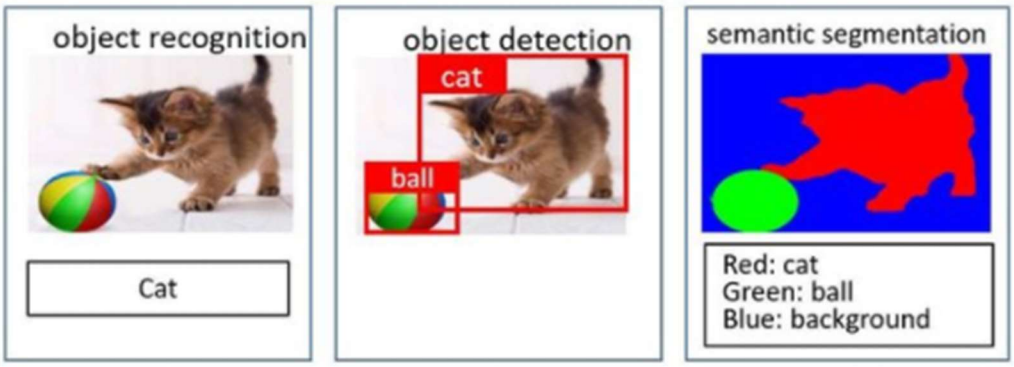
- **Με βάση τα όρια του σχήματος (boundary-based):** Χρησιμοποιεί μονάχα τα εξωτερικά όρια που έχουν τεθεί στο σχήμα.
- **Με βάση την περιοχή του σχήματος (region-based):** Χρησιμοποιεί ολόκληρη την περιοχή του σχήματος**.**

Αντιπροσωπευτική παρουσίαση αυτών των δυο κατηγοριών αποτελεί το Fourier Descriptor, το οποίο εφαρμόζεται για την περιγραφή του σχήματος οποιουδήποτε αντικειμένου και παρουσιάζει τα όρια σχήματος σε ένα τμήμα της εικόνας. Το κύριο πλεονέκτημα του είναι ότι παραμένει αμετάβλητο στις διάφορες διαδικασίες μετάφρασης (translation), περιστροφής (rotation) και κλιμάκωσης (scaling). Έτσι η περιγραφή του σχήματος γίνεται ανεξάρτητη από το μέγεθος του αντικειμένου και της σχετικής θέσης του αντικειμένου.

Spatial Location features

Είναι η αναγνώριση σημείων σε μία εικόνα με βάση την θέση τους. Για παράδειγμα, ένας άνθρωπος κάθεται πάνω σε ένα παγκάκι με φόντο ένα κτήριο. Είναι δύσκολα στην εκπαίδευση καθώς απαιτείται ο ορισμός των αντικειμένων στην εικόνα με βάση την σχετική τοποθεσία. Ο χρήστης του συστήματος δεν χρειάζεται να γνωρίζει τις ακριβείς συντεταγμένες των οντοτήτων. Θα πρέπει ο υπολογιστής να αναφέρει την θέση του κάθε αντικειμένου ως προς το άλλο και να εμφανίσει φράσεις που είναι κατανοητές για τον άνθρωπο ("δίπλα", "πίσω", "πάνω") Για να επιτευχτεί αυτό, το σύστημα χρησιμοποιεί εκπαιδευτικές εικόνες μαζί με προγραμματισμένα πρότυπα.

Οι σχέσεις μεταξύ των οντοτήτων ονομάζονται spatial relations. Οι σχέσεις αυτές χωρίζονται σε 3 κατηγορίες: βασικές, δεικτικές και συγγενικές. Για την εικόνα, οι δεικτικές σχέσεις είναι οι πιο χρήσιμες, αφού περιγράφουν την σχέση με βάση την οπτική γωνία του χρήστη. Για την σωστή αναγνώριση, ο υπολογιστής ακολουθεί πρότυπα τα οποία έχουν σχεδιαστεί να προβλέπουν την σχέση των αντικειμένων. Με την ανάπτυξη των νευρωνικών δικτύων, η αναγνώριση των αντικείμενων και η σχέση μεταξύ οντοτήτων γίνεται όλο και πιο αποτελεσματική (Haldekar, 2017).



Εικόνα 17. Διαχωρισμός αντικειμένων. https://www.researchgate.net/figure/The-differences-among-six-popular-computer-vision-tasks-1-Object-recognition-sometimes_fig1_335013237

Facial recognition

Το Facial Recognition ή αναγνώριση προσώπου είναι ένας τρόπος αποτύπωσης και ταυτοποίησης ενός ανθρώπινου προσώπου. Χρησιμοποιεί βιομετρικά στοιχεία για την χαρτογράφηση των χαρακτηριστικών ενός προσώπου μέσα από μια φωτογραφία, ένα βίντεο ή απευθείας από μια κάμερα ηλεκτρονικής συσκευής. Συγκρίνει τις πληροφορίες που εντοπίζει σε μια βάση δεδομένων προσώπων για να κάνει την σωστή αντιστοίχιση.

Η αναγνώριση προσώπου χρησιμοποιεί τις τεχνολογίες ανάκτησης εικόνας. Είναι ένα αναπόσπαστο κομμάτι της, συνδυάζει τις τεχνικές της και συλλέγει χαρακτηριστικά με παρόμοιο τρόπο. Όπως η ανάκτηση πληροφοριών, έτσι και η αναγνώριση προσώπου έχει γίνει κομμάτι της καθημερινότητας μας τα τελευταία χρόνια. Είναι μια λειτουργία που αναπτύσσεται ταυτόχρονα με τις τεχνικές της ανάκτησης εικόνας, χρησιμοποιώντας τα χαρακτηριστικά της εικόνας και το pattern recognition (Karnila, 2019).

Η εμφάνιση της θεωρίας ενός συστήματος αναγνώρισης προσώπου ξεκίνησε το 1964 από Αμερικάνους ερευνητές, υπό τον Woodrow Bledsoe. Με την ανάπτυξη της τεχνητής νοημοσύνης το 1988, η ανάπτυξη ενός τέτοιου συστήματος έγινε πραγματικότητα. Το 1991, το πρώτο παράδειγμα της τεχνολογίας της αναγνώρισης προσώπου είναι γεγονός, χάρη στον Alex Pentland και στον Matthew Turk. Από τότε, πολλοί οργανισμοί αναπτύσσουν συστήματα αναγνώρισης προσώπων και παίρνουν μέρος σε διαγωνισμούς αξιολόγησης τους. Η ανακάλυψη του deep learning και άλλων τεχνολογιών βελτίωσε σημαντικά την πρόοδο του facial recognition με την ακρίβεια και την αποτελεσματικότητα των συστημάτων συνεχώς να αυξάνεται (Adjabi, 2020).

Είναι ένα ισχυρό εργαλείο που μπορεί να βοηθήσει τους ανθρώπους ώστε να επαληθεύσουν την ταυτότητά τους ή να αποτρέψουν οποιαδήποτε απατή φέρει το πρόσωπο και το οποιοδήποτε ενδεχόμενο πλαστοπροσωπίας.

Η τεχνολογία αυτή είναι διαθέσιμη και στο διαδίκτυο. Μέσα κοινωνικής δικτύωσης όπως το Facebook και το Instagram, εντοπίζουν αυτόματα τα πρόσωπα που εμφανίζονται στις φωτογραφίες των χρηστών.

Η πιο συνήθης εφαρμογή του face recognition, είναι η χρήση του για το ξεκλείδωμα κινητού τηλεφώνου με την άμεση αναγνώριση του προσώπου. Το iPhone X χάρη στη τεχνολογία Face ID, μπορεί να εκμεταλλευτεί την κάμερα και να δημιουργήσει ένα αποτύπωμα προσώπου, σκανάρωντας 30.000 σημεία και αποθηκεύοντας τα στο κινητό.

Μια από τις πιο χρήσιμες ιδιότητες για τις οποίες έχει αναγνωριστεί είναι η ικανότητα του για την βοήθεια στην εύρεση αγνοουμένων και θυμάτων human trafficking. Βέβαια, υπάρχουν πολλοί οργανισμοί και φορείς (ιατρεία, αεροδρόμια, τράπεζες, αυτοκινητοβιομηχανίες) που στηρίζονται σε αυτή την τεχνολογία για την βελτίωση και προστασία των υπηρεσιών που προσφέρουν.

3.1.1 Face Recognition System

Τα συστήματα αναγνώρισης προσώπου ξεκίνησαν με την αναγνώριση σε μορφή 2D. Σήμερα, οι νέες τεχνολογίες επιτρέπουν τα συστήματα αυτά να αναγνωρίζουν τα πρόσωπα σε πραγματικό χρόνο σε μορφή 3D.

Το κάθε σύστημα αναγνώρισης προσώπου έχει διαφορετική δομή. Ανάλογα με τον τρόπο που συλλέγονται τα χαρακτηριστικά του προσώπου, αλλάζει και η δομή του. Όμως ως προς την λειτουργία ενός συστήματος, η δομή παραμένει σε γενικές γραμμές η ίδια. Το facial recognition λειτουργεί σε 4 στάδια.

1. **Εντοπισμός:** Η εικόνα με το πρόσωπο του χρήστη συλλέγεται άμεσα από φωτογραφίες ή βίντεο είτε απευθείας από την κάμερα κάποιας συσκευής .
2. **Ανάλυση:** Μετά, το σύστημα ελέγχει και συλλέγει τα πιο σημαντικά γεωμετρικά χαρακτηριστικά του προσώπου. Αυτό επιτυγχάνεται αναλύοντας τις αποστάσεις των σημείων αυτών πάνω στο πρόσωπο (απόσταση των ματιών, απόσταση των ματιών με το πιγούνι)
3. **Μετατροπή:** Η πληροφορία αυτή τότε μετατρέπεται σε δεδομένα και μέσω ενός αλγόριθμου δημιουργείται το αποτέλεσμα που είναι το μοναδικό αποτύπωμα του κάθε προσώπου και στην συνέχεια αποθηκεύεται στην βάση δεδομένων του συστήματος.

4. **Σύγκριση:** Τώρα το σύστημα έχει την δυνατότητα να συγκρίνει τα αποτυπώματα των προσώπων με άλλα και να φέρει ένα αποτέλεσμα.

Εικόνα 18. Αναγνώριση προσώπου στα πιο χαρακτηριστικά σημεία.

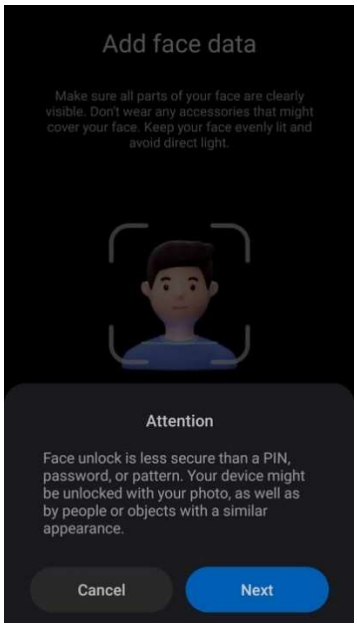
3.1.2 Εκπαίδευση

Τα συστήματα αναγνώρισης προσώπου, όπως τα συστήματα ανάκτησης εικόνας, έχουν την δυνατότητα να εκπαιδεύονται από τα dataset. Η πρώτη εμφάνιση ενός dataset έγινε το 1994, με το όνομα ORL Dataset. Αναπτύχθηκε από το Cambridge University Computer Laboratory και ήταν μια συλλογή από μόλις 400 εικόνες 40 διαφορετικών ατόμων με διαφορετικές εκφράσεις (Adjabi, 2020). Με τον καιρό, κυκλοφόρησαν νέα dataset που περιέχουν όλο και περισσότερες εικόνες και βοήθησαν στην ανάπτυξη νέων μεθόδων αναγνώρισης. Εκτός της εκπαίδευσης, πραγματοποιείται και η αξιολόγηση του κάθε συστήματος. Μερικά από αυτά είναι:

- **Flickr-Faces-HQ Dataset**: Είναι συλλογή από φωτογραφίες προσώπων του Flickr. Περιέχει 70.000 εικόνες διάφορες μεταβολές στην ηλικία και εθνικότητα του κάθε προσώπου. Δημιουργήθηκε το 2019.
- **Google Facial Expression Comparison Dataset**: Dataset της Google που βοηθάει στην αναγνώριση των προσώπων ακόμα και όταν πρόκειται για διαφορετικές εκφράσεις του προσώπου.

- **CASIA WebFace:** Χρησιμοποιείται για εργασίες επαλήθευσης και ταυτοποίησης προσώπων. Περιέχει 494.414 εικόνες 10.575 πραγματικών ταυτοτήτων που συλλέχθηκαν από το web.
- **Face Images With Marked Landmark Points**: Κυκλοφόρησε το 2018 και χρησιμοποιείται κυρίως για συστήματα ασφαλείας.
- **VGGFACE2:** Αποτελείται από περίπου 3,31 εκατομμύρια εικόνες χωρισμένες σε 9.131 κλάσεις, καθεμία από τις οποίες αντιπροσωπεύει μια διαφορετική ταυτότητα προσώπου.

Υπάρχουν αρκετοί παράγοντες που μπορούν να μειώσουν την ακρίβεια. Για παράδειγμα, η αλλαγή στη έκφραση και στην πόζα, η γήρανση, ο φωτισμός κλπ. Οι δυσκολίες αυτές μειώθηκαν με την ανάπτυξη 3D τεχνικών ανάκτησης (Adjabi, 2020).



Εικόνα 19. Face unlock κινητής συσκευής.

Η ανάκτηση εικόνας σε εφαρμογή

Παρακάτω θα αναλύσουμε μερικές από τις πιο γνωστές μηχανές αναζήτησης και ανάκτησης εικόνας που εντοπίσαμε.

Google Images

Το Google είναι η πιο διάσημη μηχανή αναζήτησης με τους περισσοτέρους ενεργούς χρήστες. Ιδρύθηκε το 1998 από Larry Page και Sergey Brin. Στις αρχές της δημιουργίας του η αναζήτηση γινόταν με κείμενο που εμφάνιζε αποτελέσματα μόνο για ιστοσελίδες. Το 2000, οι ιδρυτές ανταποκρίθηκαν στα αιτήματα των χρηστών τους για την δυνατότητα της ανάκτησης των εικόνων και ξεκίνησαν να δουλεύουν στην επέκταση των δυνατοτήτων της Google στον τομέα αυτό. Έτσι, τον Ιούλιο του 2001 πραγματοποιήθηκε η κυκλοφορία του Google Image Search, σήμερα γνωστό και ως Google Images.

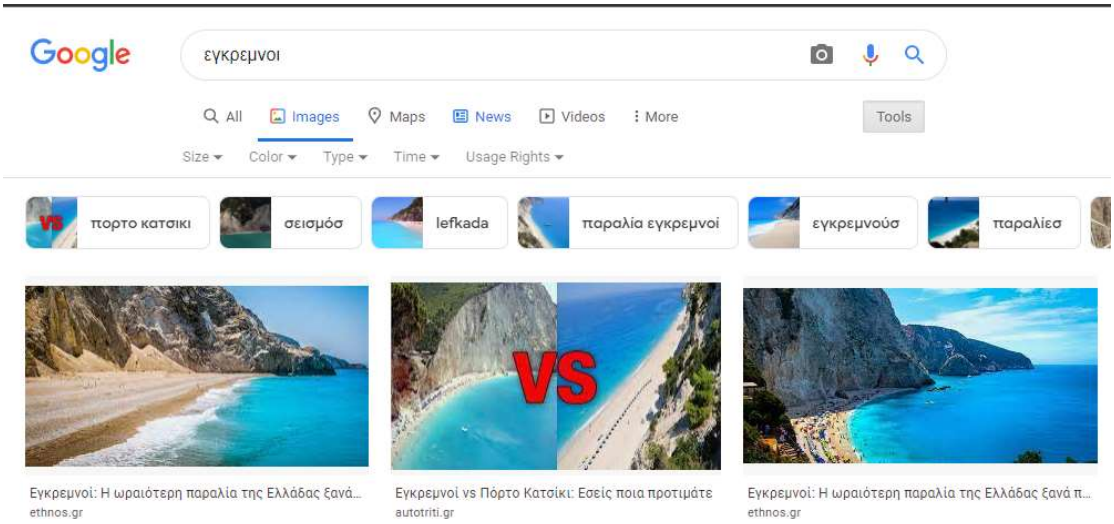
Εκτελεί τις ίδιες βασικές λειτουργίες αποστολής ερωτημάτων με την αρχική μηχανή αναζήτησης της Google. Ενώ η αναζήτηση Google φέρνει αποτέλεσμα με ιστοσελίδες σαρώνοντας το περιεχόμενο της ιστοσελίδας απευθείας, στο Google Images η διαδικασία φέρεται να είναι λίγο διαφορετική, καθώς επιστρέφει εικόνες με βάση το κείμενο ή τις λέξεις-κλειδιά που έχουν εισαχθεί (TBIR).

Αυτό όμως δεν ήταν από μόνο του αρκετό. Το Google Images ξεκίνησε την στήριξή του στις συναφείς πληροφορίες που αντλούσε από το κείμενο που βρισκόταν στην ίδια σελίδα αποτελεσμάτων με μία εικόνα.

Έτσι, ως τελικό συστατικό, ο αλγόριθμος αξιοποιεί machine learning, μέσα από το οποίο το Google Images μαθαίνει να συσχετίζει ορισμένες εικόνες μεταξύ τους για τη δημιουργία συμπλεγμάτων. Καταφέρνει επίσης να πετύχει την δυνατότητα του reverse image search, δηλαδή το Content-based Image Retrieval (CBIR).

Λειτουργία

Με τα διάφορα εργαλεία (Tools), υπάρχει η δυνατότητα να περιοριστούν τα αποτελέσματα επιλέγοντας να εμφανίζονται εικόνες συγκεκριμένου χρώματος, μεγέθους, τύπου αρχείου, ημερομηνίας δημιουργίας και πνευματικών δικαιωμάτων.



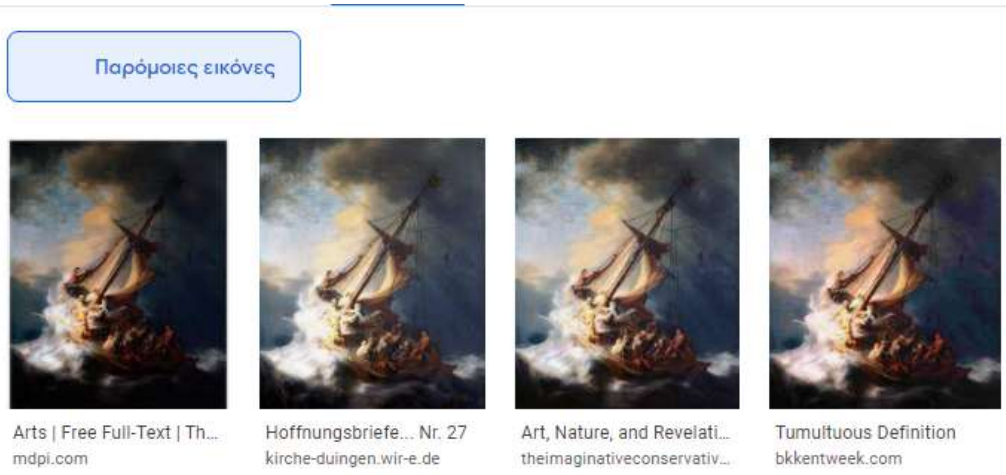
Εικόνα 20. Αναζήτηση με την χρήση κειμένου.

Google reverse image search

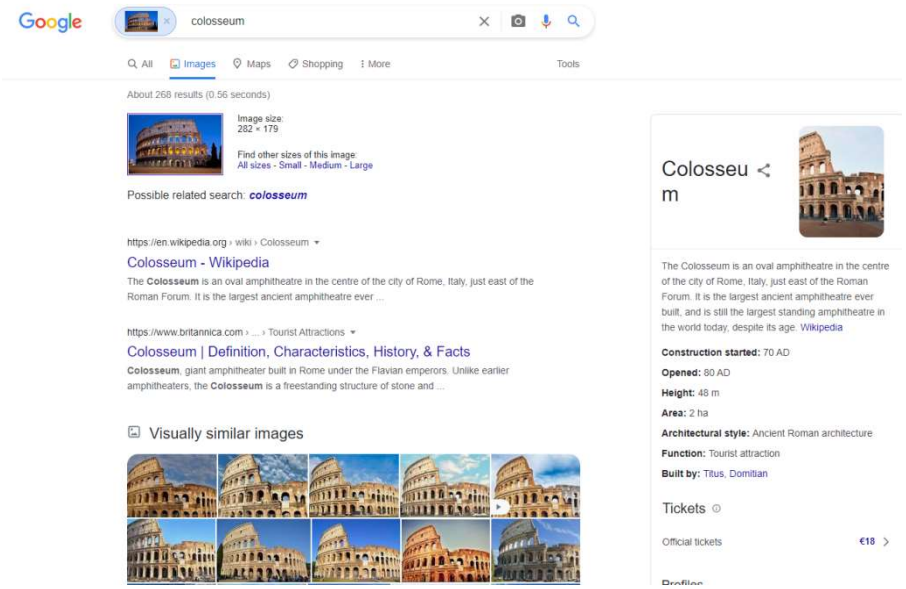
Με την μέθοδο αναζήτησης Google Search by Images, ο χρήστης έχει την δυνατότητα, να χρησιμοποιήσει εικόνες για την αναζήτηση αντί του γραπτού κειμένου. Αυτό γίνεται με την αποστολή της εικόνας ως ερώτημα στην μπάρα αναζήτησης ή ως link από κάποια ιστοσελίδα.

Όπως και στο CBIR έτσι και στο Google Search by Images, η εικόνα αναλύεται ως προς τα συγκεκριμένα χαρακτηριστικά, όπως το χρώμα, την ύφη και τα σχήματα που υπάρχουν σε αυτή, επιστρέφοντας μέχρι και δισεκατομμύρια αποτελέσματα οπτικά παρόμοιων εικόνων.

Εκτός από την εμφάνιση αυτών των εικόνων, ο χρήστης μπορεί να μάθει όλες τις ιστοσελίδες που εμφανίζουν το περιεχόμενο ή την περιγραφή της εικόνας και να μάθει την αρχική πηγή της, αν αυτή περιλαμβάνεται στην βάση της Google. Αυτή η λειτουργία είναι ειδικά χρήσιμη για του κατόχους των πνευματικών δικαιωμάτων, ώστε να παρακολουθούν την οποιαδήποτε αναπαραγωγή από τρίτους. Βέβαια αυτό δεν είναι απόλυτα έμπιστο ακόμα, γιατί φαίνεται πως για να βρει την ίδια εικόνα στο διαδίκτυο, πρέπει να περιέχει ολόκληρη την εικόνα και όχι μέρος της.



Εικόνα 21. Οι παρόμοιες εικόνες που βρέθηκαν είχαν την ίδια ανάλυση.



Εικόνα 22. Αναζήτηση με την χρήση εικόνας.

Περιορισμοί

Στην διάρκεια των πειραματισμών μας, παρατηρήσαμε κάποιους περιορισμούς στην αναζήτηση. Ανακαλύψαμε ότι δεν δέχεται εικόνες με πολύ μεγάλο αριθμό pixel και συγκεκριμένα οτιδήποτε μεγαλύτερο από 8000x6000 pixels. Ο περιορισμός αυτός δεν κοστίζει αρκετά στην ανάκτηση, καθώς ελάχιστες εικόνες στο διαδίκτυο ξεπερνούν αυτές τις διαστάσεις. Επίσης δοκιμάσαμε διάφορους τύπους αρχείων με την ίδια εικόνα και το μόνο είδος που δεν αναγνώριζε ήταν το TIFF. Τέλος, ανακαλύψαμε ότι με την ίδια εικόνα υπάρχει η μικρή πιθανότητα να εμφανίσει ένα εντελώς διαφορετικό αποτέλεσμα. Αυτό συνέβαινε σε περιπτώσεις που η εικόνα ήταν ένα σχέδιο ή ένας πίνακας.

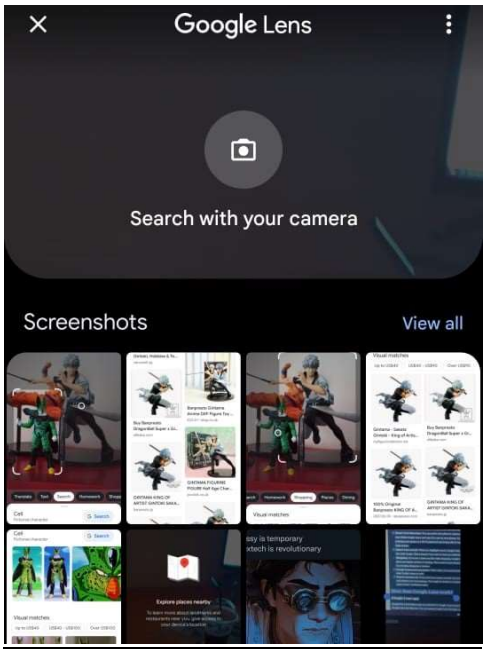
Η εικόνα είναι πολύ μεγάλη (πάνω από 8.000 επί 6.000 pixel) ή το Google δεν μπορεί να διαβάσει κωδικοποίησή της.

Google Lens

Το Google image search, αφορά κυρίως τους χρήστες ηλεκτρονικών υπολογιστών. Με την διάδοση των κινητών συσκευών, η Google ως βάση την μηχανή αναζήτησης της, ανέπτυξε μία εφαρμογή ικανή να δουλεύει στις κινητές συσκευές και να είναι πιο φιλική στην χρήση, το Google Lens.

Το Google Lens κυκλοφόρησε και έγινε διαθέσιμο τον Οκτώβριο του 2017, οπού ήταν προ εγκατεστημένο στην συσκευή Google Pixel 2. Έγινε διαθέσιμο ως αυτόνομο app για όλους στο Google Play στα τέλη του 2018.

Η εφαρμογή μπορεί να λειτουργήσει παράλληλα με το Google Images, Google Photos, το Google Assistant και την ενσωματωμένη εφαρμογή κάμερας του Android. Προαπαιτούμενο για την λειτουργία της εφαρμογής είναι η συσκευή που χρησιμοποιείται να είναι συνδεδεμένη σε κάποιο λογαριασμό Google.



Εικόνα 23. Google Lens App preview.

Λειτουργία

Σε γενικές γραμμές το Lens είναι μια οπτική μηχανή αναζήτησης που αναλύει συνεχώς τα δεδομένα που βρίσκονται μπροστά της ή σε μια εικόνα, για την επιτέλεση συγκεκριμένων εργασιών.

Σύμφωνα με την επίσημη σελίδα της εφαρμογής, το Google Lens εκμεταλλεύεται τα νευρωνικά δίκτυα και αναλύει την εικόνα για να προσδιορίσει τι περιέχει. Εντοπίζει τα αντικείμενα ή τις οντότητες που εμφανίζονται σε μία φωτογραφία ή απευθείας από την κάμερα της συσκευής και τα συγκρίνει με άλλες φωτογραφίες του διαδικτύου (μέσα από την Google), με βάση την ομοιότητα και το ποσό σχετική είναι η αρχική φωτογραφία ή το αντικείμενο που εμφανίζεται.

Για την εμφάνιση των σχετικών αποτελεσμάτων, το Lens χρησιμοποίει την γλώσσα, τις λέξεις κλειδιά και τα μεταδεδομένα από την ιστοσελίδα που προέρχεται η εικόνα. Κατά την διάρκεια ανάλυσης εικόνας, το Lens παράγει πολλά πιθανά αποτελέσματα και τα κατατάσσει ανάλογα με το ποσό σχετικό είναι το αποτέλεσμα, περιορίζοντας μερικές φορές τις πιθανότητες σε ένα μονάχα αποτέλεσμα.

Μπορεί να αναγνωρίσει και να αναζητήσει το περιεχόμενο μίας εικόνας, μέσα από την συλλογή εικόνων είτε απευθείας από την κάμερα της συσκευής που χρησιμοποιείται. Είναι σχεδιασμένο έτσι ώστε να εμφανίζει πληροφορίες σχετικές με το αντικείμενο που αναγνωρίζει. Χρησιμοποιώντας την αναγνώριση εικόνας (βασισμένη στο νευρωνικό δίκτυο (neural network)), την οπτική αναγνώριση και την τεχνητή νοημοσύνη (Artificial Intelligence), λειτουργεί καλύτερα και ταχύτερα από παλαιότερες εφαρμογές αναγνώρισης εικόνας**.**

Όταν κατευθύνετε την κάμερα του τηλεφώνου σε ένα αντικείμενο, το Google Lens θα προσπαθήσει να αναγνωρίσει το αντικείμενο, να διαβάσει barcodes, κωδικούς QR, ετικέτες και κείμενο και θα εμφανίσει σχετικά αποτελέσματα αναζήτησης, φωτογραφίες, ιστοσελίδες και πληροφορίες .

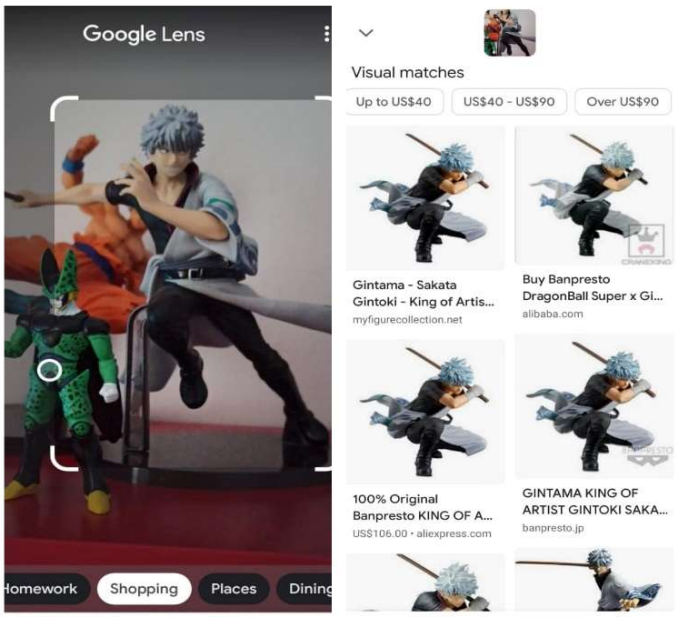
Ιδιότητες και χρήση

Η κάμερα του τηλεφώνου αναλύει συνεχώς αυτό που βλέπει για να δείξει τα αποτελέσματα σε πραγματικό χρόνο. Σύμφωνα με την google, η εφαρμογή προσφέρει:

- **Αναγνώριση αντικειμένου και εμφάνιση παρόμοιων αποτελεσμάτων για ηλεκτρονικές αγορές:**

  Όταν είναι σίγουρο ότι καταλαβαίνει ποιο προϊόν ενδιαφέρει τον χρήστη στη συγκεκριμένη φωτογραφία, το Lens θα επιστέψει αποτελέσματα αναζήτησης σχετικά με εκείνο το προϊόν. Για προϊόντα ή αντικείμενα που μπορεί να είναι ρούχα, ηλεκτρονικές συσκευές, βιβλία, κλπ., εμφανίζονται περεταίρω πληροφορίες και αποτελέσματα αγορών για το συγκεκριμένο προϊόν. Για την εμφάνιση των πιο

επιθυμητών αποτελεσμάτων, μπορεί να στηριχτεί και στις εν λόγο κρητικές από άλλους χρήστες του διαδικτύου.



