ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Στερεοσκοπική όραση με χρήση τεχνητού νευρωνικού δικτύου

Εξεταζόμενος φοιτητής:
Βασίλειος Γκολεμης

Επιβλέπων καθηγητής:
Αναστάσιος Ντελοπουλος

Διπλωματική εργασία που υποβλήθηκε στα πλαίσια της ολοκλήρωσης του διπλώματος Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών

Ομάδα κατανόησης πολυμέσων
Εργαστήριο Επεξεργασίας Πληροφορίας
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

3 Νοεμβρίου 2017

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

Περίληψη

Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Δίπλωμα Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών

Βασίλειος Γκολέμης

Στερεοσκοπική όραση με χρήση τεχνητού νευρωνικού δικτύου

Η εξαγωγή πληροφορίας βάθους από ένα στερεοσκοπικό ζεύγος είναι ένα ανοικτό πρόβλημα της υπολογιστικής όρασης. Στην παρούσα εργασία προτείνεται η χρήση μηχανικής μάθησης και συγκεκριμένα τεχνητού νευρωνικού δικτύου για την ποιοτικότερη εύρεση των «αντίστοιχων σημείων» στις δύο λήψεις. Προσεγγίζουμε το πρόβλημα της επιλογής κατάλληλης τιμής παράλλαξης ως πρόβλημα ταξινόμησης πολλαπλών κατηγοριών και εκπαιδεύουμε το νευρωνικό δίκτυο σε κατάλληλα διαμορφωμένο σετ εκπαίδευσης. Μετά την αρχικοποίηση του πίνακα κόστους από το νευρωνικό δίκτυο, εφαρμόζουμε την στερεοσκοπική μέθοδο που συνίσταται στα εξής βήματα: άθροιση κόστους, ημικαθολική αντιστοίχηση, εντοπισμό κι αναπροσαρμογή εξωκείμενων τιμών. Αξιολογούμε τα αποτελέσματα της μεθόδου στις γνωστότερες στερεοσκοπικές συλλογές: KITTI 2012, KITTI 2015 και Middlebury.

ARISTOTLE UNIVERSITY OF THESSALONIKI

Abstract

Faculty of Engineering

Department of Electrical and Computer Engineering

Diploma of Electrical and Computer Engineering

Vasilis Gkolemis

Stereo vision using artificial neural network

Extracting depth information from a stereo pair is an open problem of the computer vision domain. In the present project, we propose a method for the accurate detection of corresponding points in the rectified views, using artificial neural network. We approach the problem of choosing the correct disparity of every point in the reference image as a multi-label classification problem and we train the neural net in a suitably configured training set. Initialization of cost matrix from the neural net is followed by the stereoscopic method which is composed of the following steps: cost aggregation, semi-global matching, detection and readjustment of outliers. The whole method is evaluated at well-known stereo datasets, such as KITTI 2012, KITTI 2015 and Middlebury.

Ευχαριστίες

Με την εργασία αυτή ολοκληρώνονται και τυπικά τα φοιτητικά μου χρόνια. Μακάρι όσα ακολουθούν να έχουν την ένταση και την αισιοδοξία τους.

Εκτός από τους εμφανείς πρωταγωνιστές αυτών των χρόνων, τους φίλους μου, οφείλω ένα ειλικρινές ευχαριστώ στα μέλη της οικογένειάς μου, Γιάννη, Χριστίνα, Νάγια, Βασιλική και Ματίνα γιατί προσφέρουν όσα μπορούν για να κάνω τη ζωή μου ενδιαφέρουσα, δίχως καν την απαίτηση να βρίσκονται στο προσκήνιό της.

Τέλος, ευχαριστώ τον κ. Ντελόπουλο γιατί υπήρξε συνεπής βοηθός και καλή παρέα τον τελευταίο χρόνο, όσο δηλαδή κράτησε αυτή η εργασία.

Περιεχόμενα

Στις όμορφες στιγμές παρέας ...

