

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ  
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΠΙΣΤΗΜΩΝ  
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

**Ανίχνευση Προσώπων Βασισμένη σε Συννελικτικά  
Νευρωνικά Δίκτυα**

Μανόλης Δελάκης

Μεταπτυχιακή Εργασία

Ηράκλειο, Ιούλιος 2003



*Αφιερώνεται στους γονείς μου και στα αδέρφια μου*



# *Ανίχνευση Προσώπων Βασισμένη σε Συνελικτικά Νευρωνικά Δίκτυα*

Μανόλης Δελάκης

Μεταπτυχιακή Εργασία

Τμήμα Επιστήμης Υπολογιστών  
Πανεπιστήμιο Κρήτης

## **Περίληψη**

Η ανίχνευση ανθρώπινων προσώπων αν και επιτελείται στιγμιαία, αβίαστα και με ενδεικτική ακρίβεια από τον ανθρώπινο εγκέφαλο, για την έρευνα της υπολογιστικής όρασης είναι ακόμα ένα θέμα υπό ανάπτυξη. Επιπρόσθετα, το μεγάλο εύρος πρακτικών εφαρμογών της όπως η αυτόματη προετοιμασία δεδομένων για την αναγνώριση προσώπου, η ανάκτηση εικόνας με βάση το περιεχόμενο ή η προχωρημένη αλληλεπίδραση μεταξύ ανθρώπου και μηχανής, την καθιστούν ως ένα πρόβλημα με θεωρητική αλλά και πρακτική αξία. Ο σκοπός της παρούσας εργασίας ήταν η εισαγωγή των Συνελικτικών Νευρωνικών Δικτύων σαν ένας αποτελεσματικός και ταχύς ανιχνευτής προσώπων, ικανός να λειτουργεί σε μη ελεγχόμενα περιβάλλοντα και χωρίς καμία προεπεξεργασία.

Μία συνελικτική νευρωνική τοπολογία προτείνεται, σχεδιασμένη ώστε να είναι σθεναρή σε μεταβλητές συνθήκες εικόνας και έκφρασης προσώπου ή σε άλλες δυνατές παραμορφώσεις της εισόδου. Το δίκτυο εκπαιδεύτηκε με ένα αρκετά μεγάλο σύνολο εκπαίδευσης, άμεσα προερχόμενο από φυσικά δεδομένα, μέσω του αλγορίθμου backpropagation. Χρησιμοποιώντας τα εκπαιδευμένα φίλτρα του δικτύου, επινοήθηκε μία γρήγορη διαδικασία για την σάρωση της εικόνας, βασισμένη ολοκληρωτικά σε απλές λειτουργίες επεξεργασίας εικόνας.

Το σύστημα δοκιμάστηκε σε μία σειρά από μεγάλα και δύσκολα σύνολα δοκιμής, επιδεικνύοντας πολύ υψηλά ποσοστά ανίχνευσης με λίγες και σποραδικές εσφαλμένες ειδο-

ποιήσεις. Η σύγκριση με τα τρέχοντα πρότυπα συστήματα σε κοινά αναφερόμενα σύνολα απεκάλυψε ότι το προτεινόμενο σύστημα είναι ο καλύτερης απόδοσης ανιχνευτής προσώπων γενικής χρήσης της βιβλιογραφίας. Επιπλέον, η ανοχή του δικτύου σε μία σειρά από δυνατές παραμορφώσεις της εισόδου μετρήθηκε και επιβεβαιώθηκε σε πειράματα ανάλυσης της ενασθησίας.

Επόπτης: Γεώργιος Τζιρίτας  
Καθηγητής  
Τμήμα Επιστήμης Υπολογιστών  
Πανεπιστήμιο Κρήτης

Επιβλέπων: Christophe Garcia  
Επικεπτης καθηγητής  
Τμήμα Επιστήμης Υπολογιστών  
Πανεπιστήμιο Κρήτης

# *Face Detection Based on Convolutional Neural Networks*

Manolis Delakis

Master of Science Thesis

Computer Science Department  
University of Crete

## **Abstract**

Human face detection even though is performed instantly, effortlessly and with indicative accuracy by the human brain, for machine vision research is a matter still under development. Moreover, its wide range of applications like automatic data preparation for face recognition, content-based image retrieval or advanced human and computer interaction, make it a problem with both theoretical and practical values. The aim of this study was to introduce Convolutional Neural Networks as an efficient and fast face detector, able to operate in un-controlled image environments and without preprocessing.

A convolutional neural topology is proposed, designed to be robust in varying image conditions and facial expression or other input deformations. The network was trained over a large enough training set of face patterns, coming directly from natural data, via the backpropagation algorithm. Using the trained convolutional filters of the network, a fast procedure for image scanning and face localization was devised, based purely on basic image processing operations.

The system was tested in a series of large and difficult test sets exhibiting very high detection rates with a few and sporadic false alarms. The comparison with the current state-of-the-art systems in common benchmark sets revealed that the proposed system is the best performing general-purposed face detector of the reported literature. Furthermore, the tolerance of the network in a series of possible input deformations was measured and verified

in conducted sensitivity analysis experiments.

Responsible Professor: George Tziritas  
Professor  
Computer Science Department  
University of Crete

Supervisor: Christophe Garcia  
Visiting associate professor  
Computer Science Department  
University of Crete

# Ευχαριστίες

Καταρχήν, θα ήθελα να ευχαριστήσω τον επιβλέποντα της εργασίας κ. Christophe Garcia. Τον ευχαριστώ για την μύηση που μου προσέφερε στον μαγικό κόσμο της αναγνώρισης προτύπων, όσο και για την απλόχερη βοήθειά του κατά τη διάρκεια της συνεργασίας μας. Τέλος, θα ήθελα να τον ευχαριστήσω και για τις αρκετές, θα έλεγα, ευκαιρίες που μου έδωσε να αποκτήσω εμπειρία δύο μόνο σε θέματα υλοποίησης, αλλά και σε θέματα δημοσίευσης της δουλειάς που έχει γίνει.

Στην συνέχεια, θα ήθελα να ευχαριστήσω τον επόπτη αυτής της εργασίας κ. Γεώργιο Τζιρίτα για την αμέριστη συμπαραγάνταση και ενθάρρυνση που επέδειξε αυτά τα δύο χρόνια, όσο και για την πάσης φύσεως βοήθειας που μου προσέφερε. Γνωρίζοντάς τον τόσο ως Καθηγητή όσο και ως Επόπτη, μπορώ να πω ότι λυπάμαι που τελικά δεν είχα την τύχη να συνεργαστώ μαζί του πιο στενά σε κάποιο συγκεκριμένο θέμα.

Επίσης, θα ήθελα να ευχαριστήσω και τα δύο άλλα μέλη της επιτροπής εξέτασης, τους κ.κ. Σ. Ορφανουδάκη και Π. Τσακαλίδη για τον χρόνο που αφιέρωσαν και για τις παρατηρήσεις τους.

Η εργασία αυτή πραγματοποιήθηκε σε ένα άριστο περιβάλλον εργασίας, απαρτιζόμενο από μία σειρά φίλων και συνεργατών. Θα ήθελα λοιπόν να μνημονεύσω και να ευχαριστήσω για την καλή παρέα (και υπομονή) τους, τους Ηλία Γκρίνια, Γιώργο Σημαντήρη, Κώστα Παναγιωτάκη, Παναγιώτη Κουτσουράκη, Γιάννη Μαυρικάκη, Γιώργο Παγώνη και Νίκο Κομοντάκη. Εύχομαι καλή συνέχεια σε όσους έχουν υπό εξέλιξη κάποια εργασία και καλή αρχή σε αυτούς που αρχίζουν μία νέα.

Τέλος, δεν μπορώ να παραλείψω να ευχαριστήσω τους γονείς μου για την υπομονή και ενθάρρυνσή τους, τόσο κατά την διάρκεια των σπουδών μου, όσο και κατά την κρίσιμη περίοδο πριν από αυτές. Επίσης, θα ήθελα ιδιαίτερα να ευχαριστήσω τα αδέρφια μου,

Μαρία και Κώστα, για την εξαιρετικά πολύτιμη βοήθειά τους σε στιγμές που πριν από αρκετά χρόνια άρχισα να αναζητώ αρχές και ιδανικά.

Μανόλης Δελάκης

# Περιεχόμενα

Περιληψη . . . . .	I
Abstract . . . . .	III
Ευχαριστίες . . . . .	V
<b>1 Εισαγωγή</b>	<b>1</b>
1.1 Το Πρόβλημα της Ανίχνευσης Προσώπων . . . . .	2
1.2 Φορμαλισμός . . . . .	5
1.3 Στόχοι . . . . .	6
1.4 Περίγραμμα της Εργασίας . . . . .	7
<b>2 Σχετικές Εργασίες</b>	<b>9</b>
2.1 Προσεγγίσεις Βασισμένες σε Χαρακτηριστικά . . . . .	9
2.2 Προσεγγίσεις Βασισμένες στην Εμφάνιση . . . . .	11
2.2.1 Μέθοδοι Προβολής . . . . .	13
2.2.2 Στατιστικές Προσεγγίσεις . . . . .	17
2.2.3 Νευρωνικά Δίκτυα . . . . .	18
2.3 Συζήτηση . . . . .	20
<b>3 Συνελικτικά Νευρωνικά Δίκτυα</b>	<b>23</b>
3.1 Αρχές Αρχιτεκτονικής . . . . .	23
3.2 Ένα Παραδειγμα Συνελικτικής Τοπολογίας . . . . .	25
3.3 Μία Πρώτη Εφαρμογή στην Ανίχνευση Προσώπων . . . . .	28
3.4 Συζήτηση . . . . .	29
<b>4 Προτεινόμενη Τοπολογία και Μεθοδολογία Εκπαίδευσης</b>	<b>31</b>
4.1 Τοπολογία του Δικτύου . . . . .	31

4.2 Συλλογή και Προετοιμασία των Δεδομένων . . . . .	35
4.3 Εκπαίδευση του Δικτύου . . . . .	38
4.4 Αποτελέσματα της Εκπαίδευσης . . . . .	42
4.4.1 Εξέλιξη της Εκπαίδευσης . . . . .	42
4.4.2 Σύγκριση με Άλλες Τοπολογίες . . . . .	45
<b>5 Σάρωση της Εικόνας και Εντοπισμός του Προσώπου</b>	<b>47</b>
5.1 Συμπεριφορά του Δικτύου Γύρω από Ένα Στόχο . . . . .	47
5.2 Εντοπισμός του Προσώπου . . . . .	49
5.3 Επιτάχυνση της Σάρωσης . . . . .	52
5.4 Υπολογιστικές Απαιτήσεις . . . . .	55
5.4.1 Σύγκριση με Άλλες Μεθόδους . . . . .	57
<b>6 Πειραματικά Αποτελέσματα</b>	<b>59</b>
6.1 Αποτίμηση και Σύγκριση . . . . .	59
6.1.1 Περιγραφή των Συνόλων Δοκιμής . . . . .	59
6.1.2 Συνολική Απόδοση . . . . .	61
6.1.3 Σύγκριση στο Σύνολο CMU . . . . .	66
6.1.4 Σύγκριση στο Σύνολο DiVAN . . . . .	71
6.1.5 Σύγκριση Μεταξύ Διάφορων Στρατηγικών Αναζήτησης . . . . .	72
6.2 Ανάλυση Εναισθησίας . . . . .	74
<b>7 Συμπεράσματα</b>	<b>83</b>
7.1 Ανασκόπηση των Ευρημάτων . . . . .	83
7.2 Περιορισμοί . . . . .	85
7.3 Χώροι Εφαρμογής . . . . .	85
7.4 Σχετικές Δημοσιεύσεις . . . . .	86
7.5 Μελλοντικές Προεκτάσεις . . . . .	86
<b>Α Επιδείξεις</b>	<b>89</b>
<b>Β Ο Αλγόριθμος Ανάστροφης Διάδοσης του Σφάλματος</b>	<b>93</b>
<b>Βιβλιογραφία</b>	<b>97</b>

# Κατάλογος Σχημάτων

1.1	Ένα τυπικό αποτέλεσμα της διαδικασίας ανίχνευσης προσώπων . . . . .	3
1.2	Προκλήσεις στην ανίχνευση προσώπων . . . . .	4
2.1	Τα διάφορα βήματα πριν δώσουμε την είσοδο στον ταξινομητή . . . . .	12
2.2	Το σύστημα των Sung και Poggio . . . . .	14
2.3	Το σύστημα των Rowley <i>et al.</i> . . . . .	18
3.1	Η τοπολογία του LeNet-1 . . . . .	25
3.2	Η λειτουργία της συνέλιξης . . . . .	26
3.3	Η λειτουργία της υποδειγματοληψίας . . . . .	27
4.1	Η προτεινόμενη συνελικτική τοπολογία . . . . .	32
4.2	Κάποια από τα παραδείγματα προσώπων που συλλέχτηκαν . . . . .	35
4.3	Η διαδικασία εξαγωγής του προσώπου . . . . .	36
4.4	Κάποια από τα μετασχηματισμένα παραδείγματα . . . . .	37
4.5	Εικόνες φόντου και εσφαλμένες ειδοποιήσεις . . . . .	40
4.6	Η εξέλιξη της προτεινόμενης διαδικασίας εκπαίδευσης-bootstrapping . . . . .	43
4.7	Η εξέλιξη του διαχωρισμού των κλάσεων . . . . .	44
5.1	Απαντήσεις του δικτύου σε κλίμακα-θέση γύρω από έναν στόχο . . . . .	48
5.2	Τα βήματα της διαδικασίας εντοπισμού . . . . .	50
5.3	Κοινή περιοχή εισόδου ανάμεσα σε δύο διαδοχικές ενεργοποιήσεις του δικτύου	53
5.4	Οι παραγόμενες εικόνες της σωλήνωσης . . . . .	54
6.1	Η καμπύλη ROC για τα σύνολα δοκιμής CMU, WEB και Σινεμά . . . . .	62

6.2 Ποσοστά ανίχνευσης και αριθμός εσφαλμένων ειδοποιήσεων έναντι του <i>ThrVol</i>	62
6.3 Αριθμός σφαλμάτων έναντι <i>ThrVol</i>	64
6.4 Αποτελέσματα του συστήματος στο σύνολο CMU	67
6.5 Αποτελέσματα του συστήματος στο σύνολο DiVAN	68
6.6 Αποτελέσματα του συστήματος στο σύνολο WEB	69
6.7 Αποτελέσματα του συστήματος στο σύνολο Σινεμά	70
6.8 Οι εικόνες που χρησιμοποιήθηκαν στην ανάλυση ευαισθησίας	75
6.9 Ανάλυση ευαισθησίας στην περιστροφή της εισόδου	76
6.10 Ανάλυση ευαισθησίας στην θόλωση της εισόδου	77
6.11 Ανάλυση ευαισθησίας στην μεταβολή της φωτεινής αντίθεσης της εισόδου	78
6.12 Ανάλυση ευαισθησίας στην προσθήκη θορύβου της εισόδου	80
6.13 Απόδοση του συστήματος στην ακολουθία MPEG Foreman	81
A.1 Επιδείξεις του συστήματος	90
B.1 Μία τυπική τοπολογία πολυστρωματικού, μη αναδραστικού νευρωνικού δικτύου	94
B.2 Ένας τυπικός μη γραμμικός νευρώνας	94



# Κατάλογος Πινάκων

4.1	Πλάνο σύνδεσης μεταξύ των χαρτών χαρακτηριστικών των στρωμάτων C2 και S1	33
4.2	Το προτεινόμενο σχήμα bootstrapping . . . . .	39
4.3	Ρυθμός λαθών για διάφορες συνελικτικές τοπολογίες . . . . .	45
5.1	Αριθμός πράξεων ανά εικονοστοιχείο εικόνας για την πρόχειρη σάρωση . . . . .	55
5.2	Αριθμός πράξεων ανά εικονοστοιχείο εικόνας για την λεπτή σάρωση . . . . .	56
6.1	Κάποια στιγμιότυπα των καμπύλων ROC για τα σύνολα δοκιμής CMU, DiVAN, WEB και Σινεμά . . . . .	65
6.2	Σύγκριση στα σύνολα CMU, CMU-125, MIT and MIT-20 . . . . .	71
6.3	Αποτελέσματα στο σύνολο δοκιμής DiVAN . . . . .	71
6.4	Σύγκριση μεταξύ διάφορων εναλλακτικών επιλογών αναζήτησης . . . . .	72
A.1	Ρυθμοί επεξεργασίας σε περιβάλλοντα βίντεο . . . . .	91



## Εισαγωγή

Ένα από τα πιο ενδιαφέροντα χαρακτηριστικά του ανθρώπινου συστήματος όρασης είναι η ικανότητα να ανιχνεύει και να εντοπίζει με εξαιρετική ακρίβεια πρόσωπα μέσα στο οπικό πεδίο του ματιού. Μετά από αυτό το στάδιο, ο εγκέφαλος είναι ελεύθερος να εξακοιβώσει την ταυτότητα του ατόμου που κοιτάζουμε ή να αποκωδικοποιήσει τα συναισθήματα που μας αποστέλλονται μέσω της έκφρασης του προσώπου του. Η ανθρώπινη οπική επικοινωνία θα ήταν μια πολύ πιο δύσκολη διαδικασία χωρίς αυτήν την ικανότητα, αν και δεν δίνουμε μεγάλη προσοχή σε αυτό το σημείο αφού λειτουργεί στιγμαία και με χαρακτηριστική ευκολία. Ενώ αυτό είναι αληθές για να ανθρώπινα όντα, για την έρευνα της υπολογιστικής όρασης είναι ένα θέμα ακόμα υπό ανάπτυξη. Στα πρόσφατα χρόνια, καθώς περισσότερο απαιτητικές και πολύπλοκες εφαρμογές αναδύονται όπως η αναγνώριση προσώπων και η προχωρημένη αλληλεπίδραση ανθρώπου-μηχανής, η αυτόματη ανίχνευση προσώπων έχει προσελκύσει με τη σειρά της το ενδιαφέρον της επιστημονικής κοινότητας. Εκτός του ότι αποτελεί ένα ζωτικό εργαλείο για άλλες προχωρημένες εφαρμογές, προσφέρεται και σαν μια αρκετά ενδιαφέρουσα περίπτωση της έρευνας πάνω στην ανίχνευση αφηρημένων αντικειμένων σε μία εικόνα.

Από τις πρώτες εργασίες στην ανίχνευση προσώπων, η κοινή αίσθηση ήταν η απόπειρα της ρητής μοντελοποίησης του προσώπου χρησιμοποιώντας ανθρωπομετρικές τεχνικές. Το αντικείμενο αναζήτησης ήταν το καλύτερο δυνατό ταίριασμα στο μοντέλο του προσώπου, βασισμένο πάνω σε χαρακτηριστικά χαμηλού επιπέδου που εξαγόντουσαν από αλγορίθμους χαμηλού επιπέδου της υπολογιστικής όρασης [15, 56]. Η αποτελεσματικότητα αυτών των μεθόδων βρέθηκε να είναι ισχυρά εξαρτημένη από τις συνθήκες της εικόνας και έτσι περιορισμένη σε στενά και ελεγχόμενα περιβάλλοντα. Η ανάδειξη καινούργιων και περισσότερο ικανών αλγορίθμων αναγνώρισης προτύπων, όπως τα Τεχνητά Νευρωνικά Δίκτυα, επέτρεψε την δημιουργία νέων τεχνικών οι οποίες δεν στηρίζονται πάνω σε ρητά και προδηλωμένα

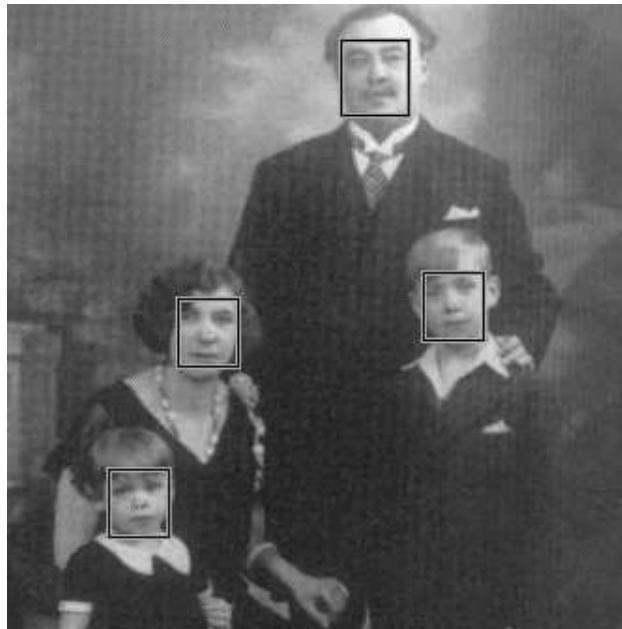
μοντέλα προσώπων. Κοινή ιδιότητα αυτών των βασισμένων σε μάθηση μεθόδων είναι η χρήση ενός μεγάλου συνόλου εκπαίδευσης με πρόσωπα, από το οποίο η μοντελοποίηση του προσώπου σχηματίζεται αυτόματα και χωρίς καμία ανθρώπινη μεσολάβηση. Μια από τις πρώτες αυτές τεχνικές ήταν το σύστημα των Sung και Poggio [48], όπου ένας αλγόριθμος αυτόματης οργάνωσης χρησιμοποιήθηκε για την κατασκευή εν δυνάμει μοντέλων προσώπων ('πρότυπα προσώπων'), ακολουθούμενος από έναν νευρωνικό ταξινομητή. Οι Rowley *et al.* [41] εισηγήθηκαν την πρώτη προχωρημένη προσέγγιση βασισμένη σε νευρωνικά δίκτυα. Χρησιμοποιήθηκε ένα νευρωνικό δίκτυο με συνδεσμολογία ειδικά προσαρμοσμένη στην είσοδο του, έτσι ώστε κάποια γνώση για την δομή του προσώπου ενσωματώθηκε απευθείας στη δομή του δικτύου. Άλλες σχετικές μέθοδοι περιλαμβάνουν τον Γραμμικό Διαχωριστή του Fisher [55], SVMs [33] και αναγνώριση της υφής του προσώπου με βάση Κυματιδιακή Ανάλυση [9, 45].

Οι Le Cun *et al.* [26] πρώτοι εισήγαγαν τα Συνελικτικά Νευρωνικά Δίκτυα στο πεδίο της αυτόματης αναγνώρισης χειρόγραφων χαρακτήρων. Βασισμένη σε μια ειδικά σχεδιασμένη τοπολογία, αυτή η κλάση νευρωνικών δικτύων περικλείει εκ κατασκευής τις έννοιες της εξαγωγής, σταθεροποίησης και συνδυασμού χαρακτηριστικών με ένα μοναδικό τρόπο. Παρ' όλη την μεγάλη τους επιτυχία στην αναγνώριση χαρακτήρων, καμία εκτεταμένη εργασία πάνω στην εφαρμογή τους στην ανίχνευση προσώπων δεν έχει αναφερθεί μέχρι σήμερα. Σκοπός της παρούσας εργασίας ήταν να επιβεβαιώσει τα Συνελικτικά Νευρωνικά Δίκτυα ως ένα αποτελεσματικό και ταχύ ανιχνευτή προσώπων και να μελετήσει την ικανότητά τους να λειτουργούν σε μη ελεγχόμενα περιβάλλοντα εικόνων.

## 1.1 Το Πρόβλημα της Ανίχνευσης Προσώπων

Το πρόβλημα της ανίχνευσης προσώπων μπορεί να διατυπωθεί επίσημα ως: "Δεδομένου μιας εικόνας βίντεο ή μεμονωμένης, ζητείται η ανίχνευση και ο ακριβής εντοπισμός της θέσης ενός αγνώστου εκ των προτέρων αριθμού προσώπων που πιθανόν να υπάρχουν στην εικόνα" [15]. Κάποιες μέθοδοι προϋποθέτουν ότι υπάρχει μόνο ένα πρόσωπο στην εικόνα ή ότι αυτό δεν είναι μερικά επικαλυπτόμενο από άλλα αντικείμενα όπως γυαλιά, ή ότι ακόμα και μια αρχική εκτίμηση της θέσης του είναι γνωστή. Όταν τα περιεχόμενα της εικόνας είναι άγνωστα και καμία εξωτερική πληροφορία δε δίνεται (όσον αφορά π.χ. σημασιολογική πληροφορία σχετική με την εικονιζόμενη σκηνή), μιλάμε για ανίχνευση προσώπων σε μη ελεγχόμενα περιβάλλοντα. Γενικά, καμία υπόθεση πάνω στην εμφάνιση και την κατάσταση των προσώπων στην εικόνα δεν πρέπει να γίνεται, εκτός του ότι τα πρόσωπα θα πρέπει να είναι ορατά και εντοπίσιμα από τον ανθρώπινο παρατηρητή.

Στην πράξη είναι πολύ δύσκολο για ένα μοναδικό ανιχνευτή να χειριστεί με επιτυχία



**Σχήμα 1.1**

Ένα τυπικό αποτέλεσμα της διαδικασίας ανίχνευσης προσώπων.

όλες τις περιπτώσεις που μπορεί να βρεθεί ένα πρόσωπο σε μία εικόνα και, έτσι, κάποιοι δεδομένοι περιορισμοί τίθενται. Στο προτεινόμενο σύστημα, η περίπτωση των ολικά προφίλ προσώπων καθώς και αυτών που είναι περιστραμμένα σε έναν οποιοδήποτε βαθμό γύρω από τον οπτικό άξονα της κάμερας δεν λαμβάνονται υπ' όψη. Αυτές οι περιπτώσεις μπορούν να θεωρηθούν ως πρόβλημα ανίχνευσης ενός διαφορετικού αντικειμένου. Μια σημαντική διαφοροποίηση μεταξύ των πολυάριθμων προσεγγίσεων της ανίχνευσης προσώπων είναι η χρησιμοποίηση ή όχι της χρωματικής πληροφορίας που μπορεί να εμπεριέχεται στην εικόνα. Το προτεινόμενο σύστημα λειτουργεί σε εικόνες διαβαθμίσεων του γκρι, ολοκληρωτικά αγνοώντας οποιαδήποτε πληροφορία χρώματος. Πρακτικά, το χρώμα θα μπορούσε να χρησιμοποιηθεί για εξοικονόμηση υπολογισμών σε ένα προ-φίλτροισμα της εικόνας που απορρίπτει περιοχές που έχουν χρώμα απόμακρο από αυτό της ανθρώπινης επιδερμίδας. Άλλα, καθώς η πιθανότητα απόρριψης περιοχών που περιέχουν πρόσωπα (π.χ. όταν αυτά φωτίζονται με έναν όχι συνηθισμένο τρόπο) είναι πάντα μεγαλύτερη του μηδενός, αυτή η προσέγγιση δεν ακολουθήθηκε.

Μια τυπική έξοδος της διαδικασίας ανίχνευσης προσώπων δίνεται στο σχήμα 1.1, όπου και τα τέσσερα πρόσωπα έχουν ανιχνευτεί και εντοπιστεί με ακρίβεια ενώ καμία άλλη περιοχή στην εικόνα δεν έχει χαρακτηριστεί ως ‘πρόσωπο’ (εσφαλμένη ειδοποίηση - *false alarm*). Τα εντοπισμένα πρόσωπα αυτής της εικόνας είναι σχετικά τετραμμένες περιπτώσεις



### Σχήμα 1.2

Προκλήσεις στην ανίχνευση προσώπων.

με καθαρά μετωπική θέα, ομοιόμορφη φωτεινότητα, χωρίς την επιρροή θιρύβου κτλ. Η πρόκληση στο παρόν πεδίο έρευνας είναι η ικανότητα χειρισμού περιπτώσεων όπως αυτών του σχήματος 1.2, όπου συναντάμε πρόσωπα με ποικίλες εκφράσεις προσώπου, μπορεί να είναι μερικά επικαλυπτόμενα ή ο φωτισμός μπορεί να μην είναι ομοιόμορφα διάχυτος στην περιοχή του προσώπου. Όλες αυτές οι περιπτώσεις θα μπορούσαν να θεωρηθούν ως παραμορφώσεις του σήματος εισόδου, τις οποίες το σύστημα ανίχνευσης θα πρέπει να είναι αρκετά ικανό να χειριστεί με επιτυχία. Υπάρχουν πολυάριθμες τέτοιες παραμορφώσεις που μπορούν να απαντηθούν σε μη ελεγχόμενα περιβάλλοντα έτσι ώστε να είναι πρακτικά ανέφικτο να συγκεντρωθούν όλες σε ένα σύνολο εκπαίδευσης πεπερασμένου μεγέθους (ή ακόμα πιο δύσκολα, σε ένα ή περισσότερα μοντέλα προσώπου σχεδιασμένα με το χέρι). Με αυτήν την παρατήρηση αναφερόμαστε στην ικανότητα γενίκευσης του συστήματος ανίχνευσης, δηλαδή στην ικανότητα του να δρα άθικτο σε νέες καταστάσεις, μη συμπεριλαμβανόμενες όπως είναι στο σύνολο εκπαίδευσης (ή στα μοντέλα του προσώπου). Μία άλλη πρόκληση σε αυτό το πεδίο είναι η ικανότητα απόρριψης περιοχών μη προσώπων του ταξινομητή όταν το φόντο της εικόνας είναι αρκετά διαταραγμένο και όχι τόσο μονότονο όπως αυτό του σχήματος 1.1. Είναι προφανές ότι δεν μπορούμε να εμπιστευτούμε έναν ταξινομητή ο οποίος θα μας δώσει αρκετές εσφαλμένες ειδοποιήσεις έτσι ώστε να ανιχνεύσει τα ένα ή δύο πρόσωπα που μπορεί να υπάρχουν στην εικόνα. Γενικεύοντας, πάντα υπάρχει ένας συμψηφισμός (trade) μεταξύ του ρυθμού των εσφαλμένων ειδοποιήσεων και του ποσοστού αληθών ανιχνεύσεων, στην οποία ένας ικανοποιητικός συμβιβασμός απαιτείται. Μέθοδοι που αποδίδουν το υψηλότερο δυνατό ποσοστό ανίχνευσης ενώ διατηρούν το ρυθμό εσφαλμένων ειδοποιήσεων σε ένα λογικό επίπεδο είναι περισσότερο προτιμητέες.

Υπάρχουν πολλά προβλήματα σχετικά με την ανίχνευση προσώπων, όπως η παρακολούθηση τους, η εκτίμηση πόζας και η μοντελοποίησή της ή η ανάλυση της έκφρασης του προσώπου. Η παρακολούθηση προσώπων δεν είναι παρά η φυσική προέκταση της ανίχνευσης προσώπων σε περιβάλλοντα βίντεο. Μία πρώτη λύση θα μπορούσε να δοθεί ως ένα σύστημα ανίχνευσης εφαρμογόμενο σε κάθε καρέ του βίντεο ξεχωριστά. Συνήθως, οι μέθοδοι παρακολούθησης στηρίζονται στις εκτιμήσεις που έχουν βρεθεί από τα προηγούμενα καρέ και από την κατάσταση του τρέχοντος για να δώσουν την νέα τους εκτίμηση. Η ανίχνευση προσώπων μπορεί να χρησιμοποιηθεί για την αυτόματη αρχικοποίηση της παραπάνω διαδικασίας και για την περιοδική επιβεβαίωση των αποτελεσμάτων παρακολούθησης καθώς η ακολουθία εξελίσσεται. Η μοντελοποίηση του προσώπου και η ανάλυση της έκφρασης [34] λειτουργούν απευθείας πάνω στην περιοχή του προσώπου και σκοπός τους είναι η εξαγωγή πληροφορίας υψηλού επιπέδου (όπως αν ο εικονιζόμενος γελάει, είναι θυμωμένος κτλ.). Η ανίχνευση προσώπων μπορεί να εξυπηρετήσει στην αυτόματη προετοιμασία των δεδομένων που θα χρησιμοποιηθούν για τέτοιους είδους ανάλυση. Μία άλλη υψηλού επιπέδου πληροφορία που εμπεριέχεται στην περιοχή του προσώπου είναι η ταυτότητα του εικονιζόμενου, η οποία είναι το αντικείμενο της αναγνώρισης προσώπων. Όταν ένα σύστημα αυτόματης αναγνώρισης προσώπων λειτουργεί σε εικόνες με τα πρόσωπα να βρίσκονται σε άγνωστες θέσεις και κλίμακες, η ανίχνευσή τους είναι, προφανώς, ένα απαραίτητο βήμα προ-επεξεργασίας. Εκτός αυτών των σχετικών με πρόσωπα προβλημάτων, η ανίχνευση προσώπων είναι απλώς μια ειδική περίπτωση της πιο γενικής ανίχνευσης αφηρημένων αντικειμένων σταθερού σώματος. Θεωρητικά όπως και πρακτικά ευρήματα σε αυτό το πεδίο μπορούν να μεταφερθούν αρκετά εύκολα σε άλλα πεδία όπως αυτά της ανίχνευσης του ανθρώπινου σώματος, της παρακολούθησης χεριών κτλ.

## 1.2 Φορμαλισμός

Για την επίλυση το προβλήματος της ανίχνευσης προσώπων προτείνεται μία τοπολογία Συνελικτικών Νευρωνικών Δικτύων. Δανειζόμενοι ιδέες αρχιτεκτονικής από την δουλειά των Le Cun *et al.* [26] στην αναγνώριση χαρακτήρων, η προτεινόμενη τοπολογία είναι η απλούστερη δυνατή, αν και αρκετά αποτελεσματική, εφαρμόσιμη για ανίχνευση προσώπων. Είναι ικανή να λειτουργεί κατευθείαν πάνω στα εικονοστοιχεία, μη κάνοντας καμία υπόθεση πάνω στις συνθήκες της εικόνας. Ένα νέο, μεγάλο και αποτελεσματικό σύνολο εκπαίδευσης με πρόσωπα κατασκευάστηκε, καθώς καμία από τις υπάρχουσες προσβάσιμες συλλογές με πρόσωπα δεν είναι ικανοποιητικά πλούσια για να αντιπροσωπεύσει τα φυσικά, μη κανονικοποιημένα δεδομένα στα οποία η προτεινόμενη τοπολογία αναμενόταν να λειτουργήσει. Επιπρόσθετα, αυτό το σύνολο εμπλουτίστηκε με νέα παραδείγματα, τεχνητά δημιουργημένα

από τα αυθεντικά, για να ενισχυθεί η απόδοση του συστήματος σε μία σειρά από δυνατές παραμορφώσεις της εισόδου. Ο πλούτος του συνόλου εκπαίδευσης και οι εκ κατασκευής ιδιότητες σθεναρότητας της προτεινόμενης τοπολογίας απέδωσαν ένα σύστημα ανίχνευσης προσώπων που είναι εντελώς ελεύθερο από οποιαδήποτε προεπεξεργασία της εισόδου. Η τεχνική της αυτοδύναμης μάθησης (bootstrapping) που εισάχθηκε από τους Sung *et al.* [48] βελτιώθηκε ώστε να εκπαιδευτεί το δίκτυο με έναν πιο αποτελεσματικό τρόπο πάνω στην ικανότητα απόριψης εσφαλμένων ειδοποιήσεων.

Χάρη στην ειδική συνελικτική φύση του δικτύου, επινοήθηκε ένας τρόπος για την σάρωση ολόκληρης της εικόνας σε ένα βήμα, επιτρέποντας μία πολύ σημαντική επιτάχυνση των υπολογισμών που απαιτούνται. Με αυτόν τον τρόπο, η συνολική λειτουργία που επιτελείται πάνω στην εικόνα είναι το πέρασμα από μία σειρά από γραμμικά και μη-γραμμικά φίλτρα. Βλέποντάς την σαν εξωτερικοί παρατηρητές θα διαπιστώσουμε ότι πρόκειται για μία λειτουργία αρκετά απλή στην υλοποίησή της, αλλά και εξαιρετικά ισχυρή στης απόδοσή της. Έχοντας όλες τις αποκρίσεις του δικτύου μετά από την διαδικασία φιλτραρίσματος, μία νέα μέθοδος προτείνεται για την ομαδοποίηση των αποκρίσεων και την λεπτή σάρωση γύρω από έναν στόχο που βελτιώνει την ικανότητα του συστήματος να διαχωρίζει μεταξύ προσώπων και μη.

Η σύγκριση με άλλες μεθόδους διεκπαιραιώθηκε πάνω σε δύο γνωστά σύνολα συγκρίσεων. Η απόδοση του παρόντος συστήματος αποτιμήθηκε επιπλέον σε δύο νέα σύνολα δοκιμών με διαφορετικές στατιστικές ιδιότητες για τον έλεγχο της καθολικότητάς του. Παρουσιάζονται λεπτομερή αποτελέσματα, συμπεριλαμβανομένου και των καμπύλων ROC όπου επιδεικνύεται ποσοτικά και με ακρίβεια ο συμβιβασμός μεταξύ ποσοστών επιτυχών ανιχνεύσεων και ρυθμού εσφαλμένων ειδοποιήσεων. Επιπλέον, το σύστημα δοκιμάστηκε σε μία σειρά από παραμορφώσεις της εισόδου για να διερευνηθούν οι ιδιότητες σθεναρότητάς του και πώς αυτό επηρεάζεται από το γεγονός ότι λειτουργεί χωρίς προεπεξεργασία. Αυτές οι παραμορφώσεις συμπεριλαμβάνουν μεταβλητές συνθήκες φωτεινής αντίθεσης (κοντράστ), εξομάλυνσης ακμών (smoothing), προσθήκης θιορύβου, περιστροφής και αλλαγής πόζας.

### 1.3 Στόχοι

Δεδομένου της πρακτικής αλλά και της θεωρητικής αξίας του πεδίου της ανίχνευσης προσώπων, οι στόχοι της παρούσας εργασίας είναι :

- Η δημιουργία ενός μεγάλου και αποτελεσματικού συνόλου εκπαίδευσης με παραδείγματα προσώπων.
- Η βελτίωση της διαδικασίας εκπαίδευσης που συχνά χρησιμοποιείται για την ανίχνευση προσώπων με αλγορίθμους μάθησης μέσω παραδειγμάτων.

- Ένα σύστημα ανίχνευσης προσώπων απαλλαγμένο πλήρως από την ανάγκη προφίλτραρισμάτος και προεπεξεργασίας της εισόδου, επιτρέποντάς το να λειτουργεί απευθείας στα εικονοστοιχεία της εικόνας.
- Ένας μεγάλος βαθμός σθεναρότητας σε μεταβλητή πόζα, έκφραση προσώπου, περιστροφή, φωτισμό και άλλες δυνατές παραμορφώσεις που μπορεί να απαντηθούν. Το σύστημα πρέπει να είναι αρκετά ικανό να χειρίζεται περιπτώσεις περιστροφής στο διάστημα [-20, 20] μοίρες, καθώς και ημι-προφίλ περιπτώσεις.
- Η απόδοση του συστήματος θα πρέπει να είναι ανεξάρτητη από τις συνθήκες του περιβάλλοντος της εικόνας και από την πιθανή ύπαρξη διαταραγμένου φόντου.
- Η ανάπτυξη μίας απλής και γρήγορης διαδικασίας για την ανίχνευση και τον εντοπισμό των προσώπων, χωρίς ευρητικές και προϋποθέσεις που πιθανόν να βλάψουν την γενικότητά της.
- Θεωρητικές διαπιστώσεις, χρήσιμες για την γενικευμένη ανίχνευση αντικειμένου σταθερού σώματος.

## 1.4 Περίγραμμα της Εργασίας

Στο κεφάλαιο 2 δίνεται μία επισκόπηση της μέχρι τώρα βιβλιογραφίας της ανίχνευσης προσώπων. Θα εξετάσουμε τις τρέχουσες τάσεις στην ανίχνευση προσώπων και θα δούμε κάποια συχνά εμφανιζόμενα προβλήματα. Το κεφάλαιο 3 παρέχει μία περιγραφή των Συνελικτικών Δικτύων όπως αυτά εισήχθησαν από τους Le Cun *et al.* [26] και με έμφαση στην κατανόηση των εννοιών και των αρχών τους. Το κεφάλαιο 4 περιγράφει με λεπτομέρεια την σχεδίαση της προτεινόμενης τοπολογίας και την μεθοδολογία της εκπαίδευσής της. Η διαδικασία σάρωσης της εικόνας εισόδου και ο εντοπισμός των προσώπων περιγράφεται στο κεφάλαιο 5. Η απόδοση του προτεινόμενου συστήματος εξετάζεται σε βάθος στο κεφάλαιο 6, συνοδευόμενη με συγκρίσεις με άλλες προσεγγίσεις. Στο ίδιο κεφάλαιο μελετάται και η σθεναρότητα του συστήματος σε σχέση με κάποιες πτυχές της μεταβλητότητας του προσώπου. Στο κεφάλαιο 7 κλείνουμε την εργασία αυτή με μία ανακεφαλαίωση των πιο σημαντικών της πορισμάτων, τους χώρους πρακτικής εφαρμογής και με κάποιες προτάσεις για μελλοντική έρευνα. Στο παράρτημα A παρουσιάζονται συνοπτικά δύο δημόσιες επιδείξεις της μεθόδου. Τέλος, στο παράρτημα B περιφράφεται συνοπτικά ο αλγόριθμος ανάστροφης διάδοσης του σφάλματος (backpropagation).