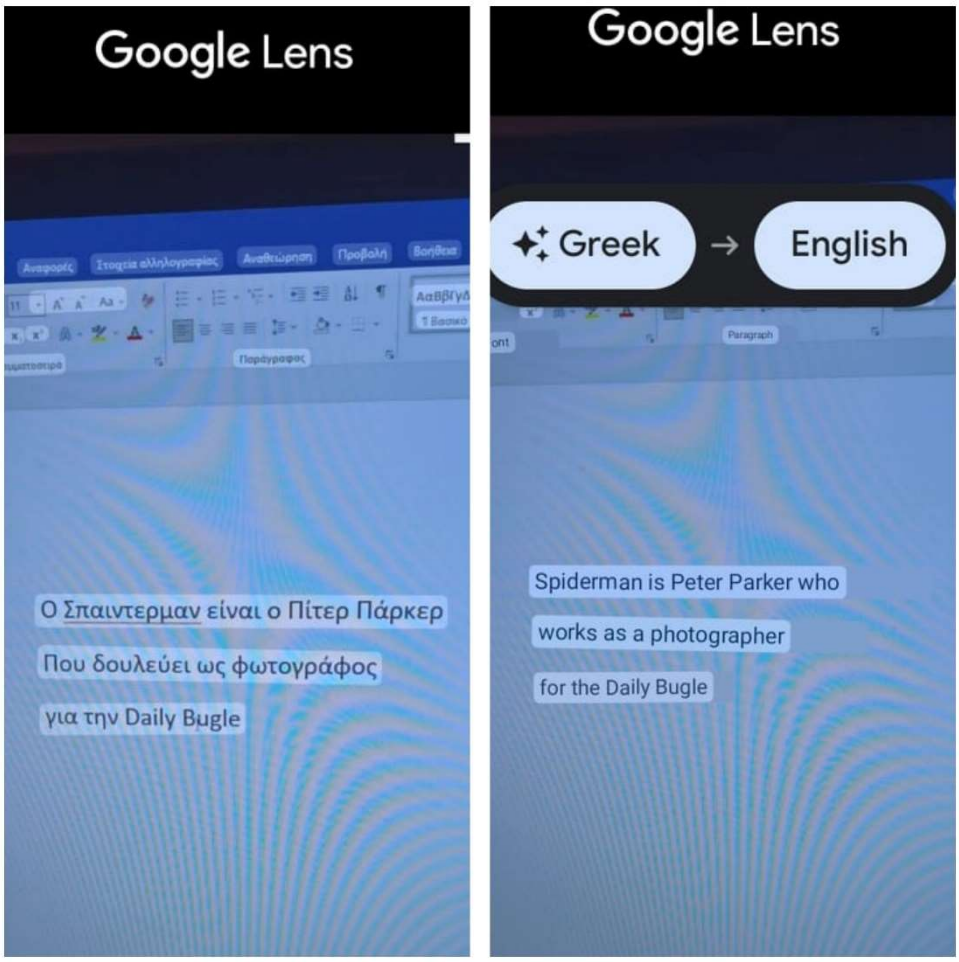
Εικόνα 24. Αποτελέσματα του Shopping.

Παρατηρήσεις

Δοκιμάζοντας τις διάφορες εργασίες του, παρατηρήσαμε ποσό ευέλικτη και γρήγορη είναι η εφαρμογή στον εντοπισμό του εν λόγο προϊόντος, προσφέροντας άμεσα μια γκάμα από αποτελέσματα που αρμόζουν σε αυτό το προϊόν (δίνοντας και ενδεικτικά την τιμή που μπορεί κάνεις να το βρει). Συμπεράναμε επίσης ότι η ανάκτηση του προϊόντος ήταν ακριβής ακόμα και όταν αφορά το χρώμα και την όψη του προϊόντος, κάνοντας μονάχα την παρατήρηση ότι στάθηκε μόνον σε αυτά και όχι σε πιο βαθιά χαρακτηριστικά που ίσως επέφεραν την ακριβής λήψη του προϊόντος και όχι παρεμφερών.

- **Οπτική αναγνώριση κειμένου και μετάφραση κειμένου:**

Παρέχει αναγνώριση και σάρωση κειμένου δίνοντας την δυνατότητα πρόσθετων εργαλείων, όπως αντιγραφής, επικόλλησης ακόμη και μετάφρασης σε οποιαδήποτε γλώσσα.

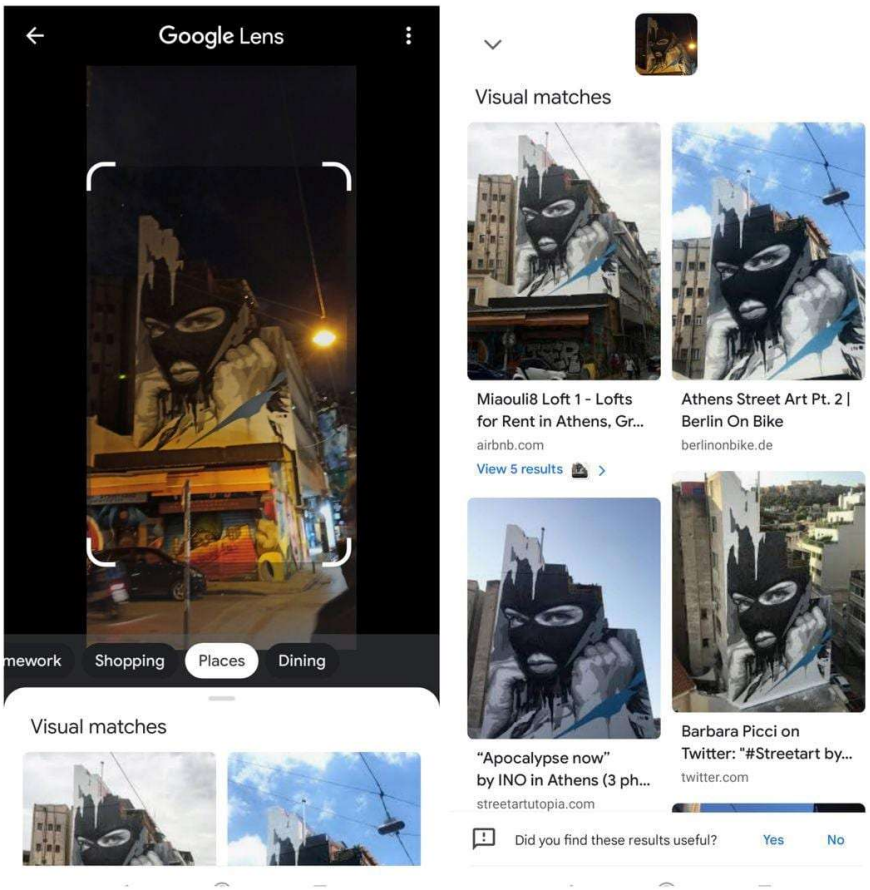


Εικόνα 25. Μετάφραση σε ζωντανό χρόνο.

Περαιτέρω, παρατηρήσαμε ότι και η αναγνώριση κειμένου αλλά και η μετάφραση λειτούργησε με αρκετή ευκολία, άμεσα και αρκετά αποτελεσματικά (τουλάχιστον όσον αφορά το συγκεκριμένο κείμενο).

- **Τοποθεσία**

Κατευθύνοντας την κάμερα της συσκευής στον περίχωρο σου, η εφαρμογή χρησιμοποιεί την τοποθεσία για να φέρει ακριβής πληροφορίες για τυχόν αξιοθέατα, τοπία και συνταγές φαγητών που αντικρίζει.



Εικόνα 26. Αντιστοίχηση του περιεχομένου της εικόνας με την τοποθεσία.

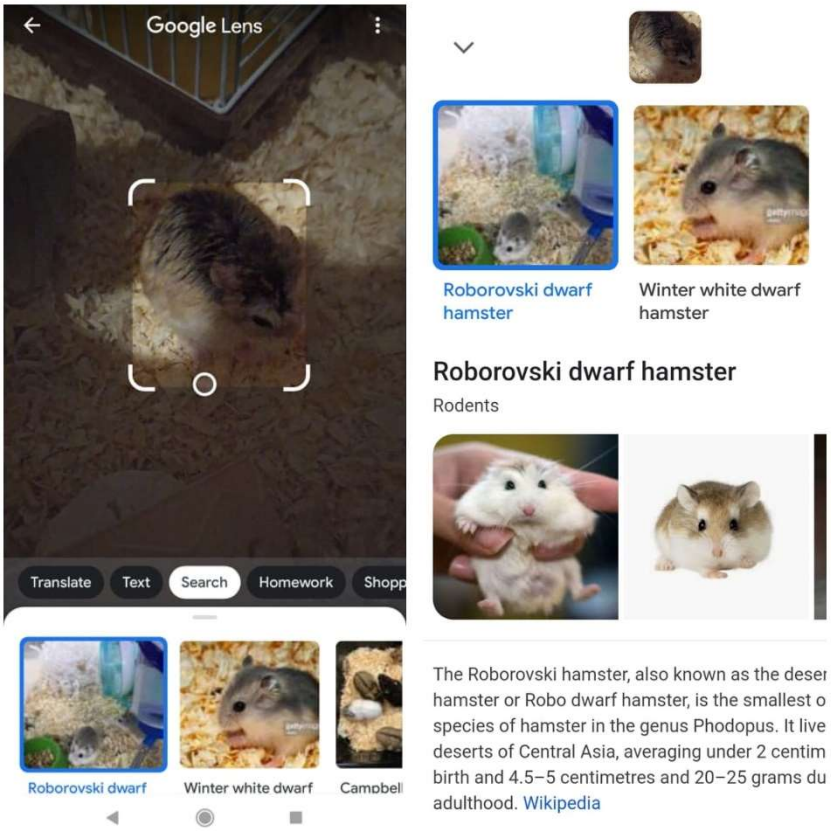
Παρατηρήσεις

Ήταν αρκετά εύκολο να εντοπίσει τοποθεσίες οι οποίες είναι ιδιαίτερες. Όμως σε αρκετές δοκιμές παρατηρήσαμε 2 βασικά θέματα σχετικά με την αναγνώριση τοποθεσίας:

1. Όταν ένας συγκεκριμένος χρηστής της εφαρμογής βρίσκεται σε μια τοποθεσία που είναι παραλία, βουνό, πεδιάδα, δάσος, κλπ. και γενικά σε μια τοποθεσία που δεν έχει κάτι χαρακτηριστικό ώστε να ξεχωρίζει ως περιοχή, ο αλγόριθμος του Lens δυσκολεύεται αρκετά να ανταποκριθεί και απλά παρουσιάζει και προτείνει παρόμοιες οπτικά τοποθεσίες.
2. Όταν ένα μέρος παρουσιάζει τις ίδιες ιδιότητες ή έχει το ίδιο οπτικό μοτίβο, ο αλγόριθμος και σε αυτήν την περίπτωση δυσκολεύεται να βγάλει ακριβής αποτέλεσμα για μια συγκεκριμένη τοποθεσία. (Κάτι τέτοιο μπορεί να προκύψει με ορισμένες αλυσίδες καταστημάτων που έχουν την ιδία λογική δομή (McDonalds, Goodys, Cinema κλπ.)).

- **Αναγνώριση του είδους μιας οντότητας.**

Προσδιορίζει το είδος του φυτού ή του ζώου που εμφανίζεται σε μια εικόνα. (π.χ. να προσδιορίσει ακριβώς την ράτσα ενός σκύλου χωρίς μεγάλη δυσκολία).



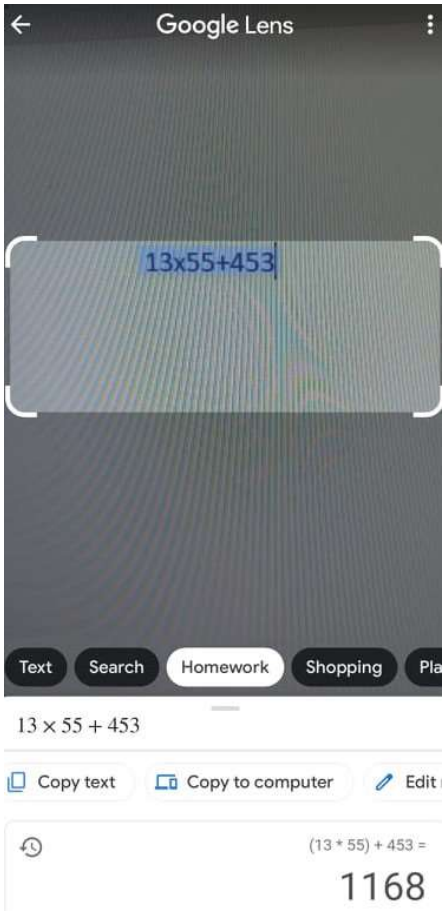
Εικόνα 27. Αναγνώριση του είδους ζώου.

Παρατηρήσεις

Κατάφερε όχι μόνο να κατανοήσει ότι πρόκειται για ένα χάμστερ αλλά και να εντοπίσει με ευκολία το είδος που παρουσιάζεται στην φωτογραφία, παρέχοντας περαιτέρω πληροφορίες για το συγκεκριμένο είδος.

- **Σχολική βοήθεια**

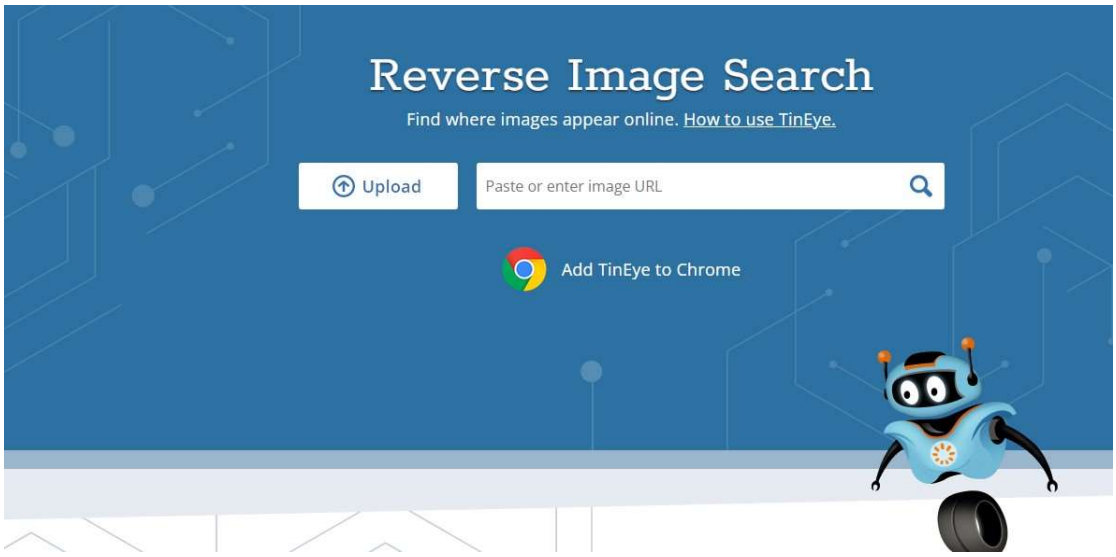
Αναγνώριση αριθμητικών εξισώσεων ή προβλημάτων και εμφάνιση των αποτελεσμάτων και πιθανών λύσεων.



Εικόνα 28. Υπολογισμός πράξεων.

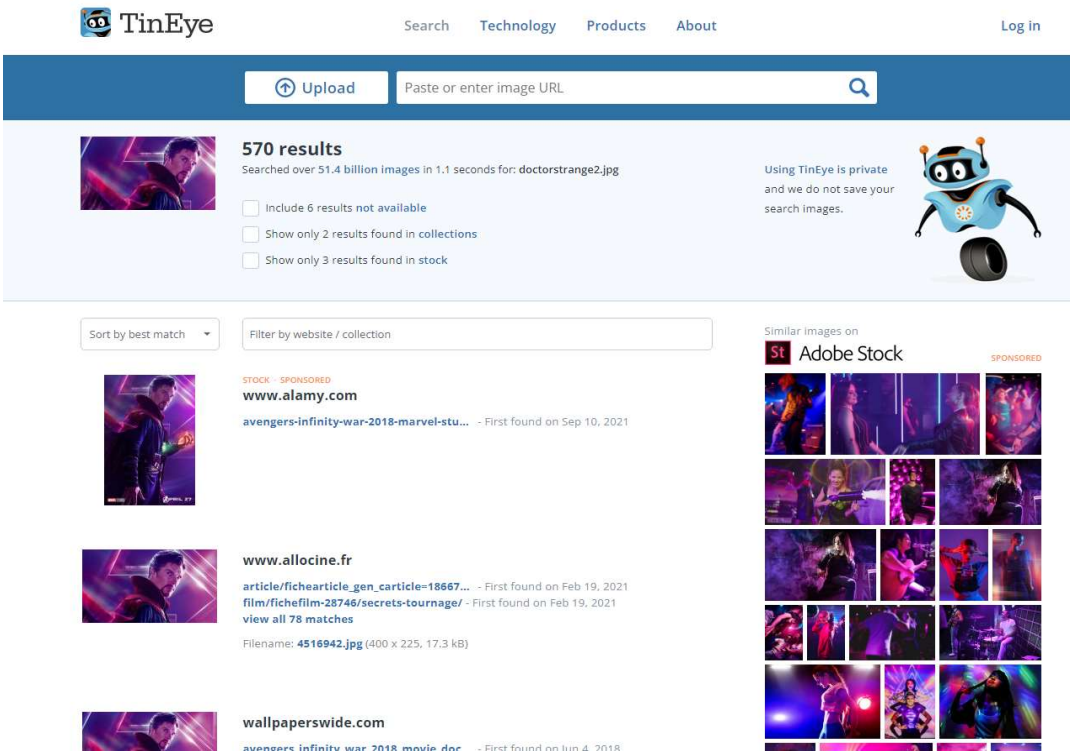
TinEye

Το TinEye είναι μια δωρεάν μη εμπορικής χρήσης μηχανή αναζήτησης που χρησιμοποιεί το reverse image search για την ανάκτηση αποτελεσμάτων. Προσφέρει μια σειρά από υπηρεσίες computer vision, image recognition και λύσεις που βοηθούν στο να γίνουν οι εικόνες που καταχωρούνται ανακτήσιμες. Κυκλοφόρησε τον Μάιο του 2008 από την Idée, Inc. του Καναδά. Είναι ελεύθερο στην χρήση και χρησιμοποιείται από επιχειρήσεις αλλά και απλούς χρήστες. Έχει την δυνατότητα να συλλέγει εικόνες από ιστοσελίδες με την μέθοδο web crawling και να τις καταχωρεί στην βάση δεδομένων. Η συλλογή αυτή την στιγμή είναι στις 51,4 δισεκατομμύρια εικόνες, με τον αριθμό αυτό να αυξάνεται συνεχώς.



Εικόνα 29. TinEye.

Η χρήση του είναι αρκετά απλή καθώς δεν διαφέρει από τον τρόπο λειτουργίας των υπόλοιπων μηχανών αναζήτησης. Εδώ δεν χρησιμοποιείται καθόλου η μπάρα αναζήτησης για την ανάκτηση με την χρήση του κειμένου, αλλά μονό με την χρήση της εικόνας ή του URL που παραπέμπει σε κάποια εικόνα.



Εικόνα 30. Το περιβάλλον του TinEye.

Η βασική μηχανή αναζήτησης του TinEye ψάχνει άλλες καταχωρήσεις της εικόνας στο διαδίκτυο, συμπεριλαμβανομένων τροποποιημένων εικόνων που βασίζονται σε αυτή την εικόνα, δηλαδή μικρότερες, μεγαλύτερες και περικομμένες (cropped) εκδοχές της, αναφέροντας την ημερομηνία και την ώρα κατά την οποία αναρτήθηκαν.

Σύμφωνα με τις πληροφορίες που αντλήσαμε από την επίσημη ιστοσελίδα του TinEye, ο τρόπος με τον οποίο δουλεύει είναι ότι από την στιγμή που ο χρήστης στέλνει μία εικόνα ως ερώτημα, το TinEye δημιουργεί ένα ηλεκτρονικό αποτύπωμα και υστέρα το συγκρίνει με άλλες εικόνες που είναι καταγεγραμμένες στην βάση του αναζητώντας τις ομοιότητες. Σε γενικές γραμμές το TinEye:

- Δεν είναι ικανό να βρει παρόμοιες θεματικά εικόνες καθώς η χρήση του δεν είναι για αυτόν τον σκοπό αλλά προσπαθεί να εντοπίσει την ιδία ακριβώς εικόνα ακόμα και εάν έχει επεξεργαστεί, έχει υποστεί περικοπή ή έχει αλλάξει μέγεθος.
- Μπορεί παρακολουθήσει την πλήρη δραστηριότητα μιας εικόνας στο διαδίκτυο και να την παρουσιάσει στον χρήστη. Τα ερωτήματα που απαντάει είναι "πόσες φορές έχει αναρτηθεί η εικόνα", "που έχει αναρτηθεί" και "πότε αναρτήθηκε".
- Να βοηθήσει τον χρήστη να εντοπίσει εικόνες υψηλής και καλής ανάλυσης ανεξάρτητα της ποιότητας του ερωτήματος.

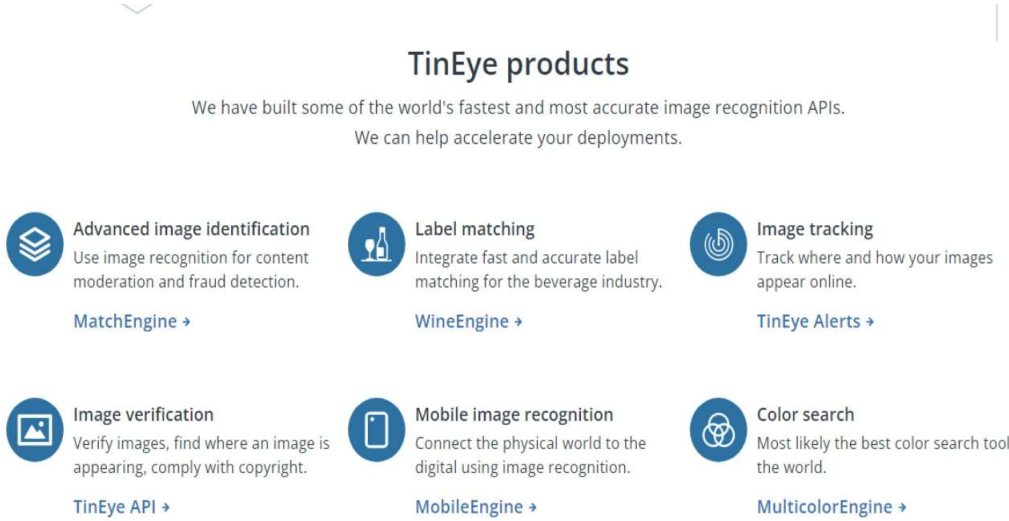
- Να εντοπίσει ιστοσελίδες που χρησιμοποιούν την εικόνα που έχει δημιουργήσει ένας χρήστης και να αναγνωρίσει εάν όντως πρόκειται για την ιδία εικόνα.
- Να αναγνωρίσει με ακρίβεια τα κυρίαρχα χρώματα στην εικόνα σας και το ποσοστό που καθένα από αυτά καταλαμβάνει.

Εικόνα 31. Αναγνώριση χρώματος.

Η μηχανή αναζήτησης TinEye είναι σε μεγάλο βαθμό δωρεάν, προσφέροντας όλες τις παραπάνω χρήσεις χωρίς να χρειαστεί κάποια είδους πληρωμή. Υπάρχει όμως και η επί πληρωμή έκδοση για εμπορική χρήση, η οποία διευρύνει την δυνατότητα αναζήτησης του TinEye.

Η εμπορική έκδοση του TinEye περιλαμβάνει μια διεπαφή χρήστη (user interface) για εύκολη αναζήτηση, καθώς και μια Διεπαφή Προγραμματισμού Εφαρμογών (Application Programming Interface ή API) ή για συντομία διεπαφή ή διασύνδεση για την ενσωμάτωση του TinEye σε έναν ιστότοπο ή σύστημα του χρήστη.

Προϊόντα κα υπηρεσίες που προσφέρει



Εικόνα 32. Προϊόντα του TinEye.

Οι αναφερόμενες υπηρεσίες σχετίζονται άμεσα με την ανάκτηση της εικόνας και η κάθε μία προσφέρει χρήσιμα εργαλεία. Οι υπηρεσίες και τα προϊόντα αυτά προορίζονται σε μεγάλο βαθμό για επιχειρήσεις οι οποίες θέλουν να ενσωματώσουν την τεχνολογία της ανάκτησης εικόνας, στην ιστοσελίδα ή στην εφαρμογή στο κινητό τους ή να έχουν την δυνατότητα να παρακολουθούν την οποιαδήποτε ανάρτηση των εικόνων τους στο διαδίκτυο.

Περιορισμοί

Βέβαια υπάρχουν κάποιοι περιορισμοί:

- Το αρχείο της εικόνας που χρησιμοποιείται ως ερώτημα δεν πρέπει να ξεπερνάει τα 20 MB και μπορεί να δεχτεί εικόνες μόνο πάνω από 100 pixels ανά διάσταση, γιατί το ηλεκτρονικό αποτύπωμα δεν θα μπορέσει να παρέχει αρκετές πληροφορίες.
- Το TinEye δεν μπορεί να αναγνωρίσει το περιεχόμενο της εικόνας ούτε τα περιγράμματα των αντικειμένων. Ως αποτέλεσμα, δεν εκτελεί facial recognition. Αναγνωρίζει ολόκληρη την εικόνα και ορισμένες τροποποιημένες εκδόσεις αυτής της εικόνας. Αυτό σημαίνει ότι το TinEye δεν μπορεί να βρει διαφορετικές εικόνες με τους ίδιους ανθρώπους ή τα ίδια πράγματα ούτε εικόνες με κοινό θέμα.
- Μπορεί να αναζητήσει σχεδόν όλους τους δυνατούς τύπους εικόνας, όπως JPEG, PNG, GIF, BMP, TIFF και WebP. Όμως, κάποιοι τύποι αρχείων που περιέχουν εικόνες στο διαδίκτυο, όπως το Adobe Flash, δεν μπορούν να αναζητηθούν
- Εικόνες που περιέχουν εμφανή watermarks θα πρέπει να αποφεύγονται διότι το TinEye μπορεί να αναζητήσει το watermark αντί για την ίδια την εικόνα.
- Με βάση την επίσημη ιστοσελίδα του TinEye δεν έχει την δυνατότητα να ανακτήσει εικόνες από συγκεκριμένες σελίδες. Πιο συγκεκριμένα, εικόνες από μέσα κοινωνικής δικτύωσης (Facebook, Instagram, Twitter) δεν γίνεται να ανακτηθούν λόγο του περιορισμού στο web crawling σε αυτές τις ιστοσελίδες. Όπως και εμφανώς εικόνες σε προστατευμένες ιστοσελίδες δεν γίνεται να ανακτηθούν αφού δεν είναι δημόσιες.

Εικόνα 33. Watermark Παράδειγμα.

Bing

Το Bing είναι η τρίτη μεγαλύτερη μηχανή αναζήτησης παγκοσμίως. Η πρώτη του εμφάνιση έγινε από τον CEO της Microsoft, Steve Ballmer το 2009.

Οι πρωτοποριακές για την εποχή του δυνατότητες περιλαμβάνουν την σημασιολογική τεχνολογία της Powerset, η οποία προσπαθεί να χρησιμοποιήσει επεξεργασία φυσικής γλώσσας για να κατανοήσει τη φύση της ερώτησης και να επιστρέψει σελίδες που περιέχουν την απάντηση δηλαδή να βρίσκει στοχευόμενες απαντήσεις στις ερωτήσεις των χρηστών. Το 2011 προσπάθησε να επικεντρωθεί στην παροχή ταχύτερων και πιο σχετικών αποτελεσμάτων αναζήτησης των χρηστών της.

Μετά το πρώτο βήμα που έκανε η Google, έτσι και το Bing επικεντρώθηκε έμπρακτα στην αναζήτηση εικόνων. Παρέχει επεκτάσεις (extension) αναζήτησης με την εμφάνιση αποτελεσμάτων εικόνας. Αυτό στην συνέχεια επεκτάθηκε και στην αναζήτηση στο διαδίκτυο, χρησιμοποιώντας εικόνα αντί για κείμενο, μέσα από ένα είδος ανάκτησης εικόνας με βάση το περιεχόμενο (CBIR), ονομάζοντας το Bing Visual Search. Η χρήση του είναι απλοϊκή και δεν διαφέρει με τις άλλες μηχανές αναζήτησης καθώς το μόνο που πρέπει να κάνει ο χρηστής είναι να πατήσει το εικονίδιο της κάμερας και να αναρτήσει την εικόνα που επιθυμεί να αναζητήσει.

Χρησιμοποιώντας το ο χρήστης μπορεί να βρει παρόμοιες εικόνες, προϊόντα, σελίδες που περιλαμβάνουν μια εικόνα, ακόμη και συνταγές.

Κεφάλαιο 4. Ανάκτηση ήχου

Υπάρχουν διάφορα πεδία έρευνας στην ανάκτηση ήχου αλλά δύο καθορίζονται ως οι βασικές κατηγορίες της, η **ανάκτηση της μουσικής** και η **αναγνώριση ομιλίας**.

4.1 Ανάκτηση μουσικής (Music Information Retrieval)

Η μουσική είναι ένα σημαντικό κομμάτι στην καθημερινή ζωή του ανθρώπου. Το τραγούδι ξεχωρίζει από τον ρυθμό, τους στοίχους, ένα όργανο, την εικόνα από ένα άλμπουμ ή το μέλος ενός συγκροτήματος. Αν και είναι εύκολο ένας άνθρωπος σε λίγα δευτερόλεπτα να αναγνωρίσει αυτά τα χαρακτηριστικά και να ονομάσει το τραγούδι, μία μηχανή θα χρειαστεί κάτι παραπάνω.

Η ανάκτηση ενός τραγουδιού περιλαμβάνει την συλλογή των μεταδεδομένων (όνομα, δημιουργό, έτος κυκλοφορίας κλπ.). Όμως υπάρχει και η δυνατότητα ανάκτησης με την χρήση του ήχου μέσα από την συλλογή των ηχητικών κυμάτων (soundaves).