Κεφάλαιο 1

Εισαγωγή

Στερεοσκοπική όραση

Στερεοσκοπική όραση ονομάζουμε τη μέθοδο που ακολουθούν πολλοί ζωντανοί οργανισμοί, ανάμεσα τους και ο άνθρωπος, για την απόκτηση τρισδιάστατης αντίληψης του χώρου στον οποίο βρίσκονται. Η στερεοσκοπική όραση προϋποθέτει την συνεργασία ενός αισθητηρίου οργάνου που είναι υπεύθυνο για την πρόσληψη της ορατής πληροφορίας (οφθαλμοί) και ενός τμήματος επεξεργασίας υπεύθυνο για την παραγωγή ουσιωδών συμπερασμάτων (εγκέφαλος). Το οπτικό σήμα (ηλεκτρομαγνητικό κύμα) προσπίπτει στην ίριδα του ματιού, διαθλάται στον κρυσταλλοειδή φακό ώστε να προσανατολιστεί κατάλληλα και καταλήγει στον αμφιβληστροειδή χιτώνα όπου μετατρέπεται σε νευρικό παλμό, μέσω των ραβδίων και των κονίων. Οι νευρικοί παλμοί μεταβιβάζονται στον εγκέφαλο μέσω των νευρικών κυττάρων του οπτικού νεύρου, όπου θα ακολουθήσει η κατάλληλη επεξεργασία. Ο εγκέφαλος μελετά την πληροφορία που έχει συλλεχθεί από τον κάθε οφθαλμό, δηλαδή τις δύο αμφιβληστροειδικές εικόνες, και συγκρίνει την οριζόντια διαφορά των απεικονιζόμενων αντικειμένων ανάμεσα στις δύο αποτυπώσεις. Όσο μεγαλύτερη απόκλιση εμφανίζεται τόσο κοντύτερα στους οφθαλμούς βρίσκεται το απεικονιζόμενο αντικείμενο και τούμπαλιν. Έτσι, αποκτάται αίσθηση του βάθους.

Σε πλήρη αναλογία λειτουργεί η στερεοσκοπική όραση στα τεχνητά υπολογιστικά συστήματα. Το ανάλογο των οφθαλμών είναι δύο ψηφιακές κάμερες σε στερεοσκοπική διάταξη και του εγκεφάλου το λογισμικό που επεξεργάζεται την πληροφορία του στερεοσκοπικού ζεύγους.

Σε αντίθεση με τους ανθρώπους, για τα τεχνητά υπολογιστικά συστήματα η στερεοσκοπική όραση δεν αποτελεί μονόδρομο για την αντίληψη του χώρου που τα περιβάλλει. Έχουν αναπτυχθεί τεχνικές βασισμένες σε εργαλεία όπως το lidar[1] που απεικονίζουν άμεσα το τρισδιάστατο περιβάλλον, χωρίς την ενδιάμεση μετατροπή του σε δισδιάστατη πληροφορία (εικόνα) και την ακόλουθη ανακατασκευή του σε τρεις διαστάσεις. Αυτές οι τεχνικές όμως είναι ιδιαίτερα δαπανηρές και κατά περιπτώσεις μη εφαρμόσιμες, όπως για παράδειγμα όταν το υπό εξερεύνηση περιβάλλον είναι σε πολύ μεγάλη απόσταση (χαρτογράφηση περιοχών από αέρα, αποστολή STEREO της NASA κ.α.) ή περιέχει διαφανείς επιφάνειες (γυαλί, θάλασσα). Η στερεοσκοπική όραση παραμένει εξαιρετικά επίκαιρη μεθοδολογία, πολλές φορές σε αλληλοσυμπλήρωση με τις παραπάνω τεχνικές.

Η στερεοσκοπική όραση στα υπολογιστικά συστήματα χωρίζεται σε δύο μεγάλες υποκατηγορίες. Ενεργή στερεοσκοπική όραση (active stereo vision) ονομάζεται όταν επιλύει

[1] σύμπτυξη των λέξεων light και radar

το πρόβλημα με υποβοήθηση από κατάλληλα στοχευμένη εξωτερική πηγή δομημένου φωτός[2] και παθητική (passive stereo vision) σε αντίθετη περίπτωση. Στην παρούσα εργασία ασχολούμαστε με την παθητική.