## ΚΕΦΑΛΑΙΟ 2

# Σχετικές Εργασίες

Σε αυτό το κεφάλαιο δίνεται μια έρευνα της σχετικής βιβλιογραφίας ανίχνευσης προσώπων. Οι προσεγγίσεις που έχουν απαντηθεί ως σήμερα μπορούν να διαιρεθούν σε δύο κατηγορίες: σε αυτές που στηρίζονται σε χαρακτηριστικά (feature-based) και σε αυτές που βασίζονται στην εμφάνιση (appearance-based). Στην πρώτη κατηγορία, εκμεταλλευόμαστε τη γνώση που διαθέτουμε για τη γεωμετρία του προσώπου για να φέρουμε σε πέρας την ανίχνευση. Τυπικά, τοπικά χαρακτηριστικά (όπως μάτια, μύτη κτλ) ανιχνεύονται και στην συνέχεια επεξεργάζονται με βάση γνωστολογική (knowledge-based) ανάλυση. Είναι αξιοπρόσεκτο ότι η συντριπτική πλειοψηφία της σχετικής βιβλιογραφίας ανήκει σε αυτήν την κατηγορία. Όσον αφορά την δεύτερη κατηγορία, στην οποία ανήκει και η παρούσα εργασία, μία περιοχή της εικόνας δίνεται κατευθείαν σε έναν αλγόριθμο που στηρίζεται σε μάθηση για να χαρακτηριστεί ως πρόσωπο ή όχι, αντιμετωπίζοντας την ανίχνευση προσώπων ως ένα πρόβλημα αναγνώρισης προτύπων. Για μία πιο λεπτομερή έρευνα της σχετικής βιβλιογραφίας από αυτήν που παρουσιάζεται εδώ και ιδιαίτερα για τις προσεγγίσεις βασισμένες σε χαρακτηριστικά, προτείνονται δύο πρόσφατα δημοσιευμένες επισκοπήσεις [15, 56] της ανίχνευσης προσώπων.

### 2.1 Προσεγγίσεις Βασισμένες σε Χαρακτηριστικά

Ένα από τα πρώτα χαρακτηριστικά που μπορεί να χρησιμοποιηθεί για ανίχνευση προσώπων είναι η πληροφορία από τις ακμές της εικόνας. Μάλιστα, ήταν το χαρακτηριστικό που χρησιμοποιήθηκε σε μία από τις πολύ πρώιμες προσεγγίσεις της βιβλιογραφίας από τους Sakai *et al.* [44]. Ένα τυπικό σενάριο μιας βασισμένης σε ακμές προσέγγισης μπορεί να βρεθεί στην εργασία του Govindajaru [12]. Πρώτα, οι ακμές εντοπίζονται και λειαίνονται από κάποια επιτηδευμένη τεχνική εύρεσης ακμών (ο Marr-Hildreth τελεστής ακμών στην περίπτωση [12]). Ακολουθούν φίλτραρισμα των ακμών και αφαίρεση θορύβου, σύμφωνα

με κάποιες μετρημένες ιδιότητες των (αληθινών) ακμών του προσώπου. Το τελευταίο βήμα είναι ο χαρακτηρισμός των εναπομεινάντων στοιχείων και η ελαχιστοποίηση τους κόστους μιας συνθήκης η οποία βασίζεται στη χωρική διάταξη των ακμών της περιοχής του προσώπου. Αυτό το τελευταίο βήμα μπορεί να θεωρηθεί και σαν συσχέτιση των εξαγόμενων χαρακτηριστικών με τη φόρμα (template) του προσώπου. Άλλες μέθοδοι που βασίζονται στις ακμές περιλαμβάνουν τις [14, 52].

Μπορεί επίσης να χρησιμοποιηθεί και η πληροφορία που περικλείεται στην φωτεινότητα των διαβαθμίσεων του γκρι της εικόνας, καθώς κάποια ελάχιστα (σκοτεινές κοιλάδες) εμφανίζονται συνήθως στις περιοχές των ματιών και του στόματος. Στην περίπτωση των εικόνων π.χ. ταυτότητας (μικρές εικόνες με απλό φόντο), η προβολή (άθροιση) όλων των εικονοστοιχείων πάνω στους δύο άξονες μπορεί να χρησιμεύσει στην εξαγωγή του συνόρου του προσώπου [19]. Αναμένεται ότι η προβολή στον κάθετο άξονα θα αποκαλύψει δύο τοπικά ελάχιστα που αντιστοιχούν στις περιοχές του στόματος και των ματιών. Η αναζήτηση επιπλέον τέτοιων χαρακτηριστικών κάτω από την ίδια φιλοσοφία (κάθετες και οριζόντιες προβολές επιλεγμένων περιοχών) μπορεί να οδηγήσει σε μία διαδικασία εντοπισμού του προσώπου. Είναι όμως προφανές ότι η διαδικασία αυτή δεν αναμένεται να δώσει καλά αποτελέσματα σε διαταραγμένα φόντα ή όταν ο φωτισμός δεν είναι τετριμμένος. Στην εργασία [53], προτείνεται μία τεχνική βασισμένη σε πολλαπλές αναλύσεις της εικόνας (multiresolution analysis). Στις χαμηλές αναλύσεις, τα παραπάνω τοπικά ελάχιστα αναμένεται να είναι πιο ορατά και εξάγονται με απλούς κανόνες που εκμεταλλεύονται την γεωμετρία του προσώπου. Χρησιμοποιώντας τις περιοχές που ικανοποιούν αυτούς τους κανόνες σαν αρχική εκτίμηση, η μέθοδος προχωράει σε πιο υψηλές αναλύσεις όπου άλλοι κανόνες χρησιμοποιούνται, κατάλληλοι για την τρέχουσα ανάλυση. Μία προέκταση αυτής της εργασίας παρουσιάζεται στην [20].

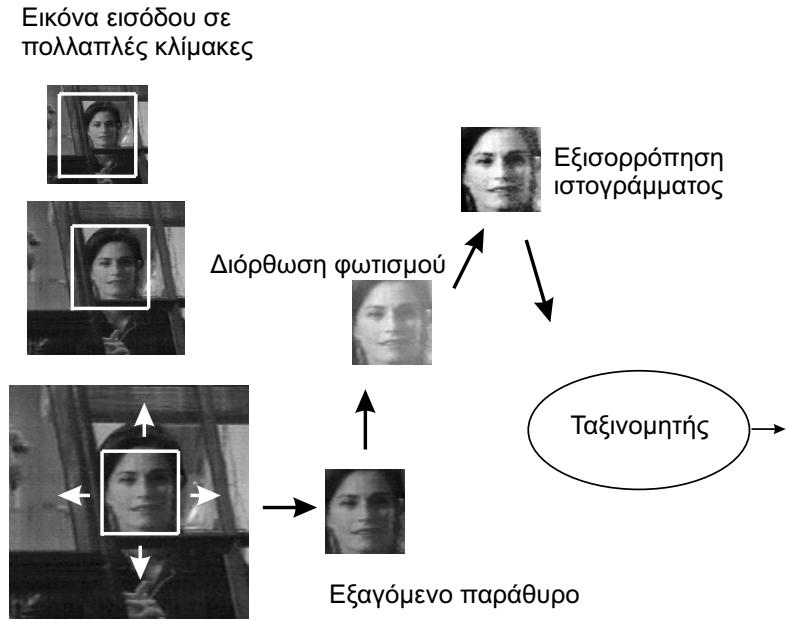
Η χρωματική πληροφορία της ανθρώπινης επιδερμίδας μπορεί να εξυπηρετήσει ως μία σημαντική νέα για την παρουσία ή όχι ενός προσώπου στην εικόνα. Βρέθηκε ότι η χρωματικότητα της ανθρώπινης επιδερμίδας μένει αναλλοίωτη μεταξύ των φυλών, σχηματίζοντας ένα στενό σύμπλεγμα στον χώρο χρώματος [18, 54]. Υπάρχουν πολυάριθμες προσεγγίσεις για το ποιος είναι ο πιο κατάλληλος χώρος χρώματος (ή ένας μετασχηματισμός αυτού). Μία συγκριτική μελέτη σε αυτό το ζήτημα παρουσιάζεται στην [49], όπου το τελικό συμπέρασμα ήταν ότι ο πιο κατάλληλος χώρος είναι ο κανονικοποιημένος TSL. Άλλα, καθώς οι χώροι YCbCr και HSV είναι ευρύτατα διαδεδομένοι σε πολλές πρακτικές εφαρμογές, οι περισσότερες εργασίες στηρίζονται σε αυτούς τους δύο χώρους. Ο διαχωρισμός μεταξύ χρώματος επιδερμίδας και μη μπορεί να επιτευχθεί χρησιμοποιώντας μία πολυδιάστατη γκαουσσιανή κατανομή [54] ή με ένα σύνολο από απλούς κανόνες που ενεργούν απευθείας πάνω στο έγχρωμα εικονοστοιχεία [9, 16]. Έχοντας μία λίστα από περιοχές της εικόνας με χρώμα

κοντά σε αυτό της επιδερμίδας, το επόμενο βήμα είναι η απόρριψη των εσφαλμένων ειδοποιήσεων και μία τελική απόφαση για την παρουσία προσώπου και την τοποθεσία του μέσα στην εικόνα. Στην εργασία [16], τα μάτια και το στόμα εντοπίζονται με βάση κατάλληλους μετασχηματισμούς και συνδυασμούς της ίδια χρωματικής πληροφορίας που χρησιμοποιήθηκε στο προηγούμενο βήμα. Αντί να παραμείνουμε σε βασισμένες σε χαρακτηριστικά προσεγγίσεις, σε αυτό το βήμα μπορεί να χρησιμοποιηθούν μέθοδοι εντελώς διαφορετικής φύσης, όπως στατιστική ανάλυση [9] ή νευρωνικές προσεγγίσεις [41]. Με αυτόν το τρόπο, η χρωματική πληροφορία χρησιμεύει σαν ένα βήμα προ-φιλτραρίσματος για την μείωση του χώρου αναζήτησης (και επίσης και του υπολογιστικού κόστους) και για την προ-απόρριψη πιθανών εσφαλμένων ειδοποιήσεων.

Άλλο ένα πρωτεύων χαρακτηριστικό που μπορεί να ληφθεί υπ' όψη είναι η πληροφορία κίνησης που εμπεριέχεται σε ακολουθίες βίντεο. Στην εργασία [30], η διαφορά μεταξύ των καρέ του βίντεο χρησιμοποιήθηκε για να εντοπιστούν χαρακτηριστικά προσώπου. Στην [28], χρησιμοποιήθηκαν κινούμενες ακμές ως πρωτεύοντα χαρακτηριστικά, παντρεύοντας το πρόβλημα της ανίχνευσης προσώπων με αυτό της εκτίμησης της οπτικής ροής. Γενικά, το προ-φιλτράρισμα κίνησης [3] μπορεί να χρησιμοποιηθεί με τον ίδιο τρόπο όπως το προ-φιλτράρισμα χρώματος που είδαμε παραπάνω. Επίσης υπάρχει η δυνατότητα να παραταχθούν σε σειρά πολλαπλοί ανιχνευτές χαρακτηριστικών (π.χ. [2, 3]), με σκοπό την επίτευξη πιο σθεναρών αποτελεσμάτων. Η χρήση πιο γενικευμένων χαρακτηριστικών όπως ο προσανατολισμός των ακμών [36] και ιδιότητες συμμετρίας [29] προτείνονται σε κάποιες άλλες προσεγγίσεις. Στα πρόσφατα χρόνια, κάποιες άλλες περισσότερο σθεναρές προσεγγίσεις βασισμένες σε χαρακτηριστικά αναπτύχθηκαν όπως ανάλυση αστερισμού (constellation) [58], τα ενεργά περιγράμματα (snakes) [57] ή οι μη στερεές φόρμες (deformable templates) [8, 59].

## 2.2 Προσεγγίσεις Βασισμένες στην Εμφάνιση

Σ' αυτήν τη δεύτερη κλάση προσεγγίσεων για την ανίχνευση προσώπων, ο διαχωρισμός μεταξύ προσώπων και μη δεν στηρίζεται σε ένα σύνολο από προκαθορισμένα χαρακτηριστικά, αλλά στην εικόνα (εμφάνιση) του προσώπου ως σύνολο (έτσι και το όνομα “βασισμένες στην εμφάνιση”). Τα εικονοστοιχεία της εικόνας δίνονται απευθείας σε μία εκπαιδευμένη μηχανή ταξινόμησης, χωρίς την παρεμβολή εξωτερικών παραγόντων. Για να ανιχνευτούν πρόσωπα σε διαφορετικές τοποθεσίες και μεγέθη (κλίμακες), πολλαπλές περιοχές (παραθυρα) εξάγονται από την εικόνα και δίνονται στην μηχανή ταξινόμησης. Αυτό το σενάριο, που το μοιράζονται σχεδόν όλες οι προσεγγίσεις αυτής της κατηγορίας, απεικονίζεται στο σχήμα 2.1. Το μέγεθος του παραθύρου αυτού διαφέρει από μέθοδο σε μέθοδο, αλλά μία ευρύτατα αποδεκτή επιλογή είναι περίπου  $20 \times 20$  εικονοστοιχεία. Είναι ένας καλός συμβι-



**Σχήμα 2.1**

Τα διάφορα βήματα πριν δώσουμε την είσοδο στον ταξινομητή. Ένα παράθυρο σαρώνει την εικόνα εισόδου σε διάφορες κλίμακες και τοποθεσίες. Σε όλες τις περιπτώσεις που εξάγουμε το παράθυρο, τα εικονοστοιχεία του υφίστανται διόρθωση φωτισμού και εξισορρόπηση ιστογράμματος.

βασιμός ανάμεσα σε ένα χώρο εισόδου με (σχετικά) χαμηλή διάσταση και της ικανοποιητικής διαύγειας του προσώπου μέσα στο παράθυρο [41, 48]. Τελικά, έχοντας συλλέξει όλες τις καταφατικές απαντήσεις του ταξινομητή γύρω από έναν στόχο, εφαρμόζεται ομαδοποίηση που θα δώσει και την τελική απάντηση για την παρουσία ή όχι ενός προσώπου καθώς επίσης και την ακριβή του έκταση μέσα στην εικόνα.

Στις περισσότερες από τις προσεγγίσεις που ακολουθούν, το παράθυρο εισόδου υφίσταται πρώτα προεπεξεργασία πριν δοθεί στον ταξινομητή. Αυτό το στάδιο περιλαμβάνει συνήθως διόρθωση φωτισμού μέσα στο παράθυρο εισόδου, ακολουθούμενη από εξισορρόπηση ιστογράμματος (αυτή η προσέγγιση εισάχθηκε πρώτα στην εργασία [48]). Σαν πρώτο βήμα, ανακτάται ένα (γραμμικό) επίπεδο που αρμόζει καλύτερα στην φαινόμενη φωτεινότητα και στην συνέχεια αφαιρείται από αυτή σε όλα τα σημεία. Με αυτό τον τρόπο, μεγάλες διαφορές στην ένταση της φωτεινότητας μέσα σε αυτό το παράθυρο απαλείφονται προσεγγιστικά. Μετά από αυτήν την χωρική κανονικοποίηση της φωτεινότητας, ακολουθεί η εξισορρόπηση του ιστογράμματος της για να βελτιώσει την ποιότητα της εισόδου. Γενικά, αυτό το βήμα κάνει τις ακμές και τις σκοτεινές περιοχές του προσώπου πιο ορατές (παρουσιάζοντας μεγαλύτερη αντίθεση με το φόντο), βοηθώντας έτσι τον ταξινομητή στο έργο του.

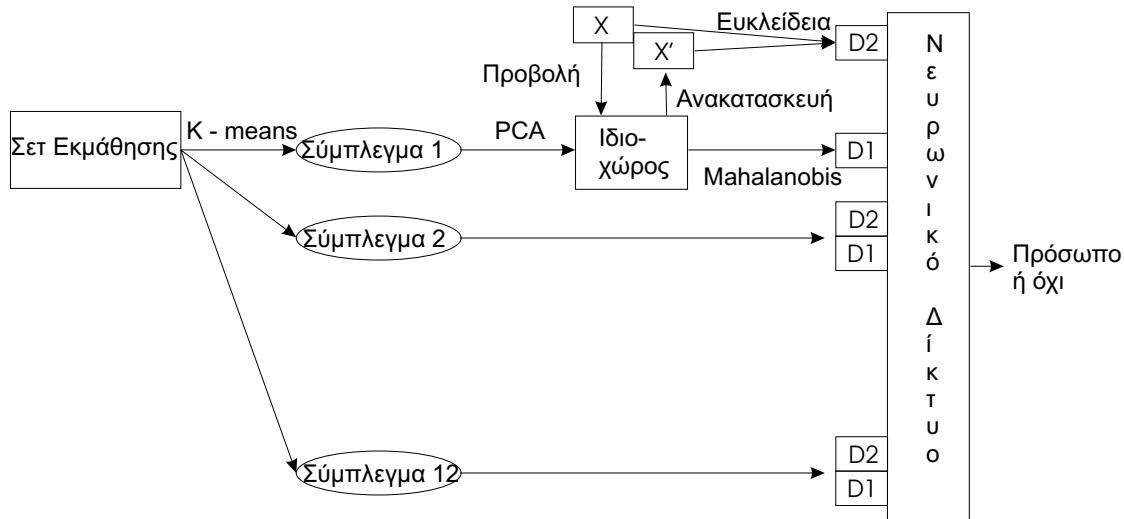
Άλλο ένα δυνατό βήμα προεπεξεργασίας της εισόδου είναι η εξομάλυνση ακμών (χρησιμοποιήθηκε π.χ. στην [3]), η οποία επιπλέον μπορεί να αφαιρέσει κάποιον πιθανό ενοχλητικό θόρυβο.

Οι πιο πολλές μέθοδοι αυτής της κατηγορίας παρουσιάζουν αποτελέσματα σε ένα κοινό σύνολο δοκιμών, το λεγόμενο σύνολο δοκιμής *CMU*, ή σε κάποιο υποσύνολό του. Αρχικά, οι Sung και Poggio [48] δρισαν το σύνολο *MIT* με 23 εικόνες και 155 μετρημένα πρόσωπα. Αυτό το σύνολο αλιμακώθηκε σε 130 εικόνες και 507 πρόσωπα από τους Rowley *et al.* [41] για πιο έμπιστες αποτυμήσεις. Οι εικόνες που προσαρτήθηκαν φέρουν κάποια πιο διαταραγμένα φόντα και κάποιες λιγότερο τετριμένες πόζες, υποθέτοντας ότι το σύστημα υπό εξέταση μπορεί να λειτουργήσει σε μη ελεγχόμενα περιβάλλοντα. Επίσημα αποτελέσματα που έχουν δοθεί σε αυτά τα σύνολα παρουσιάζονται στην ενότητα 6.1.3, σε αντιπαραβολή με αυτά της προτεινόμενης μεθόδου.

### 2.2.1 Μέθοδοι Προβολής

Αυτές οι προσεγγίσεις προβάλλουν το σήμα εισόδου σε ένα άλλον χώρο χαμηλότερης διάστασης (υποχώρος - subspace), όπου πιστεύεται ότι η ταξινόμηση μπορεί να διεξαχθεί πιο εύκολα. Ένας ευρύτατα διαδεδομένος μετασχηματισμός για την δημιουργία αποτελεσματικών και αντιποσωπευτικών υποχώρων είναι η Ανάλυση Κυρίων Συνιστώσων (Principal Component Analysis - PCA). Οι Turk και Pentland [50] εισήγαγαν τον PCA στο πεδίο της αναγνώρισης προσώπων. Δεδομένου ενός συνόλου από εικόνες με το πρόσωπο του ίδιου ατόμου, ανακτώνται οι κύριες συνιστώσες της κατανομής (τα περισσότερο ισχυρά ιδιοδιανύσματα ή “ιδιοπρόσωπα” - “eigenfaces”), σχηματίζοντας την βάση του νέου υποχώρου, του ιδιοχώρου ή “προσωποχώρου” (“facespace”). Όταν μία καινούργια εικόνα πρέπει να ταξινομηθεί, προβάλλεται στον προσωποχώρο μέσω των ιδιοπρόσωπων και στην συνέχεια ανακατασκευάζεται πίσω στον χώρο εισόδου. Το σφάλμα αυτής της ανακατασκευής μπορεί να χρησιμοποιηθεί σαν μία μέτρηση της εγγύτητας του εικονιζόμενου προσώπου με αυτό από του οποίου κατασκευάστηκε ο προσωποχώρος. Αυτή η τεχνική επεκτάθηκε στην ανίχνευση των χαρακτηριστικών του προσώπου και στην καθ' αυτή ανίχνευση προσώπων [35], χρησιμοποιώντας αυτά τα χαρακτηριστικά ως φόρμες συσχέτισης. Στην εργασία [31], εκτός από τις κύριες συνιστώσες της κατανομής του προσώπου, λήφθηκαν υπ' όψη και τα ορθογώνια συμπληρώματα για την αντιμετώπιση της μεταβολής του φωτισμού.

Μία από τις πρώτες προχωρημένες μεθόδους βασισμένες στην εμφάνιση που παρουσιάστηκαν στην βιβλιογραφία είναι αυτή των Sung και Poggio [46, 47, 48]. Οι συγγραφείς χρησιμοποίησαν πολλαπλούς υποχώρους για να αναπαραστήσουν το πολυσύνθετο της κατανομής του προσώπου και δύο μετρικές απόστασης από αυτούς τους υποχώρους για τις ανάγκες της ταξινόμησης. Το σύστημα τους αποτελείται από δύο συστατικά: ένα σύνολο

**Σχήμα 2.2**

Το σύστημα των Sung και Poggio [48].

από πολυδιάστατους γκαουσσιανούς πυρήνες, που αντιπροσωπεύουν τα “πρωτότυπα προσώπων” και “μη-προσώπων” και έναν νευρωνικό ταξινομητή (σχήμα 2.2). Ένα σύνολο εκπαίδευσης από 4.150 παραδείγματα προσώπων (τα 1.067 από αυτά αυθεντικά, ενώ τα υπόλοιπα κατασκευάστηκαν τεχνητά με μετασχηματισμούς των αυθεντικών) απέδωσε στην δημιουργία 6 συμπλεγμάτων (clusters) προσώπων, αναπαριστώμενα από 6 γκαουσσιανούς πυρήνες με μία κεντροειδή τοποθεσία και έναν πίνακα συνδιασπορών. Το μέγεθος της εικόνας εισόδου ορίστηκε να είναι  $19 \times 19$ , το οποίο μας δίνει και τις διαστάσεις του γκαουσσιανού πυρήνα. Η δομή του πίνακα συνδιασπορών περιορίστηκε σε διαγώνια, περιορίζοντας έτσι τους βαθμούς ελευθερίας καθώς ο αριθμός των παραδειγμάτων δεν θεωρήθηκε ικανοποιητικά μεγάλος. Για την δημιουργία των συμπλεγμάτων, μία τροποποιημένη έκδοση του αλγορίθμου *k-means* εφαρμόσθηκε, προσαρμοσμένη ειδικά στις ανάγκες της ελλειπτικής δομής των πυρήνων. Πιο συγκεριμένα, χρησιμοποιήθηκε η κανονικοποιημένη απόσταση Mahalanobis σαν μετρική αποστάσεων, η οποία διαφέρει από την κλασική απόσταση Mahalanobis στο ότι βαρύνει την απόσταση από ένα σύμπλεγμα ανεξάρτητα από την “έκτασή” του (νόρμα του πίνακα συνδιασπορών). Οι συγγραφείς κατέληξαν ότι ένας αριθμός 6 τέτοιων συμπλεγμάτων ήταν ικανοποιητικός λόγο του ανεπαρκή μεγέθους του συνόλου εκπαίδευσης, ενώ υπέθεσαν μία “ομαλή” συμπεριφορά των δεδομένων μεταξύ αυτών των συμπλεγμάτων. Για την αντιπροσώπευση των μη-προσώπων, ειδικά αυτών που βρίσκονται “κοντά” στα πρωτότυπα των προσώπων, άλλα 6 συμπλέγματα σχηματίστηκαν με τον ίδιο τρόπο, βασισμένα σε ένα σύνολο από εικόνες που μοιάζουν με πρόσωπα.

Έχοντας τα 12 συμπλέγματα με τα κεντροειδή τους και τους πίνακες συνδιασποράς τους,

ο PCA εφαρμόσθηκε σε καθένα από αυτά. Με αυτόν τον τρόπο, μόνο οι πιο σημαντικές συνιστώσες των πινάκων συνδιασπορών αριθμήθηκαν σαν ένα επιπλέον μέτρο για την αποφυγή υπέρ-προσαρμογής (overfitting) στα δεδομένα λόγω του μικρού μεγέθους του συνόλου εκπαίδευσης. Μετά την αφαίρεση όλων των λιγότερο σημαντικών ιδιοδιανυσμάτων, 75 από αυτά διατηρήθηκαν για κάθε σύμπλεγμα. Δεδομένης μίας εικόνας εισόδου να χαρακτηρισθεί πρόσωπο ή όχι, σαν πρώτο βήμα προβάλλεται σε αυτούς τους 12 υποχώρους. Μέσα σε αυτούς, υπολογίζεται η κανονικοποιημένη απόσταση Mahalanobis από το κέντρο του κάθε συμπλέγματος. Έτσι παίρνουμε τις 12 αποστάσεις D1 που βλέπουμε στο σχήμα 2.2. Κάθε τιμή D1 μπορεί να γίνει αντιληπτή ως η απόσταση από το κέντρο του συμπλέγματος και προς την κατεύθυνση των πιο σημαντικών συνιστωσών του συμπλέγματος, καθώς υπολογίζεται εντός του ιδιοχώρου. Οι υπόλοιπες τιμές D2 του σχήματος είναι κατά κάποιον τρόπο τα συμπληρώματα των τιμών D1 καθώς μας δίνουν το σφάλμα ανακατασκευής της εικόνας εισόδου, αφότου αυτή έχει προβληθεί στους ιδιοχώρους. Αυτό το σφάλμα αναμένεται να λάβει υπ' όψη αποστάσεις στις κατεύθυνσεις των λιγότερο σημαντικών συνιστωσών του κάθε συμπλέγματος. Η απλή ευκλείδεια νόρμα χρησιμοποιήθηκε για τον υπολογισμό αυτών των αποστάσεων. Το τελευταίο βήμα της διαδικασίας ταξινόμησης ήταν η διέγερση ενός νευρωνικού δικτύου με αυτές τις 12 τιμές D1 και D2. Ένα κλασικό πολυστρωματικό perceptron (Multilayer Perceptron) χρησιμοποιήθηκε σε αυτό το βήμα με 24 κρυμμένες μονάδες και άλλη μία μονάδα εξόδου που έδειχνε την ύπαρξη ή όχι προσώπου μέσα στην εικόνα εισόδου.

Το πρόβλημα της εύρεσης ενός ικανοποιητικού συνόλου από παραδείγματα μη πρόσωπων, δηλ. ενός συνόλου με αρκετά ‘διδακτικά’ και χρήσιμα παραδείγματα, επίσης αντιμετωπίστηκε. Σχεδιάστηκε μία τεχνική bootstrapping, με την οποία το σύστημα επανα-εκπαιδεύοταν επαναληπτικά με τις εσφαλμένες ειδοποιήσεις του προηγούμενου βήματος. Ξεκινώντας από μία αρχική υπόθεση, το σύστημα εκπαιδεύοταν και στην συνέχεια εφαρμοζόταν σε μία σειρά από εικόνες που δεν περιείχαν πρόσωπα. Όλες οι παραγόμενες εσφαλμένες ειδοποιήσεις συγκεντρώνονταν και προστίθεντο στο σύνολο εκπαίδευσης μη προσώπων για την επόμενη εκπαίδευση του συστήματος, στην επόμενη επανάληψη της διαδικασίας. Είναι αρκετά αξιοσημείωτο ότι η αυτή η ίδια διαδικασία bootstrapping υιοθετήθηκε σε σχεδόν όλες τις μεθόδους που ακολουθούν παρακάτω, ανεξάρτητα της συγκεκριμένης φύσης του κάθε συστήματος.

Η χρήση πολλαπλών υποχώρων για μία πιο αποδοτική αντιπροσώπευση της ολικής κατανομής του προσώπου υιοθετήθηκε επίσης και από τους Yang *et al.* [55]. Σε αυτήν την εργασία, δύο μέθοδοι προτάθηκαν και δοκιμάστηκαν: μία μέθοδος βασισμένη στην παραγωγική (generative) προσέγγιση και μία μέθοδος βασισμένη στην προσέγγιση διάκρισης (discriminative). Εξ ορισμού, αρνητικά παραδείγματα δεν χρησιμοποιούνται καθόλου με

την πρώτη προσέγγιση, αλλά αυτό που αναζητείται είναι η κατασκευή ενός καλού μοντέλου του προσώπου αυτού καθ' αυτού με τεχνικές όπως η Maximum Likelihood Estimation (MLE). Με την δεύτερη προσέγγιση, το όριο διαχωρισμού (decision boundary) μεταξύ των δύο κλάσεων, πρόσωπα και μη, αναζητείται άμεσα και μόνο για τους σκοπούς της ταξινόμησης. Μέθοδοι όπως τα νευρωνικά δίκτυα ανήκουν σε αυτήν την κατηγορία. Η παραγωγική μέθοδος που προτάθηκε είναι μία τεχνική παρόμοια με αυτή του PCA, η Factor Analysis, με την οποία τα δεδομένα επίσης προβάλλονται σε ένα χώρο μικρότερης διάστασης. Τα παραδείγματα δοκιμής επίσης προβάλλονται σε αυτόν τον χώρο και η πιθανότητα να ανήκουν στην κλάση των προσώπων υπολογίζοταν. Αν ήταν πάνω από κάποιο συγκεκριμένο κατώφλι, τότε χαρακτηρίζοταν ως πρόσωπο. Χρησιμοποιώντας μία μίξη από Factor Analyzers [4] εκπαιδευμένους μέσω του αλγορίθμου EM [11], τα δεδομένα προβάλλονται σε πολλαπλούς υποχώρους. Η διάσταση των εικόνων εισόδου ορίστηκε σε  $20 \times 20$  εικονοστοιχεία, η οποία οδηγούσε σε ένα κάπως μεγάλο χώρο όπου κάποιοι υπολογισμοί που απαιτούνται από τον αλγόριθμο EM (όπως αντιστροφή πίνακα) δεν μπορούσαν να εκτελεστούν αποδοτικά. Για την επίλυση αυτού του προβλήματος, εφαρμοζόταν αρχικά ο κανονικός PCA για να μειωθεί η διάσταση του χώρου εισόδου στις 80 διαστάσεις.

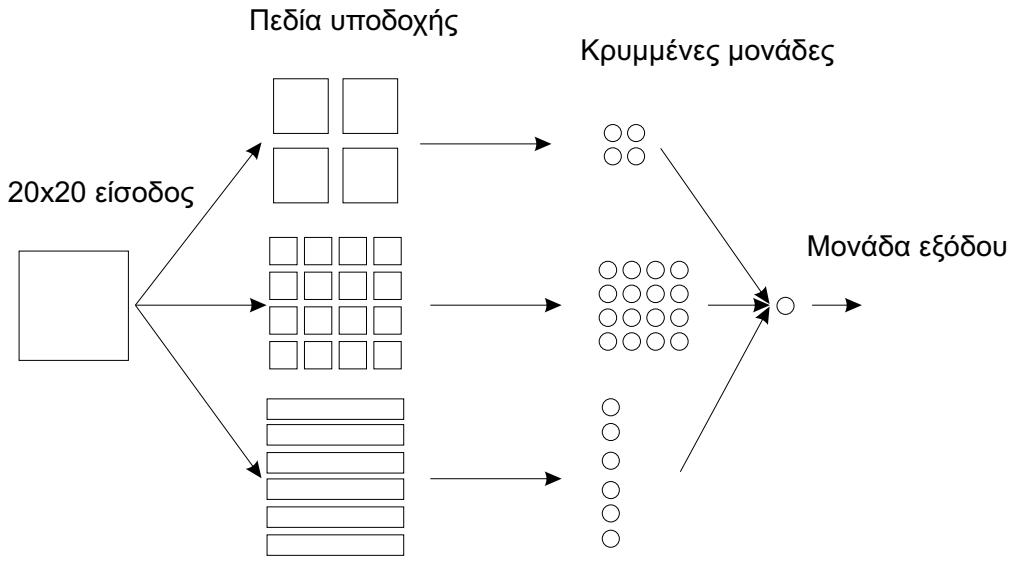
Η προτεινόμενη μέθοδος διάκρισης ήταν η Linear Discrimination Analysis (LDA), η οποία βασίζεται στον Γραμμικό Διαχωριστή του Fisher (Fisher Linear Discriminant). Το πλεονέκτημα του LDA σε σύγκριση με τον PCA είναι ότι αναζητεί μία προβολή του χώρου εισόδου σε ένα νέο, όπου ο γραμμικός διαχωρισμός μεταξύ δύο κλάσεων είναι ο βέλτιστος δυνατός. Ο PCA αναζητεί μία προβολή του χώρου εισόδου η οποία διατηρεί όσο το δυνατό την διασπορά του (variance) και γι' αυτό είναι κατάλληλος κυρίως σε προβλήματα συμπίεσης. Αρχικά, ο αλγόριθμος SOM (Self Organizing Map) εφαρμόστηκε σε όλα τα παραδείγματα προσώπων και μη, διαμοιράζοντας των χώρο εισόδου σε 25 κλάσεις προσώπων και 25 κλάσεις μη πρόσωπων. Στην συνέχεια, ο LDA εφαρμόστηκε σε αυτό το πρόβλημα των 50 κλάσεων, δίνοντας τον επιθυμητό 49-διάστατο χώρο όπου αυτές οι 50 κλάσεις είναι γραμμικά διαχωρίσιμες. Εκεί, γκαουσσιανές κατανομές χρησιμοποιήθηκαν για να μοντελοποιηθούν τα δεδομένα. Όταν ένα καινούργιο παραδειγμα δοκιμής έπρεπε να ταξινομηθεί, η πιθανότητα να ανήκει σε μία από τις παραπάνω κατανομές υπολογίζοταν και τελικά το παραδειγμα χαρακτηρίζοταν ανάλογα. Για την εκπαίδευση και των δύο παραπάνω μεθόδων, ένα σύνολο εκπαίδευσης με 1.681 αυθεντικά πρόσωπα συλλέχθηκε αρχικά και κλιμακώθηκε σε 16.810 παραδείγματα με τεχνητούς μετασχηματισμούς. Τα απαιτούμενα αρνητικά παραδείγματα για την εκπαίδευση της δεύτερης μεθόδου συλλέχθηκαν με την τεχνική bootstrapping [48] που αναφερθήκαμε. Τα πειράματα που εκτελέστηκαν έδειξαν ότι η μέθοδος βασισμένη στον LDA δίνει ελάχιστα καλύτερα αποτελέσματα από αυτήν που στηρίζεται στην Factor Analysis.

### 2.2.2 Στατιστικές Προσεγγίσεις

Οι Osuna *et al.* [32, 33] πρότειναν ένα σύστημα ανίχνευσης προσώπων βασισμένο σε ένα ταξινομητή SVM (Support Vector Machines - Μηχανές Διανυσμάτων Υποστήριξης). Οι SVMs είναι εκπαιδευόμενες μηχανές βασισμένες στην αρχή της *Ελαχιστοποίησης του Δομικού Ρίσκου* (Structural Risk Minimization), σύμφωνα με την οποία ένα βέλτιστο σφάλμα γενίκευσης αναζητείται, σε αντιπαραβολή με την βελτιστοποίηση του σφάλματος εκπαίδευσης που επιδιώκεται από τα MLPs. Η λειτουργία ενός SVM περιλαμβάνει μία μη γραμμική απεικόνιση από τον χώρο εισόδου σε ένα χώρο χαρακτηριστικών (feature space) μέσω κάποιων πυρήνων (kernel functions). Σε αυτόν τον χώρο, μπορεί να ευρεθεί ένα υπερ-επίπεδο που να διαχωρίζει γραμμικά και με έναν βέλτιστο τρόπο τις κλάσεις. Αυτή η λύση θα δώσει μία σειρά από τα παραδείγματα εισόδου τα οποία βρίσκονται κοντά στο υπερ-επίπεδο διαχωρισμού (Διανύσματα Υποστήριξης), τα οποία υπηρετούν ως κέντρα για τα kernel functions. Παρ' όλη την θεωρητική ελκυστικότητα, στην πράξη απαιτείται η επίλυση ενός προβλήματος βελτιστοποίησης τετραγωνικού βαθμού και μεγάλης διάστασης που είναι πρόκληση να λυθεί αποδοτικά. Οι Osuna *et al.* πρότειναν ένα σχήμα διάσπασης για την επίλυση του παραπάνω προβλήματος χρησιμοποιώντας έναν SVM με πολυώνυμα δευτέρου βαθμού ως kernel functions. Η εκπαίδευση από ένα μεγάλο σύνολο παραδειγμάτων έδωσε 2.500 διανύσματα υποστήριξης (η χρήση παραπάνω διανυσμάτων θα οδηγούσε τους πόρους της μηχανής σε εξάντληση). Τέλος, ως ένα σύστημα ανίχνευσης προσώπων, διαμοιράζεται αρκετά χαρακτηριστικά με αυτό των Sung και Poggio [48] (μέγεθος εισόδου, προεπεξεργασία κτλ).

Οι Colmenarez και Huang [1] πρότειναν ένα σύστημα βασισμένο στη σχετική πληροφορία του Kullback (Kullback divergence) για να μετρήσουν τη διαφορά μεταξύ συνδυασμένων ιστογραμμάτων, υπολογισμένα για κάθε ζεύγος εικονοστοιχείων στις εικόνες του συνόλου εκπαίδευσης για τις κλάσεις των προσώπων και μη. Επίσης χρησιμοποίησαν μία αλυσίδα Markov πρώτης τάξης για να μοντελοποιήσουν την χωρική συσχέτιση των εικονοστοιχείων. Ανάμεσα στη σχετική βιβλιογραφία, χρησιμοποίησαν το μικρότερο μέγεθος εισόδου ( $11 \times 11$ ) που έχει αναφερθεί μέχρι στιγμής.

Οι Garcia και Tziritas [9] πρότειναν έναν ταξινομητή προσώπων βασισμένο στην τμηματοποίηση του χρώματος επιδερμίδας και σε στατιστική ανάλυση της υφής του προσώπου. Η τελευταία περιγράφηκε από διανύσματα σχηματισμένα από απλές στατιστικές μετρήσεις (διασπορές) που είχαν εξαχθεί από κάθε υπο-ζώνη (subband) ενός διακριτού τρι-επίπεδου πακέτου ανάλυσης κυματιδίων της εικόνας του προσώπου. Το πακέτο ανάλυσης κυματιδίων, το οποίο συλλαμβάνει πληροφορία σχετική με τις ορατές ιδιότητες σε χώρο, κλίμακα και κατεύθυνση, βρέθηκε αρκετά αποδοτικό για την περιγραφή των χαρακτηριστικών του ανθρώπινου προσώπου. Χρησιμοποίησαν ένα ζεύγος κατάλληλα επιλεγμένων συνηγών τετραγωνικών βαθυπερατών και υψηπερατών φίλτρων τα οποία επίσης είχαν συμπεριλάβει και

**Σχήμα 2.3**Το σύστημα των Rowley *et al.* [41].

σε ένα σύστημα αναγνώρισης προσώπων [10]. Τα εξαγόμενα διανύσματα χαρακτηριστικών χαρακτηριζόντουσαν στην συνέχεια ως πρόσωπα ή μη με βάση την απόσταση Bhattacharyya και κάποια πρότυπα προσώπων κατασκευασμένα μέσω εκπαίδευσης.

Οι Schneiderman και Kanade [45] πρότειναν πιο πρόσφατα έναν ταξινομητή προσώπων επίσης βασισμένο σε ένα τοπικά δειγματοληπτημένο τρι-επίπεδο πακέτο ανάλυσης wavelet. Πολλαπλά σύνολα από τους συντελεστές wavelet εξάχθηκαν από επιλεγμένες υπο-ξώνες του δέντρου wavelet. Αυτοί οι συντελεστές επανα-κβαντοποιήθηκαν σε τρία επίπεδα και πιθανοκρατικές συναρτήσεις πυκνότητας κατασκευάστηκαν από ιστογράμματα. Τέλος, ο κανόνας του Bayes χρησιμοποιήθηκε για την ταξινόμηση μεταξύ προσώπων και μη.