Ο όρος "Music Information Retrieval" χρησιμοποιήθηκε για πρώτη φορά από τον Michael Kassler στο βιβλίο του το 1966 (Kassler, άνοιξη-καλοκαίρι 1966). Περιέγραψε το MIR ως γλώσσα προγραμματισμού που χρησιμοποιείται για την εξαγωγή συγκεκριμένων πληροφοριών από τα μουσικά δεδομένα.

Το Music Information Retrieval (MIR), σύμφωνα με τον Downie, είναι επιστημονικό πεδίο που αφορά την εξαγωγή χαρακτηριστικών από την μουσική, την ευρετηρίαση της μουσικής με βάση τα χαρακτηριστικά αυτά και την ανάπτυξη διαφορετικών μεθόδων αναζήτησης και ανάκτησης. Η ερευνά πάνω στο MIR είχε αρχίσει ήδη από την δεκαετία του 1990 και οι παράγοντες που βοήθησαν στην ανάπτυξη του να είναι οι εξής:

- Η ανάπτυξη των μεθόδων συμπίεσης του ήχου.
- Η αναβάθμιση των ηλεκτρονικών υπολογιστών.
- Η αυξημένη δημοτικότητα των κινητών music player.
- Η κυκλοφορία εφαρμογών και ηλεκτρονικών υπηρεσιών, που έκαναν δυνατή την αναπαραγωγή μουσικής οπουδήποτε και πάντα διαθέσιμη.

Είναι ένας αναπτυσσόμενος τομέας που ξεκίνησε την άνοδο του το 2000. Είναι σε μεγάλο βαθμό εστιασμένος στους χρήστες και εκμεταλλεύεται τις νέες τεχνολογίες της ανάκτησης. Ο τομέας αυτός μπορεί μέσα από την έρευνα να αναπτυχθεί ταυτόχρονα με το γενικό πεδίο της ανάκτησης του ήχου.

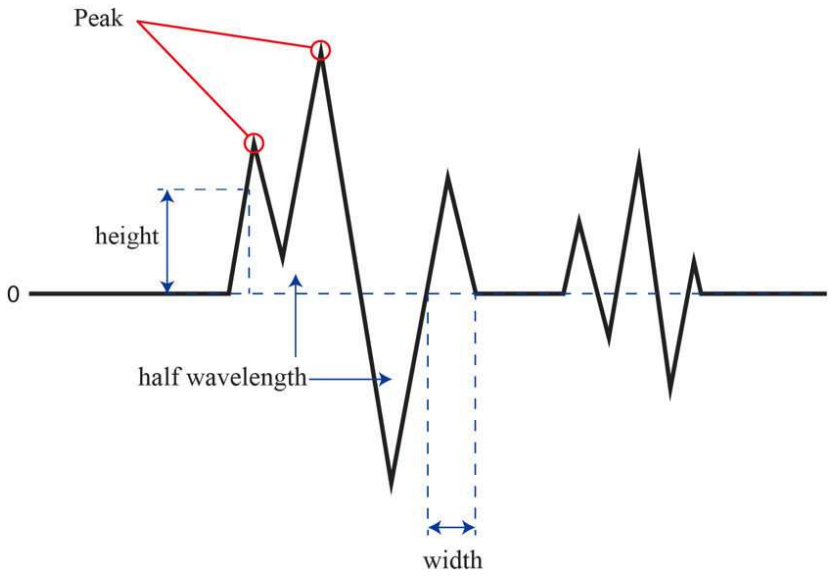
Περιλαμβάνει έναν αριθμό διαφορετικών προσεγγίσεων με στόχο την διαχείριση της μουσικής, την εύκολη πρόσβαση και την απόλαυση της από τους ακροατές και απώτερα να γίνει πιο προσιτή η δημιουργία της από τους μουσικούς και επαγγελματίες.

Οι εφαρμογές που ασχολούνται με την ανάκτηση της μουσικής έχουν ως στόχο την εύρεση τραγουδιών σε μεγάλες συλλογές. Για την δημιουργία μιας τέτοιας εφαρμογής χρειάζεται μεγάλος αριθμός ακρίβειας και μικρός βαθμός λεπτομέρειας (Yu & Deng, 2016), όπως και η κατανόηση των βασικών εννοιών της μουσικής θεωρίας.

4.1.1 Περιγραφή Μουσικού Περιεχόμενου

Για να γίνει κατανοητή η διαδικασία της περιγραφής του μουσικού περιεχομένου, απαιτείται η βασική γνώση λειτουργίας του ήχου.

Ο ήχος είναι ό,τι αισθανόμαστε με την ακοή (Μητσόπουλος, 2007). Είναι κύματα που μετακινούνται στον αέρα ή στο νερό. Τα κύματα που παράγονται μπορούν να μετρηθούν με την συχνότητα και το πλάτος τους.



Εικόνα 34. Γραφική αναπαράσταση της συχνότητας και των κυμάτων (Seo, Yeong-Seok & Huh, Jun-Ho, 2019).

Η συχνότητα είναι ο αριθμός των κυμάτων που περνούν από ένα σταθερό σημείο σε ένα δεδομένο χρονικό διάστημα (ανά δευτερόλεπτο). Μετριέται με την μονάδα των Hertz (Hz). Πχ. 50Hz = 50 επαναλήψεις το δευτερόλεπτο. Το πλάτος του κύματος αναφέρεται στο μέγεθος κάθε επανάληψης.

4.1.2 Οι ιδιότητες του ήχου

Ο ήχος και πιο συγκεκριμένα η μουσική, αποτελείται από δεκάδες ιδιότητες, ικανές να την περιγράψουν όσο πιο αναλυτικά γίνεται. Στον βασικό του ορισμό, υπάρχουν τρεις ιδιότητες που κάνουν κάθε ήχο μοναδικό:

1. **Το εύρος της έντασης του ήχου (amplitude).** Είναι το μέγεθος της δόνησης, το οποίο αντιλαμβανόμαστε ως την ένταση (loudness) του ήχου.
2. **Η συχνότητα (frequency).** Είναι ο ρυθμός με τον οποίο πραγματοποιείται η δόνηση. Η συχνότητα ενός ήχου είναι αυτό που αντιλαμβανόμαστε ως ύψος τόνου (pitch).
3. **Ο χρόνος (time).**

Όταν μιλάμε για ήχο που δημιουργείται από ένα μουσικό όργανο, τότε παράγεται ένα σύνολο από κύματα, δηλαδή ένα σύνολο ήχων. Ο κάθε ήχος έχει την δική του συχνότητα. Όλες οι συχνότητες προέρχονται από την χαμηλότερη συχνότητα, η οποία αυτή ονομάζεται θεμελιώδης συχνότητα (fundamental frequency). Όλες οι υπόλοιπες συχνότητες πάνω από την θεμελιώδη ονομάζονται αρμονικές (overtones).

Σημαντικό είναι επίσης να ξεχωρίσουμε το γεγονός ότι στην περίπτωση των μουσικών οργάνων, πολλά από τα όργανα μπορούν να παίζουν τις ίδιες νότες. Δηλαδή μια νότα στο βιολί και μία στο πιάνο ακούγεται διαφορετικά. Η διαφορά που υπάρχει σε κάθε όργανο, εκφράζεται με τα παρακάτω χαρακτηριστικά:

- **Η χροιά (Timbre):** Είναι η ξεχωριστή ιδιότητα του τόνου του κάθε μουσικού οργάνου και της φωνής. Το κάθε όργανο έχει την δική του χροιά, και για αυτό αναγνωρίζουμε το καθένα ξεχωριστά.
- **Το ύψος τόνου (Pitch):** Είναι η συχνότητα του κάθε τόνου σε ένα μουσικό όργανο. Συχνά περιγράφεται ως υψηλό ή χαμηλό.
- **Η ένταση (Intensity):** Σχετίζεται με το εύρος του κύματος (amplitude), και συνεπώς με την ενέργεια της δόνησης. Το εύρος της έντασης περιγράφεται από ασθενές έως ισχυρό. Μονάδα μέτρησης της έντασης είναι τα dB (decibel).

Πέρα από τις ιδιότητες αυτές και για την καλύτερη κατανόηση των συστημάτων αναγνώρισης και ανάκτησης ήχου, υπάρχουν και άλλες έννοιες και γνωρίσματα που απαρτίζουν το σύνολο των μουσικών κομματιών που έχουν δημιουργηθεί. Μερικές από αυτές είναι:

- **Ακουστική (Acoustics):** Αποτελεί επέκταση της χροιάς που αποδίδεται από εξωτερικά χαρακτηριστικά όπως η ακουστική δωματίου, θόρυβος υποβάθρου, μετα-επεξεργασία ήχου, φιλτράρισμα και εξισορρόπηση(equalization).

- **Ρυθμός (Rhythm):** Είναι η περιοδική τοποθέτηση επαναλαμβανόμενων ήχων στο χρόνο. Ο κάθε ρυθμός ενώνεται με μικρές διαμορφώσεις μέσα στο τραγούδι, ανάλογα με την ταχύτητά του, μπορεί να είναι αργός ή γρήγορος. Διαφορετικοί ρυθμοί γίνονται αντιληπτοί ταυτόχρονα στην περίπτωση της πολυρυθμικής μουσικής.
- **Μελωδία (Melody):** Μία ακολουθία τόνων που αποτελείται από ρυθμό, εντάσεις και κλίμακες με σκοπό να τις λαμβάνει ως σύνολο ο ανθρώπινος νους.
- **Αρμονία (Harmony):** Ταυτόχρονοι ήχοι με αναγνωρίσιμο ύψος τόνου.
- **Τέμπο (Tempo):** Η ταχύτητα με την οποία ένα μουσικό έργο παίζεται σε ένα συγκεκριμένο χρονικό διάστημα. Συνήθως μετριέται σε χτύπους ανά λεπτό.

Το σύνολο όλων αυτών των γνωρισμάτων και πολλών άλλων είναι που δημιουργούν την δομή και τα χαρακτηριστικά (επαναλήψεις, η εναλλαγή θεμάτων και ρεφραίν, η παρουσία διαλειμμάτων, οι αλλαγές των χρονικών υπογραφών), αποτελούν την περιγραφή ενός μουσικού έργου. Τα χαρακτηριστικά στο σύνολο τους συλλέγονται και στην συνέχεια εκτιμούνται για το αν θα συνεισφέρουν στην επιτυχή ανάκτηση.

Μέθοδοι ανάκτησης στα MIR

Ένα σύστημα ανάκτησης της μουσικής MIR συνήθως αποτελείται από κάποιες βασικές μεθόδους και τεχνολογίες που επιτρέπουν σε κάθε σύστημα την δυνατότητα της εύρεσης της κατάλληλης πληροφορίας. Κάθε εφαρμογή και τεχνολογία MIR μπορεί να αντιμετωπίζει διαφορετικές προσεγγίσεις για να επιτύχει την ανάκτηση (Choi, Fazekas, Cho & Sandler, 2018). Το ευρύ πλαίσιο των χαρακτηριστικών του ήχου έχει οδηγήσει σε έναν αρκετά μεγάλο αριθμό τεχνολογιών, για να καταφέρουν να ανταπεξέλθουν σε καθένα από αυτά τα χαρακτηριστικά της μουσικής.

Προσπαθήσαμε να στηριχτούμε στις πιο διαδεδομένες τεχνολογίες που στην πλειοψηφία τους τα μέσα ανάκτησης μουσικής χρησιμοποιούν, χωρίς απαραίτητα να περιορίζονται μόνο σε αυτές, περιγράφοντας τες περιεκτικά.

Audio identification

Η μεγαλύτερη δυσκολία στην ανάκτηση των χαρακτηριστικών της μουσικής και του ήχου, όπως και στης εικόνας, είναι ο τρόπος που θα πρέπει να περιγράφει το καθένα από αυτά ώστε να είναι κατανοητά για ένα υπολογιστικό σύστημα.

Η ιδιαιτερότητα του ήχου απαιτεί διαφορετικές τεχνικές στην ανάκτηση από την εικόνα. Για να πετύχουμε την συγκεκριμένη ανάκτηση, χρειάζεται να συλλέξουμε τον ήχο και να τον μετρήσουμε για ένα συγκεκριμένο χρονικό διάστημα (Grosche, Müller & Serrà, 2012).

Σε αντίθεση με την εικόνα, που έχουμε ολόκληρο ή ένα αντιπροσωπευτικό μέρος της για ανάκτηση, δεν είναι πρακτικό να χρησιμοποιήσουμε ολόκληρο το τραγούδι. Τα ηχητικά κύματα είναι συνεχή, όμως για την μετατροπή όλων των κυμάτων σε ψηφιακά χρειάζονται υπερβολικά μεγάλοι πόροι και χρήματα. Πέρα από την ανάκτηση ενός μόνο τραγουδιού, έχουν γίνει προσπάθειες ώστε τα τραγούδια να μπορούν να κατηγοριοποιηθούν ανάλογα με το είδος, να συλλέγονται οι στοίχοι και να εμφανίζεται μία λίστα με τα τραγούδια που ο χρήστης πιθανόν να ενδιαφέρεται να ακούσει.

Στην πάροδο των χρόνων έχουν υπάρξει πολλοί μέθοδοι αναπαράστασης της μουσικής, όμως λίγες έχουν επικρατήσει. Το κάθε ηχητικό κύμα που παράγεται μπορεί να αποτυπωθεί με διάφορους τρόπους μέσω γραφικών αναπαραστάσεων.

Αναφορικά, οι πιο γνωστές αναπαραστάσεις για τον ήχο είναι:

- **Short-Time Fourier Transform (STFT)**
- **Constant-Q Transform (CQT)**
- **Chromagram**

Track Separation

Το track separation είναι η διαχώριση ενός μουσικού κομματιού σε υποενότητες με βάση τις διαφοροποιήσεις του ήχου σε αυτές.

Ο διαχωρισμός του ήχου (Track separation) παρέχει λύσεις για την εξαγωγή ενός συγκεκριμένου οργάνου ή ήχου από ένα αρχείο ήχου (audio file) το οποίο είναι πολυφωνικό και είναι δυνατόν να περιέχει πολλές μελωδίες. Ο αλγόριθμος διαχωρίζει το αρχείο ήχου σε επαναλαμβανόμενο υπόβαθρο (background) και μη επαναλαμβανόμενο προσκήνιο (foreground) (Grosche, Müller & Serrà, 2012).

Η επανάληψη είναι μέρος της μουσικής. Οι ήχοι που επαναλαμβάνονται κατά την διάρκεια του μουσικού κομματιού δημιουργούν ένα υπόβαθρο που κάνει το τραγούδι και την αναπαραγωγή των μουσικών οργάνων ευκολότερο και γενικά πιο ευχάριστο στην ακρόαση.

Η ιδιαιτερότητα του κάθε είδους μουσικής παρέχει σημαντικές πληροφορίες για την ανάκτηση. Τα περισσότερα έργα του track separation, επικεντρώνονται στη συλλογή δεδομένων σε μονάχα ένα μουσικό όργανο (σε μεγάλο βαθμό τα ντραμς), το οποίο ανάλογα με το είδος μουσικής, μπορεί να οδηγήσει στην καλύτερη αναγνώριση του. Όσο περισσότερα είναι τα μουσικά όργανα που χρησιμοποιούντα στο track separation, τόσο πιο αποτελεσματικός θα είναι ο διαχωρισμός του. Ωστόσο, αυξάνεται η πολυπλοκότητα και χρειάζεται περισσότερη εκπαίδευση. Οποιαδήποτε διαφορά στην μελωδία μπορεί να σημαίνει την διαχώριση του από το υπόλοιπο τραγούδι (Grosche, Müller & Serrà, 2012).

Τέλος, μία νέα μέθοδος που ήδη εφαρμόζεται σήμερα είναι ο διαχωρισμός της μουσικής και των φωνητικών μέσα στο τραγούδι. Είναι μια αρκετά χρήσιμη τεχνολογία που η συμβολή της είναι αναμφίβολα εξαρτώμενη από το MIR και οποιοδήποτε άλλο παρακλάδι της ανάκτησης του ήχου.

Audio Fingerprinting

Ο όρος Audio Fingerprinting ή ηχητικά αποτυπώματα προέρχεται από τη μέθοδο για τον εντοπισμό των ανθρώπων μέσα από το δακτυλικό αποτύπωμα (fingerprint). Όπως και στην περίπτωση του ανθρώπου, έτσι και τα ηχητικά αποτυπώματα έχουν λίγες πληροφορίες για την προέλευσή τους, αλλά πληροφορίες που είναι τόσο μοναδικές που επιτρέπουν τον ακριβή εντοπισμό της προέλευσης τους (Grosche, Müller & Serrà, 2012).

Η χρήση δακτυλικών αποτυπωμάτων μειώνει τον απαιτούμενο χώρο αποθήκευσης, αυξάνει την αποτελεσματικότητα της σύγκρισης και επιταχύνει την αναζήτηση μέσω της βάσης δεδομένων. Συνήθως το δακτυλικό αποτύπωμα συνδέεται μόνο με μεταδεδομένα, όπου

όλες οι άσχετες πληροφορίες παραλείπονται. Όπως και στα συστήματα που χρησιμοποιούν δακτυλικά αποτυπώματα, έτσι και στα συστήματα που στηρίζονται στα ηχητικά αποτυπώματα εμπεριέχουν μια μέθοδο εξαγωγής αποτυπωμάτων και μια μέθοδο αναγνώρισής τους. Ένα ηχητικό αποτύπωμα είναι μια ψηφιακή σύνοψη που μπορεί να χρησιμοποιηθεί για την ταυτοποίηση ενός δείγματος ήχου ή για τον γρήγορο εντοπισμό παρόμοιων στοιχείων σε μια βάση δεδομένων ήχου.

Η κύρια χρήση των αλγορίθμων ηχητικών αποτυπωμάτων είναι η αναγνώριση μουσικών κομματιών. Αναλύοντας το αποτύπωμα του τραγουδιού, το σύστημα ανακτά δεδομένα για τον καλλιτέχνη, τον τίτλο, τον εκδότη, την ημερομηνία κυκλοφορίας και άλλα. Για παράδειγμα, όταν σιγοτραγουδάτε (humming) ένα τραγούδι σε κάποιον, δημιουργείται ένα δακτυλικό αποτύπωμα, επειδή εξάγεται από τη μουσική αυτό που θεωρείται ουσιώδες.

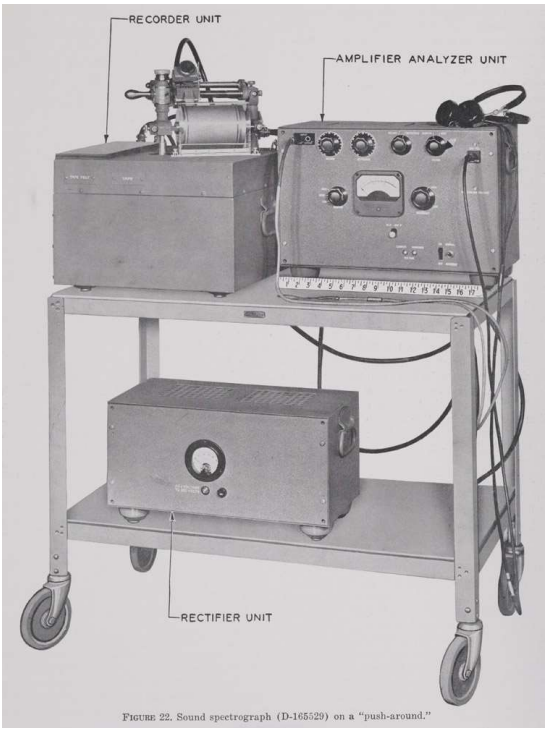
Το audio fingerprinting είναι κάτι από το οποίο μπορεί να επωφεληθούν οι επαγγελματικοί φορείς που ασχολούνται με την οργάνωση της μουσικής, καθώς μεγάλες ψηφιακές συλλογές ήχου μπορούν να οργανωθούν αυτόματα, αγνοώντας λανθασμένες ετικέτες και μεταδεδομένα (Grosche, Müller & Serrà, 2012).

Υπάρχουν διάφορες εταιρείες που προσφέρουν υπηρεσίες ταυτοποίησης μουσικής και χρησιμοποιούν το ηχητικό αποτύπωμα, όπως για την παρακολούθηση ραδιοφωνικών εκπομπών με σκοπό την προστασία των καλλιτεχνών ως προς τα πνευματικά δικαιώματα. Με αυτόν τον τρόπο ένας καλλιτέχνης, χορηγοί ή διαφημιστές μπορούν να παρακολουθήσουν εάν ένας ραδιοφωνικός σταθμός πληροί τους όρους συμβολαίου ή της σύμβασης.

Ένα αποτελεσματικό εργαλείο για να παρουσιάσουμε και να συγκρίνουμε τα ηχητικά δεδομένα είναι το φασματογράφημα (Spectrogram), το οποίο θεωρείται και η βάση του ηχητικού αποτυπώματος.

Φασματογράφημα

Το κάθε ηχητικό κύμα που παράγεται μπορεί να αποτυπωθεί στο φασματόγραμμα. Κατά την διάρκεια του 2ου παγκοσμίου πολέμου, η Bell Telephone Company, σχεδίαζε μία συσκευή η οποία μπορεί να αποτυπώνει συνεχή ηχητικά κύματα σε ένα κομμάτι χαρτί. Η πρώτη αναφορά αυτής της αναλογικής συσκευής έγινε το 1946 και είχε μεγάλη επιτυχία στην επιστημονική κοινότητα (Feaster, 2018).

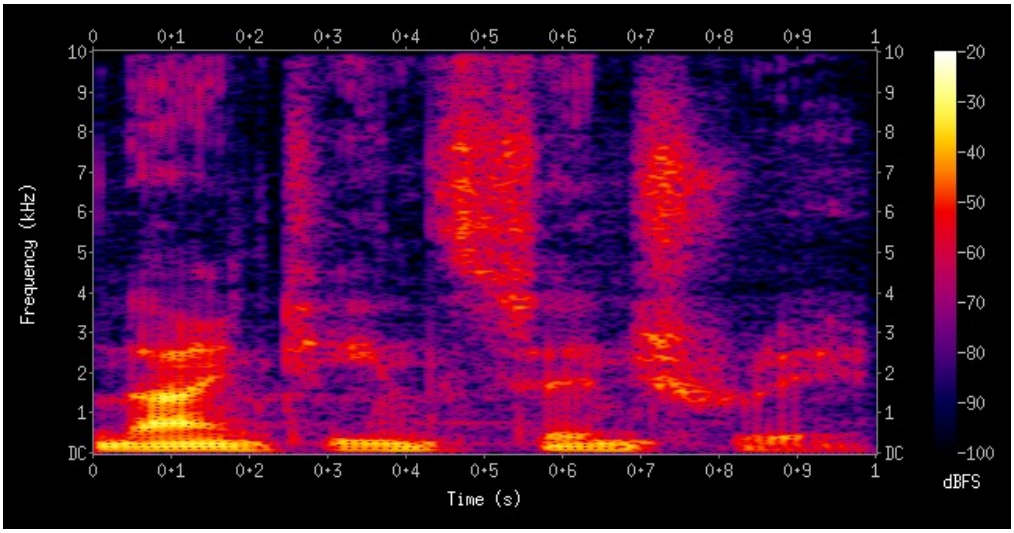


Εικόνα 35. Η πρώτη συσκευή φασματογράμματος.

https://griffonagedotcom.wordpress.com/2018/07/26/the-secret-military-origins-of-the-sound-spectrograph/

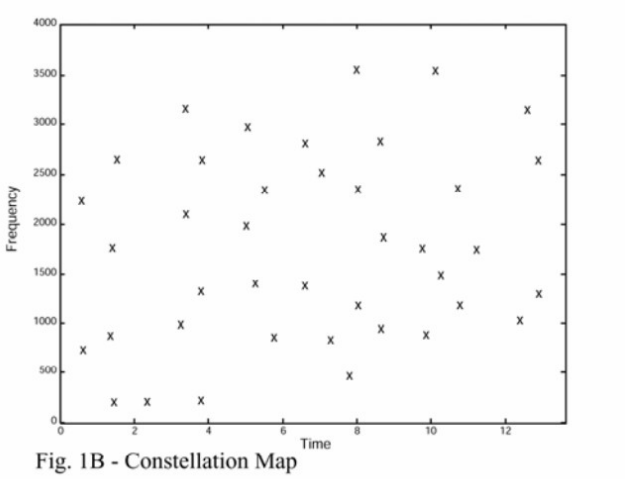
Από τότε, το υλικό αυτό έχει ενσωματωθεί στα ψηφιακά μέσα και μπορεί να καταγράφει τη συχνότητα του κάθε κύματος την δεδομένη χρονική στιγμή. Συνήθως υπάρχει ο κάθετος (y) άξονας που δείχνει την χρονική στιγμή και ο οριζόντιος (x) που δείχνει την συχνότητα. Στην περίπτωση που εμφανιστεί στο διάγραμμα ένα χρώμα, αυτό απεικονίζει την ένταση της κάθε συχνότητας (db). Χάρη στον μετασχηματισμό Fourier, το οποίο είναι μια μαθηματική τεχνική που μετασχηματίζει μια συνάρτηση του χρόνου σε συνάρτηση της συχνότητας, είμαστε ικανοί να εξάγουμε την μουσική πληροφορία. Έτσι, είναι δυνατόν να αποθηκεύσουμε την πληροφορία αυτή και να την συγκρίνουμε με άλλες.

Οι επιλογές που συνθέτουν το spectrogram κατάφεραν να είναι ανθεκτικές σε ορισμένα ζητήματα γνωστά και ως παραμορφώσεις, όπως το white noise (λευκός θόρυβος), έναν θόρυβο που παράγεται συνδυάζοντας πολλές διαφορετικές συχνότητες μαζί. Επειδή ο λευκός θόρυβος περιέχει όλες τις συχνότητες, χρησιμοποιείται συχνά για τη συγκάλυψη άλλων ήχων, χωρίς αυτό να έχει τεράστιο αντίκτυπο στις ιδιαίτερα ισχυρές κορυφές (Choi, Fazekas, Cho & Sandler, 2018).



Εικόνα 36. Το φασματόγραμμα. https://commons.wikimedia.org/w/index.php?curid=5544473

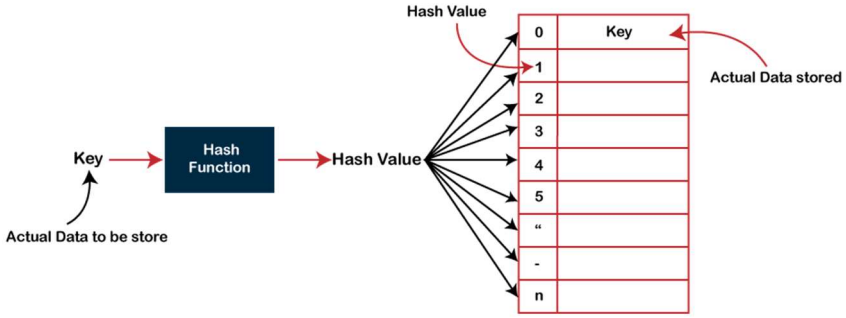
Το φασματογράφημα αποτελεί την βάση αρκετών συστημάτων ανάκτησης μουσικής. Το Shazam είναι μία από τις πρώτες εταιρίες που δημιούργησαν έναν σχετικό αλγόριθμο. Περιείχε τη χρήση ενός φασματογραφήματος για τον εντοπισμό των ισχυρότερων κορυφών έντασης (peaks) και την αποθήκευση αυτών των κορυφών στην βάση δεδομένων τους. Από τις υψηλότερες τιμές που προέκυψαν από το φασματογράφημα, είμαστε σε θέση να δημιουργήσουμε το διάγραμμα διασποράς (Scatter Plot ή Constellation Map).



Εικόνα 37. Διάγραμμα διασποράς.

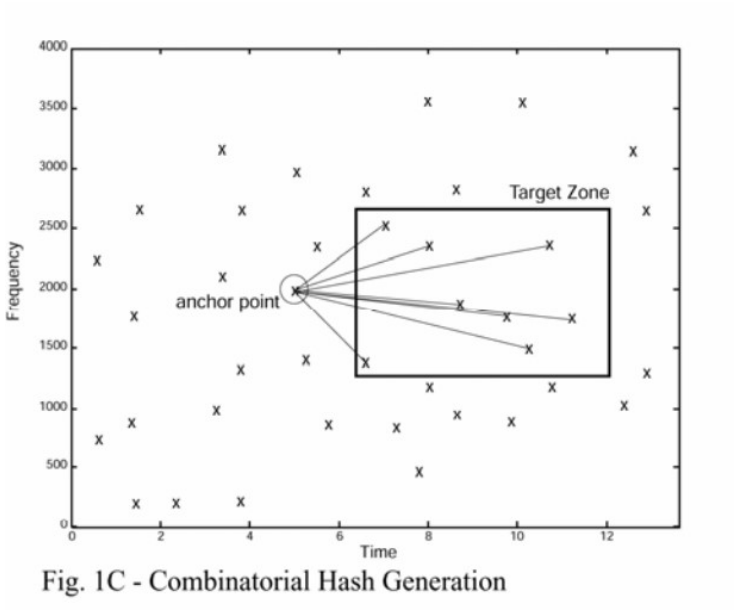
https://miro.medium.com/max/1292/0*goypmWSJuXKRChgx.

Το διάγραμμα διασποράς (Scatter Plot ή Constellation Map), παίζει σημαντικό ρόλο στην καταγραφή του αποτυπώματος και στην συνέχεια στην σύγκριση των μουσικών κομματιών. Η αποθήκευση του διαγράμματος γίνεται μέσω του πίνακα κατακερματισμού (Hash Table). Χρησιμοποιεί μια τεχνική γνωστή ως την συνάρτηση κατακερματισμού (Hash Function) για την αποθήκευση δεδομένων (data structure) που αντιστοιχεί τις κατάλληλες τιμές με τα κλειδιά τους (keys) (Shahrior, 2021).



Εικόνα 38. Η δομή ενός Hash table. https://www.javatpoint.com/hash-table

Ο πίνακας είναι μονάχα ένα μέρος της επιτυχίας της αντιστοίχισης. Μέσω μιας διαδικασίας που ονομάζεται συνδυαστικός κατακερματισμός (combinatorial hashing), επιλέγεται μία κορυφή του διαγράμματος διασποράς ως επίκεντρο (anchors) και μετά συνδέεται με άλλες κορυφές του διαγράμματος ώστε να σχηματιστεί το σημείο επίκεντρου (anchor point). Το σημείο επίκεντρου μιας συγκεκριμένης χρονικής περιόδου και συχνότητας, είναι γνωστό και ως target zone. Ο σκοπός ενός target zone είναι η ένωση των διαφορετικών σημείων μεταξύ τους και η διευκόλυνση του συστήματος στην αναζήτηση.