Μηχανική μάθηση

Ως μηχανική μάθηση ορίζουμε τον τομέα της τεχνητής νοημοσύνης που δίνει τη δυνατότητα σε ένα υπολογιστικό σύστημα να μαθαίνει πως να περατώνει έναν σκοπό, χωρίς να έχει εκ των προτέρων προγραμματιστεί ρητά γι΄ αυτό.[3] Για την αυστηρότερη θεμελίωση του όρου «μαθαίνει», ο Tom Mitchell (1998) πρότεινε: «Όταν λέμε ότι ένα υπολογιστικό σύστημα μαθαίνει εννοούμε ότι από μια δεδομένη επίδοση P, με δεδομένη εμπειρία E σε ένα πρόβλημα T, η επίδοσή του P στο ίδιο πρόβλημα T βελτιώνεται καθώς αυξάνεται η εμπειρία του E». Ο όρος εμπειρία αναφέρεται στην ποσότητα παραδειγμάτων που έχει προσλάβει. Η διαδικασία της μάθησης ονομάζεται επιτηρούμενη ή επιβλεπόμενη (supervised) όταν το υπολογιστικό σύστημα δέχεται ως είσοδο παραδείγματα τα οποία εμπεριέχουν και την επιθυμητή έξοδο, λειτουργούν δηλαδή ως «δάσκαλος» και προσπαθεί μέσα από αυτά να δημιουργήσει έναν γενικό κανόνα πρόβλεψης κατάλληλης εξόδου ανάλογα με την είσοδο. Η μηχανική μάθηση χρησιμοποιείται όταν δεν είναι σαφές στον προγραμματιστή το «πως» ακριβώς περατώνεται ένας σκοπός.

Τα συνελικτικά νευρωνικά δίκτυα (convolutional neural networks) έχουν εμφανίσει εξαιρετικές επιδόσεις σε προβλήματα που λαμβάνουν εικόνα ως αρχική πληροφορία. Όπως περιγράφηκε παραπάνω, η στερεοσκοπική όραση απαιτεί την σύγκριση της σχετικής θέσης των προβεβλημένων αντικειμένων στις δύο λήψεις. Η σύγκρισή αυτή προϋποθέτει την απάντηση στο ερώτημα που βρίσκεται το ίδιο αντικείμενο στην κάθε λήψη. Η προσέγγιση του ερωτήματος με χρήση μηχανικής μάθησης αποδίδει αρκετά ποιοτικότερα αποτελέσματα.

Για την επίλυση του προβλήματος της υπολογιστικής στερεοσκοπικής όρασης συμπυκνώνεται γνώση από πολλά διαφορετικά επιστημονικά πεδία. Χαρακτηριστικά αναφέρουμε:

- **Φυσική:** Το φως είναι ηλεκτρομαγνητική ακτινοβολία συγκεκριμένου φάσματος[4] που διαδίδεται σε μορφή κύματος. Τα φαινόμενα που το περιγράφουν όπως η διάθλαση, η ανάκλαση και η διάχυση μελετώνται από την φυσική οπτική.
- **Μαθηματικά:** Η προοπτική γεωμετρία ορίζει και περιγράφει τον σχηματισμό της εικόνας. Η γεωμετρία πολλαπλών προβολών μελετάει τους περιορισμούς που εισάγονται κατά την προβολή του ίδιου τοπίου σε πολλές λήψεις. Η γραμμική άλγεβρα και ο λογισμός πολλών μεταβλητών περιγράφουν δομές μηχανικής μάθησης, όπως τα τεχνητά νευρωνικά δίκτυα, και επιλύουν αποτελεσματικά προβλήματα βελτιστοποίησης. Η στατιστική κι η επεξεργασία σήματος μελετούν τα χαρακτηριστικά της εικόνας κι αναζητούν τρόπους εξαγωγής χρήσιμων συμπερασμάτων.
- **Νευροεπιστήμες:** Το κύκλωμα διασυνδεδεμένων βιολογικών νευρώνων που αποτελεί τον νευρικό ιστό ενέπνευσε την δημιουργία των αντίστοιχων αλγορίθμων τεχνητών νευρωνικών δικτύων που αποτελούν βασικό εργαλείο της τεχνητής νοημοσύνης. Η τεχνητή νοημοσύνη αλληλεπιδρά αμφίδρομα με την ανθρώπινη, εμπνεόμενη από τους τρόπους με τους οποίους ο άνθρωπος λειτουργεί για να παράξει