### 2.2.3 Νευρωνικά Δίκτυα

Η πρώτη προχωρημένη προσέγγιση βασισμένη σε νευρωνικά δίκτυα που έδωσε αποτελέσματα σε ένα μεγάλο και δύσκολο σύνολο δοκιμής παρουσιάστηκε από τους Rowley *et al.* [39, 40, 41]. Το σύστημά τους ενσωμάτωσε γνώση για την δομή του προσώπου σε ένα νευρωνικό δίκτυο με ειδική συνδεσμολογία στην είσοδό του των  $20 \times 20$  εικονοστοιχείων, σχήμα 2.3. Στην έκδοση της υλοποίησής τους που χρησιμοποιεί ένα μόνο νευρωνικό δίκτυο (αναφερόμενο ως system 5) υπήρχαν δύο αντίγραφα ενός κρυμμένου στρώματος με 26 μονάδες, δίνοντας συνολικά 52 κρυμμένες μονάδες. Τα πεδία υποδοχής του στρώματος εισόδου ορίστηκαν ως εξής: 4 μονάδες που λαμβάνουν είσοδο από  $10 \times 10$  υπο-περιοχές, 16 από  $5 \times 5$

υπο-περιοχές και 6 από  $20 \times 5$  επικαλυπτόμενες οριζόντιες λωρίδες. Αυτή η τοπολογία φέρει έναν μεγάλο αριθμό από προσαρμόσιμα βάρη (2.905), τα οποία εκπαιδεύτηκαν μέσω του κλασικού αλγορίθμου backpropagation. Το σύνολο εκπαίδευσης αποτελούνταν από 1.050 αυθεντικά παραδείγματα προσώπων και κλιμακώθηκε σε 15 φορές πιο μεγάλο με τεχνητούς μετασχηματισμούς των αυθεντικών. Οι συγγραφείς υιοθέτησαν την ίδια προεπεξεργασία της εισόδου όπως στους Sung και Poggio [48] και την ίδια διαδικασία bootstrapping. Το δίκτυο εφαρμόστηκε για να σαρώσει την εικόνα εισόδου με ένα κινητό  $20 \times 20$  παράθυρο σε κάθε δυνατή τοποθεσία και σε κάθε δυνατή κλίμακα με ένα παράγοντα υποδειγματοληψίας 1,2. Οι απαντήσεις συγκεντρώνονταν και ομαδοποιούνταν για τον σχηματισμό μιας τελικής απάντησης για την ύπαρξη ή όχι ενός προσώπου και την εύρεση της θέσης του και του ύψους του. Για να μειωθεί ο ρυθμός εσφαλμένων ειδοποιήσεων, προτάθηκε ένα σχήμα με πολλαπλά νευρωνικά δίκτυα διαφορετικά εκπαίδευμένα και συνδυασμένα με μία μορφή διαιτησίας. Όσον αφορά θέματα επιτάχυνσης των υπολογισμών, πρότειναν μία μέθοδο με ένα γρήγορο (με μικρή τοπολογία) δίκτυο να λειτουργεί ως στάδιο προ-φιλτραρίσματος για την πρόχειρη σάρωση της εικόνας εισόδου και να δίνει υποψήφιες θέσεις για περισσότερη επεξεργασία. Τέλος, επινοήθηκε μια στρατηγική για την χειρισμό αυθαίρετων εντός-του-πλάνου περιστροφών των προσώπων [42].

Οι Roth *et al.* [38] πρότειναν ένα ανιχνευτή προσώπων βασισμένο σε μία καινούργια αρχιτεκτονική μάθησης, την SNoW (Sparse Network of Winnows - Αραιό Δίκτυο Λιχνισμάτων), η οποία αποτελείται από δύο γραμμικές μονάδες κατωφλίου (αντιπροσωπεύοντας τις δύο κλάσεις πρόσωπα και μη) που λειτουργούν πάνω σε ένα χώρο εισόδου δυαδικών χαρακτηριστικών. Χαρακτηριστικά όπως φωτεινότητα, μέση φωτεινότητα, και διασπορά πρώτα εξάγονταν από μία σειρά υπο-περιοχών του παραθύρου εισόδου και στην συνέχεια διακριτοποιούνταν σε έναν προκαθορισμένο αριθμό κλάσεων για να δώσουν τα δυαδικά χαρακτηριστικά σε έναν 135.424-διάστατο χώρο χαρακτηριστικών. Το σύστημα αυτό εκπαιδεύτηκε με έναν απλό κανόνα εκπαίδευσης που προάγει και υποβιβάζει βάρη σε περιπτώσεις λάθους ταξινόμησης, έτσι ώστε να ταξινομεί δυαδικά χαρακτηριστικά προσώπων και μη.

Οι Féraud *et al.* [3] πρότειναν μία διαφορετική νευρωνική προσέγγιση, βασισμένη σε ελεγχόμενα παραγωγικά μοντέλα (Constrained Generative Models - CGMs), τα οποία είναι αυτο-συσχετιζόμενα πλήρως συνδεμένα MLPs με τοία μεγάλα στρώματα με βάρη εκπαίδευμένα να εκτελούν έναν μη γραμμικό PCA. Η ταξινόμηση στηρίζεται στο σφάλμα επανακατασκευής των CGMs. Τα καλύτερα αποτελέσματα αναφέρθηκαν με έναν συνδυασμό των CGMs με μία υπό όρους μίξη (conditional mixture) και ένα με ένα MLP δίκτυο στο ρόλο της θύρας (gate) αυτής της μίξης. Καθώς το υπολογιστικό κόστος αυτής της μεθόδου ήταν αρκετά υψηλό λόγω της πολύ μεγάλης τοπολογίας των 141.741 βαρών, κάποια προ-φιλτραρίσματα χρησιμοποιήθηκαν, όπως ανίχνευση χρώματος επιδερμίδας και τιμηματοποίηση κίνησης. Το

σύνολο εκπαίδευσης περιείχε 8.000 παραδείγματα προσώπων και για την κατασκευή των επιπλέον παραδειγμάτων μη προσώπων, η τεχνική bootstrapping [48] υιοθετήθηκε επίσης και εδώ. Έγινε επίσης bootstrapping και πάνω στα θετικά παραδείγματα.

### 2.3 Συζήτηση

Είναι κοινή διαπίστωση [15, 56] ότι οι περισσότερες από τις μεθόδους βασισμένες σε χαρακτηριστικά δεν μπορούν να χρησιμοποιηθούν ως ανιχνευτές προσώπων γενικής χρήσεως. Η αποτύμηση αυτών των συστημάτων έγινε σε ένα σύνολο δοκιμής προτεινόμενο από τους ίδιους τους συγγραφείς, που συνήθως περιείχε σκηνές εργαστηρίου, και χωρίς να δίνεται κάποια σύγκριση με άλλες μεθόδους (π.χ. στο σύνολο CMU). Αυτό κάνει πολύ δύσκολη την ποσοτική επιβεβαίωση της δυναμικής ή/και των περιορισμών που έχουν αυτά τα συστήματα. Οι μέθοδοι που στηρίζονται στην χρωματική πληροφορία, οι οποίες έχουν προσελκύσει το ενδιαφέρον της επιστημονικής κοινότητας τελευταία, δεν μπορούν να αποτυμηθούν στο σύνολο CMU επειδή αυτό περιέχει ασπρόμαυρες εικόνες. Παρ' όλα αυτά, δεν έχει καθιερωθεί ακόμα κάποιο κοινό σύνολο δοκιμών που να περιέχει έγχρωμες εικόνες.

Όσον αφορά τις μεθόδους βασισμένες στην εμφάνιση, κάποιες κοινές και συγκεκριμένες δυσκολίες έχουν παρατηρηθεί, στην προσπάθεια αυτών των μεθόδων να εφαρμόσουν κλασικούς αλγορίθμους της αναγνώρισης προτύπων στην ανίχνευση προσώπων. Αρχικά, η διάσταση του χώρου εισόδου είναι αρκετά υψηλή, όντας ένα τμήμα μιας 2-διάστατης εικόνας. Χρησιμοποιώντας ένα παράθυρο  $20 \times 20$ , μας δίνει έναν 400-διάστατο χώρο εισόδου και  $256^{400}$  διακριτές περιπτώσεις να ταξινομηθούν ως πρόσωπα ή μη. Κανονικά, ένας αλγόριθμος που πρέπει να λειτουργήσει σε έναν τέτοιο χώρο εισόδου θα έχει εντελώς φυσικά και έναν μεγάλο αριθμό βαθμών ελευθερίας (ελεύθερες παραμέτρους), το οποίο με την σειρά του απαιτεί μία εξαιρετικά πυκνή σάρωση του χώρου εισόδου για την κατασκευή ενός αποδοτικού συνόλου εκπαίδευσης. Αυτό το πρόβλημα είναι ευρύτατα γνωστό στην βιβλιογραφία της αναγνώρισης προτύπων ως η κατάρα της διάστασης (curse of dimensionality). Η πρώτη λύση, όπως ήδη αναφέρθηκε, είναι ένα τεράστιο σύνολο εκπαίδευσης που να αντιπροσωπεύει αξιόπιστα των χώρο εισόδου, αλλά το οποίο είναι πρακτικά ανέφικτο να κατασκευαστεί καθώς ίσως εκατομμύρια παραδείγματα θα χρειαστούν. Η δεύτερη λύση, υιοθετημένη από τις περισσότερες περιπτώσεις της βιβλιογραφίας όπως είδαμε, είναι αυτής της ορητής μείωσης της διάστασης του χώρου εισόδου (με την χρήση π.χ. του PCA ή του SOM) πριν τροφοδοτήσουμε τον ταξινομητή. Αυτή η λύση έχει το μειονέκτημα ότι ο καινούργιος χώρος στον οποίο τα δεδομένα θα προβληθούν μπορεί να μην είναι βέλτιστος για τις ανάγκες διαχωρισμού μεταξύ προσώπων και μη (δείτε [55]). Έτσι είναι πιθανό να υποβαθμίζει τις ικανότητες διαχωρισμού του συστήματος πριν ακόμα το σήμα εισόδου φθάσει στον ταξινο-

μητή. Μία τρίτη λύση είναι να ενσωματωθεί στο σύστημα εκ των προτέρων γνώση που έχουμε γύρω από το εκάστοτε πρόβλημα (π.χ. ιδιότητες του χώρου εισόδου), με απώτερο σκοπό να μειωθούν οι βαθμοί ελευθερίας του. Αυτή η λύση θα εξεταστεί στο επόμενο κεφάλαιο.

Μία άλλη δυσκολία, που συναντιέται επίσης και στις μεθόδους βασισμένες σε χαρακτηριστικά, είναι ο πλούτος των φυσικών δεδομένων που απαντιέται συνήθως σε περιβάλλοντα εικόνων του πραγματικού κόσμου. Για την μερική αντιμετώπιση αυτού του προβλήματος, χρησιμοποιείται αρκετά συχνά η τεχνική προεπεξεργασίας που είδαμε για την κανονικοποίηση της εισόδου από πολλές δυνατές παραμορφώσεις. Αυτό έχει το παρόπλευρο μειονέκτημα της απεικόνισης κάποιων περιπτώσεων, που τουλάχιστον αρχικά απέχουν πολύ από το να χαρακτηριστούν πρόσωπα, σε προσωποειδείς καταστάσεις.

Τέλος, άλλη μία σοβαρή δυσκολία που συναντιέται είναι το πρόβλημα σχετικά με το ποιο μπορεί να είναι ένα αρκετά αντιπροσωπευτικό/διδακτικό σύνολο με παραδείγματα μη προσώπων για την εκπαίδευση του εκάστοτε ταξινομητή, πρόβλημα έμφυτο σε κάθε ταξινομητή μάθησης με παραδείγματα. Στην πράξη, υπάρχουν αναρίθμητες εικόνες-παραδείγματα μη προσώπων, αλλά μόνο ένα μικρό υποσύνολο αυτών μπορούμε και πρέπει να χρησιμοποιήσουμε για την εκπαίδευση. Αυτό εξαρτάται αρκετά από την φύση του ταξινομητή και επίσης από τον χώρο παραμέτρων μέσα στον οποίον ενεργεί και αναζητεί την λύση. Έτσι, αυτό το ‘βέλτιστο’ σύνολο εκπαίδευσης είναι αρκετά πιθανό να διαφέρει μεταξύ των διάφορων ταξινομητών (π.χ. βασισμένους σε PCA, SVMs, ή ακόμα και σε διαφορετικές τοπολογίες MLPs). Για την αντιμετώπιση αυτού του προβλήματος, η τεχνική bootstrapping επινοήθηκε στην [48], όπου ξητήθηκε από τον ίδιο τον ταξινομητή να διαλέξει αρνητικά παραδείγματα, προκαλώντας τον να μάθει από τα λάθη του. Μία σημαντική συνθήκη για την επιτυχία αυτής της τεχνικής είναι η καλή δυνατότητα γενίκευσης του ταξινομητή. Πρέπει να μάθει γρήγορα από αυτές τις εσφαλμένες ειδοποιήσεις και ιδεατά να σταματήσει να παράγει περισσότερες καθώς η εκπαίδευση εξελίσσεται. Διαφορετικά, θα χρειαστεί και πάλι μία πυκνή σάρωση του χώρου εισόδου.



## ΚΕΦΑΛΑΙΟ 3

# Συνελικτικά Νευρωνικά Δίκτυα

Οι αρχές της αρχιτεκτονικής σχεδίασης των Συνελικτικών Νευρωνικών Δικτύων περιγράφονται σε αυτό το κεφάλαιο. Σε αυτήν την τάξη των πολυστρωματικών perceptrons, ενσωματώνεται απευθείας στην τοπολογία του δικτύου η εκ των προτέρων γνώση που έχει ο μηχανικός πάνω στο συγκεκριμένο πρόβλημα με την μορφή περιορισμών. Η τοπολογία του δικτύου αποκτά με αυτόν τον τρόπο κάποια δομή, προικίζοντάς το με πολλές αρετές όπως θα δούμε. Τα συνελικτικά δίκτυα εφαρμόσθηκαν με εξαιρετική επιτυχία στο πεδίο της Οπτικής Αναγνώρισης Χαρακτήρων (Optical Character Recognition) [26], αποτελώντας το τρέχων πρότυπο σύστημα σε αυτό το πεδίο. Μία πρώτη προσπάθεια που έγινε για την εφαρμογή των συνελικτικών δικτύων στην ανίχνευση προσώπων [51] θα παρουσιαστεί επίσης.

Η πρώτη εργασία η οποία εισήγαγε κάποιες ιδέες αρχιτεκτονικής που οδήγησαν στα συνελικτικά δίκτυα ήταν αυτή του Fukushima [5, 6]. Ο συγγραφέας εμπνεύστηκε από την εργασία των Hubel και Weisel [17], στην οποία ανακαλύφθηκε κάποια δομή στις συνδέσεις του φλοιού του εγκεφάλου της γάτας. Μία τοπικότητα των πεδίων υποδοχής των νευρώνων του οπτικού συστήματος αποκαλύφθηκε, επιτρέποντας ευαισθησία σε τοπικούς και στοιχειώδης ερεθισμούς (χαρακτηριστικά) όπως κατευθυνόμενες ακμές. Ο Fukushima έχτισε μία πολυστρωματική νευρωνική τοπολογία με εκτεταμένη χρήση τοπικών συνδέσεων και συνελιξεων. Προτάθηκε ένα σχήμα εκπαίδευσης στρώμα-σε-στρώμα, βασισμένο σε ανταγωνιστική μάθηση (competitive learning). Οι Le Cun *et al.* [24, 25, 26] γενίκευσαν από αυτά τα δεδομένα και τα ενσωμάτωσαν, μεταξύ άλλων, σε πολυστρωματικές τοπολογίες MLPs, εκπαιδευμένες με τον αλγόριθμο backpropagation.

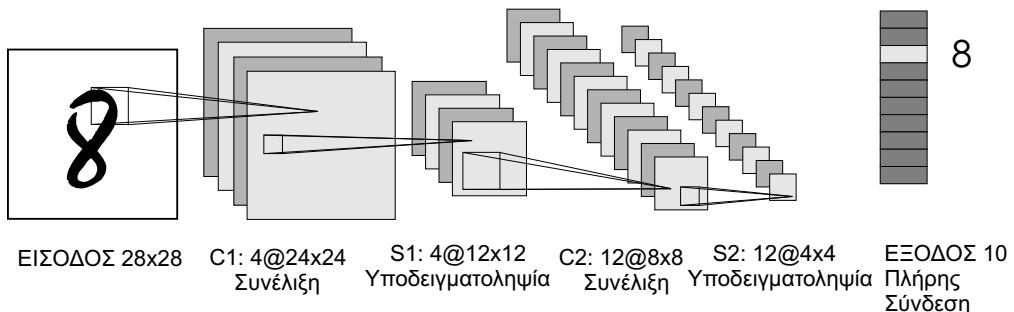
### 3.1 Αρχές Αρχιτεκτονικής

Για να εισάγουμε την λογική πίσω από τις επιλογές τις αρχιτεκτονικής σχεδίασής τους, ας σημειώσουμε πρώτα ότι στις πολυστρωματικές νευρωνικές τοπολογίες μία (μη γραμμική)

εξαγωγή χαρακτηριστικών στα κρυμμένα στρώματα προηγείται της πραγματικής ταξινόμησης που λαμβάνει χώρα στο τελευταίο στρώμα (στρώμα εξόδου). Αναφέρθηκε πριν ότι, το οπτικό σύστημα της γάτας προτιμά να εξαγάγει στοιχειώδη χαρακτηριστικά στα πρώτα επίπεδα της ενεργοποίησης του εγκεφάλου. Η ίδια στρατηγική μπορεί να ακολουθηθεί και στην αναγνώριση χαρακτήρων, καθώς πράγματι αυτοί αποτελούνται από κάποια σύνολα από στοιχειώδη χαρακτηριστικά όπως κάθετες ή οριζόντιες γραμμές, καμπύλες κτλ. Αυτά τα χαρακτηριστικά μπορούν να εξαχθούν από τα κρυμμένα στρώματα ενός νευρωνικού δικτύου, αλλά με έναν ειδικό τρόπο καθώς φέρουν κάποιες ενδιαφέρουσες ιδιότητες. Εκτός του ότι είναι τοπικά από την φύση τους, η συγκεκριμένη θέση που έχουν μπορεί να αλλάξει μέσα στην εικόνα εισόδου π.χ. περιστρέφοντας τον χαρακτήρα ή αλλάζοντάς τον, καθώς διαφορετικοί χαρακτήρες μπορούν να έχουν κάποια κοινά χαρακτηριστικά αλλά σε διαφορετικές θέσεις. Μάλιστα, αυτό που είναι πραγματικά χρήσιμο δεν είναι η ακριβής θέση ενός χαρακτηριστικού, αλλά η σχετική του θέση ανάμεσα στα υπόλοιπα χαρακτηριστικά. Για παράδειγμα, ο χαρακτήρας ‘8’ μπορεί να αναγνωριστεί προσεγγιστικά σαν δύο κύκλοι (ή οβάλ), περίπου ο ένας πάνω στον άλλο. Δεν παίζει ιδιαίτερη σημασία αν βρίσκονται σε επαφή, ή η τυχόν κατεύθυνσή τους. Τέλος, πρέπει να μεριμνήσουμε και για πιθανές παραμορφώσεις του σήματος εισόδου καθώς αυτό προέρχεται συνήθως από τον πραγματικό κόσμο. Γενικότερα, οι χειρόγραφοι χαρακτήρες ποικίλουν αρκετά σε σχέση με τον χαρακτήρα ‘πρότυπο’ καθώς ο κάθε γράφων έχει το δικό του προσωπικό γραφικό χαρακτήρα, ο χαρακτήρας μπορεί να είναι κακογραμμένος κτλ.

Έχοντας συζητήσει πάνω στην γενικότερη φύση του σήματος εισόδου, μπορούμε τώρα να συνοψίσουμε τις αρχές σχεδίασης των συνελικτικών νευρωνικών δικτύων ως εξής [26]:

- **Τοπικότητα των πεδίων υποδοχής** (local receptive fields). Το πεδίο υποδοχής ενός νευρώνα πρέπει να περιορισθεί σε μία τοπική γειτονιά της εισόδου ή του προηγούμενου στρώματος γενικότερα. Αυτό επιτρέπει την εξαγωγή τοπικών και στοιχειωδών χαρακτηριστικών.
- **Διαμοιρασμός των βαρών** (weight sharing). Τα βάρη των νευρώνων πρέπει να αντιγράφονται χωρικά μέσα σε ένα δεδομένο στρώμα. Με αυτόν τον τρόπο, το ίδιο τοπικό χαρακτηριστικό μπορεί να ανιχνευτεί σε διαφορετικές τοποθεσίες της εισόδου, δίνοντας στο δίκτυο την εκ κατασκευής δυνατότητα σταθερότητας (μη ευαισθησίας) στην μεταπότιση της εισόδου. Η συνδυασμένη λειτουργία αυτών των δύο αρχών, της τοπικότητας των πεδίων υποδοχών και του διαμοιρασμού των βαρών, ισοδυναμεί με την καθαρή συνέλιξη της εισόδου με έναν εκπαιδεύσιμο πυρήνα (έτσι και το όνομα “συνελικτικά” νευρωνικά δίκτυα).
- **Χωρική υποδειγματοληψία** (spatial subsampling). Έχοντας εξάγει τα τοπικά χαρακτηριστικά, οι διαδικασίες που ακολουθούν θα πρέπει να είναι οι μετατόπιση και η υποδειγματοληψία των πεδίων υποδοχής.



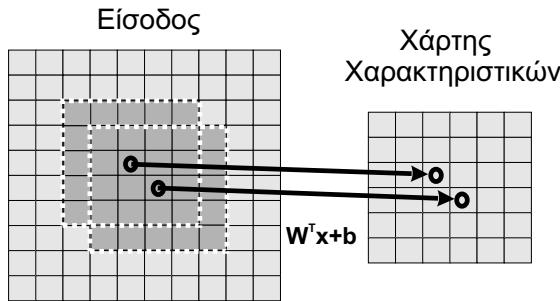
**Σχήμα 3.1**  
Η τοπολογία του LeNet-1.

κτηριοτικά, η ανάλυση του σήματος μπορεί να μειωθεί με την τοπική λήψη του μέσου όρου και την υποδειγματοληψία αυτού. Με αυτόν το τρόπο, μόνον μία προσεγγιστική τοποθεσία των χαρακτηριστικών διατηρείται και σε σχετική θέση με αυτές των άλλων χαρακτηριστικών. Αυτό οδηγεί σε έναν δεδομένο βαθμό εκ κατασκευής σταθερότητας του δικτύου στην αλλοίωση της εισόδου. Πρέπει επίσης να σημειωθεί ότι αυτή η λειτουργία μειώνει την διάσταση του σήματος, άρα και την πολυπλοκότητα της ταξινόμησης, ενώ διατηρεί τα χρήσιμα και κατατοπιστικά χαρακτηριστικά.

- **Συνδυασμός χαρακτηριστικών** (feature combination). Στα μεσαία στρώματα του δικτύου πρέπει να εξάγονται χαρακτηριστικά υψηλότερης τάξης. Αυτό μπορεί να γίνει με τον συνδυασμό του τελικού αποτελέσματος των διάφορων εξαγωγών χαρακτηριστικών των πρώτων στρωμάτων.

## 3.2 Ένα Παράδειγμα Συνελικτικής Τοπολογίας

Σε αυτήν την ενότητα παρουσιάζεται μία απλή συνελικτική τοπολογία, το δίκτυο “LeNet-1” [25], για την επίδειξη του πως εφαρμόζονται οι παραπάνω αρχές στην πράξη και για το πρόβλημα της οπτικής αναγνώρισης ψηφίων. Μία λεπτομερής περιγραφή μιας πιο προχωρημένης τοπολογίας (δηλ. με περισσότερα κρυμμένα στρώματα κτλ) μπορεί να βρεθεί στην [26] (“LeNet-5”). Η τοπολογία του LeNet-1 απεικονίζεται στο σχήμα 3.1. Η είσοδος του δικτύου είναι μία  $28 \times 28$  εικόνα που περιέχει τον χαρακτήρα προς αναγνώριση προσεγγιστικά στο κέντρο της. Τα κρυμμένα στρώματα C1 μέχρι S2 αποτελούνται από πολλαπλά επίπεδα στα οποία αποτυπώνονται τα αποτελέσματα των συνελίξεων και των υποδειγματοληψιών. Αυτά τα επίπεδα ονομάζονται χάρτες χαρακτηριστικών (feature maps) στην ορολογία των συνελικτικών δικτύων. Η ύπαρξη πολλαπλών χαρτών χαρακτηριστικών ανά στρώμα επιτρέπει την εξαγωγή διαφορετικών ειδών από χαρακτηριστικά στο ίδιο στρώμα. Ο αριθμός



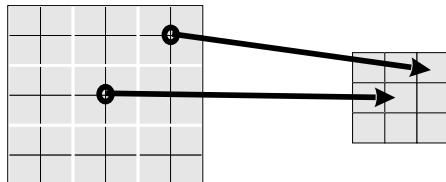
### Σχήμα 3.2

Η λειτουργία της συνέλιξης. Κάθε κελί της εισόδου αντιστοιχεί σε ένα εικονοστοιχείο της εικόνας ενώ κάθε κελί του χάρτη χαρακτηριστικών αντιστοιχεί σε έναν νευρώνα. Στα αριστερά μπορούμε να διακρίνουμε τα πεδία υποδοχής των νευρώνων και πως αυτά αλληλοεπικαλύπτονται. Κάθε νευρώνας υπολογίζει το εσωτερικό γινόμενο  $W^T x + b$ , όπου τα  $W, b$  είναι τα διαμοιρασμένα βάρη αυτού του χάρτη χαρακτηριστικών.

χαρτών χαρακτηριστικών είναι μία καθαρή επιλογή του μηχανικού και μπορεί να τεθεί μετά από πειραματισμό. Στο τελευταίο στρώμα (έξοδος), υπάρχουν 10 νευρώνες οι οποίοι δίνουν και το τελικό αποτέλεσμα της αναγνώρισης. Αυτό το στρώμα είναι πλήρως συνδεμένο με το προηγούμενό του, όπως στις κλασικές τοπολογίες των MLPs.

Κάθε νευρώνας του στρώματος C1 δέχεται είσοδο μόνον από την αντίστοιχη  $5 \times 5$  τοπική γειτονιά της εικόνας εισόδου, εφαρμόζοντας την αρχή της τοπικότητας των πεδίων υποδοχής (σχήμα 3.2). Έτσι, έχει 25 εκπαιδεύσιμα βάρη και επιπλέον ένα προσθετικό εκπαιδεύσιμο bias. Τα πεδία υποδοχής διαδοχικών νευρώνων αλληλοεπικαλύπτονται κατά  $4 \times 4$  εικονοστοιχεία, με μία έννοια που θυμίζει αρκετά την λειτουργία της συνέλιξης. Επιπρόσθετα, εφαρμόζοντας την αρχή των διαμοιρασμένων βαρών, όλοι οι νευρώνες κατά πλάτος ενός χαρτη χαρακτηριστικών περιορίζονται να έχουν το ίδιο σύνολο από 26 βάρη. Για την εκπαίδευση αυτών των βαρών απαιτείται μία μικρή τροποποίηση των εξισώσεων του αλγορίθμου backpropagation η οποία θα εξεταστεί αργότερα σε αυτήν την ενότητα. Τέλος, τα αποτελέσματα των τεσσάρων συνελίξεων θα δώσουν τέσσερις χάρτες χαρακτηριστικών διάστασης  $24 \times 24$  καθώς δεν λαμβάνονται υπ' όψη τα συμπτώματα ορίων των συνελίξεων. Σημειώστε ότι η λειτουργία αυτού του στρώματος είναι καθαρά γραμμική.

Το στρώμα S1 απαρτίζεται από τέσσερις χάρτες χαρακτηριστικών, ένα για κάθε χάρτη του στρώματος C1. Τα πεδία υποδοχής της κάθε μονάδας (σχήμα 3.3) είναι μία  $2 \times 2$  περιοχή του αντίστοιχου χάρτη χαρακτηριστικών του προηγούμενου στρώματος. Κάθε μονάδα υπολογίζει τον μέσο όρο των τεσσάρων εισόδων του, πολλαπλασιάζει με αυτό έναν εκπαιδεύσιμο συντελεστή, προσθέτει ένα εκπαιδεύσιμο bias και, τελικά, περνάει το αποτέλεσμα από μία σιγμοειδή συνάρτηση. Διαδοχικοί νευρώνες δεν φέρουν αλληλοεπικαλυπτόμενα πεδία υπο-



### Σχήμα 3.3

Η λειτουργία της υποδειγματοληψίας. Σε κάθε περιοχή τεσσάρων σημείων υπολογίζεται ο μέσος όρος, πολλαπλασιάζεται με έναν συντελεστή, αθροίζεται με ένα bias και διέρχεται μέσα από μία σιγμοειδή συνάρτηση.

δοχών εδώ πέρα. Έτσι έχουμε σαν αποτέλεσμα την υποδειγματοληψία της εισόδου με ένα παράγοντα 2 και στις δύο κατευθύνσεις, παίρνοντας χάρτες χαρακτηριστικών που έχουν τον μισό αριθμό γραμμών και στηλών σε σύγκριση με αυτούς του προηγούμενου στρώματος. Η αρχή του διαμοιρασμού των βαρών εξακολουθεί να ισχύει και εδώ. Σε αυτό το στρώμα, το σήμα καθαρίζεται από πιθανό θόρυβο, οι παραμορφώσεις του εξομαλύνονται και, τελικά, διατηρούνται μόνο τα χρήσιμα χαρακτηριστικά για τους σκοπούς της ταξινόμησης.

Στο στρώμα C2 υπάρχουν 12 χάρτες χαρακτηριστικών οι οποίοι είναι συνδεμένοι στο στρώμα S1 με μία έννοια συνδυασμού χαρακτηριστικών. Πιο συγκεκριμένα, κάποιοι από τους χάρτες αυτού του επιπέδου μπορούν να συγχωνεύσουν τα αποτελέσματα ενός οι περισσότερων χαρτών του προηγούμενου στρώματος. Για να επιτευχθεί αυτό, ένας χάρτης του C2 μπορεί να έχει δύο ή περισσότερους πυρήνες συνέλιξης εφαρμοζόμενους σε δύο ή περισσότερους χάρτες του στρώματος S1. Τα αποτελέσματα αυτών των συνέλιξεων αθροίζονται στην συνέχεια στον χάρτη του C2, μαζί με το αθροιστικό bias όπως πάντα. Εκτός από αυτήν την συγχώνευση χαρακτηριστικών, η λειτουργία των επιπέδων C2 και S2 είναι ακριβώς η ίδια με αυτήν των επιπέδων C1 και S1, αντίστοιχα. Στην τοπολογία LeNet-1, οι πρώτοι τέσσερις χάρτες του C2 είναι συνδεμένοι μόνον με τους αντίστοιχους τους χάρτες του S1, χωρίς να επιτελούν κάποια συγχώνευση χαρακτηριστικών. Οι υπόλοιποι οκτώ χάρτες συγχωνεύουν τα αποτελέσματα όλων των δυνατών ζευγαριών χαρτών του S1. Γενικά, το σχήμα σύνδεσης μεταξύ αυτών των δύο στρωμάτων είναι καθαρά επιλογή του μηχανικού. Μία σημαντική παρατήρηση [26] είναι ότι αυτές οι συνδέσεις πρέπει κατά κάποιον τρόπο να δημιουργούν μία ασυμμετρία στο δίκτυο, επιτρέποντας μία πιο αποδοτική διασπορά των χαρακτηριστικών διαμέσου του δικτύου (πρόβλημα απόδοσης ευθυνών - credit-assignment problem του αλγορίθμου backpropagation).

Τέλος, το στρώμα εξόδου περιλαμβάνει 10 νευρώνες που λειτουργούν όπως στις κλασικές τοπολογίες MLP, καθώς είναι πλήρως συνδεμένοι με όλους του χάρτες χαρακτηριστικών του στρώματος S2, και σε όλα τα σημεία τους. Είναι αρκετά ενδιαφέρον, ότι τα περισσότερα εκπαιδεύσιμα βάρη του δικτύου φιλοξενούνται από αυτό το στρώμα (κάθε νευρώνας του

έχει  $12 \times 4 \times 4 = 192$  βάροη, δίνοντας συνολικά 1.920 ενώ ο συνολικός αριθμός βαρών του δικτύου είναι 2.578).

Για την διαμόρφωση των διαμοιρασμένων βαρών κατά την διάρκεια της εκπαίδευσης, χρειάζεται μία καινούργια φόρμα του αλγορίθμου backpropagation. Αυτό που αλλάζει τώρα είναι η έκφραση που δίνει την τοπική παραγωγή του διαδιδόμενου προς τα πίσω σήματος του σφάλματος  $E$ , σε συνάρτηση των διαμοιρασμένων βαρών  $w_s$  (εξίσωση B.4). Με την απλή σκέψη ότι κάθε χάρτης χαρακτηριστικών περιέχει στην πραγματικότητα έναν μοναδικό νευρώνα αλλά με πολλαπλές υπάρξεις, η τοπική παραγωγής  $\partial E / \partial w_s$  για αυτόν τον νευρώνα είναι απλώς το άθροισμα των τοπικών παραγώγων σε όλες τις επιμέρους υπάρξεις του. Έτσι, η ξητούμενη έκφραση δίνεται από [26]:

$$\frac{\partial E}{\partial w_s} = \sum_{k \in S} \frac{\partial E}{\partial w_k}$$

όπου  $w_k$  είναι το βάρος σαν να μην ήταν διαμοιρασμένο και  $S$  είναι το σύνολο όλων των συνδέσεων που χρησιμοποιούν το διαμοιρασμένο βάρος  $w_s$ .

### 3.3 Μία Πρώτη Εφαρμογή στην Ανίχνευση Προσώπων

Οι Vaillant *et al.* [51] χρησιμοποίησαν συνελικτικά δίκτυα για ανίχνευση αντικειμένων σε εικόνες και ασχολήθηκαν ειδικά με την περίπτωση της ανίχνευσης προσώπων. Σε αυτήν την ενδιαφέρουσα προκαταρκτική προσέγγιση, οι συγγραφείς πρότειναν μία απλή αρχιτεκτονική διαμοιρασμένων βαρών με τέσσερις συνελικτικούς και τέσσερις υποδειγματοληπτικούς χάρτες χαρακτηριστικών (όπως στα στρώματα C1 και S1 της τοπολογίας του LeNet-1). Ένα ακόμα κρυμμένο στρώμα με νευρώνες συνδεμένους πλήρως με τους χάρτες χαρακτηριστικών ακολουθούσε και, τέλος, ένα στρώμα εξόδου με ένα νευρώνα που δήλωνε την παρουσία ή όχι προσώπου στην εικόνα εισόδου. Σαν είσοδος στο δίκτυο δινόταν μία εικόνα  $20 \times 20$  αφού διέρθει πρώτα από ένα λαπλασιανό φίλτρο. Αυτή η τοπολογία είχε 1.157 εκπαίδευσιμα βάροη. Ένα σύνολο εκπαίδευσης από 1.792 παραδείγματα προσώπων χρησιμοποιήθηκε για την εκπαίδευση του δικτύου, συνεπικουρούμενο από ένα ίσο αριθμό παραδειγμάτων μη προσώπων, επιλεγμένων με το χέρι (δηλ. δεν χρησιμοποιήθηκε κάποια διαδικασία bootstrapping).

Οι συγγραφείς θεώρησαν επίσης και το πρόβλημα της εύρεσης μιας αποδοτικής στρατηγικής ανίχνευσης. Συμπέραναν ότι είναι πιο αποδοτικό να χρησιμοποιηθεί αρχικά ένα δίκτυο για να ανιχνεύει και να εντοπίζει προσεγγιστικά υποψήφια πρόσωπα, με έννοια προφίλ προσώπων. Για την εκπαίδευση αυτού του δικτύου, μετατόπισαν τεχνητά τα παραδείγματα προσώπων μέσα στην εικόνα εισόδου, κατά την διάρκεια της εκπαίδευσης. Όταν ένα πρόσωπο βρισκόταν ακριβώς στο κέντρο της, η επιθυμητή απάντηση γινόταν ίση με 1, ενώ στις υπόλοιπες περιπτώσεις έπαιρνε κάποια μικρότερη τιμή. Ακολουθούσε ένα δεύτερο

δίκτυο (με την ίδια τοπολογία αλλά εκπαιδευμένο έτσι ώστε να εντοπίζει με ακρίβεια ένα πρόσωπο), για την σάρωση της εικόνας γύρω από το υποψήφιο πρόσωπο. Οι απαντήσεις χρησιμοποιούντουσαν για τον σχηματισμό μιας τελικής απόφασης αν είναι τελικά πρόσωπο ή όχι. Κάποια ομαδοποίηση εφαρμόζοταν για την δημιουργία μιας τελικής απάντησης για την θέση και την κλίμακα του προσώπου μέσα στην συνολική εικόνα, με έναν τρόπο αρκετά άμιον με αυτό που προτάθηκε αργότερα από τους Rowley *et al.* [41]. Τέλος, δεν δόθηκαν λεπτομερή αποτελέσματα για την απόδοση της μεθόδου σε κάποιο σύνολο δοκιμής.

Η χρήση συνελικτικών νευρωνικών δικτύων επίσης προτάθηκε και για το πρόβλημα της αναγνώρισης προσώπων. Οι Lawrence *et al.* [21, 22] χρησιμοποίησαν συνελικτικά δίκτυα ως ένα μέρος ενός ταξινομητή επιφορτισμένου να ταυτοποιεί ένα άτομο από την 2Δ εικόνα του. Οι συγγραφείς προτίμησαν να μην τροφοδοτήσουν το συνελικτικό δίκτυο απευθείας με τα εικονοστοιχεία της εικόνας εισόδου αλλά να εφαρμόσουν πρώτα μείωση της διάστασης της εισόδου μέσω του αλγορίθμου SOM. Μεταξύ άλλων, συνέκριναν την απόδοση των συνελικτικών δικτύων με αυτήν των κλασικών τοπολογιών MLPs, αναφέροντας ότι τα πρώτα υπερέχουν καθαρά.

## 3.4 Συζήτηση

Τα συνελικτικά νευρωνικά δίκτυα είναι ικανά να λύσουν δύο σημαντικά προβλήματα ταυτόχρονα: το πρόβλημα της κατάρας της διάστασης και το πρόβλημα της εκ κατασκευής σταθερότητας στις παραμορφώσεις της εισόδου. Το πρώτο πρόβλημα παρουσιάστηκε στην ενότητα 2.3 και είχε ειπωθεί ότι μπορεί να λυθεί χρησιμοποιώντας εκ των προτέρων γνώση πάνω στην φύση του προβλήματος. Τα συνελικτικά δίκτυα είναι γενικά μεγάλες νευρωνικές τοπολογίες (η ‘απλή’ τοπολογία του LeNet-1 φέρει 98.441 συνδέσεις), όπως απαιτείται από τον εξαιρετικά μεγάλης διάστασης χώρου της εισόδου, και που θα μπορούσαν να υποφέρουν από το παραπάνω πρόβλημα. Η συνεισφορά τους είναι οι περιορισμοί που τίθενται στο δίκτυο, οι οποίοι μειώνουν τον συνολικό αριθμό των ελεύθερων παραμέτρων κατά ένα σημαντικό παράγοντα (από 98.442 συνδέσεις σε 2.578 βάροη, στην περίπτωση του LeNet-1). Έτσι η υπόθεση (hypothesis) αναζητείται σε ένα χώρο πολύ μικρότερης διάστασης. Επιπλέον, αυτοί οι περιορισμοί είναι αρκετά χρήσιμοι και πετυχημένοι καθώς προέρχονται από την κοινή μας αίσθηση πάνω στην επεξεργασία εικόνας. Με αυτό τον τρόπο δεν εκφυλίζονται οι ικανότητες μάθησης του δικτύου. Τέλος, αυτοί οι περιορισμοί παρέχουν στο δίκτυο κάποιες αρκετά χρήσιμες ιδιότητες σταθερότητας στην μετατόπιση, στην περιστροφή και, γενικότερα, στις παραμορφώσεις οποιασδήποτε φύσης της εισόδου, όπως εξετάσαμε παραπάνω.

Τα συνελικτικά δίκτυα εφαρμόσθηκαν με μεγάλη επιτυχία σε ένα από τα πιο δύσκολα προβλήματα της αναγνώρισης προτύπων που έχουν αντιμετωπισθεί ως τώρα, αυτό της ανα-

γνώρισης χειρόγραφων ψηφίων. Οι Le Cun *et al.* [26] συνέλεξαν ένα πολύ μεγάλο σύνολο εκπαίδευσης από 60.000 παραδείγματα εκπαίδευσης και 10.000 δοκιμής και εφάρμοσαν πάνω σε αυτά μία σειρά από αλγορίθμους μάθησης, συμπεριλαμβάνοντας φυσικά και τα συνελικτικά δίκτυα. Μερικοί από τους άλλους αλγορίθμους είναι τα MLPs, ο ταξινομητής Nearest Neighbor, ο PCA, οι SVMs και η Tangent Distance. Τα συνελικτικά δίκτυα έδωσαν τα καλύτερα αποτελέσματα με λιγότερο από 1% ποσοστό λάθους στο σύνολο δοκιμής. Οι SVMs βρέθηκαν πολύ κοντά από άποψη απόδοσης, αλλά με το μειονέκτημα του αρκετά μεγάλου κόστους που απαιτούν σε υπολογιστικούς πόρους. Είναι αρκετά αξιοσημείωτο ότι αυτά τα ποσοστά επιτυχίας είναι αρκετά κοντά στα ανθρώπινα, καθώς οι περισσότεροι από τους χαρακτήρες που ταξινομήθηκαν εσφαλμένα ήταν εξαιρετικά παραμορφωμένοι και δύσκολα αναγνώσιμοι ακόμα και από το ανθρώπινο μάτι. Ένα άλλο αποφασιστικό πλεονέκτημα που διαθέτουν τα συνελικτικά δίκτυα είναι η οικονομία υπολογισμών που μπορεί να επιτευχθεί όταν σαρώνουμε μία ολόκληρη εικόνα. Όπως θα δούμε στην ενότητα 5.3, η όλη διαδικασία μπορεί να ελαχιστοποιηθεί σε μία σειρά από γραμμικά και μη γραμμικά φιλτραρίσματα της εικόνας, κάνοντάς την παρά πολύ γρήγορη και εύκολη στην υλοποίηση.

