Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Αυτόματη προσθήκη επισημάνσεων για πρόσωπα και αντικείμενα σε πολυμεσικό περιεχόμενο

Διπλωματικη Εργασια

του

Μαρινέλλη Δ. Γεώργιου

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2018

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής
Εργαστήριο Διαχείρισης και Βέλτιστου Σχεδιασμού Δικτύων Τηλεματικής

Αυτόματη προσθήκη επισημάνσεων για πρόσωπα και αντικείμενα σε πολυμεσικό περιεχόμενο

Διπλωματικη Εργασια

του

Μαρινέλλη Δ. Γεώργιου

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 17η Οκτωβρίου, 2018.

.....	.....	.....
Συμεών Παπαβασιλείου	Ιωάννα Ρουσσάκη	Θεοδώρα Βαρβαρίγου
Καθηγητής Ε.Μ.Π.	Επίκ. Καθηγήτρια Ε.Μ.Π.	Αν. Καθηγήτρια Ε.Μ.Π.

Αθήνα, Οκτώβριος 2018

.....

Μαρινελλης Γεωργιος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικος Υπολογιστών Ε.Μ.Π.

Copyright © -- All rights reserved Μαρινέλλης Γεώργιος, 2018.
Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Στις μέρες μας, ο όλο και μεγαλύτερος όγκος πολυμεσικών δεδομένων που δημιουργείται καθιστά τα δεδομένα μη διαχειρίσιμα και αναξιοποίητα. Εμφανίζεται έτσι η ανάγκη της επισήμανσης και ταξινόμησης του πολυμεσικού αυτού περιεχομένου για την καλύτερη αξιοποίησή του τόσο από απλούς χρήστες όσο και από άλλους επαγγελματικούς κλάδους. Η προσθήκη αυτόματων επισημάνσεων στο περιεχόμενο είναι κάτι παραπάνω από απαραίτητη καθώς ,όπως είπαμε, ο όγκος των δεδομένων δε αφήνει χώρο και χρόνο για χειροκίνητες προσθήκες.

Στα πλαίσια της συγκεκριμένης εργασίας, επιλέξαμε να παράγουμε αυτόματες επισημάνσεις χρησιμοποιώντας αλγορίθμους του κλάδου της όρασης των υπολογιστών. Πιο συγκεκριμένα υλοποιήθηκε μια εφαρμογή η οποία πραγματοποιεί ανίχνευση και αναγνώριση προσώπων και αντικειμένων. Στο κείμενό παρουσιάζουμε όλες τις σύγχρονες μεθόδους για την ανίχνευση και αναγνώριση προσώπων και αντικειμένων, αναλύουμε τα πλεονεκτήματα και τα μειονεκτήματα της καθεμίας καθώς και τους λόγους που επιλέξαμε να χρησιμοποιήσουμε συγκεκριμένες από αυτές. Χρησιμοποιούμε επίσης και μια παραλλαγή της μεθόδου Local Binary Patterns Histogram για την αναγνώριση προσώπων η οποία σχεδιάστηκε και τροποποιήθηκε από εμάς.

Στο τέλος γίνεται μια αποτίμηση της παραλλαγμένης αυτής μεθόδου χρησιμοποιώντας κατάλληλες μετρικές πάνω σε εικόνες από τις βάσεις εικόνων προσώπων AT&T Facedatabe, Yale Facedatabase A, Extended Yale Facedatabase B και MyLucce Facedatabase. Οι εικόνες για την τελευταία βάση συλλέχθηκαν από εμάς. Τα αποτελέσματα που συλλέξαμε παρουσιάζουν ενδιαφέρον και δίνουν μια συνολικότερη αντίληψη πάνω στο γενικότερο πρόβλημα της αναγνώρισης προσώπων.

Λέξεις Κλειδιά

artificial intelligence, machine learning, convolutional neural networks, object detection, face detection, face recognition, recognition systems, viola jones, LBPH, Mobilenet, SSD, Faster R-CNN, cascade classifier, Local Binary Patterns

Abstract

Nowadays, the continuously growing amount of multimedia data makes them unmanagable and useless. Shows up thus the need to mark and classify this multimedia content for better use by both simple and professional users. The extraction of automatic annotations though from this content is an essential process because the volume of data is such that there is no space to add manual annotations.

In the context of this thesis, we decided to extract automatic annotations using computer vision algorithms. More specifically, we implemented an application which performs face and object detection and recognition. In the following text we present all the state-of-the art methods for face and object detection and recognition, we analyse each method's pros and cons and we explain the reasons we choosed to use the particular ones. We also use a modified version of the Local Binary Patterns Histogram method for face recognition.

Finally, we perform an evaluation of the modified method using well known face databases such as AT&T Facedatabase, Yale Facedabase A, Extended Yale Facedatabase B and MyLucce Facedatabase. The last one was created by us. The results are promising and provide a complete overview over the general problem of face recognition.

Keywords

artificial intelligence, machine learning, convolutional neural networks, object detection, face detection, face recognition, recognition systems, viola jones, LBPH, Mobilenet, SSD, Faster R-CNN, cascade classifier, Local Binary Patterns

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Συμεών Παπαβασιλείου για την ευκαιρία που μου έδωσε να ασχοληθώ με το συγκεκριμένο τομέα της επιστήμης των υπολογιστών στο Εργαστήριο Διαχείρισης και Βέλτιστου Σχεδιασμού Δικτύων Τηλεματικής, καθώς και για τη θετική συμβολή του καθόλη τη διάρκεια των σπουδών μου.

Η εκπόνηση της διπλωματικής αυτής εργασίας, δε θα μπορούσε να ολοκληρωθεί χωρίς την καίρια συμβολή του επί χρόνια συναδέλφου μου και υποψήφιου διδάκτορα Γιώργου Μήτση καθώς και ολόκληρου του εργαστηρίου Διαχείρισης και Βέλτιστου Σχεδιασμού Δικτύων Τηλεματικής και του εργαστηρίου Ευφυών Συστημάτων, Περιεχομένου και Αλληλεπίδρασης.

Τέλος ένα μεγάλο ευχαριστώ στην οικογένειά μου και στους φίλους μου για τη συνεχή τους στήριξη όλο αυτό το διάστημα των σπουδών μου και για τα ωραιότερα συναισθήματα και αναμνήσεις που μου προσφέρουν όλα αυτά τα χρόνια.

Μαρινέλλης Γιώργος

Περιεχόμενα

Περίληψη **v**

Abstract **vii**

Ευχαριστίες **ix**

Περιεχόμενα **xii**

Κατάλογος σχημάτων **xiii**

Κατάλογος πινάκων **xv**

Κατάλογος Αλγορίθμων **xvii**

1 Εισαγωγή **1**
1.1 Ανίχνευση προσώπων και αντικειμένων 1
1.2 Αναγνώριση προσώπων 2
1.3 Συνεισφορά της διπλωματικής 2
1.4 Οργάνωση κειμένου 2

2 Ανίχνευση προσώπων **5**
2.1 Συνοπτική παρουσίαση βασικών μεθόδων αναγνώρισης προσώπων 6
2.1.1 Η μέθοδος Rowley, Baluja και Kanade [32] 6
2.1.2 Η μέθοδος SNow(Sparse Network of Winnows) [17] 7
2.2 Η μέθοδος Viola-Jones 7
2.2.1 Χαρακτηριστικά τύπου Haar 8
2.2.2 Ολοκλήρωμα εικόνας 9
2.2.3 Ο αλγόριθμος AdaBoost 13
2.2.4 Χρήση ενός Cascade Classifier 15
2.3 Άλλες μέθοδοι που έχουν χρησιμοποιηθεί 17

3 Ανίχνευση αντικειμένων **19**
3.1 Εισαγωγή 19
3.2 Ανίχνευση αντικειμένων χρησιμοποιώντας Περιφερειακά Συνελικτικά Νευρωνικά Δίκτυα (R-CNN) 21
3.2.1 R-CNN 21
3.2.2 Fast R-CNN 21
3.2.3 Faster R-CNN 22
3.3 Ανίχνευση αντικειμένων χρησιμοποιώντας ανιχνευτές μονής λήψης 23

3.3.1 Single Shot Multibox Detector (SSD) 25
3.3.2 Η μέθοδος You Only Look Once (YOLO) 25
3.4 Άλλες μέθοδοι για την ανίχνευση αντικειμένων 26
3.4.1 Πλήρως Συνελικτικά Δίκτυα βασιζόμενα σε μια περιοχή της εικόνας (Region-based Fully Convolutional Networks - R-FCN [10]) 26
3.4.2 Δίκτυα Πυραμίδας Χαρακτηριστικών (Feature Pyramid Networks FPN [22]) 26
3.5 Συνολική αποτίμηση 28

4 Αναγνώριση προσώπων 31
4.1 Η μέθοδος EigenFaces 32
4.1.1 Περιγραφή του αλγορίθμου 32
4.2 Η μέθοδος FisherFaces 34
4.2.1 Περιγραφή του αλγορίθμου 34
4.3 Η μέθοδος Local Binary Patterns Histograms (LBPH) 36
4.3.1 Τυπική υλοποίηση 38
4.3.2 Υλοποίηση με τη χρήση του αλγορίθμου k-Nearest Neighbor 38

5 Ανάλυση αποτελεσμάτων 41
5.1 Υλοποίηση 41
5.2 Πειραματική μεθοδολογία 41
5.2.1 Βάσεις δεδομένων 41
5.2.2 Πρωτόκολλο αξιολόγησης 42
5.3 Αποτελέσματα 43

6 Επίλογος 47
6.1 Συμπεράσματα 47
6.2 Προτάσεις για μελλοντική έρευνα 48

Βιβλιογραφία 49

Κατάλογος σχημάτων

2.1 α) σε επίπεδο εικόνας, β) σε επίπεδο περιγράμματος γ) σε επίπεδο παραθύρου. 6
2.2 Χαρακτηριστικά τύπου Haar 8
2.3 Υπολογισμός ολοκληρώματος εικόνας 11
2.4 Υπολογισμός 45°-περιστραμμένου ολοκληρώματος εικόνας 12
2.5 Τα επιμέρους βήματα ενός Cascade Classifier 15

3.1 Η μέθοδος R-CNN 21
3.2 Η μέθοδος Fast R-CNN 22
3.3 Η μέθοδος Faster R-CNN 22
3.4 Regional Proposal Network 23
3.5 Regional Proposal Network 23
3.6 Η μέθοδος Single Shot Multibox Detector 25
3.7 Η τεχνική R-FCN 26
3.8 Το FPN του σχήματος τροφοδοτεί έναν Object Detector 27
3.9 α) Χρόνος εκτέλεσης κ mAP β) Μέγεθος αντικειμένου γ) Μνήμη 28

4.1 EigenFaces για 10 πρόσωπα της βάσης δεδομένων AT & T Facedatabase . . 34
4.2 FisherFaces για 16 πρόσωπα της βάσης δεδομένων Yale Facedatabase A . . . 35
4.3 Υπολογισμός των Local Binary Patterns 37
4.4 α) αρχική εικόνα σε grayscale, β) ύστερα από το μετασχηματισμό LBP . . . 37
4.5 Εξαγωγή ιστογραμμάτων από κάθε υποπεριοχή του πλέγματος 37
4.6 Extended Local Binary Patterns 38
4.7 Η διαφορά μεταξύ 1-Nearest Neighbor και k-Nearest Neighbor 39

5.1 Η μέθοδος k-fold Cross Validation 43
5.2 k-4 cross validation 44
5.3 k-10 cross validation 44
5.4 k-4 cross validation 45
5.5 k-10 cross validation 45

Κατάλογος πινάκων

2.1 Πλήθος χαρακτηριστικών Haar ανά παράθυρο 10
2.2 Μείωση του κόστους υπολογισμού χαρακτηριστικών με 3 ορθογώνια 13

Κατάλογος Αλγορίθμων

3.1 Regional Proposal Methods 24
3.2 Single Shot Method 24

Κεφάλαιο 1

Εισαγωγή

Στις μέρες μας, η τεχνητή νοημοσύνη είναι ένας ταχύτατα αναπτυσσόμενος κλάδος της επιστήμης των υπολογιστών. Η όραση υπολογιστών είναι ένα επιστημονικό πεδίο της τεχνητής νοημοσύνης που δε θα μπορούσε να μείνει ανεπηρέαστο από αυτή την εξέλιξη.

Η όραση των υπολογιστών (από και στο εξής ΟτΥ) ασχολείται με την απόκτηση, ανάλυση και κατανόηση εικόνων, βίντεο και γενικά πολυμεσικού περιεχομένου πολλών διαστάσεων από τον πραγματικό κόσμο. Έχει ως σκοπό να δώσει στα υπολογιστικά συστήματα μια εποπτεία και κατανόηση του τετραδιάστατου πραγματικού κόσμου.

Για να επιτευχθεί ο άνωθεν σκοπός χρειάζεται να αυτοματοποιηθεί μέσω μιας υπολογιστικής αλγοριθμικής διαδικασίας η μέθοδος της ανθρώπινης όρασης. Έτσι η πραγματική ν-διάσταση αναπαράσταση που απεικονίζει και αναγνωρίζει ο ανθρώπινος εγκέφαλος, αναπαρίσταται με συμβολικό και αριθμητικό τρόπο.

Σε ένα υπολογιστικό σύστημα, μια εικόνα διαθέτει μια ψηφιακή αναπαράσταση. Στην πιο απλή και συνηθισμένη της μορφή μιας δισδιάστατης εικόνας, αναπαρίσταται με ένα ψηφιακό σήμα δυο διαστάσεων. Η τιμή του σήματος σε κάθε σημείο του επιπέδου αφορά την τιμή του χρώματος της εικόνας στη θέση αυτή. Τα σημεία που αποτελούν το σύνολο μια εικόνας είναι ευρέως γνωστά ως εικονοστοιχεία (pixels).

1.1 Ανίχνευση προσώπων και αντικειμένων

Η ανίχνευση προσώπων και γενικότερα αντικειμένων σε μια εικόνα συνίσταται στη διαδικασία εύρεσης των χαρακτηριστικών εκείνων που καθιστούν ένα συγκεκριμένο αντικείμενο μέλος μια κλάσης αντικειμένων. Είναι μια καθημερινή, αυτοματοποιημένη και τετριμμένη διαδικασία για τον άνθρωπο. Ο ανθρώπινος εγκέφαλος είναι εκπαιδευμένος με τέτοιο τρόπο ώστε να μπορεί να αναγνωρίζει αντικείμενα ακαριαία. Η αναγνώριση αφορά την αναγνώριση μεμονωμένων και συγκεκριμένων αντικειμένων αλλά γενικότερα και την αναγνώριση της κλάσης η των κλάσεων στην/στις οποίες ανήκει Αντίθετα, η διαδικασία αυτή δεν εκτελείται το ίδιο εύκολα και από ένα υπολογιστικό σύστημα. Είναι απαραίτητη η εξαγωγή ενός μεγάλου πλήθους χαρακτηριστικών για την ταυτοποίηση της κλάσης ή πιθανών κλάσεων του αντικειμένου. Ο υπολογισμός όλων αυτών των χαρακτηριστικών είναι μια αρκετά χρονοβόρα -σε κύκλους του επεξεργαστή- διαδικασία. Για την επίτευξή της γίνεται χρήση τεχνικών μηχανικής μάθησης και των νευρωνικών δικτύων

1.2 Αναγνώριση προσώπων

Η αναγνώριση προσώπων αποτελεί ένα επιπλέον στάδιο της ανίχνευσης ενός προσώπου, όπου ένα εντοπισμένο πρόσωπο επιχειρείται να ταυτιστεί με κάποιο ήδη γνωστό πρόσωπο Είναι μια διαδικασία που εκτελείται και πάλι σχεδόν ακαριαία από τον ανθρώπινο εγκέφαλο αλλά εν αντίθεση υπολογιστικά χρειάζεται αρκετή υπολογιστική ισχύς και προεκπαιδευση των συστημάτων με τα πρόσωπα που θέλει κανείς να αναγνωρίσει

1.3 Συνεισφορά της διπλωματικής

Στην παρούσα διπλωματική δημιουργήσαμε μια διαδικτυακή εφαρμογή η οποία μπορεί να παράγει αυτόματες επισημάνσεις στο πολυμεσικό περιεχόμενο (video) που δέχεται ως είσοδο. Οι τεχνικές για την εξαγωγή αυτών των επισημάνσεων χρησιμοποιούν ως βάση state-of-art τεχνολογίες όσον αφορά την ανίχνευση και αναγνώριση προσώπων και κλάσεων αντικειμένων σε μια εικόνα, προσαρμοσμένες στο περιβάλλον ενός βίντεο Επιπρόσθετα, στο κομμάτι της αναγνώρισης προσώπων , επιχειρήσαμε να βελτιώσουμε την ακρίβεια και την ταχύτητα εξαγωγής της πρόβλεψης, τροποποιώντας ένα μέρος της μεθόδου Linear Binary Pattern Histograms.

Στο πλαίσιο της παρούσας διπλωματικής, επιχειρείται ,ως εκ τούτου, μια παρουσίαση των υπάρχουσων τεχνικών για την αναγνώριση και ανίχνευση προσώπων και αντικειμένων και μια συγκριτική αξιολόγηση των τεχνικών για την αναγνώριση προσώπων.

1.4 Οργάνωση κειμένου

Το παρόν κείμενο έχει την εξής δομή:

Κεφάλαιο 2:
Στο κεφάλαιο αυτό θα κάνουμε μια γενική επισκόπηση των μεθόδων που χρησιμοποιούνται για την ανίχνευση κλάσεων αντικειμένων. Θα γίνει μια παρουσίαση του αλγορίθμου των Viola-Jones. και ο τρόπος με τον οποίο τον χρησιμοποιήσαμε. Πρόκειται για τον αλγόριθμο πάνω στον οποίο βασίζεται η ανίχνευση προσώπων στις εικόνες και συνεπακόλουθα στα βίντεο. Θα συζητήσουμε πως σχετίζεται το αποτέλεσμα της μεθόδου με την μετέπειτα αναγνώριση των προσώπων.

Κεφάλαιο 3:
Θα κάνουμε μια γενική επισκόπηση των σύγχρονων τεχνικών ανίχνευσης κλάσεων αντικειμένων με τη χρήση νευρωνικών δικτύων. Θα συγκρίνουμε τα μεταξύ τους πλεονεκτήματα και μειονεκτήματα, ενώ τέλος θα μιλήσουμε αναλυτικότερα για τους ανιχνευτές μονής λήψης και τους λόγους που επιλέχθηκαν να χρησιμοποιηθούν στην υλοποίηση της παρούσας διπλωματικής.

Κεφάλαιο 4:
Θα παρουσιάσουμε τις μεθόδους που χρησιμοποιούνται για την αναγνώριση προσώπων. Θα αναλύσουμε τον τρόπο και τους λόγους για τον οποίο επιλέξαμε να τροποποιήσουμε μία από αυτές.

Κεφάλαιο 5:

Στο κομμάτι αυτό θα δούμε κάποια συγκριτικά αποτελέσματα των τεχνικών αναγνώρισης προσώπων. Θα αναλύσουμε τα σημεία και τους λόγους διαφοροποίησης από άλλες μεθόδους και θα προταθούν τροποποιήσεις που πιθανόν να οδηγήσουν σε ποιο εύστοχα αποτελέσματα.

Κεφάλαιο 6:

Τέλος θα κάνουμε μια ανασκόπηση της διπλωματικής. Θα αναφερθούμε σε διάφορες άλλες τεχνικές και θα γίνει μια πρόταση για παραπέρα έρευνα

Κεφάλαιο 2

Ανίχνευση προσώπων

Η επιστημονική περιοχή της ανίχνευσης αντικειμένου μπορεί να χωριστεί σε τρεις επιμέρους ανεξάρτητες διαδικασίες.

- Εντοπισμός πιθανών περιοχών αντικειμένου/ων
- Ταξινόμηση αντικειμένου
- Προσδιορισμός θέσης αντικειμένου

Η διαδικασία του εντοπισμού των περιοχών αντικειμένων αφορά την εύρεση των περιοχών εντός μιας εικόνας όπου είναι πιθανό να υπάρχουν αντικείμενα. Πιο συγκεκριμένα, αφορά τον προσδιορισμό του συνόλου των εικονοστοιχείων εκείνων τα οποία με μια γρήγορη επεξεργασία δίνουν μεγάλη πιθανότητα να εμπεριέχουν ένα αντικείμενο.

Η ταξινόμηση αντικειμένου αφορά τον προσδιορισμό της κλάσης του αντικειμένου που έχει προσδιοριστεί. Η διαδικασία της ταξινόμησης απαιτεί την εξαγωγή διάφορων χαρακτηριστικών από την εικόνα και επεξεργασία αυτών με κάποιο αλγόριθμο ταξινόμησης για τον συμπερασμό της κλάσης του αντικειμένου

Ο τελικός προσδιορισμός της θέσης του αντικειμένου εντός της εικόνας μπορεί να γίνει σε τρία επίπεδα: α) σε επίπεδο εικόνας, β) σε επίπεδο περιγράμματος γ) σε επίπεδο παραθύρου (βλ. Σχήμα 5.5).