[2] παραδείγματα τεχνικών δομημένου φωτός: Conventional structured-light vision (SLV), Conventional active stereo vision (ASV), Structured-light stereo (SLS)
[3] Ορισμός κατά τον Arthur Samuel (1959).
[4] Το φάσμα του ορατού φωτός κυμαίνεται στο διάστημα $[400nm, 700nm]$

συμπέρασμα, γνώση, αντίληψη και σκέψη, αλλά παρέχοντας ταυτόχρονα εργαλεία ώστε να μελετηθούν αναλυτικότερα οι τρόποι λειτουργίας του ανθρώπινου εγκεφάλου.

- **Επιστήμη Υπολογιστών:** Οι μέθοδοι αποτύπωσης της ηλεκτρομαγνητικής ακτινοβολίας σε ψηφιακή εικόνα, η ακόλουθη επεξεργασία της, η ανάπτυξη αλγορίθμων για την παραγωγή πληροφορίας και συμπερασμάτων από την ψηφιακή εικόνα, με ή χωρίς τη χρήση τεχνητής νοημοσύνης, η παραγωγή υλικού (αισθητήρες, επεξεργαστές, κάρτες γραφικών) για την ταχεία και ευσταθή περάτωση αυτών των αλγορίθμων είναι περιληπτικά κάποια από τα αντικείμενα της επιστήμης υπολογιστών.

1.1 Παλαιότερες προσεγγίσεις στο πρόβλημα της στερεοσκοπικής όρασης

Μέτρηση ομοιότητας χωρίων

Η στερεοσκοπική όραση απαιτεί την αναγνώριση του απεικονιζόμενου αντικειμένου στις δύο λήψεις. Για την επίτευξη αυτού του σκοπού, συγκρίνουμε κάθε σημείο (x, y) της εικόνας αναφοράς με όλα τα υποψήφια σημεία στα οποία μπορεί να αντιστοιχίζεται στην έτερη λήψη και αποθηκεύουμε μια τιμή ομοιότητας. Το σύνολο των μετρήσεων αποθηκεύονται σε έναν τρισδιάστατο πίνακα:

$$C(d, x, y) = \text{ομοιότητα}[\text{εικόνα\_αναφοράς}(x, y), \text{έτερη\_εικόνα}(x - d, y)]$$

Οι πρώτοι αλγόριθμοι που υλοποιήθηκαν χρησιμοποίησαν ως μετρική σύγκρισης το «άθροισμα των απόλυτων διαφορών» (sum of absolute differences) [1] [26], το «άθροισμα των τετραγώνων των διαφορών» (sum of square differences) [16] και την κανονικοποιημένη ετεροσυσχέτιση ή ομοιότητα συνημιτόνου [10]. Οι Zabih et. Woodfill (1994) [43] πρότειναν την μέθοδο Census που συγκρίνει τα περιφερειακά pixels του χωρίου με το κεντρικό, αποθηκεύοντας την τιμή 1 αν είναι φωτεινότερα και 0 αντίστροφα, δημιουργώντας έτσι μια δυαδική συμβολοσειρά από bits. Αυτή η συμβολοσειρά αποτελεί τον τοπικό περιγραφέα του χωρίου κι η μετρική σύγκρισης είναι ακολούθως η απόσταση Hamming των συμβολοσειρών. Οι Birchfield et. Tomasi (1998) [2] συγκρίνουν κάθε pixel της εικόνας αναφοράς με μια συνάρτηση γραμμικής παρεμβολής της έτερη εικόνας. Τέλος, οι Mei et. al (2011) [28] πρότειναν την μέθοδο AD-Census που συνδυάζει την πληροφορία από την σύγκριση μέσω «αθροίσματος απόλυτων διαφορών» και του μετασχηματισμού Census.