Εικόνα 39. Combinatorial Hash Generation.

https://miro.medium.com/max/630/1*bMp__nztMgJJJNWQbPGFLg.png

Κάθε ζεύγος επίκεντρου (anchor-point) αποθηκεύεται σε έναν πίνακα που περιέχει την συχνότητα του επίκεντρου, την συχνότητα της κορυφής (peak) και το χρόνο μεταξύ επίκεντρου και κορυφής, σχηματίζοντας το ενδεικτικό hash. Από την στιγμή που ο χρήστης στείλει το δείγμα προς αναζήτηση και εφόσον έχει ολοκληρωθεί η διαδικασία του audio fingerprinting των ηχητικών αρχείων, τότε είναι δυνατή η διαδικασία αντιστοίχισης. (Pubudumal, 2018). Τα δεδομένα αυτά συνδέονται με έναν πίνακα (hash table) όπου περιέχει τον χρόνο μεταξύ του επίκεντρου (anchor) και της αρχής του αρχείου του δείγματος.

Τα σημεία του target zone που συλλέχθηκαν από την ηχογράφηση του χρήστη μεταφέρονται στη βάση δεδομένων του Shazam. Από εκεί, δημιουργείται μια βαθμολογία αντιστοίχισης (score) μεταξύ των ζευγών των σημείων επίκεντρου (anchor-point) και προκύπτουν τα τραγούδια με το μεγαλύτερο score. Στην συνέχεια υπολογίζει κάποιες συγκεκριμένες χρονικές στιγμές, ώστε να γίνει η σωστή χρονική αντιστοιχία μεταξύ όλων των αποτελεσμάτων ώστε να βρεθεί και να επιστέφει το κατάλληλο αποτέλεσμα.

4.1.3 Λοιπές λειτουργίες MIR

Πέρα από την γνωστή μέθοδο ανάκτησης μουσικής μέσα από το audio fingerprinting, υπάρχουν διάφορες άλλες λειτουργίες και εννοείς που επικεντρώνονται σε διαφορετικές πτυχές ή ιδιότητες ενός μουσικού κομματιού. Αποσπούν ένα χαρακτηριστικό του ήχου ή μια νέα συγκεκριμένη μέθοδος και εστιάζουν σε αυτό, είτε για να βελτιώσουν, είτε να συνεισφέρουν στην ανάκτηση.

Παρακάτω θα αναφέρουμε κάποιες από αυτές που εφαρμόζονται στα σύγχρονα συστήματα.

4.1.3.1 Query by humming and query by tapping

Αναζήτηση με βάση την εισαγωγή της μελωδίας ή ρυθμού. Το query by tapping λειτουργεί κυρίως με τραγούδια που είναι αποθηκευμένα σε μορφή midi, λόγω της απλοϊκότητας των δεδομένων. Χαρακτηριστικό παράδειγμα είναι η ιστοσελίδα Musipedia.

Το query by humming είναι η αναζήτηση ενός τραγουδιού με την χρήση μιας φωνητικής μελωδίας από τον χρήστη και την αντιστοίχιση του με το πιο κοντινό σε αυτό. Η τεχνολογία αυτή είναι σε πρώιμο στάδιο ακόμα, αλλά εταιρίες όπως η Google, η Soundhound φαίνεται να πιστεύουν στην προοπτική τους.

4.1.3.2 Tempo Estimation (εκτίμηση τέμπο)

Το τέμπο της μουσικής σε γενικές γραμμές αντιδρά στην συχνότητα των κυμάτων, δηλαδή στην ταχύτητα που οι άνθρωποι αντιδρούν στα χτυπήματα της μουσικής. Το tempo estimation (εκτίμηση τέμπο) σπεύδει να απλοποιήσει και να επιταχύνει την εύρεση της τιμής του τέμπο, είναι μία από τις θεμελιώδεις διαδικασίες στο music information retrieval καθώς

Είναι ένα αναπόσπαστο κομμάτι στην αναγνώριση της μουσικής και αποτελεί τη βάση άλλων τύπων ανάλυσης του ήχου, όπως η ανίχνευση του beat (beat tracking) και η ανίχνευση μοτίβων (pattern recognition). Υπάρχει μεγάλος όγκος εργασίας στον τομέα της εκτίμησης του τέμπο χρησιμοποιώντας μια ποικιλία διαφορετικών προσεγγίσεων που διαφέρουν ως προς την ακρίβειά τους καθώς και ως προς την πολυπλοκότητα τους. Η εκτίμηση του τέμπο (tempo estimation) έδειξε ότι αυτό το έργο μπορεί να είναι δύσκολο λόγω της ιδιαιτερότητας κάποιων ειδών μουσικής. Συγκεκριμένα το σφάλμα είναι πολύ ορατό στα ειδή με μεταβαλλόμενο τέμπο όπως το thrash metal ή grindcore που είχαν σχεδόν πέντε φορές μεγαλύτερο μέσο σφάλμα από πολύ απλούστερα είδη όπως το rock ή black metal.

4.1.3.3 Chords Recognition (Αναγνώριση συγχορδιών)

Ένα από τα προβλήματα στην ανάκτηση μουσικών πληροφοριών ήταν η αναγνώριση συγχορδιών. Ένας από τους καλύτερους τρόπους για να πραγματοποιηθεί το chord recognition, είναι με την χρήση των νευρωνικών δικτύων.

Το pitch class profile vector (διάνυσμα προφίλ κλάσης του ύψους του τόνου) χρησιμοποιείται στη μέθοδο του νευρωνικού δικτύου. Το διάνυσμα παρέχει μόνο 12 στοιχεία για τιμές των ημιτόνων, αλλά είναι επαρκής για το έργο της αναγνώρισης. Φυσικά υπάρχουν και άλλες μέθοδοι για την εκτέλεση αυτού του έργου, και οι περισσότερες από αυτές βασίζονται στη κατηγορία προφίλ του ύψους του τόνου (pitch class profiling) για να αλλάξουν τη συγχορδία σε μια πιο αντιληπτή μορφή, που θα βοηθήσει διαδικασία της αναγνώρισης, η οποία αναγνώριση βασίζεται σε πολύ σύνθετες μεθόδους που χρησιμοποιούν αρκετή μνήμη συστήματος.

Οι αναπτυσσόμενες τεχνολογίες και σύγχρονα μέσα έχουν κάνει δυνατή την αναγνώριση συγχορδιών ακόμα και σε συνθήκες που αλλοιώνουν τον ήχο ή επικρατεί θόρυβος. Οι υψηλές επιδόσεις που προκύπτουν είναι δυνατές χάρη στην εκπαίδευση του δικτύου (neural network) μέσα από αθόρυβες ηχογραφήσεις, που βοηθάνε στην εξοικείωση των συστημάτων αυτών. Για αυτό, στην αναγνώριση χορδών (chord recognizing) το πιο σημαντικό είναι ένα σύνολο δεδομένων dataset καλής ποιότητας.

4.1.3.4 Audio Alignment (Ευθυγράμμιση ήχου).

Το εργαλείο ευθυγράμμισης ήχου αναλύει τον ήχο που θέλετε να επεξεργαστείτε. Επιτρέπει να συγχρονίσετε διαφορετικά όργανα ή φωνητικά κομμάτια και γενικά την αυτόματη αντιστοίχιση ήχων που ο χρήστης θέλει να αναπαράγονται ταυτόχρονα. Μπορεί επίσης να επιλύσει προβλήματα phasing που εμφανίζονται όταν χρησιμοποιούνται διαφορετικά μικρόφωνα στην ίδια λήψη ("Audio Alignment", n.d.).

4.1.4 Προβλήματα στην ανάκτηση μουσικής

Οι σχεδιαστές των συστημάτων ανάκτησης πρέπει να αντιμετωπίσουν αρκετούς παράγοντες που μπορεί να αλλοιώσουν το αποτέλεσμα της ανάκτησης. Μερικά από τα προβλήματα που εμφανίζονται είναι:

- **Το τραγούδι-δείγμα δεν είναι αντιπροσωπευτικό:** Είναι πολύ πιθανό το δείγμα που θα στείλει ως ερώτημα ο χρήστης να μην είναι αρκετά μεγάλο ή να μην είναι σε σημείο που να μπορεί το σύστημα να αναγνωρίσει το τραγούδι. Πολλές φορές, ο χρήστης δεν έχει στην διάθεση του ολόκληρο το μουσικό κομμάτι και αναγκάζεται να χρησιμοποιήσει αυτό που έχει με ανάμεικτα αποτελέσματα.

- **Ο θόρυβος στο περιβάλλον:** Η αναγνώριση είναι ιδιαίτερα δύσκολη όταν υπάρχει εξωτερικός θόρυβος και βαβούρα που μπερδεύεται με την μουσική σύνθεση την στιγμή που προσπαθεί να αναγνωρίσει το σύστημα το δείγμα.

- **Κακή ποιότητα του ήχου:** Υπάρχουν περιπτώσεις που η ίδια η ποιότητα του ήχου είναι κακή και σε κατάσταση που δεν μπορεί το σύστημα να κάνει την οποιαδήποτε ανάκτηση. Πχ. Όταν πρόκειται για μουσική που δεν έχει γίνει ποιοτική ηχογράφηση ή όταν η ιδία η συσκευή που αναπαράγει μουσική έχει κακή ποιότητα στο ηχείο. Δοκιμές που έγιναν με τα ηχητικά αποτυπώματα audio fingerprinting έδειξαν ότι είναι πολύ σημαντικό η καταγραφή και η ανάλυση να γίνεται στις καλύτερες δυνατές συνθήκες για μεγαλύτερα ποσοστά επιτυχίας. Τα κακώς καταγεγραμμένα μέρη δεν μπορούσαν να αναγνωριστούν με ακρίβεια.

- **Διαφορετική αποτελεσματικότητα ανάλογα με το είδος της μουσικής:** Μερικά είδη μουσικής, όπως το grindcore, thrashcore, thrash metal, κ.α., λόγω του ρυθμού ή της πολυπλοκότητας της μουσικής εμφανίζουν ποικίλα αποτελέσματα μικρότερης ακρίβειας από τα υπόλοιπα.

Χάρη στην τεχνολογία των ηλεκτρονικών συσκευών, μερικά από αυτά τα προβλήματα έχουν λυθεί σε μεγάλο βαθμό. Με τα νέα πρότυπα που υπάρχουν στην αγορά των κινητών τηλεφώνων και γενικά στα περιφερειακά των ηλεκτρονικών συσκευών, η ποιότητα των μικρόφωνων είναι καλύτερη σε σχέση με τα προηγούμενα χρόνια και η εμφάνιση των νέων τεχνολογιών μείωσης του θορύβου (noise cancellation), κάνουν την ανάκτηση πιο αποτελεσματική.

4.2 Παραδείγματα ανάκτησης μουσικής

Παρακάτω θα μιλήσουμε για τις εφαρμογές που είναι στην κορυφή της λίστας ως προς τους ενεργούς χρήστες και την σημασία που έχουν στην ανοδική πορεία της ανάκτησης μουσικής τα τελευταία χρόνια.

4.2.1 Shazam

Το Shazam είναι πλέον μια από τις πιο διαδεδομένες εφαρμογές αναγνώρισης της μουσικής. Όπως σε κάθε άλλη εφαρμογή ανάκτησης μουσικής, έτσι και στο Shazam συναντάμε την ίδια δομή, όπου με ένα σύντομο δείγμα που συλλέγεται χρησιμοποιώντας το μικρόφωνο της συσκευής, καταφέρνει να δώσει στον χρήστη το όνομα και τον καλλιτέχνη πίσω από τη μουσική που παίζει.

Εικόνα 40. Η απλοϊκή και γρήγορη σχεδίαση του Shazam, επιτρέπει στον χρήστη να μάθει περισσότερα για το τραγούδι που ψάχνει.

Η ιστορία του Shazam ξεκινάει το 1999, όταν ιδρύεται από τους Dhiraj Mukherjee, Chris Barton, Philip Inghelbrecht και Avery Wang. Οι ιδρυτές έγραψαν το λογισμικό ανάλυσης ήχου που επέτρεπε στο πρόγραμμα να αναγνωρίζει οποιοδήποτε τραγούδι. Όπως πολλές άλλες επιχειρήσεις, το Shazam ξεκίνησε από δύο φίλους που συζητούσαν την μεγάλη ιδέα.

Η μουσική υπηρεσία ήταν μπροστά από την εποχή της όταν ξεκίνησε το 1999 αλλά το σωστό επιχειρηματικό μοντέλο δεν ήταν εύκολο να βρεθεί.

Στα τέλη της δεκαετίας του '90, δεν υπήρχαν smartphones, δεν υπήρχαν εφαρμογές και το iTunes δεν είχε ακόμη εφευρεθεί. Για να χρησιμοποιήσουν το Shazam, οι άνθρωποι καλούσαν έναν αριθμό, έβαζαν το τηλέφωνό τους στο ραδιόφωνο και στη συνέχεια λάμβαναν ένα μήνυμα που αναγνώριζε το τραγούδι. Ήταν δύσκολο, αλλά η τεχνολογία αργούσε να φτάσει το όραμα του Shazam για το μέλλον.

Το 2002, το Shazam είχε 1 εκατομμύριο τραγούδια στη βάση δεδομένων του και χρειαζόταν 15 δευτερόλεπτα για να επεξεργαστεί το αίτημα ενός χρήστη.

Σήμερα, χρειάζεται μόλις 2 δευτερόλεπτα για να εντοπίσει το τραγούδι αναμεσά σε πάνω από 30 εκατομμύρια τραγούδια. Ακόμα και σε περιπτώσεις που μπορεί να υπάρχει θόρυβος στο παρασκήνιο (background noise) ή στην περίπτωση διασκευών (remix,cover songs).

4.2.1.1 Λειτουργία

Το Shazam δεν αναζητεί τα ίδια τα αρχεία ήχου. Αντιθέτως, έχει ένα ηχητικό αποτύπωμα για κάθε αρχείο ήχου στη βάση δεδομένων του.

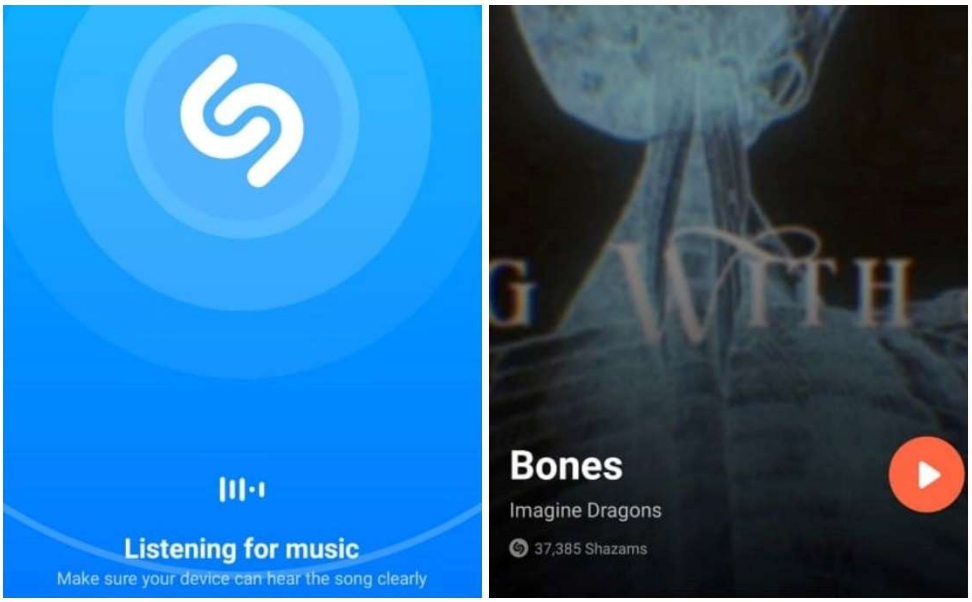
Εφόσον το αποτύπωμα του μουσικού κομματιού έχει δημιουργηθεί **(βλέπε κεφ. 4.2.3)** είναι έτοιμο να αποθηκευτεί στην βάση δεδομένων του Shazam και να συγκριθεί με το δείγμα.

Η ηχογράφηση που υποβάλλει ένας χρήστης του Shazam αποτελεί το δείγμα, αυτό μετατρέπεται σε ηχητικό αποτύπωμα για να γίνει η σύγκριση και να επιστραφεί το κατάλληλο αποτέλεσμα. Τα αποτυπώματα ήχου αποτελούνται από συλλογές αριθμητικών δεδομένων που επιτρέπουν στο Shazam να κάνει συγκρίσεις με ακρίβεια και ταχύτητα. (Cooper,2018).

Δηλαδή καταγράφοντας από οποιαδήποτε συσκευή τουλάχιστον 20 δευτερόλεπτα από ένα τραγούδι, ανεξάρτητα από το αν πρόκειται για την εισαγωγή, το ρεφρέν ή οποιοδήποτε άλλο μέρος του τραγουδιού, δημιουργείται ένα αποτύπωμα (audio fingerprinting) για το καταγεγραμμένο δείγμα. Εκμεταλλεύεται την βάση δεδομένων και χρησιμοποιεί τον

αλγόριθμο αναγνώρισης μουσικής που εμπεριέχει για να βρει ακριβώς οποιοδήποτε τραγούδι αναζητεί ο χρήστης . (Jovanovic,2015)

Η μέθοδος αναζήτησης ήχου του Shazam είναι αρκετά ακριβής ώστε να βρίσκει αντιστοιχίες παρά τον εξωτερικό θόρυβο, όπως ομιλίες ανθρώπων, θόρυβο του περιβάλλοντος, ακόμη και άλλα τραγούδια.



Εικόνα 41. Παράδειγμα αναζήτησης.

Δοκιμάσαμε την εφαρμογή του Shazam πάνω σε διάφορα είδη μουσικής και καταλήξαμε στο γεγονός ότι μπόρεσε να βρει αυτό που αναζητούσαμε με μεγάλη επιτυχία και ταχύτητα. Ακόμα και με διαφορετικές κινητές συσκευές, τα αποτελέσματα ήταν τα ίδια. Παρατηρήσαμε ότι το Shazam διαθέτει έναν άρτιο και αρκετά εξυπηρετικό αλγόριθμο και ότι το μόνο που θα μπορούσε να περιορίσει τα αποτελέσματα του Shazam είναι το τραγούδι να μην υπάρχει στην βάση δεδομένων του, κάτι το οποίο όμως δεν προέκυψε στις δοκιμές μας.

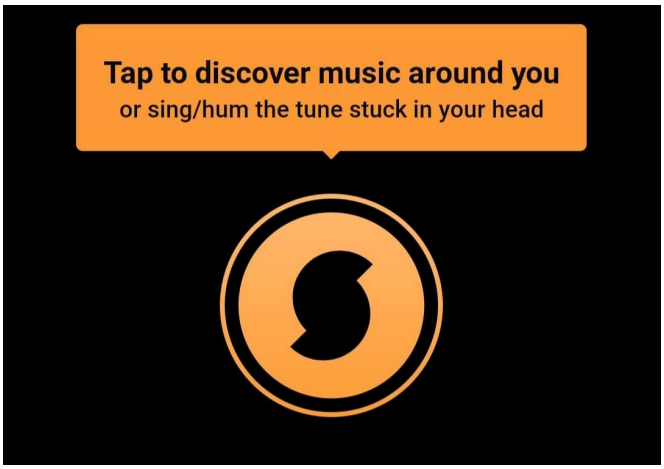
4.2.2 SoundHound

Στην άλλη πλευρά του ίδιου νομίσματος με το Shazam έχουμε και το SoundHound. Ήταν γνωστό ως Midomi μέχρι το 2009, μια εφαρμογή για τις κινητές συσκευές που ονομάστηκε από την εταιρεία δημιουργίας του (**SoundHound Inc.**). Ιδρύθηκε το 2005 από τον Keyvan Mohajer, έναν επιστήμονα της πληροφορικής.

Η εταιρία, από την δημιουργία της μέχρι και σήμερα, αποσκοπεί στο να αναπτύξει τεχνολογίες αναγνώρισης ομιλίας, κατανόησης της φυσικής γλώσσας, αναγνώρισης και ανάκτησης ήχου.

Για πάνω από 10 χρονιά είχαν αφιερώσει την έρευνά τους πάνω σε μία τεχνολογία θα που είχε τη δυνατότητα να φέρει επανάσταση στη φωνητική αλληλεπίδραση μεταξύ ανθρώπου και υπολογιστή. Μέχρι τότε όλες οι φωνητικές αλληλεπιδράσεις στηρίζονταν στο "speech to text" και "text to meaning" (ομιλία σε κείμενο και κατανόηση του από το σύστημα), όμως οι ιδρυτές γνώριζαν ότι για να ευδοκιμήσει μια πραγματική μηχανή κατανόησης φωνής, έπρεπε να κατανοεί άμεσα την ομιλία, όπως ακριβώς κάνουν και οι άνθρωποι.

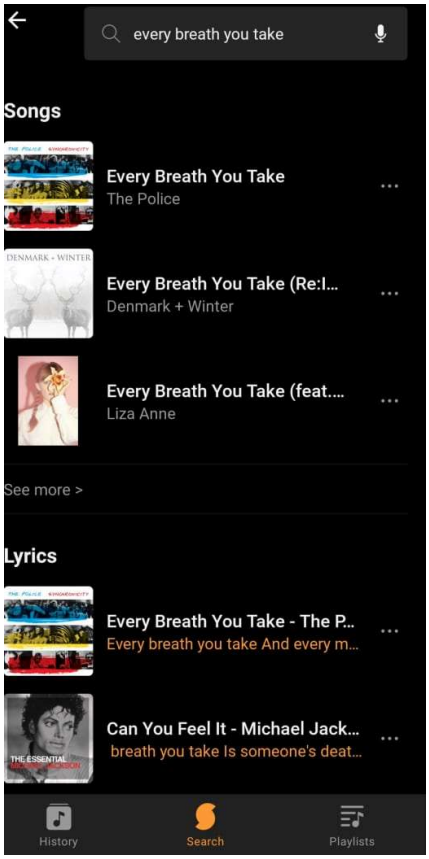
Μέσα από την πλατφόρμα Houndify Voice AI που ευνοήθηκε από τις επαναστατικές τεχνολογίες "Speech-to-Meaning" και "Deep Meaning Understanding", το Houndify προσέφερε ασύγκριτη φωνητική αλληλεπίδραση και οι δυνατότητές του εξακολουθούν να είναι στην κορυφή, μαζί με άλλα συστήματα ακόμα και σήμερα (Size, Size & WIRE, 2015).



Εικόνα 42. Η εφαρμογή του SoundHound.

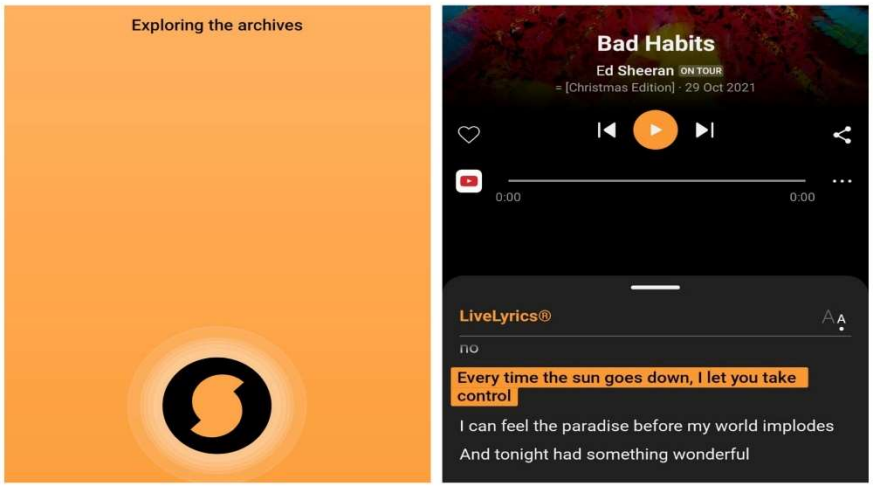
Τροφοδοτείται από τη δική του τεχνολογία αντιστοίχισης της μουσικής. Όπως γίνεται με τις περισσότερες εφαρμογές ανάκτησης μουσικής, το σύστημα αντιστοιχεί το ηχητικό αποτύπωμα από το δείγμα που δημιουργείται εκείνη την στιγμή, με μια βάση δεδομένων που περιέχει το σύνολο από τα αποτυπώματα μουσικής, δίνοντάς το καλύτερο δυνατό αποτέλεσμα, για κάθε είδος μουσικού κομματιού.

Η εφαρμογή απευθύνεται σε λάτρεις της μουσικής και όχι μόνο. Ο χρήστης μπορεί εύκολα να αναζητήσει ένα τραγούδι που έχει κολλήσει στο μυαλό του, πληκτρολογώντας (text based retrieval), ακόμα και τραγουδώντας το τραγούδι που θέλει με την ίδια του την φωνή (humming). Η τεχνολογία αντιστοίχισης του humming είναι ικανή να ταιριάξει τη μελωδία και το ρυθμό με τις εκατομμύρια ηχογραφήσεις χρηστών. Αν η αναζήτησή σας περιλαμβάνει λέξεις εκμεταλλεύεται και τους στίχους. Αν και είναι ευπρόσδεκτη η παρουσία της συγκεκριμένης δυνατότητας, τα αποτελέσματα τείνουν να είναι άστοχα όταν χρησιμοποιείται η φυσική ομιλία (Size, Size & WIRE, 2015).



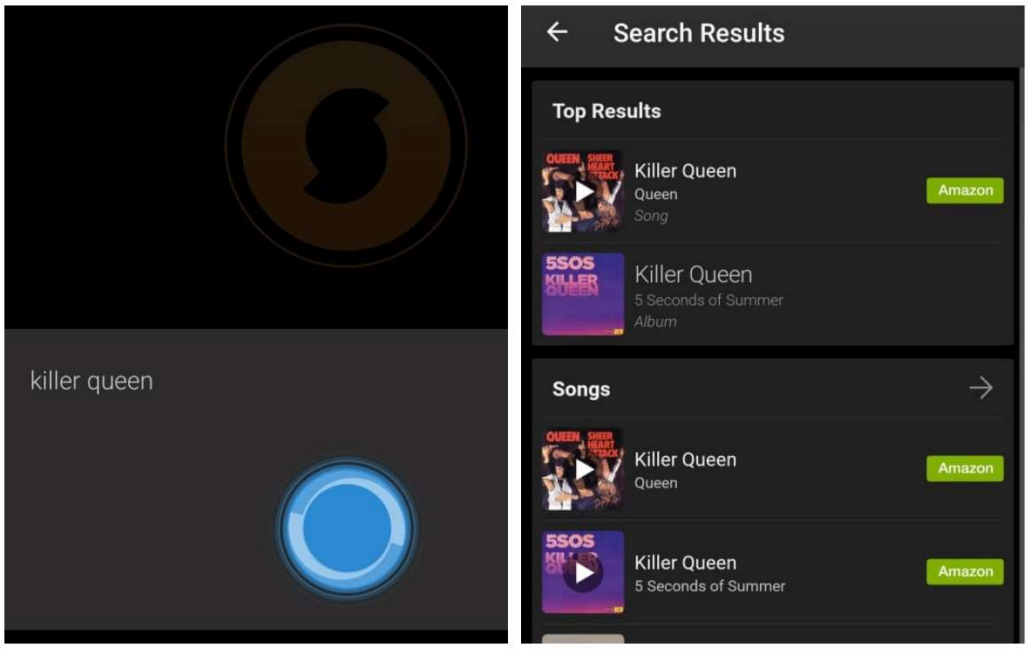
Εικόνα 43. Αναζήτηση με βάση το κείμενο στο Soundhound.

Το Soundhound χρειάζεται μόνο λίγα δευτερόλεπτα για να εμφανίσει αποτελέσματα με ζωντανούς στίχους (live lyrics) να συνοδεύουν παράλληλα το τραγούδι. Ο χρήστης μπορεί να βάλει σε σελιδοδείκτη τους αγαπημένους καλλιτέχνες και τραγούδια του.



Εικόνα 44. Παράδειγμα χρήσης Soundhound και εμφάνιση των Ζωντανών στίχων (Live lyrics).

Η δυνατότητα της αντιστοίχισης ηχητικών δεδομένων λειτουργεί και όταν πατάμε το κουμπί “search” για τις απλές φωνητικές αναζητήσεις. Χρησιμοποιεί μια συμπαγή και ευέλικτη αναπαράσταση της φωνής σας, που μπορεί να την αντιστοιχήσει άμεσα με ονόματα των τραγουδιών και καλλιτεχνών. Αυτή η δυνατότητα αντιστοίχισης περιλαμβάνεται επίσης στην αναζήτηση κειμένου του SoundHound, όπου η ικανότητά του να κατανοεί την προφορά, του επιτρέπει να δίνει σωστά αποτελέσματα για ανορθόγραφες αναζητήσεις.



Εικόνα 45. Audio search παράδειγμα και εμφάνιση αποτελεσμάτων.

Το SoundHound έχει εισχωρήσει με μεγάλη επιτυχία στην καθημερινότητα αρκετών ανθρώπων βοηθώντας, τους στην αναζήτηση των τραγουδιών που ψάχνουν. Από το 2016,

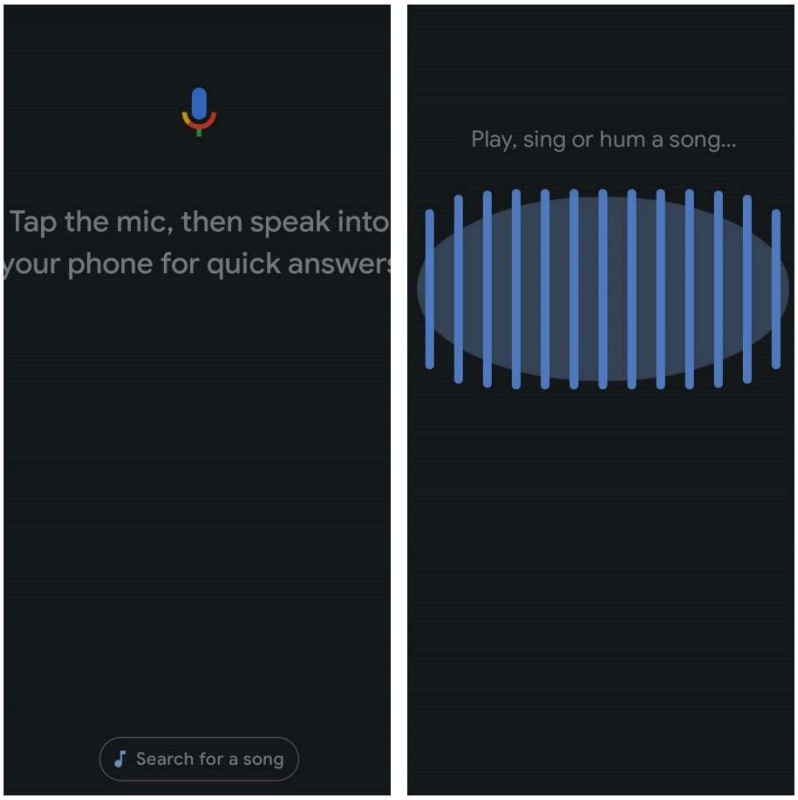
υπάρχουν πάνω από 300 εκατομμύρια χρήστες παγκοσμίως και με τον αριθμό μέχρι και σήμερα να αυξάνεται.