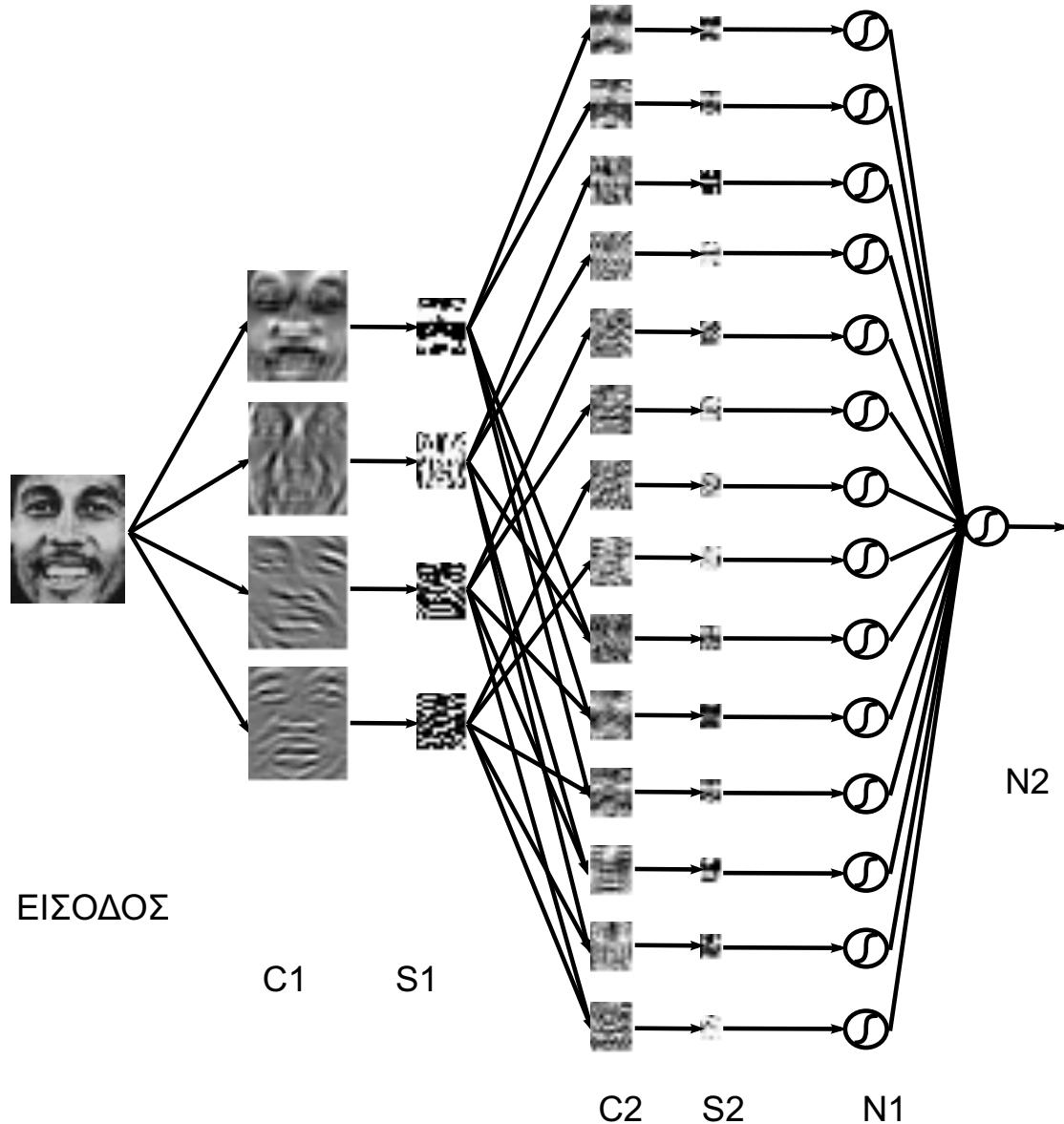
## ΚΕΦΑΛΑΙΟ 4

# Προτεινόμενη Τοπολογία και Μεθοδολογία Εκπαίδευσης

Σε αυτό το κεφάλαιο παρουσιάζεται λεπτομερώς η προτεινόμενη συνελικτική τοπολογία για την ανίχνευση προσώπων και η τακτική bootstrapping που χρησιμοποιήθηκε για την εκπαίδευση του δικτύου. Η τοπολογία αυτή μοιράζεται τις ιδέες αρχιτεκτονικής που παρουσιάστηκαν στον προηγούμενο κεφάλαιο και δεν θα συζητηθούν και πάλι εδώ. Στην συνέχεια δίνεται μία σύντομη περιγραφή του συνόλου εκπαίδευσης από παραδείγματα προσώπων. Κατά τις γνώσεις του συγγραφέα, είναι το μεγαλύτερο (3.702 αυθεντικά παραδείγματα) ποτέ που επιστρατεύτηκε στην σχετική βιβλιογραφία, όπως και το πλουσιότερο καθώς περιέχει καθαρά φυσικά δεδομένα. Ακολουθεί η παρουσίαση της διαδικασίας bootstrapping, η οποία βασίζεται σε αυτήν της [48], αλλά κατάλληλα διαμορφωμένη σε κάποια σημεία για την βελτίωση των αποτελεσμάτων. Τέλος, θα δοθούν τα αποτελέσματα της εκπαίδευσης όπως και μία σύγκριση της απόδοσης μεταξύ διαφορετικών συνελικτικών τοπολογιών και του γραμμικού ταξινομητή.

### 4.1 Τοπολογία του Δικτύου

Η τοπολογία του δικτύου δίνεται στο σχήμα 4.1. Αποτελείται από εξι στρώματα, S1 εως N2, όπου εκτελούνται διαδοχικά συνελίξεις και υποδειγματοληψίες με έναν τρόπο παρόμοιο με αυτόν που είδαμε στην ενότητα 3.2. Η είσοδος του δικτύου είναι μία  $32 \times 36$  περιοχή μιας εικόνας η οποία θα χαρακτηριστεί ως πρόσωπο ή μη από το δίκτυο. Οι τιμές των εικονοστοιχείων έχουν γραμμικά προσαρμοστεί στο διάστημα  $[-1, 1]$ , το οποίο έχει σχέση με το διάστημα τιμών που οι νευρώνες του δικτύου λειτουργούν, όπως θα δούμε αργότερα. Καμία βελτίωση δεν γίνεται στο σήμα εισόδου, όπως διόρθωση φωτισμού ή εξισορρόπηση ιστογράμματος, αφήνοντας το δίκτυο να τροφοδοτείται κατευθείαν από τα εικονοστοιχεία



**Σχήμα 4.1**

Η προτεινόμενη συνελικτική τοπολογία. Τα περιεχόμενα των χαρτών χαρακτηριστικών απεικονίζουν τα πραγματικά χαρακτηριστικά που έχουν εξαχθεί από ένα πραγματικό παράδειγμα.

-	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	X	X							X	X	X			
2			X	X					X			X	X	
3				X	X				X		X		X	
4					X	X				X		X	X	

**Πίνακας 4.1**

Πλάνο σύνδεσης μεταξύ των χαρτών χαρακτηριστικών των στρωμάτων C2 (οριζόντιοι) και S1 (κάθετοι).

της εικόνας. Κατά τις γνώσεις του συγγραφέα, η εργασία αυτή είναι η πρώτη νευρωνική προσέγγιση κατά την οποία δεν επιστρατεύεται προεπεξεργασία της εισόδου με κανένα τρόπο.

Το στρώμα C1 περιέχει τέσσερις χάρτες χαρακτηριστικών διάστασης  $28 \times 32$ , φέροντες από ένα  $5 \times 5$  συνελικτικό πυρήνα ο καθένας. Η λειτουργία αυτών των χαρτών είναι ακριβώς η ίδια όπως αυτή των αντίστοιχων χαρτών του στρώματος C1 της τοπολογίας LeNet-1 (σχήμα 3.2). Υποθέτοντας ότι αυτό στρώμα να είναι πλήρως συνδεμένο στην είσοδο, όπως στις κλασικές MLP τοπολογίες, θα είχε σαν αποτέλεσμα  $4.132.352$  συνδέσμους. Η εφαρμογή της αρχής της τοπικότητας των πεδίων υποδοχής περιορίζει τον αριθμό των συνδέσμων αυτού του στρώματος σε 93.184. Επιπλέον, η εφαρμογή της τεχνικής διαμοιρασμού των βαρών μειώνει τον αριθμό των ελεύθερων παραμέτρων ακόμα περισσότερο σε μόλις 104, δεδομένου ότι κάθε χάρτης χαρακτηριστικών έχει  $5 \times 5 + 1 = 26$  εκπαιδεύσιμα βάροντα. Στο στρώμα S1, επιτελείται η λειτουργία της υποδειγματοληψίας του σχήματος 3.3. Η διάσταση του σήματος θα μειωθεί σε αυτό το σημείο σε  $14 \times 16$ . Αυτό το στρώμα περιέχει μόλις 8 εκπαιδεύσιμα βάροντα (τέσσερις συντελεστές για τους πολλαπλασιασμούς και τέσσερα προσθετικά bias), παρά τους 4.480 συνδέσμους που διαθέτει με το στρώμα S1.

Στο στρώμα C2 υπάρχουν 14 χάρτες χαρακτηριστικών που χρησιμοποιούν  $3 \times 3$  συνελικτικούς πυρήνες, αντί για  $5 \times 5$  του στρώματος C1 ή του αντίστοιχου στρώματος C2 του LeNet-1. Κάποιοι από αυτούς τους χάρτες συνδέονται με ένα μόνο χάρτη του στρώματος S1 ενώ οι υπόλοιποι με δύο, επιτελώντας συνδυασμό χαρακτηριστικών. Ο τρόπος σύνδεσης των δύο αυτών στρωμάτων δίνεται στον πίνακα 4.1. Κάθε ένας από τους τέσσερις υποδειγματοληπτικούς χάρτες του στρώματος S1 δίνει είσοδο σε δύο χάρτες του στρώματος C2. Αυτό δίνει σαν αποτέλεσμα τους πρώτους οκτώ χάρτες χαρακτηριστικών του στρώματος C2. Κάθε ένας από τους έξι υπόλοιπους χάρτες του C2 δέχεται είσοδο από έναν εκ των δυνατών συνδυασμών ανά ζεύγη των χαρτών του S1. Έτσι, αυτό το στρώμα έχει συνολικά 20 συνελικτικούς πυρήνες και 30.254 συνδέσεις με το στρώμα S1, αλλά μόνον 194 εκπαιδεύσιμα βάροντα. Ακολουθείται από το στρώμα S2 το οποίο λειτουργεί ακριβώς με τον ίδιο τρόπο με το στρώμα

S1. Η διάσταση του σήματος μειώνεται σε αυτό το σημείο σε  $6 \times 7$  για νάθε έναν από τους 14 χάρτες χαρακτηριστικών. Αυτό το στρώμα έχει 2.940 συνδέσεις και 28 εκπαιδεύσιμα βάροη.

Σαν τελικό αποτέλεσμα των συνελικτικών/υποδειγματοληπτικών στρωμάτων C1 ως S2, μία σειρά από, ελπίζουμε σταθερά και μη συσχετίσιμα, χαρακτηριστικά μικρής διάστασης  $6 \times 7$  έχουν εξαχθεί για να χρησιμοποιηθούν από τα στρώματα N1 και N2, φέροντας σε πέρας την ταξινόμηση. Αυτά τα στρώματα περιέχουν ακλασικές νευρωνικές μονάδες οι οποίες λειτουργούν σαν ταξινομητές, έχοντας τα προηγούμενα στρώματα να λειτουργούν σαν εξαγωγείς χαρακτηριστικών. Κάθε νευρώνας του στρώματος N1 είναι πλήρως συνδεμένος σε όλα τα σημεία ενός και μόνον χάρτη χαρακτηριστικών του στρώματος S2. Τέλος, ο μοναδικός νευρώνας του στρώματος N2, που μας δίνει και το τελικό αποτέλεσμα της ταξινόμησης, είναι πλήρως συνδεμένος με όλους τους νευρώνες του στρώματος N1. Οι μονάδες αυτών των τελευταίων δύο στρωμάτων εκτελούν το ακλασικό εσωτερικό γινόμενο μεταξύ των διανυσμάτων εισόδου τους και των διανυσμάτων των βαρών τους, στο οποίο προστίθεται και ένα bias. Στην συνέχεια αυτό το σταθμισμένο άθροισμα διέρχεται από την σιγμοειδή συνάρτηση ενεργοποίησης του νευρώνα. Τα στρώματα N1 και N2 φέρουν αντίστοιχα 602 και 15 εκπαιδεύσιμα βάροη, όσες και οι συνδέσεις τους.

Σχετικά με την τακτική εκπαίδευσης, όλα τα βάροη προσαρμόστηκαν με μάθηση βάση-της-παραγώγου (gradient-based), μέσω του τροποποιημένου αλγορίθμου backpropagation (ενότητα 3.2). Κατά την διάρκεια της εκπαίδευσης, οι επιθυμητές έξοδοι του δικτύου τέθηκαν ίσες με  $-1, 0$  για μη πρόσωπα και  $+1, 0$  για πρόσωπα. Αυτή η παρατήρηση δίνει και την λογική της προσαρμογής της εισόδου στο πεδίο τιμών  $[-1, 1]$ .

Συνολικά, η προτεινόμενη τοπολογία έχει μόλις 951 εκπαιδεύσιμα βάροη, παρά τις 131.475 συνδέσεις που χρησιμοποιεί. Ο διαμοιρασμός των βαρών προσφέρει το πλεονέκτημα της μείωσης του αριθμού των ελευθέρων παραμέτρων, δίνοντας a priori μία ισχυρότερη δυνατότητα γενίκευσης, όπως είδαμε και στην ενότητα 3.4. Η τοπικότητα των πεδίων υποδοχής, ο διαμοιρασμός των βαρών και η υποδειγματοληψία παρέχουν πολλά προτερήματα για την επίλυση δύο σημαντικών προβλημάτων με μία κίνηση: το πρόβλημα της σθεναρότητας του δικτύου και το πρόβλημα της ικανοποιητικής γενίκευσης, το οποίο είναι αρκετά κρίσιμο δεδομένου της μη δυνατότητας στην πράξη της συλλογής σε ένα πεπερασμένου μεγέθους συνόλου εκπαίδευσης όλων των δυνατών διαφοροποιήσεων του προτύπου του προσώπου. Καθώς όλα τα βάροη σε όλα τα στρώματα έχουν υπολογιστεί μέσω μάθησης, το προτεινόμενο σύστημα μπορεί να θεωρηθεί ότι κατασκευάζει το δικό του σύνολο εξαγωγέων χαρακτηριστικών - σχετικών πάντα με το συγκεκριμένο πρόβλημα. Καμία δύσκολη και επίπονη επιλογή του μηχανικού δεν χρειάζεται καθώς αυτοί οι εξαγωγείς χαρακτηριστικών βασίζονται αυθεντικά σε μάθηση. Επιπρόσθετα, διαφορετικά από άλλες προσεγγίσεις που στηρίζονται σε ένα ξεχωριστό στάδιο εξαγωγής χαρακτηριστικών προηγούμενο της πραγματικής ταξινόμησης,



**Σχήμα 4.2**

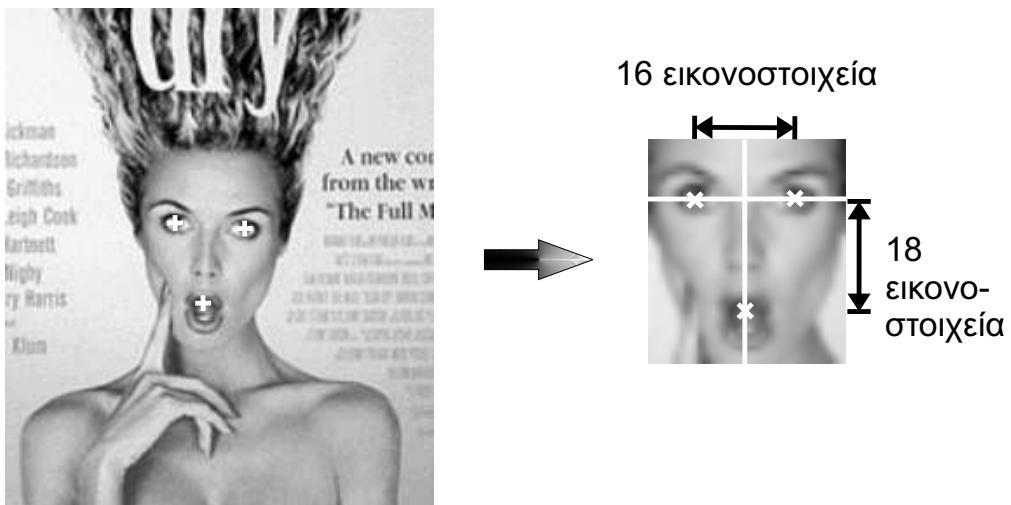
Κάποια από τα παραδείγματα προσώπων που συλλέχθηκαν. Η διαφοροποίηση σε πόζα, φωτισμό, έκφραση προσώπου, ποιότητα εικόνας κτλ είναι προφανής. Σημειώστε ότι κάποια πρόσωπα είναι μερικώς επικαλυπτόμενα από άλλα αντικείμενα όπως γυαλιά.

στην προτεινόμενη προσέγγιση αυτά τα δύο στάδια ενοποιούνται σε ένα και αδιαχώριστο σύστημα.

## 4.2 Συλλογή και Προετοιμασία των Δεδομένων

Τα παραδείγματα προσώπων που χρησιμοποιήθηκαν για την εκπαίδευση του δικτύου συλλέχθηκαν από διάφορες πηγές του Internet ή από εικόνες από εφημερίδες. Ο στόχος αυτής της συλλογής είναι η αποδοτική σύλληψη της διαφορετικότητας και του πλούτου των φυσικών δεδομένων, δίνοντάς μας ένα σύστημα εκπαίδευμένο να λειτουργεί σε μη ελεγχόμενα φυσικά περιβάλλοντα. Επιλέχθηκε η κατασκευή ενός νέου συνόλου εκπαίδευσης από την αρχή καθώς οι περισσότερες από τις προηγούμενες προσεγγίσεις χρησιμοποιούσαν δεδομένα εκπαίδευσης από βάσης εικόνων διαβατηρίων ή από σύνολα δεδομένων που προορίζονταν για αναγνώριση προσώπων (δείτε τις [55, 56] για κάποιες αναφορές για σύνολα δεδομένων). Γενικά, αυτές οι συλλογές περιέχουν έναν μεγάλο αριθμό από τετριμμένες πόζες, δεν παρέχουν διαφοροποίηση στις συνθήκες φωτισμού, έχουν εικόνες σε καλή ποιότητα, κτλ. Αυτό ίσως να μην είναι πολύ σημαντικό για άλλες μεθόδους που βασίζονται σε προεπεξεργασία της εισόδου, αλλά στην περίπτωσή μας είναι επιθυμητό ένα σύνολο εκπαίδευσης περισσότερο πλούσιο και γενικευμένο. Κάποια από τα 3.702 αυθεντικά παραδείγματα προσώπων που συλλέχτηκαν φαίνονται στο σχήμα 4.2.

Για την εξαγωγή των παραδειγμάτων προσώπου, αρχικά οι θέσεις των δύο ματιών και του στόματος για όλα τα πρόσωπα μαρκαρίστηκαν με το χέρι. Χρησιμοποιώντας αυτά τα σημεία, έγινε η εξαγωγή των περιοχών των προσώπων και στην συνέχεια η διαμόρφωσή τους στο μέγεθος  $32 \times 36$  εικονοστοιχείων, μετά από απο-περιστροφή και αλλαγή κλίμακας. Η κατάσταση απεικονίζεται στο σχήμα 4.3. Ο σκοπός αυτής της διαδικασίας ήταν να τοπο-



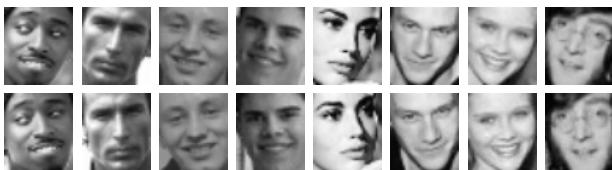
### Σχήμα 4.3

Η διαδικασία εξαγωγής του προσώπου. Στα αριστερά, οι θέσεις των δύο ματιών και του στόματος σημειώνονται με λευκούς σταυρούς. Μετά από απο-περιστροφή, εξαγωγή του τμήματος και προσαρμογή του μεγέθους έχουμε την εικόνα στα δεξιά.

Θετήσει τα δύο μάτια περίπου στις ίδιες θέσεις για όλα τα πρόσωπα, ενώ η απόσταση του στόματος από την γραμμή των ματιών χρησιμοποιήθηκε για να διατηρήσει προσεγγιστικά τον αυθεντικό λόγο διαστάσεων (aspect ratio) των προσώπων. Πιο συγκεκριμένα, το ευθύγραμμο τμήμα ανάμεσα στα δύο μάτια έπρεπε να γίνει περίπου 16 εικονοστοιχεία, ενώ η απόστασή του από το στόμα περίπου 18 εικονοστοιχεία. Σύμφωνα με τον πρώτο κανόνα, το πλάτος της εξαγόμενης περιοχής (εξάπλωση στην διεύθυνση της γραμμής των ματιών) μπορεί ένυκλα να οριστεί. Για τον υπολογισμό του ύψους της εξαγόμενης περιοχής (εξάπλωση στην κάθετη διεύθυνση), κάποιες στρογγυλοποιήσεις πρέπει να γίνουν καθώς η εξαγόμενη περιοχή πρέπει να έχει τελικά λόγο διαστάσεων ακριβώς  $32 \times 36$ . Διαφορετικά, τα πρόσωπα πρέπει να ταιριάζουν στο  $32 \times 36$  παράθυρο της εισόδου.

Οι περισσότερες από τις προσεγγίσεις βασισμένες στην εμφάνιση της βιβλιογραφίας [33, 38, 41, 48] χρησιμοποιούν παράθυρο εισόδου διάστασης  $20 \times 20$  εικονοστοιχεία, αναφέροντας ότι είναι το μικρότερο δυνατό παράθυρο που μπορεί κάποιος να χρησιμοποιήσει χωρίς να χάνει κρίσιμη πληροφορία. Συνήθως αυτό το παράθυρο περιέχει την αρκετά κεντρική περιοχή του προσώπου, αποκλείοντας τα όριά του και περιοχές φόντου. Σε αυτήν την εργασία, έχει επιλεχθεί περίπου η ίδια κλίμακα για την κεντρική περιοχή του προσώπου, η οποία όμως βρίσκεται στο κέντρο ενός  $32 \times 36$  παραθύρου στο οποίο τα όρια του προσώπου επίσης περικλείονται. Υπάρχουν δύο λόγοι κάτω από αυτήν την επιλογή. Πρώτα, το δίκτυο τροφοδοτείται με κάποια επιπλέον πληροφορία πάνω στο σχήμα του προσώπου η οποία

### Περιστροφή



### Εξομάλυνση



### Μείωση κοντράστ



### Σχήμα 4.4

Κάποια από τα μετασχηματισμένα παραδείγματα. Για τον μετασχηματισμό της εξομάλυνσης χρησιμοποιήθηκε ένας γκαουσσιανός πυρήνας με τυπική απόκλιση  $\sigma = 1$ . Στην μείωση του κοντράστ, η φωτεινότητα του εικονοστοιχείου  $I_p$  μεταβάλλεται σε  $I_p = 0.5I_m + 0.5I_p$ , όπου  $I_m$  είναι η μέση φωτεινότητα του παραθύρου. Σημειώστε ότι σε κάποια από τα πρόσωπα των δύο τελευταίων γραμμών έχει ήδη διεξαχθεί περιστροφή πρίν υποστούν τον αντίστοιχο μετασχηματισμό.

ενδέχεται να βοηθήσει στην μείωση του αριθμού εσφαλμένων ειδοποιήσεων που είναι πιθανό να εμφανιστούν όταν μόνο το κέντρο του προσώπου λαμβάνεται υπ' όψη. Κατά δεύτερο λόγο, κάποια συμπτώματα ορίων των συνελίξεων εξαλείφονται.

Ας σημειώσουμε επίσης ότι το χειρωνακτικό μαρκάρισμα των σημείων των προσώπων αναπόφευκτα εισάγει στοιχεία φυσικού θορύβου (λάθους/αστοχίας μαρκαρίσματος) στην διαδικασία εξαγωγής των προσώπων, τα οποία τελικά επηρεάζουν την ακριβή τοποθεσία των παραδειγμάτων εντός της εισόδου. Αυτό το λάθος έρχεται αρκετά φυσικά καθώς είναι πολύ δύσκολο για ένα ανθρώπινο χέρι να μαρκάρει με εξαιρετική ακρίβεια  $3.702 \times 3 = 11.106$  θέσεις, ειδικά όταν η αυθεντική κλίμακα των προσώπων είναι αρκετά μικρή. Όπως ειπώθηκε νωρίτερα, τα συνελικτικά δίκτυα είναι αρκετά σθεναρά στην μετάλλαξη της κλίμακας και της τοποθεσίας, οπότε αυτό το φυσικό λάθος -ελπίζοντας βέβαια να είναι διαδικασία λευκού θορύβου ώστε να μην εισαγάγει κάποια ισχυρή προκατάληψη- ενισχύει αυτήν την ικανότητα με το να παρέχει παραδείγματα όχι επακριβώς κανονικοποιημένα ή τοποθετημένα. Επιπλέον, η συλλογή των παραδειγμάτων εκπαίδευσης είναι μία αρκετά λιγότερο επίπονη διαδικασία όταν δεν απαιτείται η αυστηρή ευθυγράμμισή τους.

Καμία μορφή προεπεξεργασίας δεν εφαρμόστηκε στα εξαγόμενα πρόσωπα, όπως συνολική διόρθωση φωτισμού ή εξομάλυνση ιστογράμματος όπως στις [3, 33, 38, 41, 48]. Επι-

πλέον, για την δημιουργία περισσότερων παραδειγμάτων και για την βελτίωση της αντοχής σε περιστροφή και διαφοροποίησης στην ποιότητα της εισόδου, εφαρμόστηκε μία σειρά μετασχηματισμών στο παραπάνω αρχικό σύνολο παραδειγμάτων προσώπων. Αρχικά, κάποια από αυτά καθορίσθηκαν και στην συνέχεια όλα από αυτά περιστράφηκαν κατά  $\pm 10$  μοίρες. Σαν τελικό βήμα, σε κάποια από αυτά έγινε εξομάλυνση (smoothing) ενώ σε κάποια άλλα εφαρμόστηκε μείωση κοντράστ. Οι δύο τελευταίοι μετασχηματισμοί βοηθούν για την εκπαίδευση του συστήματος σε περιπτώσεις όπου η είσοδος φέρει ένα αρκετά αδύνατο σήμα (π.χ. υπερ-εξομαλυμένο ή με αδύνατες ακμές). Κάποια παραδείγματα αυτών των μετασχηματισμένων παραδειγμάτων φαίνονται στο σχήμα 4.4. Γενικά, υψηλά διαφοροποιημένα παραδείγματα είναι ζωτικά για την απόδοση του συστήματος γιατί ούτε διόρθωση φωτισμού ούτε εξομάλυνση ιστογράμματος δεν θα εφαρμόζεται στο σήμα εισόδου. Τελικά, το σύνολο εκπαίδευσης έφτασε τον αριθμό των 25.212 παραδειγμάτων προσώπων, συμπεριλαμβάνοντας περιπτώσεις μερικής επικάλυψης, ανισόρροπου φωτισμού, στροφής<sup>1</sup> ως  $\pm 60$  μοίρες και με εντάσεις φωτεινότητας να διαφέρουν από σκοτεινές σε φωτεινές.

### 4.3 Εκπαίδευση του Δικτύου

Η συλλογή ενός αντιπροσωπευτικού και βέλτιστου, κατά κάποια έννοια, συνόλου παραδειγμάτων μη προσώπων είναι πολύ πιο δύσκολη καθώς δεν μπορούμε να το γνωρίζουμε αυτό εκ των προτέρων. Είδαμε στην επισκόπηση της σχετικής βιβλιογραφίας ότι αυτό είναι ένα εγγενές πρόβλημα για όλους τους ταξινομητές βασισμένους σε παραδείγματα. Επίσης είδαμε και την στρατηγική bootstrapping η οποία φαίνεται να δίνει στην πράξη ικανοποιητικά αποτελέσματα. Πράγματι, καθώς πιστεύεται ότι τα συνελικτικά δίκτυα έχουν, εκ φύσεως, μία πολύ αναπτυγμένη ικανότητα γενίκευσης, ένα μικρό και προσεκτικά επιλεγμένο σύνολο από εικόνες φόντου μπορεί να βοηθήσει το δίκτυο να μην παράγει πολλές εσφαλμένες ειδοποιήσεις. Στην προτεινόμενη προσέγγιση, αυτή η στρατηγική υιοθετήθηκε με κάποιες βελτιώσεις σε επιμέρους σημεία.

Πριν την εφαρμογή του bootstrapping, ένα αρχικό σύνολο από 6.422 παραδείγματα μη προσώπων κατασκευάστηκε με επιλεκτική εξαγωγή τους από κάποιες εικόνες. Τα περισσότερα από αυτά τα παραδείγματα περιείχαν υπο-περιοχές προσώπων καθώς παρατηρήθηκε σε κάποια πρώιμα πειράματα ότι εικόνες τέτοιου είδους είναι μία σοβαρή πηγή εσφαλμένων ειδοποιήσεων. Ένα σύνολο εικόνων φόντου (μη περιέχοντας πρόσωπα) χρειαζόταν επίσης, από τις οποίες θα εξάγονταν οι εσφαλμένες ειδοποιήσεις και θα χρησιμοποιόντουσαν στην συνέχεια στην εκπαίδευση. Κάποιες από αυτές τις εικόνες φαίνονται στο σχήμα 4.5. Περιέ-

<sup>1</sup>Ένα πρόσωπο χαρακτηρίζεται στραμμένο όταν δεν κοιτάει ακριβώς προς την κάμερα. Οι ακραίες περιπτώσεις είναι οι ολικώς προφίλ θέσεις, δηλ. πρόσωπα στραμμένα κατά  $\pm 90$  μοίρες.

1. Δημιουργία ενός συνόλου επιβεβαίωσης, μη τεμνόμενο με το σύνολο εκπαίδευσης, από 400 πρόσωπα και 400 μη πρόσωπα εξαγόμενα και εξαιρούμενα από το αρχικό σύνολο εκπαίδευσης. Αυτό το σύνολο θα υποδείξει την καλύτερα αποδιδόμενη διαμόρφωση των βαρών κατά την διάρκεια των βημάτων 3 και 8.
2. Ανάθεση  $BooIter = 0$ ,  $ThrFa = 0, 8$ .
3. Εκπαίδευση του δικτύου για 60 εποχές μάθησης. Χρησιμοποίηση ενός ίσου αριθμού θετικών και αρνητικών παραδειγμάτων σε κάθε εποχή. Ανάθεση  $BooIter = BooIter + 1$ .
4. Συγκέντρωση όλων των εσφαλμένων ειδοποιήσεων από ένα σύνολο 692 εικόνων φόντου με έξοδο δικτύου μεγαλύτερη από  $ThrFa$ . Τερματισμός όταν 5000 έχουν συλλεχθεί ή όταν όλες οι εικόνες φόντου έχουν επεξεργαστεί.
5. Προσθήκη των νέων παραδειγμάτων στο σύνολο των μη προσώπων.
6. Αν  $ThrFa \geq 0, 2$  ανάθεση  $ThrFa = ThrFa - 0, 2$ .
7. Αν  $BooIter < 6$  μεταφορά στο βήμα 3.
8. Εκπαίδευση του δικτύου για 60 ακόμα εποχές και έξοδος.

**Πίνακας 4.2**

Το προτεινόμενο σχήμα bootstrapping.



(i)



(ii)

#### Σχήμα 4.5

(i) κάποιες από τις 692 εικόνες φόντου (μη περιέχοντας πρόσωπα) που χρησιμοποιήθηκαν κατά το bootstrapping. (ii) κάποιες από τις εσφαλμένες ειδοποιήσεις που προέκυψαν.

χουν μία μεγάλη ποικιλία από υφές και αντικείμενα όπως γράμματα, φυσικά αντικείμενα, σκηνές εντός και εκτός εστίας κτλ. Έχοντας σαν βάση αυτές τις εικόνες και το αρχικό σύνολο εκπαίδευσης, η προτεινόμενη διαδικασία bootstrapping εξελίσσεται σύμφωνα με τον πίνακα 4.2.

Στο βήμα 1, ένα σύνολο επιβεβαίωσης (validation set) κατασκευάζεται για την δοκιμή της ικανότητας γενίκευσης του δικτύου κατά την διάρκεια της εκπαίδευσης και υποδεικνύει την διαμόρφωση των βαρών (κατάσταση του δικτύου) που αποδίδει τα μέγιστα σε αυτό ως αυτήν που θα πρέπει να κρατήσουμε τελικά. Αυτό το σύνολο διατηρείται σταθερό καθ’ όλη την διάρκεια του bootstrapping, σε αντίθεση με το σύνολο εκπαίδευσης που συνεχώς αυξάνει. Περιέχει 400 τυχαίως επιλεγμένα παραδείγματα από όλο το σύνολο προσώπων και 400 επίσης τυχαίως επιλεγμένα παραδείγματα από το (αρχικό) σύνολο των 6.422 αρνητικών παραδειγμάτων.

Στο βήμα 3, ο αλγόριθμος backpropagation χρησιμοποιήθηκε, τροποποιημένος όπως στην ενότητα 3.2 και με την προσθήκη ενός όρου ορμής για τους νευρώνες των στρωμάτων N1 και N2. Η στοχαστική μάθηση (stochastic learning) προτιμήθηκε αντί της μαζικής (batch learning). Σε κάθε εποχή μάθησης, ένας ίσος αριθμός παραδειγμάτων και από τις δύο κλάσεις παρουσιάζονται στο δίκτυο, μη δίνοντας προκατάληψη προς κάποια από τις δύο. Σε αυτό το περιεχόμενο, μια εποχή μάθησης είναι η παρουσίαση  $N$  παραδειγμάτων προσώπων (από συνολικά  $N_p$ ) και  $N$  παραδειγμάτων μη προσώπων (από συνολικά  $N_n$ ) στο δίκτυο, όπου  $N$  είναι ο μικρότερος εκ των  $\{N_p, N_n\}$ .

Η παραγωγή των καινούργιων παραδειγμάτων που θα προστεθούν στο σύνολο των μη προσώπων διεξάγεται στο βήμα 4. Οι παραγόμενες εσφαλμένες ειδοποιήσεις αυτού του βήματος πιέζουν το δίκτυο, στην επόμενη επανάληψη του bootstrapping, να αναθεωρήσει την τρέχουσα μοντελοποίησή του για την κλάση των προσώπων και να βελτιώσει το όριο διαχωρισμού του μεταξύ των προσώπων και μη προσώπων. Σε κάθε επανάληψη, επιλέγονται εσφαλμένες ειδοποιήσεις μέσα στα όρια της κλάσης προσώπων (προκαλώντας δηλαδή εξόδους δικτύου μεγαλύτερους από *ThrFa*). Στην συνέχεια το κατώφλι *ThrFa* μειώνεται σταδιακά μέχρι να φτάσει στο 0. Η διαδικασία τερματίστηκε στην έκτη επανάληψη, όπου παρατηρήθηκε η σύγκλιση της, όταν δηλ. το δίκτυο έπαψε να δίνει πλέον έναν σημαντικό αριθμό εσφαλμένων ειδοποιήσεων. Αυτή η διαδικασία βοηθάει στην διόρθωση προβλημάτων που ανακύπτουν στον αυθεντικό αλγόριθμο των Sung και Poggio, όπου οι εσφαλμένες ειδοποιήσεις επιλέγονταν ανεξάρτητα της ισχύος της εξόδου του δικτύου.

Τα σημεία-κλειδιά του προτεινόμενου σχήματος bootstrapping μπορούν να ανακεφαλαθούν ως εξής:

- Ένας ίσος αριθμός από παραδείγματα και των δύο κλάσεων παρουσιάζεται στο δίκτυο. Γενικά, αυτό είναι κοινή επιλογή στην εκπαίδευση νευρωνικών δικτύων όταν έχουμε ήδη δειγματοληπτήσει τον χώρο εισόδου σε ένα δεδομένο σύνολο εκπαίδευσης που περιέχει a priori ισάριθμα παραδείγματα απ' όλες τις κλάσεις. Με αυτόν τον τρόπο, όλες οι κλάσεις τίθενται a priori ισοπίθανες, μη δίνοντας προκατάληψη προς κάποια συγκεκριμένη κατά την διάρκεια της εκπαίδευσης. Στην περίπτωση μας, το σύνολο των μη προσώπων δυναμικά αυξάνεται έτσι ώστε αυτή η συνθήκη πρέπει να τεθεί ζητά.
- Επιλέγονται τα φέροντα περισσότερη πληροφορία παραδείγματα. Αυτές οι εσφαλμένες ειδοποιήσεις αναμένεται να βρίσκονται βαθιά μέσα στο όριο διαχωρισμού των προσώπων, όπως γίνεται αυτό αντιληπτό από το δίκτυο. Έτσι, είναι και οι πρώτοι υποψήφιοι για να βοηθήσουν το δίκτυο να αναθεωρήσει γρήγορα το όριό του προς την σωστή κατεύθυνση. Καθώς το δίκτυο γενικεύει απ' αυτές τις εσφαλμένες ειδοποιήσεις, το κατώφλι *ThrFa* μπορεί να μειωθεί με ασφάλεια για να φτάσει στο επιθυμητό επίπεδο του 0,0.
- Αποτρέπεται κάποιος πλεονασμός στο σύνολο των αρνητικών παραδειγμάτων. Θα δούμε στην ενότητα 5.1 ότι η κατανομή των εξόδων του δικτύου γύρω από ένα στόχο (πρόσωπο ή εσφαλμένη ειδοποίηση) σχηματίζει ένα καμπανοειδές σχήμα. Καθώς η συνελικτική τοπολογία παρουσιάζει κάποια σταθερότητα στην μετατόπιση της εισόδου, κάποιες εσφαλμένες ειδοποιήσεις γύρω απ' αυτήν που προκαλεί την μεγαλύτερη έξοδο μπορούν κάλλιστα να γενικευθούν μέσω της τελευταίας από το δίκτυο. Πιο γενικά, αν τα περισσότερο διδακτικά παραδείγματα προστίθενται στην αρχή της διαδικασίας,

το λιγότερο διδακτικά και πλεονάζοντα μπορούν αργότερα να απορριφθούν, έχοντας ήδη γενικευθεί από το δίκτυο. Με αυτόν τον τρόπο, μόνο μία κρίσιμη μάζα από αυστηρέστα παραδείγματα χρησιμοποιούνται στο σύνολο εκπαίδευσης.

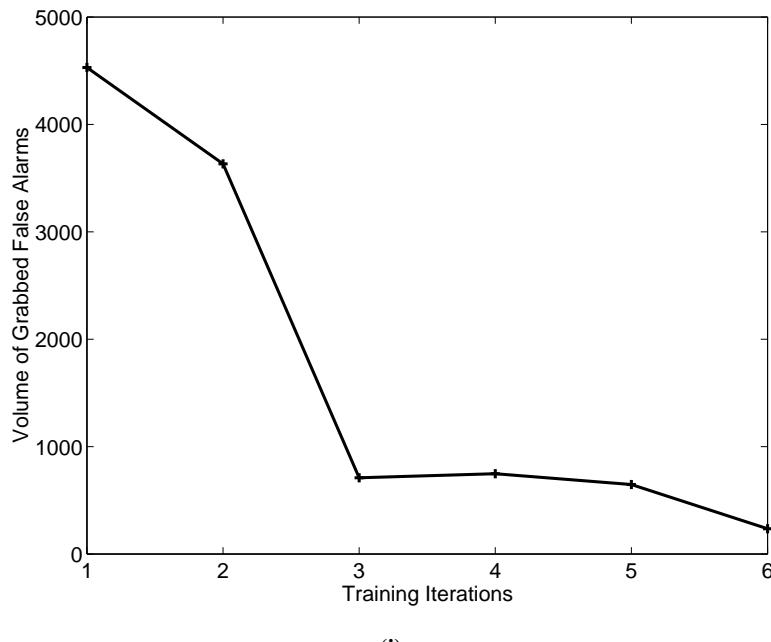
Τελικά, η ελεγχόμενη διαδικασία bootstrapping προσέθεσε 19.065 παραδείγματα μη προσώπων στο σύνολο εκπαίδευσης. Κάποια απ' αυτά δίνονται στο σχήμα 4.5(ii) με επτά παραδείγματα ανά γραμμή από κάθε μία των έξι επαναλήψεων.