Τοπική άθροιση ομοιότητας

Τοπικές μέθοδοι άθροισης εξομαλύνουν τις αρχικοποιημένες μετρήσεις ομοιότητας του προηγούμενου βήματος. Η άθροιση ή ο υπολογισμός μέσης τιμής γίνεται σε υπολογισμένες περιοχές υποστήριξης, οι οποίες μπορεί να είναι είτε δισδιάστατες (χώρος (x, y)) είτε τρισδιάστατες (χώρος (d, x, y)). Οι πρώτες μέθοδοι που δοκιμάστηκαν εφάρμοζαν σε τετράγωνα χωρία φίλτρα μέσης τιμής. Οι Kanade et. Okutomi [15] και Kang et. Szeliski [17] πρότειναν την εφαρμογή μεταβλητών περιοχών υποστήριξης. Οι Zhang et. al [46] πρότειναν τον υπολογισμό αυτών των περιοχών υποστήριξης με τη μέθοδο του

σταυρού cross based cost aggregation, πετυχαίνοντας να μην εμπεριέχουν μεταβάσεις από ένα αντικείμενο σε ένα άλλο. Οι Scharstein et Szeliski [37] πρότειναν την μέθοδο iterative diffusion που υπολογίζει σταθμισμένους μέσους όρους εντός των περιοχών υποστήριξης επαναληπτικά.

Υπολογισμός χάρτη παράλλαξης

Σε αυτό το βήμα, με δεδομένο τον πίνακα C μεταβαίνουμε στον πίνακα D που περιέχει την οριζόντια μετατόπιση d (ονομάζεται παράλλαξη) κάθε σημείου ανάμεσα στις δύο λήψεις. Η προφανής επιλογή συνίσταται στην επιλογή της παράλλαξης που εμφανίζει την μεγαλύτερη ομοιότητα, δηλαδή την εφαρμογή της πράξης $D = argmax_d(C)$.[5] Η παραπάνω λογική ονομάζεται winner takes it all.

Έχουν προταθεί μέθοδοι που αντιμετωπίζουν τον πίνακα D ως πρόβλημα καθολικής βελτιστοποίησης επιχειρώντας να δημιουργήσουν έναν λείο χάρτη παράλλαξης D που λαμβάνει υπόψιν του τις τιμές ολόκληρου του πίνακα C. Αυτές οι μέθοδοι ορίζουν μια συνάρτηση ενέργειας:

$$E_C(D) = E_{\text{ομοιότητας}}(D) + \tau E_{\text{εξομάλυνσης}}(D)$$

κι ακολούθως επιχειρούν να βρουν τον πίνακα παράλλαξης D που ελαχιστοποιεί την τιμή του E_C.

Ο όρος $E_{\text{ομοιότητας}}(D)$ «προτιμά» τις παραλλάξεις με τις καλύτερες τιμές ομοιότητας:

$$E_{\text{ομοιότητας}}(D) = \sum_{\mathbf{p}} C(\mathbf{p}, D(\mathbf{p}))$$

ενώ ο όρος $E_{\text{εξομάλυνσης}}(D)$ «προτιμά» την επιλογή ίδιων ή κοντινών τιμών παράλλαξης ανάμεσα σε γειτονικά σημεία:

$$E_{\text{εξομάλυνσης}}(D) = \sum_{\mathbf{p}} \sum_{\mathbf{q} \in N_p} g(D(\mathbf{p}) - D(\mathbf{q}))$$

όπου g μια γνησίως αύξουσα συνάρτηση.

Η καθολική εύρεση του ελάχιστου στο παραπάνω πρόβλημα, είναι υπολογιστικά αδύνατη. Προτάθηκαν μέθοδοι που αντιμετωπίζουν το πρόβλημα με πιθανοτικά γραφικά μοντέλα, όπως για παράδειγμα Markov Random Fields. Οι Boykov et. al (2001)[3] και Kolmogorov et al. (2001) [20] αντιμετώπισαν το πρόβλημα με την μέθοδο graph cuts , ενώ οι Felzenszwalb et. al (2006) [6] πρότειναν την μέθοδο belief propagation. Ο Heiko Hirscmuller (2008) πρότεινε την μέθοδο Semi-Global Matching (SGM) [12] που βρίσκει το ελάχιστο κατά μήκος 16 προκαθορισμένων κατευθύνσεων μέσω δυναμικού προγραμματισμού. Έπειτα, υπολογίζει τον μέσο όρο των 16 ελαχίστων, ως το ολικό ελάχιστο.