Παρουσιάζει έναν αρκετά αποτελεσματικό αλγόριθμο που προσπαθεί να προσαρμοστεί αρκετά σε κάθε αναζήτηση που επιδιώκει. Όμως, φαίνεται πως η μεγαλύτερη αδυναμία της εφαρμογής, σε σύγκριση με το Shazam, βρίσκεται στη βάση δεδομένων του, καθώς υπάρχουν αρκετές παραλείψεις από τραγούδια λιγότερο δημοφιλή. Η χρήση του search by humming δεν είναι τόσο εξυπηρετική καθώς ο αλγόριθμος περιορίζεται αρκετά στην φυσική ομιλία, αλλά είναι μια αρκετά ενδιαφέρουσα προσθήκη που με συνεχόμενες βελτιώσεις ίσως αποτελέσει ένα αρκετά χρήσιμο και επαναστατικό εργαλείο για την συγκεκριμένη εφαρμογή.

Google Hum to search

Από το 2020, η Google κυκλοφόρησε μία νέα λειτουργία στην αναζήτηση. Τώρα η μηχανή αναζήτησης μπορεί να φέρει αποτελέσματα μέσω humming, δηλαδή τραγουδώντας ένα μέρος από το μουσικό κομμάτι.

Η διαδικασία γίνεται πολύ απλά. Το μόνο που χρειάζεται είναι το πάτημα του κουμπιού “Search a song” δίπλα στο εικονίδιο του μικροφώνου που εμφανίζεται στην μπάρα αναζήτησης ή με την εκφώνηση της εντολής “what's this song?” στην ηχητική αναζήτηση. Προσφέρει άμεση ανταπόκριση και μπορεί να επιφέρει αποτελέσματα σε 10-15 δευτερόλεπτα.



Εικόνα 46.Google hum to search.

Λειτουργία

Η Google έχει στηριχθεί στην μέθοδο αντιστοίχισης του audio fingerprinting, χαρακτηρίζοντας την μελωδία ενός τραγουδιού σαν ένα "δακτυλικό αποτύπωμα" (fingerprint), οπού μέσα από μοντέλα machine learning μπορεί να αντιστοιχίσει το φωνητικό απόσπασμα με το κατάλληλο "δακτυλικό αποτύπωμα".

Η τεχνική και ο τρόπος αντιστοιχίας είναι παρόμοιος και με τις επακόλουθες εφαρμογές (Shazam, Soundhound). Τα μοντέλα μηχανικής μάθησης μετατρέπουν τον ήχο σε μια αριθμητική ακολουθία που αντιπροσωπεύει τη μελωδία του τραγουδιού. Συγκρίνει τις ακολουθίες αυτές με χιλιάδες τραγούδια από όλο τον κόσμο και εντοπίζει πιθανές αντιστοιχίες σε πραγματικό χρόνο. Η κύρια διαφορά όμως βρίσκεται στην διαδικασία συλλογής, καθώς τα μοντέλα αυτά (machine learning) έχουν εκπαιδευτεί να αναγνωρίζουν τραγούδια με βάση μια ποικιλία πηγών, συμπεριλαμβανομένων των ανθρώπων που τραγουδούν, σφυρίζουν ή σιγοτραγουδούν, καθώς και ηχογραφήσεων που γίνονται σε στούντιο. Οι αλγόριθμοι αφαιρούν όλες τις άλλες λεπτομέρειες, όπως τα συνοδευτικά όργανα και την χροιά μαζί με τον τόνο της φωνής, αφήνοντας μόνο να μείνει η αριθμητική ακολουθία του τραγουδιού, ή αλλιώς το δακτυλικό αποτύπωμα.

Ο πολλά υποσχόμενος αυτός αλγόριθμος δεν σταμάτησε εκεί καθώς η Google συνεχίζει να επεκτείνει τις δυνατότητες του χάρη στα νευρωνικά δίκτυα (deep neural networks). Αυτό έχει ως αποτέλεσμα αυτή τη στιγμή να μπορεί να αναγνωρίσει τραγούδια χωρίς τους στίχους ή το αρχικό τραγούδι, άλλα μονάχα με το βουητό.

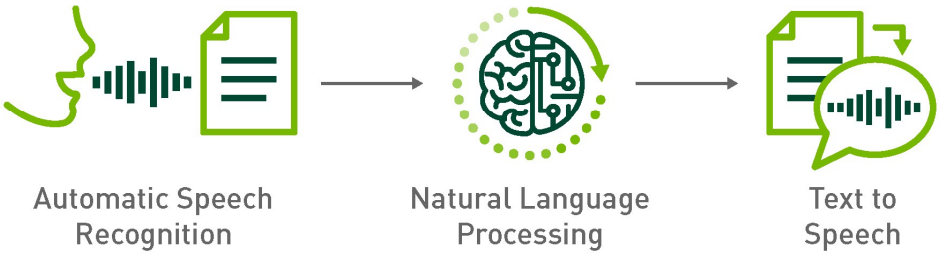
Το hum to search της google είναι ένα αρκετά ευρηματικό μέσον μουσικής ανάκτησης. Τα δυνατά του σημεία φαίνονται στον αρκετά ανεπτυγμένο αλγόριθμο του και στην ευρεία βάση δεδομένων του. Έχει ουσιαστικά δανειστεί τα δυνατότερα σημεία του Shazam και του Soundhound, αλλά βαδίζει ολοφάνερα στο δικό του ορίζοντα, προβλέπεται μίαν αρκετά ανοδική πορεία για το hum to search.

Υπάρχουν δεκάδες εφαρμογές που θα μπορούσαν να αναφερθούν, όμως η αποτελεσματικότητα του αλγορίθμου και η βάση δεδομένων τραγουδιών που είναι καταγεγραμμένα στο Shazam, η πρωτοπορία του Soundhound και το ανερχόμενο νέο μέσον αναζήτησης της Google μέσα από το hum to search δεν υπάρχουν στο ίδιο βαθμό αλλού.

Κεφάλαιο 5. Αναγνώριση ομιλίας (Speech recognition)

Η αναγνώριση ομιλίας, γνωστή και ως automatic speech recognition, computer speech recognition και speech to text, έχει ως σκοπό την αποτελεσματική επικοινωνία μεταξύ του ανθρώπου και της μηχανής.

Η αναγνώριση ομιλίας είναι μια εξελισσόμενη τεχνολογία που παρέχει την δυνατότητα σε μια μηχανή ή ένα πρόγραμμα να αναγνωρίζει λέξεις που ακούγονται από την ομιλία του ανθρώπου και να τις μετατρέπει σε αναγνώσιμο κείμενο. Επίσης, εκτός της επικοινωνίας με τον υπολογιστή, το πεδίο αυτό περιέχει και την αναγνώριση της φωνής (voice recognition), ή αλλιώς voice identification (Willyard, 2010). Η διαφορά σε αυτή την τεχνολογία, είναι πως ο υπολογιστής δεν έχει σκοπό να επικοινωνήσει με τον χρήστη, αλλά να αναγνωρίσει την φωνή και να την συγκρίνει και με μία άλλη. Η φωνητική αναγνώριση χρησιμοποιείται κυρίως σε συστήματα ασφαλείας.



Εικόνα 47. Η διαδικασία του ASR. https://developer.nvidia.com/blog/how-to-build-domain-specific-automatic-speech-recognition-models-on-gpus/

Η έρευνα πάνω στην **αναγνώριση της ομιλίας** έχει διάρκεια πάνω από 5 δεκαετίες, όμως έπρεπε να περάσουν αρκετά χρόνια ώστε να έχει κάποιο αποτέλεσμα αφού η υπολογιστική δύναμη τότε δεν είχε φτάσει στο επιθυμητό σημείο (Yu & Deng, 2016).

Οι πιο σημαντικοί παράγοντες που βοήθησαν στην πρόσφατη ανάπτυξη των συστημάτων αναγνώρισης ομιλίας είναι (Yu & Deng, 2016):

- Ο " νόμος του Moore ". Βασίζεται στην εμπειρική παρατήρηση από τον Gordon Moore που αναφέρει ότι "ο αριθμός των τρανζίστορ σε ένα μικροτσίπ (όπως ονομάζονταν το 1965) θα διπλασιαζόταν περίπου κάθε χρόνο."
- Η συνεχόμενη αύξηση της υπολογιστικής δύναμης.
- Η ανάπτυξη των νευρωνικών μοντέλων.

Η ανάπτυξη αυτή της τεχνολογίας δημιούργησε μεθόδους που βοήθησαν στην ελαχιστοποίηση των σφαλμάτων στα συστήματα φωνητικής αναγνώρισης. Η κυκλοφορία των προσωπικών υπολογιστών και των λειτουργικών συστημάτων επέτρεπαν την εγκατάσταση προγραμμάτων με την χρήση μία δισκέτας ή ενός CD, επέτρεψαν στους προγραμματιστές να σχεδιάσουν το δικό τους πρόγραμμα και να μοιραστούν τις δυνατότητές τους σε άλλους υπολογιστές. Έτσι, πολλές εταιρείες δημιούργησαν προγράμματα για διάφορες ανάγκες. Η πρώτη σημαντική προσπάθεια για την προσβάσιμη χρήση της τεχνολογίας ανάκτησης στον υπολογιστή, έγινε το 1996 από την IBM, με την κυκλοφορία του VoiceType Simply Speaking. Επιστήμες όπως η γλωσσολογία, σε συνδυασμό με την πληροφορική συνέβαλλαν για την περαιτέρω ανάπτυξη των συστημάτων αυτών.

Η λειτουργικότητα των συστημάτων δεν περιορίζεται στην αποστολή εντολών και στην εξυπηρέτηση πελατών από ένα αυτοματοποιημένο σύστημα. Έχει την προοπτική να γίνει ένα χρήσιμο εργαλείο στην επικοινωνία, από τα ψηφιακά μέσα μεταξύ των ανθρώπων, αλλά και μεταξύ ανθρώπων και μηχανών. Μπορεί να χρησιμοποιηθεί για την αποστολή μηνυμάτων ή email, την ζωντανή μετάφραση και την απάντηση διάφορων ερωτημάτων, διότι σε σύγκριση με τα συμβατικά μέσα επικοινωνίας των ηλεκτρονικών συσκευών (ποντίκι, πληκτρολόγιο, οθόνη αφής) η επικοινωνία με την χρήση της φωνής είναι πιο ανθρώπινη και γρήγορη.

Τα συστήματα αναγνώρισης ομιλίας χρησιμοποιούν αλγόριθμους για την επεξεργασία και ερμηνεία προφορικών λέξεων και τη μετατροπή τους σε κείμενο. Η διαδικασία αυτή ξεκινάει με την εισαγωγή και έπειτα ανάλυση του ήχου, με την αποκοπή του σε κομμάτια, την ψηφιοποίηση όλων των κομματιών σε αναγνώσιμη από τον υπολογιστή μορφή (computer-machine readable) και εν τέλη στην χρήση ενός αλγόριθμου για την αντιστοίχιση του με την καταλληλότερη αναπαράσταση κειμένου (Fu, 2020).

Ένα απλό λογισμικό αναγνώρισης ομιλίας έχει περιορισμένο λεξιλόγιο και μπορεί να αναγνωρίζει λέξεις και φράσεις μόνο όταν εκφωνούνται με σαφήνεια. Ακόμα και τώρα, οι πιο εξελιγμένες τεχνολογίες αναγνώρισης ομιλίας με πολύπλοκο λογισμικό μπορούν να χειριστούν τον φυσικό λόγο, διαφορετικές προφορές και ποικίλες γλώσσες με μεγαλύτερη επιτυχία, αλλά όχι σε ικανοποιητικό επίπεδο.

Το λογισμικό αναγνώρισης ομιλίας πρέπει να προσαρμόζεται στην ιδιαίτερα μεταβλητή και εξειδικευμένη φύση της ανθρώπινης ομιλίας. Οι αλγόριθμοι που επεξεργάζονται και οργανώνουν τον ήχο σε κείμενο μέσα από datasets εκπαιδεύονται σε διαφορετικά μοτίβα ομιλίας, διαφορετικές γλώσσες, διαλέκτους, προφορές και φράσεις, μαθαίνοντας να διαχωρίζει επίσης στον ήχο τον προφορικό λόγο από τον περιβάλλοντα θόρυβο που μπορεί να επικρατεί.

Οι εφαρμογές ή υπηρεσίες που χρησιμοποιούν την τεχνολογία της αναγνώρισης ομιλίας είναι παντού στην καθημερινότητά μας. Μπορεί να είναι η δυνατότητα του κινητού τηλέφωνου να ανταποκρίνεται σε φωνητικές εντολές, ενός αυτοκίνητο να ακούει την διεύθυνση από τον οδηγό και να την επιλέγει στο GPS, ή να είναι η αυτόματη εξυπηρέτηση σε τηλεφωνικό κέντρο. Κάποια από τα παραπάνω είναι παραδείγματα ενός spoken languange system.

Οι δυνατότητες όμως των συστημάτων αυτών είναι περιορισμένες, καθώς συνήθως μπορούν να απαντήσουν μονάχα σε μία ή δύο διαφορετικές εργασίες (π.χ. εύρεση μιας διεύθυνσης ή ενός τμήματος τεχνικής υποστήριξης). Σήμερα, η πρόοδος στην τεχνολογία αναγνώρισης ομιλίας είναι πιο εμφανής στους γνωστούς οικιακούς βοηθούς, όπως είναι η Alexa της Amazon, το Google Now της Google, η Siri της Apple κ.α., συσκευές οι οποίες είναι ικανές να αναγνωρίσουν και να ανταποκριθούν σε πιο σύνθετες εντολές. Στόχος για το μέλλον είναι ανεξάρτητα με την διαφορά στην ομιλία του κάθε ανθρώπου, να υπάρχουν συστήματα που να μπορούν να εκτελούν εντολές με μεγάλη πολυπλοκότητα και με απόλυτη ακρίβεια, χωρίς να ανησυχεί ο χρήστης για οποιαδήποτε αστοχία σε αυτό που είπε.

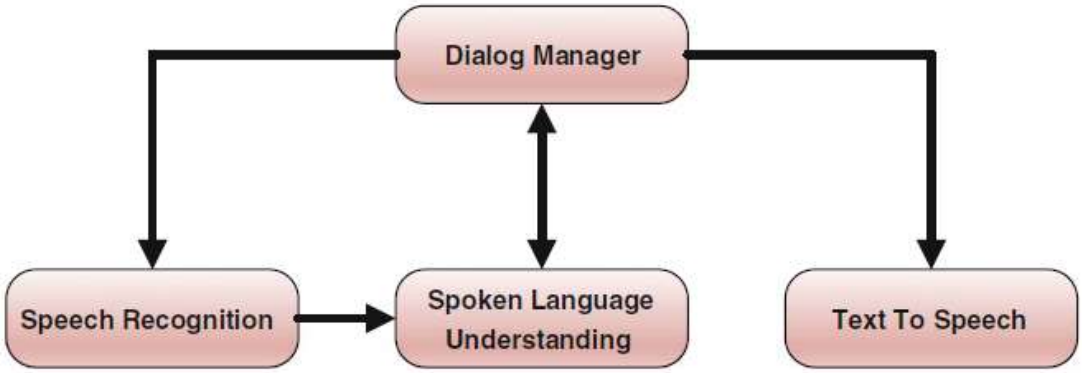
5.1 Δομή ενός συστήματος αναγνώρισης ομιλίας

Η αναγνώριση της ομιλίας είναι μέρος ενός ευρύτερου συστήματος, γνωστό και ως σύστημα ομιλούμενης γλώσσας (Spoken language system). Το σύστημα αυτό αποτελείται από τέσσερα μέρη (Yu & Deng, 2016):

- Αναγνώριση φωνής (Speech recognition component)
- Κατανόηση της ομιλούμενης γλώσσας (Spoken language understanding component)

- Κείμενο σε ομιλία (Text-to-speech component)
- Διαχειριστής Διαλόγου (Dialog Manager)

Το Speech recognition component μετατρέπει την ομιλία σε κείμενο, το Spoken language understanding component ανακτά σημασιολογική πληροφορία στις ομιλούμενες λέξεις, το Text-to-speech component εκφράζει την ομιλούμενη πληροφορία και το Dialog Manager είναι μεσολαβητής ανάμεσα στην εφαρμογή και στα υπόλοιπα τρία μέρη (Yu & Deng, 2016).



Εικόνα 48. Σύστημα ανάκτησης ήχου (Yu & Deng, 2016).

Τα συστήματα αναγνώρισης ομιλίας (Speech Recognition Systems), στην πλειοψηφία τους, χρησιμοποιούν 4 μέρη.

1. **Signal processing and feature extraction**
2. **Acoustic model**
3. **Language model**
4. **Hypothesis search**

5.1.1 Signal processing and feature extraction

Εδώ συμβαίνει η εισροή των δεδομένων. Δέχεται τον ήχο και τον επεξεργάζεται ώστε να δοθεί έμφαση στην ομιλία. Για να επιτευχθεί αυτό, αφαιρούνται οι θόρυβοι και οι οποιεσδήποτε διαστρεβλώσεις στον ήχο, μετατρέπεται το σήμα ώστε να υπολογίζεται με την συχνότητα και όχι με τον ήχο (μετασχηματισμός Fourier) και τέλος εφαρμόζει τα πιο σημαντικά διανύσματα για την εξαγωγή ηχητικών δεδομένων. (feature extraction)

5.1.2 Acoustic model

Τα επεξεργασμένα χαρακτηριστικά του ήχου που συλλέχτηκαν από το signal processing και την εξαγωγή των χαρακτηριστικών (feature extraction), δηλαδή η ίδια η ομιλία, μεταφέρεται στο acoustic model. Το μοντέλο αυτό προσπαθεί να καταλάβει τις συλλαβές που "ακούει" ξεχωριστά και να τις καταγράψει ως αποτέλεσμα. Η χρησιμότητά του είναι ότι στην ανάπτυξη ενός συστήματος μίας γλώσσας, μπορεί να αναγνωρίσει διαφορετικές προφορές.

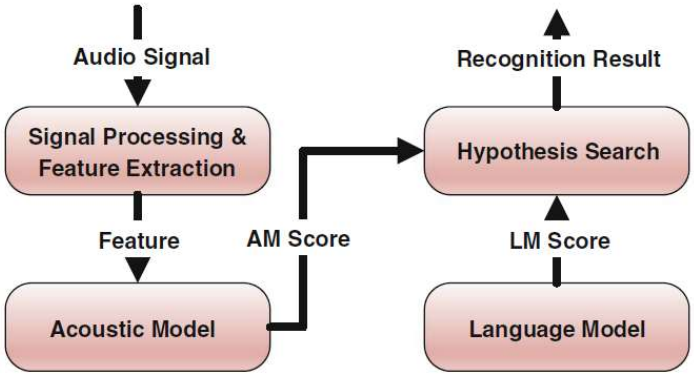
5.1.3 Language model

Το language model προβλέπει τις υποθετικές λέξεις που σχηματίζονται από τον ήχο και καταγράφει το πιο πιθανό αποτέλεσμα. Το αποτέλεσμα αυτό είναι πιο ακριβείς αν οι λέξεις αυτές είναι προκαθορισμένες μέσα στο σύστημα.

Χωρίζεται σε 2 τύπους, το στατικό και το δυναμικό μοντέλο. Το στατικό μοντέλο περιλαμβάνει συγκεκριμένες λέξεις μέσα στο λεξιλόγιο της βάσης δεδομένων και έτσι περιορίζονται οι επιλογές και τα λάθη. Είναι εύκολο στην εφαρμογή, όμως δεν μπορεί να επεκταθεί η ανάκτηση πέρα από αυτής που έχει προγραμματιστεί. Για αυτό υπάρχει το δυναμικό μοντέλο, το οποίο μπορεί να μάθει και να προσαρμόσει τις νέες λέξεις στην βάση δεδομένων. Το μειονέκτημα του δυναμικού μοντέλου είναι οι υψηλές απαιτήσεις που χρειάζεται για τους μαθηματικούς υπολογισμούς και γενικά για την λειτουργιά του. Η ακρίβεια των δύο μοντέλων είναι παρόμοια, και η επιλογή του κάθε τύπου βασίζεται στην κρίση του υπεύθυνου της ανάπτυξης ενός συστήματος. (Yu & Deng, 2015).

5.1.4 Hypothesis search

Εδώ συνδυάζονται τα αποτελέσματα των acoustic model και language model και οι συνδυασμοί λέξεων που φαίνονται πιο πιθανοί στο σύστημα, επιλέγονται.

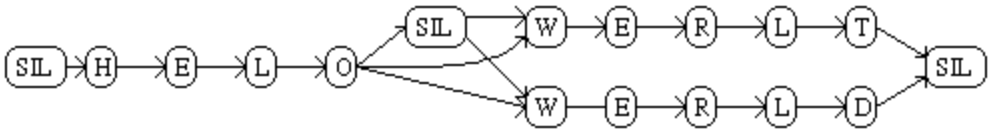


Εικόνα 49. Αυτόματη αναγνώριση ομιλίας (Yu & Deng, 2016).

Το κάθε σύστημα αναγνώρισης ομιλίας μπορεί να χρησιμοποιεί σε κάθε τμήμα του, διαφορετικά μοντέλα και τεχνικές με σκοπό να μειώσει τα περιθώρια λάθους. Για παράδειγμα, μπορεί να αλλάξει η ομιλούμενη γλώσσα, αλλά η εξαγωγή των δεδομένων του ήχου γίνεται με τον ίδιο τρόπο. Η επιλογή των κατάλληλων αυτών τεχνικών, θα κάνουν το σύστημα πιο αποτελεσματικό από τα άλλα.

5.2 Οι δυσκολίες της αναγνώρισης της ομιλίας

Οι χρήστες των αυτόματων συστημάτων αναγνώρισης ομιλίας, έχουν την απαίτηση το σύστημά να μπορεί να αναγνωρίζει την φωνή τους και να απαντούν στα διάφορα ερωτήματά τους ή εντολές με μεγάλη ευστοχία. Για να το καταφέρουν αυτό οι δημιουργοί των συστημάτων, δημιουργούν μία βάση δεδομένων με όλες τις πιθανές λέξεις που θα είναι χρήσιμες στα ερωτήματα των χρηστών. Όμως, μπορεί σε κάθε λέξη να υπάρχουν διαφορετικές προφορές σε κάθε συλλαβή. Το σύστημα σε αυτήν την περίπτωση μπορεί να προβλέψει την διαφορά στην προφορά της λέξης ακόμα και αν η λέξη που αναγνωρίζει στην αρχή δεν ταιριάζει. Για παράδειγμα, ένας Ιταλός αναζητά πληροφορίες για το Λούβρο και στέλνει το ερώτημα στην μηχανή αναζήτησης. Η προφορά του Λούβρου είναι διαφορετική με αυτή του Γάλλου, αλλά όχι η λέξη. Έτσι, η βάση δεδομένων θα πρέπει να περιέχει εναλλακτικές λέξεις ή συλλαβές για την ίδια λέξη (Willyard, 2010).



Εικόνα 50. Σενάριο διαφορετικών προφορών της φράσης "hello world".

http://i13pc106.ira.uka.de/~mthoma/asr/disc-thread/score

5.3 Κατηγορίες ASR

Υπάρχουν διάφορες κατηγορίες συστημάτων, οι οποίες χωρίζονται με πολλά κριτήρια, όπως τον τύπο της ομιλίας, την εισαγωγή των δεδομένων κτλ.

Η αναγνώριση της ομιλίας έχει 2 κατηγορίες, αυτήν **βασισμένη στον ομιλητή** (speaker dependent) και **την μη βασισμένη στον ομιλητή** (speaker independent) (Willyard, 2010).

Τα συστήματα βασισμένα στον ομιλητή εκπαιδεύονται από αυτόν που θα χρησιμοποιήσει το σύστημα. Το σύστημα έχει υψηλή ακρίβεια στην αναγνώριση των λέξεων, όμως το

μειονέκτημα αυτής την κατηγορίας είναι η χρησιμότητα του, μοναχά από το άτομο που εκπαίδευσε το σύστημα.

Τα συστήματα που δεν βασίζονται στον ομιλητή εκπαιδεύονται πάνω στην λέξη ανεξάρτητα από το ποιος την λέει. Αυτό μεγαλώνει την δυσκολία στην ανάπτυξη του συστήματος, λόγω των ποικίλων διαφορών που υπάρχουν στην ομιλία του κάθε ανθρώπου.

5.3.1 Isolated word speech recognition (IWR)

Το σύστημα αυτό αναγνωρίζει μία λέξη μέσα από την βάση δεδομένων του συστήματος. Η πρώτες προσπάθειες πάνω σε αυτό ξεκινούν από την δεκαετία του 1970 και με την τεχνολογική άνοδο του 2000, η τεχνολογία της αναγνώρισης ομιλίας έγινε πιο προσιτή. Στην συνέχεια, τα συστήματα αυτά κατάφεραν να αναγνωρίσουν τις ίδιες λέξεις με διαφορετικά φωνήματα. Αυτή η κατηγορία χρησιμοποιείται για συστήματα που δεν απαιτούν σύνθετες εντολές και αρκεί η μία λέξη. Το λεξιλόγιο στην βάση δεδομένων μπορεί να αποτελείται από μερικές δεκάδες μονάχα λέξεις.

5.3.2 Connected word recognition (CWR)

Εδώ, μπορεί να γίνει ο διαχωρισμός τον λέξεων με την χρήση των παύσεων. Το σύστημα αντιλαμβάνεται τις παύσεις στην ομιλία και δημιουργεί κενά, με αποτέλεσμα να καταλάβει μία σειρά από λέξεις. Είναι πιο αποτελεσματικό και η πολυπλοκότητα και το πλήθος των λέξεων αυξάνεται.

5.3.3 Continuous speech recognition (CSR)

Οι λέξεις δεν χωρίζονται από παύσεις, αλλά είναι ενωμένες προτάσεις χωρίς ενδιάμεσα κενά. Σε αυτήν την περίπτωση έχουν δημιουργηθεί υποκατηγορίες, λόγω των μοναδικών τεχνικών που έχουν αναπτυχθεί για το CSR. Μια πιο σύγχρονη κατηγορία του CWR είναι το large vocabulary continuous speech recognition (LVCSR). Αν και δεν είναι τέλειο, έχει καταφέρει να εφαρμοστεί σε πολλές υπηρεσίες με μεγάλη επιτυχία (Saon & Chien, 2012).

5.3.4 Spontaneous speech recognition

Το πρόβλημα που υπάρχει με τις παραπάνω κατηγορίες αναγνώρισης ομιλίας, είναι πως στην καθημερινή ομιλία δεν υπάρχει η τέλεια συνέχεια των λέξεων. Μπορεί να υπάρξουν μεγάλες παύσεις κατά την διάρκειά της, ή να βρεθούν λέξεις που να μην αντιστοιχούν σε κάποιο λεξιλόγιο. Η κατηγορία αυτή έχει σκοπό να ασχοληθεί με την κανονική ομιλία και να αντιμετωπίσει τον "θόρυβο".

5.4 Εξαγωγή των χαρακτηριστικών της ομιλίας

Τα δεδομένα που έχουν συλλεχθεί πρέπει να είναι τα κατάλληλα ώστε η αναγνώριση των συλλαβών και στην συνέχεια των λέξεων να γίνει σωστά. Πολλές τεχνικές έχουν εφευρεθεί, με αρκετά πλεονεκτήματα και μειονεκτήματα από την κάθε μία. Η εξαγωγή χαρακτηριστικών βασίζεται στην εύρεση πιθανών τιμών μέσα από τυχαίες μεταβλητές πραγματικών τιμών. Έχει την δυνατότητα να συλλέγει διακεκομμένες ή συνεχόμενες τιμές ή συνδυασμό αυτών. Η μέθοδος αυτή είναι γνωστή και ως κανονική κατανομή και το μοντέλο ονομάζεται Gaussian mixture model (GMM). Η ονομασία του προέρχεται από τον γερμανό μαθηματικό Carl Friedrich Gauss, όπου η δημιουργία των μαθηματικών και στατιστικών θεωριών του χρησιμοποιούνται ακόμα και σήμερα. Το μοντέλο έχει συμβάλει στην αποτελεσματικότητα των αυτόματων συστημάτων αναγνώρισης ομιλίας, ιδιαίτερα της αναγνώρισης ομιλίας και την αποθορυβοποίηση, δηλαδή την παράκαμψη της μη χρήσιμης πληροφορίας. Το μεγάλο μειονέκτημα να είναι ότι η εξαγωγή των δεδομένων γίνεται με στατικό τρόπο, δηλαδή οι τυχαίες μεταβλητές έχουν συγκεκριμένες τιμές σε κάθε υπολογισμό. Με την έλευση του deep learning, το GMM μοντέλο έχει συγκριτικά μικρότερη ακρίβεια και φαίνεται να χάνει έδαφος ως προς την χρησιμότητα του. (Yu & Deng, 2016).

5.5 Η εφαρμογή της αναγνώρισης ομιλίας στην καθημερινότητα

Παρακάτω θα μιλήσουμε πιο αναλυτικά για μερικά από τα έργα που σκοπό έχουν να αναπτύξουν την ποιότητα της αναγνώριση ομιλίας.

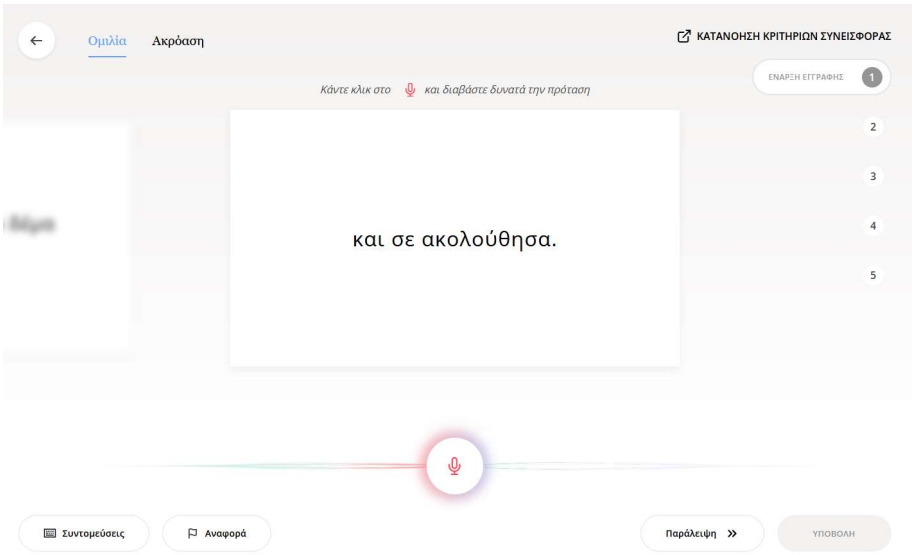
5.5.1 Common Voice

Το Common Voice είναι μη κερδοσκοπικό έργο της Mozilla και σκοπό έχει να αναπτύξει μία βάση δεδομένων, με την βοήθεια εθελοντών, που να εξυπηρετεί όσες περισσότερες γλώσσες είναι δυνατόν. Κυκλοφόρησε το 2017 και από τότε, περιέχει 87 διαφορετικές γλώσσες, συμπεριλαμβανομένου και των ελληνικών, πάνω από 18.000 ώρες ηχογραφημένου υλικού, με ανθρώπους από κάθε ηλικία και φύλο, με τις 14.000 να είναι επαληθευμένες στην βάση δεδομένων. Εκτός από την σωστή αναγνώριση της ομιλίας σε οποιαδήποτε διαθέσιμη γλώσσα, το Common Voice μπορεί να γίνει και ένα χρήσιμο εργαλείο για την μετάφραση μεταξύ των γλωσσών.