## 4.4 Αποτελέσματα της Εκπαίδευσης

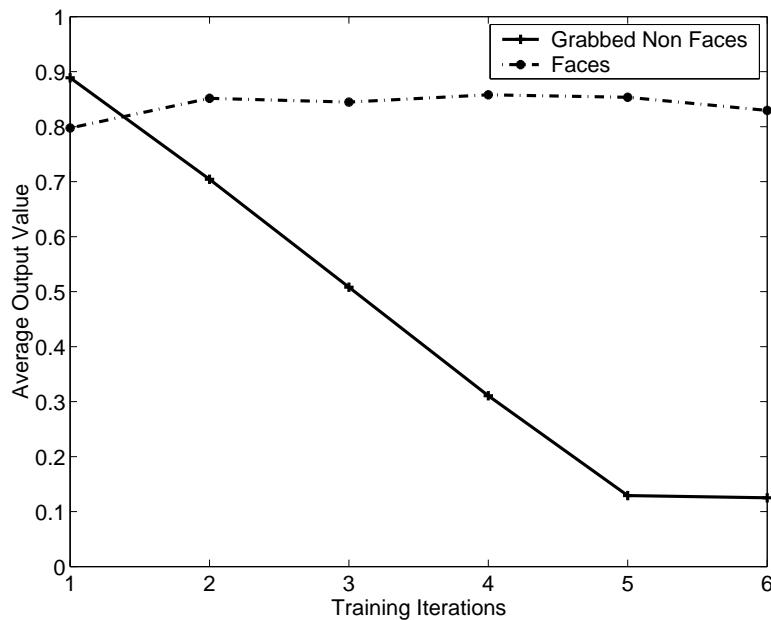
### 4.4.1 Εξέλιξη της Εκπαίδευσης

Η απόδοση της προτεινόμενης διαδικασίας bootstrapping παρουσιάζεται στο σχήμα 4.6. Συγκεκριμένα, ενδιαφερόμαστε περισσότερο στο κατά πόσο το δίκτυο ενισχύθηκε από άποψη απόρριψης των εσφαλμένων ειδοποιήσων, ο οποίος είναι και ο σκοπός του bootstrapping. Το πρώτο γράφημα (σχήμα 4.6(i)) δίνει των όγκων των συλλεγμένων εσφαλμένων ειδοποιήσεων σε σχέση με τις επαναλήψεις εκπαίδευσης-bootstrapping. Ο όγκος ορίζεται ως το άθροισμα όλων των (θετικών) απαντήσεων του δικτύου που αντιστοιχούν στις εσφαλμένες ειδοποιήσεις που συλλέχθηκαν κατά την διάρκεια της σάρωσης των εικόνων φόντου. Αυτή η μετρική εκτιμά την *ισχύ* αυτών των εσφαλμένων ειδοποιήσεων, έναντι π.χ. του αριθμού τους, ως μία πιο ενδιαφέρουσα μέθοδος αποτίμησης της εκπαίδευσης του δικτύου. Βλέπουμε σ' αυτό το σχήμα ότι η πρώτη επανάληψη παρήγαγε πολύ ισχυρές εσφαλμένες ειδοποιήσεις, το οποίο όμως ήταν αναμενόμενο γιατί το δίκτυο σάρωσε για πρώτη φορά τις εικόνες φόντου. Ο συλλεγμένος όγκος είναι σχεδόν στα δριά του, στο 5.000 περίπου, το οποίο είναι η έξοδος του δικτύου περίπου στο +1,0 επί το μέγιστο επιτρεπόμενο αριθμό εσφαλμένων ειδοποιήσεων. Καθώς το bootstrapping προχωράει, βλέπουμε στο σχήμα ότι το δίκτυο έμαθε γρήγορα να μην δίνει ισχυρές εσφαλμένες ειδοποιήσεις από την απότομη πτώση του όγκου στην επανάληψη 3. Από εκεί και έπειτα, η συμπεριφορά μένει περίπου σταθερή, το οποίο μας πληροφορεί ότι η διαδικασία bootstrapping μπορεί να τελειώσει με ασφάλεια.

Το σχήμα 4.6(ii) εικονογραφεί την συμπεριφορά του δικτύου πάνω στα θετικά παραδείγματα εκπαίδευσης. Εδώ βλέπουμε ότι η μέση ανταπόκριση του δικτύου σε αυτά παραμένει περίπου σταθερή γύρω από το 0,85, μετά την πρώτη επανάληψη. Έτσι, είναι προφανές ότι το δίκτυο έμαθε πολύ γρήγορα να ανταποκρίνεται στα θετικά παραδείγματα με υψηλές θετικές τιμές, από τα πρώτα κιόλας βήματα της διαδικασίας εκπαίδευσης. Επιπρόσθετα, αυτές οι ανταποκρίσεις διατηρήθηκαν υψηλές μέχρι τέλους, ενώ το σύνολο των αρνητικών παραδειγμάτων αυξάνονταν συνέχεια. Στο ίδιο σχήμα, η σταθερά υψηλή μέση ανταπόκριση στα πρόσωπα αντιπαραβάλλεται με την βαθμαία μείωση της μέσης ανταπόκρισης στις εσφαλμένες ειδοποιήσεις, συνοψίζοντας τα αποτελέσματα της εκπαίδευσης. Είναι αξιοσημείωτο το



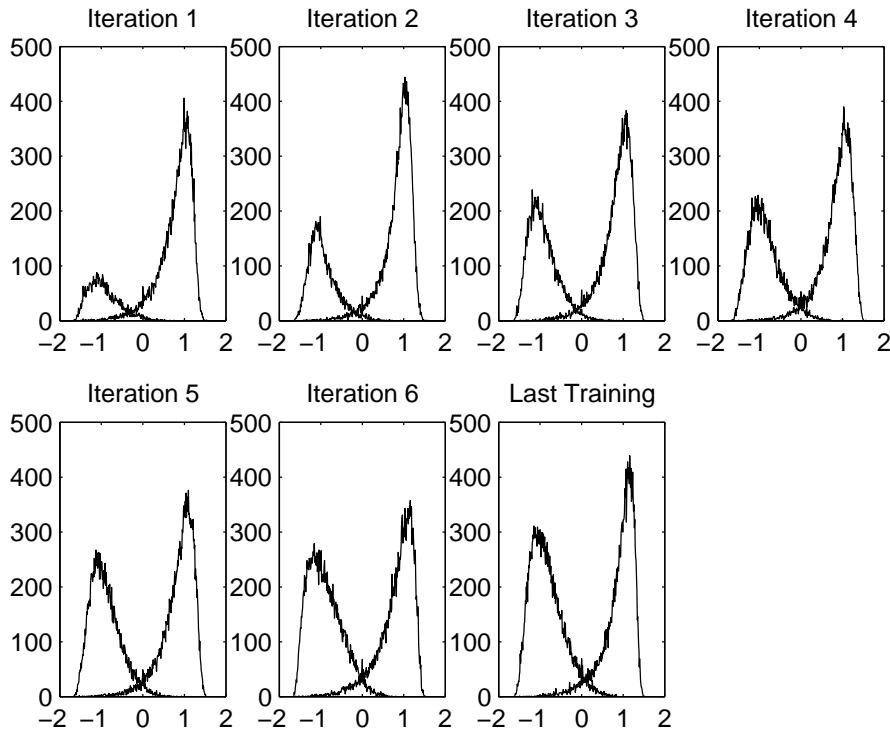
(i)



(ii)

#### Σχήμα 4.6

Η εξέλιξη της προτεινόμενης διαδικασίας εκπαίδευσης-bootstrapping. (i) ο όγκος των συλλεγμένων εσφαλμένων ειδοποιήσων ανά επανάληψη του bootstrapping. (ii) η μέση τιμή των συλλεγμένων εσφαλμένων ειδοποιήσων και η μέση τιμή των απαντήσεων του δικτύου για τα πρόσωπα του συνόλου εκπαίδευσης ανά επανάληψη του bootstrapping.

**Σχήμα 4.7**

Η εξέλιξη του διαχωρισμού των κλάσεων.

γεγονός ότι, η μέση ανταπόκριση στις εσφαλμένες ειδοποιήσεις ακολουθεί αυστηρά την γραμμική μείωση των τιμών του *ThrFa*.

Ένας ευθύς τρόπος για να αντιληφθούμε το κατά πόσο καλά διαχωρίζει το δίκτυο τις δύο κλάσεις είναι να υπολογίσουμε το ιστόγραμμα των ανταποκρίσεων του δικτύου πάνω σε ολόκληρο το σύνολο εκπαίδευσης. Η εξέλιξη αυτού του ιστογράμματος φαίνεται στο σχήμα 4.7<sup>2</sup>. Βλέπουμε καθαρά σε όλα τα ιστογράμματα δύο κατανομές που αντιστοιχούν στην κλάση των προσώπων και στην κλάση των μη προσώπων, με κορυφές πάνω στις αντίστοιχες επιθυμητές εξόδους +1 για τα πρόσωπα και -1 για τα μη πρόσωπα. Αν και ο διαχωρισμός δεν είναι απόλυτος (κάτι το οποίο θα σήμαινε και μηδενικό λάθος πάνω στο σύνολο εκπαίδευσης), είναι αρκετά ικανοποιητικός καθώς ένα πολύ μικρό ποσοστό των κατανομών αλληλοεπικαλύπτεται. Η κατανομή των προσώπων παραμένει περίπου σταθερή, το οποίο συμφωνεί με τα δεδομένα του σχήματος 4.6(ii). Είναι αρκετά ενδιαφέρον να σημειωθεί ότι, προχωρώντας από την κορυφή της (+1) προς τα αριστερά, η κατανομή φθίνει εκθετικά,

<sup>2</sup>Τα ιστογράμματα πάνω στα αρνητικά παραδείγματα δεν αναφέρονται στις εσφαλμένες ειδοποιήσεις που συλλέχθηκαν στην αντίστοιχη επανάληψη, αλλά στα αρνητικά παραδείγματα όντας ήδη εντός του συνόλου εκπαίδευσης από τις προηγούμενες επαναλήψεις.

Τοπολογία	Σύνολο εκπαίδευσης		Σύνολο Επιβεβαίωσης		Εσφαλμένες Ειδοποιήσεις
	Λάθη	MSE	Λάθη	MSE	
N1	2607	0,17	60	0,26	37287
<b>N4</b>	<b>1921</b>	<b>0,17</b>	<b>32</b>	<b>0,18</b>	<b>19065</b>
N5	1540	0,15	24	0,17	26133
Γραμμική	11558	0,65	185	0,66	-

#### Πίνακας 4.3

Ρυθμός λαθών για διάφορες συνελικτικές τοπολογίες και για τον γραμμικό ταξινομητή. Η προτεινόμενη τοπολογία (σχήμα 4.1) είναι η N4. Για κάθε τοπολογία δίνεται ο αριθμός των λαθών (αριθμός παραδειγμάτων που προκαλούν έξοδο δικτύου διαφορετικά προσημασμένη με την επιθυμητή έξοδο), το MSE (Mean Square Error - Μέσο Τετραγωνικό Σφάλμα) και ο αριθμός εσφαλμένων ειδοποιήσεων που μετρήθηκε κατά το bootstrapping.

αφήνοντας την κεντρική τής μάζα γύρω από την κορυφή. Αυτό υποδηλώνει ότι μπορούμε να έχουμε σημεία διαχωρισμού στα οποία διατηρούμε τις περισσότερες από τις ανταποκρίσεις σε πρόσωπα και λίγες, αν όχι καθόλου, ανταποκρίσεις σε μη πρόσωπα. Όσον αφορά την κατανομή των μη προσώπων, φυσιολογικά διογκώνεται καθώς περισσότερα παραδείγματα μη προσώπων προστίθενται στο σύνολο εκπαίδευσης σε κάθε επανάληψη. Δυστυχώς, δεν παρατηρείται εδώ μια εκθετική πτώση, όπως στην περίπτωση των προσώπων. Ειδικά στα τελευταία ιστόγραμματα, παρατηρούμε μια περισσότερο γραμμική πτώση από την κορυφή προς τα δεξιά. Άλλα συνυπολογίζοντας ότι συνεχώς προσθέτουμε εσφαλμένες ειδοποιήσεις και άρα προσκείμενες στα όρια της κατανομής, αυτή η γραμμική πτώση είναι ένα μάλλον ενθαρρυντικό γεγονός, υποδηλώνοντας ότι το δίκτυο έμαθε με επιτυχία αυτές τις εσφαλμένες ειδοποιήσεις. Επιπρόσθετα, και πιο σημαντικά, σχεδόν όλη η κατανομή βρίσκεται αυστηρά κάτω του μηδενός.

#### 4.4.2 Σύγκριση με Άλλες Τοπολογίες

Στον πίνακα 4.3 βλέπουμε τις μετρήσεις σφαλμάτων της προτεινόμενης τοπολογίας (αναφερόμενης ως 'N4') στο σύνολο εκπαίδευσης και στο σύνολο επιβεβαίωσης, μαζί με τις αντίστοιχες άλλων τοπολογιών, συνελικτικών ή όχι. Η εκπαίδευση της N4 συντέλεσε σε ένα ποσοστό περίπου 4% λάθος ταξινομήσεων και στα δύο σύνολα. Η τοπολογία N1 είναι ένα συνελικτικό δίκτυο με ένα συνελικτικό/υποδειγματοληπτικό στρώμα, ενώ εκτός αυτού είναι σχεδιασμένη με την ίδια φιλοσοφία όπως και η N4. Μας θυμίζει επίσης την τοπολογία του συνελικτικού δικτύου που χρησιμοποιήθηκε από τους Vaillant *et al.* [51]. Αυτή η τοπολογία έχει 1.009 βάροη, περισσότερα από αυτά της N4, αν και είναι πιο 'έλαφριά' αφού φέρει λιγότερες συνδέσεις. Η τοπολογία N5 έχει και αυτή κατασκευαστεί με την ίδιο

φιλοσοφία όπως και η N4, αλλά με την διαφορά ότι έχει προστεθεί ένας ακόμα χάρτης χαρακτηριστικών στα στρώματα C1 και S1, το οποίο, με ένα παρόμοιο σχήμα συνδέσεων με αυτό του πίνακα 4.1, θα δώσει 6 περισσότερους χάρτες χαρακτηριστικών στα στρώματα C2/S2. Αυτή η τοπολογία φέρει 1.346 βάρη. Οι δύο παραπάνω τοπολογίες είναι μικρές διαφοροποιήσεις της προτεινόμενης, στην κατεύθυνση της προσθήκης “δομής” στο δίκτυο (N5) ή, αντίθετα, στην αφαίρεσή της (N1). Τέλος, δίνονται αποτελέσματα και για τον γραμμικό ταξινομητή (ένα μονο-στρωματικό perceptron με σιγμοειδή συνάρτηση ενεργοποίησης και  $32 \times 36 + 1 = 1.153$  βάρη) πάνω στο τελικό σύνολο εκπαίδευσης που δημιουργήθηκε από το bootstrapping της N4. Αυτές οι μετρήσεις μας υποδεικνύουν πόσο δύσκολο είναι να επιλυθεί γραμμικά ο διαχωρισμός των παραδειγμάτων, προσώπων και εσφαλμένων ειδοποιήσεων του bootstrapping.

Αρχικά, βλέπουμε ότι οι τοπολογίες N4 και N5 έχουν την καλύτερη απόδοση, με ελάχιστη διαφορά μεταξύ τους. Η N5 έχει ελαφρώς καλύτερη απόδοση στα δύο σύνολα, αλλά έδωσε περισσότερες εσφαλμένες ειδοποιήσεις κατά το bootstrapping, καθώς χρειάστηκαν δύο ακόμα επαναλήψεις μέχρι να παρατηρηθεί σύγκλιση. Καθώς η τοπολογία N4 έχει λιγότερα βάρη και συνδέσεις, άρα είναι και λιγότερο ακριβή στους υπολογισμούς, μπορούμε τελικά να επιλέξουμε αυτή ως πιο οικονομική και απλή λύση. Επιπλέον, θα δούμε στην ενότητα 6.1.5 ότι η N4 αποδίδει λίγο καλύτερα στα σύνολα δοκιμής. Η τοπολογία με ένα συνελικτικό/υποδειγματοληπτικό στρώμα προκάλεσε πολύ περισσότερες εσφαλμένες ειδοποιήσεις απ' ότι οι δύο παραπάνω. Μπορούμε να πούμε ότι απέτυχε τελικά να επεξεργαστεί σύγκλιση κατά το bootstrapping, με την έννοια ότι δεν κατάφερε τελικά να επεξεργαστεί όλες τις εικόνες φόντου με λιγότερες εσφαλμένες ειδοποιήσεις του ανώτατου ορίου (5.000) σε όλες τις επαναλήψεις. Αποδίδει επίσης φτωχά και στο σύνολο επιβεβαίωσης, έχοντας διπλάσιο περίπου αριθμό σφαλμάτων σε σύγκριση με τις N4 και N5. Τέλος, μπορούμε αρκετά εύκολα να αντιληφθούμε την πολύ φτωχή απόδοση του γραμμικού ταξινομητή. Δίνει περίπου 23% ποσοστό λάθους στο σύνολο εκπαίδευσης και 23% στο σύνολο επιβεβαίωσης, σε αντιπαράθεση με τα αντίστοιχα 4% και 4% της N4. Αυτό επιδεικνύει πόσο μακριά είναι το πρόβλημά μας για να επιλυθεί γραμμικά.

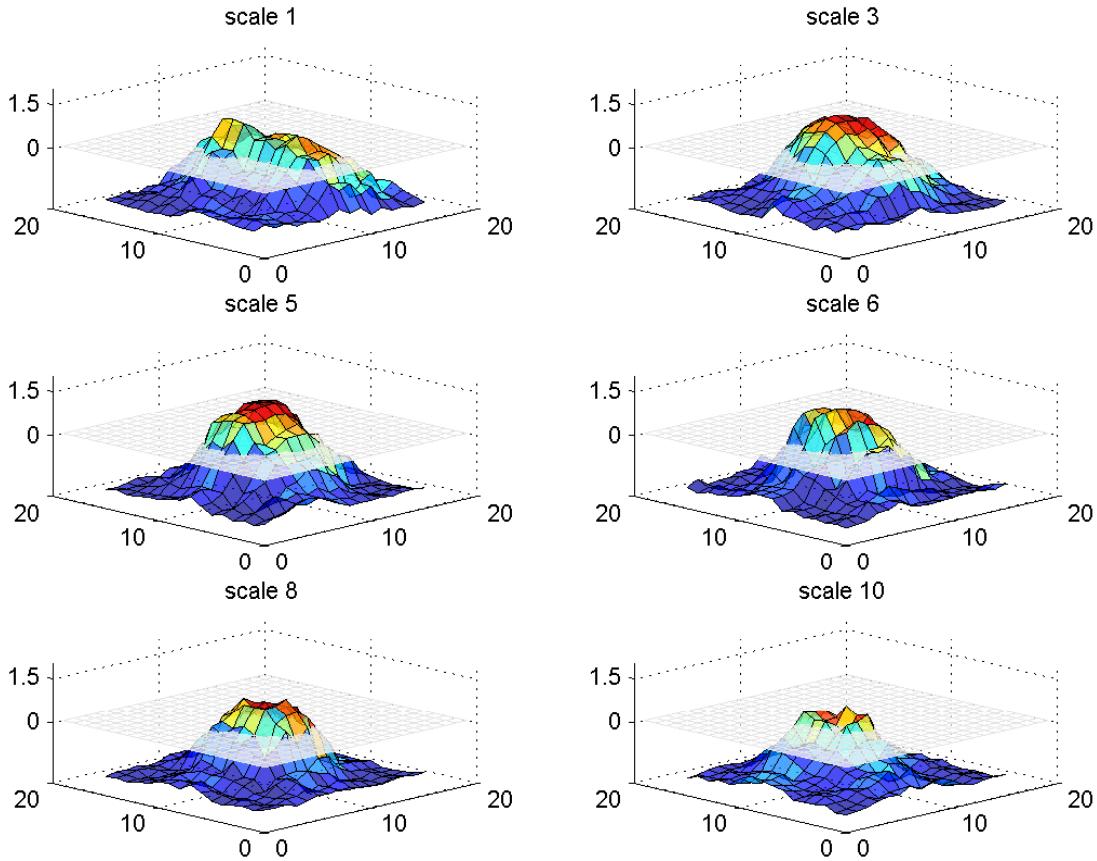
## ΚΕΦΑΛΑΙΟ 5

# Σάρωση της Εικόνας και Εντοπισμός του Προσώπου

Σε αυτό το κεφάλαιο θα δούμε πως το εκπαιδευμένο συνελικτικό δίκτυο χρησιμοποιείται για να σαρώσουμε την εικόνα εισόδου και να εξάγουμε όλες τις περιοχές με πρόσωπα που μπορεί να υπάρχουν σ' αυτήν. Αυτή η διαδικασία διαχωρίζεται σε δύο επιμέρους βήματα: στην σάρωση της εικόνας για τον εντοπισμό εν δυνάμει στόχων και στην τοπική αναζήτηση γύρω από αυτούς τους στόχους, όπου και τελικά αυτοί χαρακτηρίζονται ως πρόσωπα ή όχι. Θα δούμε επίσης πως μπορεί να αποφευχθεί πλεονασμός υπολογισμού κατά την σάρωση της εικόνας με το συνελικτικό δίκτυο που μας δίνει τελικά ένα πολύ γρήγορο και απλό κανάλι φίλτρων, ισοδύναμο στο αποτέλεσμα. Τέλος, δίνεται μία λεπτομερής ανάλυση του υπολογιστικού κόστους της προτεινόμενης μεθόδου και σε σύγκριση με άλλες νευρωνικές μεθόδους.

### 5.1 Συμπεριφορά του Δικτύου Γύρω από Ένα Στόχο

Καθώς ο στόχος της διαδικασίας εντοπισμού είναι ένας αποτελεσματικός τρόπος ανίχνευσης προσώπων σε διαφορετικές κλίμακες και θέσεις, είναι χρήσιμο να εξετάσουμε πρώτα την συμπεριφορά του δικτύου γύρω από ένα στόχο. Περιμένουμε ότι οι απαντήσεις του δικτύου γύρω του θα σχηματίζουν καμπανοειδής λοβούς με την κορυφή τους στο σημείο που πραγματικά ο στόχος βρίσκεται, σε θέση και κλίμακα. Πράγματι, αυτό είναι προφανές στο σχήμα 5.1. Μία εικόνα που περιέχει ένα  $32 \times 36$  πρόσωπο μαζί με κάποιες περιοχές φόρντου έχει σαρωθεί από το δίκτυο σε κάθε σημείο της και σε διάφορες κλίμακες γύρω από την πραγματική. Πιο συγκεκριμένα, η εικόνα διαμορφώθηκε 10 φορές (κλίμακες 1 ως 10), αρχίζοντας από 0,5 της πραγματικής κλίμακας και καταλήγοντας σε 1,5 αυτής και με σταθερό βήμα μετάβασης. Οι συλλεγμένες απαντήσεις απεικονίζονται στο σχήμα 5.1, για τις κλίμακες 1,



### Σχήμα 5.1

Απαντήσεις του δικτύου σε κλίμακα-θέση γύρω από έναν στόχο. Σε κάθε γράφημα, απεικονίζεται η απάντηση του δικτύου (άξονας z) έχοντας την θέση (άξονες x και y) να μεταβάλλεται. Οι απαντήσεις περισυλλέγησαν σε ένα  $16 \times 16$  πλέγμα με το πρόσωπο περίπου στο κέντρο.

3, 5, 6, 8 και 10. Σε κάθε κλίμακα, βλέπουμε ένα καμπανοειδή λοβό με το κέντρο του γύρω από την ίδια (απόλυτη) τοποθεσία, αλλά με μεταβλητή ισχή. Στις μεσαίες κλίμακες 5 και 6, η ισχύς των απαντήσεων είναι στο μέγιστό τους και το καμπανοειδή σχήμα είναι πιο απλωμένο στον χώρο. Είναι εδώ που το πρόσωπο έχει ύψος περίπου 36 εικονοστοιχεία έτσι ώστε να ταιριάζει ακριβώς στο  $32 \times 36$  παραθύρο του δικτύου. Ο λοβός των θετικών απαντήσεων είναι εκτάσεως περίπου  $5 \times 5$  εικονοστοιχείων. Έξω απ' αυτόν, οι απαντήσεις είναι καθαρά και σταθερά αρνητικές. Καθώς απομακρυνόμαστε από τις μεσαίες κλίμακες, η ισχύς και το άνοιγμα των λοβών μειώνεται σταδιακά, αλλά χωρίς να εξαφανίζεται εντελώς.

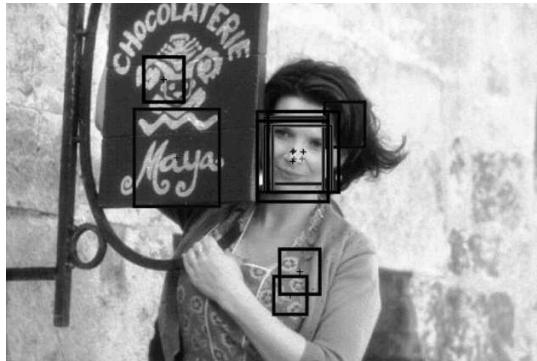
Έχοντας αυτές τις παρατηρήσεις, η στρατηγική εντοπισμού μπορεί να διαιρεθεί σε δύο ξεχωριστά βήματα:

- Σε ένα βήμα πρόχειρου εντοπισμού, όπου κάποια τμήματα αυτών των λοβών θα πρέπει να ανιχνευτούν. Καθώς έχουν κάποια έκταση σε κλίμακα και θέση, μπορούν να ανιχνευτούν χωρίς εξονυχιστική σάρωση σε όλες τις δυνατές κλίμακες και θέσεις.
- Σε ένα βήμα ακριβή εντοπισμού, όπου στόχος είναι η μέτρηση του όγκου της δραστηριότητας, δηλ. του αθροίσματος όλων των θετικών απαντήσεων του δικτύου γύρω από τον στόχο. Σε αυτό το βήμα, γίνεται μία λεπτομερή σάρωση η οποία θα καλύψει, κατ' ευχήν, όλη την έκταση του προσώπου σε κλίμακα και θέση και θα δώσει μία καλή εκτίμηση του πραγματικού όγκου της δραστηριότητας. Αυτή η μετρική είναι αρκετά σημαντική και θα είναι αρκετή για να δώσει την τελική απάντηση για το αν ο στόχος είναι πρόσωπο ή όχι. Επιπρόσθετα, η καθολική κορυφή των καμπανοειδών λοβών θα μας αποκαλύψει και την ακριβή τοποθεσία και κλίμακα του προσώπου.

## 5.2 Εντοπισμός του Προσώπου

Σαν ένα πρώτο βήμα της διαδικασίας εντοπισμού, η εικόνα εισόδου υποδειγματοληπτείται επανειλημμένα κατά ένα παράγοντας 1,2 δίνοντας μία πυραμίδα κλιμάκων όπως αυτήν της εικόνας 2.1. Η αρχική και η τελική κλίμακα καθορίζουν το εύρος κλιμάκων στις οποίες πρέπει να βρίσκονται τα πρόσωπα για να ανιχνευθούν. Μια τυπική επιλογή είναι η εκκίνηση από την πραγματική κλίμακα της εικόνας (δηλ. ανιχνεύονται πρόσωπα ύψους τουλάχιστον 36 εικονοστοιχείων) και ο τερματισμός όταν η εικόνα έχει μειωθεί (υποδειγματοληφθεί) σε περίπου  $32 \times 36$  (δηλ. ανιχνεύονται πρόσωπα που καταλαμβάνουν το πολύ όλη την εικόνα).

Δεδομένου των διαφόρων κλιμάκων για την αναζήτηση, χρησιμοποιούμε το συνελικτικό δίκτυο για την σάρωση όλων αυτών για εν δυνάμει περιοχές προσώπων. Υπάρχουν δύο δυνατότητες: να σαρώσουμε εξονυχιστικά σε κάθε θέση ή να σαρώσουμε σποραδικά, εφαρμόζοντας το δίκτυο με ένα σταθερό βήμα μετατόπισης. Στην πρώτη περίπτωση, είναι περισσότερο πιθανό να εντοπιστεί ένα πρόσωπο ειδικά όταν αυτό βρίσκεται σε κάποιες ακραίες συνθήκες (π.χ. περιστραμμένο ή πολύ θρυψβώδες). Τα μειονεκτήματα της λεπτομερούς σάρωσης είναι το αυξημένο υπολογιστικό κόστος και ο κίνδυνος να αυξηθεί ο ρυθμός των εσφαλμένων ειδοποιήσεων, καθώς είναι πολύ πιθανό ότι περισσότεροι στόχοι θα δοθούν για επιπλέον ανάλυση στο δεύτερο βήμα της διαδικασίας. Από εδώ και έπειτα σε αυτό το κείμενο, αυτή η μέθοδος θα αναφέρεται ως λεπτή σάρωση. Η άλλη δυνατότητα είναι να σαρώσουμε την εικόνα εφαρμόζοντας το δίκτυο όχι σε όλες τις θέσεις αλλά μ' ένα σταθερό βήμα και στις δύο κατευθύνσεις (x και y). Όπως είδαμε, οι απαντήσεις του δικτύου δεν φθιστούνται εντελώς καθώς απομακρυνόμαστε κάποια εικονοστοιχεία από την (πραγματική) θέση του στόχου. Έτσι, εν δυνάμει περιοχές προσώπων μπορούν ακόμα να ανιχνευθούν εφαρμόζοντας το δίκτυο μ' αυτόν τον τρόπο. Αυτή η μέθοδος θα αναφέρεται από εδώ και



(i)



(ii)



(iii)



(iv)

### Σχήμα 5.2

Τα βήματα της διαδικασίας εντοπισμού. (i) ανίχνευση στόχου. (ii) ομαδοποίηση. (iii) λεπτομερής σάρωση γύρω του στόχου. (iv) απόρριψη εσφαλμένων στόχων.

έπειτα ως πρόχειρη σάρωση. Μία καλή επιλογή για το βήμα είναι 4 εικονοστοιχεία σε κάθε κατεύθυνση. Τέλος, τα αποτελέσματα αυτού του βήματος απεικονίζονται στο σχήμα 5.2(i), όπου έχει γίνει επεξεργασία με λεπτή σάρωση. Όλες οι απαντήσεις του δικτύου άνω του μηδενός συλλέχθηκαν και προβλήθηκαν πάνω στην αυθεντική εικόνα, ανάλογα την κλίμακα στην οποία εντοπίστηκαν.

Στη συνέχεια, τα υποψήφια πρόσωπα ομαδοποιούνται επαναληπτικά σύμφωνα με την εγγύτητά τους σε θέση και κλίμακα. Από κάθε ομάδα συντίθεται ένα αντιπροσωπευτικό πρόσωπο. Το κέντρο και το μέγεθος του υπολογίζονται ως οι μέσοι όροι των κέντρων και των κλιμάκων των προσώπων της ομάδας, σταθμισμένοι σύμφωνα με τις αντίστοιχες απαντήσεις του δικτύου. Η κατάσταση απεικονίζεται στο σχήμα 5.2(ii), όπου έχουμε τελικά ένα αληθινό υποψήφιο πρόσωπο και πέντε εσφαλμένους στόχους. Μετά την εφαρμογή αυτού του αλγορίθμου ομαδοποίησης, το σύνολο των αντιπροσωπευτικών υποψηφίων προσώπων μας δίνει την βάση για το επόμενο στάδιο του αλγορίθμου, επιφορτισμένου για τον ακριβή εντοπισμό των προσώπων και για την απόρριψη των εσφαλμένων στόχων. Σε άλλες νευρωνικές προσεγγίσεις, όπως στην [41], το παρόν στάδιο είναι και το τελικό της διαδικασίας. Πράγματι, ήδη έχουμε κάποια στοιχεία για την παρουσία ή όχι προσώπων στην εικόνα, όπως επίσης και κάποιες εκτιμήσεις για τις θέσεις τους και τις κλίμακές τους. Στην προτεινόμενη μέθοδο, η διαδικασία εντοπισμού επεκτείνεται με περαιτέρω ανάλυση καθώς δεν εκτιμάται ότι τα ευρήματα ως τα τώρα είναι καλές εκτιμήσεις του πραγματικού όγκου δραστηριότητας. Στην ενότητα 6.1.5 θα δοθούν πειραματικά αποτελέσματα για την απόδοση των διάφορων στρατηγικών εντοπισμού.

Στο βήμα 4, μία τοπική αναζήτηση γίνεται στην περιοχή γύρω από το υποψήφιο πρόσωπο στον χώρο θέσης και κλίμακας με ένα σενάριο παρόμοιο του σχήματος 5.1. Ορίζεται ένας περιορισμένος χώρος αναζήτησης με κέντρο την τοποθεσία του υποψήφιου προσώπου για την ακριβή μέτρηση του όγκου δραστηριότητας. Αντιστοιχεί σε μία μικρή αυτήν την φορά πυραμίδα εικόνων που καλύπτει δέκα ισόρροπα κατανεμημένες κλίμακες αρχιζόντας από την κλίμακα 0,8 της αυθεντικής του υποψήφιου και τελειώνοντας στην κλίμακα 1,5. Για κάθε μία από αυτές, η παρουσία του προσώπου αποτιμάται σε ένα πλέγμα διαμορφωμένο σε  $16 \times 16$  εικονοστοιχεία. Βρέθηκε ότι αυτή η περιοχή είναι αρκετή για την κάλυψη όλου του όγκου δραστηριότητας ενός αληθινού προσώπου ακόμα και δεδομένης μίας φτωχής αρχικής εκτίμησης. Καθώς ο στόχος αυτού του βήματος είναι να εντοπίσει με ακρίβεια, η λεπτή σάρωση χρησιμοποιείται κατά κανόνα εδώ. Το υπολογιστικό κόστος αυτού του βήματος είναι καθαρά οικονομικότερο, καθώς απαιτούνται μόνο  $16 \times 16 \times 10 = 2.560$  ενεργοποιήσεις του δικτύου ανά υποψήφιο. Τα αποτελέσματα αυτού του βήματος για την εικόνα του παραδείγματός μας φαίνονται στο σχήμα 5.2(iii). Η δραστηριότητα γύρω από το αληθή υποψήφιο πρόσωπο (απεικονιζόμενη σαν μία πληθώρα από μαύρα περιγράμματα

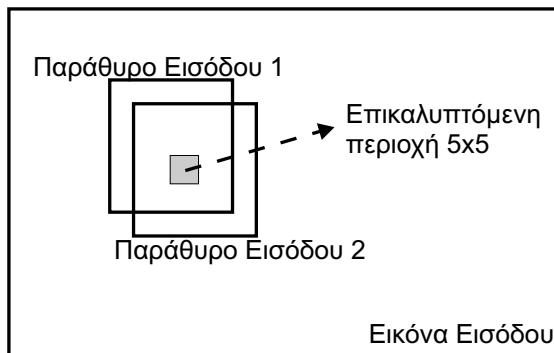
γύρω του) είναι πολύ πιο ισχυρή απ' ό,τι αυτές των εσφαλμένων υποψηφίων. Αυτό θα μας επιτρέψει να τους απορρίψουμε εύκολα.

Το τελικό βήμα της διαδικασίας εντοπισμού δίνεται στο σχήμα 5.2(iv). Βασιζόμενοι στα πειραματικά δεδομένα που δίνονται στο επόμενο κεφάλαιο, ένα υποψήφιος στόχος χρίζεται ως πρόσωπο αν ο αποτιμώμενος όγκος δραστηριότητάς του είναι μεγαλύτερος από  $ThrVol = 22,0$ . Στο παρόντο παρόπλιτα με την εντοπισμένη υποψήφια, η αποτιμώμενη ακριβής θέση και έκταση του προσώπου δίνονται απ' αυτές της μεγαλύτερης απογεγραμμένης απάντησης του δικτύου κατά την διάρκεια του προηγούμενου βήματος. Αυτή η εκτίμηση θέσης και έκτασης (κλίμακας) δίνεται στην εικόνα σαν ένα παράθυρο που περιβάλλει το πρόσωπο. Σαν ένα βήμα μετα-επεξεργασίας, γίνεται απαλοιφή αλληλοεπικαλύψεων πάνω στα ανιχνευμένα πρόσωπα, καθώς γνωρίζουμε εκ των προτέρων ότι δύο περιοχές προσώπων δεν μπορούν να αλληλοεπικαλύπτονται μέχρι ένα ποσοστό, το οποίο ορίστηκε στο 20% σε αυτήν την εργασία. Στην περίπτωση του σχήματος 5.2(iv), η απαλοιφή αλληλοεπικαλύψεων δεν επηρεάζει το τελικό αποτέλεσμα, καθώς πρωτίστως έχει ανιχνευτεί ένα μόνο πρόσωπο.

### 5.3 Επιτάχυνση της Σάρωσης

Η προτεινόμενη τοπολογία έχει και ένα επιπλέον αποφασιστικό πλεονέκτημα. Στις νευρωνικές τοποθετήσεις της βιβλιογραφίας [3, 33, 38, 41, 48], για τον εντοπισμό των προσώπων σε μία δεδομένη κλίμακα, το δίκτυο πρέπει να επαναλαμβάνεται σε όλες τις θέσεις της εικόνας εισόδου. Όλες αυτές οι προσεγγίσεις ακολουθούν την τακτική προεπεξεργασίας των Sung και Poggio [48], η οποία κανονικοποιεί τοπικά το παράθυρο εισόδου πριν την πραγματική τροφοδότηση της εισόδου του δικτύου. Έτσι, κάθε δυνατό παράθυρο εισόδου αντιμετωπίζεται ξεχωριστά. Στην παρούσα εργασία, καμία τοπική προεπεξεργασία δεν εφαρμόζεται πάνω σε αυτά τα παράθυρα αλλά δίνονται κατευθείαν οι τιμές των εικονοστοιχείων της εικόνας, καθολικά και εκ των προτέρων διαμορφωμένες στο πεδίο τιμών  $[-1, 1]$ . Με αυτόν τον τρόπο είμαστε καταρχήν ελεύθεροι από την ανάγκη να επεξεργαζόμαστε αρχικά την προσώπου εισόδου ξεχωριστά.

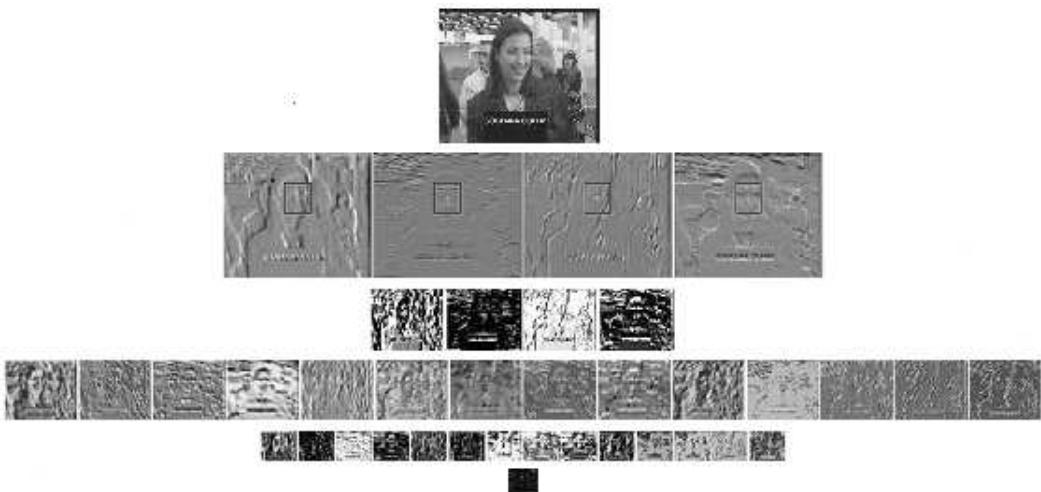
Επιπλέον, καθώς κάθε στρώμα του δικτύου στην ουσία πραγματοποιεί συνελίξεις, για κάθε μία απ' αυτές ένα πολύ μεγάλο μέρος του υπολογισμού είναι κοινός ανάμεσα σε δύο διαδοχικά παράθυρα εισόδου. Το σχήμα 5.3 δίνει ένα τέτοιο παρόπλιτο παράθυρο για την εντοπισμό της εικόνας,  $x_{common}$ , είναι από κοινού μεταξύ των δύο παραθύρων εισόδου του δικτύου. Χάρις την καθαρά συνελικτική φύση του πρώτου στρώματος, ακριβώς ο ίδιος υπολογισμός  $w^T x_{common}$  πρέπει να πραγματοποιηθεί δύο φορές κατά την διάρκεια των δύο ενεργοποήσεων του δικτύου. Το ίδιο φαινόμενο συναντιέται επίσης και στα

**Σχήμα 5.3**

Κοινή περιοχή εισόδου ανάμεσα σε δύο διαδοχικές ενεργοποιήσεις του δικτύου.

επόμενα στρώματα, καθώς η τοπολογία είναι υψηλά δομημένη. Έτσι, πιο γενικά, υπάρχει ένας αρκετά μεγάλος υπολογιστικός πλεονασμός ανάμεσα σε δύο διαδοχικές και διακριτικές εφαρμογές του δικτύου σε μία εικόνα που είναι εύκολο να αποφευχθεί. Είναι αξιοσημείωτο ότι η ίδια παρατήρηση δεν εξακολουθεί να ισχύει για άλλες πλήρως συνδεμένες νευρωνικές τοπολογίες, ακόμα και όταν έχουν μία απλή δομή όπως αυτή της [41] (σχήμα 2.3). Αν και κάποια τοπικότητα των πεδίων υποδοχής χρησιμοποιήθηκε εκεί, η επιτελούμενη λειτουργία δεν είναι μία συνέλιξη πάνω σε όλο το παράθυρο εισόδου αλλά συσχέτιση φορμών (template matching) σε σταθερές θέσεις της εισόδου.