Στη συνέχεια του κειμένου της παρούσας διπλωματικής όταν αναφερόμαστε στον όρο ανίχνευση αντικειμένου θα αναφερόμαστε στην ανίχνευση σε επίπεδο παραθύρου.

Η ανίχνευση προσώπου αποτελεί μια υποπεριοχή της ανίχνευσης αντικείμενου. Επομένως η μεθοδολογία που ακολουθείται είναι όμοια με αυτή που περιγράψαμε ανωτέρω. Ειδικότερα, η ανίχνευση προσώπου παρουσιάζει μια ευκολία σε σχέση με την ανίχνευση κλάσης αντικειμένου με την έννοια ότι κάθε πρόσωπο έχει συγκεκριμένα και καθολικά χαρακτηριστικά στο σύνολο των ανθρώπων. Συνεπώς, η ανίχνευση προσώπου δεν αφορά το γενικότερο πρόβλημα του εντοπισμού της κλάσης ενός αντικειμένου αλλά το υποπρόβλημα του εντοπισμού -εντός μια εικόνας- μιας και μόνον κλάσης αντικειμένου (του προσώπου) με αυστηρώς προσδιορισμένα χαρακτηριστικά.

Παρακάτω θα κάνουμε μια σύντομη περιγραφή των βασικότερων μεθόδων αναγνώρισης προσώπου που έχουν χρησιμοποιηθεί έως σήμερα.

Σχήμα 2.1: α) σε επίπεδο εικόνας, β) σε επίπεδο περιγράμματος γ) σε επίπεδο παραθύρου.

2.1 Συνοπτική παρουσίαση βασικών μεθόδων αναγνώρισης προσώπων

2.1.1 Η μέθοδος Rowley, Baluja και Kanade [32]

Το 1998 οι *Rowley, Baluja και Kanade* περιέγραψαν μια μέθοδο για την αναγνώριση προσώπων βασισμένη στο συνδυασμό αποτελεσμάτων από νευρωνικά δίκτυα. Στην εικόνα δοκιμάζεται ένα σετ από φίλτρα βασισμένα σε νευρωνικά δίκτυα. Στη συνέχεια τα αποτελέσματα επεξεργάζονται από ένα διαμεσολαβητή και συνδυάζονται Πιο συγκεκριμένα η μέθοδος αποτελείται από 2 βήματα:

Βήμα 1

Στο βήμα αυτό επιλέγεται αρχικά ένα παράθυρο μεγέθους 20x20 από την αρχική εικόνα. Το παράθυρο αυτό υπόκειται μια προεργασία για την ισοστάθμιση της έντασης του χρώματος, την αντίθεση κ.α. και στη συνέχεια δίνεται ως είσοδο σε ένα ή και περισσότερα (προεκπαιδευμένα) νευρωνικά δίκτυα. Τα δίκτυα αυτά υπολογίζουν κάποια συγκεκριμένα χαρακτηριστικά που αφορούν τα πρόσωπα (πχ μύτη, μάτια, στόμα) και αποφασίζουν αν στο παράθυρο αυτό υπάρχει ή όχι πρόσωπο. Η εικόνα σαρώνεται ολόκληρη σε παράθυρα μεγέθους 20x20 τα οποία ακολουθούν την παραπάνω ροή. Το μέγεθος του παραθύρου αυξάνεται σταδιακά με ένα προκαθορισμένο βάρος για εντοπίσει πρόσωπα μεγαλύτερα του προαναφερθέντος μεγέθους.

Βήμα 2

Το παρόν βήμα αποτελείται από δύο υπο-βήματα. Αφενός συλλέγονται τα επιμέρους αποτελέσματα του κάθε νευρωνικού δικτύου και με βάση μια ευριστική απορρίπτονται κάποιες λανθασμένες (false positive) προβλέψεις. Αφετέρου, συνδυάζονται τα αποτελέσματα από όλα τα επιμέρους νευρωνικά δίκτυα και χρησιμοποιώντας κάποια άλλη ευριστική παράγονται οι τελικές προβλέψεις του αλγορίθμου για τα πρόσωπα.

Βάση μετρήσεων η παραπάνω μέθοδος μπορεί να επιτύχει ποσοστά επιτυχούς αναγνώρισης προσώπων μεταξύ $77.9\% - 90.3\%$ -με ένα ασφαλή αριθμό λανθασμένων προβλέψεων- ανάλογα με τις ευριστικές μεθόδους που χρησιμοποιούνται.

2.1.2 Η μέθοδος SNow(Sparse Network of Winnows) [17]

Αργότερα, το *2000* οι *Yang, Roth* και *Ahuja* πρότειναν μια νέα μέθοδο βασισμένη στην αρχιτεκτονική εκμάθησης *SNoW* [31] [6]. Η αρχιτεκτονική αυτή -*SNoW*- (Sparse Network of Winnows) είναι στην πραγματικότητα ένα δίκτυο από γραμμικές συναρτήσεις που χρησιμοποιούν τον κανόνα Winnow [23] και επιτρέπει την ανίχνευση προσώπων με διαφορετικά χαρακτηριστικά, σε διαφορετικές θέσεις και στάσεις και με διαφορές στις φωτιστικές συνθήκες. Το σύστημα αυτό είναι ειδικά σχεδιασμένο για μάθηση σε τομείς στους οποίους ο δυνητικός αριθμός των χαρακτηριστικών που συμμετέχουν στις αποφάσεις είναι πολύ μεγάλο, αλλά μπορεί να είναι άγνωστο a priori.

Η διαδικασία της αναγνώρισης προσώπου είναι όμοια με αυτή των *Rowley, Baluja και Kanade* όπου η εικόνα σαρώνεται αρχικά σε παράθυρα μεγέθους 20x20 και το μέγεθος αυτό αυξάνεται σε κάθε επανάληψη κατά μια σταθερή, προκαθορισμένη τιμή (1.2). Η διαδικασία αυτή επαναλαμβάνεται 10 φορές. Σε κάθε επανάληψη το τρέχον παράθυρο από την εικόνα δίνεται ως είσοδο στο προεκπαιδευμένο δίκτυο SNoW το οποίο λαμβάνει την απόφαση για την ύπαρξη ή όχι προσώπου υπολογίζοντας κάθε φορά ένα διαφορετικό αριθμό χαρακτηριστικών.

Η μέθοδος αυτή μπορεί να επιτυγχάνει ποσοστά επιτυχίας κοντά στο 93%.

2.2 Η μέθοδος Viola-Jones

Ο αλγόριθμος των Viola-Jones [41] υπήρξε τομή στο πεδίο της ανίχνευσης αντικειμένων σε εικόνες. Ουσιαστικά δημιούργησε την πρώτη υποδομή ανίχνευσης αντικειμένων η οποία παρήγαγε ανταγωνιστικά αποτέλεσμα σε πραγματικό χρόνο. Παρόλο που σχεδιάστηκε ώστε να μπορεί να εκπαιδευτεί με σκοπό να αναγνωρίζει οποιαδήποτε κλάση αντικειμένων, ουσιαστικά ο αρχικός σχεδιασμός του αλγορίθμου έγινε με γνώμονα την αναγνώριση προσώπων σε εικόνες. Εν τέλει εκεί εντοπίζεται και το μεγαλύτερο ποσοστό της χρήσης του.

Τα βασικά χαρακτηριστικά του ανωτέρω αλγορίθμου είναι:

Αξιοπιστία
: Ο αλγόριθμος έχει πάντα υψηλό ποσοστό σωστών ανιχνεύσεων (true-positives) και χαμηλό ποσοστό λάθος ανιχνεύσεων (false-positives)

Ταχύτητα πραγματικού χρόνου
: Σε εφαρμογές πραγματικού περιβάλλοντος επεξεργάζονται τουλάχιστον 2 εικόνες (frames) ανά δευτερόλεπτο

Αναγνώριση προσώπων
: Ο αλγόριθμος έχει στόχο να ανιχνεύει πρόσωπα από μη-πρόσωπα. Δεν έχει σκοπό να αναγνωρίζει τα πρόσωπα που ανιχνεύει

Για να το πετύχει αυτό ο αλγόριθμος χρησιμοποιεί κατά σειρά τα παρακάτω στάδια τα οποία θα αναλύσουμε εκτενέστερα στη συνέχεια:

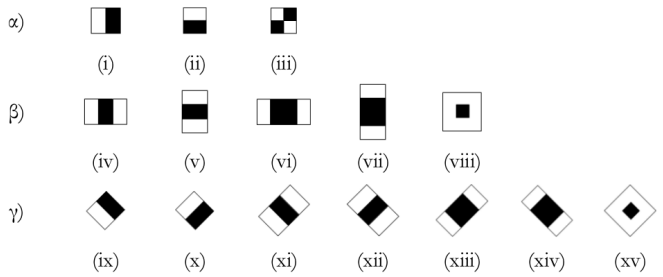
- Επιλογή χαρακτηριστικών τύπου Haar (Haar features selection)
- Κατασκευή ολοκληρώματος εικόνας (Integral image creation)
- Εκπαίδευση του αλγορίθμου AdaBoost (AdaBoost training)
- Χρήση ενός διαδοχικά διασυνδεδεμένου ταξινομητή (Cascade classifier)

2.2.1 Χαρακτηριστικά τύπου Haar

Για να ανιχνεύσουμε αντικείμενα σε εικόνες απαιτείται κατάλληλη επεξεργασία και αναπαράσταση του περιεχομένου τους. Για την αναπαράσταση του περιεχομένου της εικόνας στη μέθοδο που εξετάζουμε, χρησιμοποιούμε τα χαρακτηριστικά τύπου Haar, τα οποία προκύπτουν από την εφαρμογή του μετασχηματισμού Wavelet σε μια εικόνα με χρήση των συναρτήσεων τύπου Haar [41].

Η χρησιμοποίηση των συναρτήσεων Haar στο μετασχηματισμό Wavelet ξεκινά από την παρατήρηση ότι η τιμή της φωτεινότητας κάθε εικονοστοιχείου επηρεάζεται έντονα από τις αλλαγές στο φωτισμό της σκηνής [26]. Αυτή η αλλαγή όμως, επηρεάζει ομοιόμορφα όλα τα pixel της εικόνας. Έτσι, η τιμή μιας συνάρτησης που εξετάζει τη μέση διαφορά ανάμεσα σε δύο ή τρεις περιοχές της ίδιας εικόνας, θα παραμένει σε μεγάλο βαθμό ανεπηρέαστη. Χρησιμοποιώντας, λοιπόν, τις συναρτήσεις Haar, η διαδικασία της ανίχνευσης αντικειμένων δε θα επηρεάζεται από τις διαφορές στη φωτεινότητα από εικόνα σε εικόνα.

Οι συναρτήσεις Haar υπολογίζουν τη διαφορά ανάμεσα στους μέσους όρους των τιμών των εικονοστοιχείων δύο (ή τριών) περιοχών. Ας θεωρήσουμε τη συνάρτηση Haar που παριστάνεται με το ορθογώνιο a.i από το Σχήμα 2.2. Υπολογίζεται ο μέσος όρος των εικονοστοιχείων που βρίσκονται μέσα στο άσπρο ορθογώνιο, καθώς και αυτών που βρίσκονται μέσα στο μαύρο ορθογώνιο. Έπειτα, ο μέσος όρος του μαύρου ορθογωνίου αφαιρείται από τον μέσο όρο του άσπρου. Η τιμή που προκύπτει αποτελεί την τιμή του Haar χαρακτηριστικού.



Σχήμα 2.2: Χαρακτηριστικά τύπου Haar

Εφαρμόζοντας τον μετασχηματισμό Wavelet με τη συναρτησιακή βάση Haar, προκύπτει ένας περιορισμένος αριθμός χαρακτηριστικών [26]. Στο μονοδιάστατο μετασχηματισμό, η απόσταση ανάμεσα σε δύο γειτονικά κυματίδια (wavelets), σε επίπεδο n , θα είναι $2n$. Η απόσταση αυτή είναι πολύ μεγάλη, κι έτσι δεν λαμβάνουμε όσες πληροφορίες θέλουμε από μια εικόνα ώστε να την περιγράψουμε λεπτομερώς. Για να έχουμε, λοιπόν, μια πιο λεπτομερή, χωρικά, αναπαράσταση του περιεχομένου της εικόνας χρειαζόμαστε ένα σύνολο από πλεονάζουσες συναρτήσεις βάσης. Για να το πετύχουμε αυτό, εφαρμόζουμε τις συναρτήσεις Haar με μεταξύ τους απόσταση ένα εικονοστοιχείο κάθε φορά. Έτσι, θα έχουμε μια πολύ πιο πυκνή

αναπαράσταση. Επίσης, στο μετασχηματισμό Wavelet, το μέγεθος των συναρτήσεων Haar, κανονικά διπλασιάζεται σε κάθε επανάληψη. Για να αυξήσουμε ακόμα περισσότερο την λαμβανόμενη πληροφορία από την εικόνα, ορίζουμε ότι το μέγεθος των συναρτήσεων Haar θα αυξάνει κάθε φορά κατά ένα μόνο εικονοστοιχείο. Έτσι, το σύνολο των χαρακτηριστικών Haar σε μία εικόνα γίνεται υπερπολλαπλάσιο του αρχικού. Αυξάνουμε, δηλαδή, την ποσότητα της πληροφορίας που αντλούμε από μια εικόνα, αυξάνοντας τα χαρακτηριστικά τύπου Haar που θα υπολογιστούν σε αυτήν.

Τα κλασσικά Haar χαρακτηριστικά φαίνονται στο Σχήμα 2.2.a (Edge features) [26]. Είναι σχετικά απλά και μπορούν να εντοπίσουν ακμές οριζόντια και κατακόρυφα καθώς και διαγώνιες γραμμές. Για να μπορέσουμε να αναπαραστήσουμε γραμμές, ράβδους και τετράγωνα καλύτερα, προσθέτουμε τα χαρακτηριστικά που φαίνονται στο Σχήμα 2.2.b (Line features)(τα χαρακτηριστικά iv και v εμφανίζονται στο [41], ενώ τα υπόλοιπα στο [21], τα οποία υπολογίζονται χωρίς να αυξάνεται ιδιαίτερα η πολυπλοκότητα, όπως θα δούμε στην ενότητα 2.2.2. Μια μεγάλη προσθήκη είναι τα χαρακτηριστικά που είναι περιστραμμένα κατά 45 °και φαίνονται στο Σχήμα 2.2.c [21]. Με τη χρήση αυτών βελτιώνεται σημαντικά η αναπαράσταση των διαγώνιων σχημάτων. Με την προσθήκη όλων αυτών των χαρακτηριστικών, το σύνολο γίνεται υπερπλήρες και αναπαριστά πολύ καλύτερα την πληροφορία που περιέχεται σε μία εικόνα.

Για τον υπολογισμό του πλήθους των Haar χαρακτηριστικών σε κάποιο παράθυρο εικόνας πλάτους W και ύψους H, ακολουθούμε την παρακάτω διαδικασία [21]. Έστω ότι w και h είναι το πλάτος και ύψος του ορθογωνίου της συνάρτησης Haar που εξετάζουμε. Το μέγεθος του ορθογωνίου θα αυξάνεται κατά ένα σε κάθε βήμα. Άρα, οι μέγιστοι συντελεστές μεγέθυνσης των ορθογωνίων σε πλάτος και ύψος θα είναι $X = \lfloor \frac{W}{w} \rfloor$ και $Y = \lfloor \frac{H}{h} \rfloor$, αντίστοιχα. Το πλήθος των χαρακτηριστικών που προκύπτουν από την εφαρμογή ενός κατακόρυφου Haar χαρακτηριστικού στο παράθυρο εικόνας, είναι:

$$XY \cdot \left(W + 1 - w\frac{X+1}{2} \right) \cdot \left(H + 1 - h\frac{Y+1}{2} \right)$$

Μπορούμε τώρα να εφαρμόσουμε τους παραπάνω τύπους για ένα παράθυρο μεγέθους 30x30. Τότε θα έχουμε $W = 20$ και $= 20$ και το πλήθος των χαρακτηριστικών στο παράθυρο φαίνεται στον παρακάτω πίνακα 2.1:

Αυτό που παρατηρούμε είναι ότι σε ένα παράθυρο μεγέθους 20x20 δηλαδή 400 pixel μπορούν να υπολογιστούν 116.358 χαρακτηριστικά. Το πλήθος αυτό είναι εντελώς δυσανάλογο με τον αριθμό των pixel σε εκείνο το παράθυρο. Η εξήγηση είναι ότι το σύνολο των χαρακτηριστικών που εξετάσαμε είναι υπερπλήρες για την περιγραφή ενός παραθύρου. Στη συνέχεια θα δούμε πως μπορούμε να υπολογίσουμε τα χαρακτηριστικά αυτά πολύ γρήγορα (σε γραμμικό χρόνο) με τη βοήθεια του ολοκληρώματος εικόνας (Integral Image)

2.2.2 Ολοκλήρωμα εικόνας

Όπως είδαμε το πλήθος των χαρακτηριστικών για ένα παράθυρο 20x20 είναι περίπου 120.000. Για να υπολογίσουμε ένα χαρακτηριστικό ο προφανής τρόπος είναι να αθροίσουμε τις τιμές των pixel που απαρτίζουν κάθε ορθογώνιο. Ένας τέτοιος υπολογισμός όμως για 120.000 χαρακτηριστικά είναι αρκετά χρονοβόρος. Σε αυτή την περίπτωση μπορούμε να χρησιμοποιήσουμε μια ενδιάμεση αναπαράσταση της εικόνας, το ολοκλήρωμα εικόνας (*Integral Image*) [1]

[1] Στη βιβλιογραφία συχνά αναφέρεται και ως Πίνακας Προστιθέμενου Εμβαδού (Summed Area Table)

χαρακτηριστικό	w	h	X	Y	πλήθος
	2	1	10	20	21.000
	1	2	20	10	21.000
	2	2	10	10	10.000
	3	1	6	20	13.230
	1	3	20	6	13.230
	4	1	5	20	9.450
	1	4	20	5	9.450
	3	3	6	6	3.969
	1	2	6	6	3.969
	2	1	6	6	3.969
	3	1	5	5	2.025
	1	3	5	5	2.025
	4	1	4	4	1.156
	1	4	4	4	1.156
	3	3	3	3	729
Σύνολο					116.358

Πίνακας 2.1: Πλήθος χαρακτηριστικών Haar ανά παράθυρο

[21] [20]. Χρησιμοποιώντας το ολοκλήρωμα εικόνας μπορούμε να υπολογίσουμε κάθε Haar feature σε σταθερό χρόνο

Υπολογίζουμε το ολοκλήρωμα εικόνας σε ένα συγκεκριμένο σημείο (x, y) χρησιμοποιώντας την ακόλουθη σχέση:

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y')$$

όπου $ii(x, y)$ είναι η τιμή του ολοκληρώματος εικόνας και $i(x, y)$ είναι η τιμή της πραγματικής εικόνας (δλδ η τιμή χρώματος του εικονοστοιχείου)στο σημείο (x, y). Ουσιαστικά η τιμή $ii(x, y)$ είναι το άθροισμα των τιμών των εικονοστοιχείων που βρίσκονται από πάνω και αριστερά του σημείου (x, y).

Χρησιμοποιώντας τις ακόλουθες σχέσεις:

$$s(x, y) = s(x, y - 1) + i(x, y) \qquad (1)$$

$$ii(x, y) = ii(x - 1, y) + s(x, y) \qquad (2)$$

(όπου $s(x, y)$ είναι το άθροισμα των στοιχείων (pixel) μιας γραμμής και $s(x, -1) = 0, ii(-1, y) = 0$) καταλήγουμε στη σχέση:

$$ii(x, y) = i(x, y) + ii(x, y - 1) + ii(x - 1, y) - ii(x - 1, y - 1)$$

Μπορούμε λοιπόν να υπολογίσουμε το ολοκλήρωμα εικόνας για όλα τα σημεία σε γραμμικό χρόνο κάνοντας μόνο ένα πέρασμα από όλα τα στοιχεία της εικόνας.