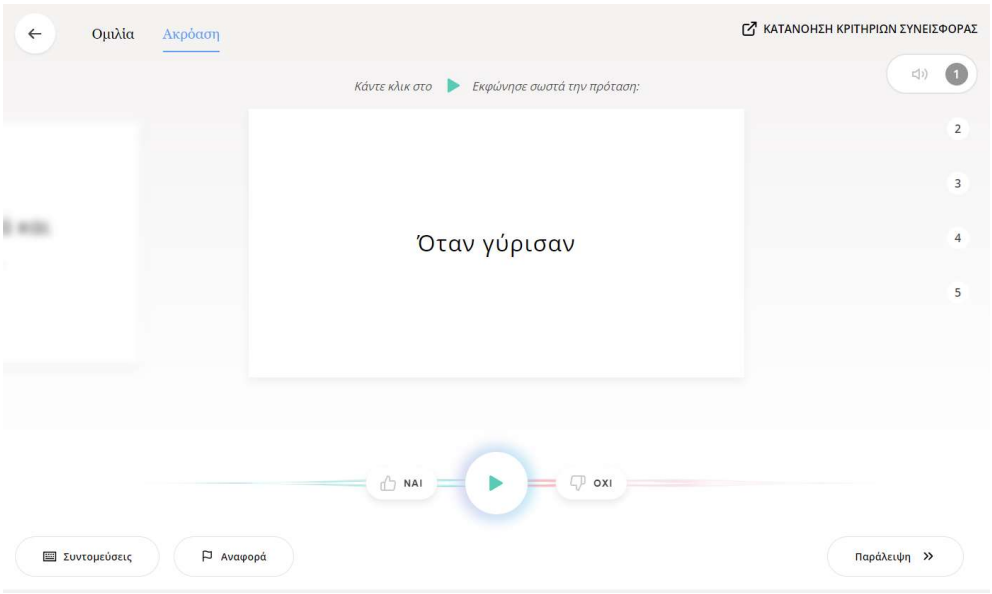
Είναι ένα χρήσιμο εργαλείο για όσους επιθυμούν να δουλέψουν πάνω στην τεχνολογία της ανάκτησης της ομιλίας, χρησιμοποιώντας ένα σύνολο dataset που διαρκώς αναπτύσσεται χάρη στην βοήθεια των εθελοντών.

Εικόνα 51. Το Common Voice. https://commons.wikimedia.org/w/index.php?curid=69338756

Ο τρόπος με τον οποίο μπορεί κάποιος να συνεισφέρει είναι με δύο τρόπους. Ο πρώτος είναι με την ηχογράφηση μίας σειράς από προτάσεις οι οποίες το Common Voice εμφανίζει στην οθόνη για να εκφωνήσει ο εθελοντής,. Στην συνέχεια, το Common Voice αποθηκεύει την ηχογράφηση στην βάση δεδομένων, όμως χωρίς να απαιτείται η αξιολόγηση της ηχογράφησης.



Ο δεύτερος τρόπος που κάποιος μπορεί να συνεισφέρει, είναι με την αξιολόγηση της κάθε ηχογράφησης άλλων εθελοντών. Πιο αναλυτικά, παρατηρώντας για αστοχίες στην πρόταση, όπως ο θόρυβος, η λάθος προφορά, η κακή ποιότητα κτλ., με σκοπό την απόρριψη αυτών των προτάσεων από την βάση δεδομένων.



Εκτός από την συνεισφορά στο έργο αυτό, όποιος θέλει μπορεί να επωφεληθεί από την δουλειά των εθελοντών. Το μόνο που έχει να κάνει είναι να κατεβάσει την βάση δεδομένων σε κάθε γλώσσα μέσα στην λίστα, με όλα τα επαληθευμένα ηχητικά αποσπάσματα σε μορφή MP3. Ανάλογα με την γλώσσα, η βάση ανανεώνεται αρκετά συχνά.

5.5.2 Watson Speech to text

Είναι μία από τις εφτά υπολογιστικές υπηρεσίες Watson Services της IBM. Στόχο έχει την ενσωμάτωση της τεχνολογίας αναγνώρισης ομιλίας, στις εφαρμογές των χρηστών. Η πρώτη έκδοση κυκλοφόρησε το 2015, μαζί με το Text to Speech το οποίο δουλεύει με παρόμοιο τρόπο, και από τότε έχει εφαρμοστεί σε χιλιάδες προϊόντα. Με την χρήση των machine learning και deep learning τεχνολογιών, το Watson™ Speech to Text έχει αυξήσει σημαντικά την ακρίβεια στην αναγνώριση ομιλίας. Υποστηρίζει την συνεχόμενη ομίλα (Continuous speech) (Alibegovic, Prljaca, Kimmel & Schultalbers, 2020). Οι υποστηριζόμενες γλώσσες είναι 14 (χωρίς να περιλαμβάνονται τα Ελληνικά), με ένα μέρος από αυτές να είναι ακόμα σε στάδιο ανάπτυξης.

Η υπηρεσία προσφέρεται δωρεάν, όμως με λίγες λειτουργίες να διαθέσιμες. Μέσω συνδρομής μπορούν να αποκτηθούν και άλλες δυνατότητες, όπως είναι η σχεδίαση ενός περιβάλλοντος επικοινωνίας με τον χρήστη, την προσθήκη ενός αυτοματοποιημένου ψηφιακού βοηθού κτλ. Σε αντίθεση με το Common Voice, η προσθήκη νέων δεδομένων στην βάση δεδομένων γίνεται επί πληρωμής.

Το Watson Speech to Text μπορεί να δεχτεί τα δεδομένα του ήχου και να τα μετατρέψει σε κείμενο. Η χρήση μικροφώνου δεν είναι απαραίτητη, αφού η υπηρεσία μπορεί να κάνει την αναγνώριση της ομιλίας και μέσα από διάφορα είδη αρχείων. Μέσα από ένα περιβάλλον

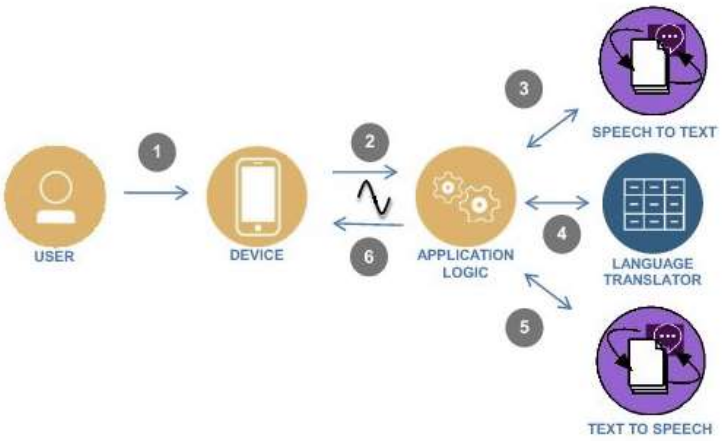
επαφής ή απευθείας από τον κώδικα, ο χρήστης μπορεί να επεξεργαστεί τις εντολές που θα δέχεται το σύστημα.

Εικόνα 52. Μία σειρά από βήματα για την εκτέλεση μιας εντολής.

Οι πρόσθετες δυνατότητες που προκύπτουν από την υπηρεσία είναι:

- Personality Analyzer: Με την προσθήκη του Watson Personality Insights, το σύστημα δημιουργεί ένα προφίλ προσωπικότητας για διαφορετικά πρόσωπα. Συλλέγει την ομιλία του χρήστη και με την χρήση αλγόριθμων εμφανίζει τα ποσοστά του κάθε χαρακτηριστικού προσωπικότητας.
- Real-time transcription: Με την προσθήκη του Watson Speech to Text και Watson Language translator, γίνεται δυνατή η μετάφραση και εκφώνηση του κειμένου σε πραγματικό χρόνο.

Είναι ανάμεσα στα συστήματα με την μεγαλύτερη ακρίβεια και θεωρείται ως μία καλή εναλλακτική του Google talk (Filippidou & Moussiades, 2020).



Εικόνα 53. Δομή του Real-time transcription.

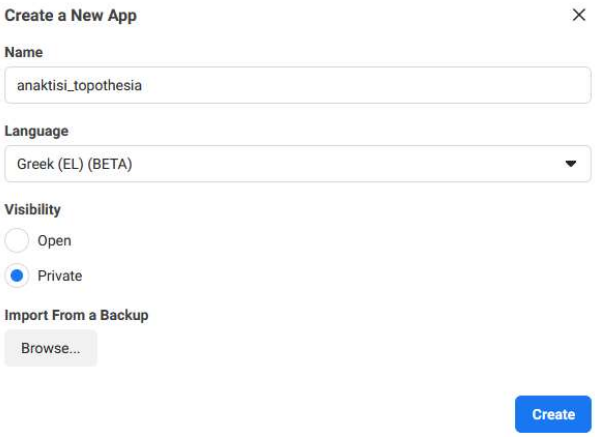
5.5.3 Wit.Ai

Ιδρύθηκε το 2013 από τους Alex Lebrun, Laurent Landowski και Willy Blandin ως μία πλατφόρμα που θα έδινε την δυνατότητα σε προγραμματιστές να δημιουργήσουν εφαρμογές που να χρησιμοποιούν την αναγνώριση της ομιλίας. Το 2015 εξαγοράστηκε από

το Facebook. Είναι ελεύθερη στη χρήση και έχει χιλιάδες ενεργούς χρήστες. Υποστηρίζει πάνω από 130 γλώσσες και μπορεί να προγραμματιστεί με την Python, την Node.js, την Ruby, την Unity και την Go. Είναι απλό στην χρήση και αρκετά γρήγορο στην εκπαίδευση του συστήματος. Το μόνο που χρειάζεται για να χρησιμοποιήσει κάποιος την πλατφόρμα είναι να συνδεθεί με το λογαριασμό του στο Facebook.

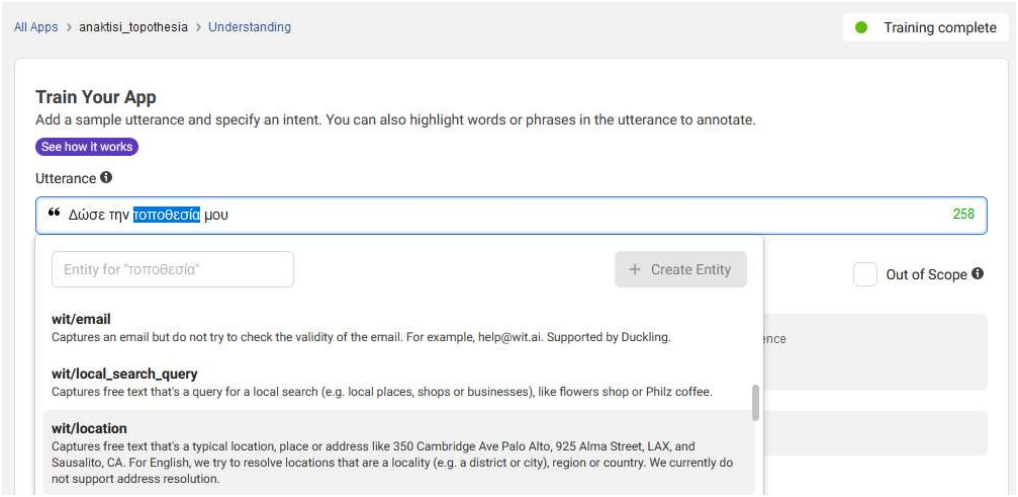
Παρακάτω θα δημιουργήσουμε ένα παράδειγμα μίας εντολής πάνω στην πλατφόρμα.

Σε αυτό το παράδειγμα θα χρησιμοποιήσουμε τα Ελληνικά, που είναι ακόμα υπό ανάπτυξη. Μέσα από την κεντρική ιστοσελίδα του Wit.ai, πρέπει πρώτα να δημιουργήσουμε μία εφαρμογή και να την ονομάσουμε. Στην συγκεκριμένη περίπτωση θα το ονομάσουμε anaktisi_topothesia



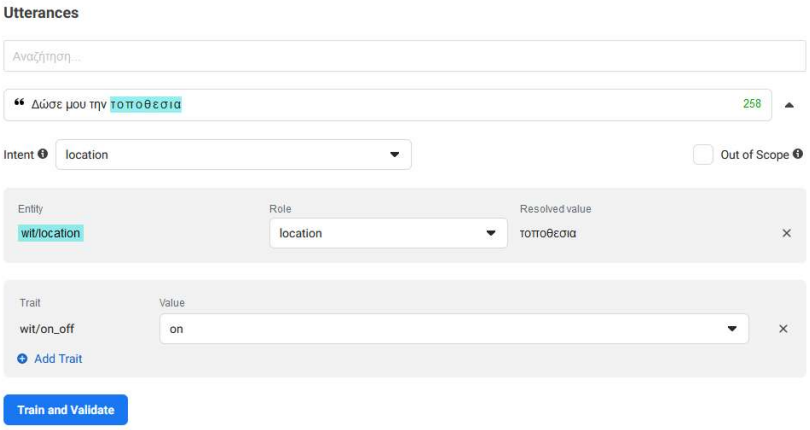
Εικόνα 54. Δημιουργία της εφαρμογής.

Τώρα μπορούμε να δουλέψουμε πάνω στην εφαρμογή και να προσθέσουμε την εντολή ή τις εντολές που θέλουμε. Ο μοναδικός σκοπός της εφαρμογής που θα δημιουργήσουμε είναι να εμφανίζει την τοποθεσία του χρήστη. Το επόμενο βήμα είναι να γράψουμε σε κείμενο την εντολή και σε οποιαδήποτε μορφή επιθυμούμε. Πρώτα γράφουμε την εντολή "Δώσε την τοποθεσία μου". Στην συνέχεια, σημειώνουμε την λέξη που θέλουμε να ορίσουμε ως οντότητα (entity) μέσα στο κείμενο που θα εκφωνήσει ο χρήστης (utterance). Ο σκοπός της σημείωσης αυτής είναι η πιο εύκολη κατανόηση του συστήματος και η μεγαλύτερη αποτελεσματικότητα στην ανάκτηση. Από εκεί, θα εμφανιστεί μία σειρά από επιλογές τύπους εννοιών.



Εικόνα 55. Επιλογή των εννοιών.

Μπορούμε να δημιουργήσουμε τους δικούς μας τύπους οντοτήτων, όμως δεν είναι απαραίτητο αφού υπάρχει το wit/location, μία οντότητα που επιστρέφει την διεύθυνση ή την τοποθεσία ενός μέρους. Τώρα μπορούμε να εκπαιδεύσουμε το σύστημά μας με την εντολή αυτή πατώντας το "Train and validate". Καταφέραμε να δημιουργήσουμε μία πρόταση που μπορεί να φέρει κάποιο αποτέλεσμα, όμως είναι αδύνατο να πετύχουμε την συγκεκριμένη πρόταση από κάθε χρήστη. Για αυτό έχουμε την δυνατότητα να δημιουργήσουμε νέες προτάσεις που να αντιστοιχούν στην εντολή αυτή.



Εικόνα 56. Δημιουργία μιας πρότασης ως εντολή.

Τα δεδομένα που δημιουργήσαμε μπορούμε οποιαδήποτε στιγμή να τα κατεβάσουμε και να τα χρησιμοποιήσουμε.

Κεφάλαιο 6. Σύνοψη και συμπεράσματα

Εικόνα

Η ανάκτηση εικόνας απασχολεί τόσο ερευνητές όσο και απλούς χρήστες. Με τη διαθεσιμότητα της τεχνολογίας του διαδικτύου, με βάσεις δεδομένων, datasets, λογισμικά για διάφορα είδη εφαρμογών και την συνεχομένη ανάπτυξη χαμηλού κόστους μηχανήματων αναγνώρισης, έχουν φέρει τις μηχανές αναζήτησης σε επίπεδο που είναι κοντά στον καθένα μας.

Η δημιουργία ενός τέτοιου αντίστοιχου συστήματος, όμως σε λογικά πλαίσια αποτελεί μια διαδικασία αρκετά δαπανηρή, χρονοβόρα και δύσκολη.

Οι μηχανές αναζήτησης ποικίλουν, ωστόσο για την ανάκτηση εικόνας θα πρέπει να γίνει η κατάλληλη έρευνα η οποία θα αντεπεξέρχεται στα ζητήματά μας.

Πολλές από τις τεχνικές που αναφέρθηκαν είτε είναι παρωχημένες και δεν μπορούν να ανταπεξέλθουν στις απαιτήσεις του συγχρόνου χρήστη, είτε βρίσκονται σε στάδιο ανάπτυξης.

Στο διαδίκτυο συναντάμε δεκάδες μηχανές αναζήτησης που χρησιμοποιούν παρόμοιες τεχνολογίες. Στην πλειοψηφία τους είναι δωρεάν η πρόσβαση, όμως αυτό ισχύει κυρίως για επιστημονικά έργα (πρότζεκτ). Ταυτόχρονα, ορισμένες έχουν μικρή βάση δεδομένων και δεν δύναται να ανταπεξέλθουν στην ακριβής ανάκτηση, με φανερό αντίκτυπο όπως το να αγνοούν τυχών ομοιότητες με άλλες φωτογραφίες και να ανακτούν ανακριβής δεδομένα. Οι πιο ολοκληρωμένες φαίνεται να είναι εκείνες που προορίζονται για εμπορική χρήση, όμως σε αυτή την περίπτωση απαιτείται άδεια χρήσης ή συνδρομή επί πληρωμή.

Το διαδίκτυο στην ανάκτηση εικόνας

Κατά την προσωπική περιήγηση μας στο διαδίκτυο για την εύρεση μηχανών αναζήτησης που χρησιμοποιούν τις νέες τεχνολογίες ανάκτησης της εικόνας, διαπιστώσαμε πως:

- η πλειοψηφία τους είναι projects με συγκεκριμένο κοινό (επιστήμονες, φοιτητές, ομοεθνείς).
- πολλές μηχανές αναζήτησης και ανάκτησης εικόνας, δεν αναπτύσσουν τις νέες τεχνικές ή τεχνολογίες, με αποτέλεσμα να καταλήξουν ξεπερασμένες και ξεχασμένες σε σύντομο χρονικό διάστημα.

- ορισμένες έχουν στηριχτεί στην ευρύ χρησιμότητα των κινητών ηλεκτρονικών συσκευών.
- οι εταιρείες αρκετές φορές περιορίζουν τις δυνατότητες ανάκτησης στον βωμό του κέρδους.

Οι εφαρμογές ανάκτησης εν μέρη προσπαθούν να παρέχουν το καλύτερο δυνατό των υπηρεσιών τους, αξιοποιώντας κάθε νέα τεχνολογία αλλά δεν καταφέρνουν να ανταποκριθούν με άνεση στις ανάγκες του κοινού. Οι πολυμεσικές μηχανές αναζήτησης, για να έχουν μακροχρόνια χρησιμότητα, εκτός από την εφαρμογή των συγχρόνων τεχνολογιών (όπως deep learning, machine learning, κλπ.), επιβάλλεται να παρέχουν φιλικότητα και προσβασιμότητα.

Οι χρήστες πρέπει να έχουν ενεργό ρόλο στην ανάπτυξη της μηχανής αναζήτησης. Ειδικότερα, απαραίτητη προϋπόθεση είναι να γίνει η ελεύθερη διάθεση και επεξεργασία του υλικού του μέσα στο σύστημα που χρησιμοποιεί. Η περιγραφή εικόνων από το άνθρωπο είναι μέχρι τώρα η αποτελεσματικότερη μορφή image classification. Οι ίδιοι οι χρήστες εμπλουτίζουν την συλλογή του συστήματος βάζοντας πεδία που συνοδεύουν την εικόνα (π.χ. η εικόνα είναι ένα σκίτσο) ή με την περιγραφή μεταδεδομένων (τίτλος εικόνας, περιγραφή εικόνας, χρήση ετικετών κλπ.). Σχεδόν όλες οι διαδεδομένες πολυμεσικές μηχανές αναζήτησης έχουν ένα τεράστιο κοινό, που μπορεί να προσφέρει το επιθυμητό feedback και βελτιώσεις στο του υπάρχοντος συστήματος.

Σημαντικό είναι επίσης ότι ανάλογα με το τύπο των εικόνων που θα διαχειριστεί το σύστημα, θα πρέπει να γίνουν οι κατάλληλες ρυθμίσεις από τους διαχειριστές ή δημιουργούς του συστήματος για την σωστή περιγραφή και ταξινόμηση των εικόνων. Εκτός αυτού, η συμβατότητα είναι κάτι που πρέπει να λαμβάνεται υπόψη καθώς αποτελεί, σχεδόν αναπόσπαστο κομμάτι ενός "ολοκληρωμένου" συστήματος ανάκτησης.

Ακόμα και αν η τεχνολογία που χρησιμοποιείται είναι αποτελεσματική, θα πρέπει να έχει την δυνατότητα να προσφέρει στους χρήστες όσες περισσότερες τεχνικές γίνεται, ώστε η αναζήτηση να καλύπτει όλες τις δυνατές πτυχές της περιγραφής. Μερικές μηχανές αναζήτησης μένουν στάσιμες επειδή δεν αναπτύσσουν νέες τεχνικές ανάκτησης ή η τεχνική αυτή είναι εύχρηστη για περιορισμένο κομμάτι χρηστών.

Η ανάκτηση με βάση το κείμενο (text-based) μπορεί να θεωρηθεί ξεπερασμένη με τα χρόνια, όμως έχει αποδείξει την σταθερότητα της και ότι ακόμα είναι σε θέση να προσφέρει στο κόσμο της ανάκτησης. Ο συνδυασμός της text based ανάκτησης με την

ανάκτηση του ήχου και πιο συγκεκριμένα την αναγνώριση της μουσικής, είναι ένα χαρακτηριστικό παράδειγμα.

Εκμεταλλεύοντας τις μουσικές επισημάνσεις (music annotations) μέσα από ετικέτες, τίτλους κλπ., που χρησιμοποιούνται για την περιγραφή των τραγουδιών, επιτρέπουν στο text-based retrieval, να στηριχθεί στην απλότητα των δυνατοτήτων του και να αποτελέσει ένα πολύ χρήσιμο εργαλείο ανάκτησης στην συγκεκριμένη περίπτωση.

Αυτήν τη στιγμή, η αναζήτηση και ανάκτηση με βάση το περιεχόμενο (CBIR) γνωρίζει μεγάλη διάδοση χάρη στην πληθώρα των πετυχημένων τεχνικών που υπάρχουν και που εφαρμόζονται.

Με το deep learning και γενικά τα νευρωνικά δίκτυα, να αποτελούν το επίκεντρο γύρω από την ανάπτυξη των συστημάτων που χρησιμοποιούν το CBIR, με τα ποσοστά επιτυχίας τη ανάκτησης εικόνων να έχουν φτάσει σε ένα ικανοποιητικό επίπεδο, σύμφωνα με τις προτιμήσεις των χρηστών. Ακόμα και σε αυτό το ποσοστό, δείχνει πως υπάρχει περιθώριο για περαιτέρω εμπλουτισμό και εξειδίκευση. Καθοριστικό ρόλο σε αυτό έχει η σωστή επιλογή των dataset για την εκπαίδευση των συστημάτων και η συνέχιση της έρευνας για νέες μεθόδους και τεχνολογίες.

Οι κινητές συσκευές αποτελούν έναν ριζικό παράγοντα, καθώς το 60% του παγκόσμιου πληθυσμού κατέχει τουλάχιστον μία κινητή συσκευή. Η χρήση των τεχνολογιών που υπάρχουν σε μία κινητή συσκευή (εύρεση τοποθεσίας (gps), κάμερα), βοηθούν σημαντικά στην καλύτερη εμπειρία του χρήστη. Στις διάφορες υπηρεσίες και ιστοσελίδες, υπάρχουν δεκάδες εφαρμογές κινητών που λειτουργούν ως εργαλείο για να πραγματοποιήσουν οι χρήστες την λειτουργία του reverse image search και κάθε λογής τεχνικής ανάκτησης εικόνας. Οι περισσότερες υπηρεσίες που κάνουν χρήση της ανάκτησης εικόνας στα κινητά, βασίζονται στο CBIR. Πιο συγκεκριμένα, δυστυχώς λίγες από αυτές είναι συμβατές ως υπηρεσίες εκτός των κινητών συσκευών.

Όλες οι αναφερόμενες τεχνικές και οι τύποι συστημάτων είναι χρήσιμοι και μπορούν να ταιριάξουν με όλες τις ιδιαιτερότητες στους πόρους, στο κοινό και στις ικανότητες των προγραμματιστών, αρκεί να γίνει η καλή προετοιμασία.

Ήχος

Ο αντίκτυπος της ανάκτησης μουσικής φαίνεται στις δεκάδες εφαρμογές που έχουν γίνει προσβάσιμες στον καθημερνό άνθρωπο τα τελευταία χρόνια. Οι εφαρμογές, όπως το Shazam, Soundhound και το Google hum to search είναι στο επίκεντρο του MIR, συνδυάζουν και στηρίζονται στις βασικές τεχνολογίες ανάκτησης και αναγνώρισης του ήχου

(όπως το audio fingerprinting και το spectrogram) θέτοντας τις βάσεις του μέλλοντος στον τομέα αυτό.

Τα συστήματα αναγνώρισης ομιλίας είναι συγκριτικά λιγότερα με τα συστήματα ανάκτησης μουσικής και εικόνας. Ένας λόγος θα μπορούσε να είναι η πολυπλοκότητα ενός τέτοιου συστήματος, η ανάγκη για την δημιουργία των τεχνικών ανάκτησης της φωνής και στην συνέχεια η μετατροπή της ομιλίας σε γλώσσα κατανοητή για τον υπολογιστή.

Η τεχνολογία της αναγνώρισης ομιλίας υπάρχει σε όλες τις σύγχρονες συσκευές και λειτουργικά συστήματα. Πολλά από αυτά διαθέτουν αναγνώριση ομιλίας, η οποία είναι ενσωματωμένη ήδη με την εγκατάσταση τους. Ειδικότερα, η Microsoft παρέχει τον προσωπικό οδηγό Microsoft Cortana σε κάθε έκδοση Windows 10, η Google σε κάθε εγκατεστημένο σύστημα android έχει το Google Voice Search κλπ. Η πλειοψηφία τους δεν είναι open-source, αλλά είναι διαθέσιμα δωρεάν. Η χρησιμότητα τους δεν περιορίζεται σε ένα τύπο συσκευής, αφού δουλεύουν ανεξάρτητα αν μιλάμε για υπολογιστή ή κινητή συσκευή. Τα περισσότερα συστήματα αναγνώρισης ομιλίας αφορούν την Αγγλική γλώσσα χωρίς να περιορίζονται απαραίτητα σε αυτήν, με τους ερευνητές να προσπαθούν να δώσουν ποικιλία και σε αυτό το στοιχείο.

Η διαφορά της αναγνώρισης ομιλίας με την ανάκτηση της μουσικής και εικόνας, έγκειται στο γεγονός πως η πρώτη λειτουργεί ως μία βοήθεια για την ανάκτηση πληροφοριών παρά ως μία μεμονωμένη μηχανή αναζήτησης που εξυπηρετεί μονάχα αυτό το σκοπό. Λόγω αυτού, ορισμένοι οργανισμοί προσφέρουν σύνολα δεδομένων dataset και τον κώδικά τους, ώστε όλοι οι ενδιαφερόμενοι να χρησιμοποιήσουν και να ενσωματώσουν την υπηρεσία όπου επιθυμούν.

Η ανάκτηση ήχου λειτουργεί παρόμοια ως προς τον σχεδιασμό με την ανάκτηση εικόνας, όμως έχει διαφορετικές απαιτήσεις και πρόκειται για μία πολυσύνθετη διαδικασία. Απαιτείται πολύ μεγάλη εξειδίκευση στον προγραμματισμό του αλγόριθμου του, στον σχεδιασμό της δομής του συστήματος και την συλλογή του τεράστιου όγκου της (ηχητικής) πληροφορίας που προϋποθέτει την επιτυχία τους.

Παρά τις αρχικές δυσκολίες, τα σύγχρονά μέσα ανάκτησης και κυρίως τα deep learning δίκτυα έχουν καταφέρει να ανταπεξέλθουν με μεγάλο ποσοστό επιτυχίας στις υψηλές απαιτήσεις των συστημάτων τους, προσφέροντας όσο το δυνατόν καλύτερες επιδόσεις στους χρήστες.

Στην διάρκεια του πειραματισμού μας με τα συστήματα, εντοπίσαμε εφαρμογές και τεχνολογίες που είχαν την βέλτιστη χρησιμότητα για τον μέσο χρήστη.

Σημαντικό είναι να αναφέρουμε πως το MIR είναι αναπόσπαστο κομμάτι της μουσικής βιομηχανίας, αφού το απαραίτητο κόστος για τα δικαιώματα χρήσης των αποτυπωμάτων στην βάση δεδομένων του συστήματος, είναι μεγάλο βάρος για έναν άνθρωπο. Για αυτό εταιρίες με προϋπολογισμό εκατομμυρίων έχουν κυριαρχήσει στον τομέα αυτό . Υπάρχουν βέβαια και μικρότερα πρότζεκτ από λάτρεις της μουσικής, που προσπαθούν να ξεχωρίσουν μέσα από το πεδίο της ανάκτησης μουσικής. Χάρη στην συνεισφορά τους, που είναι ευδιάκριτη μέσα στην διεθνή κοινωνία του MIR, ιστοσελίδες όπως το ISMIR καταφέρνουν να φέρουν στην δημοσιότητα νέους και ευέξαπτους επιστήμονες, που ελπίζουν ότι στο μέλλον η σχεδίαση ενός συστήματος ανάκτησης της μουσικής και γενικότερα η ερευνά πάνω στην μουσική θα είναι πιο προσιτή για όλους.