Αυτός ο πλεονασμός εξαλείφεται με το να εκτελούμε τις συνελίξεις ενός στρώματος σε όλη την εικόνα εισόδου και σε ένα βήμα. Ο συνολικός υπολογισμός όλων των στρωμάτων τότε απλώς μειώνεται σε μια σωλήνωση συνελίξεων και μη γραμμικών μετασχηματισμών, καθώς τα αποτελέσματα τους ενός στρώματος πάνω σε ολόκληρη την εικόνα περνάνε για επεξεργασία στο επόμενο. Πιο συγκεκριμένα, η λειτουργία του πρώτου στρώματος (C1) είναι ισοδύναμη με τέσσερις συνελίξεις της εικόνας εισόδου με τους τέσσερις  $5 \times 5$  συνελικτικούς πυρήνες των αντίστοιχων χαρτών χαρακτηριστικών αυτού του στρώματος. Στην συνέχεια, η λειτουργία του στρώματος S1 είναι ισοδύναμη με τοπικό υπολογισμό μέσου όρου όλων των μη επικαλυπτόμενων  $2 \times 2$  μπλοκ των αποτελεσμάτων των συνελίξεων, ακολουθούμενο από την μη γραμμικότητα των σιγμοειδών. Αυτές οι λειτουργίες έχουν πραγματοποιηθεί σε ολόκληρη την εικόνα χωρίς κανένα υπολογιστικό πλεονασμό, δίνοντας τέσσερις εικόνες της μισής διάστασης της εισόδου που μπορούν να θεωρηθούν σαν χάρτες χαρακτηριστικών πάνω σε ολόκληρη την εικόνα εισόδου. Ακολουθούμενοι την ίδια λογική, μπορούμε εύκολα να επεκτείνουμε τέτοιου είδους λειτουργίες και για τα επόμενα στρώματα του δικτύου. Το τελικό αποτέλεσμα αυτής της σωλήνωσης είναι μία εικόνα  $4 \times 4$  φορές μικρότερη από την εικόνα εισόδου, και η οποία φέρει τις απαντήσεις της σάρωσης με το δίκτυο σαν αυτό να



#### Σχήμα 5.4

Οι παραγόμενες εικόνες της σωλήνωσης. Οι χάρτες χαρακτηριστικών του σχήματος 4.1 τώρα επεκτείνονται και καλύπτουν ολόκληρη την εικόνα εισόδου. Το τελικό αποτέλεσμα είναι μία εικόνα που φέρει όλες τις απαντήσεις του δικτύου με βήμα 4 και στις δύο κατευθύνσεις. Η τοποθεσία του προσώπου δίνεται με ένα ορθογώνιο στους τέσσερις πρώτους χάρτες.

εφαρμοζόταν με βήμα 4 εικονοστοιχεία και προς τις δύο κατευθύνσεις. Με άλλα λόγια, αυτή η σωλήνωση υλοποιεί την μέθοδο της πρόχειρης σάρωσης, στην οποία αναφερθήκαμε παραπάνω. Αυτή η σωλήνωση μπορεί να απεικονιστεί όπως στο σχήμα 5.4.

Για την υλοποίηση της λεπτής σάρωσης, χρειαζόμαστε μία λίγο διαφορετική αρχιτεκτονική σωλήνωσης η οποία θα μας δίνει τις απαντήσεις του δικτύου για όλες τις θέσεις της εικόνας εισόδου και όχι μόνο αυτές με το βήμα των 4 εικονοστοιχείων. Για αυτό τον σκοπό, ας παρατηρήσουμε πρώτα ότι μετά το πρώτο συνελικτικό στρώμα, κάθε χάρτης χαρακτηριστικών διαχωρίζεται σε μη επικαλυπτόμενα τοπικά μπλοκ  $2 \times 2$  εικονοστοιχείων, τα οποία και θα δώσουν μία τιμή εξόδου στον αντίστοιχο χάρτη του πρώτου υποδειγματοληπτικού στρώματος. Υπάρχουν όμως τέσσερις τέτοιοι διαφορετικοί διαχωρισμοί που δεν προκαλούν επικαλυπτόμενα μπλοκ. Αποκομίζονται με την μετατόπιση του πρώτου  $2 \times 2$  μπλοκ κατά  $(0, 0)$ ,  $(1, 0)$ ,  $(0, 1)$  και  $(1, 1)$  εικονοστοιχεία. Στην περίπτωση της πρόχειρης σάρωσης, μόνο ο πρώτος τρόπος διαχωρισμού λαμβάνονταν υπ' όψη. Τώρα, για τις ανάγκες της λεπτής σάρωσης, θα χρησιμοποιήσουμε και τους υπόλοιπους. Το πρώτο συνελικτικό στρώμα εφαρμόζεται, όπως και πριν, και οι εικόνες εξόδου του θα επεξεργαστούν τώρα από τέσσερα διαφορετικά υποδειγματοληπτικά στρώματα S1, αντίστοιχα των τεσσάρων διανυσμάτων μετατόπισης. Στην συνέχεια, κάθε μία από αυτές τις τέσσερις υποδειγματοληπτημένες εικόνες θα υποστεί επεξεργασία από το στρώμα C2 ξεχωριστά. Οι εικόνες εξόδου του θα υποδειγ-

	+	$\times$	Σιγμοειδής
Στρώμα C1 - S1	109	101	1
Στρώμα C2 - S2	$\frac{846}{16}$	$\frac{734}{16}$	$\frac{14}{16}$
Στρώμα N1	$\frac{602}{16}$	$\frac{588}{16}$	$\frac{14}{16}$
Στρώμα N2	$\frac{15}{16}$	$\frac{14}{16}$	$\frac{1}{16}$
Σύνολο	$\frac{3207}{16}$	$\frac{2952}{16}$	$\frac{45}{16}$
Περίπου	200	185	3

**Πίνακας 5.1**

Αριθμός πράξεων ανά εικονοστοιχείο εικόνας για την πρόχειρη σάρωση και για δεδομένη κλίμακα.

ματοληπτηθούν και πάλι από τέσσερα διαφορετικά στρώματα S2, όπως και στην περίπτωση του S1. Τέλος, οι παραγόμενες εικόνες τροφοδοτούν τα στρώματα N1 και N2. Έτσι, παράγονται τελικά 16 εικόνες εξόδου, οι οποίες συγχωνεύονται με βάση το τι μετατοπίσεις φέρει η καθεμία για την δημιουργία ενός τελικού αποτελέσματος με το ίδιο μέγεθος με την εικόνα εισόδου.

## 5.4 Υπολογιστικές Απαιτήσεις

Το υπολογιστικό κόστος για την σάρωση της εικόνας εισόδου όταν το δίκτυο εφαρμόζεται διακριτικά είναι απλό να βρεθεί: είναι απλώς το υπολογιστικό κόστος μίας μοναδικής ενεργοποίησης του δικτύου επί τις διαστάσεις της εικόνας. Το προτεινόμενο δίκτυο έχει 131.475 συνδέσεις, έτσι απαιτείται προσεγγιστικά<sup>1</sup> αυτός ο αριθμός πολλαπλασιασμών και αυτός ο αριθμός προσθέσεων για την ενεργοποίησή του. Επιπλέον, απαιτούνται και 1.499 πράξεις σιγμοειδούς (στρώματα S1, S2, N2 και N1). Για την πρόχειρη σάρωση με βήμα 4 εικονοστοιχεία απαιτείται ο ίδιος αριθμός πράξεων διαιρεμένος με 16, δηλ. περίπου 16.400 πολλαπλασιασμοί/προσθέσεις και 94 σιγμοειδής, ανά εικονοστοιχείο εισόδου.

Το υπολογιστικό κόστος της σωλήνωσης χωρίς υπολογιστικό πλεονασμό της πρόχειρης σάρωσης δίνεται αναλυτικά στον πίνακα 5.1, συνολικά και ανά στρώμα. Για παράδειγμα,

<sup>1</sup>Σε κάθε λειτουργία υποδειγματοληψίας, υπάρχουν 5 συνδέσεις αλλά απαιτούνται 4 προσθέσεις και μόνον ένας πολλαπλασιασμός. Άλλα καθώς ο αριθμός των συνδέσεων υποδειγματοληψίας είναι μικρός (7.420) σε σύγκριση με τον συνολικό του δικτύου, μπορούμε να προσεγγίσουμε τον αριθμό προσθέσεων και τον αριθμό πολλαπλασιασμών να είναι ίσος με τον αριθμό συνδέσεων του δικτύου.

	+	×	Σιγμοειδής
Στρώμα C1 - S1	124	104	4
Στρώμα C2 - S2	264	194	14
Στρώμα N1	602	588	14
Στρώμα N2	15	14	1
Σύνολο	1005	900	33

### Πίνακας 5.2

Αριθμός πράξεων ανά εικονοστοιχείο εικόνας για την λεπτή σάρωση και για δεδομένη κλίμακα.

μία συνέλιξη του στρώματα C1 απαιτεί 26 προσθέσεις (συμπεριλαμβανομένου και του bias) και 25 πολλαπλασιασμούς ανά εικονοστοιχείο. Η υποδειγματοληψία απαιτεί 5 προσθέσεις και ένα πολλαπλασιασμό ανά εικονοστοιχείο, αλλά διαιρεμένα με 4 καθώς η διάσταση έχει μειωθεί τώρα. Τέλος, όλοι αυτοί οι αριθμοί θα πρέπει να πολλαπλασιαστούν με τέσσερα καθώς έχουμε 4 χάρτες χαρακτηριστικών σε αυτά τα δύο στρώματα. Έτσι παίρνουμε τις 109 προσθέσεις και τους 101 πολλαπλασιασμούς που βλέπουμε στον πίνακα. Ο υπολογισμός και των υπόλοιπων αριθμών είναι το ίδιο απλός. Παρατηρήστε ότι, οι παρενέργειες ορίων των συνελίξεων δεν λήφθηκαν υπ' όψη για λόγους απλότητας της παρουσίασης. Για παράδειγμα, το στρώμα S1 λαμβάνει εικόνες  $(N-4) \times (M-4)$  εικονοστοιχείων και όχι  $N \times M$  (ολόκληρη την εικόνα εισόδου), όπως υποθέσαμε παραπάνω. Αυτές οι διορθώσεις μπορούν να μην ληφθούν υπ' όψη, ειδικά όταν οι εικόνες εισόδου είναι μεγάλου μεγέθους.

Ο συνολικός αριθμός πράξεων που απαιτείται για την πρόχειρη σάρωση είναι περίπου 200 προσθέσεις, 185 πολλαπλασιασμοί και 3 σιγμοειδής ανά εικονοστοιχείο εισόδου. Έτσι έχουμε περίπου 400 βασικές αριθμητικές πράξεις, σε σύγκριση με τις περίπου 16.500 βασικές πράξεις που απαιτούνται για την σάρωση της εικόνας με επιμέρους ενεργοποιήσεις του δικτύου. Με άλλα λόγια, έχουμε μία επιτάχυνση υπολογισμών γύρω στις 40 φορές. Οι υπολογιστικές απαιτήσεις για την λεπτή σάρωση δίνονται στον πίνακα 5.2. Αυτή η σωλήνωση τελειώνει με περίπου 2.000 βασικές αριθμητικές πράξεις. Θα περιμέναμε ότι θα χρειαζόταν περίπου 16 φορές περισσότεροι υπολογισμοί γιατί η λεπτή σάρωση δίνει 16 φορές περισσότερες απαντήσεις του δικτύου. Τελικά, χρειαζόμαστε μόλις 5 φορές περισσότερες πράξεις, καθώς με την λεπτή σάρωση επικάλυψη των υπολογισμών είναι πολύ μεγαλύτερη, άρα έχουμε μεγαλύτερο υπολογιστικό κέρδος. Σε σύγκριση με την σάρωση με επιμέρους ενεργοποιήσεις, έχουμε τώρα υπολογιστική επιτάχυνση της τάξεως του 130.

Τέλος, πρέπει να προσθέσουμε στα παραπάνω αποτελέσματα των πινάκων 5.1 και 5.2

και το κόστος του βήματος ακριβούς εντοπισμού. Δυστυχώς, αυτό το κόστος είναι απρόβλεπτο, καθώς εξαρτάται (γραμμικά) από τον αριθμό των στόχων που ανιχνεύτηκαν στο βήμα πρόχειρου εντοπισμού. Επιπλέον, μπορεί να θεωρηθεί ως αμελητέο καθώς χρησιμοποιεί το δίκτυο για την σάρωση πολύ μικρών περιοχών.

#### 5.4.1 Σύγκριση με Άλλες Μεθόδους

Στις μεθόδους που χρησιμοποιούν την στρατηγική προεπεξεργασίας των Sung και Poggio [48], κάθε δυνατό παράθυρο εισόδου πρέπει να υποστεί προεπεξεργασία, άρα και να θεωρηθεί ξεχωριστά. Υποθέτουμε ότι έχουμε να προεπεξεργαστούμε ένα  $M \times N$  παράθυρο με γραμμική διόρθωση φωτισμού και στην συνέχεια με εξισορρόπηση ιστογράμματος. Για την πρώτη εργασία, προσεγγίζουμε κάθε τιμή εικονοστοιχείου  $p_{ij}$  σε αυτό το παράθυρο ως:

$$p_{ij} = ai + bj + c$$

όπου  $(i, j)$  είναι οι συντεταγμένες του εικονοστοιχείου και  $a, b, c$  είναι οι συντελεστές της γραμμικής μάσκας φωτισμού που ψάχνουμε. Η λύση εμπεριέχει την επίλυση με ελάχιστα τετράγωνα του συστήματος  $Ax = p$ , όπου  $A$  είναι ο  $(MN) \times 3$  πίνακας των συντεταγμένων,  $x = [a \ b \ c]^T$  και  $p$  είναι το διάνυσμα με τις τιμές των εικονοστοιχείων. Η λύση αυτή και η τελική αφαίρεση αυτής της μάσκας από το παράθυρο θα στοιχίσει τελικά περίπου  $12 \times (MN)$  προσθέσεις και  $7 \times (MN)$  πολλαπλασιασμούς. Η εξισορρόπηση ιστογράμματος θα προσθέσει στα πραπάνω  $(MN) + 256$  προσθέσεις και  $(MN)$  πολλαπλασιασμούς. Έτσι, έχοντας ένα  $20 \times 20$  παράθυρο ([41, 48]) θα απαιτούσε περίπου 8.500 βασικές αριθμητικές πράξεις, ενώ ένα παράθυρο  $15 \times 20$  ([3]) θα απαιτούσε περίπου 6.500. Έτσι, το απαιτούμενο κόστος προεπεξεργασίας είναι ήδη περίπου 3 με 4 φορές μεγαλύτερο από το να εφαρμόσουμε την προτεινόμενο μέθοδο με λεπτή σάρωση.

Οι Rowley *et al.* [41] δοκίμασαν διάφορες τοπολογίες και στρατηγικές διαιτησίας. Τα καλύτερα αποτελέσματα προέκυψαν με το αναφερόμενο Σύστημα 11, το οποίο χρησιμοποιεί δύο νευρωνικά δίκτυα με 2 και 3 αντίγραφα των κρυμμένων μονάδων, αντίστοιχα. Αυτή η τοπολογία έχει σαν σύνολο 7.262 βάρη. Ένας πρόχειρος υπολογισμός του υπολογιστικού κόστους αυτού του MLP θα μας δώσει περίπου δύο φορές τον αριθμό των βαρών βασικές πράξεις, καθώς κάθε νευρώνας εκτελεί τον υπολογισμό  $w^T x$ . Έτσι, ένας προσεγγιστικός αριθμός των 15.000 βασικών αριθμητικών πράξεων επιτελείται ανά εικονοστοιχείο (εξαιρουμένης φυσικά της προεπεξεργασίας).

Οι Féraud *et al.* [3] χρησιμοποιούν μία νευρωνική προσέγγιση βασισμένη στο constrained generative model (CGM). Η ιδέα πίσω απ' αυτό το μοντέλο είναι η επιδιώξη ενός μη γραμμικού PCA. Η ταξινόμηση επιτυγχάνεται με την θεώρηση των σφάλματος επανακατασκευής

του CGM. Το CGM τους είναι αυτοσυσχετιστικό πλήρως συνδεμένο MLP με τρία επίπεδα βαρών, με 300 μονάδες εισόδου και 300 μονάδες εξόδου (το μέγεθος του παραθύρου εισόδου τους είναι  $15 \times 20$  και η έξοδος είναι του ίδιου μεγέθους μια και επιχειρείται ανακατασκευή της εισόδου). Το πρώτο αρχιμένο στρώμα έχει 35 μονάδες, ενώ το δεύτερο έχει 50. Τα καλύτερα αποτελέσματα επιτεύχθηκαν με μία τοπολογία συνδυασμού τεσσάρων CGMs με δεσμευμένη μίξη. Αυτή η τοπολογία έχει, εν τέλει, τον αρκετά μεγάλο αριθμό 140.741 βαρών. Άρα, περισσότερες από 280.000 πράξεις επιτελούνται ανά εικονοστοιχείο για μία δεδομένη κλίμακα. Για την μείωση αυτού του εξαιρετικά υψηλού υπολογιστικού κόστους, κάποιες στρατηγικές προφιλτραρίσματος επιτάσσονται, όπως φιλτράρισμα χρώματος ή κίνησης, όταν φυσικά αυτά τα στοιχεία είναι διαθέσιμα. Στην γενική περίπτωση των στατικών μονόχρωμων εικόνων, ένα προ-δίκτυο εφαρμόζεται σε κάθε θέση της εισόδου για την επιλογή υποψηφίων. Αυτό το προ-δίκτυο έχει 300 μονάδες εισόδου, 20 αρχιμένες και μία εξόδου, έτσι απαιτεί προσεγγιστικά 12.000 αριθμητικές πράξεις ανά εικονοστοιχείο. Σύμφωνα με τους συγγραφείς, δίνει ένα μέσο ρυθμό εσφαλμένων ειδοποιήσεων 7% των παραθύρων εισόδου. Έτσι, το σύστημά τους απαιτεί με το φιλτράρισμα προ-δικτύου γύρω στις 32.000 αριθμητικές πράξεις ανά εικονοστοιχείο.

Καθώς και οι τρεις μέθοδοι ακολουθούν την ίδια στρατηγική όσον αφορά την κλίμακα, είμαστε έτοιμοι να συγκρίνουμε τα υπολογιστικά κόστη τους. Μπορεί να παρατηρηθεί αρκετά εύκολα ότι η προτεινόμενη μέθοδος είναι λιγότερο ακριβή στον χρόνο υπολογισμού που χρειάζεται, 16 ή 8 φορές σε σύγκριση με αυτήν των Féraud *et al.* ή των Rowley *et al.*, αντίστοιχα και όταν δεν συνυπολογίζεται το κόστος προεπεξεργασίας. Στην αντίθετη περίπτωση, η προτεινόμενη μέθοδος είναι 19 ή 12 φορές ταχύτερη απ' αυτήν των Féraud *et al.* ή των Rowley *et al.*, αντίστοιχα.

## ΚΕΦΑΛΑΙΟ 6

# Πειραματικά Αποτελέσματα

Σε αυτό το κεφάλαιο θα αποτιμήσουμε την απόδοση του προτεινόμενο συστήματος σε διάφορα και δύσκολα σύνολα δοκιμής. Κάποια απ' αυτά έχουν επίσης χρησιμοποιηθεί και από άλλους συγγραφείς και εξυπηρετούν σαν πηγή σύγκρισης με άλλες προσεγγίσεις. Επιπλέον, δύο νέα σύνολα δοκιμής εισάγονται σε αυτήν την εργασία, με διαφορετικές στατιστικές ιδιότητες για την δοκιμή της καθολικότητας του προτεινόμενου συστήματος. Κυριαρχούνται από μη ελεγχόμενες καταστάσεις αληθινού κόσμου, στις οποίες είναι πιθανό να ζητηθεί να λειτουργήσει ένας ανιχνευτής προσώπων γενικής χοήσεως. Θα δοθούν αναλυτικά ποσοστά ανίχνευσης και εσφαλμένων ειδοποιήσεων το οποία και θα αποσαφηνίσουν την δυναμική της χρησιμοποίησης συνελικτικών νευρωνικών δικτύων για την ανίχνευση προσώπων. Τα αποτελέσματα στα σύνολα δοκιμής που ακολουθούν αναφέρονται στην μέθοδο λεπτής σάρωσης ακολουθούμενη από ακριβή εντοπισμό γύρω από κάθε στόχο (ενότητα 5.2). Θα δοθεί επίσης και σύγκριση μεταξύ των διάφορων στρατηγικών αναζήτησης. Τέλος, θα παρουσιαστεί και μια ανάλυση γύρω από την ευαισθησία του δικτύου σε μία σειρά από πιθανούς μετασχηματισμούς (παραμορφώσεις) της εισόδου, συμπεριλαμβανομένου μεταβλητών συνθηκών εικόνας, περιστροφής και αλλαγής έκφρασης του προσώπου. Είναι ένα πολύ χρήσιμο εργαλείο για την ποσοτική εκτίμηση των ορίων της προτεινόμενης προσέγγισης.

### 6.1 Αποτίμηση και Σύγκριση

#### 6.1.1 Περιγραφή των Συνόλων Δοκιμής

Τα τέσσερα σύνολα δοκιμής που θα χρησιμοποιηθούν σε αυτό το κεφάλαιο είναι τα εξής:

- Το σύνολο δοκιμής *CMU*. Εισάχθηκε από τους Rowley *et al.* [41] και είναι ως τώρα το πιο συχνά αναφερόμενο σύνολο δοκιμής της βιβλιογραφίας. Αποτελείται από 130 ασπρόμαυρες εικόνες με ένα σύνολο από 507 καταμετρημένα πρόσωπα, περιστραμ-

μένα το πολύ  $\pm 15$  μοίρες. Στα ακόλουθα πειράματα, η πυραμίδα εικόνων για τον χειρισμό των των διάφορων κλιμάκων αρχίζει από 24 και τελειώνει σε 360 εικονοστοιχεία. Δεδομένου αυτών των κλιμάκων, ένα σύνολο από 215.312.256 υπο-παράθυρα των εικόνων θα ερευνηθούν από το δίκτυο και θα χαρακτηριστούν ως ‘πρόσωπα’ ή όχι. Αυτό το σύνολο περιέχει 23 εικόνες από αυτό των Sung και Poggio [48], αναφερόμενο ως σύνολο *MIT*. Ένα υποσύνολο του συνόλου CMU, αναφερόμενο ως σύνολο *CMU-125*, έχει επίσης χρησιμοποιηθεί από πολλούς ερευνητές. Στο σύνολο αυτό αποκλείστηκαν κάποια πρόσωπα σχεδιασμένα με το χέρι ή καρτούν, αφήνοντας τελικά 483 από τα 507. Το αντίστοιχο υποσύνολο του *MIT* στο οποίο αποκλείστηκαν οι παραπάνω περιπτώσεις αναφέρεται ως *MIT-20*. Αν και το προτεινόμενο σύστημα μπορεί να δοκιμαστεί πάνω σε ολόκληρο το σύνολο CMU, θα δοθούν επίσης αποτελέσματα και πάνω στα υποσύνολά του προς χάριν της σύγκρισης. Κάποιες από τις εικόνες αυτού του συνόλου δίνονται στο σχήμα 6.4 (με απεικονισμένα πάνω σε αυτές τα αποτελέσματα της προτεινόμενης μεθόδου). Μπορούμε να παρατηρήσουμε ότι κάποιες εικόνες, που ανήκουν κατά κυριότητα στο υποσύνολο *MIT*, αυτού του συνόλου είναι σε κακή ποιότητα έχοντας τα πρόσωπά τους δύσκολα ορατά.

- Το σύνολο *DiVAN*. Περιέχει έγχρωμες εικόνες με προέλευση το Institut National Audiovisuel (INA), Γαλλία και τη Τηλεόραση EPT, Ελλάδα. Αυτές οι εικόνες συλλέχθηκαν για την αποτίμηση της απόδοσης του συστήματος των Garcia και Tziritas [8, 9] που αναπτύχθηκε ως μέρος του Ευρωπαϊκού προγράμματος *DiVAN*. Αυτό το σύνολο δοκιμής περιέχει 100 εικόνες με 104 πρόσωπα μεγέθους μεγαλύτερου από  $48 \times 80$ , το οποίο ήταν και το μικρότερο μέγεθος που ο ανιχνευτής τους μπορούσε να χειριστεί. Δυστυχώς κάποια από αυτά είναι περιστραμμένα περισσότερο από  $\pm 20$  μοίρες, κανονούντας την σύγκριση δύσκολη. Κάποιες εικόνες από αυτό το σύνολο φαίνονται στο σχήμα 6.5.
- Το σύνολο *WEB*, που εισάγεται με αυτήν την εργασία. Είναι ένα τυχαία επιλεγμένο υποσύνολο των εικόνων που έχουν αποσταλεί στην δικτυακή<sup>1</sup> επίδειξη του συστήματος, επιτρέποντας στον επισκέπτη να αποστέλει εικόνες και να βλέπει αμέσως το αποτέλεσμα. Αυτό το σύνολο περιέχει μια μεγάλη ποικιλία από παραδείγματα, και όχι προκατειλημμένα ή ευνοϊκά προς μία συγκεκριμένη επιλογή. Μπορεί να θεωρηθεί ως ένα αντιπροσωπευτικό σύνολο των ‘μέσων’ περιβαλλόντων που συναντά κάποιος στο *WEB*, την μεγαλύτερη βιβλιοθήκη πολυμέσων που έχει κατασκευαστεί ποτέ. Περιέχει 215 εικόνες, 499 πρόσωπα με κλίμακες (ύψος) να διαφέρουν από 24 σε 360 εικονοστοιχεία και ένα σύνολο από 256.887.936 υπο-παράθυρα προς διερεύνηση από

<sup>1</sup><http://www.csd.uoc.gr/~cgarcia/FaceDetectDemo.html>, παράρτημα A.

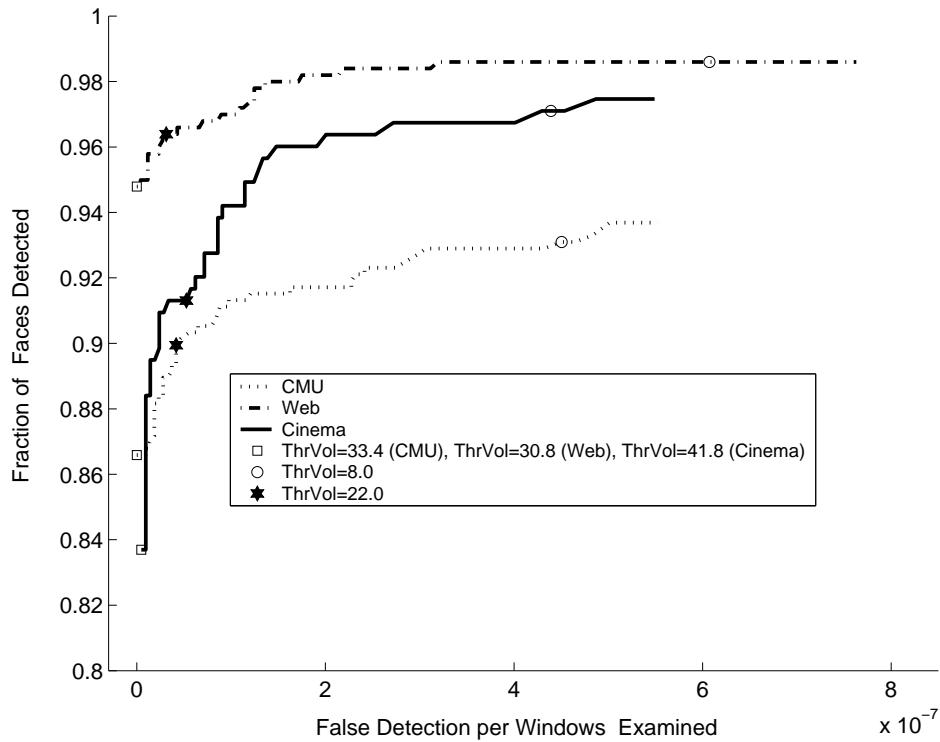
τον ανιχνευτή. Ο στόχος αυτού του συνόλου είναι η επίδειξη της απόδοσης του συστήματος σε ‘κανονικές’ συνθήκες λειτουργίας. Κάποιες εικόνες από αυτό το σύνολο φαίνονται στο σχήμα 6.6.

- Το σύνολο *Σινεμά*, που επίσης εισάγεται με αυτήν την εργασία. Αποτελείται από 162 εικόνες που βρέθηκαν σε συλλογές εικόνων σχετικά με κινηματογραφικά φιλμ και με 276 πρόσωπα. Αυτό το σύνολο θεωρείται και ως το πιο προκλητικό από άποψη απόδοσης, καθώς πολλές απ' αυτές τις εικόνες έχουν επίτηδες επιλεγεί για την δοκιμή των ορίων του συστήματος. Περιέχει ένα μεγάλο αριθμό από πρόσωπα με ακραίες εκφράσεις ή πόζες, καθώς και καταστάσεις μερικής επικάλυψης ή δύσκολης σκίασης. Επιπλέον, το επίπεδο δυσκολίας αυξάνεται και από την κατά κανόνα παρουσία φόντων με πολύπλοκη υφή. Τα πρόσωπα διαφέρουν σε κλίμακα μεταξύ των 36 και των 360 εικονοστοιχείων. Οι πυραμίδες των εικόνων θα δώσουν τελικά 209.450.880 παράθυρα υπό εξέταση. Ο στόχος αυτού του συνόλου είναι η επίδειξη της απόδοσης του συστήματος σε ‘εχθρικές’ συνθήκες λειτουργίας. Κάποιες εικόνες από αυτό το σύνολο φαίνονται στο σχήμα 6.7.

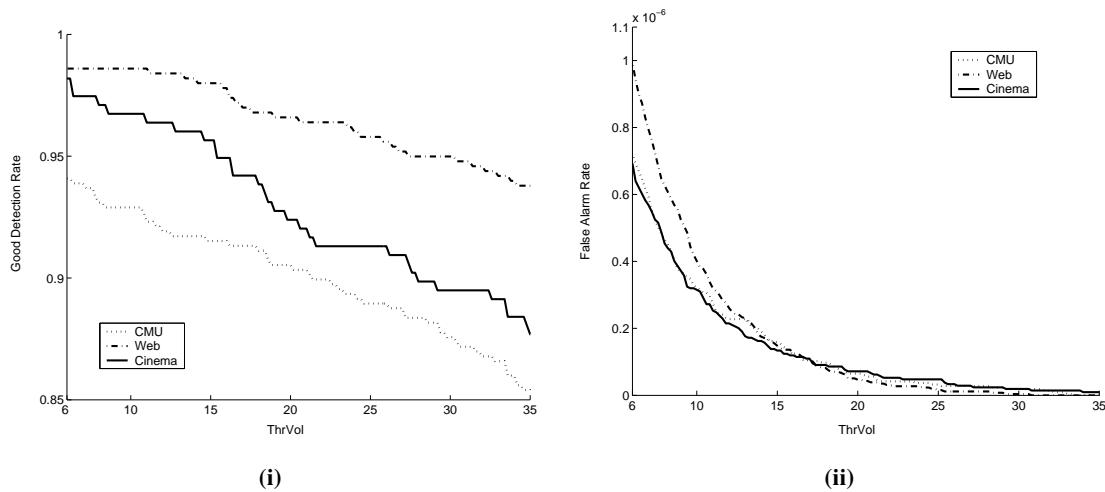
Όλες οι εικόνες των παραπάνω συνόλων δεν χρησιμοποιήθηκαν κατά την εκπαίδευση του δικτύου, ούτε με την μορφή παραδειγμάτων προσώπων και ούτε με την μορφή εικόνων φόντων για την ενίσχυση της ικανότητάς του απόρριψης. Οι εικόνες των συνόλων WEB και *Σινεμά* είναι ελεύθερα διαθέσιμες στον δικτυακό τόπο της επίδειξης του συστήματος. Επίσης δίνεται και η καταμέτρηση αυτών των προσώπων (αληθιοφάνεια), ελπίζοντας ότι αυτά τα δεδομένα θα χρησιμοποιηθούν στο μέλλον από άλλους συγγραφείς για αποτίμηση και σύγκριση. Το σύνολο *Σινεμά* αποτελείται κυρίως από έγχρωμες εικόνες, έτσι είναι δυνατή η χρήση και από προσεγγίσεις που βασίζονται στο χρώμα.

### 6.1.2 Συνολική Απόδοση

Σε κάποιες εργασίες, αναφερόμενα αποτελέσματα μεθόδων σε σύνολα προσώπων είναι κάπως δύσκολο να κατανοηθούν. Για μία δεδομένη προσέγγιση, κάποια αποτελέσματα αντιστοιχούν σε μέγιστα ποσοστά ανίχνευσης με ένα μεγάλο αριθμό εσφαλμένων ειδοποιήσεων, ενώ άλλα σε ένα χαμηλότερο ποσοστό ανίχνευσης, συνοδευόμενο από ένα μικρό αριθμό εσφαλμένων ειδοποιήσεων. Πράγματι, στους περισσότερους ανιχνευτές προσώπων μπορούμε να ρυθμίσουμε κάποιες παραμέτρους (συνήθως ένα κατώφλι), ανάλογα το κατά πόσο συντηρητικός στις αποφάσεις του θέλουμε να είναι ο ανιχνευτής. Αυτές οι ρυθμίσεις επηρεάζουν τα ποσοστά (ή ρυθμούς) ανίχνευσης και εσφαλμένων ειδοποιήσεων, ανταλλάσσοντας τα μεν για τα δε. Ο ρυθμός ανίχνευσης ορίζεται ως ο λόγος μεταξύ των επιτυχημένων (αληθών) ανιχνεύσεων και των αριθμού των καταμετρημένων προσώπων στο δεδομένο σύνολο

**Σχήμα 6.1**

Η καμπύλη ROC για τα σύνολα δοκιμής CMU, WEB και Σινεμά.

**Σχήμα 6.2**

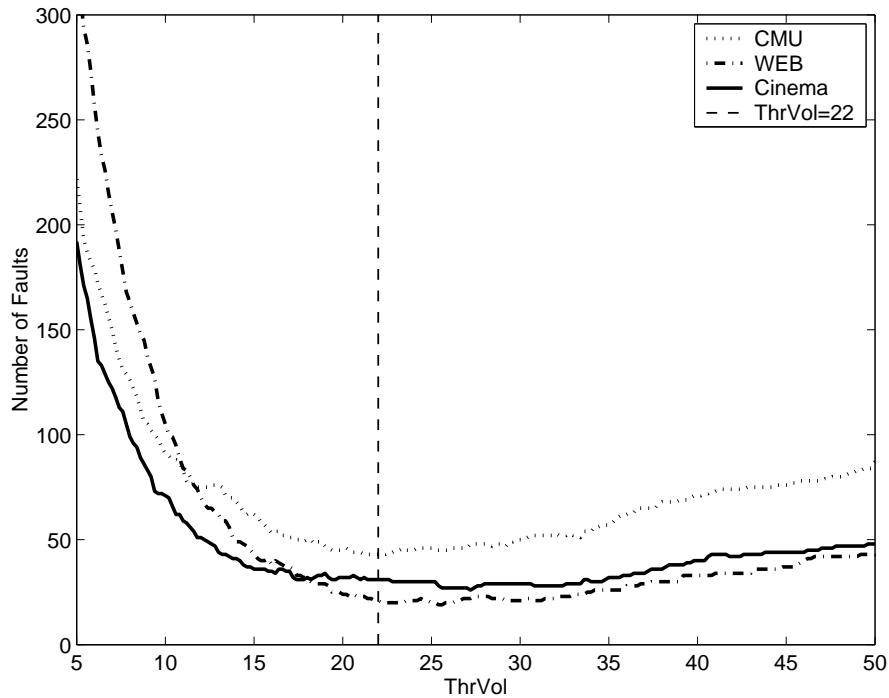
Ποσοστά ανίχνευσης και αριθμός εσφαλμένων ειδοποιήσεων έναντι του  $ThrVol$ .

δοκιμής. Ο ρυθμός εσφαλμένων ειδοποιήσεων ορίζεται ως ο λόγος μεταξύ του αριθμού των εσφαλμένων αποκρίσεων του δικτύου και του συνολικού αριθμού των παραθύρων στα οποία έγινε η αναζήτηση των πρόσωπων.

Αυτή η ανταλλαγή μπορεί να εκφραστεί ποσοτικά μέσω της καμπύλης ROC (Receiver Operating Characteristic). Οι καμπύλες ROC για τα σύνολα δοκιμής WEB, CMU και Σινεμά<sup>2</sup> δίνονται στο σχήμα 6.1. Κάθε σημείο σ' αυτές τις καμπύλες αντιστοιχεί σε μία συγκεκριμένη τιμή του *ThrVol*, του κατωφλίου για τον όγκο δραστηριότητας που χρησιμοποιούμε για να χαρακτηρίσουμε έναν στόχο ως πρόσωπο ή όχι. Για κάθε τέτοια τιμή, λαμβάνουμε από ένα ποσοστό ανίχνευσης, το οποίο βρίσκεται με την προβολή πάνω στον άξονα  $y$  του γραφήματος (εκφρασμένο σαν πιθανότητα ανίχνευσης). Επίσης λαμβάνουμε και από ένα ρυθμό εσφαλμένων ειδοποιήσεων, που βρίσκεται με την προβολή του σημείου πάνω στον άξονα  $x$  (εξεφρασμένο σαν πιθανότητα εσφαλμένης ειδοποίησης). Καθώς ο χώρος υπόθεσης είναι αρκετά μεγάλος με εκατομμύρια παραθύρων προς εξέταση, αυτή η πιθανότητα είναι της τάξεως του  $10^{-7}$ . Μία καμπύλη ROC αρχίζει από σημεία υψηλών τιμών κατωφλίου, δίνοντας τα αντίστοιχα ποσοστά ανίχνευσης με λίγες ή καθόλου εσφαλμένες ειδοποιήσεις. Με την επιλογή τέτοιων τιμών για το κατώφλι, απαιτούμε από τον ανιχνευτή να είναι όσο το δυνατόν περισσότερο συντηρητικός στις αποφάσεις του. Στο τέλος της καμπύλης βλέπουμε τα μεγαλύτερα δυνατά ποσοστά ανίχνευσης, αλλά ταυτόχρονα ανταλλαγμένα με πολύ υψηλά ποσοστά εσφαλμένων ειδοποιήσεων. Σε αυτά τα σημεία ζητάμε από τον ανιχνευτή να είναι όσο το δυνατόν ανοιχτός, αγνοώντας εντελώς τον κίνδυνο από τα υψηλά ποσοστά εσφαλμένων ειδοποιήσεων. Γενικά, θα επιθυμούσαμε μια καμπύλη ROC να είναι όσο το δυνατόν περισσότερο ‘αρθρωτή’ (εκτός βέβαια του να αναφέρει υψηλά ποσοστά ανίχνευσης!). Αυτό σημαίνει ότι υπάρχουν κάποια σημεία, στο ‘γόνατο’ της καμπύλης, στα οποία μπορούμε να διατηρήσουμε υψηλά ποσοστά ανίχνευσης χωρίς να κάνουμε σημαντικές παραχωρήσεις πάνω στην ικανότητα απόρριψης του συστήματος. Βλέπουμε στις καμπύλες του σχήματος ένα αντιπροσωπευτικό σημείο από κάθε μία των παραπάνω περιπτώσεων (χωρίς εσφαλμένες ειδοποιήσεις,  $ThrVol = 22$ ,  $ThrVol = 8$ ).

Κοιτάζοντας στο σχήμα 6.1, παρατηρούμε πρώτα ότι οι καμπύλες αρχίζουν σε σημεία που ήδη έχει επιτευχθεί ένα πολύ καλό ποσοστό ανίχνευσης συνοδευόμενο από μηδενικές εσφαλμένες ειδοποιήσεις. Πιο συγκεκριμένα, βλέπουμε ποσοστά 94,8%, 86,6% και 83,7% για τα σύνολα WEB, CMU και Σινεμά έχουν ήδη επιτευχθεί. Μειώνοντας την τιμή του *ThrVol*, εισερχόμαστε στην περιοχή ‘ανταλλαγής’ των καμπύλων. Είναι εύκολα αντιληπτή μία αρθρωτή συμπεριφορά σε όλες τις καμπύλες, όντας πιο δυνατή σ' αυτές των συνόλων WEB και CMU. Όσον αφορά το σύνολο Σινεμά, παρατηρούμε μία περισσότερο γραμμική

<sup>2</sup>Η καμπύλη ROC για το σύνολο DiVAN δεν αποτιμήθηκε γιατί το σύνολο αυτό είναι αρκετά μικρότερο και όσον αφορά τα καταμετρημένα πρόσωπα και όσον αφορά τα εξεταζόμενα παράθυρα.



**Σχήμα 6.3**

Αριθμός σφαλμάτων έναντι  $ThrVol$ . Ένα σφάλμα αντιστοιχεί σε μία εσφαλμένη ειδοποίηση ή σε ένα μη ανιχνευμένο πρόσωπο.

συμπεριφορά, υποδηλώνοντας την ειδική δυσκολία αυτού του συνόλου. Τελικά, δίνοντας τιμές στο  $ThrVol$  πολύ κοντά στο μηδέν, κερδίζουμε ποσοστά ανίχνευσης της τάξεως του 5% και στις τρείς περιπτώσεις. Η επιρροή του μεταβλητού  $ThrVol$  σε ποσοστά ανίχνευσης και εσφαλμένων ειδοποιήσεων μπορεί να γίνει καλύτερα αντιληπτή στα σχήματα 6.2(i) και 6.2(ii). Στο πρώτο βλέπουμε καθαρά ότι τα ποσοστά ανίχνευσης φθίνουν γραμμικά με αυξανόμενο κατώφλι. Στο δεύτερο αυτή η συμπεριφορά αλλάζει δραματικά σε μία εκθετική παρακμή των εσφαλμένων ειδοποιήσεων, καλύπτοντας ταυτόχρονα και ένα πολύ μικρό μέρος της δραστηριότητας (0 σε περίπου 35, ενώ για τα πρόσωπα είναι 0 σε περίπου 400).