[5] Αν ο πίνακας ^ μετρούσε αντίθεση (κόστος) αντί για ομοιότητα, η αντίστοιχη πράξη θα ήταν $D = argmin_d(C)$.

Χρήση μηχανικής μάθησης

Πριν την δημιουργία συλλογών στερεοσκοπικών δεδομένων με πληροφορία χάρτη παράλλαξης, λίγες προσεγγίσεις χρησιμοποιούσαν εκπαιδεύσιμα μοντέλα για την στερεοσκοπική αντιστοίχηση. Οι Kong, Tao (2004) [21] εκπαίδευσαν ένα μοντέλο που αντιστοιχούσε σε κάθε αρχικό υπολογισμό ομοιότητας μια πιθανότητα: η πρόβλεψη να είναι σωστή, η πρόβλεψη να είναι λάθος λόγω αντικειμένου στο προσκήνιο, η πρόβλεψη να είναι λάθος για οποιονδήποτε άλλο λόγο. Έπειτα ακολουθούσε κατάλληλη επεξεργασία ανάλογα με την κατηγορία που ανήκε η κάθε πρόβλεψη.

Οι Zhang, Seitz (2007) [47], οι Scharstein, Pal (2007) [34] και οι Li, Huttenlocher (2008) [22] χρησιμοποίησαν μοντέλα για την εκμάθηση των βέλτιστων παραμέτρων των αντίστοιχων πιθανοτικών μοντέλων που χρησιμοποίησαν[6].

Οι Haeusler et al (2003) [9] χρησιμοποίησαν έναν ταξινομητή random forest για την αξιολόγηση της ευστάθειας των αρχικών προβλέψεων ομοιότητας, ενώ οι Spyropoulos et. al (2014) [39] χρησιμοποίησαν αυτές τις αξιολογήσεις για την ρύθμιση των παραμέτρων του markov random field που εφάρμοσαν στη συνέχεια. Σε αντίστοιχη λογική οι Park and Yoon (2015) [31] χρησιμοποίησαν εκτιμήσεις ποιότητας των αρχικών προβλέψεων για την ρύθμιση του Semi-global matching.

1.2 Προσεγγίσεις παρακείμενες στην προτεινόμενη μέθοδο

Οι Zbontar and Lecun (2016) [45] χρησιμοποίησαν ένα βαθύ νευρωνικό δίκτυο για την σύγκριση τετράγωνων περιοχών υποστήριξης κι αρχικοποιώντας κατά αυτό τον τρόπο τον πίνακα ομοιότητας. Εκπαίδευσαν το νευρωνικό δίκτυο σε τετράγωνα χωρία διάστασης $[9, 9]$ pixels.Ακολούθως, εφάρμοσαν αρκετές τεχνικές για την βελτίωση των αποτελεσμάτων (όπως άθροιση κόστους, semi-global matching κ.α.). Οι μέθοδοι τους πέτυχαν κορυφαία αποτελέσματα για μεγάλο χρονικό διάστημα.

Οι Luo et. al (2016) [24] ανέπτυξαν μια μέθοδο παρόμοια με αυτή των Zbontar and Lecun με δύο βασικές διαφορές. Εκπαίδευσαν το νευρωνικό δίκτυο με μεθόδους ταξινόμησης πολλαπλών κατηγοριών (το αντίστοιχο των Zbontar and Lecun είχε εκπαιδευτεί σε δυαδική ταξινόμηση), ενώ βελτίωσαν έντονα τους χρόνους εκτέλεσης επιλέγοντας αρχιτεκτονική λιγότερων παραμέτρων.