αρχική εικόνα

3	8	2	1
6	3	9	7
5	2	4	9
6	0	1	8

ολοκλήρωμα εικόνας

3	11	13	14
9	20	31	39
14	27	A 42	59 B
20	33	49 D	74 C

Σχήμα 2.3: Υπολογισμός ολοκληρώματος εικόνας

Αν το δούμε σχηματικά, έχοντάς το Ολοκλήρωμα Εικόνας (Integral Image), κάθε κάθετο ορθογώνιο πάνω στην εικόνα μπορεί να υπολογιστεί χρησιμοποιώντας 4 αναφορές στον πίνακα (βλ Σχήμα 2.3). Αντίστοιχα, η διαφορά μεταξύ δύο ορθογωνίων μπορεί να υπολογιστεί με 8 αναφορές στον πίνακα. Παρακάτω θα δούμε πως αυτά εφαρμόζονται για τον υπολογισμό

των χαρακτηριστικών τύπου Haar σε σταθερό χρόνο με όσο το δυνατόν λιγότερες πράξεις. Παραδείγματος χάρη από το Σχ. 2.3

$$S_{ABCD} = ii_C + ii_(A) - ii_(B) - ii_(D)$$

Το περιστραμμένο κατά 45°ολοκλήρωμα εικόνας [21] [20] στη θέση (x, y) περιλαμβάνει το άθροισμα των τιμών όλων των στοιχείων (pixel) που βρίσκονται στο περιστραμμένο κατά 45°ορθογώνιο που έχει το κατώτερο σημείο του στο σημείο (x, y) (βλ Σχήμα 2.4). Έτσι το 45°Rotated Integral Image ορίζεται ως:

$$rii(x, y) = \sum_{y' \leq y, y' \leq y - |x - x'|} i(x', y')$$

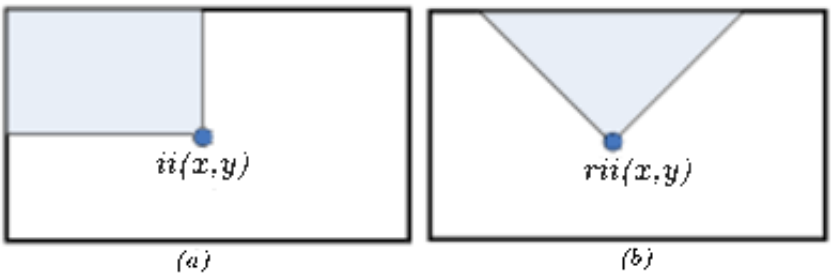
Αντίστοιχα χρησιμοποιώντας τις σχέσεις:

$$rii(x, y) = rii(x - 1, y - 1) + rii(x + 1, y - 1) - rii(x, y - 2) + i(x, y) + i(x, y - 1)$$

και

$$rii(-1, y) = rii(x, -1) = rii(x, -2) = rii(-1, -1) = rii(-1, -2) = 0$$

μπορούμε να υπολογίσουμε τον πίνακα περιστραμμένου προστιθέμενου εμβαδού με ένα πέρασμα όλων των στοιχείων της εικόνας, από αριστερά προς τα δεξιά και από πάνω προς τα κάτω



Σχήμα 2.4: Υπολογισμός 45°-περιστραμμένου ολοκληρώματος εικόνας

Επομένως όπως μπορούμε να καταλάβουμε και από το Σχήμα 2.4 για να υπολογίσουμε το εμβαδόν ενός περιστραμμένου ορθογωνίου εντός μιας εικόνας χρειαζόμαστε μόνο 4 αναφορές στον ανωτέρω πίνακα, και άρα σταθερό χρόνο.

Ας εφαρμόσουμε όμως όλη την παραπάνω λογική για να υπολογίσουμε ένα χαρακτηριστικό τύπου Haar.

Ας δούμε τώρα, το κόστος υπολογισμού κάθε χαρακτηριστικού που χρησιμοποιούμε, με χρήση του ολοκληρώματος εικόνας. Για τον υπολογισμό των χαρακτηριστικών που αποτελούνται από δύο γειτονικά ορθογώνια (από τον Πίνακα 2.1 τα i, ii, ix, x) θα χρειαστούμε 6 αναφορές σε πίνακα και 6 αριθμητικές πράξεις για τον υπολογισμό των δύο ορθογωνίων και 1 αριθμητική πράξη για τη μεταξύ τους αφαίρεση. Άρα, συνολικά 6 αναφορές σε πίνακα και 7 αριθμητικές πράξεις. Για τον υπολογισμό των χαρακτηριστικών που αποτελούνται από τρία γειτονικά ορθογώνια (από τον Πίνακα 2.1 τα iv-vii, xi-xiv) θα χρειαστούμε 8 αναφορές σε πίνακα και 11 αριθμητικές πράξεις (9 για τον υπολογισμό των ορθογωνίων και 2 για τις μεταξύ τους πράξεις). Οι πράξεις τελικά μειώνονται σε 8 ακολουθώντας το σκεπτικό που

φαίνεται στον Πίνακα 2.2 [21]. Τα χαρακτηριστικά viii και xv που φαίνονται στον Πίνακα 2.1, παρότι αποτελούνται από δύο ορθογώνια, αυτά δεν είναι γειτονικά μεταξύ τους. Έτσι, για τον υπολογισμό τους, σύμφωνα με τον Πίνακα 2.2, χρειάζονται 8 αναφορές σε πίνακα και 8 αριθμητικές πράξεις. Για το χαρακτηριστικό iii που αποτελείται από τέσσερα ορθογώνια, απαιτούνται 9 αναφορές σε πίνακα και 12 αριθμητικές πράξεις. Βλέπουμε, λοιπόν, ότι όλα τα χαρακτηριστικά που περιγράφηκαν στην ενότητα 2.2.1 μπορούν να υπολογιστούν σε σταθερό χρόνο, ανεξάρτητα από το μέγεθος του χαρακτηριστικού, με τη χρήση των δύο αυτών πινάκων. Το γεγονός αυτό, επιταχύνει δραστικά το σύστημα ανίχνευσης.

	3 ορθογώνια → 8 αναφορές και 11 πράξεις
	2 ορθογώνια → 8 αναφορές και 6 πράξεις
	8 αναφορές και 8 πράξεις

Πίνακας 2.2: Μείωση του κόστους υπολογισμού χαρακτηριστικών με 3 ορθογώνια

Η ανίχνευση αντικειμένων, θα πρέπει να γίνει σε κάθε δυνατή θέση της εικόνας, καθώς επίσης και σε κάθε δυνατή κλίμακα. Για τον έλεγχο σε κάθε θέση, ο ανιχνευτής κινείται μέσα στην εικόνα, διατρέχοντάς την ολόκληρη, και εφαρμόζοντας τη μέθοδο ανίχνευσης σε κάθε υποπαράθυρο. Για τον έλεγχο σε κάθε κλίμακα, άλλες μέθοδοι δημιουργούν μια πυραμίδα από σμικρύνσεις της εικόνας, και εφαρμόζουν σε κάθε κλίμακα της πυραμίδας τον ανιχνευτή, διατηρώντας σταθερό το μέγεθός του. Με αυτή τη μέθοδο, πέρα από το κόστος της ίδιας της ανίχνευσης, προστίθεται και το χρονικό κόστος της κατασκευής της πυραμίδας εικόνων, το οποίο είναι αρκετά σημαντικό. Στη μέθοδο που εξετάζουμε, αντί να αλλάζουμε το μέγεθος της εικόνας όπου γίνεται η ανίχνευση, αλλάζουμε το μέγεθος του ίδιου του ανιχνευτή. Αυτό είναι εφικτό, καθώς τα χαρακτηριστικά τύπου Haar μπορούν να μεταβληθούν σε μέγεθος. Επίσης, με τη χρήση των δύο πινάκων που είδαμε προηγουμένως, ο υπολογισμός ενός χαρακτηριστικού δεν επηρεάζεται χρονικά από το μέγεθός του. Έτσι, η εφαρμογή της μεθόδου μπορεί να γίνει στον ίδιο ακριβώς χρόνο για οποιαδήποτε κλίμακα του παραθύρου ανίχνευσης. Από μετρήσεις που έχουν γίνει, έχει βρεθεί ότι ο χρόνος που χρειάζεται για την κατασκευή της πυραμίδας εικόνων που χρησιμοποιούν άλλες μέθοδοι, είναι παραπλήσιος με το χρόνο που χρειάζεται η μέθοδος που εξετάζουμε, για όλη τη διαδικασία ανίχνευσης [ViJo02]. Έτσι, βλέπουμε ότι οποιαδήποτε μέθοδος χρησιμοποιεί πυραμίδες εικόνων για την ανίχνευση αντικειμένων, θα είναι αναγκαστικά πιο χρονοβόρα από την εξεταζόμενη μέθοδο.

2.2.3 Ο αλγόριθμος AdaBoost

Στην προηγούμενη ενότητα παρουσιάσαμε το σύνολο των Haar features. Τα χαρακτηριστικά αυτά μας βοηθούν να δούμε αν στο τρέχον παράθυρο εντός της εικόνας που εξετάζουμε υπάρχει κάποιο αντικείμενο (στη συγκεκριμένη περίπτωση πρόσωπο). Μας βοηθάνε δηλαδή να *ταξινομήσουμε* τα επιμέρους παράθυρα επιλέγοντας εκείνα που περιέχουν πρόσωπα. Η διαδικασία αυτή ονομάζεται ταξινόμηση (*Classification*) και ο αλγόριθμος που υλοποιεί, ταξινομητής (*Classifier*).

Όπως είδαμε, τα χαρακτηριστικά που χρησιμοποιούμε μπορούν να υπολογιστούν πάρα πολύ γρήγορα. Το σύνολο όμως των χαρακτηριστικών παραμένει μεγάλο (περίπου 120.000 για μια εικόνα 20x20).Συνεπώς, η διαδικασία υπολογισμού του συνόλου των χαρακτηριστικών

για όλα τα παράθυρα μέσα σε μια εικόνα παραμένει χρονοβόρα [41]. Χρειάζεται λοιπόν, να επιλέξουμε ένα υποσύνολο των χαρακτηριστικών αυτών τα οποία θα είναι ικανά να μας παρέχουν την πληροφορία για την ύπαρξη αντικειμένου.

Ο αλγόριθμος AdaBoost είναι ένας αλγόριθμος μηχανικής μάθησης ο οποίος χρησιμοποιείται τόσο για την επιλογή του υποσυνόλου των χαρακτηριστικών που θα χρησιμοποιηθούν από τον classifier, όσο και για την εκπαίδευση του ταξινομητή. Ανήκει στην κατηγορία των boosting algorithms, χρησιμοποιείται δηλαδή για να αυξήσει την απόδοση ενός οποιουδήποτε απλού αλγορίθμου ταξινόμησης (*weak classifier*). Στην πραγματικότητα, αυτό που κάνει ο AdaBoost είναι να φτιάξει μια αλυσίδα από από weak classifiers χρησιμοποιώντας ένα άπληστο αλγόριθμο, ώστε να σχηματίσει τελικά από αυτούς έναν ισχυρότερο classifier.

Η βελτίωση του ασθενούς αλγορίθμου ταξινόμησης πραγματοποιείται, καλώντας τον αλγόριθμο να επιλύσει μια αλληλουχία προβλημάτων ταξινόμησης. Αρχικά, όλα τα παραδείγματα (θετικά και αρνητικά) παίρνουν μια τιμή βάρους, η οποία είναι ίδια για όλα. Δίνονται στον αλγόριθμο και πραγματοποιείται ο πρώτος κύκλος εκμάθησης, όπου ο αλγόριθμος ταξινομεί όλα τα παραδείγματα με κάθε διαθέσιμη συνάρτηση ταξινόμησης. Έπειτα, οι συναρτήσεις ταξινόμησης διατάσσονται σύμφωνα με τα αποτελέσματά τους, λαμβάνοντας υπόψη το βάρος κάθε παραδείγματος. Επιλέγεται ένας μικρός αριθμός συναρτήσεων ταξινόμησης, από αυτές με τα καλύτερα αποτελέσματα, που αποτελούν τον πρώτο ασθενή ταξινομητή. Ο πρώτος κύκλος εκμάθησης ολοκληρώνεται και τα βάρη των παραδειγμάτων ισοσταθμίζονται, δίνοντας μεγαλύτερο βάρος στα παραδείγματα που ταξινομήθηκαν λανθασμένα από τον πρώτο ασθενή ταξινομητή. Έτσι, στον δεύτερο κύκλο εκμάθησης ο αλγόριθμος ταξινόμησης θα θεωρήσει πιο σημαντικά τα παραδείγματα που ταξινομήθηκαν λανθασμένα από τον προηγούμενο ταξινομητή. Τα βήματα επαναλαμβάνονται διαδοχικά, μέχρι να φτάσουμε στο επίπεδο του συνολικού λόγου λανθασμένης ταξινόμησης που επιθυμούμε. Τελικά, ο ισχυρός ταξινομητής προκύπτει από τον συνδυασμό των ασθενών ταξινομητών που επιλέχθηκαν και ένα κατώφλι. Κατά την διαδικασία της ταξινόμησης ενός υποπαραθύρου εικόνας από τον ισχυρό ταξινομητή, εφαρμόζονται στο υποπαράθυρο όλοι οι ασθενείς ταξινομητές. Τα αποτελέσματα των ασθενών ταξινομητών αθροίζονται, και αν το άθροισμα ξεπερνά το κατώφλι του ταξινομητή, το υπό εξέταση αντικείμενο ταξινομείται ως θετικό, αλλιώς ως αρνητικό.

Ο αλγόριθμος AdaBoost διαθέτει 4 διαφορετικές εκδοχές [15]. Η αρχική εκδοχή ονομάζεται Διακριτός AdaBoost (*Discrete AdaBoost - DAB*) καθώς η συνάρτηση ταξινόμησης κάθε weak classifier παίρνει 2 διακριτές τιμές $-1, 1$ ανάλογα με το αν ένα δείγμα ταξινομείται ως θετικό ή αρνητικό. Η δεύτερη εκδοχή ονομάζεται Πραγματικός AdaBoost (*Real AdaBoost - RAB*), καθώς η συνάρτηση ταξινόμησης έχει ως πεδίο τιμών ολόκληρο το διάστημα $[0, 1]$. Με τη χρήση του RAB, μπορούμε να έχουμε μια ένδειξη εμπιστοσύνης για τα αποτελέσματα της ταξινόμησης, χρησιμοποιώντας τις τιμές που επιστρέφονται από τον αλγόριθμο και όχι μόνο το αποτέλεσμα της θετικής ή αρνητικής ταξινόμησης. Άλλη εκδοχή είναι ο LogitBoost, ο οποίος έχει δύο παραλλαγές, αυτή που χρησιμοποιεί δύο κλάσεις και αυτή που χρησιμοποιεί J κλάσεις Τέλος υπάρχει και ο Gentle AdaBoost, οποίος ουσιαστικά βασίζεται στον Real AdaBoost παράγει όμως τα επιμέρους βήματα χρησιμοποιώντας τη μέθοδο Newton αντί να χρησιμοποιεί ακριβή υπολογισμό σε κάθε βήμα.

Στη μέθοδο που χρησιμοποιήσαμε για την ανίχνευση προσώπων, κάθε ασθενής αλγόριθμος εκμάθησης παράγει τιμές από το περιορισμένο σύνολο συναρτήσεων ταξινόμησης που αποτελούνται από ένα μόνο χαρακτηριστικό τύπου Haar. Προφανώς, από ένα μόνο χαρακτηριστικό δε μπορούμε να περιμένουμε ιδιαίτερα χαμηλό λόγο σφάλματος. Σε κάθε στάδιο του αλγορίθμου AdaBoost επιλέγεται το χαρακτηριστικό που διαχωρίζει καλύτερα τα θετικά από τα αρνητικά δείγματα. Για κάθε χαρακτηριστικό, ο weak classifier προσδιορίζει ένα κατώφλι της

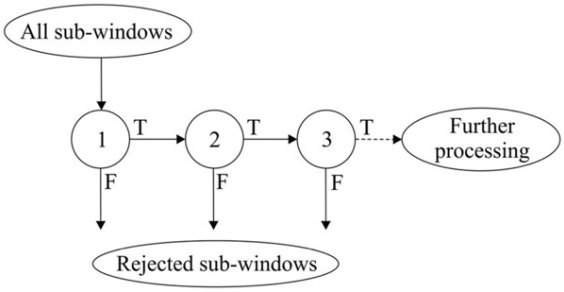
τιμής του χαρακτηριστικού, που ελέγχοντάς το περιορίζονται οι λανθασμένες ταξινομήσεις από το συγκεκριμένο χαρακτηριστικό στις ελάχιστες δυνατές. Έπειτα, επιλέγεται ως ασθενής ταξινομητής το χαρακτηριστικό τύπου Haar, που, για το δεδομένο κατώφλι του, κάνει τη συνολικά καλύτερη ταξινόμηση. Ο AdaBoost συνεχίζει εκπαιδεύοντας όλους τους ασθενείς ταξινομητές, μέχρι το σημείο που ο ισχυρός συνολικός ταξινομητής επιτυγχάνει το επίπεδο ταξινόμησης που ζητάμε.

Ο αλγόριθμος AdaBoost παρέχει αρκετά ισχυρές εγγυήσεις για την ορθότητά του. Έχει αποδειχθεί, ότι το σφάλμα ταξινόμησης του ισχυρού ταξινομητή που προκύπτει από την εφαρμογή του αλγορίθμου, τείνει προς το μηδέν εκθετικά ως προς τον αριθμό των κύκλων εκπαίδευσης [34]. Εξίσου σημαντικό αποτελεί το γεγονός ότι η όλη διαδικασία της εκμάθησης πραγματοποιείται σχεδόν σε πραγματικό χρόνο. Για να κατασκευάσουμε έναν Classifier από τον αλγόριθμο AdaBoost με M επιμέρους Weak Classifiers από ένα συγκεκριμένο πλήθος K χαρακτηριστικών Haar για N εικόνες χρειαζόμαστε χρόνο $O(MKN)$. Αντίθετα άλλοι αλγόριθμοι χρειάζονται $O(MNKN)$ βήματα.

2.2.4 Χρήση ενός Cascade Classifier

Σε αυτή την ενότητα παρουσιάζεται η μέθοδος ταξινόμησης εικόνων με τη χρήση ενός Cascade Classifier [41]. Η μέθοδος αυτή βοηθά στη επίτευξη υψηλού λόγου ανίχνευσης, μειώνοντας σημαντικά τον απαιτούμενο χρόνο. Η γενική ιδέα είναι ότι αντί για έναν μεγάλο και χρονοβόρο classifier μπορούμε να χρησιμοποιήσουμε διαδοχικά πολλούς μικρότερους (άρα και γρηγορότερους) classifiers. Στη συγκεκριμένη περίπτωση χρησιμοποιούμε απλούστερους και πολύ γρήγορους ταξινομητές (ελέγχουν λιγότερα haar features) αρχικά, οι οποίοι θα απορρίπτουν γρήγορα την πλειονότητα των αρνητικών υποπαραθύρων, και πιο σύνθετους και χρονοβόρους που ελέγχουν περισσότερα haar features αργότερα , ώστε να μειώσουμε το λόγο λανθασμένων ανιχνεύσεων.

Κάθε υποπαράθυρο της εικόνας εισέρχεται στον πρώτο ταξινομητή. Αν ο ταξινομητής το κατατάξει ως θετικό, αυτό περνά ως είσοδος στον δεύτερο ταξινομητή. Αν και αυτός το κατατάξει ως θετικό, τότε περνά στον τρίτο κ.ο.κ. Αν σε αυτή τη διαδοχή κάποιος ταξινομητής κατατάξει το υποπαράθυρο ως αρνητικό, τότε αυτό απορρίπτεται και δεν εξετάζεται από κανένα άλλο ταξινομητή (βλέπε Σχήμα 2.5). Θα μπορούσαμε να παρομοιάσουμε αυτή τη διαδικασία με έναν μεγάλο ταξινομητή που αποτελείται από το σύνολο των χαρακτηριστικών Haar, όπου όμως δεν περιμένουμε να υπολογιστεί όλο το πλήθος των χαρακτηριστικών. Αντίθετα, σε κάθε στάδιο ελέγχονται μερικά χαρακτηριστικά και ανάλογα με το άθροισμα των τιμών τους ο ταξινομητής αποφασίζει αν το υπό εξέταση υποπαράθυρο απορρίπτεται ή όχι.



Σχήμα 2.5: Τα επιμέρους βήματα ενός Cascade Classifier

Κάθε ταξινομητής του Cascade Classifier εκπαιδεύεται χρησιμοποιώντας ένα σύνολο θετικών και ένα σύνολο αρνητικών παραδειγμάτων. Το σύνολο των θετικών παραδειγμάτων είναι το ίδιο κατά την εκπαίδευση κάθε ταξινομητή. Το σύνολο των αρνητικών παραδειγμάτων όμως, μεταβάλλεται. Συγκεκριμένα, κάθε ταξινομητής εκπαιδεύεται χρησιμοποιώντας ως αρνητικά παραδείγματα, τα παραδείγματα που ταξινομούνται από τους προηγούμενους ταξινομητές ως θετικά [41]. Αυτό αυξάνει σε πολύ μεγάλο βαθμό τα αρνητικά παραδείγματα τα οποία θα εξεταστούν συνολικά. Για να φτάσει ένας συγκεκριμένος αριθμός αρνητικών παραδειγμάτων στον τρέχοντα ταξινομητή, θα πρέπει τα παραδείγματα αυτά να ταξινομηθούν από όλους τους προηγούμενους ταξινομητές (λανθασμένα) ως θετικά. Ας θεωρήσουμε ότι σε κάθε στάδιο ενός Cascade Claasifier θέλουμε να εξετάζονται 1.000 αρνητικά παραδείγματα και κάθε στάδιο έχει λόγο λανθασμένης θετικής ανίχνευσης 0,5. Τότε, για να περάσουν στο δέκατο στάδιο 1.000 αρνητικά παραδείγματα, αυτά θα χρειαστεί να έχουν ταξινομηθεί από τα προηγούμενα εννιά στάδια ως θετικά. Έτσι, συνολικά θα πρέπει να εξεταστούν περίπου 512.000 αρνητικά παραδείγματα.