Από την άλλη πλευρά, η αναγνώριση ομιλίας είναι ένας ρεαλιστικός στόχος για τους σχεδιαστές συστημάτων. Πολλές εταιρίες μοιράζονται τα εργαλεία τους με πελάτες ή με χρήστες και μπορούν να δουλέψουν πάνω σε αυτά. Πολλές εφαρμογές κινητών τηλεφώνων έχουν ένα κομμάτι της τεχνολογία αυτής, αφού οποιαδήποτε εντολή που απαιτεί φωνητική αναγνώριση ανήκει στο πεδίο αυτό. Τα διάφορα είδη και οι τεχνικές αναγνώρισης ομιλίας δίνουν ευελιξία στους δημιουργούς ώστε να προσαρμόσουν την τεχνολογία αυτή όπως θέλουν. Τα οφέλη αυτά της αναγνώρισης ομιλίας είναι πολύ σημαντικά για τους χρήστες, και πρέπει να αξιοποιηθούν κατάλληλα από όλους.

Αξιολόγηση τεχνικών πολυμεσικής ανάκτησης

Κρίνεται απαραίτητο να αξιολογήσουμε τις τεχνικές και τεχνολογίες που συνδράμουν στην ανάκτηση. Μέσα από την ακρίβεια (Dong, Hussain & Chang, 2011), την ταχύτητα (Han, Zhang, Cheng, Liu & Xu, 2018) και την πρακτικότητα (Gediga, Hamborg, & Düntsch, 2002) θα προκύψει η βαθμολογία για την κάθε υπηρεσία, με απώτερο στόχο να εντοπίσουμε τα χαρακτηριστικά της ιδανικής μηχανής ανάκτησης.

Κριτήρια

Ο τελικός βαθμός υπολογίζεται από τα ακόλουθα κριτήρια, τα όποια είναι ισοδύναμα και η βαθμολογία αυτών βγαίνει με άριστα το 10.

Ακρίβεια: Είναι μια επέκταση της μεθόδου αξιολόγησης της ακρίβειας (precision) **(βλέπε 3.3).** Η βαθμολογία της βασίστηκε στην ποιότητα των αποτελεσμάτων μας (Dong, Hussain & Chang, 2011). Τα βήματα που ακολουθήσαμε είναι τα εξής:

- Στέλνουμε κάθε ένα από τα δείγματα στην μηχανή αναζήτησης (η λίστα των δειγμάτων βρίσκεται στο **Παράρτημα**).
- Ελέγχουμε αν το αποτέλεσμα που επιστρέφει είναι σχετικό με το δείγμα.
- Το σχετικό αποτέλεσμα παίρνει μία (1) μονάδα και το μη σχετικό μηδέν (0) μονάδες. Στα μη σχετικά γίνεται επανάληψη της αναζήτησης, όπου αν φέρει σχετικό αποτέλεσμα, παίρνει μισή (0,5) μονάδα.
- Συλλέγονται όλες οι μονάδες και συγκρίνονται με το σύνολο των δειγμάτων από την μηχανή αναζήτησης (π.χ. 25 δείγματα και 20 μονάδες). Υπολογίζεται η ακρίβεια (20/25=0,8) και τέλος γίνεται η αναγωγή για να γίνει η βαθμολόγηση με άριστα το 10 (0,8x10=8).

Κριτήριο	Περιγραφή	Μέγιστη βαθμολογία
Ακρίβεια	1. Υπολογισμός Σχετικότητας : - Σχετικά αποτελέσματα = 1 μονάδα - Σχετικά αποτελέσματα μετά από 2η προσπάθεια = 0,5 μονάδα - Μη σχετικά αποτελέσματα = 0 μονάδες 2. Ακρίβεια = Σχετικά αποτελέσματα/Σύνολο δειγμάτων 3. Βαθμολογία = Ακρίβεια x 10	10

Πίνακας 1. Περιγραφή της ακρίβειας.

Ταχύτητα: Εκτός από την ποιότητα των αποτελεσμάτων, σημαντική είναι και η ταχύτητα με την οποία ανταποκρίνεται η υπηρεσία. Σε αυτή την περίπτωση η βαθμολογία προκύπτει από τα δευτερόλεπτα που χρειάζονται για να εμφανιστεί το αποτέλεσμα (Han, Zhang, Cheng, Liu & Xu, 2018). Αναλυτικότερα:

Δευτερόλεπτα (s)	Βαθμολογία
<1	10
1-1,4	9
1,5-2,4	8
2,5-4,9	7
5-7,9	6
8-11,9	5
12-17,9	4
18-24,9	3
25-40	2
>=40	1

Πίνακας 2. Βαθμολόγηση της ταχύτητας των αποτελεσμάτων.

Πρακτικότητα : Σε τι ποσοστό είναι ικανό το σύστημα ανάκτησης να αφήσει ικανοποιημένο τον χρήστη σε λειτουργίες πέραν της απόδοσης; Η βαθμολογία βασίζεται κατά βάση στην εμπειρία μας κατά την διάρκεια χρήσης των συστημάτων που θα αναφερθούν . Για να επιτευχθεί αυτό, πήραμε ένα μέρος του προτύπου αξιολόγησης ISO 9126 (Gediga, Hamborg, & Düntsch, 2002). Πιο συγκεκριμένα:

Κριτήριο	Χαρακτηριστικό	Μέγιστη βαθμολογία
Χρησιμότητα	Εμφάνιση: Η ποιότητα του περιβάλλοντος διεπαφής χρήστη (χρώματα και αισθητική).	2,5
	Χειρισμός: Ο βαθμός δυσκολίας που απαιτείται για να αποσταλεί ένα ερώτημα στην υπηρεσία (πλοήγηση).	2,5
Φορητότητα	Προσαρμοστικότητα: Η επίδοση και διάθεση των υπηρεσιών μεταξύ ηλεκτρονικών υπολογιστών και κινητών συσκευών.	2,5
	Ευστάθεια: Η πιθανότητα να εμφανιστούν προβλήματα μη σχετικά με την ανάκτηση. (προβλήματα με διακομιστή, διακοπή λειτουργίας κλπ.)	2,5

Πίνακας 3. Κριτήρια της πρακτικότητας.

Υπηρεσίες

Για την εικόνα, οι τεχνολογίες δοκιμάστηκαν από τις έξης υπηρεσίες και μηχανές αναζήτησης:

Τεχνολογία	Υπηρεσία
CBIR	Google Lens
	Google Image Search
	Bing

Για την μουσική:

Τεχνολογία	Υπηρεσία
Music Identification	Soundhound
	Shazam
	Google Hum to search

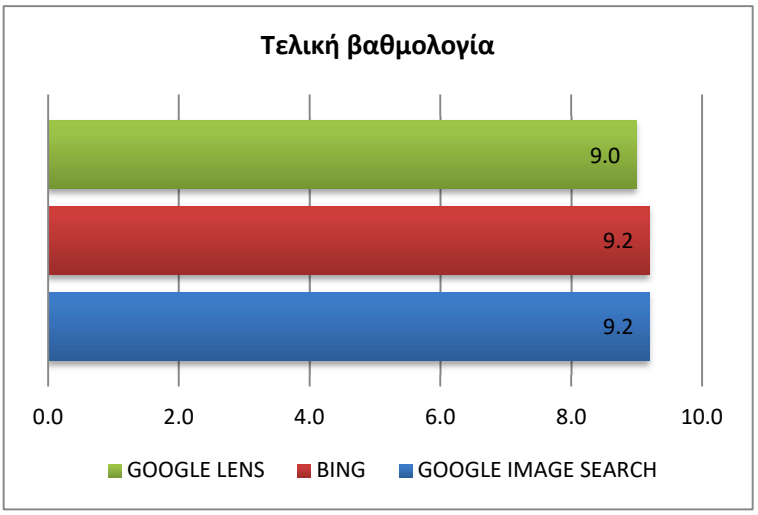
και για την ομιλία:

Τεχνολογία	Υπηρεσία
Speech recognition	Google Now
	Bing

Βαθμολόγηση

Με βάση τα αποτελέσματα, καταλήξαμε στις τελικές βαθμολογίες. Οι οποίες παρουσιάζονται μαζί με αναλυτικούς πίνακες και γραφήματα.

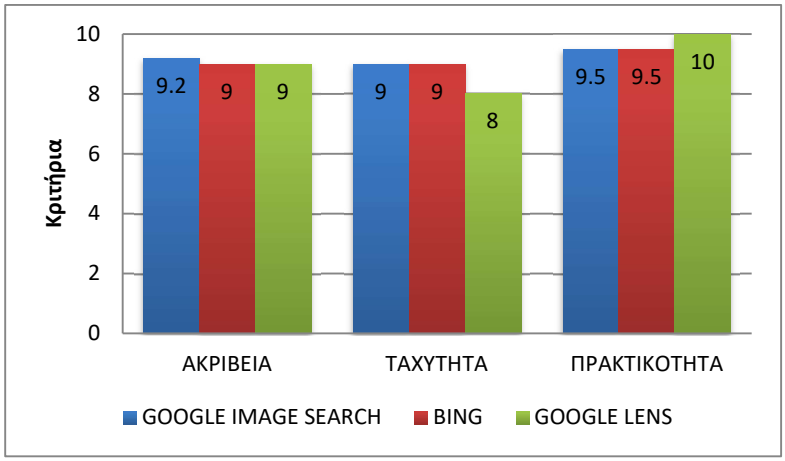
Συστήματα ανάκτησης εικόνας



Γράφημα 1Α. Τελική βαθμολογία συστημάτων ανάκτησης εικόνας.

ΜΕΘΟΔΟΣ	ΣΥΣΤΗΜΑ	ΑΚΡΙΒΕΙΑ (Σχετικά/Σύνολο)	ΤΑΧΥΤΗΤΑ (Δευτερόλεπτα)	ΠΡΑΚΤΙΚΟΤΗΤΑ (ΕΜΦ/ΧΕΙΡ/ΠΡΟ/ΕΥΣ)	ΤΕΛΙΚΗ ΒΑΘΜΟΛΟΓΙΑ
CBIR	Google Image Search	**9,2** (46/50)	**9** (1 s)	**9,5** (9/9/10/10)	**9,2**
	Bing	**9** (45/50)	**9** (1,2 s)	**9,5** (9/9/10/10)	**9,2**
	Google Lens	**9** (45/50)	**8** (1,5 s)	**10** (10/10/10/10)	**9**

Πίνακας 4. Βαθμολόγηση των συστημάτων ανάκτησης εικόνας.

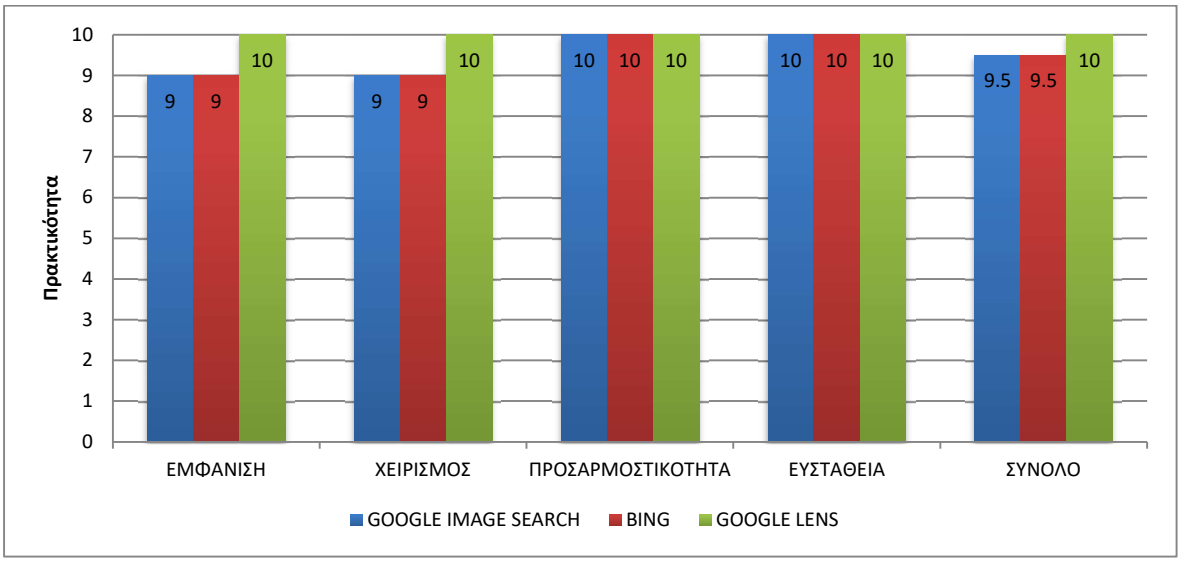


Γράφημα 1Β. Βαθμολόγηση των συστημάτων ανάκτησης εικόνας.

Οπού η βαθμολογία της πρακτικότητας προκύπτει από:

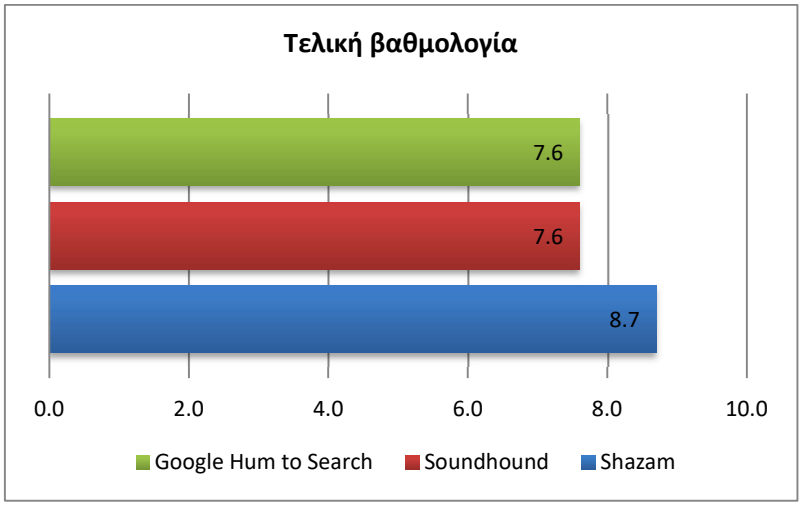
ΜΕΘΟΔΟΣ	ΣΥΣΤΗΜΑ	ΕΜΦΑΝΙΣΗ	ΧΕΙΡΙΣΜΟΣ	ΠΡΟΣΑΡΜΟΣΤΙΚΟΤΗΤΑ	ΕΥΣΤΑΘΕΙΑ	ΣΥΝΟΛΟ
CBIR	Google Image Search	**9**	**9**	**10**	**10**	**9,5**
	Bing	**9**	**9**	**10**	**10**	**9,5**
	Google Lens	**10**	**10**	**10**	**10**	**10**

Πίνακας 5. Βαθμολόγηση της πρακτικότητας.



Γράφημα 1Γ. Βαθμολόγηση της πρακτικότητας.

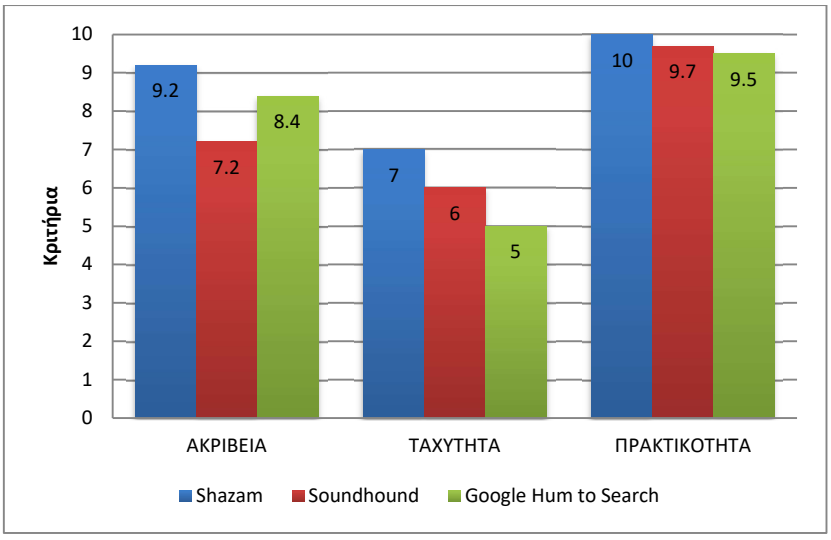
Συστήματα ανάκτησης μουσικής



Γράφημα 2Α. Τελική βαθμολογία συστημάτων ανάκτησης μουσικής.

ΜΕΘΟΔΟΣ	ΣΥΣΤΗΜΑ	ΑΚΡΙΒΕΙΑ (Σχετικά/Σύνολο)	ΤΑΧΥΤΗΤΑ (Δευτερόλεπτα)	ΠΡΑΚΤΙΚΟΤΗΤΑ (ΕΜΦ/ΧΕΙΡ/ΠΡΟ/ΕΥΣ)	ΤΕΛΙΚΗ ΒΑΘΜΟΛΟΓΙΑ
Music Identification	Shazam	**9,2** (23/25)	**7** (2,5 s)	**10** (10/10/10/10)	**8,7**
	Soundhound	**7,2** (18/25)	**6** (5 s)	**9,8** (10/10/10/9)	**7,6**
	Google Hum to Search	**8,4** (21/25)	**5** (11 s)	**9,5** (10/8/10/10)	**7,6**

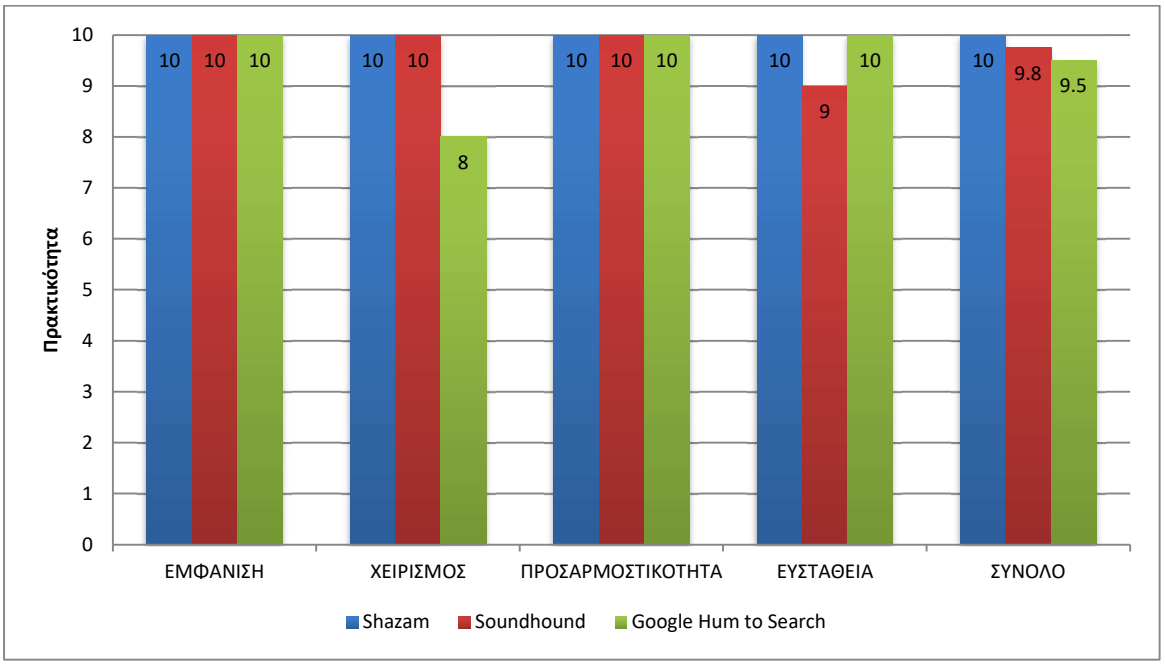
Πίνακας 6. Βαθμολόγηση των συστημάτων ανάκτησης μουσικής.



Γράφημα 2Β. Βαθμολόγηση των συστημάτων ανάκτησης μουσικής.

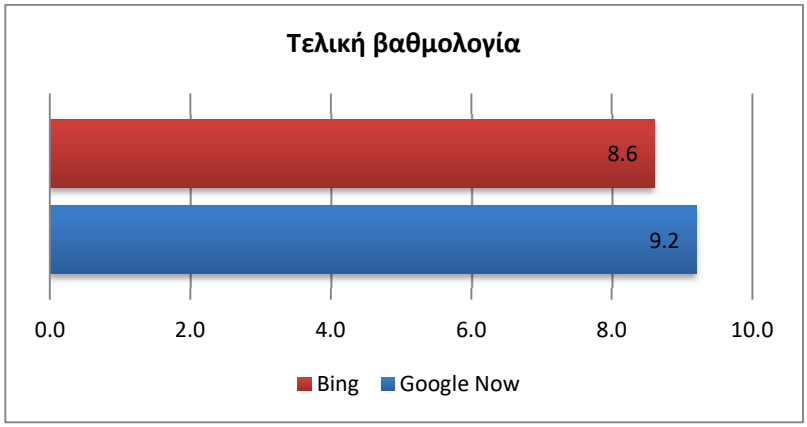
ΜΕΘΟΔΟΣ	ΣΥΣΤΗΜΑ	ΕΜΦΑΝΙΣΗ	ΧΕΙΡΙΣΜΟΣ	ΠΡΟΣΑΡΜΟΣΤΙΚΟΤΗΤΑ	ΕΥΣΤΑΘΕΙΑ	ΣΥΝΟΛΟ
Music Identification	Shazam	**10**	**10**	**10**	**10**	**10**
	Soundhound	**10**	**10**	**10**	**9**	**9,8**
	Google Hum to Search	**10**	**8**	**10**	**10**	**9,5**

Πίνακας 7. Βαθμολόγηση της πρακτικότητας.



Γράφημα 2Γ. Βαθμολόγηση της πρακτικότητας.

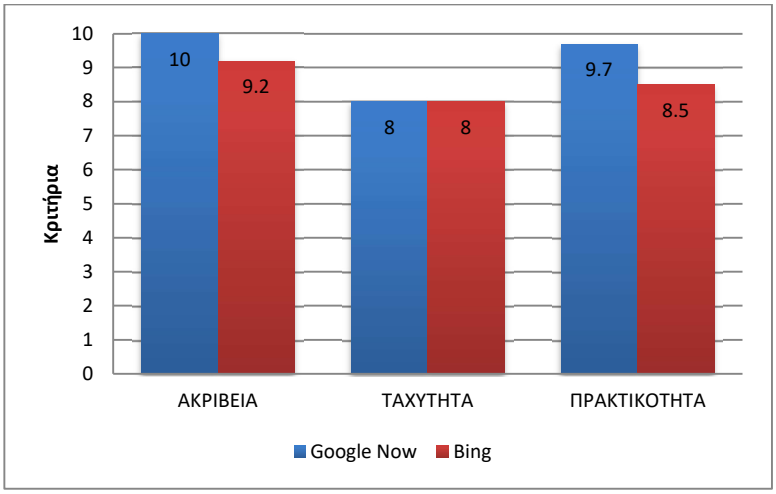
Συστήματα αναγνώρισης ομιλίας



Γράφημα 3Α. Τελική βαθμολογία συστημάτων αναγνώρισης ομιλίας.

ΜΕΘΟΔΟΣ	ΣΥΣΤΗΜΑ	ΑΚΡΙΒΕΙΑ (Σχετικά/Σύνολο)	ΤΑΧΥΤΗΤΑ (Δευτερόλεπτα)	ΠΡΑΚΤΙΚΟΤΗΤΑ (ΕΜΦ/ΧΕΙΡ/ΠΡΟ/ΕΥΣ)	ΤΕΛΙΚΗ ΒΑΘΜΟΛΟΓΙΑ
Speech Recognition	Google Now	**10** (25/25)	**8** (1,5 s)	**9,8** (9/10/10/10)	**9,2**
	Bing	**9,2** (23/25)	**8** (1,5 s)	**8,5** (7/7/10/10)	**8,6**

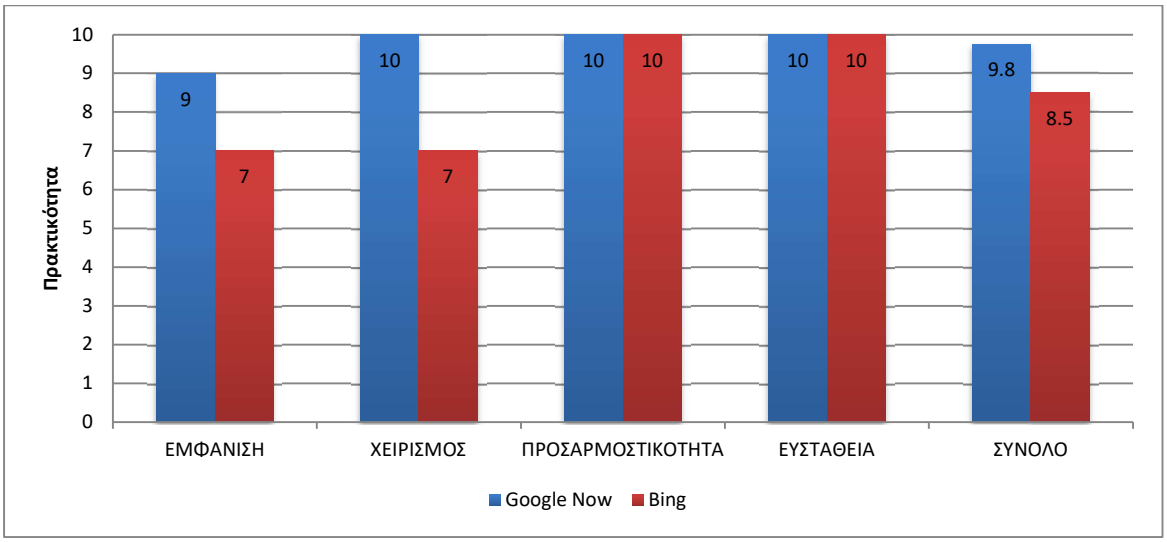
Πίνακας 8. Βαθμολόγηση των συστημάτων αναγνώρισης ομιλίας.



Γράφημα 3Β. Βαθμολόγηση των συστημάτων αναγνώρισης ομιλίας.

ΜΕΘΟΔΟΣ	ΣΥΣΤΗΜΑ	ΕΜΦΑΝΙΣΗ	ΧΕΙΡΙΣΜΟΣ	ΠΡΟΣΑΡΜΟΣΤΙΚΟΤΗΤΑ	ΕΥΣΤΑΘΕΙΑ	ΣΥΝΟΛΟ
Speech Recognition	Google Now	**9**	**10**	**10**	**10**	**9,8**
	Bing	**7**	**7**	**10**	**10**	**8,5**

Πίνακας 9. Βαθμολόγηση της πρακτικότητας.



Γράφημα 3Γ. Βαθμολόγηση της πρακτικότητας.

Με βάση τα παραπάνω, καταλήξαμε στο συμπέρασμα ότι:

- Η Google έχει ένα μικρό προβάδισμα στην πολυμεσική αναζήτηση εικόνας και ομιλίας. Φαίνεται πως είναι κατάλληλα εφοδιασμένο με τις σύγχρονες τεχνολογίες και τεχνικές ανάκτησης και αναζήτησης παρέχοντας άμεση ανταπόκριση με εύστοχο περιεχόμενο αποτελεσμάτων.
- Η Microsoft μέσα από το Bing, έχει αναβαθμίσει σε μεγάλο βαθμό τις τεχνικές ανάκτησης, τόσο που να μπορεί να ανταγωνιστεί ισάξια την Google.
- Το Shazam προσφέρει την καλύτερη εφαρμογή ανάκτησης μουσικής και η διαφορά μεγέθους στην βάση δεδομένων των μουσικών κομματιών είναι εμφανής συγκριτικά με το Soundhound.
- Η τεχνολογία Google Hum to Search παρά το γεγονός ότι κυκλοφόρησε πρόσφατα φαίνεται να ανταποκρίνεται αρκετά αποτελεσματικά ως προς την αξιοποίηση των φωνητικών δεδομένων.
- Η ευκολία στην κατανόηση της ομιλίας είναι αξιοσημείωτη στα συστήματα που δοκιμάσαμε, όμως το Google Now μέσα από τις ποικίλες διαθέσιμες γλώσσες φαίνεται να επικρατεί ως η καλύτερη επιλογή από αυτή του Bing.

Το πληρέστερο σύστημα ανάκτησης εικόνας

Σύμφωνα με τα παραπάνω και υποθέτοντας ότι έχουμε τους διαθέσιμους πόρους, είμαστε σε θέση να προτιμήσουμε τεχνικές και τεχνολογίες για την κατασκευή του ιδανικού συστήματος ανάκτησης εικόνας.

Η κατάλληλη προσέγγιση για την δημιουργία του, θα ήταν εστιασμένη σε ένα εύκολο σε περιήγηση περιβάλλον, με βολικό πλαίσιο συστήματος και ολοκληρωτικά να είναι όσο το δυνατόν πιο εξυπηρετικό για τον χρήστη. Επιφανειακά αυτό φαίνεται να είναι επαρκές, αλλά οι τεχνολογίες που απορρέουν από το σύστημα θα πρέπει να τηρούν όλες τις κατάλληλες προδιαγραφές ως προς τις προαναφερόμενες απαιτήσεις. Πρέπει να συνδυάζει τα πλεονεκτήματα από όλους τους ενεργούς τύπους ανάκτησης και αναζήτησης εικόνας. Από τον πιο απλό τύπο του TBIR, στον επικρατέστερο τύπο του CBIR παρέχοντας εδραιωμένο αποτέλεσμα ανεξαρτήτως του τρόπου αναζήτησης του χρηστή.

Θα επικεντρωνόμασταν σε όσο πιο ακριβή ανάκτηση μέσω TBIR καθώς φαίνεται να είναι ο πιο κατανοητός και χρήσιμος τρόπος αναζήτησης για τον χρήστη. Ταυτόχρονα δίνει την δυνατότητα εναλλαγής του τρόπου ερωτήματος χρησιμοποιώντας το CBIR και των διάφορων εργαλείων του.

Η εκμετάλλευση των νευρωνικών δικτύων, και πιο συγκεκριμένα των CNN δεν γίνεται να παραλειφθεί. Ένα σύστημα είναι απαραίτητο να διαθέτει τεχνολογίες που να έχουν εκπαιδευτεί πάνω στα πιθανά ερωτήματα και να μπορεί να ανταποκριθεί σε όλες τις αναζητήσεις του κοινού. Για αυτό θα εστιάζαμε σε ένα δυνατό dataset που θα μπορούσε ανεξάρτητα από το ποσό μεγάλη είναι η ζήτηση, να μπορεί να ανταπεξέλθει με ευελιξία στις απαιτήσεις των χρηστών. Αν το σύστημα αφορά μια τεράστια αγορά, οι απαιτήσεις είναι περισσότερες. Όμως αν τα ερωτήματα είναι συγκεκριμένα και λιγότερο απαιτητικά, τότε μέσα από τα κατάλληλα dataset, οι χρήστες μπορούν και πάλι να εξυπηρετηθούν.