Οι Gidaris and Komodakis (2016)[8] υπολόγισαν έναν αρχικό χάρτη παράλλαξης μέσω του νευρωνικού δικτύου που εκπαίδευσαν οι Luo et. al κι ακολούθως χρησιμοποίησαν νέα επεξεργασία μέσω νευρωνικού δικτύου για την βελτίωση του αρχικού χάρτη παράλλαξης. Ουσιαστικά, προσπάθησαν να αντικαταστήσουν τις κλασικές μεθόδους άθροισης κόστους, semi-global matching, κλπ με μηχανική μάθηση. Το συνολικό μοντέλο τους χωρίζει το συνολικό πρόβλημα υπολογισμού του χάρτη παράλλαξης σε τρία μικρότερα υποπροβλήματα, καθένα εκ των οποίων λύνεται μέσω νευρωνικών δικτύων.

Τέλος, οι Kendall et .al (2017) [18] εκπαίδευσαν ένα βαθύ νευρωνικό δίκτυο πολλών παραμέτρων, προσεγγίζοντας με μηχανική μάθηση το σύνολο της διαδικασίας υπολογισμού του χάρτη παράλλαξης. Η προσέγγισή τους αποτελεί αυτή τη στιγμή το state-of-the-art στην στερεοσκοπική συλλογή KITTI.

[6] markov random field και conditional random field

1.3 Προτεινόμενη μέθοδος

Στην παρούσα εργασία, αρχικά επιχειρούμε μια ποιοτική ανάλυση των αρχών και των περιορισμών της στερεοσκοπικής όρασης. Ακολούθως, επιλέγουμε συγκεκριμένα σημεία των παραπάνω μεθόδων και τα συνδυάζουμε σε μια ενιαία μεθοδολογία. Αναλύουμε σε ποιες βασικές στερεοσκοπικές αρχές βασίζεται η κάθε μεθοδολογία.

Αντιμετωπίζουμε το πρόβλημα της αρχικοποίησης του πίνακα ομοιότητας ως πρόβλημα ταξινόμησης πολλαπλών κατηγοριών. Κάθε σημείο της εικόνας αναφοράς οφείλει να κατηγοριοποιηθεί σε μια εκ του συνόλου των πιθανών οριζόντιων μετατοπίσεων. Εκπαιδεύουμε ένα βαθύ νευρωνικό δίκτυο ώστε να συγκρίνει τετράγωνα χωρία διάστασης $[19, 19]$ pixels και να υπολογίζει μια πιθανοτική κατανομή ομοιότητας.

Στη συνέχεια, εφαρμόζουμε διαδοχικά βήματα ώστε να βελτιώσουμε τις αρχικές προβλέψεις ομοιότητας. Χρησιμοποιούμε την μέθοδο άθροισης κόστους σε περιοχές υποστήριξης υπολογισμένες με την μέθοδο του σταυρού (cross-based cost aggregation), προσπαθώντας να δημιουργήσουμε περιοχές υποστήριξης που θα περιορίζονται στην επιφάνεια ενός αντικειμένου. Έτσι, αποφεύγουμε την άθροιση πληροφορίας σε περιοχές μεταβάσεων που δημιουργούν μεγάλα σφάλματα.

Εφαρμόζουμε επίσης την μέθοδο semi-global matching με δυναμικό προγραμματισμό, περιορίζοντας τις διευθύνσεις βελτιστοποίησης σε 2[7], ώστε να μειωθεί η υπολογιστική πολυπλοκότητα και να επιταχυνθεί η εκτέλεση.

Τέλος εφαρμόζουμε τις μεθόδους outlier detection και subpixel enhancement, όπως προτάθηκαν στην εργασία των Mei et al [28].

Αξιολογούμε το σύνολο της μεθόδου στις γνωστότερες συλλογές στερεοσκοπικών δεδομένων KITTI και Middlebury. Αποδεικνύουμε ότι η ακρίβεια της μεθόδου οφείλεται στην χρήση του νευρωνικού δικτύου και τις συγκρίσεις ομοιότητας που αυτό υλοποιεί και όχι στην ακόλουθη επεξεργασία. Η ακόλουθη επεξεργασία, αν και βελτιώνει αισθητά τα αποτελέσματα, είναι πολύ πιο αδύναμη αν δεχτεί ως είσοδο τον παραγόμενο πίνακα C μιας συμβατικής μεθόδου σύγκρισης, όπως της μέσης απόλυτης διαφοράς. Αποδεικνύουμε επίσης ότι το νευρωνικό δίκτυο με κατάλληλη εκπαίδευση καταφέρνει να μάθει έναν ποιοτικό κανόνα σύγκρισης, που έχει επιτυχία ακόμα κι αν εφαρμοστεί σε εικόνες διαφορετικής στατιστικής από αυτές που εκπαιδεύτηκε.