Η εξέταση πολύ μεγαλύτερου αριθμού αρνητικών παραδειγμάτων αυξάνει την τελική απόδοση του CC. Αντίθετα, κάθε ταξινομητής καλείται να πραγματοποιήσει μια πιο δύσκολη ταξινόμηση από αυτές των προηγούμενων ταξινομητών. Τα αρνητικά παραδείγματα που θα έχει στη διάθεσή του θα είναι πιο δύσκολα στην ταξινόμηση από τα παραδείγματα που είχαν τα προηγούμενα από αυτό στάδια. Έχοντας, λοιπόν, πιο δύσκολο σύνολο εκπαίδευσης ένας ταξινομητής που βρίσκεται σε προχωρημένο στάδιο, θα παρουσιάσει αυξημένες λανθασμένες ταξινομήσεις, θετικές και αρνητικές.

Οι απλοί επιμέρους ταξινομητές, θα πρέπει να έχουν πολύ χαμηλό λόγο λανθασμένων αρνητικών ταξινομήσεων, ώστε να μην χάνονται τα πραγματικά αντικείμενα στη συνολική ταξινόμηση. Για να διασφαλίσουμε τη σωστή λειτουργία του CC, θα πρέπει να αυξήσουμε περαιτέρω τις θετικές ταξινομήσεις (είτε αφορούν πραγματικά αντικείμενα είτε όχι) [41]. Μια τεχνική για να πετύχουμε αυτό το αποτέλεσμα είναι να επέμβουμε στις τιμές των κατωφλιών των ταξινομητών που προσδιόρισε ο αλγόριθμος AdaBoost κατά την εκπαίδευση. Το κατώφλι ενός ταξινομητή ορίζει την ελάχιστη τιμή του σταθμισμένου με βάρη αθροίσματος των τιμών των χαρακτηριστικών που θα πρέπει να έχει ένα υποπαράθυρο για να ταξινομηθεί ως θετικό. Ένα υποπαράθυρο ταξινομείται, δηλαδή, ως θετικό, όταν το (σταθμισμένο) άθροισμα των τιμών των χαρακτηριστικών που υπολογίστηκαν για αυτό ξεπερνά το κατώφλι του ταξινομητή. Έτσι, μειώνοντας τις τιμές των κατωφλιών θα αυξηθεί ο αριθμός των παραθύρων που ταξινομούνται ως θετικά, άρα και ο λόγος θετικών ταξινομήσεων.

Ο ολοκληρωμένος CC που χρησιμοποιήθηκε για την ανίχνευση προσώπων αποτελείται από 38 επιμέρους στάδια και ένα σύνολο 6000 χαρακτηριστικών. Ο χρόνος που απαιτείται για να τρέξει ο CC είναι άμεσα συνδεδεμένος με τον συνολικό αριθμό των χαρακτηριστικών που υπολογίζονται για κάθε υποπαράθυρο που εξετάζεται. Με βάση αξιολογήσεις που έχουν γίνει [41] [42] πάνω στο MIT-CMU test set [32] υπολογίζονται κατά μέσο όρο 10 Haar features από τα συνολικά 6061 ανά υποπαράθυρο. Αυτό οφείλεται στο γεγονός ότι το μεγαλύτερο ποσοστό από τα υποπαράθυρα απορρίπτεται σε πρώτα στάδια.

Η απόδοση του CC εξαρτάται επίσης από το πλήθος των επιμέρους σταδίων και την απόδοση του καθενός από αυτά.

2.3 Άλλες μέθοδοι που έχουν χρησιμοποιηθεί

Στο κεφάλαιο αυτό κάναμε μια σύντομη αναφορά στις πιο βασικές μεθόδους αναγνώρισης προσώπου σε εικόνα με βάση την απόδοση που επιτυγχάνουν και το χρόνο εκτέλεσής τους. Οι τεχνικές αυτές βασίζονται κατά κύριο λόγο στον υπολογισμό κάποιων χαρακτηριστικών σκίασης από την εικόνα σε συνδυασμό με κάποιο προεκπαιδευμένο μοντέλο αναγνώρισης. Γενικά, στο επιστημονικό πεδίο αυτό έχουν προταθεί αρκετές μέθοδοι με βάση διαφορετικά χαρακτηριστικά της εικόνας. Μερικά από αυτά είναι το χρώμα, η κίνηση καθώς και συνδυασμός αυτών. Ενδεικτικά αναφέρουμε μερικές τεχνικές.

Τεχνικές σε ελεγχόμενο περιβάλλον (πχ μονοχρωματικό φόντο)

Με βάση το χρώμα

- Explanation of basic color extraction for face detection [2]
- Face detection in color images using PCA [3]

Με βάση την κίνηση

- Explanation of basic motion detection for face finding [4]
- Blink detection: human eyes are simultaneously blinking [5]

Με συνδυασμό των δυο προηγουμένων

- A mixture of color and 3D [6]
- A mixture of color and background removal [7]

[2] http://web.archive.org/web/20040815172250/http:/www.dcs.qmul.ac.uk/research/vision/publications/papers/bmvc97/node2.html

[3] http://web.archive.org/web/20070621092425/http:/www.ient.rwth-aachen.de/forschung/bebi/facedetection/publi/men99b.pdf

[4] http://web.archive.org/web/20080522171806/http:/www.ansatt.hig.no/erikh/papers/hig98_6/node2.html

[5] http://www-prima.imag.fr/ECVNet/IRS95/node13.html

[6] http://people.eecs.berkeley.edu/~trevor/papers/1998-021/

[7] http://web.archive.org/web/20030821151710/http:/atwww.hhi.de/blick/Head_Tracker/head_tracker.html

Κεφάλαιο 3

Ανίχνευση αντικειμένων

Στο κεφάλαιο 2 είδαμε ότι η σύγχρονες μέθοδοι ανίχνευσης αντικειμένων αποτελούνται από 3 επιμέρους ανεξάρτητες διαδικασίες.

- Εντοπισμός πιθανών περιοχών αντικειμένου/ων
- Ταξινόμηση αντικειμένου
- Προσδιορισμός θέσης αντικειμένου

Κάθε διαδικασία αποτελεί και ένα ξεχωριστό επιστημονικό πεδίο.

Τα συνελικτικά νευρωνικά δίκτυα άρχισαν να εμπλέκονται στο πεδίο της ανίχνευσης και αναγνώρισης αντικειμένων από 2012. Τη χρονιά αυτή το νευρωνικό δίκτυο AlexNet κέρδισε τον διαγωνισμό ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[1]. Στο διαγωνισμό αυτό, διάφορες ερευνητικές ομάδες αξιολογούν τους αλγορίθμους τους πάνω σε ένα δεδομένο σετ εικόνων, το ImageNet[2], και προσπαθούν να επιτύχουν τη μεγαλύτερη δυνατή ακρίβεια. Τη χρονιά εκείνη το υψηλότερο ποσοστό είχε επιτευχθεί από την ερευνητική ομάδα των Alex Krizhevsky, Ilya Sutskever και Geoffrey E. Hinton χρησιμοποιώντας το συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network) AlexNet [19] επιτυγχάνοντας ένα ποσοστό top-5 error 15.3%, 10.8 ποσοστιαίες μονάδες πάνω από το δεύτερο.

Παρακάτω θα κάνουμε μια σύντομη αναδρομή στο πρόσφατο παρελθόν σχετικά με την εξέλιξη των μεθόδων ανίχνευσης αντικειμένων με τη χρήση νευρωνικών δικτύων.

3.1 Εισαγωγή

Μέχρι πριν λίγα χρόνια, οι πιο επιτυχημένες μέθοδοι για ανίχνευση αντικειμένων χρησιμοποιούσαν για τον εντοπισμό πιθανόν περιοχών αντικειμένου την τεχνική κυλιόμενου παραθύρου [42, 27, 14], με την οποία αποδοτικοί ταξινομητές ελέγχουν για παρουσία αντικειμένου σε κάθε παράθυρο της εικόνας. Με αυτόν τον τρόπο ελέγχονται 10^4 με 10^5 παράθυρα. Με την αύξηση των εικονοστοιχείων της εικόνας έχουμε αντίστοιχα τάξεις μεγέθους μεγαλύτερο αριθμό παραθύρων, αφού τα παράθυρα επιλέγονται σε διάφορα μεγέθη, ενώ οι σύγχρονες

[1] http://www.image-net.org/challenges/LSVRC/
[2] https://en.wikipedia.org/wiki/ImageNet

βάσεις δεδομένων απαιτούν και ανίχνευση της διάστασης του αντικειμένου, κάτι που αυξάνει ακόμα περισσότερο τον χώρο αναζήτησης σε 10^6 με 10^7 παράθυρα.

Η αύξηση της πολυπλοκότητας των αλγορίθμων για την ανίχνευση αντικειμένων οδήγησε σε αύξηση της ποιότητας ανίχνευσης αλλά ταυτόχρονα αυξήθηκε σημαντικά και ο χρόνος που απαιτείται σε κάθε υποψήφιο παράθυρο [36, 9]. Ένας τρόπος αντιμετώπισης του προβλήματος έτσι ώστε να διατηρηθεί ο χρόνος σε λογικά επίπεδα και να παραμείνει υψηλή η ποιότητα της ανίχνευσης είναι η χρήση των υποψήφιων θέσεων αντικειμένων [11, 40]. Θεωρώντας ότι όλα τα αντικείμενα μοιράζονται παρόμοια χαρακτηριστικά για να ξεχωρίζουν από το περιβάλλον τους, σχεδιάστηκαν αλγόριθμοι οι οποίοι παίρνοντας μια εικόνα σαν είσοδο, δίνουν σαν έξοδο ένα σύνολο από θέσεις της εικόνας στις οποίες υπάρχει αυξημένη πιθανότητα να υπάρχει αντικείμενο. Στόχος των μεθόδων αυτών είναι να επιστρέψουν όσο το δυνατόν περισσότερα αντικείμενα που περιέχει η εικόνα, σε όσο το δυνατών μικρότερο αριθμό παραθύρων. Έτσι λοιπόν, πιο εξειδικευμένοι και πολύπλοκοι αλγόριθμοι τρέχουν σε πολύ μικρότερο χρόνο αφού τρέχουν σε πολύ μικρότερο αριθμό παραθύρων.

Η συγκεκριμένη μεθοδολογία προτάθηκε αρκετά πρόσφατα, μόλις το 2011, από τους B. Alexe, T. Deselaers και V. Ferrari [2]. Το μέτρο που χρησιμοποίησαν για το κατά πόσο ένα σημείο της εικόνας είναι αντικείμενο ή όχι ονομάζεται `objectness` (αντικειμενικότητα). Σύμφωνα με εκείνους ένα αντικείμενο χρειάζεται να πληρεί μία τουλάχιστον από τις παρακάτω προϋποθέσεις:

- να είναι ορισμένο εντός ενός κλειστού ορίου
- η αναπαράστασή του εντός της εικόνας να είναι διαφορετική από το περιβάλλον του
- να είναι μοναδικό ή να ξεχωρίζει εντός της εικόνας

Σε αντίθεση με τους ανιχνευτές αντικειμένων που εξειδικεύονται σε μια κλάση αντικειμένων, όπως αυτοκίνητα ή άνθρωποι, οι υποψήφιες θέσεις αντικειμένων έχουν γενικότερη δράση και εντοπίζουν όλα τα είδη αντικειμένων. Αυτό έχει σαν αποτέλεσμα την εύκολη γενίκευση της εφαρμογής της μεθόδου ακόμα και σε αντικείμενα που δεν έχουν ανιχνευθεί ξανά στο παρελθόν

Από την στιγμή που το πρότειναν οι Alexe, Deselaers και Ferrari μέχρι σήμερα, έχουν προταθεί πολλές μέθοδοι για τον υπολογισμό των υποψήφιων θέσεων αντικειμένων. Η βασική μετρική για την ποιότητα των αλγορίθμων αυτών είναι η ανάκληση (recall). Ως ανάκληση ορίζουμε το ποσοστό των πραγματικών θέσεων αντικειμένων της εικόνας που επιστρέφονται ως προτεινόμενα από την εκάστοτε μέθοδο. Στις μέρες μας, έχουν αναπτυχθεί πολλοί καλοί αλγόριθμοι, ο καθένας με επιτυχία σε διαφορετικούς τομείς. Υπάρχουν αλγόριθμοι που έχουν επιτύχει πολύ καλή ανάκληση (98% Selective Search [39]) αλλά επιστρέφουν μεγάλο αριθμό υποψήφιων θέσεων, άλλοι πολύ γρήγοροι (0.2/ Bing [8]) αλλά με χαμηλή ανάκληση και άλλοι με πολύ καλά αποτελέσματα, λίγες υποψήφιες θέσεις αλλά πολύ αργοί (CPMC [7]).

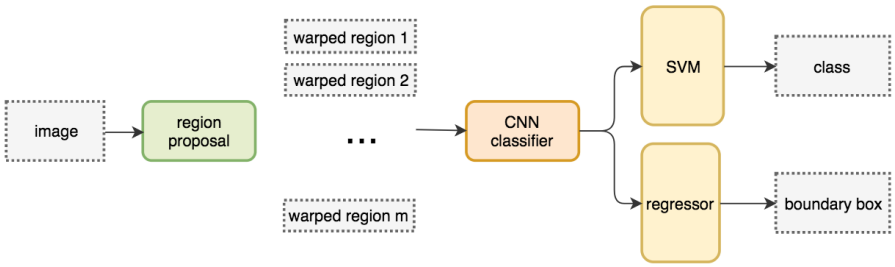
Όλοι οι σύγχρονοι βέλτιστοι ανιχνευτές αντικειμένων για τις βάσεις δεδομένων PASCAL [12] και ImageNet [33], χρησιμοποιούν όλοι υποψήφιες θέσεις αντικειμένων. Η χρήση των υποψήφιων θέσεων αντικειμένων αλλάζει τα δεδομένα τα οποία επεξεργάζεται ο ταξινομητής. Αυτό μπορεί να βελτιώσει και την ποιότητα των αποτελεσμάτων μειώνοντας τις εσφαλμένες ανιχνεύσεις (false positives).

3.2 Ανίχνευση αντικειμένων χρησιμοποιώντας Περιφερειακά Συνελικτικά Νευρωνικά Δίκτυα (R-CNN)

Η λογική των R-CNNs είναι η παραγωγή περιοχών ενδιαφέροντος (Regions Of Interest - ROIs) ,χρησιμοποιώντας μια από τις μεθόδους για υποψήφιες θέσεις αντικειμένων, εξαγωγή των χαρακτηριστικών εικόνας για κάθε ROI και στη συνέχεια ταξινόμηση (classification) του ROI και εξαγωγή του περιγράμματος του αντικειμένου.

3.2.1 R-CNN

Η τεχνική αυτή αυτή χρησιμοποιεί με μέθοδο παραγωγής περιοχών αντικειμένων για να εξάγει 2000 περιοχές ενδιαφέροντος (ROIs). Οι περιοχές αυτές ,αφού πρωτίστως ανασχηματιστούν ώστε να έχουν όλες το ίδιο μέγεθος, δίνονται ως είσοδο σε ένα CNN για την εξαγωγή χαρακτηριστικών από αυτές. Στη συνέχεια ο χάρτης των χαρακτηριστικών δίνεται ως είσοδο στα πλήρως συνδεδεμένα επίπεδα του νευρωνικού δικτύου (Fully Connected Layers) για την ταξινόμηση που αντικειμένου και την εύρεση του περιγράμματός του.



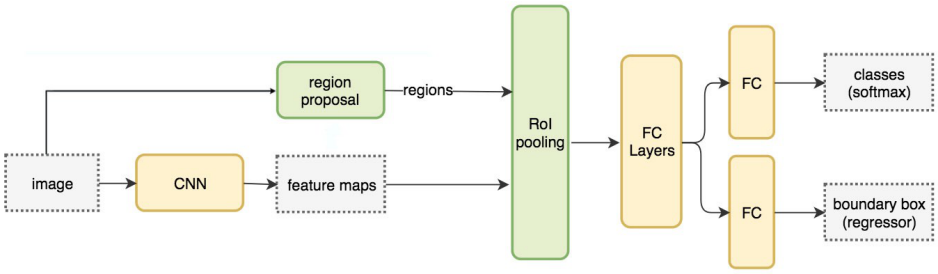
Σχήμα 3.1: Η μέθοδος R-CNN

Η ανίχνευση αντικειμένου χρησιμοποιώντας μια μέθοδο παραγωγής προτεινόμενων περιοχών ενδιαφέροντος και το νευρωνικό δίκτυο R-CNN επιτυγχάνει μικρότερο χρόνο εκτέλεσης και μεγαλύτερη ακρίβεια σε σχέση με τις μεθόδους κυλιόμενου παραθύρου. Όμως επειδή κάθε μία από τις 2000 προτεινόμενες περιοχές ενδιαφέροντος επεξεργάζεται ξεχωριστά, η μέθοδος συνολικά παραμένει αργή τόσο στην εκπαίδευση όσο και στην εξαγωγή αποτελέσματος.

3.2.2 Fast R-CNN

Το νευρωνικό Fast R-CNN επιλύει το παραπάνω πρόβλημα εφαρμόζοντας τη μέθοδο εξαγωγής χαρακτηριστικών κατευθείαν πάνω στην αρχική εικόνα. Παράλληλα, με χρήση μιας μεθόδου παραγωγής προτεινόμενων περιοχών ενδιαφέροντος, εξάγονται οι περιοχές αυτές οι οποίες συνδυάζονται με το χάρτη χαρακτηριστικών της εικόνας. Το αποτέλεσμα είναι η παραγωγή των χαρτών χαρακτηριστικών όλων των ROIs σε μόλις δύο βήματα. Οι χάρτες αυτοί υπόκεινται πρώτα σε μια τεχνική που ονομάζεται ομοιομορφοποίηση (ROI pooling) για να αποκτήσουν ίσες διαστάσεις και στη συνέχεια δίνονται ως είσοδο στα πλήρως συνδεδεμένα επίπεδα του νευρωνικού δικτύου για την ταξινόμηση και την εύρεση του περιγράμματος.

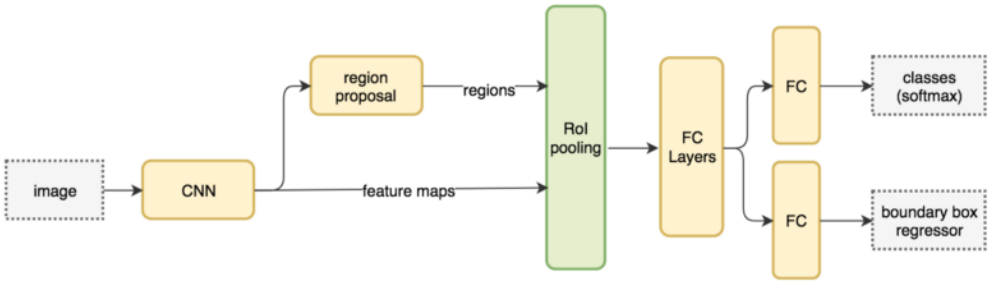
Το Fast R-CNN ,με τη παραπάνω διαδικασία, αποφεύγει την εφαρμογή της συνάρτησης εξαγωγής χαρακτηριστικών για κάθε ROI με αποτέλεσμα ο συνολικός χρόνος εκτέλεσης της μεθόδου για την ανίχνευση του αντικειμένου να ελαττώνεται σημαντικά.



Σχήμα 3.2: Η μέθοδος Fast R-CNN

3.2.3 Faster R-CNN

Όπως ήδη έχουμε αναφέρει, η διαδικασία εξαγωγής προτεινόμενων περιοχών ενδιαφέροντος είναι μια αρκετά ακριβή χρονικά διαδικασία. Πιο συγκεκριμένα, ο συνολικός χρόνος εκτέλεσης της μεθόδου Fast R-CNN για την εξαγωγή της κλάσης και του περιγράμματος των αντικειμένων εντός μια εικόνας, διαρκεί κατά μέσο όρο $2,3$ δευτερόλεπτα. Από αυτό το χρόνο περίπου τα 2 δευτερόλεπτα χρειάζονται για την εξαγωγή των 2000 περιοχών ενδιαφέροντος από την αρχική εικόνα.