Σχετικά τώρα με το ποια είναι η τιμή του  $ThrVol$  που μας δίνει τον καλύτερο συμβιβασμό μεταξύ ποσοστών ανίχνευσης και εσφαλμένων ειδοποιήσεων, βρέθηκε ότι αυτή είναι η τιμή  $ThrVol = 22$ . Αυτή η τιμή μαρκαρίστηκε στις καμπύλες ROC, ευρισκόμενη περίπου στην αρχή της περιοχής ανταλλαγής. Η λογική πάνω στην συγκεκριμένη επιλογή μπορεί να γίνει πιο εύκολα κατανοητή στο σχήμα 6.3. Κάθε εσφαλμένη ειδοποίηση και κάθε μη εντοπισμένο πρόσωπο προκαλούμενο από κάποια συγκεκριμένη επιλογή του  $ThrVol$  χαρακτηρίστηκε ως ένα ‘σφάλμα’ του συστήματος (υπονοώντας ότι δίνουμε ίδιο βάρος σε κάθε

Σύνολο Δοκιμής	Κατώφλι Όγκου	Ανιχνευμένα Πρόσωπα	Ποσοστό Ανίχνευσης	Εσφαλμένες Ειδοποιήσεις	Ποσοστό Ακρίβειας
<b>CMU</b>	1,0	490	96,65%	1067	31,4%
	<b>22,0</b>	<b>456</b>	<b>89,95%</b>	<b>8</b>	<b>98,3%</b>
	33,4	439	86,6%	0	100%
<b>DiVAN</b>	1,0	102	98%	28	78,5%
	<b>22,0</b>	<b>96</b>	<b>92,3%</b>	<b>6</b>	<b>78,5%</b>
	49,8	91	87,5%	0	100%
<b>WEB</b>	1,0	494	99%	1279	27,9%
	<b>22,0</b>	<b>481</b>	<b>96,4%</b>	<b>8</b>	<b>98,4%</b>
	30,8	473	94,8%	0	100%
<b>Σινεμά</b>	1,0	272	98,55%	705	27,8%
	<b>22,0</b>	<b>252</b>	<b>91,3%</b>	<b>11</b>	<b>95,8%</b>
	41,8	231	83,7%	0	100%

Πίνακας 6.1

Κάποια στιγμότυπα των καμπύλων ROC για τα σύνολα δοκιμής CMU, DiVAN, WEB και Σινεμά. Γι κάθε ένα απ' αυτά, δίνονται αποτελέσματα για το καλύτερο μετρημένο ποσοστό ανίχνευσης ( $ThrVol = 1$ ), ένας καλός συμβιβασμός ( $ThrVol = 22$ ) και το ποσοστό ανίχνευσης χωρίς εσφαλμένες ειδοποιήσεις.

χαμένο πρόσωπο και σε κάθε εσφαλμένη ειδοποίηση). Στο σχήμα βλέπουμε τον αριθμό σφαλμάτων σε συνάρτηση των τιμών  $ThrVol$ . Μία πρώτη παρατήρηση είναι ότι βλέπουμε την ίδια συμπεριφορά και στις τρείς καμπύλες: ένας μεγάλος αριθμός σφαλμάτων στην αρχή, κυριαρχούμενος από τον υψηλό αριθμό εσφαλμένων ειδοποιήσεων σε αυτήν την περιοχή. Στο διάστημα περίπου [15, 35] οι καμπύλες βρίσκονται στα χαμηλότερα σημεία τους. Τέλος, ο αριθμός σφαλμάτων αρχίζει να αυξάνει αργά μετά από αυτήν την περιοχή, καθώς τώρα κάποια πρόσωπα χάνονται σταδιακά. Επιπλέον βλέπουμε με ευχαρίστηση ότι, κάθε καμπύλη έχει το ολικό της ελάχιστο περίπου στην ίδια θέση, η οποία αντιστοιχεί σε  $ThrVol = 22$  ακριβώς. Αυτό δεν είναι αυστηρά σωστό για το σύνολο Σινεμά, αλλά το ολικό του ελάχιστο είναι πολύ κοντά σε αυτό το σημείο.

Ο πίνακας 6.1 παρουσιάζει λεπτομερώς τα αποτελέσματα που προέκυψαν στα σύνολα δοκιμής για διάφορες τιμές του  $ThrVol$ . Οι πιο δεικτικές περιπτώσεις είναι τα καλύτερα μετρημένα<sup>3</sup> ποσοστά ανίχνευσης, ο συμβιβασμός  $ThrVol = 22$  και χωρίς εσφαλμένες ειδο-

<sup>3</sup> Είναι πολύ δύσκολος ο διαχωρισμός και η καταμέτρηση μεταξύ αληθών και εσφαλμένων ειδοποιήσεων του δικτύου, ειδικά όταν το  $ThrVol$  παίρνει πολύ μικρές τιμές προκαλώντας εκαποντάδες (εσφαλμένες) ειδοποιήσεις. Έτσι οι μετρήσεις άρχισαν με  $ThrVol = 1$ .

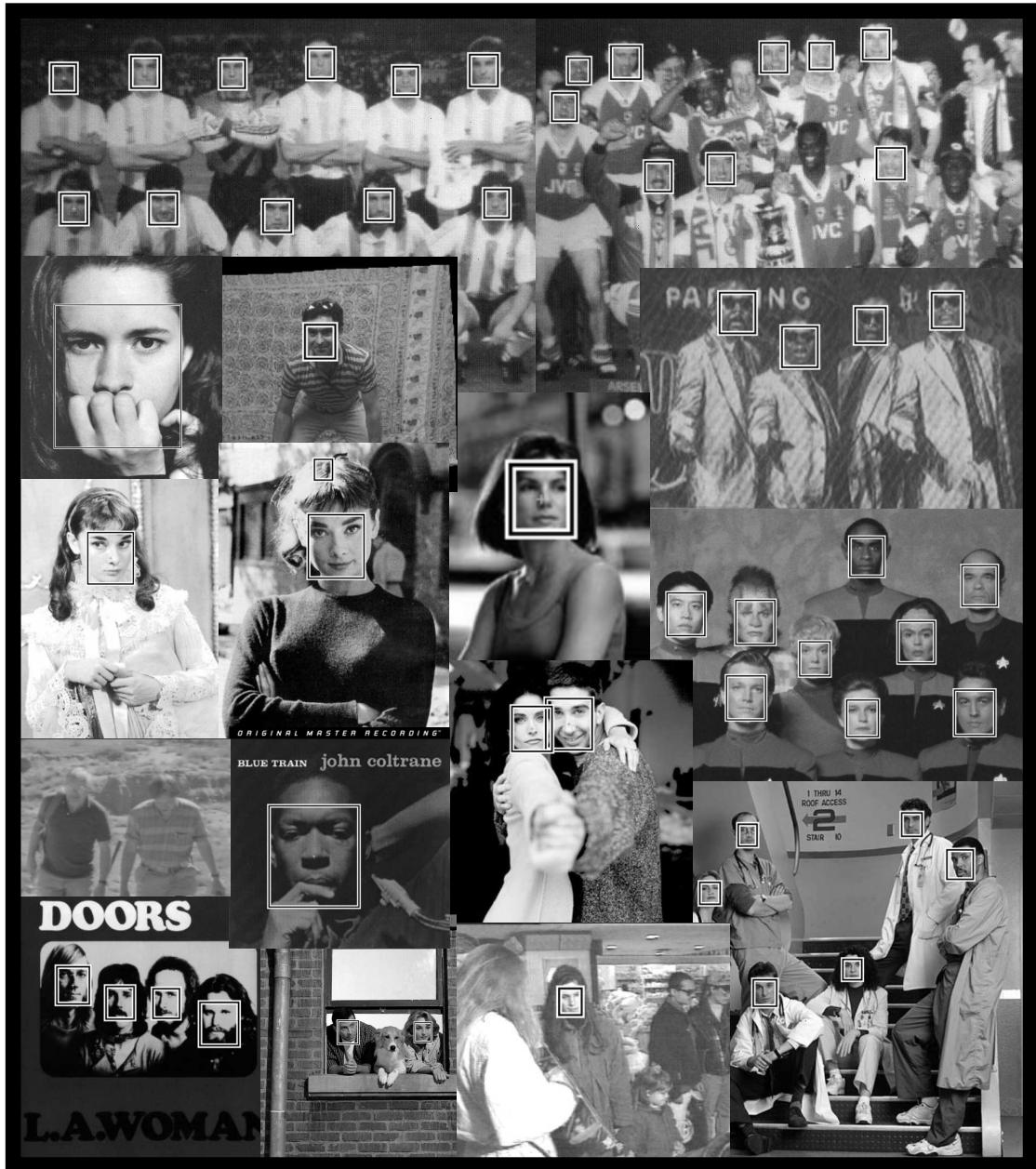
ποιήσεις. Στην πρώτη περίπτωση, τα ποσοστά ανίχνευσης πλησιάζουν το 100%, εκτός στην περίπτωση του CMU που το ανώτατο είναι 96,65%. Ο υψηλότερος αριθμός εσφαλμένων ειδοποιήσεων απαντήθηκε στα σύνολα WEB και CMU (1.279 και 1.067, αντίστοιχα), αλλά σημειώστε επίσης και τον αριθμό του συνόλου Σινεμά (705), το οποίο φέρει ένα μικρότερο αριθμό εικόνων, άρα διαθέτει λιγότερα παραθυρά προς εξέταση. Όταν  $ThrVol = 22$ , διατηρούμε ακόμα ένα πολύ υψηλό ποσοστό ανίχνευσης με ένα λογικό επίπεδο ρυθμού εσφαλμένων ειδοποιήσεων. Τα ποσοστά ανίχνευσης διαφέρουν από 90% σε 96%, υποδεικνύοντας την δυσκολία του καθ' ενός συνόλου. Τέλος, μπορούμε να παρατηρήσουμε ότι τα ποσοστά ανίχνευσης μειώνονται μόλις περίπου 3% όταν θέλουμε να απορρίψουμε όλες τις εσφαλμένες ειδοποιήσεις πάνω στα σύνολα CMU και WEB.

Η τελευταία στήλη του πίνακα 6.1 δίνει την ακρίβεια της δραστηριότητας, για κάθε σύνολο δοκιμής και για τα δεδομένα κατώφλια. Η ακρίβεια ορίζεται ως ο λόγος ανάμεσα στον αριθμό των επιτυχημένων ειδοποιήσεων και στον συνολικό αριθμό ειδοποιήσεων που καταμετρήθηκαν, επιτυχημένων ή εσφαλμένων. Αυτή η μετρική, λοιπόν, μας δίνει την πιθανότητα να είναι επιτυχημένη μία ειδοποίηση του συστήματος, η οποία, φυσικά, εξαρτάται από την συγκεκριμένη επιλογή κατωφλίου. Θεωρώντας την τιμή  $ThrVol = 22$ , βλέπουμε ότι με πιθανότητα πάνω από 94% θα λάβουμε ένα πραγματικό πρόσωπο από μία ειδοποίηση του δικτύου, επιπρόσθετα της πάνω από 90% πιθανότητας να εντοπίσουμε ένα πρόσωπο, όπως είδαμε προηγουμένως.

Κάποια από τα αποτελέσματα στα παραπάνω σύνολα δοκιμών δίνονται στα σχήματα 6.4, 6.5, 6.6 και 6.7. Τα αποτελέσματα αυτά παρήγαγαν με  $ThrVol = 22$ , αρχική και τελική κλίμακα αναζήτησης όπως ειπώθηκε για κάθε σύνολο στην ενότητα 6.1.1 και με λεπτή σάρωση για τον εντοπισμό υποψήφιων στόχων.

### 6.1.3 Σύγκριση στο Σύνολο CMU

Ο πίνακας 6.2 δίνει τα αποτελέσματα που έχουν αναφερθεί στην σχετική βιβλιογραφία ως τώρα στο σύνολο CMU και στα υποσύνολά του και τα συγκρίνει με τα αντίστοιχα της προτεινόμενης μεθόδου. Η σύγκριση στο σύνολο CMU ως όλον υποδεικνύει ότι η προτεινόμενη μέθοδος δίνει ποσοστά μεγαλύτερα κατά περίπου 4% και με ίδιο αριθμό εσφαλμένων ειδοποιήσεων ([2]) ή μικρότερο ([41]). Καλύτερη απόδοση βρίσκεται επίσης και στο σύνολο MIT ως όλον, σε σύγκριση με τις [33, 41, 48]. Αυτές οι συγκρίσεις δείχνουν ότι τα συνελικτικά νευρωνικά δίκτυα υπερέχουν όλων των άλλων νευρωνικών μεθόδων της βιβλιογραφίας. Όσον αφορά τα υποσύνολα CMU-125 και MIT-20, είναι δυσκολότερο να βγουν συμπεράσματα καθώς οι περισσότερες προσεγγίσεις του πίνακα δίνουν περίπου τα ίδια ή ελάχιστα καλύτερα αποτελέσματα σε σύγκριση με τα τις προτεινόμενης, αλλά όλα σε πολύ υψηλά ποσοστά ανίχνευσης (πάνω από 90%). Σε αυτό το σημείο πρέπει να σημειωθεί ότι οι εξαιρούμενες



## Σχήμα 6.4

Αποτελέσματα του συστήματος στο σύνολο CMU.



**Σχήμα 6.5**

Αποτελέσματα του συστήματος στο σύνολο DiVAN.

**Σχήμα 6.6**

Αποτελέσματα του συστήματος στο σύνολο WEB.

**Σχήμα 6.7**

Αποτελέσματα του συστήματος στο σύνολο Σινεμά.

Σύστημα Ανίχνευσης	CMU	CMU-125	MIT	MIT-20
Schneiderman <i>et al.</i> [45]		90,2%/110		
Yang <i>et al.</i> [55]		93,6%/74		91,5%/1
Roth <i>et al.</i> [38]		94,8%/78		94,1%/3
Rowley <i>et al.</i> [41]	86,2%/23		84,5/8	
Féraud <i>et al.</i> [3]	86,0%/8			
Colmenarez <i>et al.</i> [1]	93,9%/8122			
Sung <i>et al.</i> [48]			79,9%/5	
Osuna <i>et al.</i> [33]			74,2%/20	
<b>Προτεινόμενη Προσέγγιση</b>	<b>89,95%/8</b>	<b>89,2%/8</b>	<b>87,8%/4</b>	<b>86,5%/4</b>

**Πίνακας 6.2**

Σύγκριση στα σύνολα CMU, CMU-125, MIT and MIT-20, με βάση τα αναφερόμενα ποσοστά ανίχνευσης και αριθμού εσφαλμένων ειδοποιήσεων.

Σύστημα Ανίχνευσης	DiVAN
Garcia και Tziritas [9]	94,2%/20
Garcia και Delakis [7]	97,5%/3
Rowley <i>et al.</i> [41]	85,6%/9
<b>Προτεινόμενη Προσέγγιση</b>	<b>92,3%/6</b>

**Πίνακας 6.3**

Αποτελέσματα στο σύνολο δοκιμής DiVAN.

εικόνες περιέχουν πρόσωπα καρτούν ή χειρόγραφα<sup>4</sup>, άρα αυτές οι μέθοδοι δεν είναι σε θέση να χειριστούν παρόμοιες καταστάσεις. Έτσι, η γενικότητα τους είναι περιορισμένη, σε αντιπαράθεση με αυτήν της προτεινόμενης μεθόδου που μπορεί να χειριστεί πρόσωπα ανεξάρτητα της πηγής από την οποία προέρχονται (δηλ. μία κάμερα που βλέπει σε αληθινά πρόσωπα, σχεδιασμένα σε υπολογιστή ή ολοκληρωτικά σχεδιασμένα με το χέρι).

#### 6.1.4 Σύγκριση στο Σύνολο DiVAN

Τα αποτελέσματα στο σύνολο DiVAN παρουσιάζονται στον πίνακα 6.3 μαζί με αυτά των [7, 9, 41]. Οι Garcia και Delakis [7] παρουσίασαν μία προηγούμενη έκδοση του παρόντος συστήματος. Βασισμένη πάνω στην ίδια προσέγγιση, περιλαμβάνει ένα συνελικτικό δίκτυο, εκπαιδευμένο με τον ίδιο τρόπο bootstrapping και σαρώνει την εικόνα με την ίδια στρατηγική, εκτός κάποιων διαφορών δευτερεύουσας σημασίας. Η κύρια διαφορά είναι ότι

<sup>4</sup>Αν και κάποια άλλα πρόσωπα της ίδιας φύσης παρέμειναν στα σύνολα CMU-125 και MIT-20.

	-	268	254	244	226	209
<b>Λ4Α</b>	272/705	90	13	2	0	0
<b>ΠΙ4Α</b>	262/270	-	22	4	1	0
<b>ΠΙ5Α</b>	256/50	-	33	4	0	0
<b>Λ4</b>	268/836	836	12	3	1	0
<b>ΠΙ4</b>	255/75	-	56	26	9	4

**Πίνακας 6.4**

Σύγκριση μεταξύ διάφορων εναλλακτικών επιλογών αναζήτησης στο σύνολο Σινεμά. Η πρώτη στήλη υποδεικνύει την στρατηγική αναζήτησης. Η δεύτερη το σημείο από το σημείο άρχισαν οι μετρήσεις (χαμηλότερο κατώφλι). Οι υπόλοιπες στήλες δίνουν τις εσφαλμένες ειδοποιήσεις της κάθε στρατηγικής για ένα δεδομένο αριθμό ανιχνευμένων προσώπων από τα 276 που περιέχονται στο σύνολο Σινεμά.

σύστημα [7] εκπαιδεύτηκε με πρόσωπα περιστραμμένα ως  $\pm 20$  μοίρες και επίσης με ένα λιγότερο γενικευμένο σύνολο εκπαίδευσης, δηλ. με ένα μικρότερο αριθμό παραδειγμάτων εκπαίδευσης και χωρίς όλο το σύνολο των μετασχηματισμών που είδαμε στην ενότητα 4.2. Έτσι είναι ικανό να ανιχνεύει κάποια περισσότερα πρόσωπα που κατά κυριότητα είναι περιστραμμένα περισσότερο από 20 μοίρες και επίσης να δίνει λιγότερες εσφαλμένες ειδοποιήσεις η οποίες προκαλούνται στο παρόν σύστημα σαν συμπτώματα του δυσκολότερου συνόλου εκπαίδευσης. Τα συνελικτικά νευρωνικά δίκτυα έχουν και εδώ υψηλότερα ή πολύ υψηλότερα ποσοστά ανίχνευσης σε σύγκριση με την [41] που είδαμε στην σύγκριση στο σύνολο CMU, και επίσης με λιγότερες εσφαλμένες ειδοποιήσεις. Ένα άλλο ενδιαφέρον σημείο είναι ότι οι ικανότητες απόρριψης των συνελικτικών δικτύων είναι υψηλότερες σε σύγκριση με την [9], η οποία βασίζεται σε πληροφορία χρώματος για προφίλτραρισμα.

### 6.1.5 Σύγκριση Μεταξύ Διάφορων Στρατηγικών Αναζήτησης

Στα δύο προηγούμενα κεφάλαια είδαμε έναν αριθμό επιλογών σχετικά την στρατηγική αναζήτησης για την ανίχνευση και τον εντοπισμό των προσώπων. Αυτές οι επιλογές περιλαμβάνουν το ποια τοπολογία να χρησιμοποιήσουμε (ενότητα 4.4.2), την χρήση ή όχι του βήματος ακριβούς εντοπισμού και, τέλος, την επιλογή ανάμεσα στην λεπτή ή την

πρόχειρη σάρωση πριν το βήμα του ακριβούς εντοπισμού (ενότητες 5.1, 5.2). Ένας ευθύς τρόπος για να συγκρίνουμε την απόδοση που λαμβάνουμε από αυτές τις επιλογές είναι να υπολογίσουμε τις καμπύλες ROC και να δούμε ποια δίνει τα καλύτερα αποτελέσματα. Άλλα καθώς είναι δύσκολο να μετρήσουμε ποσοτικά την διαφορά που έχουν στην απόδοση, ειδικά όταν αυτή είναι μικρή, ένας πιο απλός τρόπος δίνεται στον πίνακα 6.4. Τα αποτελέσματα δίνονται στην μορφή δεδομένων ποσοστών ανίχνευσης έναντι του αριθμού εσφαλμένων ειδοποιήσεων που δίνει κάθε επιλογή, με άλλα λόγια μας δίνει το κόστος που απαιτείται για να βρούμε έναν συγκεκριμένο αριθμό προσώπων. Οι μετρήσεις έγιναν στο σύνολο Σινεμά, όντας μικρό σε αριθμό προσώπων και εικόνων αλλά, ταυτόχρονα, και αρκετά αντιπροσωπευτικό για την έκδοση ασφαλών συμπερασμάτων.

Η μέθοδος που χρησιμοποιήθηκε μέχρι στιγμής σε αυτό το κεφάλαιο δηλώνεται στον πίνακα ως ‘Λ4Α’, η οποία χρησιμοποιεί λεπτή σάρωση για την ανίχνευση στόχων (γράμμα ‘Λ’), την τοπολογία του σχήματος 4.1 (γράμμα ‘4’), και το βήμα ακριβούς εντοπισμού (γράμμα ‘Α’). Η δεύτερη επιλογή (Π4Α) είναι να χρησιμοποιήσουμε την πρόχειρη σάρωση για τον εντοπισμό προσώπων, η οποία είναι και πιο οικονομική. Με ‘Π5Α’ δηλώνεται η τοπολογία N5 της ενότητας 4.4.2, η οποία έχει περισσότερους χάρτες χαρακτηριστικών απ’ ότι η προτεινόμενη. Είδαμε ότι υπερείχε στην απόδοσή της στο σύνολο εκπαίδευσης και τώρα θα την αποτιμήσουμε και σε ένα σύνολο δοκιμής. Τέλος, οι δύο τελευταίες επιλογές (Λ4 και Π4) είναι οι Λ4Α και Π4Α χωρίς το τελικό βήμα ακριβούς εντοπισμού. Για την τελική κρίση πάνω στην παρουσία ή όχι ενός προσώπου χρησιμοποιήθηκε πάντα το ίδιο κριτήριο, δηλ. ο σύγκρισης δραστηριότητας (άθροισμα όλων των θετικών απαντήσεων γύρω από έναν στόχο).

Συγκρίνοντας την απόδοση που λαμβάνουμε αν χρησιμοποιήσουμε λεπτή ή πρόχειρη σάρωση (Λ4Α έναντι Π4Α και Λ4 έναντι Π4) βρίσκουμε ότι η πρώτη μέθοδος δίνει καλύτερα αποτελέσματα καθώς απαιτούνται λιγότερες εσφαλμένες ειδοποιήσεις για την ανίχνευση του ίδιου αριθμού προσώπων, όπως άλλωστε αναμενόταν. Με την πρόχειρη σάρωση της εισόδου είναι πιο δύσκολο να ανιχνευτούν κάποια πρόσωπα, ειδικά όταν αυτά βρίσκονται σε κάποιες ακραίες συνθήκες που το σύνολο Σινεμά διαθέτει. Έτσι απαιτείται ένα πιο χαλαρό κριτήριο (χαμηλότερο κατώφλι) για την ανίχνευση ίδιου αριθμού προσώπων με την λεπτή σάρωση, το οποίο ισοδυναμεί με έναν αυξημένο αριθμό εσφαλμένων ειδοποιήσεων. Παρ’ όλη την μικρή διαφορά στην απόδοση, η στρατηγική Π4Α πρέπει να θεωρηθεί ως μια καλή και πιο οικονομική εναλλακτική της Λ4Α, καθώς απαιτεί 5 φορές λιγότερο υπολογιστικό κόστος (ενότητα 5.4). Επιπλέον, δίνει και έναν πολύ μικρότερο αριθμό στόχων για ακριβή εντοπισμό, όπως βλέπουμε αν συγκρίνουμε την Π4 με την Λ4 καθώς κάθε θετική δραστηριότητα αυτών δίνει και από έναν στόχο.

Για την εκτίμηση του κέρδους του βήματος ακριβούς εντοπισμού, θα συγκρίνουμε την απόδοση των Λ4Α έναντι Λ4 και Π4Α έναντι Π4. Στην δεύτερη περίπτωση, το κέρδος αυ-

τού του βήματος είναι προφανές καθώς ο αριθμός εσφαλμένων ειδοποιήσεων της Π4 είναι αρκετά πιο μεγάλος απ' ότι της Π4A. Η Λ4 αποδίδει κοντά της Λ4A εκτός όταν το κριτήριο είναι υπερ-χαλαρωμένο, όπου και βλέπουμε μια σημαντική διαφορά στον αριθμό εσφαλμένων ειδοποιήσεων. Αυτό υπονοεί ότι η λεπτή σάρωση είναι ικανή να καλύψει μία σημαντική περιοχή του όγκου δραστηριότητας, αλλά πάντα έχουμε ένα μικρό κέρδος όταν ακολουθεί το βήμα ακριβούς εντοπισμού. Επιπλέον, βρέθηκε εμπειρικά ότι η ποιότητα εντοπισμού (ακριβής τοποθεσία και έκταση του προσώπου) βελτιώνεται σε πολλές περιπτώσεις.

Τέλος, συγκρίνοντας τον αριθμό εσφαλμένων ειδοποιήσεων των Π4A και Π5A, βρίσκουμε ότι οι δύο διαφορετικές τοπολογίες αποδίδουν πολύ κοντά, με την πρώτη να δίνει καλύτερα αποτελέσματα στις πρώτες στήλες και την δεύτερη στις τελευταίες. Πιο γενικά, φαίνεται καθαρά ότι η μεγαλύτερη τοπολογία σχημάτισε ένα πιο συντηρητικό όριο διαχωρισμού μεταξύ προσώπων και μη, καταλήγοντας σε μια υπερ-απορριπτική συμπεριφορά. Συνολικά, μπορούμε να την απορρίψουμε σαν μία βαρύτερη τοπολογία με ίδια ή χαμηλότερη απόδοση.

## 6.2 Ανάλυση Ευαισθησίας

Από τα αναλυτικά ποσοστά ανίχνευσης σε ποικίλες συνθήκες περιβάλλοντος έχουμε ήδη κάποια στοιχεία γύρω από τις ικανότητες σθεναρότητας του προτεινόμενου συστήματος. Σε αυτήν την ενότητα θα μελετήσουμε διεξοδικά πως επηρεάζεται η εξόδος του δικτύου από τις διακυμάνσεις (παραμορφώσεις) των προσώπων που συχνά συναντάμε σε περιβάλλοντα πραγματικού κόσμου. Ιστορικά, οι Rowley *et al.* [41] παρουσίασαν μία ανάλυση της εξόδου του δικτύου κάτω από την προσθήκη τοπικού θορύβου στην είσοδο. Αυτή η ανάλυση υπέδειξε ποιες είναι οι πιο σημαντικές και χρήσιμες για την ταξινόμηση περιοχές της εισόδου προσώπου, καταλήγοντας στις δύο περιοχές των ματιών και στην περιοχή του στόματος, με σειρά προτεραιότητας. Οι Féraud *et al.* [3] παρουσίασαν συνοπτικά την ευαισθησία του συστήματός τους σε σχέση με την μετατόπιση των προσώπων. Σε αυτήν την εργασία προτείνεται μία πιο διεξοδική ανάλυση, με την θεώρηση της ευαισθησίας του συστήματος σε σχέση με την περιστροφή των προσώπων εντός του επιπέδου της εικόνας, την θόλωση (blurring), την μεταβολή του κοντράστ, την προσθήκη λευκού θορύβου και, τέλος, την μεταβολή πόζας και έκφρασης των προσώπων.

Για τους σκοπούς αυτής της ανάλυσης, επιλέχτηκε ένα σύνολο από 20 εικόνες, καθεμία φέροντας από ένα πρόσωπο περίπου στο κέντρο. Όλα αυτά τα πρόσωπα ανήκουν στο σύνολο δοκιμής WEB και φαίνονται στο σχήμα 6.8. Μπορεί να παρατηρηθεί ότι αυτά τα πρόσωπα είναι περίπου μετωπικής θέας, είναι απο-περιστραμμένα και δεν βρίσκονται σε κάποια παθολογική κατάσταση. Κάθε μία από τις παρακάτω αναλύσεις έγινε με την εφαρμογή ενός

**Σχήμα 6.8**

Οι εικόνες που χρησιμοποιήθηκαν στην ανάλυση ευαισθησίας.

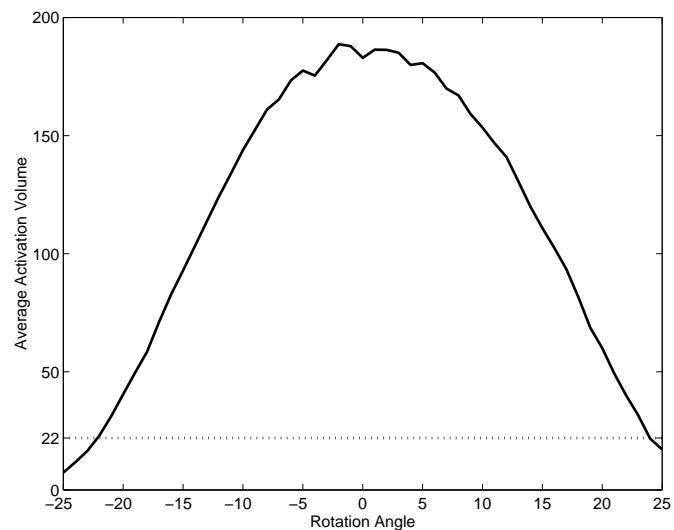
δεδομένου μετασχηματισμού σε κάθε μία από τις 20 εικόνες και την καταγραφή του μέσου όγκου δραστηριότητας που έδωσε το σύστημα χρησιμοποιώντας την στρατηγική αναζήτησης Λ4Α του πίνακα 6.4. Η περίπτωση της αλλαγής πόζας και έκφρασης αντιμετωπίστηκε ξεχωριστά, αφού είναι αδύνατο να μελετηθεί με τεχνητούς μετασχηματισμούς.

Η αντοχή στην περιστροφή μελετήθηκε περιστρέφοντας σταδιακά τις εικόνες σε γωνίες από -25 μοίρες σε +25 και με βήμα μίας μοίρας. Κάποια παραδείγματα του μετασχηματισμού αυτού δίνονται στο σχήμα 6.9(i). Κάθε εικόνα δόθηκε στον ανιχνευτή για επεξεργασία και καταγράφηκε ο τελικός μέσος όρος δραστηριότητας. Το σχήμα 6.9(ii) δίνει πως αυτή η μέση δραστηριότητα επηρεάζεται από την γωνία περιστροφής. Μπορεί να παρατηρηθεί ότι η καμπύλη έχει ένα καμπανοειδές σχήμα, με την μέση δραστηριότητα να μειώνεται με αργούς ρυθμούς από την μέγιστη τιμή της (γωνία περιστροφής μηδέν) σε τιμές μικρότερες του σημείου  $ThrVol = 22$  όταν η γωνία περιστροφής βρίσκεται στα όρια ( $\pm 25$  μοίρες). Από αυτό το γράφημα μπορούμε να δούμε ότι τι σύστημα ανέχεται γωνίες περιστροφής ως και  $\pm 20$  μοίρες, παρ' όλο που είχε εκπαιδευτεί με παραδείγματα περιστροφαμένα ως και  $\pm 10$  μοίρες.

Για την ανάλυση της επίδρασης της θόλωσης στην απόδοση του συστήματος ανίχνευσης, σε κάθε εικόνα εφαρμόστηκε ένα γκαουσσιανό φίλτρο εξομάλυνσης με τυπική απόκλιση να διαφέρει από 0,0 σε 2,5. Κάποια παραδείγματα του μετασχηματισμού αυτού δίνονται στο σχήμα 6.10(i). Τα αποτελέσματα της ανάλυσης δίνονται στο σχήμα 6.10(ii). Ο ανιχνευτής



(i)



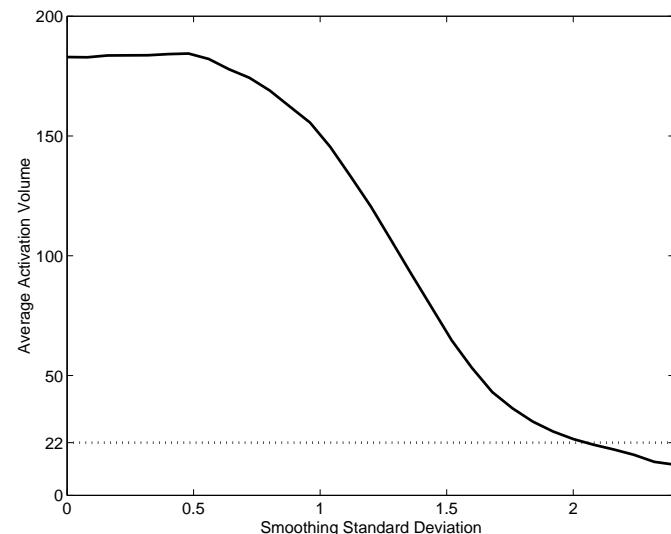
(ii)

### Σχήμα 6.9

Ανάλυση ευαισθησίας στην περιστροφή της εισόδου. (i) παραδείγματα μετασχηματισμών.  
(ii) επίδραση στον μέσο όγκο δραστηριότητας.



(i)



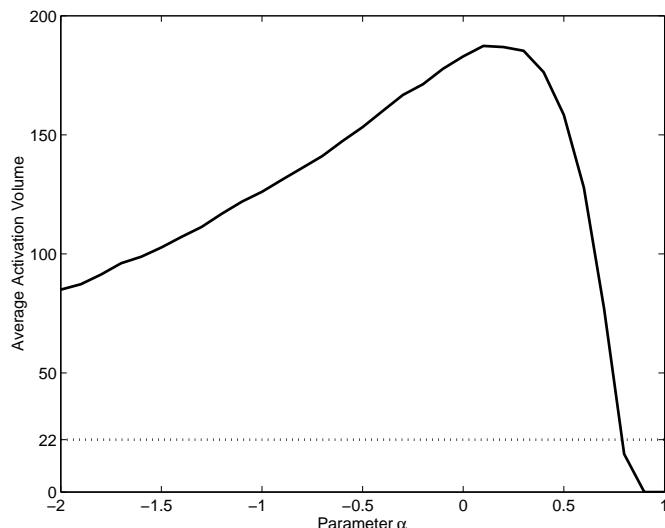
(ii)

**Σχήμα 6.10**

Ανάλυση ευαισθησίας στην θόλωση της εισόδου. (i) παραδείγματα μετασχηματισμών. (ii) επίδραση στον μέσο όγκο δραστηριότητας.



(i)



(ii)

### Σχήμα 6.11

Ανάλυση ευαισθησίας στην μεταβολή της φωτεινής αντίθεσης της εισόδου. (i) παραδείγματα μετασχηματισμών. (ii) επίδραση στον μέσο όγκο δραστηριότητας.

παράγει εξόδους που μειώνονται ομαλά καθώς η εικόνα θολώνεται βαθμιαία. Στην αρχής της καμπύλης, όταν τα αποτελέσματα της εξομάλυνσης είναι μαλακά, το σύστημα μένει ανεπηρέαστο. Αυτό το γεγονός μας υποδηλώνει ότι δεν θα έχουμε κάποιο κέρδος από άποψη ποσοστών ανίχνευσης, αν κάποια εξομάλυνση μικρής κλίμακας εφαρμοζόταν στην είσοδο του δικτύου. Στην συνέχεια, καθώς η τυπική απόκλιση του φίλτρου μεγαλώνει, παρατηρούμε μία πτώση της εξόδου του δικτύου μέχρι ενός σημείου σταθεροποίησης. Έτσι, μπορούμε να συμπεράνουμε ότι το δίκτυο δεν βασίζεται σε κάποια στοιχειώδη χαρακτηριστικά του προσώπου, όπως οι σκοτεινές κοιλάδες των ματιών και του στόματος κτλ., για την ταξινόμηση, αλλά χρησιμοποιεί κυρίως πληροφορία από την υφή του προσώπου. Τέλος, όταν τα πρόσωπα της δοκιμής εκφυλίζονται σε απλώς τρεις σκοτεινές κοιλάδες που αντιστοιχούν στα μάτια και στο στόμα, η εξόδος του συστήματος αρχίζει να είναι κάτω του σημείου  $ThrVol = 22$ .

Η ανοχή στις διακυμάνσεις της φωτεινής αντίθεσης της εικόνας μελετήθηκε με την με-

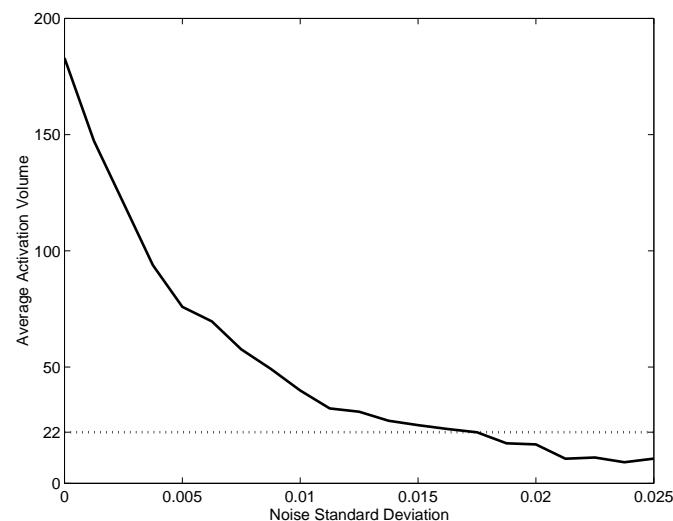
ταβολή της στις εικόνες δοκιμής. Σε κάθε εικόνα, η φωτεινότητα ενός εικονοστοιχείου  $I_p$  μεταβλήθηκε ως  $I_p = \alpha I_m + (1 - \alpha)I_p$ , όπου  $I_m$  είναι η μέση φωτεινότητα της εικόνας και  $\alpha$  είναι μία παράμετρος κυμαινόμενη από -2,0 σε 1,0. Όταν αυτή είναι μικρότερη του μηδενός, επιτυγχάνεται αύξηση της φωτεινής αντίθεσης, ενώ διαφορετικά επιτυγχάνεται η μείωσή της. Κάποια παραδείγματα του μετασχηματισμού αυτού δίνονται στο σχήμα 6.11(i). Οι παρενέργειες του πρώτου μετασχηματισμού είναι η παραγωγή καταστάσεων σαν να έχει εφαρμοστεί εξισορρόπηση ιστογράμματος ή όταν τα πρόσωπα είναι καρτούν. Οι παρενέργειες της μείωσης της αντίθεσης είναι μία εικόνα ‘μέσου όρου’, της οποίας το ιστόγραμμα συμπυκνώνεται γύρω από την μέση τιμή του. Παρατηρήστε ότι αυτός ο μετασχηματισμός είναι πολύ πιο ισχυρός της αύξησης της αντίθεσης, καθώς τα πρόσωπα ‘εξαφανίζονται’ γρήγορα από την εικόνα όταν το  $\alpha$  προσεγγίζει το 1,0. Το σχήμα 6.11(ii) δείχνει πως μεταβάλλεται η μέση δραστηριότητα σε σχέση με την τιμή του  $\alpha$ . Μπορούμε να δούμε ότι μειώνεται αρκετά αργά για αυξανόμενη βελτίωση αντίθεσης ( $\alpha < 0$ ), μένοντας πάντα σε τιμές αρκετά πάνω του σημείου  $ThrVol = 22$ . Όταν εφαρμόζεται μείωση φωτεινής αντίθεσης, η μέση δραστηριότητα μειώνεται απότομα λόγο του δυνατού εφέ αυτού του μετασχηματισμού. Πρέπει βέβαια να σημειωθεί ότι η μέση δραστηριότητα παραμένει ικανοποιητικά υψηλή (γύρω στο 40) για μείωση της αντίθεσης της τάξεως του 75%.

Για την μελέτη της επίδρασης του θορύβου, εφαρμόστηκε στις εικόνες λευκός γκαουσιανός θόρυβος με τυπική απόκλιση κυμαινόμενη από 0,0 σε 0,025. Κάποια παραδείγματα του μετασχηματισμού αυτού δίνονται στο σχήμα 6.12(i). Το σχήμα 6.12(ii) δίνει την μέση δραστηριότητα σε σχέση με την τυπική απόκλιση του θορύβου. Το σύστημα ανέχεται θόρυβο εικόνας μέχρι τα χαρακτηριστικά του προσώπου να καταστραφούν εντελώς με την απόκλιση μεγαλύτερη του 0,015, όπως φαίνεται και από τις δύο τελευταίες εικόνες των παραδειγμάτων.