Είναι γεγονός πως οι σωστές επιλογές των τεχνικών ανάκτησης, καθιστούν τη μηχανή αναζήτησης πετυχημένη. Όμως δεν έχει βρεθεί η χρυσή τομή ακόμα. Αν επιθυμούμε να υπάρχει τεράστιο κοινό, τότε πρέπει να έχουμε στο νου μας τις δυνατότητες των ηλεκτρονικών συσκευών. Μία νέα τεχνολογία που θα επιβάρυνε τους χρήστες, θα ήταν μονάχα μία δυσμενής προσθήκη.

Το πληρέστερο σύστημα ανάκτησης μουσικής

Η δημιουργία ενός συστήματος ανάκτησης μουσικής, στην προκείμενη περίπτωση, πέρα από το λειτουργικό κομμάτι, απαιτεί την χρήση των αναφερομένων τεχνολογιών, εφαρμογών και τεχνικών. Δηλαδή, την αντιστοίχιση των μουσικών κομματιών μέσω audio fingerprinting, σε συνδυασμό με το spectrogram και όλων των συναφών τεχνολογιών (chord recognition, track separation, tempo estimation και audio alignment), για την βελτίωση των αποτελεσμάτων. Εξυπηρετεί την συνεχή ανάπτυξη αυτών και της τεχνολογίας αντιστοίχισης και ανάκτησης μέσω deep learning, νευρωνικών δικτύων και datasets.

Η διεπαφή του χρήστη με την βάση δεδομένων θα γίνει αποκλειστικά από μία εφαρμογή για κινητά τηλέφωνα, οπού φαίνεται κιόλας να επικεντρώνεται μελλοντικά και η ανάκτηση της μουσικής.

Η ιδανική προσέγγιση της δομής αυτής θα στηρίζονταν στις ήδη υπάρχουσες εφαρμογές του Shazam και του Soundhound και τον ενσωματωμένο τύπο μουσικής αναζήτησης της Google (Hum to search), καθώς λόγο της απλοϊκής τους δομής και της ευκολίας της χρήσης τους αποτελούν ιδανικό παράδειγμα προς μίμηση.

Στηριζόμενοι όμως στον τρόπο λειτουργίας του Shazam, θα πρέπει να επικεντρωθούμε στην ανάπτυξη μίας ευρείας και συνεχόμενα αναβαθμιζόμενης βάσης δεδομένων των μουσικών κομματιών και των αποτυπωμάτων τους, για να μην υπάρξουν κενά στις αναζητήσεις και ανακτήσεις των χρηστών.

Το query by humming είναι επίσης ένας τύπος αναζήτησης που πρέπει να αξιοποιηθεί, παρότι βρίσκεται σε ένα σχετικά πρώιμο στάδιο, η Google έχει αναδείξει τις δυνατότητες του. Βρίσκεται σε ένα ικανοποιητικό επίπεδο που μπορεί να προσφέρει ιδιαίτερα στην ποιότητα των ερωτημάτων. Με την ραγδαία εξέλιξη των τεχνολογιών η συνεισφορά του μπορεί να αποβεί μοιραία ακόμη και μελλοντικά.

Το πληρέστερο σύστημα ανάκτησης αναγνώρισης ομιλίας

Αν πρέπει να επιλέξουμε το τύπο ενός συστήματος αναγνώρισης ομιλίας για την προσθήκη του σε κάποια μηχανή αναζήτησης, τότε αυτό δεν θα είναι βασισμένο στον ομιλητή (speaker independent). Ο λόγος είναι πως η μηχανή αυτή έχει στόχο να χρησιμοποιηθεί από διαφορετικούς χρήστες και πρέπει να λειτουργεί με γνώμονα αυτό. Ένα ερώτημα σε μια μηχανή αναζήτησης περιλαμβάνει μία σειρά από προσδιορισμένες έννοιες που περιγράφουν το υλικό της.

Η επίδοση των συστημάτων είναι περίπου η ίδια, όμως οι λειτουργίες είναι διαφορετικές και αλλάζουν ανάλογα με αυτό που επιθυμούμε. Το Spontaneous speech recognition δεν είναι εύλογη επιλογή, επειδή η ανάπτυξη του είναι πολύπλοκη και χρειάζεται περισσότερους πόρους σε σύγκριση με τα άλλα συστήματα. Ένα Connected word recognition ή Isolated word speech recognition σύστημα θα ήταν ιδανικό για εμάς, αφού είναι κατάλληλο για την απάντηση ερωτημάτων με την χρήση λέξεων-κλειδιών.

Στην περίπτωση του λεξιλογίου, καλό θα ήταν να το δανειστούμε από κάποια βάση δεδομένων, γιατί η δημιουργία ενός από την αρχή είναι απαιτητική. Απαραίτητη προϋπόθεση είναι να διαθέτει τις γλώσσες που θα χρησιμοποιούνται και να υπάρχει η επαρκής ανταπόκριση στην αναγνώριση των ιδιαιτεροτήτων της φωνής. Συνήθως οι πλατφόρμες που τις παρέχουν δωρεάν, είναι υπό ανάπτυξη, όμως στην περίπτωση του Common Voice, οι υπηρεσίες του είναι υπεραρκετές.

Επίλογος

Ολοένα και περισσότεροι επενδύουν στην τεχνητή νοημοσύνη με στόχο να μειώσουν το χρόνο σε καθημερινές εργασίες δίχως να απασχολούν τον άνθρωπο. Η πραγματοποίηση της αυτονομίας αποτελεί έργο δύσκολο, καθώς τα εκάστοτε δομικά στοιχεία ενός συστήματος πρέπει να λειτουργούν σε απόλυτη αρμονία. Από την στιγμή που δόθηκε σε περισσότερους ανθρώπους η δυνατότητα πειραματισμού, η εξέλιξη ήταν το φυσικό επακόλουθο.

Η ανάκτηση πολυμεσικού υλικού, όλες αυτές τις δεκαετίες βασίστηκε σε αυτή την ιδέα και κάνει σταθερά βήματα προόδου. Αυτή μας επιτρέπει να οργανώσουμε και να αναζητήσουμε υλικό το οποίο θα ήταν αδύνατο να βρεθεί στο αχανές διαδίκτυο. Η πληθώρα των ηλεκτρονικών συσκευών ενισχύει την προβολή της πολυμεσικής ανάκτησης. Η μοναδικότητα του κάθε υλικού δεν αποτελεί εμπόδιο για την περιγραφή του, αλλά μία νέα δοκιμασία για την επιστημονική κοινότητα. Πιο συγκεκριμένα, η αναγνώριση αντικειμένων σε μία εικόνα και η κατηγοριοποίησή τους, ο διαχωρισμός των μουσικών οργάνων και η αντιστοίχιση του ρυθμού τους με των διασκευών και όλα τα συναφή επιτεύγματα είναι το παρόν και το μέλλον της ανάκτησης πληροφοριών. Όλοι μπορούν να αντιληφθούν τον αντίκτυπο που έχει και ελπίζουμε να συνεχιστεί αυτή η ανοδική πορεία και να δούμε νέες δυνατότητες στην αναζήτησή του πολυμεσικού υλικού.

Βιβλιογραφικές Αναφορές

1. Meharban, M., & Priya, D. (2021). *A Review on Image Retrieval Techniques*. Academia.edu. Retrieved 12 November 2021, from https://www.academia.edu/31727733/A_Review_on_Image_Retrieval_Techniques?auto=citations&from=cover_page.
2. Elmogy, M., Alkhawlani, M., & M El-Bakry, H. (2021). *Text-based, Content-based, and Semantic-based Image Retrievals: A Survey*. Reasearch Gate. Retrieved 3 December 2021, from https://www.researchgate.net/publication/273258916_Text-based_Content-based_and_Semantic-based_Image_Retrievals_A_Survey.
3. Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence And Neuroscience*, *2018*, 1-13. https://doi.org/10.1155/2018/7068349 .
4. Chen, W., Liu, Y., Wang, W., M. Bakker, E., & Georgiou, T. (2021). *Deep Image Retrieval: A Survey* [Ebook]. Retrieved 4 December 2021, from 2101.11282_deep%20oimage%20retrieval.pdf.
5. Wang, Y., Liu, T., Bu, Z., Huang, Y., Gao, L., & Wang, Q. (2021). *A Semantic Indexing Structure for Image Retrieva* [Ebook]. School of Information Science and Engineering, Southeast University, Nanjing, China. Retrieved 2 December 2021, from A%20Semantic.pdf.
6. Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. Retrieved 10 December 2021, from https://www.cs.toronto.edu/~kriz/cifar.html
7. Haldekar, M., Ganesan, A., & Oates, T. (2017). *Identifying Spatial Relations in Images using Convolutional Neural Networks* [Ebook]. Retrieved 5 December 2021, from Spatial%20Relations.pdf.
8. Zenggang, X., Zhiwen, T., Xiaowen, C., Xue-min, Z., Kaibin, Z., & Conghuan, Y. (2019). Research on Image Retrieval Algorithm Based on Combination of Color and Shape Features. *Journal Of Signal Processing Systems*, *93*(2-3), 139-146. https://doi.org/10.1007/s11265-019-01508-y
9. Mutlag, W., Ali, S., Aydam, Z., & Taher, B. (2020). Feature Extraction Methods: A Review. *Journal Of Physics: Conference Series*, *1591*(1), 012028. https://doi.org/10.1088/1742-6596/1591/1/012028
10. Ali Jafar Zaidi, S., Buriro, A., Noman Riaz, M., Mahboob, A., & Riaz, M. (2019). *Implementation and Comparison of Text-Based Image Retrieval Schemes* [Ebook]. IJACSA. Retrieved 9 November 2021, from http://Paper_77-Implementation_and_Comparison_of_Text_based_Image.pdf.
11. Chandra, A. (2021). McCulloch-Pitts Neuron — Mankind's First Mathematical Model Of A Biological Neuron. Medium. Retrieved 17 November 2021, from https://towardsdatascience.com/mcculloch-pitts-model-5fdf65ac5dd1 .
12. Adjabi, I., Ouahabi, A., Benzaoui, A., & Taleb-Ahmed, A. (2020). Past, Present, and Future of Face Recognition: A Review. Electronics, 9(8), 1188. https://doi.org/10.3390/electronics9081188

13. Erik Cheever, S. (2022). Linear Physical Systems - Erik Cheever. Retrieved 15 January 2022, from https://lpsa.swarthmore.edu/Fourier/Xforms/FXformIntro.html
14. Datasets. (2021). Retrieved 4 January 2022, from https://ismir.net/resources/datasets/
15. Education, I. (2020). What is Machine Learning?. Retrieved 19 December 2021, from https://www.ibm.com/cloud/learn/machine-learning
16. Willyard, W. (2010). What is an Acoustic Model in Speech Recognition?. Retrieved 23 March 2022, from https://www.rev.com/blog/resources/what-is-an-acoustic-model-in-speech-recognition
17. Acoustic Score Computation. Retrieved 16 March 2022, from http://i13pc106.ira.uka.de/~mthoma/asr/disc-thread/score
18. Fu, W. (2020). Application of an Isolated Word Speech Recognition System in the Field of Mental Health Consultation: Development and Usability Study. JMIR Medical Informatics, 8(6), e18677. doi: 10.2196/18677
19. Speaker Dependent / Speaker Independent. (2022). Retrieved 9 March 2022, from https://www.imagesco.com/articles/hm2007/SpeechRecognitionTutorial02.html
20. Merriam-Webster. (n.d.). Timbre. In Merriam-Webster.com dictionary. Retrieved January 9, 2022, from https://www.merriam-webster.com/dictionary/timbre
21. Pubudumal, D. (2018). Demystifying Shazam (or Soundhound). Retrieved 14 March 2022, from https://medium.com/swlh/demystifying-shazam-or-soundhound-c85d0afd5910
22. Shahrior, T. (2021). Hash Tables, Hashing and Collision Handling. Retrieved 6 March 2022, from https://medium.com/codex/hash-tables-hashing-and-collision-handling-8e4629506572
23. Local Feature Detection and Extraction. Retrieved 17 December 2021, from https://www.mathworks.com/help/vision/ug/local-feature-detection-and-extraction.html
24. Audio Alignment. Retrieved 15 March 2022, from https://steinberg.help/nuendo/v8/en/cubase_nuendo/topics/sample_editor_tempo_matching_audio/sample_editor_tempo_matching_audio_alignment_c.html
25. Ebermann, T., Helfenstein, P., Ebermann, T., & Ebermann, T. (2022). Speech recognition with wit.ai Blog Liip. Retrieved 17 March 2022, from https://www.liip.ch/en/blog/speech-recognition-with-wit-ai
26. Size, F., Size, F., & WIRE, B. (2015). SoundHound First Music Discovery Service to Be Live on Wearables. Retrieved 6 March 2022, from https://www.businesswire.com/news/home/20150313005592/en/SoundHound-Music-Discovery-Service-Live-Wearables#.VXFFhVxVikq

27. Jovanovic, J. (2015). How does Shazam work | Coding Geek. Retrieved 11 March 2022, from http://coding-geek.com/how-shazam-works/
28. Cooper, T. (2018). How Shazam Works. Retrieved 13 March 2022, from https://medium.com/@treycoopermusic/how-shazam-works-d97135fb4582
29. Han, J., Zhang, D., Cheng, G., Liu, N., & Xu, D. (2018). Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey. IEEE Signal Processing Magazine, 35(1), 84-100. doi: 10.1109/msp.2017.2749125
30. Dong, H., Hussain, F., & Chang, E. (2011). A Service Search Engine for the Industrial Digital Ecosystems. IEEE Transactions On Industrial Electronics, 58(6), 2183-2196. doi: 10.1109/tie.2009.2031186
31. Imbriaco, R., Sebastian, C., Bondarev, E., & de With, P. (2019). Aggregated Deep Local Features for Remote Sensing Image Retrieval. Remote Sensing, 11(5), 493. doi: 10.3390/rs11050493
32. Gediga, G., Hamborg, K. C., & Düntsch, I. (2002). Evaluation of software systems. Encyclopedia of computer science and technology, 45(supplement 30), 127-53.
33. Saon, G., & Chien, J. (2012). Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances. IEEE Signal Processing Magazine, 29(6), 18-33. doi: 10.1109/msp.2012.2197156
34. Buckland, M., & Gey, F. (1994). The relationship between recall and precision. Journal of the American society for information science, 45(1), 12-19.
35. Shorten, C., & Khoshgoftaar, T. (2019). A survey on Image Data Augmentation for Deep Learning. Journal Of Big Data, 6(1). doi: 10.1186/s40537-019-0197-0
36. Grosche, P., Müller, M., & Serrà, J. (2012). Audio Content-Based Music Retrieval. doi: 10.4230/DFU.Vol3.11041.157
37. Feaster, P. (2018). The Secret Military Origins of the Sound Spectrograph. Retrieved 6 May 2022, from https://griffonagedotcom.wordpress.com/2018/07/26/the-secret-military-origins-of-the-sound-spectrograph/

Πρόσθετη Βιβλιογραφία

Σε αυτή τη παράγραφο προστίθεται η πρόσθετη βιβλιογραφία η οποία δεν έχει παραπομπή εντός του κειμένου της πτυχιακής εργασίας.

1. Sachendra, S. (2020). Springer Link. Retrieved 3 November 2021, from https://doi.org/10.1007/s11042-019-08401-7.
2. TAREK, M. (2021). *Image Retrieval: Color Coherence Vector*. Owlcation. Retrieved 29 November 2021, from https://owlcation.com/stem/Image-Retrieval-Color-Coherence-Vector.
3. *Computer Vision*. Google Books. (2021). Retrieved 29 November 2021, from https://books.google.gr/books?id=OEsgAwAAQBAJ&pg=PA541&dq=Computer+vision&hl=el&sa=X&ved=2ahUKEwjinJmEydn0AhVQSPEDHaeuCDUQ6AF6BAgGEAI#v=onepage&q&f=false .
4. Goodrum, A. (2021). *Image Information Retrieval: An Overview of Current Research* [Ebook]. Retrieved 29 October 2021, from IMAGE.pdf.
5. Sandeep, K. (2021). *How Many Images Are On The Internet 2021?*. 16 Best. Retrieved 30 November 2021, from https://www.16best.net/how-many-images-are-on-the-internet.
6. Szeliski, R. (2010). *Computer Vision Algorithms and Applications* [Ebook]. Baya Lina. Retrieved 10 December 2021, from https://www.academia.edu/35928048/Computer_Vision_Algorithms_and_Applications_pdf?from=cover_page.
7. Desai, P., Pujari, J., & Parvatikar, S. (2011). Image Retrieval Using Shape Feature: A Study. *Communications In Computer And Information Science*, 817-821. https://doi.org/10.1007/978-3-642-25734-6_146
8. Hirwane, R. (2017). *Semantic based Image Retrieval* ,2017.https://DOI10.17148/IJARCCE.2017.6423
9. Kalantidis, Y., Tolias, G., pyrou, E., Mylonas, P., & Avrithis, Y. *Visual Image Retrieval and Localization* [Ebook]. Retrieved 29 November 2021.
10. Difference between Image Processing and Computer Vision - GeeksforGeeks. (2021). Retrieved 21 December 2021, from https://www.geeksforgeeks.org/difference-between-image-processing-and-computer-vision/
11. A Brief History of Computer Vision (and Convolutional Neural Networks) | Hacker Noon. (2021). Retrieved 23 December 2021, from https://hackernoon.com/a-brief-history-of-computer-vision-and-convolutional-neural-networks-8fe8aacc79f3
12. Panda, S. (2008). Image Mining, Spatial. *Encyclopedia Of GIS*, 475-479. doi: 10.1007/978-0-387-35973-1_585
13. Dey, N., Karâa, W., Chakraborty, S., Banerjee, S., Salem, M., & Azar, A. (2015). Image mining framework and techniques: a review. *International Journal Of Image Mining*, *1*(1), 45. doi: 10.1504/ijim.2015.070028
14. CHANDREN, S. (2008). TinEye - Searching For Images With An Image. Retrieved 28 December 2021, from https://www.makeuseof.com/tag/tineye-searching-for-images-with-image/
15. TinEye Reverse Image Search. (2021). Retrieved 27 December 2021, from https://tineye.com/

16. Terrasi, J. (2019). Ever Wanted to Search for Google Images? Google Has a Tool for You. Retrieved 26 December 2021, from https://www.lifewire.com/what-is-google-images-4585165
17. McCamy, L. (2021). How to reverse image search on Google to find information related to a specific photo. Retrieved 22 December 2021, from https://www.businessinsider.com/google-reverse-image-search
18. Karnila, S., Irianto, S., & Kurniawan, R. (2019). Face Recognition using Content Based Image Retrieval for Intelligent Security. International Journal Of Advanced Engineering Research And Science, 6(1), 91-98. doi: 10.22161/ijaers.6.1.13
19. Google Lens: Όλες οι δυνατότητες του Google μέσω μίας κάμερας - LAB - Σύγχρονο Εργαστήριο Υπολογιστών. (2021). Retrieved 3 January 2022, from https://www.lab.com.gr/google-lens-%CF%84%CE%BF-google-%CE%BC%CE%AD%CF%83%CF%89-%CE%BC%CE%AF%CE%B1%CF%82-%CE%BA%CE%AC%CE%BC%CE%B5%CF%81%CE%B1%CF%82/
20. Shapovalov, Viktor & Shapovalov, Yevhenii & I., Bilyk. (2020). The Google Lens analyzing quality: an analysis of the possibility to use in the educational process.
21. Laukkonen, J. (2021). What is Google Lens, and How Does it Work?. Retrieved 29 December 2021, from https://www.lifewire.com/google-lens-4153383
22. Yu, D., & Deng, L. (2016). Automatic Speech Recognition. London: Springer.
23. Schedl, M., Gómez, E., & Urbano, J. (2014). Music Information Retrieval: Recent Developments and Applications. *Foundations And Trends® In Information Retrieval*, *8*(2-3), 127-261. doi: 10.1561/1500000042
24. About ImageNet. (2022). Retrieved 9 December 2021, from https://www.image-net.org/about.php
25. Machine Learning in Computer Vision. (2019). Retrieved 23 January 2022, from https://fullscale.io/blog/machine-learning-computer-vision/
26. 15 Ways Machine Learning Will Impact Your Everyday Life. (2019). Retrieved 12 January 2022, from https://elitedatascience.com/machine-learning-impact
27. Deep Learning: GoogLeNet Explained. (2020). Retrieved 7 January 2022, from https://towardsdatascience.com/deep-learning-googlenet-explained-de8861c82765
28. Deep Learning: Understand The Inception Module. (2020). Retrieved 13 January 2022, from https://towardsdatascience.com/deep-learning-understand-the-inception-module-56146866e652
29. Szymon Janusz, S. (2014). Retrieved 5 January 2022, from http://www.diva-portal.org/smash/get/diva2:830004/FULLTEXT01.pdf
30. FEDEWA, J. (2021). Howtogeek.com. Retrieved 12 May 2022, from https://www.howtogeek.com/696191/how-to-hum-to-search-for-a-song-using-google/.
31. Neves, Cláudio & Veiga, Arlindo & Sá, Luís & Perdigão, Fernando. (2009). Audio Fingerprinting System for Broadcast Streams.

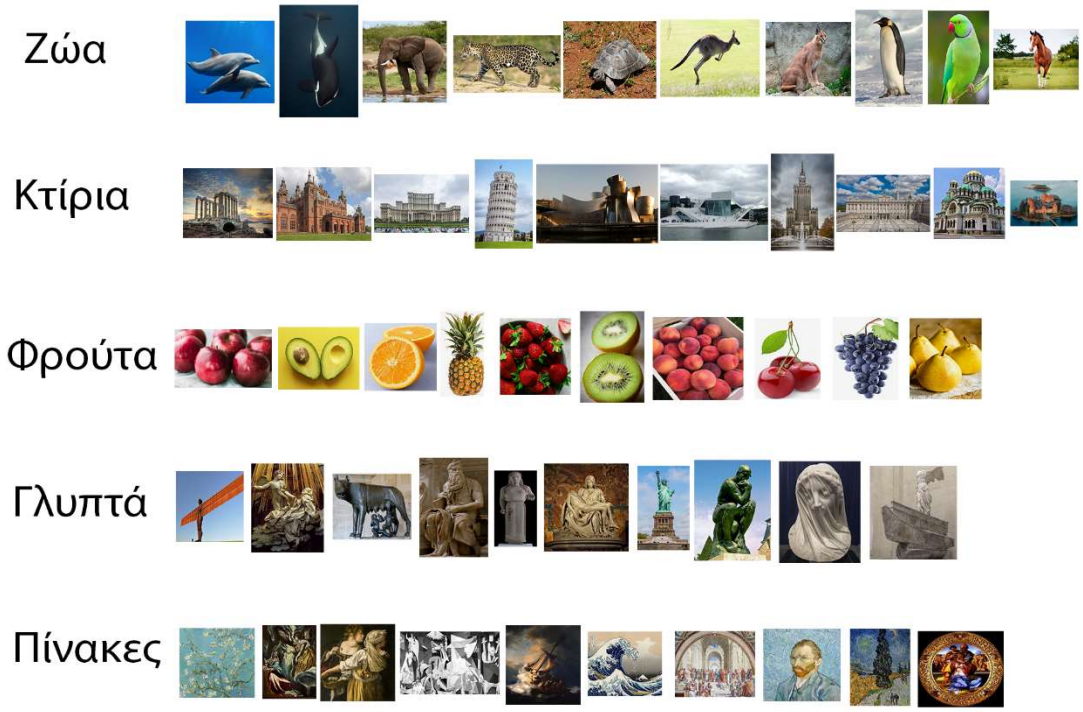
32. KYLE, M. Comb Filtering Explained: What Does a Comb Filter Sound Like? – Audio University. Retrieved 19 February 2022, from https://audiouniversityonline.com/comb-filtering-explained/
33. Cross Correlation. (2019). Retrieved 15 February 2022, from https://www.statisticshowto.com/cross-correlation/
34. Musical Terms and Concepts | SUNY Potsdam. Retrieved 16 March 2022, from https://www.potsdam.edu/academics/crane-school-music/departments-programs/music-theory-history-composition/musical-terms
35. OnMusic Dictionary -. (2015). Retrieved 21 March 2022, from https://dictionary.onmusic.org/
36. Wang, A. (2004). An Industrial-Strength Audio Search Algorithm [Ebook]. Retrieved from http://Wang03-shazam.pdf
37. Mozilla Common Voice. (2022). Retrieved 11 March 2022, from https://commonvoice.mozilla.org/el/speak
38. Artificial Intelligence Applications and Innovations. (2020). IFIP Advances In Information And Communication Technology. doi: 10.1007/978-3-030-49161-1
39. Shazam co-founder: 'We were growing a business in a collapsing market'. (2016). Retrieved 7 March 2022, from https://www.theguardian.com/small-business-network/2016/dec/07/shazam-co-founder-we-were-growing-a-business-in-a-collapsing-market
40. Technology for a voice-enabled world | SoundHound Inc. (2022). Retrieved 1 March 2022, from https://www.soundhound.com/

Παράρτημα

Δείγματα εικόνας

Για την αξιολόγηση των συστημάτων της εικόνας, επιλέξαμε 50 διαφορετικές εικόνες, χωρισμένες σε 5 κατηγορίες (κτίρια, ζώα, πίνακες, γλυπτά, φρούτα). Οι εικόνες περιλαμβάνονται μαζί με τον συνοδευτικό φάκελο.

Η λογική επιλογής των εικόνων βρίσκεται στις χαρακτηριστικές ιδιαιτερότητες που τις απαρτίζουν (χρώμα, υφή, σχήμα) και χωρίζονται σε 5 βασικές κατηγορίες που σχετίζονται με την καθημερινότητα ενός σύγχρονου ανθρώπου. Μία μηχανή αναζήτησης πρέπει να ανταποκρίνεται σε κάθε είδος εικόνας με αξιοπιστία, ανεξάρτητα από το περιεχόμενό του. Η κατηγοριοποίηση βοηθάει στην περίπτωση που κάποια κατηγορία έχει μεγάλη απόκλιση από τις υπόλοιπες ως προς την ποιότητα των αποτελεσμάτων. Έτσι καταλήξαμε σε μία σειρά από αντικείμενα και ζώα που ξεχωρίζουν σε μεγάλο βαθμό μεταξύ τους και καλύπτουν το σύνολο των απαιτήσεων σε ένα σύστημα ανάκτησης.



Δείγματα μουσικής

Για την αξιολόγηση των συστημάτων της μουσικής, επιλέξαμε 25 τραγούδια για να εκτιμήσουμε την μέθοδο αναγνώρισης και ανάκτησης της μουσικής, Ως δείγμα χρησιμοποιήσαμε το πρώτο λεπτό αναπαραγωγής, μέσω του YouTube.

Η λογική στις επιλογές μας στηρίζεται σε τραγούδια που έχουν απήχηση στο κοινό, με σχετικά απλά στην περιγραφή είδη μουσικής (rock, pop κλπ.). Είναι εύκολα ανακτήσιμα για να δοκιμάσουμε την επίδοση των συστημάτων στην πιο απλή τους λειτουργία. Έπειτα εστιάσαμε σε τραγούδια που θα είναι πιο δύσκολα διακριτά (trash, heavy metal). Διευρύναμε τις αναζητήσεις μας με τραγούδια από διάφορες χώρες και λιγότερα γνωστά τραγούδια με μικρότερη απήχηση, για να εκτιμήσουμε τα όρια κάθε εφαρμογής. Τέλος, για να ολοκληρώσουμε την αξιολόγηση μας, επιλέξαμε διασκευές (cover songs). Είναι η πιο σύνθετη μέθοδος αναγνώρισης, κρίνοντας τα ως τα πιο πολύπλοκα σε ανάλυση τραγούδια.

Δείγματα αναγνώρισης μουσικής

1. Bon Jovi - You Give Love A Bad Name
2. AC/DC - Thunderstruck
3. Imagine Dragons – Sharks
4. DJ Tiesto - Insomnia
5. Ethan Bortnick - engravings
6. Gorillaz - Cracker Island ft. Thundercat
7. Tears For Fears - Everybody Wants To Rule The World
8. Billy Joel – Pressure
9. Slipknot - Psychosocial
10. CORPSE - POLTERGEIST! Ft. OmenXIII
11. Moon Walker- The TV Made Me Do It
12. Σ ΑΝΑΖΗΤΩ ΣΤΗ ΣΑΛΟΝΙΚΗ – ΜΗΤΡΟΠΑΝΟΣ
13. Παντελής Παντελίδης - Γίνεται
14. Barbara Pravi – Voilà
15. Måneskin - Torna a casa
16. L'Arc-en-Ciel - Driver's High
17. Addicted to You - Utada Hikaru
18. The Real Folk Blues • Mai Yamane • Yoko Kanno
19. Joe Cocker - You Can Leave Your Hat On (Cover by Anna Vishnevskaia)
20. Michael Jackson - They Don't Care About Us (Cover by Matty Carter + Ariel)
21. Loveless - MIDDLE OF THE NIGHT

22. Glimpse of Us - Joji cover by Alex Porat
23. Valiant Hearts - Devil Trigger (Official Lyric Video) [Casey Edwards Cover]
24. IMAGINE DRAGONS: BELIEVER (Chase Holfelder Cover)
25. Do I Wanna Know (Arctic Monkeys cover) - Katerine Duska

Δείγματα αναγνώρισης ομιλίας

Η αξιολόγηση έχει γίνει με την χρήση των παρακάτω 25 φωνητικών ερωτημάτων μέσω των εφαρμογών μηχανή αναζήτησης.

Δοκιμάσαμε τις βασικές φωνητικές αναζητήσεις που επιδεικνύουν την χρησιμότητα και ευελιξία των συγκεκριμένων εφαρμογών στην καθημερινότητα (αριθμητικές πράξεις, καιρός ανά περιοχή, εύρεση κοντινότερων καταστημάτων και αναζητήσεις τραγουδιών). Συνεχίσαμε με λίγο πιο εξειδικευμένα ερωτήματα με επιστημονική, ιστορική, καλλιτεχνική βάση για να εξετάσουμε τις ικανότητες των εφαρμογών σε περιπτώσεις που εμπλέκονται πολύπλοκοι όροι, πρόσωπα και τοποθεσίες.

1. Weather in Athens
2. What is the earth’s area total
3. 1 + 10
4. 10 yen to euro conversion
5. How glass is made
6. Nearest petrol station
7. History of Japan
8. Thank you in Italian
9. Directions to Athens airport
10. Best Tv series of 2020
11. Pablo Picasso
12. Leaning Tower of Pisa
13. 15 x 15
14. 13-3
15. 18/2
16. Date of the Battle of Waterloo
17. Translate in Russian
18. Pictures with boats
19. Billy Joel – Pressure

20. Loveless - MIDDLE OF THE NIGHT
21. IMAGINE DRAGONS - BELIEVER
22. Best songs of the 20th century
23. Artificial intelligence
24. Image and music retrieval
25. Halo Theme Song