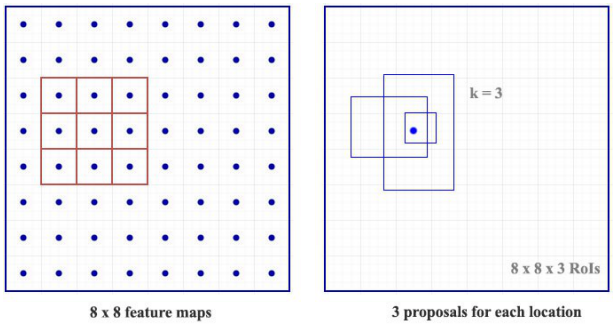
Σχήμα 3.3: Η μέθοδος Faster R-CNN

Η μέθοδος Faster R-CNN, χρησιμοποιεί την ίδια λογική με τη μέθοδο Fast R-CNN, αλλά για να βελτιώσει το χρόνο εκτέλεσης της μεθόδου εξαγωγής προτεινόμενων περιοχών ενδιαφέροντος χρησιμοποιεί εσωτερικά ένα μαθησιακό δίκτυο ώστε οι περιοχές ενδιαφέροντος να εξάγονται από το χάρτη χαρακτηριστικών της εικόνας. Το δίκτυο εξαγωγής περιοχών ενδιαφέροντος (Regional Proposal Network - RPN) είναι πιο αποτελεσματικό και χρειάζεται μόλις 10 χιλιοστά του δευτερολέπτου για να δημιουργήσει τα ROIs. Η συνέχεια της μεθόδου είναι όμοια με την παραπάνω 3.2.2. Από το χάρτη χαρακτηριστικών της αρχικής εικόνας και τις προτεινόμενες περιοχές ενδιαφέροντος σχηματίζονται οι χάρτες χαρακτηριστικών των περιοχών ενδιαφέροντος. Οι χάρτες αυτοί ,αφού πρώτα υποβληθούν σε ομοιομορφοποίηση (ROI pooling), εισάγονται στα πλήρως συνδεδεμένα επίπεδα του νευρωνικού δικτύου για την ταξινόμηση του αντικειμένου και τον προσδιορισμού του περιγράμματος.

Regional Proposal Network - Anchors

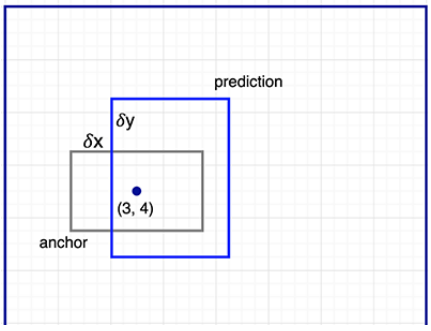
Το RPN ,όπως είδαμε, δέχεται ως είσοδο το χάρτη χαρακτηριστικών της εικόνας και πα-

ράγει για κάθε θέση k υποψήφια παράθυρα και $2 * k$ τιμές για το objectness. Για παράδειγμα σε ένα χάρτη $8x8$ και για $k = 3$ έχουμε:



Σχήμα 3.4: Regional Proposal Network

Λόγω του ότι τα αντικείμενα έχουν διάφορα σχήματα, θα θέλαμε το προτεινόμενα παράθυρα να έχουν και αυτά διαφορετικές διαστάσεις. Για το λόγο αυτό λοιπόν οι προβλέψεις των παραθύρων δεν υπολογίζονται τυχαία αλλά υπολογίζονται σε σχέση με κάποια προκαθορισμένα παράθυρα γύρω από κάθε θέση τα οποία ονομάζονται **anchors**. Για κάθε λοιπόν προτεινόμενο παράθυρο υπολογίζονται οι διαφορές x και y από την πάνω αριστερή γωνία του αντίστοιχου anchor.



Σχήμα 3.5: Regional Proposal Network

Οι anchors προσαρμόζονται ώστε να καλύπτουν τα χαρακτηριστικά αντικειμένων του πραγματικού κόσμου τόσο σε διαστάσεις, μέγεθος και κλίμακα. Με αυτόν τον τρόπο η εκπαίδευση του συνολικού δικτύου είναι ελεγχόμενη με την έννοια ότι τα προτεινόμενα παράθυρα για κάθε θέση είναι ακριβέστερα τόσο στη θέση όσο και στο σχήμα τους.

Σε εφαρμογές πραγματικού περιβάλλοντος ή μέθοδος Faster R-CNN χρησιμοποιεί 9 anchors για κάθε θέση, 3-ων διαφορετικών σχημάτων σε 3 διαφορετικές κλίμακες.

3.3 Ανίχνευση αντικειμένων χρησιμοποιώντας ανιχνευτές μονής λήψης

Οι ανιχνευτές βασιζόμενοι σε μια περιοχή της εικόνας (Regional Based Detectors) 3.2 είναι αρκετά ακριβείς, όμως ο χρόνος εκτέλεσής τους συνεχίζει να είναι υψηλός. Π.χ. η μέθοδος

Faster R-CNN επεξεργάζεται 7 εικόνες το δευτερόλεπτο (7 Frames Per Second) από το εκπαιδευτικό σετ PASCAL VOC 2007 [13] [3].

```
feature_maps = process(image)
ROIs = region_proposal(feature_maps)
for ROI in ROIs
    patch = roi_align(feature_maps, ROI)
    results = detector2(patch)
```

Listing 3.1: Regional Proposal Methods

Ένας εύκολος τρόπος για να βελτιωθεί ο χρόνος εκτέλεσης της μεθόδου, είναι να μειωθεί ο χρόνος που χρειάζεται για την επεξεργασία και εξαγωγή αποτελέσματος από κάθε περιοχή ενδιαφέροντος. Πιο συγκεκριμένα μπορούμε να εξάγουμε την κλάση και το περίγραμμα ενός αντικειμένου κατευθείαν από το χάρτη χαρακτηριστικών της εικόνας. Με αυτόν τον τρόπο παρακάμπτουμε τα βήματα παραγωγής και επεξεργασίας των ROIs, ελαττώνοντας σημαντικά το συνολικό χρόνος εκτέλεσης της μεθόδου.

```
feature_maps = process(image)
results = detector3(feature_maps) # Χωρίς επιπλέον βήμα για τα ROIs
```

Listing 3.2: Single Shot Method

Εξετάζοντας εκ νέου τη μέθοδο κυλιόμενου παραθύρου, παρατηρούμε ότι χρησιμοποιούνται παράθυρα διαφορετικού σχήματος για διαφορετικούς τύπους αντικειμένων τα οποία διατρέχουν όλη την εικόνα ή το χάρτη χαρακτηριστικών της. Το βασικό μειονέκτημα αυτής της μεθόδου είναι ότι το παράθυρο το οποίο διατρέχει την εικόνα χρησιμοποιείται από τη μέθοδο ως το τελικό περίγραμμα του εντοπιζόμενου αντικειμένου. Επομένως η μέθοδος χρησιμοποιεί όλα τα πιθανά μεγέθη και σχήματα παραθύρου για να μπορεί να εντοπίσει τα αντικείμενα όλων των διαφορετικών τύπων και μεγεθών.

Ένας πιο αποτελεσματικός τρόπος, είναι να θεωρήσουμε το παράθυρο αυτό όχι το τελικό περίγραμμα του αντικειμένου, αλλά μια αρχική πρόβλεψη για τη θέση του. Επομένως, στη συνέχεια θα χρειαστούμε έναν ανιχνευτή για τον ταυτόχρονο προσδιορισμό της κλάσης και του πραγματικού περιγράμματος του αντικειμένου.

Ο τρόπος αυτός μοιάζει αρκετά με τη λειτουργία των anchors που χρησιμοποιεί για να εξάγει τις περιοχές ενδιαφέροντος το Regional Proposal Network στη μέθοδο Faster R-CNN 3.5. Αντιθέτως όμως, ένας ανιχνευτής μονής λήψης προβλέπει τόσο το περίγραμμα (boundary box), όσο και την κλάση του αντικειμένου την ίδια χρονική στιγμή.

Η διαφορά έγκειται στο γεγονός ότι το Faster R-CNN χρησιμοποιεί ένα συνελικτικό φίλτρο για να κάνει μια πρόβλεψη 5 παραμέτρων για κάθε θέση στο χάρτη χαρακτηριστικών. Οι 4 από αυτές αναφέρονται στο προβλεπόμενο bounding box από το τρέχον anchor ενώ η άλλη μας δίνει το ποσοστό αντικειμενικότητας του anchor. Επομένως επαναφέροντας το παράδειγμα της

[3] http://host.robots.ox.ac.uk/pascal/VOC/

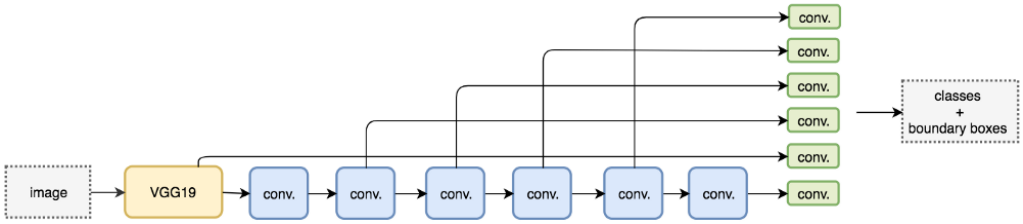
προηγούμενης ενότητας 3.2.3, εφαρμόζοντας ένα συνελικτικό φίλτρο διαστάσεων $3 \times 3 \times D \times 5$ σε ένα χάρτη χαρακτηριστικών $8 \times 8 \times D$ εκείνος τελικά μετατρέπεται σε ένα χάρτη διαστάσεων $8 \times 8 \times 5$.

Ένας ανιχνευτής μονής λήψης (Single Shot) κάνει επιπλέον C προβλέψεις για την κλάση του αντικειμένου για κάθε θέση του χάρτη χαρακτηριστικών. Με αυτόν τον τρόπο για $C = 20$ το συνελικτικό φίλτρο που χρησιμοποιείται έχει διαστάσεις $3 \times 3 \times 25(5 + C)$ και μετατρέπει τον χάρτη χαρακτηριστικών από $8 \times 8 \times D$ σε $8 \times 8 \times 25$.

Γενικά, οι ανιχνευτές μονής λήψης πετυχαίνουν αρκετά καλύτερους χρόνους εκτέλεσης υστερώντας όμως σε ακρίβεια σε σχέση με τις μεθόδους που χρησιμοποιούν περιοχές ενδιαφέροντος. Επιπλέον έχει παρατηρηθεί ότι δυσκολεύονται να ανιχνεύσουν αντικείμενα που βρίσκονται πολύ κοντά ή είναι πολύ μικρά.

3.3.1 Single Shot Multibox Detector (SSD)

Ο ανιχνευτής SSD [24] σχεδιάστηκε για την ανίχνευση αντικειμένων σε πραγματικό χρόνο. Η τυπική του υλοποίηση χρησιμοποιεί το νευρωνικό δίκτυο **VGG16** [35] για την εξαγωγή χαρακτηριστικών από μια εικόνα. Το αποτέλεσμα προωθείται σε μια αλληλουχία συνελικτικών επιπέδων για να καταλήξει τελικά σε συνελικτικά φίλτρα τα οποία πραγματοποιούν και την τελική πρόβλεψη.



Σχήμα 3.6: Η μέθοδος Single Shot Multibox Detector

Για κάθε θέση του χάρτη χαρακτηριστικών το SSD πραγματοποιεί 4 προβλέψεις. Κάθε πρόβλεψη αποτελείται από ένα περίγραμμα (τεσσάρων παραμέτρων) και 21 βαθμολογίες· μία για κάθε κλάση, όπως αυτές προσδιορίστηκαν από το διαγωνισμό PASCAL VOC [13]. Από αυτές κρατάει την υψηλότερη ως την προβλεπόμενη κλάση του αντικειμένου.

3.3.2 Η μέθοδος You Only Look Once (YOLO)

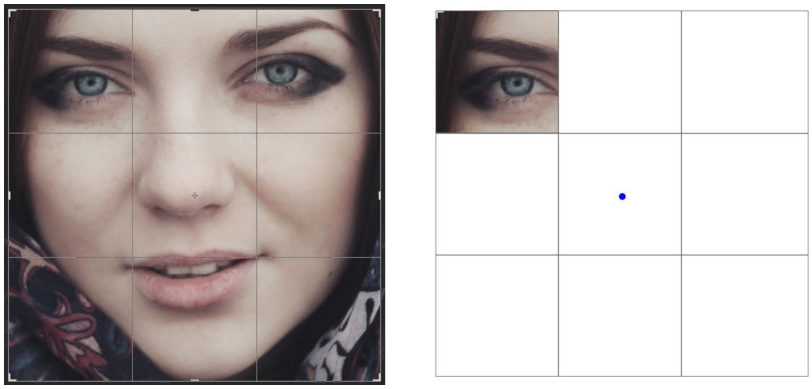
Το YOLO είναι ένας ανιχνευτής μονής λήψης. Χρησιμοποιεί το νευρωνικό δίκτυο DarkNet [28] για την εξαγωγή των χαρακτηριστικών, ακολουθούμενο από συνελικτικά επίπεδα. Σε αντίθεση με τη μέθοδο SSD, δεν χρησιμοποιεί τους χάρτες χαρακτηριστικών με διαφορετικές διαστάσεις από κάθε επίπεδο για να κάνει προβλέψεις, αλλά αντί αυτού ανασχηματίζει και συνθέτει διάφορους χάρτες πραγματοποιώντας από αυτούς τις προβλέψεις. Για παράδειγμα ανασχηματίζει ένα επίπεδο $28 \times 28 \times 512$ σε ένα $14 \times 14 \times 2048$. Στη συνέχεια παίρνει τον παραγόμενο από αυτό το επίπεδο χάρτη χαρακτηριστικών και τον συνθέτει με ένα μικρότερης ανάλυσης χάρτη $14 \times 14 \times 1024$. Στη συνέχεια εφαρμόζει στο νέο αυτό χάρτη διαστάσεων 14×3072 συνελικτικά φίλτρα για να πραγματοποιήσει στη συνέχεια τις προβλέψεις.

Το YOLOv2 [29] πετυχαίνει 63.4 μέση ακρίβεια (mean Average Precision - mAP). Με διάφορες βελτιώσεις το YOLOv2 [30] καταφέρνει να αυξήσει την mAP σε 78.6 ενώ το YOLO900 [30] καταφέρνει να ανιχνεύσει έως και 9000 διαφορετικές κατηγορίες αντικειμένων.

3.4 Άλλες μέθοδοι για την ανίχνευση αντικειμένων

3.4.1 Πλήρως Συνελικτικά Δίκτυα βασιζόμενα σε μια περιοχή της εικόνας (Region-based Fully Convolutional Networks - R-FCN [10])

Στο σχήμα 3.1 είδαμε τη γενική μέθοδο που χρησιμοποιούν οι τεχνικές συνελικτικών νευρωνικών δικτύων. Οι τεχνικές αυτές στο βήμα 5, όπου πραγματοποιείται η ανίχνευση, επεξεργάζονται τους χάρτες χαρακτηριστικών των ROIs και παράγουν αποτελέσματα μέσω πλήρως συνδεδεμένων επιπέδων. Η επεξεργασία δεδομένων από πλήρως συνδεδεμένα επίπεδα είναι μια σχετικά ακριβή διαδικασία και αν σκεφτούμε ότι στην προκείμενη περίπτωση το νευρωνικό επεξεργάζεται 2.000 ROIs, τότε η συνολική διαδικασία είναι αρκετά ακριβή.



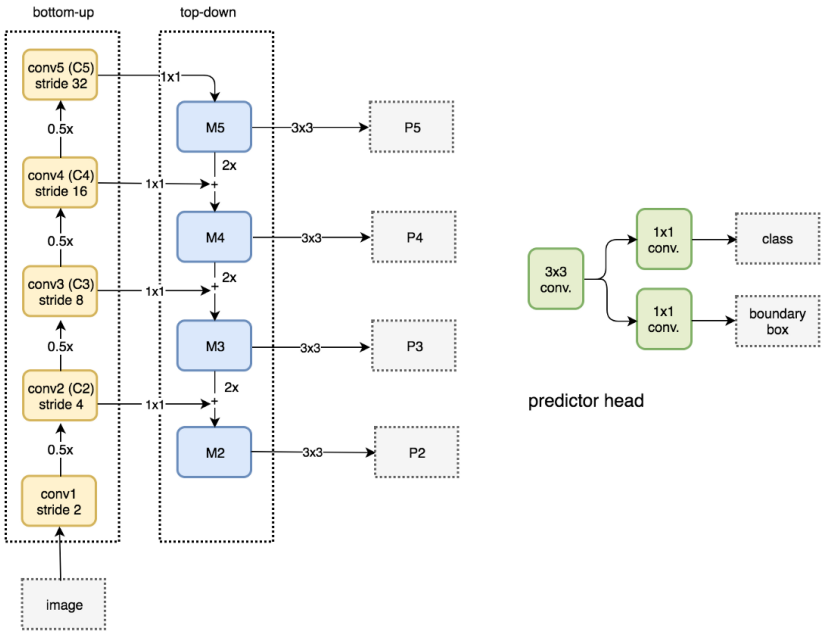
Σχήμα 3.7: Η τεχνική R-FCN

Αν λοιπόν μειωθεί ο χρόνος επεξεργασίας που χρειάζεται για κάθε ROI τότε θα βελτιωθεί και η ταχύτητα εκτέλεσης του ανιχνευτή. Η τεχνική R-FCN εκμεταλλεύεται τους χάρτες χαρακτηριστικών κάποιων υποπεριοχών του αντικειμένου και της θέσης του για να εξάγει ένα τελικό συμπέρασμα για την κλάση και τη θέση του αντικειμένου. Ένα απλοϊκό παράδειγμα έχει να κάνει με την ανίχνευση προσώπου. Σε κάθε υποψήφιο ROI ελέγχεται αν συγκεκριμένες υποπεριοχές του πληρούν τις προϋποθέσεις ενός προσώπου. Π.χ. το δεξί μάτι να είναι στην πάνω αριστερή γωνία, το αριστερό στην πάνω δεξιά κ.ο.κ. Με αυτό τον τρόπο συνδυάζοντας τις επιμέρους βαθμολογίες για κάθε υποπεριοχή ενός ROI το δίκτυο εξάγει γρηγορότερα αποτελέσματα χωρίς απώλειες σε ακρίβεια.

3.4.2 Δίκτυα Πυραμίδας Χαρακτηριστικών (Feature Pyramid Networks FPN [22])

Τα δίκτυα πυραμίδας δεν αποτελούν από μόνα τους ένα ανιχνευτή αντικειμένων, αλλά ενσωματώνονται στις υπάρχουσες τεχνικές για την ανίχνευση αντικειμένων σε διάφορες κλίμακες

(μεγέθη). Τα Δίκτυα Πυραμίδας παράγουν πυραμίδες χαρτών χαρακτηριστικών από μια εικόνα οι οποίοι βελτιώνουν την ταχύτητα και την ακρίβεια του τελικού αποτελέσματος. Στο παρακάτω σχήμα 3.8 τα $P2, P3, P4, P5$ είναι αναπαριστούν την προαναφερθείσα πυραμίδα.