Μία άλλη πτυχή της σθεναρότητας του συστήματος είναι η ανοχή στην έκφραση και στην πόζα του προσώπου. Για την επίδειξη της ανοχής του συστήματος στις διακυμάνσεις της πόζας, δόθηκαν για επεξεργασία στον ανιχνευτή όλα τα καρέ της γνωστής ακολουθίας MPEG Foreman. Περιέχει 250 καρέ μεγέθους  $288 \times 352$ , στα οποία ο ομιλητής διατρέχει μια έκρηξη αλλαγών σε πόζα και έκφραση. Το σχήμα 6.13(i) παρουσιάζει κάποια από τα επεξεργασμένα καρέ της ακολουθίας ενώ το σχήμα 6.13(ii) δίνει τον όγκο δραστηριότητας που επιτεύχθηκε σε κάθε καρέ. Η καμπύλη του σχήματος εξελίσσεται καθώς η πόζα και έκφραση μεταβάλλονται, επιδεικνύοντας μεγάλες μεταπτώσεις λόγο της εκρηκτικότητας του ομιλητή. Μπορεί κάποιος να παρατηρήσει, το οποίο βέβαια είναι και το πιο σημαντικό, ότι η εξόδος του ανιχνευτή είναι μεγαλύτερη του  $ThrVol = 22$  για την συντριπτική πλειοψηφία των καρέ, εκτός τριών υπακολουθιών (καρέ από 108 σε 111, 188-190 και από 231 σε 236). Σε αυτές τις τρεις περιπτώσεις, η πόζα του ομιλητή είναι πολύ κοντά σε ολικό προφίλ με αποτέλεσμα τα πρόσωπα να έχουν ανιχνευθεί με έναν μικρό, άρα και αναξιόπιστο, όγκο



(i)



(ii)

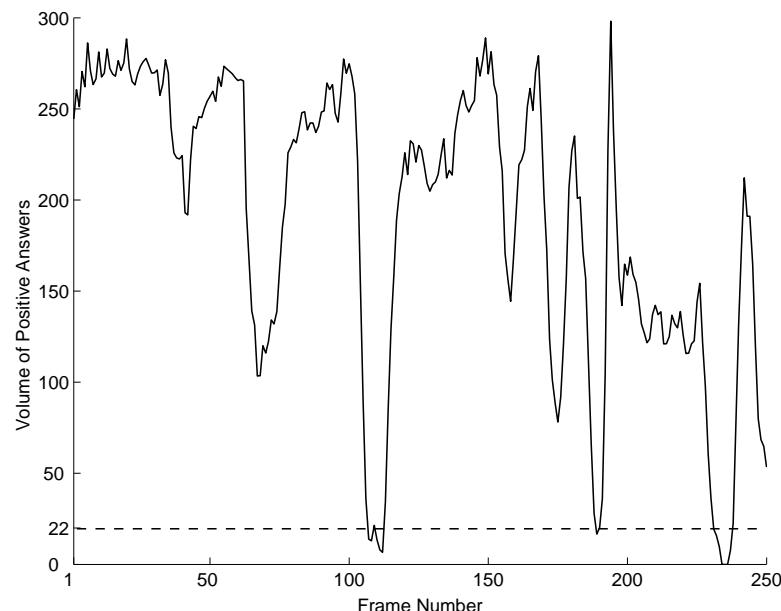
### Σχήμα 6.12

Ανάλυση ευαισθησίας στην προσθήκη θορύβου της εισόδου. (i) παραδείγματα μετασχηματισμών. (ii) επίδραση στον μέσο όγκο δραστηριότητας.

δραστηριότητας. Δεν ανιχνεύτηκαν καθόλου πρόσωπα (μηδενικός όγκος) στα καρέ 233 και 235, όπου το πρόσωπο είναι σε θέση ολικού προφίλ.



(i)



(ii)

**Σχήμα 6.13**

Απόδοση του συστήματος στην ακολουθία MPEG Foreman. (i) κάποια επεξεργασμένα καρέ. (ii) ο όγκος δραστηριότητας για κάθε καρέ.



# Συμπεράσματα

## 7.1 Ανασκόπηση των Ευρημάτων

Η υπόθεσή μας ότι τα Συνελικτικά Νευρωνικά Δίκτυα είναι μία πολλά υποσχόμενη προσέγγιση για την ανίχνευση προσώπων δοκιμάστηκε διεξοδικά και επιβεβαιώθηκε. Η προτεινόμενη συνελικτική τοπολογία του σχήματος 4.1 σχεδιάστηκε προσεκτικά έτσι ώστε να επιδεικνύει μία ωραία συμπεριφορά σταθερότητας στην μεταφορά, περιστροφή και σε άλλου είδους παραμορφώσεις της εισόδου. Αυτή η τοπολογία, έχοντας τα πρώτα στρώματά της να ενεργούν σαν εκπαιδεύσιμοι εξαγωγείς χαρακτηριστικών προσαρμοσμένοι ειδικά στις ανάγκες του συγκεκριμένου προβλήματος, δεν απαιτεί κάποια κανονικοποίηση ή προεπεξεργασία της εισόδου. Τα συνελικτικά νευρωνικά δίκτυα, με το να προσθέτουν δομή παραβάρη στο δίκτυο, μειώνουν τους βαθμούς ελευθερίας του ταξινομητή χωρίς να υπονομεύουν τις ικανότητές του. Και γνωρίζουμε από την θεωρία της Αναγνώρισης Προτύπων ότι οι μειωμένοι βαθμοί ελευθερίας οδηγούν αβίαστα σε αυξημένη ικανότητα γενίκευσης του ταξινομητή.

Αυτό το θεωρητικό εύρημα επιβεβαιώθηκε πειραματικά και σε πρώτη φάση κατά την διάρκεια εκπαίδευσης του δικτύου. Εφαρμοζόμενο πάνω σε ένα μεγάλο σύνολο εικόνων φόντου κατά την διάρκεια της ειδικής διαδικασίας εκπαίδευσης του πίνακα 4.2, το δίκτυο έμαθε γρήγορα να μην δίνει εσφαλμένες ειδοποιήσεις με υψηλής ή μεσαίας τάξης ενεργοποίηση (σχήμα 4.6), γενικεύοντας σωστά από τις αρχικές εσφαλμένες ειδοποιήσεις. Επιπλέον, η ειδική επιλογή των εσφαλμένων ειδοποιήσεων ενίσχυσε αυτήν την ικανότητα του δικτύου, βιοθώντας το να σχηματίσει γρήγορα ένα ακριβές όριο διαχωρισμού μεταξύ των προσώπων και μη. Η υπόθεση ότι οι πιο ισχυρές εσφαλμένες ειδοποιήσεις, φέροντας περισσότερη πληροφορία, θα πρέπει να επιλέγονται σε κάθε βήμα της διαδικασίας bootstrapping επιβεβαιώθηκε από την καλή συμπεριφορά σύγκλισης του δικτύου (σχήμα 4.6).

Το γεγονός ότι το δίκτυο είναι ικανό να λειτουργεί χωρίς προεπεξεργασία της εισό-

δου και, πιο σημαντικά, η ειδική συνελικτική/υποδειγματοληπτική δομή του επέτρεψαν στην κατασκευή μιας απλής και πολύ γρήγορης διαδικασίας σάρωσης της εικόνας. Βασίζεται ολοκληρωτικά σε απλές λειτουργίες φιλτραρίσματος της εικόνας που απαιτούν περίπου 400 (πρόχειρη σάρωση) ή 2.000 (λεπτή σάρωση) βασικές αριθμητικές πράξεις ανά εικονοστοιχείο της εικόνας. Αυτό το υπολογιστικό κόστος είναι μικρότερο ακόμα και από το κόστος της προεπεξεργασίας που απαιτείται από την πλειοψηφία της σχετικής βιβλιογραφίας. Χωρίς εξειδικευμένο λογισμικό ή hardware, απαιτείται περίπου 2,5 δευτερόλεπτα για την επεξεργασία ενός τυπικού καρέ βίντεο των  $352 \times 288$  εικονοστοιχείων σε έναν προσωπικό υπολογιστή με επεξεργαστή Pentium 4. Κάποιες απλές βελτιστοποιήσεις σε επίπεδο λογισμικού μπορούν να δώσουν ρυθμιούς επεξεργασίας της τάξης του ρυθμού βίντεο (παράρτημα A).

Δεδομένου ότι το δίκτυο εκπαιδεύτηκε με ένα μεγάλο σύνολο παραδειγμάτων προσώπων (3.702 αυθεντικά πρόσωπα, ενώ κλιμακώθηκε τεχνητά σε 25.212 παραδείγματα - κατά τις γνώσεις του συγγραφέα το μεγαλύτερο σύνολο εκπαίδευσης της βιβλιογραφίας) προερχόμενα κατευθείαν από φυσικά δεδομένα, η ίδια καλή συμπεριφορά εκπαίδευσης αναμενόταν και κατά την διάρκεια της φάσης δοκιμής. Επιπλέον, αυτή η φάση θα επιβεβαίωνε αποφασιστικά τις ικανότητες γενίκευσης του δικτύου. Η απόδοση του δικτύου αποτιμήθηκε σε μία σειρά από μεγάλα και δύσκολα σύνολα δοκιμής, περιέχοντας συνολικά 1.386 πρόσωπα, 607 εικόνες και μία μεγάλη ποικιλία πόζας, συνθηκών εικόνας και φόντου. Στον πίνακα 6.1 παρατηρούμε ότι τα ποσοστά ανίχνευσης σε όλα τα σύνολα είναι 90% ή περισσότερο και με μονοψήφιους αριθμούς εσφαλμένων ειδοποιήσεων. Αυτός ο πίνακας επίσης αποκαλύπτει και την καθολικότητα του προτεινόμενου συστήματος, το οποίο επιδεικνύει παρόμοια συμπεριφορά ανεξάρτητα των στατιστικών ιδιοτήτων του κάθε σύνολο δοκιμής. Επιπλέον, μπορούμε πρακτικά να απορρίψουμε όλες τις εσφαλμένες ειδοποιήσεις διατηρώντας πολύ υψηλά ποσοστά ανίχνευσης, της τάξεως του 85% ή περισσότερο. Οι συγκρίσεις στα σύνολα δοκιμής CMU (πίνακας 6.2) και DiVAN (πίνακας 6.3) υποδεικνύουν ότι το προτεινόμενο σύστημα είναι ο καλύτερης απόδοσης ανιχνευτής προσώπων γενικής χρήσεως της βιβλιογραφίας.

Καθώς το δίκτυο τροφοδοτείται με μη κανονικοποιημένα δεδομένα, κάποια πτώση της απόδοσης θα μπορούσε να σημειωθεί σε εξασθενημένη είσοδο (π.χ. υπερ-εξομαλυμένη ή αρκετά σκοτεινή). Η ανάλυση ευαισθησίας απέδειξε ότι η απόδοση του συστήματος πέφτει αργά σε μία σειρά από δυνατούς μετασχηματισμούς της εισόδου. Βρέθηκε ότι ο όγκος δραστηριότητας είναι πολύ πιο υψηλά από το κατώφλι που χαρακτηρίζει ή όχι την παρουσία ενός προσώπου ( $ThrVol = 22$ ) σε όλες τις περιπτώσεις εκτός των ακραίων των σχημάτων 6.10 ως 6.12 για τους μετασχηματισμούς θόλωσης, μεταβολής κοντράστ και προσθήκης θορύβου. Επιπρόσθετα, το σχήμα 6.9 υποδεικνύει ότι το σύστημα μπορεί να ανεχτεί με ασφάλεια περιπτώσεις περιστροφής των προσώπων ως και  $\pm 20$  μοίρες. Τέλος, το σύστημα φάνηκε

αξιοσημείωτα σταθερό στην αλλαγή έκφρασης προσώπου και πόζας στην γνωστή ακολουθία βίντεο Foreman (σχήμα 6.13).

## 7.2 Περιορισμοί

Όπως αναφέρθηκε και στην ενότητα 1.1, οι περιπτώσεις ολικού προφίλ και αυθαίρετης περιστροφής των προσώπων δεν λήφθηκαν υπ' όψη σε αυτήν την εργασία. Πρακτικά αλλά και με την βοήθεια της ανάλυσης ευαισθησίας, βρέθηκε ότι το σύστημα μπορεί να ανιχνεύσει με ασφάλεια πρόσωπα στραμμένα το πολύ κατά  $\pm 60$  μοίρες και περιστραμμένα το πολύ κατά  $\pm 20$  μοίρες. Κάτω από αυτές τις προϋποθέσεις, το σύστημα είναι εν δυνάμει ικανό να ανιχνεύσει οποιοδήποτε ανθρώπινο πρόσωπο σε οποιαδήποτε άλιμακα και τοποθεσία εντός μιας εικόνας, αν αυτό είναι ορατό και αντιληπτό χωρίς την βοήθεια αντίληψης υψηλού επιπέδου από έναν ανθρώπινο παρατηρητή.

## 7.3 Χώροι Εφαρμογής

Το προτεινόμενο σύστημα, όντας αρκετά ακριβή, γρήγορο και γενικής χρήσεως, μπορεί να χρησιμοποιηθεί σε κάθε δυνατή εφαρμογή της ανίχνευσης προσώπων. Κατά το καλύτερο των γνώσεων του συγγραφέα, η ανίχνευση προσώπων έχει δύο κύριες εφαρμογές: ανάκτηση εικόνας με βάση το περιεχόμενο (content-based image retrieval) και σαν ένα εργαλείο αυτόματης προετοιμασίας δεδομένων για την αναγνώριση του προσώπου ή την ανάλυση της έκφρασής του (ενότητα 1.1). Η πρώτη κατηγορία εφαρμογών απαιτεί την ανίχνευση προσώπων για το χαρακτηρισμό των εικόνων ως 'έχοντες' ή 'μη έχοντες πρόσωπα'. Έτσι, απαιτείται το καλύτερο δυνατό ποσοστό ανίχνευσης με τον αριθμό εσφαλμένων ειδοποιήσεων να παραμένει σε λογικά όρια. Επιπλέον, καθώς αναμένεται μία μεγάλη ποικιλία συνθηκών εικόνας και φόντου, απαιτείται ένας ανιχνευτής χωρίς περιορισμούς ή προϋποθέσεις. Πειραματικά, το προτεινόμενο σύστημα βρέθηκε να είναι ο καλύτερος δυνατός υποψήφιος για εφαρμογές τέτοιου είδους. Η άλλη κύρια εφαρμογή που χρησιμοποιεί την ανίχνευση προσώπων, αυτή της αναγνώρισης προσώπων ή έκφρασης, συνήθως προϋποθέτει μία καλή ποιότητα της εισόδου καθώς εργάζεται χυρίως σε ελεγχόμενα περιβάλλοντα. Ένας ανιχνευτής προσώπων λειτουργεί τότε περισσότερο σαν εντοπιστής προσώπων, με τα ποσοστά επιτυχών ή εσφαλμένων ειδοποιήσεων να μην είναι και τόσο μεγάλης σημασίας. Σε αυτήν την περίπτωση, το προτεινόμενο σύστημα είναι επίσης ένας φυσικός υποψήφιος, καθώς αποδείχτηκε αρκετά ακριβή όσον αφορά την εξαγωγή της περιοχής του προσώπου, ενώ είναι ταυτόχρονα και αρκετά γρήγορο χάρις την συνελικτική φύση του.

## 7.4 Σχετικές Δημοσιεύσεις

- C. Garcia and M. Delakis. Convolutional face finder: A neural architecture for fast and robust face detection. Submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- M. Delakis and C. Garcia. Training convolutional filters for robust face detection. To appear in *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, 2003.
- C. Garcia and M. Delakis. A neural architecture for fast and robust face detection. In *Proceedings of International Conference on Pattern Recognition*, volume 2, pages 44–48, 2002.
- M. Delakis and C. Garcia. Robust face detection based on convolutional neural networks. In *Proceedings of the 2nd Hellenic Conference on Artificial Intelligence*, pages 367–378, 2002.

## 7.5 Μελλοντικές Προεκτάσεις

Μία προέκταση του συστήματος ώστε να είναι σε θέση να χειριστεί πρόσωπα περιστραμμένα περισσότερο από 20 μοίρες ή πρόσωπα προφίλ βρέθηκε αρκετά χρήσιμη από πλευράς πρακτικής εφαρμογής. Πιστεύεται ότι η απόδοση του συστήματος δεν θα πέσει αισθητά για την χειρισμό αυτών των περιπτώσεων. Άλλωστε είναι αρκετά ενθαρρυντικό το γεγονός ότι είναι ήδη σε θέση να βρίσκει σποραδικά κάποια πρόσωπα προφίλ και επίσης ο μεγάλος βαθμός ανοχής που επιδεικνύει στην περιστροφή. Πρέπει να σημειωθεί εδώ ότι το υπάρχον πλαίσιο εργασίας είναι ήδη αρκετό για αυτήν την προέκταση: το μόνο που χρειάζεται να γίνει είναι η προσθήκη παραδειγμάτων προφίλ στο σύνολο εκπαίδευσης. Το ίδιο είναι αληθές, ίσως με την προσθήκη κάποιων επιπλέον αποφάσεων του μηχανικού, για την προέκταση των συνελικτικών δικτύων και σε άλλες περιπτώσεις ανίχνευσης αντικειμένων σταθερού σώματος, όπως η ανίχνευση κειμένου ή του ανθρώπινου σώματος.

Μία δεύτερη μελλοντική προέκταση αυτής της έρευνας μπορεί να είναι η επιπλέον διερεύνηση της διαδικασίας bootstrapping. Αν και βρέθηκε να δίνει αρκετά καλά αποτελέσματα από πρακτικής άποψης, η βελτίωσή της μπορεί να δώσει ίσως ακόμα καλύτερα αποτελέσματα. Επιπρόσθετα, η διαδικασία αυτή έχει σχέση με μία ειδική κατηγορία προβλημάτων της Αναγνώρισης Προτύπων, αυτής των προβλημάτων ‘μίας κλάσης’ και του ‘υπόλοιπου κόσμου’. Σε τέτοιου είδους προβλήματα, αλγόριθμοι βασισμένοι σε παραδείγματα (οπως ο backpropagation) φαίνεται να έχουν μία εγγενή δυσκολία της μοντελοποίησης του ‘υπόλοι-

που κόσμου'. Έτσι, ευρήματα προς αυτήν την κατεύθυνση μπορεί να έχουν μαζί θεωρητικές και πρακτικές αξίες.



## Επιδείξεις

Μία δημόσια προσβάσιμη επίδειξη του συστήματος είναι διαθέσιμη στον δικτυακό τόπο (σχήμα A.1(i)):

<http://www.csd.uoc.gr/~cgarcia/FaceDetectDemo.html>

Ο επισκέπτης μπορεί να αποστείλει μία εικόνα οποιασδήποτε μορφής και να δεί αμέσως τα αποτελέσματα ανίχνευσης σε αυτήν. Τα ανιχνευμένα πρόσωπα δίνονται ξεχωριστά, μαζί με τις αντίστοιχες τιμές του όγκου δραστηριότητας. Αυτά τα αποτελέσματα αποθηκεύονται σε σελίδες HTML, σχηματίζοντας χώρους έκθεσης από αποτελέσματα ανίχνευσης για οποιονδήποτε επιθυμεί να ελέγξει τα αποτελέσματα παρελθόντων αποστολών. Ως τώρα<sup>1</sup>, περίπου 3.300 εικόνες έχουν αποσταλεί και έχουν αποθηκευτεί στους χώρους έκθεσης. Ο ανιχνευτής ενεργοποιείται αυτόμata μέσω σκριπτ CGI, χωρίς καμία ανθρώπινη παρεμβολή. Η επεξεργασία όλων των εικόνων γίνεται με πρόχειρη σάρωση ακολουθούμενη από ακριβή εντοπισμό (μέθοδος Π4Α του πίνακα 6.4) και σε κλίμακες που αρχίζουν στα 36 εικονοστοιχεία και τελειώνουν στο ύψος της εικόνας (δηλ. πρόσωπα με ύψος εντός αυτών των ορίων δύναται να ανιχνευτούν).

Λόγο του πολύ μεγάλου όγκου αυτών των δεδομένων, ο οποίος μάλιστα αυξάνει διαρκώς, δεν υπάρχουν κάποιες μετρήσεις σχετικές των ποσοστών ανίχνευσης και του ρυθμού εσφαλμένων ειδοποιήσεων. Όσον αφορά μάλιστα το πρώτο, είναι πολύ δύσκολη κάποια αντικειμενική καταμέτρηση καθώς δεν υπάρχει κάποια αληθοφάνεια, καλά ορισμένη και δημόσια. Αυτό που είναι ιδιαίτερα ενδιαφέρων είναι ο αριθμός των εσφαλμένων ειδοποήσεων, όντας μία αντικειμενική μέτρηση. Ο αριθμός αυτός κυμαίνεται από 0 σε 2 ανά 50 εικόνες. Γενικά, τα αποτελέσματα που παρουσιάζονται στους χώρους έκθεσης λιγότερο ή περισσότερο επιβεβαιώνουν τις αντίστοιχες μετρήσεις μας πάνω στο σύνολο WEB, το οποίο αποτελείται φυσικά από εικόνες ακριβώς αυτών των χώρων έκθεσης.

<sup>1</sup>Τέλη Απριλίου του 2003.

**Submit an image to our Face Detector**



Image:  Browse...

**Some important notes:**

- The process of submission and the display of the results may take several seconds. It involves the image upload, the conversion to an appropriate format and the final HTML generation, apart from the actual execution of the detector.
- Images of quite large dimensions will be rejected for protecting the limited resources of a shared/multi-user machine. More precisely, if the width/height of the image is greater than 600,000 then it will be rejected.
- The online version of our system is **not able** to detect small faces (less than 30 pixel high), more than 20 degrees rotated faces, full profile faces (see examples below).

Examples: small:  rotated:  profile: 

\* Disclaimer: The owners of this site reject any responsibility regarding the content of the images of the gallery. The gallery will be checked periodically and images depicting explicit sexual contact or otherwise deemed pornographic or obscene will be removed.

(i)



(ii)

### Σχήμα A.1

Επιδείξεις του συστήματος. (i) δικτυακή επίδειξη. (ii) επίδειξη με κάμερα/βίντεο.

	<b>Adèle</b>	<b>Foreman</b>	<b>Κάμερα</b>
	512×304	352×288	320×240
<b>P4</b>	10,8 fps	14,2 fps	20 fps
<b>P4A</b>	7,9 fps	8,5 fps	14 fps

**Πίνακας Α.1**

Ρυθμοί επεξεργασίας (σε καρέ ανά δευτερόλεπτο - frames per second) σε περιβάλλοντα βίντεο. Σε όλες τις περιπτώσεις, τα αποτελέσματα την ανίχνευσης μπορούν να παρακολουθηθούν ζωντανά.

Μία άλλη μορφή επίδειξης του συστήματος αναπτύχθηκε με την μορφή επίδειξης μέσω κάμερας ή βίντεο (σχήμα A.1(ii)). Πιο συγκεκριμένα, το λογισμικό αυτό μπορεί να χειριστεί ζωντανό βίντεο που δίνεται από μία κάμερα ή ακολουθίες βίντεο MPEG, και τα δύο στον αέρα, καθώς εξελίσσεται η ακολουθία. Σχεδιάστηκε ειδικά για την επίδειξη του χαμηλού υπολογιστικού κόστους της μεθόδου, ακόμα και όταν αυτή εκτελείται από έναν υπολογιστή γενικής χρήσης των ημερών μας. Για αυτό τον σκοπό, μια σειρά από βελτιστοποιήσεις λογισμικού προστέθηκαν στην υλοποίηση. Αρχικά, η βιβλιοθήκη Image Processing Library<sup>2</sup>, η οποία διαθέτει μία σειρά από αλασικές λειτουργίες επεξεργασίας εικόνας βελτιστοποιημένες για τον επεξεργαστή Pentium III, χρησιμοποιήθηκε για να φέρει σε πέρας την διαμόρφωση των εικόνων και, φυσικά, τις συνελίξεις. Παρατηρήθηκε επιτόχυνση της τάξεως του 50% πάνω στο φιλτράρισμα της εικόνας με αυτήν την βιβλιοθήκη. Δεύτερον, παρατηρήθηκε μία αρκετά μεγάλη επιτάχυνση υπολογισμών όταν η συγμοειδής συνάρτηση που χρησιμοποιείται από τους νευρώνες αντικαταστάθηκε άμεσα από τον πίνακα τιμών της (look-up table). Τέλος, κάποιες άλλες τοπικές βελτιστοποιήσεις εφαρμόστηκαν, χυρίως στην χρήση της μνήμης. Αν και αυτή η υλοποίηση δεν μπορεί να θεωρηθεί σαν πρότυπη όσον αφορά την οικονομία και την βελτιστοποίηση του λογισμικού, μπορεί όμως να δώσει νύξεις για το τι ρυθμούς επεξεργασίας μπορούμε να επιτύχουμε. Πρέπει να σημειωθεί σε αυτό το σημείο ότι το λογισμικό επιτελεί ανίχνευση προσώπων σε κάθε καρέ ξεχωριστά, χωρίς να λαμβάνει υπ' όψη προηγούμενα ευρήματα (δηλ. δεν επιτελεί παρακολούθηση προσώπων ή κάποια παραλλαγή της).

Ο πίνακας A.1 παρουσιάζει τους ρυθμούς επεξεργασίας που επιτεύχθηκαν για δύο ακολουθίες βίντεο ('Adèle' και την γνωστή μας από την ενότητα 6.2, 'Foreman') όπως και επίσης και για ζωντανό βίντεο ('Κάμερα'). Οι αναφερόμενες μέθοδοι σάρωσης που υλοποιήθηκαν στο λογισμικό είναι η πρόχειρη σάρωση με (P4A) ή χωρίς (P4) ακριβή εντοπισμό (ενότητα 6.1.5). Οι κλίμακες αναζήτησης προσώπων αρχίζουν από τα 72 εικονοστοιχεία και τελειώ-

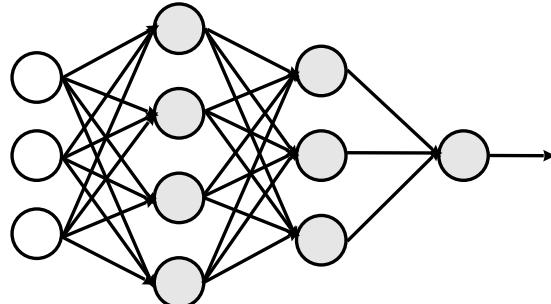
<sup>2</sup><http://www.intel.com/support/performancetools/libraries/ipl/>. Για την σύλληψη του βίντεο χρησιμοποιήθηκε η βιβλιοθήκη OpenCV <http://sourceforge.net/projects/opencvlibrary/>.

νουν στο ύψος του καρέ. Τα πειράματα έγιναν σε έναν επεξεργαστή Pentium 4 στα 1,7 GHz. Σε όλες τις περιπτώσεις, παρατηρούμε ρυθμούς επεξεργασίας της τάξεως του ρυθμού του βίντεο: περίπου στα 8 με 20 fps, ανάλογα της διάστασης του καρέ. Ο θεατής μπορεί να παρακολουθεί τα αποτελέσματα της ανίχνευσης χωρίς να αντιλαμβάνεται άμεσα την διαδικασία του φιλτραρίσματος των καρέ. Η διαφορά μεταξύ των ρυθμών των Π4 και Π4Α μπορεί να εξηγηθεί από το επιπλέον υπολογιστικό κόστος της εξαγωγής τμημάτων εικόνων και της διαμόρφωσης τους σε κατάλληλες κλίμακες που απαιτείται από την Π4Α. Το μεγάλο κενό ανάμεσα στους ρυθμούς επεξεργασίας του ζωντανού βίντεο και του βίντεο MPEG δύναται να οφείλεται στην επιβάρυνση των αναγνώσεων του αρχείου βίντεο από τον σκληρό δίσκο.

# Ο Αλγόριθμος Ανάστροφης Διάδοσης του Σφάλματος

Ο αλγόριθμος ανάστροφης διάδοσης του σφάλματος (error back-propagation algorithm, ή απλά *backpropagation*, όπως αναφέρεται σε αυτό το κείμενο) είναι ένα κανόνας μάθησης βασισμένος στην παράγωγο (gradient-based) για την εκπαίδευση των πολυστρωματικών νευρωνικών δικτύων εμπρόσθιας διάδοσης. Ένα παράδειγμα μιας τέτοιας τοπολογίας δίνεται στο σχήμα B.1. Το δίκτυο είναι οργανωμένο σε πολλαπλά στρώματα, με καθένα από αυτά να έχει έναν συγκεκριμένο αριθμό από μονάδες ή *neurons*. Η είσοδος του δικτύου παρέχεται από το στρώμα εισόδου. Στη συνέχεια, το σήμα διαδίδεται μέσω των συνδέσμων του δικτύου και υφίσταται επεξεργασία από τα *neurons* των στρώματα, μέχρι να φθάσει στο στρώμα εξόδου, όπου και υπολογίζεται η πραγματική έξοδος του δικτύου. Δεν υπάρχει καμία ανάδραση του σήματος προς τα πίσω, δηλ. το σήμα διαδίδεται αυστηρά από τα αριστερά προς τα δεξιά. Έχοντας ένα σύνολο παραδειγμάτων απεικονίσεων εισόδου-εξόδου, εκπαιδεύουμε το δίκτυο μέσω του αλγορίθμου με τελικό σκοπό να *παρεμβάλουμε* (interpolate) μεταξύ αυτών των παραδειγμάτων. Η διαδικασία εκπαίδευσης συνίσταται στην προσαρμογή των ελεύθερων παραμέτρων του δικτύου, ώστε να παράγεται από αυτό το επιθυμητό αποτέλεσμα. Αυτές οι ελεύθερες παράμετροι είναι τα *βάρη* του δικτύου, τα οποία είναι πραγματικοί αριθμοί που χαρακτηρίζουν την ισχύ των συνδέσμων του.

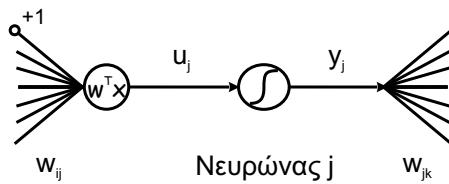
Ο αλγόριθμος *backpropagation* είναι μία προέκταση σε πολυστρωματικές τοπολογίες του κανόνα μάθησης του *Perceptron* [37], που αναπτύχθηκε για την εκπαίδευση του ταξινομητή *perceptron*. Γι' αυτό τον λόγο φέρουν και την εναλλακτική ονομασία “Πολυστρωματικά *Perceptrons*” (MultiLayer Perceptrons - MLPs). Ο αλγόριθμος εισάχθηκε αρχικά από τους Rumelhart *et al.* [43], αλλά επίσης διερευνήθηκε ανεξάρτητα και στην ίδια περίπου χρονική περίοδο από τον Le Cun [23]. Για μία πιο λεπτομερή και σφαιρική κάλυψη του αλγορίθμου,



Στρώμα Εισόδου Κρυμμένο Στρώμα Κρυμμένο Στρώμα Εξόδου

### Σχήμα B.1

Μία τυπική τοπολογία πολυστρωματικού, μη αναδραστικού νευρωνικού δικτύου.



### Σχήμα B.2

Ένας τυπικός μη γραμμικός νευρώνας.

προτείνεται το βιβλίο [13].

Η λειτουργία του backpropagation μπορεί να χωριστεί σε δύο φάσεις: στην εμπρόσθια και στην ανάστροφη μετάδοση. Στην πρώτη φάση, το σήμα διαδίδεται από αριστερά προς τα δεξιά για την παραγωγή της εξόδου του δικτύου. Είναι ισοδύναμη με την απλή ενεργοποίηση του δικτύου. Η λειτουργία ενός νευρώνα  $j$  (fig. B.1) που δέχεται σαν είσοδο το διάνυσμα  $x$  από το προηγούμενο στρώμα ή από το στρώμα εισόδου περιγράφεται από τις εξισώσεις:

$$u_j = \sum_i w_{ij} x_i \quad (\text{B.1})$$

$$y_j = \phi(u_j) \quad (\text{B.2})$$

$$\phi(u) = \begin{cases} (1 + \exp(-\alpha u))^{-1} \\ \hat{\eta} \\ \tanh(-\alpha u) \end{cases} \quad (\text{B.3})$$

όπου το  $w_{ij}$  αναφέρεται στην σύνδεση μεταξύ του νευρώνα  $i$  του προηγούμενο στρώματος και του νευρώνα  $j$ , το  $x_o = +1$  είναι σταθερό και το  $w_{oj}$  αποκαλείται *bias* αυτού του νευρώνα. Έχοντας υπολογίσει το εσωτερικό γινόμενο ανάμεσα στην είσοδο και του διανύσματος των βαρών, το αποτέλεσμα αυτό διέρχεται από μια σιγμοειδή συνάρτηση ενεργοποίησης  $\phi(u)$ , με

τυπικές επιλογές μία εκ των δύο περιπτώσεων της εξίσωσης B.3. Η παραμέτρος  $\alpha$  ωθείται στην αλίση της σιγμοειδούς. Τέλος, η εξόδος του νευρώνα  $y_j$  διαδίδεται στο επόμενο στρώμα, όπου και αυτή η διαδικασία συνεχίζεται επαναληπτικά μέχρι το στρώμα εξόδου.

Ο σκοπός της ανάστροφης μετάδοσης είναι η προσαρμογή των βαρών του δικτύου σύμφωνα με την φόρμουλα της καθόδου με βάση την αλίση (gradient descent):

$$\Delta w = -\rho \frac{\partial E}{\partial w} \quad (\text{B.4})$$

όπου  $\rho$  εκφράζει τον ρυθμό μάθησης και  $E$  είναι η συνάρτηση σφάλματος πάνω στο σύνολο εκπαίδευσης. Οι εξισώσεις που περιγράφουν την ανάστροφη μετάδοση ορίζονται ως εξής:

$$E = \frac{1}{2} \sum_o (d_o - y_o)^2 \quad (\text{B.5})$$

$$e_o = d_o - y_o \quad (\text{B.6})$$

$$\delta_o = \phi'(u_o) e_o \quad (\text{B.7})$$

$$\Delta w_{io} = \rho \delta_o y_i \quad (\text{B.8})$$

$$\delta_j = \phi'(u_j) \sum_k \delta_k w_{jk} \quad (\text{B.9})$$

$$\Delta w_{ij} = \rho \delta_j y_i \quad (\text{B.10})$$

όπου ο δείκτης  $o$  υποδηλώνει νευρώνα του στρώματος εξόδου, το  $d_o$  είναι η επιθυμητή εξόδος του αντίστοιχου νευρώνα και το  $w_{jk}$  είναι το βάρος της σύνδεσης του νευρώνα  $j$  με τον νευρώνα  $k$  του επόμενου στρώματος (fig. B.1). Ο αλγόριθμος εξελίσσεται με την παρουσίαση όλων των διανυσμάτων εισόδου στο δίκτυο πολλαπλές φορές και από την ακολουθούμενη προσαρμογή  $\Delta w$  μέχρι το σφάλμα να ελαχιστοποιηθεί σε ένα ικανοποιητικό επίπεδο, δηλ. μέχρι να παρατηρηθεί σύγκλιση.

Εκτός των παραπάνω εξισώσεων, που στην πράξη είναι απλό να υλοποιηθούν, ο αλγόριθμος backpropagation συμπεριλαμβάνει έναν μεγάλο αριθμό παραμέτρων ρυθμούς που ελέγχουν την συμπεριφορά του αλγορίθμου πάνω σε ένα συγκεκριμένο σύνολο εκπαίδευσης. Κάποια παραδείγματα τέτοιων παραμέτρων είναι η τιμή του ρυθμού μάθησης, η επιλογή της συνάρτησης ενεργοποίησης, η επιλογή τοπολογίας κτλ. Λόγω αυτού του μεγάλου αριθμού επιλογών του μηχανικού που απαιτούνται, μια αποτελεσματική εκπαίδευση μέσω του backpropagation “... μπορεί να θεωρηθεί περισσότερο τέχνη παρά επιστήμη” [27]. Σε αυτήν την εργασία, οι Le Cun *et al.* δίνουν μία σειρά από προτάσεις για το πως μπορεί να βελτιωθεί η απόδοση του backpropagation.



# Βιβλιογραφία

- [1] A.J. Colmenarez and T.S. Huang. Face detection with information-based maximum discrimination. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 782–787, 1997.
- [2] R. Féraud and O. Bernier. Ensemble and modular approaches for face detection: A comparison. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems 10*. MIT Press, 1998.
- [3] R. Féraud, O. Bernier, J.-E. Viallet, and M. Collobert. A fast and accurate face detector based on neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):42–53, 2002.
- [4] B. Frey, A. Colmenarez, and T. Huang. Mixtures of local subspaces for face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 32–37, 1998.
- [5] K. Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, 20:121–136, 1975.
- [6] K. Fukushima. Neocognitron: A model for visual pattern recognition. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.
- [7] C. Garcia and M. Delakis. A neural architecture for fast and robust face detection. In *Proceedings of International Conference on Pattern Recognition*, volume 2, pages 44–48, 2002.

- [8] C. Garcia, G. Simandiris, and G. Tziritas. A feature-based face detector using wavelet frames. In *Proceedings of International Workshop of Very Low Bit Coding*, pages 71–76, 2001.
- [9] C. Garcia and G. Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia*, 1(3):264–277, 1999.
- [10] C. Garcia, G. Zikos, and G. Tziritas. Wavelet packet analysis for face recognition. *Image and Vision Computing*, 18(4):289–297, 2000.
- [11] Z. Ghahramani and G. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto, 1996.
- [12] V. Govindaraju. Locating human faces in photographs. *International Journal of Computer Vision*, 19(2):129–146, 1996.
- [13] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1999.
- [14] R. Herpers, G. Verghese, K. Derpanis, R. McCready, J. MacLean, A. Jepson, and J. K. Tsotsos. Detection and tracking of faces in real environments. In *Proceedings IEEE International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems*, 1999.
- [15] E. Hjelmås and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83:236–274, 2001.
- [16] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706, 2002.
- [17] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962.
- [18] M. Hunke and A. Waibel. Face locating and tracking for human-computer interaction. In *28th Asilomar Conference on Signals, Systems and Computers*, 1994.
- [19] T. Kanade. *Picture Processing by Computer Complex and Recognition of Human Faces*. PhD thesis, Kyoto University, 1973.
- [20] C. Kotropoulos and I. Pitas. Rule-based face detection in frontal views. In *Proceedings of International Conference Acoustics, Speech and Signal Processing*, volume 4, pages 2537–2540, 1997.

- [21] S. Lawrence, C. Gilles, A. Tsai, and A. Back. Face recognition: A hybrid neural network approach. Technical Report UMIACS-TR-96-16, University of Maryland, 1996.
- [22] S. Lawrence, C. Gilles, A. Tsai, and A. Back. Face recognition: A convolutional neural network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997.
- [23] Y. LeCun. A theoretical framework for back-propagation. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proceedings of the 1988 Connectionist Models Summer School*, pages 21–28, CMU, Pittsburgh, Pa, 1988. Morgan Kaufmann.
- [24] Y. LeCun. Generalization and network design strategies. In R. Pfeifer, Z. Schreter, F. Fogelman, and L. Steels, editors, *Connectionism in Perspective*. Elsevier, Zurich, Switzerland, 1989.
- [25] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In David Touretzky, editor, *Advances in Neural Information Processing Systems 2 (NIPS\*89)*. Morgan Kaufman, Denver, CO, 1990.
- [26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.
- [27] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and Muller K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998.
- [28] C. H. Lee, J. S. Kim, and K. H. Park. Automatic human face location in a complex background. *Pattern Recognition*, 29:1877–1889, 1996.
- [29] C. C. Lin and W. C. Lin. Extracting facial features by an inhibitory mechanism based on gradient distributions. *Pattern Recognition*, 29:2079–2101, 1996.
- [30] B.K. Low. *Computer Extraction of Human Faces*. PhD thesis, De Montfort University, 1998.
- [31] B. Moghaddam and A. Pentland. Probabilistic visual learning for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.
- [32] E. Osuna, R. Freund, and F. Girosi. Support vector machines: Training and applications. Technical Report AI Memo 1602, Massachusetts Institut of Technology, 1997.
- [33] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997.

- [34] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.
- [35] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proceedings of Fourth IEEE International Conference on Computer Vision*, pages 84–91, 1994.
- [36] D. Reisfeld and Y. Yeshurun. Robust detection of facial features by generalised symmetry. In *Proceedings of 11th International Conference on Pattern Recognition*, pages A117–120, 1992.
- [37] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.
- [38] D. Roth, M.-H. Yang, and N. Ahuja. A SNoW-based face detector. In *Advances in Neural Information Processing Systems 12*, pages 855–861. MIT Press, 2000.
- [39] H. Rowley. *Neural Network-Based Face Detection*. PhD thesis, Carnegie Mellon University, 1999.
- [40] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–208, 1996.
- [41] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.
- [42] H. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 38–44, 1998.
- [43] D. Rumelhart, G. Hinton, and R. Williams. Learning representations of back-propagation errors. *Nature*, 323:533–536, 1986.
- [44] T. Sakai, M. Nagao, and T. Kanade. Computer analysis and classification of photographs of human faces. In *Proceedings First USA–Japan Computer Conference*, pages 2–7, 1972.
- [45] H. Schneiderman and T. Kanade. A statistical model for 3D object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 746–751, 2000.

- [46] K.-K. Sung. *Learning and Example Selection for Object and Pattern Detection*. PhD thesis, Massachusetts Institut of Technology, 1996.
- [47] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. Technical Report AI Memo 1521, Massachusetts Institut of Technology, 1994.
- [48] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.
- [49] J.-C. Terrillon, M. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 54–61, 2000.
- [50] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
- [51] R. Vaillant, C. Monrocq, and Y. LeCun. Original approach for the localisation of objects in images. *IEE Proceedings on Vision, Image, and Signal Processing*, 141(4):245–250, August 1994.
- [52] J. Wang and T. Tan. A new face detection method based on shape information. *Pattern Recognition Letters*, 21:463–471, 2000.
- [53] G. Yang and T. S. Huang. Human face detection in complex background. *Pattern Recognition*, 27(1):53–63, 1994.
- [54] J. Yang and A. Waibel. A real-time face tracker. In *IEEE Proceedings of the 3rd Workshop on Applications of Computer Vision*, pages 142–147, 2000.
- [55] M.-H. Yang, D. Kriegman, and N. Ahuja. Face detection using multimodal density models. *Computer Vision and Image Understanding*, 84:264–284, 2001.
- [56] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.
- [57] T. Yokoyama, Y. Yagi, and M. Yachida. Facial contour extraction model. In *Proceedings of IEEE Conference in Automatic Face and Gesture Recognition*, 1998.
- [58] K.C. Yow and R. Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15(9):713–735, 1997.

- [59] A. Yuille, P. Hallinan, and D. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, 1992.