Σχήμα 3.8: Το FPN του σχήματος τροφοδοτεί έναν Object Detector

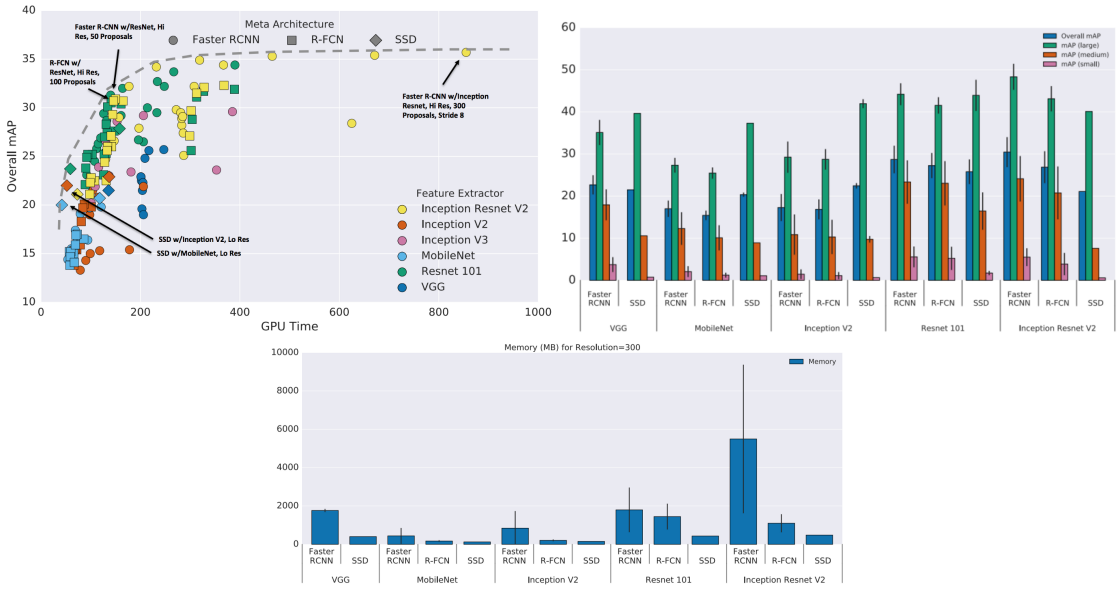
Υπάρχουν 2 πυραμιδικές διαδρομές από χάρτες, μία από κάτω προς τα πάνω και μία από πάνω προς τα κάτω. Τα χαμηλότερα επίπεδα χαρτών έχουν υψηλότερη ανάλυση αλλά μικρότερη πυκνότητα πληροφορίας σε σχέση με τα ανώτερα. Στο SSD 3.3.1 για παράδειγμα χρησιμοποιούνται μόνο τα ανώτερα επίπεδα (πλούσια σε πληροφορία) για τον εντοπισμό της θέσης ενός αντικειμένου. Παρόλο που η πρόβλεψη είναι ακριβέστερη και πιο γρήγορη, λόγω της μικρότερης ανάλυσης το SSD δυσκολεύεται να ανιχνεύσει μικρά αντικείμενα. Ένα FPN ανακατασκευάζει από τα ανώτερα επίπεδα χαρτών, νέα επίπεδα με περισσότερες διαστάσεις. Μάλιστα επειδή τα νέα αυτά επίπεδα υστερούν σε ακρίβεια -λόγω του ότι προέρχονται ύστερα από υποδειγματοληψία και υπερδειγματοληψία του αρχικού χάρτη- είναι συνδεδεμένα με τους χάρτες χαρακτηριστικών ιδίων διαστάσεων. Με αυτόν τον τρόπο ο ανιχνευτής υποψήφιων θέσεων αντικειμένων λειτουργεί καλύτερα.

Ένα FPN είπαμε είναι ένα εξαγωγέας χαρακτηριστικών από μια εικόνα. Στις διάφορες μεθόδους ακολουθείται κυρίως από ένα RPN 3.2.3 το οποίο εξάγει τα ROIs. Ανάλογα με το μέγεθος του κάθε ROI επιλέγεται ο χάρτης χαρακτηριστικών κατάλληλων διαστάσεων για να εξαχθούν οι υποπεριοχές χαρακτηριστικών. Στη συνέχεια κατά τα γνωστά κατά τα άλλα οι υποπεριοχές αυτές υπόκεινται σε ομαλοποίηση (ROI pooling) για να επεξεργαστούν στη συνέχεια από πλήρως συνδεδεμένα επίπεδα και να εξαχθεί η κλάση και το περίγραμμα του αντικειμένου.

3.5 Συνολική αποτίμηση

Στο κεφάλαιο αυτό έγινε μια καταγραφή των σύγχρονων τεχνικών ανίχνευσης αντικειμένων χρησιμοποιώντας νευρωνικά δίκτυα. Είδαμε την εξέλιξη των μεθόδων που βασίζονται σε προτεινόμενες περιοχές ενδιαφέροντος καταλήγοντας στην πλέον σύγχρονη Faster R-CNN 3.2.3. Αναλύσαμε επιπλέον την τεχνική των ανιχνευτών μονής λήψης και είδαμε ορισμένες σύγχρονες μεθόδους όπως η SSD 3.3.1 και η YOLO 3.3.2. Κάθε μια μέθοδος έχει τα δικά της χαρακτηριστικά (mAP, χρόνο εκτέλεσης, κατανάλωση μνήμης) ανάλογα με το περιβάλλον που χρησιμοποιείται. Για παράδειγμα είδαμε ότι η μέθοδος SSD αναγνωρίζει σχετικά μεγάλα αντικείμενα με μεγάλη ακρίβεια ενώ έχει αρκετά μικρότερη ακρίβεια σε μικρά.

Στα πλαίσια της παρούσας εργασίας έγινε ανάλυση των χαρακτηριστικών της κάθε μεθόδου και αυστηρός προσδιορισμός των απαιτήσεων του υλοποιηθέντος συστήματος για την επιλογή της κατάλληλης μεθόδου. Η μέθοδος η οποία χρησιμοποιήθηκε για την ανίχνευση αντικειμένων ήταν η SSD χρησιμοποιώντας το νευρωνικό δίκτυο MobileNet [16] για την εξαγωγή των χαρακτηριστικών από την εικόνα.



Σχήμα 3.9: α) Χρόνος εκτέλεσης κ mAP β) Μέγεθος αντικειμένου γ) Μνήμη

Τα παραπάνω αποτελέσματα ελήφθησαν χρησιμοποιώντας το σύνολο εικόνων MS COCO.

Ο βασικός λόγος που επιλέχθηκε η συγκεκριμένη μέθοδος ήταν η απαίτηση του συστήματός μας για μικρό χρόνος εκτέλεσης καθώς και ο περιορισμός του σε μνήμη. Πιο συγκεκριμένα, λόγω του ότι το σύστημα επεξεργάζεται βίντεο δηλαδή μια αλληλουχία εικόνων και μάλιστα κινηματογραφικού επιπέδου, σημαίνει ότι ο συνολικός αριθμός των frames που εισάγεται σε κάθε κύκλο επεξεργασίας είναι αρκετά μεγάλος. Για να επιτύχουμε ένα χρόνο επεξεργασίας που θα είναι ανταγωνιστικός χρειάζεται να επενδύσουμε σε μια γρήγορη μέθοδοι. Επίσης σε ένα βίντεο στο πραγματικό κόσμο υπάρχουν αρκετές εκατοντάδες αντικείμενα τα οποία όμως δεν εμπίπτουν όλα στο ενδιαφέρον του χρήστη. Για το χρήστη τα σημαντικά αντικείμενα είναι εκείνα που με κάποιο τρόπο προσδιορίζουν το γενικότερο θέμα του βίντεο. Για παράδειγμα ένα ντοκιμαντέρ για τα ζώα, προβάλλει πολλά είδη ζώων ένα σχετικά μεγάλο αντικείμενο. Το SSD δεν έχει πρόβλημα να αναγνωρίσει αντικείμενα τέτοιου μεγέθους, οπότε για τις συγκεκριμένες ανάγκες του συστήματός μας οι προβλέψεις του ήταν ανταγωνιστικές. Τέλος το γεγονός

της περιορισμένης μνήμης των μηχανημάτων που διαθέταμε μας οδήγησε περισσότερο στην επιλογή του δικτύου MobileNet για την εξαγωγή των χαρακτηριστικών το οποίο σχεδιάστηκε με στόχο να έχει μικρές απαιτήσεις σε μνήμη και υπολογιστική ισχύ ώστε να κάνει τις μεθόδους ανίχνευσης αντικειμένου υλοποιήσιμες σε έξυπνες συσκευές κινητής τηλεφωνίας.

Κεφάλαιο 4

Αναγνώριση προσώπων

Στο κεφάλαιο αυτό θα μιλήσουμε για τη διαδικασία της αναγνώρισης προσώπου. Η διαδικασία αυτή διαφέρει από τη διαδικασία της ανίχνευσης προσώπου όσον αφορά τα εξής:

Ανίχνευση Προσώπου
Έχει ως σκοπό τον εντοπισμό προσώπων (θέση, διαστάσεις) εντός μιας εικόνας και πιθανότητα την εξαγωγή τους για να χρησιμοποιηθούν από τη διαδικασία της αναγνώρισης προσώπων

Αναγνώριση Προσώπου
Παραλαμβάνει μια εικόνα προσώπου από την προηγούμενη διαδικασία και έχει ως σκοπό είτε **α)** να ταυτοποιήσει -κάνοντας μία $1x1$ σύγκριση- ότι το πρόσωπο αυτό ταυτίζεται με ένα συγκεκριμένο πρόσωπο που δέχθηκε ως είσοδο είτε **β)** να αναγνωρίσει το πρόσωπο πραγματοποιώντας $1xN$ συγκρίσεις με ένα σύνολο από εικόνες προσώπων.

Η αναγνώριση προσώπων είναι μια εύκολη διαδικασία για τον ανθρώπινο εγκέφαλο. Έρευνες [37] έχουν δείξει ότι ακόμη και ένα βρέφος τριών ημερών είναι σε θέση να διακρίνει μεταξύ γνωστών προσώπων.Σύμφωνα με τις εργασίες των David Hubel[1] και Torsten Wiesel[2] ο ανθρώπινος εγκέφαλος έχει εξειδικευμένους νευρώνες οι οποίοι ανταποκρίνονται σε συγκεκριμένα χαρακτηριστικά μιας οπτικής εικόνας -όπως οι γραμμές, οι ακμές, οι γωνίες και η κίνηση- και τα οποία επεξεργάζεται καταλλήλως και δημιουργεί πρότυπα. Πάνω σε αυτό το πλαίσιο λειτουργεί και η αναγνώριση προσώπων από υπολογιστές, στην εξαγωγή συγκεκριμένων χαρακτηριστικών από μια εικόνα, την αναπαράστασή τους με μια αναγνωρίσιμη από υπολογιστές μορφή και τη διενέργεια κάποιου είδους ταξινόμησης μέσα από αυτά.

Η πιο συνήθης προσέγγιση είναι η αναγνώριση προσώπων χρησιμοποιώντας κάποια γεωμετρικά χαρακτηριστικά του προσώπου. Μια πρώτη προσέγγιση πάνω στο προαναφερθείσα λογική περιγράφεται εδώ [18] όπου χαρακτηριστικά όπως τα μάτια, η μύτη, τα αυτιά κ.α. χρησιμοποιούνται για την κατασκευή διανυσμάτων σχετικά με τη θέση, την απόσταση και τη γωνίας μεταξύ τους. Στη συνέχεια υπολογίζεται η ευκλείδεια απόσταση μεταξύ των προηγουμένως υπολογισμένων διανυσμάτων και των διανυσμάτων μιας εικόνας με ένα γνωστό πρόσωπο. Συγκρίνοντας τις αποστάσεις με διάφορα δείγματα το πρόσωπο που αναγνωρίζεται θεωρείται το δείγμα με την μικρότερη απόσταση. Παρόλο που η μέθοδος αυτή δεν επηρεάζεται

[1] https://en.wikipedia.org/wiki/David_H._Hubel
[2] https://en.wikipedia.org/wiki/Torsten_Wiesel

από τις αλλαγές στο φωτισμό, σύμφωνα με μελέτες [5], τα γεωμετρικά χαρακτηριστικά ενός προσώπου δεν παρέχουν αρκετή πληροφορία για ακριβή αποτελέσματα.

Κατά καιρούς διάφορες μέθοδοι έχουν προταθεί. Οι πιο χαρακτηριστικές είναι:

1. *EigenFaces - 1991 [38]*
2. *Local Binary Patterns Histograms - 1996 [1]*
3. *FisherFaces - 1997 [4]*
4. *Scale Invariant Feature Transform (SIFT) - 1999 [25]*
5. *Speed Up Robust Features (SURF) - 2006 [3]*

Οι μέθοδοι EigenFaces και FisherFaces, όπως και οι SIFT και SURF χρησιμοποιούν τις ίδιες τεχνικές για την εξαγωγή των χαρακτηριστικών μιας εικόνας και τη σύγκρισή τους με το σύνολο των εικόνων που δέχτηκαν ως είσοδο.

Παρακάτω θα δοθεί μια περιγραφή για τις τρεις πρώτες μεθόδους και να αναλυθεί το σημείο της μεθόδου LBPH που τροποποιήθηκε στα πλαίσια αυτής της διπλωματικής.

4.1 Η μέθοδος EigenFaces

Το πρόβλημα με την αναπαράσταση εικόνας που χρησιμοποιείται είναι ότι αναπαρίσταται στο διανυσματικό χώρο με ένα διάνυσμα πολλών διαστάσεων. Για παράδειγμα μια ασπρόμαυρη εικόνα δύο (2) διαστάσεων pxq αναπαρίσταται με ένα διάνυσμα $m = pq$-θέσεων. Οπότε μια εικόνα $100x100$ pixels παράγει ένα διάνυσμα $10,000$ θέσεων. Στην πραγματικότητα όμως για να εξάγουμε συμπεράσματα στο εν λόγω πρόβλημα της αναγνώρισης προσώπων δε μας είναι απαραίτητη το σύνολο της πληροφορίας που υπάρχει σε ολόκληρη την εικόνα. Χρειάζεται να εντοπίσουμε τις υποπεριοχές της εικόνας οι οποίες περιέχουν την πληροφορία που χρειαζόμαστε για την αναγνώριση του προσώπου. Η βασική ιδέα για να επιλέξουμε την πληροφορία που χρειαζόμαστε είναι ότι ένα διάνυσμα πολλών διαστάσεων συνήθως περιγράφεται από ένα σύνολο μεταβλητών που συσχετίζονται μεταξύ τους και επομένως χρειάζονται μόνο ορισμένες για να εξαχθεί η πληροφορία που χρειαζόμαστε. Η παραπάνω ιδέα αναλύεται στη μέθοδο Principal Component Analysis (PCA) η οποία προτάθηκε τόσο από τον Karl Pearson [3] (1901) όσο και από τον Harold Hotelling [4] (1933) και περιγράφει πως ένα σύνολο πιθανώς συσχετιζόμενων μεταβλητών περιγράφεται επαρκώς από ένα μικρότερο σύνολο μη-συσχετιζόμενων μεταβλητών.

4.1.1 Περιγραφή του αλγορίθμου

Έστω $X = \{x_1, x_2, ..., x_n\}$ ένα τυχαίο διάνυσμα όπου κάθε $x_i \in \mathbb{R}^d$

1. Υπολόγισε τον μέσο μ

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

[3] https://en.wikipedia.org/wiki/Karl_Pearson
[4] https://en.wikipedia.org/wiki/Harold_Hotelling

2. Υπολόγισε την μήτρα συνδιακύμανσης S

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T \tag{2}$$

3. Υπολόγισε τις ιδιοτιμές (eigenvalues) λ_i και τα ιδιοδιανύσματα (eigenvectors) v_i

$$Sv_i = \lambda_i v_i \tag{3}$$

4. Ταξινόμησε τα ιδιοδιανύσματα σε φθίνουσα σειρά σύμφωνα με τις ιδιοτιμές τους. Τα k κυριότερα στοιχεία είναι τα ιδιοδιανύσματα που αντιστοιχούν στις k μεγαλύτερες ιδιοτιμές.

Τα k κυριότερα στοιχεία του διανύσματος x δίνονται από τον ακόλουθο τύπο

$$y = W^T(x - \mu) \tag{4}$$

όπου $W = (v_1, v_2, \ldots, v_k)$
Η ανακατασκευή του διανύσματα χρησιμοποιώντας τη μέθοδο PCA προκύπτει

$$x = Wy + \mu \tag{5}$$

Η αναγνώριση του προσώπου με τη μέθοδο Eigenface γίνεται με τον εξής τρόπο:

1. Αρχικά αναπαριστούμε με τη μέθοδο PCA όλες τις εικόνες που χρησιμοποιούμε ως δείγματα χρησιμοποιώντας την εξίσωση 4
2. Αναπαριστούμε με την μέθοδο PCA την προς αναγνώριση εικόνα με βάση την εξίσωση 5
3. Συγκρίνουμε την προηγούμενη αναπαράσταση με τις αναπαραστάσεις των δειγμάτων και υπολογίζουμε τον κοντινότερο γείτονα

Από πλευράς πολυπλοκότητας, η μέθοδος που ακολουθήσαμε εμπεριέχει ένα πρόβλημα υπολογιστικού χώρου Ας υποθέσουμε ότι μας δίνονται ως δείγματα 400 εικόνες μεγέθους 100 x 100 pixels. Ακολουθώντας τη μέθοδο PCA, πρέπει να υπολογίσουμε τη μήτρα συνδιακύμανσης $S = XX^T$, όπου *Θ(X)* = 10000 x 400. Καταλήγουμε με αυτόν τον τρόπο με έναν πίνακα 10000 x 10000, ο οποίος καταλαμβάνει χώρο περίπου *0.8GB*. Προκειμένου να μειώσουμε αυτό το χώρο, μπορούμε να χρησιμοποιήσουμε από τη γραμμική άλγεβρα τη θέση ότι ένας πίνακας MxN όπου $>$ μπορεί να έχει μόνο $N - 1$ μη μηδενικές ιδιοτιμές. Επομένως χρησιμοποιώντας την ιδιοαποσύνθεση του S, $S = X^TX$ μεγέθους NxN έχουμε:

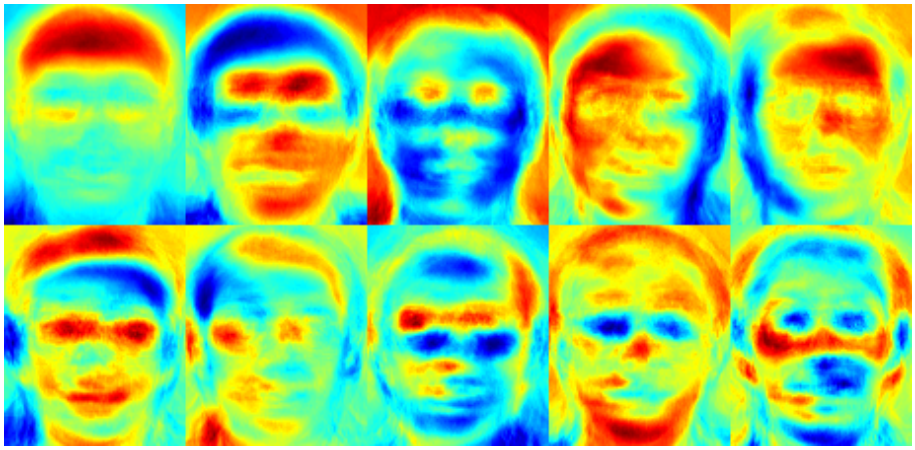
$$X^TXv_i = \mu_i v_i \tag{6}$$

και από εκεί υπολογίζουμε τα ιδιοδιανύσματα $S = XX^T$ κάνοντας αριστερό πολλαπλασιασμό πινάκων:

$$XX^T(Xv_i) = \mu_i(Xv_i) \tag{7}$$

Από τα παραγόμενα ορθογώνια ιδιοδιανύσματα, υπολογίζουμε τα αντίστοιχα ορθοκανονικά.

Τα EigenFaces οπτικοποιούνται ακολούθως:



Σχήμα 4.1: EigenFaces για 10 πρόσωπα της βάσης δεδομένων AT & T Facedatabase [5]

4.2 Η μέθοδος FisherFaces

Μία άλλη μέθοδος για τον περιορισμό της επεξεργάσιμης πληροφορία μέσα από μια εικόνα είναι η μέθοδος FisherFaces. Η μέθοδος αυτή στηρίζεται στη γραμμική διακριτή ανάλυση (Linear Discriminant Analysis) για τον περιορισμό των διαστάσεων μιας εικόνας και προτάθηκε από τον Sir R. A. Fisher [6]. Ο Fisher ,το 1936, ταξινόμησε με επιτυχία τα *λουλούδια* ως κλάση αντικειμένου. Το αρνητικό της μεθόδου PCA είναι ότι είναι ευάλωτη σε εξωτερικές πηγές. Πιο συγκεκριμένα, η μέθοδος PCA βρίσκει έναν γραμμικό συνδυασμό χαρακτηριστικών της εικόνας ο οποίος μεγιστοποιεί τη διακύμανση ωφέλιμης πληροφορίας. Ο τρόπος αυτός ενώ μας παρέχει έναν ισχυρό τρόπο αναπαράστασης της πληροφορίας, δεν λαμβάνει υπ' όψιν την κλάση του αντικειμένου με αποτέλεσμα διακεκριμένη πληροφορία που αφορά τη συγκεκριμένη κλάση αντικειμένου να χάνεται κατά τη μέθοδο. Ας υποθέσουμε ότι εισάγεται θόρυβος στην πληροφορία της εικόνας από εξωτερικό φως. Τα στοιχεία που αναγνωρίζονται από την ανάλυση PCA δεν περιέχουν απαραίτητα κάποια συγκεκριμένη πληροφορία σχετικά με το θόρυβο από το μια εξωτερική πηγή φωτός. Αυτό καθιστά την ταξινόμηση του προσώπου αδύνατη. Η μέθοδος της Linear Discriminant Analysis εστιάζει στην ανεύρεση των χαρακτηριστικών εκείνων που εντοπίζουν καλύτερα τις διαφορές μεταξύ διάφορων κλάσεων αντικειμένων. Η ιδέα βασίζεται στο ότι όμοιες κλάσεις ομαδοποιούνται κοντά ενώ διαφορετικές κλάσεις έχουν απόσταση μεταξύ τους.

4.2.1 Περιγραφή του αλγορίθμου

Έστω X τυχαίο διάνυσμα με δείγματα από C κλάσεις:

$$X = \{X_1, X_2, \ldots, X_c\}$$
$$X_i = \{x_1, x_2, \ldots, x_n\}$$

Οι διεσπαρμένοι πίνακες S_B και S_W υπολογίζονται ως εξής:

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

[6] https://en.wikipedia.org/wiki/Ronald_Fisher

$$S_W = \sum_{i=1}^{c} \sum_{x_j \in X_i} (x_j - \mu_i)(x_j - \mu_i)^T$$

όπου μ είναι ο συνολικός μέσος

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

και μ_i είναι ο μέσος της κλάσης $i \in \{1, \ldots, c\}$

$$\mu_i = \frac{1}{|X_i|} \sum_{x_j \in X_i} x_j$$

Ο κλασικός αλγόριθμος του Fisher υπολογίζει την προβολή W η οποία μεγιστοποιεί το κριτήριο διαχωριστικότητας των κλάσεων ως:

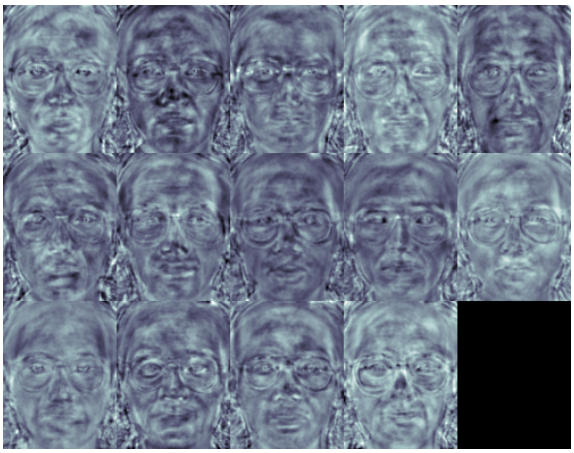
$$W_opt = \arg\max_W \frac{|W^T S_B W|}{|W^T S_W W|}$$

Μια λύση για αυτό το πρόβλημα βελτιστοποίησης δίνεται από το γενικότερο Eigenvalue Problem [4]

$$S_B v_i = \lambda_i S_w v_i$$

$$S_W^{-1} S_B v_i = \lambda_i v_i$$

Τα FisherFaces οπτικοποιούνται ακολούθως:



Σχήμα 4.2: FisherFaces για 16 πρόσωπα της βάσης δεδομένων Yale Facedatabase A
7

Στο σημείο αυτό αξίζει να σημειώσουμε ότι η μέθοδος FisherFace επικεντρώνει στην επισήμανση των διαφορών μεταξύ των προσώπων και συνεπώς είναι ευαίσθητη στις μεταβολές του φωτισμού. Ένα σύστημα αναγνώρισης με τη μέθοδο FisherFace εκπαιδευμένο με πρόσωπα εκτεθειμένα σε έντονο φωτισμό δυσκολεύεται να αναγνωρίσει πρόσωπα σε χαμηλό φωτισμό. Επομένως, η ποιότητα των αποτελεσμάτων της μεθόδου εξαρτάται σε πολύ μεγάλο βαθμό από τα δεδομένα (πρόσωπα) εισόδου με τα οποία εκπαιδεύεται ο αλγόριθμος.

4.3 Η μέθοδος Local Binary Patterns Histograms (LBPH)

Η ιδέα πίσω από τη μέθοδο LBPH έγκειται στην εξαγωγή κάποιον τοπικών χαρακτηριστικών από την εικόνα. Με αυτό τον τρόπο αποφεύγεται η αντιμετώπιση της εικόνας ως ένας πίνακα πολλών διαστάσεων. Παράλληλα το γεγονός ότι βασίζεται στην ανάλυση υφής (texture analysis) της εικόνας την καθιστά ανθεκτική σε αλλαγές στην φωτεινότητα, το μέγεθος και την γωνία (περιστροφή) της εικόνας.

Πιο συγκεκριμένα, τα Local Binary Patterns εφαρμόζονται στην αναπαράσταση της εικόνας στην κλίμακα grayscale. Υπολογίζουν μια νέα της αναπαράσταση συγκρίνοντας την ένταση κάθε εικονοστοιχείου (pixel) με αυτή των γειτονικών του. Στην καινούργια αναπαράσταση τα γειτονικά pixel λαμβάνουν την τιμή 1 ή 0 ανάλογα με το αν η ένταση τους είναι μεγαλύτερη ή μικρότερη από το κεντρικό. Τελικά η τιμή του κεντρικού pixel υπολογίζεται από την δεκαδική αναπαράσταση της παράθεσης των τιμών των γειτονικών pixel. Η διαδικασία αυτή επαναλαμβάνεται για όλα τα pixel της εικόνας.

Υπολογισμός Local Binary Pattern:

$$LBP(x_c, y_c) = \sum_{p=0}^{p-1} 2^p s(i_p - i_c)$$

όπου

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

Ας δούμε τη μέθοδο με περισσότερη ακρίβεια.

Οι βασικές παράμετροι του αλγορίθμου είναι οι εξής:

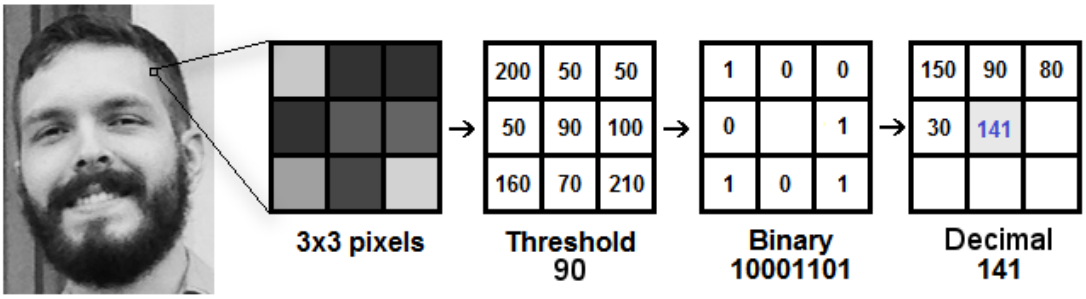
Radius (ακτίνα)
Έχει ως σκοπό τον εντοπισμό προσώπων (θέση, διαστάσεις) εντός μιας εικόνας και πιθανότητα την εξαγωγή τους για να χρησιμοποιηθούν από τη διαδικασία της αναγνώρισης προσώπων. Στην τυπική υλοποίηση της μεθόδου η τιμή **Radius** είναι 1.

Neighbors (πλήθος γειτόνων)
Παραλαμβάνει μια εικόνα προσώπου από την προηγούμενη διαδικασία και έχει ως σκοπό είτε **α)** να ταυτοποιήσει -κάνοντας μία $1x1$ σύγκριση- ότι το πρόσωπο αυτό ταυτίζεται με ένα συγκεκριμένο πρόσωπο που δέχθηκε ως είσοδο είτε **β)** να αναγνωρίσει το πρόσωπο πραγματοποιώντας $1xN$ συγκρίσεις με ένα σύνολο από εικόνες προσώπων. Στην τυπική υλοποίηση της μεθόδου η τιμή **Neighbors** είναι 8.

Grid X
Ο αριθμός των οριζόντιων κελιών στον οποίο θα χωριστεί η εικόνα. Όσα περισσότερα κελιά έχουμε τόσο μεγαλύτερη είναι η διάσταση του χάρτη χαρακτηριστικών της εικόνας. Στην τυπική υλοποίηση της μεθόδου η τιμή **Grid X** είναι 8.

Grid Y
Ο αριθμός των κάθετων κελιών στον οποίο θα χωριστεί η εικόνα. Όσα περισσότερα κελιά έχουμε τόσο μεγαλύτερη είναι η διάσταση του χάρτη χαρακτηριστικών της εικόνας. Στην τυπική υλοποίηση της μεθόδου η τιμή **Grid Y** είναι 8.

Χρησιμοποιώντας τις προκαθορισμένες τιμές για την ακτίνα και τους γείτονες έχουμε:

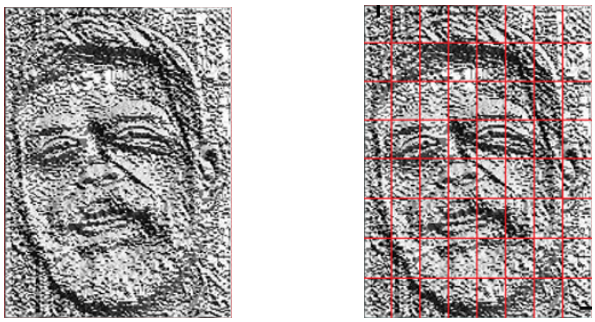


Σχήμα 4.3: Υπολογισμός των Local Binary Patterns

Το αποτέλεσμα είναι μια εικόνα που εμφανίζει σαφέστερα τα χαρακτηριστικά της αρχικής εικόνας.

Σχήμα 4.4: α) αρχική εικόνα σε grayscale, β) ύστερα από το μετασχηματισμό LBP

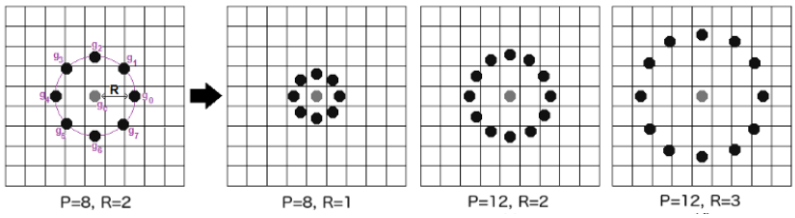
Στη συνέχεια χρησιμοποιώντας τις παραμέτρους **Grid X** και **Grid Y** η νέα εικόνα χωρίζεται σε ένα πλέγμα διαστάσεων (**Grid X** $*$ **Grid Y**).



Σχήμα 4.5: Εξαγωγή ιστογραμμάτων από κάθε υποπεριοχή του πλέγματος

Υπολογίζεται το ιστόγραμμα (histogram) κάθε υποπεριοχής που προκύπτει. Εφόσον η αρχική εικόνα ήταν σε ασπρόμαυρη κλίμακα (grayscale) το κάθε παραγόμενο ιστόγραμμα θα περιέχει 256 θέσεις ($0 - 255$). Εν τέλει συνθέτουμε όλα τα παραγόμενα ιστογράμματα σε ένα συνολικό το οποίο αναπαριστά τα χαρακτηριστικά της αρχικής εικόνας. Το τελικό ιστόγραμμα θα περιέχει (**Grid X** $*$ **Grid Y** $*$ 256 θέσεις).

Ανάλογα με την τιμή των παραμέτρων **Radius** και **Neighbors** τα pixel που συμμετέχουν στον υπολογισμό του Local Binary Pattern αλλάζουν. Για παράδειγμα:



Σχήμα 4.6: Extended Local Binary Patterns

Ο νέος αυτός τελεστής , οποίος διαφοροποιείται από το κλασσικό LBP, ονομάζεται Extended Local Binary Patterns.

4.3.1 Τυπική υλοποίηση

Στην τυπική υλοποίηση της μεθόδου LBPH, αρχικά κατασκευάζεται μια βάση από ιστογράμματα από ένα προκαθορισμένο σύνολο εικόνων (training dataset) το οποίο αποτελεί τη βάση αλήθειας της μεθόδου (ground truth). Για κάθε εικόνα στην οποία θέλουμε να αναγνωρίσουμε ένα πρόσωπο υπολογίζεται το αντίστοιχο ιστόγραμμα και στη συνέχεια με χρήση του αλγορίθμου 1-Nearest Neighbor [8] ανιχνεύεται η κλάση (πρόσωπο) καθώς και η ελάχιστη απόσταση μεταξύ των ιστογραμμάτων της αρχικής εικόνας και της κλάσης. Η εικόνα της κλάσης από την οποία προκύπτει η ελάχιστη διαφορά επιστρέφεται επίσης ως ένα μέτρο της αξιοπιστίας του αποτελέσματος.

Για να υπολογιστεί η απόσταση δύο ιστογραμμάτων μπορούν να χρησιμοποιηθούν διάφορα είδη αποστάσεων όπως `Ευκλείδεια απόσταση`, `Chi-Square`, `Απόλυτη τιμή`. Στην τυπική υλοποίηση χρησιμοποιείται η `Ευκλείδεια απόσταση`.

$$D = \sqrt{\sum_{i=1}^{n} (hist1_i - hist2_i)^2}$$

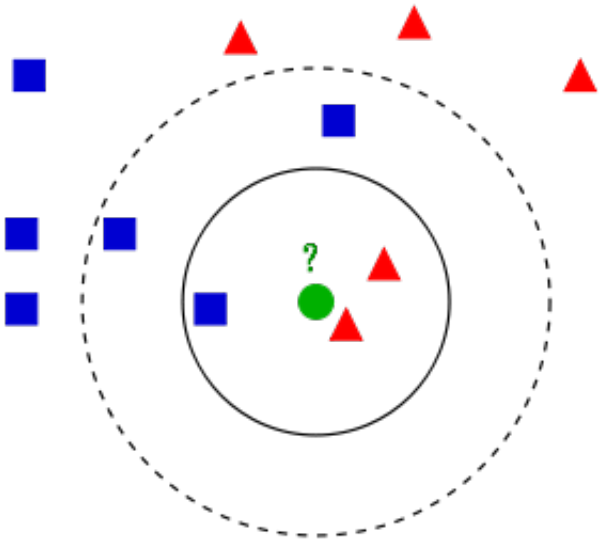
4.3.2 Υλοποίηση με τη χρήση του αλγορίθμου k-Nearest Neighbor

Στη δικιά μας υλοποίηση, αντικαταστήσαμε των αλγόριθμο προσδιορισμού του προσώπου (1-Nearest Neighbor) με τον αλγόριθμο k-Nearest Neighbor [9]. Η διαφορά των δύο μεθόδων βρίσκεται στον αριθμό των κοντινών δειγμάτων που θα χρησιμοποιηθούν

για τον προσδιορισμό της κλάσης Όπως βλέπουμε στο σχήμα 4.7 για $k = 1$ η κλάση που επιστρέφεται από τον αλγόριθμο είναι η 'κόκκινη' δηλαδή ουσιαστικά είναι η κλάση που

[8] `https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#The_1-nearest_neighbor_classifier`

[9] `https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm`



Σχήμα 4.7: Η διαφορά μεταξύ 1-Nearest Neighbor και k-Nearest Neighbor

βρίσκεται στην μικρότερη απόσταση από το δείγμα. Για $k = 3$ η κλάση που υπολογίζεται είναι και πάλι η 'κόκκινη' καθώς λαμβάνοντας υπόψιν τα 3 κοντινότερα σημεία στο δείγμα μας παρατηρούμε ότι τα 'κόκκινα' είναι περισσότερα από τα 'μπλε'. Όμως για $k = 5$ παρατηρούμε ότι τα 'μπλε' είναι περισσότερα από τα 'κόκκινα' οπότε σε αυτή την περίπτωση τη κλάση που επιστρέφει ο αλγόριθμος ταξινόμησης είναι η 'μπλε'.

Στο επόμενο κεφάλαιο θα αναλύσουμε με σαφήνεια τα αποτελέσματα της χρήσης του αλγορίθμου k-Nearest Neighbor για τον προσδιορισμό της κλάσης του προσώπου και παραθέσουμε τα συμπεράσματά μας.

Κεφάλαιο 5

Ανάλυση αποτελεσμάτων

5.1 Υλοποίηση

Η υλοποίηση της μεθόδου k-Nearest Neighbor έγινε στη γλώσσα C++ και πραγματοποιήθηκε ως προσθήκη στην υπάρχουσα υλοποίηση της μεθόδου από τη βιβλιοθήκη OpenCV [1]. Για την μεταγλώττιση του κώδικα και την παραγωγή των εκτελέσιμων αρχείων χρησιμοποιήθηκε ο μεταγλωττιστής g++ με έκδοση 4:6.3.0-4 σε περιβάλλον Debian Stretch Linux. Η πειραματική διάταξη υλοποιήθηκε στη γλώσσα Python2.7 χρησιμοποιώντας τους python wrappers της OpenCV. Όλα τα πειράματα εκτελέστηκαν σε υπολογιστή Intel(R) Core(TM) i7-2640M CPU @ 2.80GHz χωρίς χρήση GPU. Για την αξιολόγηση χρησιμοποιήθηκε το σύστημα της k-fold Cross Validation[2].

5.2 Πειραματική μεθοδολογία

5.2.1 Βάσεις δεδομένων

Για τις μετρήσεις χρησιμοποιήθηκαν οι βάσεις δεδομένων προσώπων AT&T Facedatabase [3], Yale Facedatabase A[4], Extended Yale Facedatabase B[5] καθώς και μια βάση κατασκευασμένη από το συγγραφέα με βάση κάποιο πολυμεσικό πειραματικό υλικό που δόθηκε από το Πανεπιστήμιο του Lucce[6].

AT&T Facedatabase

Η βάση AT&T Facedatabase περιέχει 10 διαφορετικές εικόνες για 40 διαφορετικά πρόσωπα. Για κάποια πρόσωπα οι εικόνες έχουν συλλεχθεί σε διάφορες χρονικές στιγμές, με ανομοιόμορφες συνθήκες φωτισμού και διαφορετικές εκφράσεις (χαμόγελο, ανοιχτά/κλειστά μάτια) και χαρακτηριστικά (πχ γυαλιά). Σε όλες τις εικόνες υπάρχει ομοιόμορφο μαύρο φόντο.

[1] https://opencv.org
[2] https://en.wikipedia.org/wiki/Cross-validation_(statistics)
[3] www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
[4] vision.ucsd.edu/content/yale-face-database
[5] http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html
[6] https://www.imtlucca.it/

Yale Facedatabase A

Περιέχει 15 πρόσωπα και 11 ασπρόμαυρες φωτογραφίες για το καθένα από αυτά. Οι εικόνες έχουν μέγεθος $320x243$ pixel ενώ υπάρχουν διαφοροποιήσεις στις συνθήκες φωτισμού, στις εκφράσεις του προσώπου και τα χαρακτηριστικά. Στη βιβλιογραφία θεωρείτε πιο αξιόπιστη από την παραπάνω.

Οι εικόνες των προσώπων των δύο αυτών βάσεων δεν είναι στις ακριβείς διαστάσεις των προσώπων και χρειάζονται μια προ-επεξεργασία για να μπορέσουν να αξιοποιηθούν.

Extended Yale Facedatabase B

Η βάση αυτή περιέχει 2414 εικόνες από 38 διαφορετικά πρόσωπα κομμένες ακριβώς στις διαστάσεις του κάθε προσώπου. Ο στόχος της βάσης είναι να μπορούν να εξαχθούν χαρακτηριστικά των προσώπων ανθεκτικά στις αλλαγές του φωτισμού. Για το λόγο αυτό όλα τα πρόσωπα έχουν σχεδόν τις ίδιες εκφράσεις και χαρακτηριστικά.

LucceFaces

Η βάση αυτή περιέχει 10 εικόνες για κάθε ένα από τα 12 πρόσωπα που περιέχει. Οι εικόνες των προσώπων εξήχθησαν από πραγματικό πολυμεσικό περιεχόμενο. Είναι μια βάση προσώπων για την αξιολόγηση του συγκεκριμένου dataset.

5.2.2 Πρωτόκολλο αξιολόγησης

Για την αξιολόγηση ενός αλγορίθμου αναγνώρισης προσώπων χρειάζεται να διαθέτουμε δύο βάσεις δεδομένων εικόνων. Μια που να αποτελεί τη βάση με την οποία εκπαιδεύεται η μέθοδος και μία πάνω στην οποία να εκτελείται. Το σχήμα που ακολουθήθηκε παρήγαγε και τις δύο βάσεις από τις παραπάνω χρησιμοποιώντας την τεχνική Cross-validation[7] Η μέθοδος εκτελείται n επαναλήψεις και σε κάθε μία αξιολογείται από τη μέθοδο αξιολόγησης k-fold Cross-validation. Στο σύνολο των επαναλήψεων υπολογίζεται η μέση ακρίβεια (accuracy). Η υλοποίηση της μεθόδου δίνει μία πρόβλεψη για κάθε πρόσωπο η οποία είτε είναι σωστή (true positive) είτε λάθος (false positive). Σε κάθε περίπτωση δεν έχουμε καθόλου true negatives και false negatives. Επομένως στη συγκεκριμένη περίπτωση το μέτρο `Accuracy` που υπολογίζουμε ταυτίζεται με το `Precision` ενώ το `Recall` θα παραμένει πάντα σταθερό και ίσο με 1. Το Precision και το Recall είναι οι τυπικές μετρικές για την αξιολόγηση αλγορίθμο που πραγματοποιούν ταξινόμηση κλάσης.

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

k-fold Cross Validation

Με τη μέθοδο αυτή χτίζονται k μη επικαλυπτόμενα σετ εκπαίδευσης και πειραματισμού από την αρχική βάση δεδομένων. Έτσι αποφεύγονται συνθήκες όπου ολόκληρη η μέθοδος είναι εκπαιδευμένη με χαρούμενα πρόσωπα ενώ το σετ εικόνων που χρησιμοποιείται στα πειράματα περιέχει λυπημένα πρόσωπα. Παρακάτω παρουσιάζεται ένα μικρό παράδειγμα της μεθόδου 4-fold Cross Validation:

[7] `https://en.wikipedia.org/wiki/Cross-validation_(statistics)`

	Model 1	Model 2	Model 3	Model 4
Fold 1	Test data	Training data	Training data	Training data
Fold 2	Training data	Test data	Training data	Training data
Fold 3	Training data	Training data	Test data	Training data
Fold 4	Training data	Training data	Training data	Test data

Σχήμα 5.1: Η μέθοδος k-fold Cross Validation

Παρατηρούμε ότι σε κάθε επανάληψη (fold) επιλέγονται καινούργια και ανεξάρτητα σετ εκπαίδευσης και πειραματισμού.

Η ακρίβεια που υπολογίζεται είναι ο μέσος όρος των τιμών της ακρίβειας που υπολογίζονται σε κάθε fold.

5.3 Αποτελέσματα

Με βάση τα χαρακτηριστικά (αριθμός προσώπων και εικόνες) των βάσεων δεδομένων που χρησιμοποιήσαμε επιλέξαμε να αξιολογήσουμε τη μέθοδο με **4-fold** και **10-fold** cross validation για μεγαλύτερη αξιοπιστία στα αποτελέσματα.

Αξιολογήσαμε την τροποποιημένη μέθοδο για διαφορετικές τιμές του **k** και για 3 διαφορετικά είδη αποστάσεων, την τυπική **Eucledian** απόσταση, την **ChiSquare** και την **Cosine** απόσταση. Παρακάτω φαίνονται τα πειράματα για τις 4 βάσεις δεδομένων.

Δίνονται οι ορισμοί των αποστάσεων:

`Ευκλείδεια απόσταση`

$$d(H_1, H_2) = \sqrt{H_1^2 + H_1^2}$$

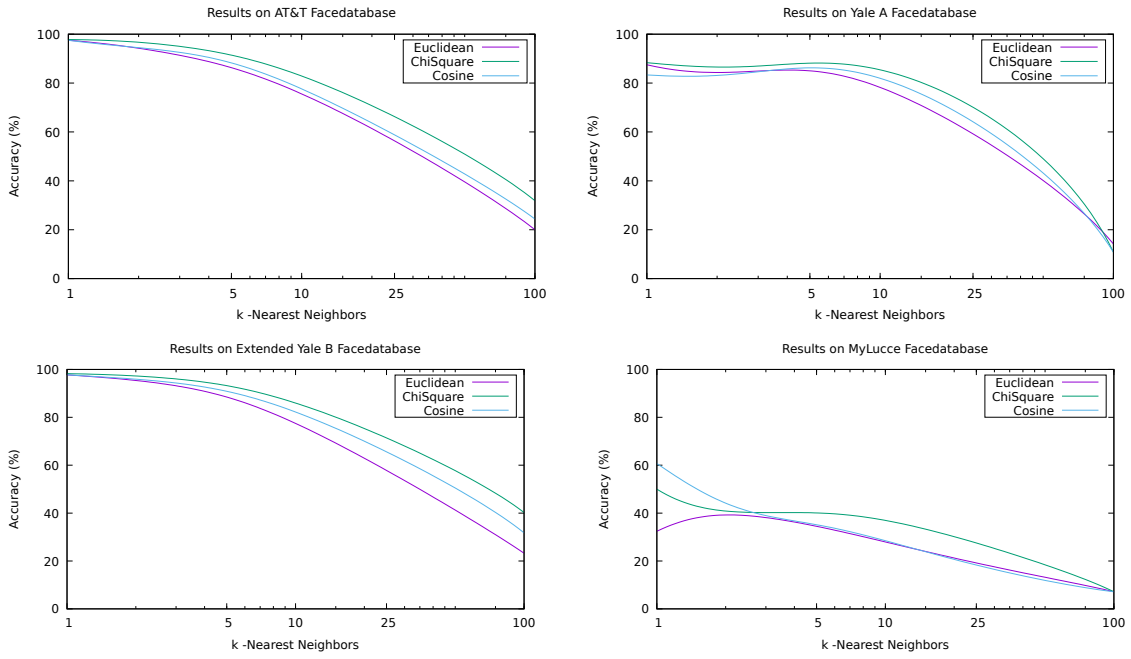
`Chisquare απόσταση`

$$d(H_1, H_2) = \sum \frac{(H_1 - H_2)^2}{H_1 + H_2}$$

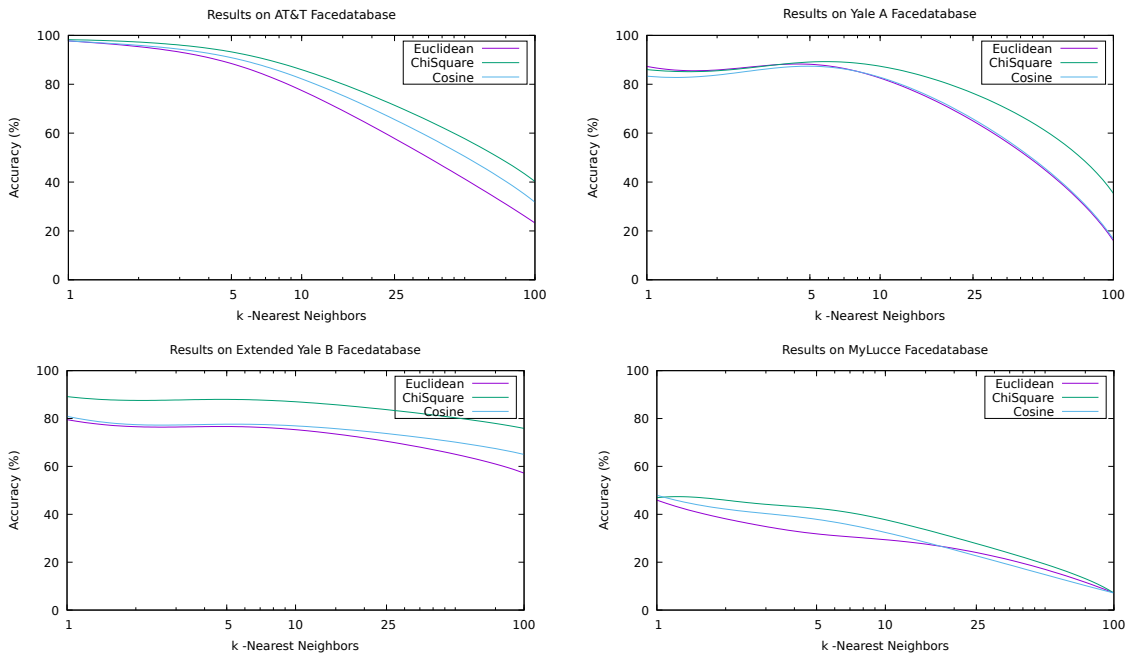
`Cosine απόσταση`

$$d(H_1, H_2) = -\frac{H_1^T * H_2}{\sqrt{(H_1 * H_1^T) x (H_2 * H_2^T))}}$$

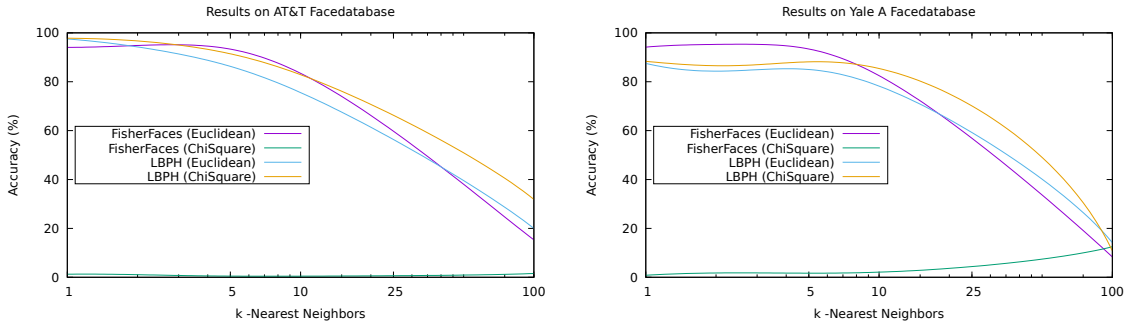
Επίσης για διαφορετικές τιμές του **k** συγκρίναμε τη μέθοδό μας για την τυπική Ευκλείδεια απόσταση και την Chisquare (που φαίνεται να υπερτερεί) με τη μέθοδο **FisherFaces** ως τον εξαγωγέα των χαρακτηριστικών από το πρόσωπο.



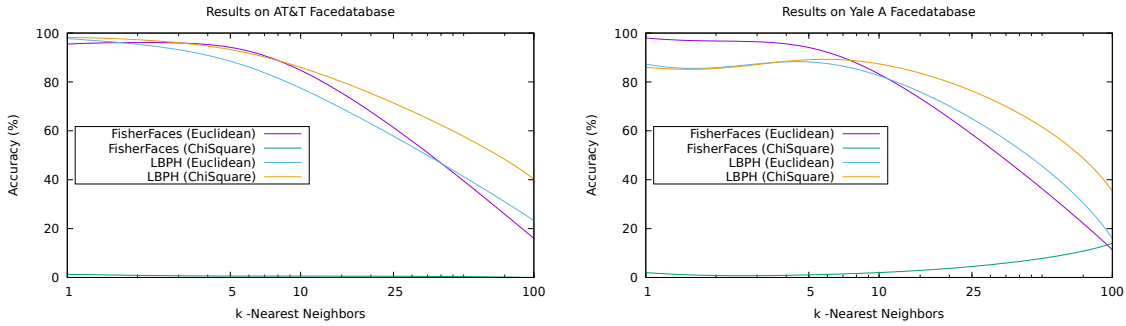
Σχήμα 5.2: k-4 cross validation



Σχήμα 5.3: k-10 cross validation



Σχήμα 5.4: k-4 cross validation



Σχήμα 5.5: k-10 cross validation

Κεφάλαιο 6

Επίλογος

6.1 Συμπεράσματα

Αυτό που εύκολα παρατηρεί κανείς είναι το γεγονός ότι η χρήσης της απόσταση *ChiSquare* για τη σύγκριση των ιστογραμμάτων στη μέθοδό μας φαίνεται να δίνει μεγαλύτερη ακρίβεια για όλες της βάσεις προσώπων που χρησιμοποιήσαμε.

Παρατηρούμε επίσης μια σημαντική διαφορά στην ακρίβεια της μεθόδου στη βάσε MyLucce. Το γεγονός αυτό οφείλεται στο ότι η βάση MyLucce δημιουργήθηκε παίρνοντας τις εικόνες των προσώπων από πραγματικό πολυμεσικό υλικό (βίντεο). Ως αποτέλεσμα δεν είχαμε καθόλου επιρροή στις συνθήκες φωτισμού, στις εκφράσεις των προσώπων και στα χαρακτηριστικά σε αντίθεση με τις άλλες βάσεις οι οποίες σχεδιάστηκαν ειδικά για σκοπούς πειραματισμού. Θα μπορούσαμε όμως να θεωρήσουμε ότι η βάση MyLucce μας δίνει μια εικόνα για την ακρίβεια της μεθόδου όταν προσπαθούμε να κάνουμε αναγνώριση χωρίς να έχουμε ένα προεκπαιδευμένο σύστημα (Unsupervised Learning[1])

Όσος αφορά την αξιολόγηση της τροποποιημένης μεθόδου, παρατηρούμε μια διαφοροποίηση στη συμπεριφορά της ανάμεσα στη βάση AT&T και τις υπόλοιπες. Στη βάση AT&T φαίνεται πως οποιαδήποτε αύξηση της τιμής του k μειώνει την ακρίβειά της συνολικά. Όμως όπως προαναφέραμε, λόγω των συνθηκών δημιουργίας της, η βάση AT&T είναι χρήσιμη για κάποια αρχικά πειράματα αλλά στην ουσία θεωρείται μια 'εύκολη' σχετικά βάση. Επομένως δεν προσφέρεται για ασφαλή συμπεράσματα. Αντίθετα στις βάσεις Yale A και Yale B ακόμη και σε ένα βαθμό και στην MyLucce παρατηρούμε ότι η αύξηση της τιμής του k δαιτηρεί σταθερή την ακρίβεια ή ακόμη και την αυξάνει μέχρι ένα σημείο καμπής όπου από εκεί και πέρα η ακρίβεια της μεθόδου πέφτει δραματικά. Είναι λοιπόν εμφανές ότι αρκετές φορές πρέπει να συμπεριλάβουμε περισσότερους από έναν κοντινότερους γείτονες στη πρόβλεψη για να έχουμε ένα καλύτερο αποτέλεσμα.

Τέλος στα τελευταία γραφήματα επιχειρήσαμε και μια σύγκριση της μεθόδου με τη μέθοδο FisherFaces ή οποία δεν βασίζεται στα local binary pattern histograms αλλά στην Linear Discriminant Analysis 4.2. Η συμπεριφορά της FisherFaces είναι ίδια όσο αυξάνεται το k. Όμως η ακρίβειά της φθίνει πολύ πιο γρήγορα από εκείνη της μεθόδου μας. Αντίθετα, στα μικρά k η ακρίβεια είναι σχετικά ίδια και για τις δύο μεθόδους.

Συμπερασματικά, θα λέγαμε ότι η τροποποίηση που πραγματοποιήσαμε εδώ κάποια ενδιαφέροντα στοιχεία. Το βασικότερο είναι ότι η τιμή του k όντως επηρεάζει την ακρίβεια της

[1] https://en.wikipedia.org/wiki/Unsupervised_learning

μεθόδου και κατά περιπτώσεις την αυξάνει, μέχρι ένα σημείο καμπής. Γίνεται λοιπόν εμφανές ότι αν θέλουμε να μειώσουμε τα false positives της μεθόδου LBPH πρέπει η τελική πρόβλεψη να μην βασίζεται μόνο στον κοντινότερο γείτονα αλλά και στους υπολοίπους. Πρέπει να γίνεται ξεκάθαρο σε κάθε βάση δεδομένων πιο είναι το σημείο καμπής της όπου από κει και πέρα η ακρίβεια πέφτει δραματικά. Αυτό εξαρτάται από της συνθήκες των προσώπων της κάθε βάσης. Και τέλος να σημειώσουμε το εύρος της διαφοράς στην ακρίβεια που παρουσιάζει η αναγνώριση με μια βάση κατασκευασμένη σε ελεγχόμενες συνθήκες και μια βάση δημιουργημένη με μη ελεγχόμενες.

6.2 Προτάσεις για μελλοντική έρευνα

Η ανίχνευση προσώπων και αντικειμένων καθώς και η αναγνώριση προσώπων, είναι προβλήματα που απασχολούν ιδιαίτερα την επιστημονική κοινότητα τα τελευταία χρόνια, με μεθόδους και τεχνικές να εμφανίζονται με εξαιρετικά μεγάλο ρυθμό. Ήδη από την στιγμή που αρχίσαμε την εκπόνηση της εργασίας μέχρι σήμερα έχουν γίνει εξαιρετικά βήματα προς την επίτευξη καλύτερων και πιο γρήγορων ανιχνεύσεων.

Ειδικότερα, οι σύγχρονες τεχνικές βασίζονται πλέον στα νευρωνικά δίκτυα. Αρχικά χρησιμοποιείται ένα νευρωνικό δίκτυο για την εξαγωγή υποψήφιων θέσεων ενός αντικειμένου και στη συνέχεια οι θέσεις αυτές ταξινομούνται στην κατάλληλη κλάση και εξάγεται το αντίστοιχο παράθυρο με τη χρήση κάποιου άλλου συνελικτικού κυρίως νευρωνικού δικτύου. Τα τελευταία χρόνια εμφανίστηκε η τάση της εξαγωγής της κλάσης και του παραθύρου ταυτόχρονα με τον προσδιορισμό των υποψήφιων θέσεων των αντικειμένων. Οι ανιχνευτές μονής λήψης όπως ονομάζονται φαίνεται να δίνουν της κατεύθυνση της ανίχνευσης αντικειμένων στο μέλλον καθώς παρουσιάζουν αρκετά καλή ακρίβεια σε μικρό χρόνο εκτέλεσης.

Παράλληλα, λόγω της αυξημένης χρήσης συσκευών περιορισμένης επεξεργαστικής ισχύς εμφανίζεται η ανάγκη περιορισμού των απαιτήσεων που χρειάζονται οι παραπάνω μέθοδοι τόσο σε μνήμη όσο και σε επεξεργαστική ισχύ. Το νευρωνικό δίκτυο MobileNet σχεδιάστηκε με τέτοιο τρόπο ώστε η μέθοδος ανίχνευσης αντικειμένου που το χρησιμοποιεί να μπορεί να τρέξει με μια συσκευή κινητής τηλεφωνίας. Επομένως φαίνεται να πως ενώ στον τομέα της ακρίβειας του αποτελέσματος της ανίχνευσης έχει επιτευχθεί ένα ικανοποιητικό ποσοστό, το μέλλον προϋποθέτει την προσαρμογή των μεθόδων σε συσκευές με μικρότερες δυνατότητες όπως κινητά τηλέφωνα και tablets με τη χρήση 'ελαφριών' νευρωνικών δικτύων.

Βιβλιογραφία

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In T. Pajdla and J. Matas, editors, *Computer Vision - ECCV 2004*, pages 469--481, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189--2202, Nov 2012.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346 -- 359, 2008. Similarity Matching in Computer Vision and Multimedia.

[4] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. 19:711--720, 07 1997.

[5] R. Brunelli and T. Poggio. Face recognition through geometrical features. In *Proceedings of the Second European Conference on Computer Vision*, ECCV '92, pages 792--800, Berlin, Heidelberg, 1992. Springer-Verlag.

[6] A. Carlson, C. Cumby, J. Rosen, and D. Roth. The snow learning architecture, 5 1999.

[7] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1312--1328, July 2012.

[8] M. Cheng, Z. Zhang, W. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3286--3293, June 2014.

[9] R. G. Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with fisher vectors. In *2013 IEEE International Conference on Computer Vision*, pages 2968--2975, Dec 2013.

[10] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016.

[11] I. Endres and D. Hoiem. Category-independent object proposals with diverse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):222--234, Feb 2014.

[12] M. Everingham, S. M. Eslami, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111(1):98--136, Jan. 2015.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627--1645, Sept 2010.

[15] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000, 1998.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[17] M. hsuan Yang, D. Roth, and N. Ahuja. A snow-based face detector. In *Advances in Neural Information Processing Systems 12*, pages 855--861. MIT Press, 2000.

[18] T. Kanade. Picture processing system by computer complex and recognition of human faces, November 1973.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097--1105. Curran Associates, Inc., 2012.

[20] R. Lienhart, A. Kuranov, and V. Pisarevsky. *Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection*, pages 297--304. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[21] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *IEEE ICIP 2002*, pages 900--903, 2002.

[22] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.

[23] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.*, 2(4):285--318, Apr. 1988.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[25] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150--1157 vol.2, Sept 1999.

[26] M. Oren, C. Papageorgiou, P. Sinha, and E. Osuna. Pedestrian detection using wavelet templates. In *Proceedings of CVPR'97, Puerto Rico*, 1997.

[27] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int. J. Comput. Vision*, 38(1):15--33, June 2000.

[28] J. Redmon. Darknet: Open source neural networks in c. `http://pjreddie.com/darknet/`, 2013--2016.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779--788, June 2016.

[30] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.

[31] D. Roth. Learning to resolve natural language ambiguities: A unified approach. In *In Proceedings of the National Conference on Artificial Intelligence. 806-813. Segond F., Schiller A., Grefenstette and Chanod F.P*, 1998.

[32] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(1):23--38, 1 1998.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[34] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, 26(5):1651--1686, 10 1998.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[36] C. Szegedy, S. E. Reed, D. Erhan, and D. Anguelov. Scalable, high-quality object detection. *CoRR*, abs/1412.1441, 2014.

[37] C. Turati, V. Macchi Cassia, F. Simion, and I. Leo. Newborns' face recognition: Role of inner and outer facial features. 77:297--311, 03 2006.

[38] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71--86, Jan. 1991.

[39] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154--171, 2013.

[40] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *2011 International Conference on Computer Vision*, pages 1879--1886, Nov 2011.

[41] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. pages 511--518, 2001.

[42] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137--154, 2004.