1.4 Δομή εργασίας

Στο κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο της στερεοσκοπικής μεθόδου. Αναλύονται οι βασικές αρχές στις οποίες βασίζεται η στερεοσκοπική μέθοδος, οι οποίες αποδεικνύονται μέσω της στερεοσκοπικής γεωμετρίας. Παρουσιάζονται οι υποθέσεις που κάνουμε για την στερεοσκοπική αντιστοίχηση και μελετώνται οι περιπτώσεις όπου αυτές οι υποθέσεις αίρονται. Γίνεται ανάλυση όλων των μεθόδων επεξεργασίας που χρησιμοποιούμε μετά την αρχικοποίηση του πίνακα ομοιότητας κι αναλύουμε σε ποιες υποθέσεις βασίζεται η κάθε μέθοδος.

Στο κεφάλαιο 3 επιλύουμε το πρόβλημα αρχικοποίησης του πίνακα ομοιότητας με χρήση μηχανικής μάθησης, μέσω τεχνητού νευρωνικού δικτύου. Αναλύουμε την αρχιτεκτονική

[7] Επιλύουμε το πρόβλημα σε 2 διευθύνσεις και 2 φορές ανά διεύθυνση. Έτσι έχουμε συνολικά 4 διαφορετικές λύσεις από τις οποίες προκύπτει ο μέσος όρος.

του δικτύου που χρησιμοποιείται, γιατί επιλέχθηκε αυτή η αρχιτεκτονική, πως δημιουργούμε το σετ εκπαίδευσης και τις ακριβείς παραμέτρους εκπαίδευσης του δικτύου.

Στο κεφάλαιο 4 παρουσιάζονται τα αποτελέσματα της μεθόδου.

Τα κεφάλαια 2 και 3 έχουν τα αντίστοιχα παραρτήματά τους στο τέλος της εργασίας. Τα παραρτήματα αξιοποιούνται κυρίως για την απόδειξη των μαθηματικών σχέσεων και την αναλυτική παράθεση αλγοριθμικών και προγραμματιστικών τεχνικών. Επιλέξαμε αυτή την τακτική ώστε η γραφή μας εντός των κεφαλαίων να παραμένει προσηλωμένη στον κεντρικό στόχο κάθε ενότητας και να μην πλατειάζει στην ανάλυση ή την απόδειξη του κάθε εργαλείου που χρησιμοποιούμε. Παρ΄ όλα αυτά, για την σφαιρική κατανόηση των μεθόδων προτείνεται στον αναγνώστη να ανατρέχει στα παραρτήματα όπου υπάρχουν αναφορές σε αυτά.

Κεφάλαιο 2

Θεωρητικό μέρος - Στερεοσκοπική όραση

2.1 Σχηματισμός εικόνας μέσω της προοπτικής προβολής

Έστω χώρος σημείων A. Εάν σε αυτόν προσδιορίσουμε μια γραμμικώς ανεξάρτητη βάση V και μια αρχή $o \in A$, τότε έχει οριστεί ένα σύστημα συντεταγμένων του χώρου σημείων $A = (o, V)$. Υπάρχει ισομορφισμός μεταξύ του $A = (o, V)$ και του συνόλου $\mathbb{R}^n$. Στην περίπτωση του τρισδιάστατου φυσικού κόσμου ο ισομορφισμός ικανοποιείται με το σύνολο $\mathbb{R}^3$. Πλέον, κάθε σημείο $p = o \bigoplus v$ του χώρου A, δηλαδή του τρισδιάστατου κόσμου, μπορεί να περιγραφεί μονοσήμαντα μέσω 3 συντεταγμένων $\mathbf{P} = (X, Y, Z)$ που είναι οι συντεταγμένες του v ως προς την προαναφερθείσα συγκεκριμένη βάση.

Η εικόνα αποτελεί την θέαση του τρισδιάστατου κόσμου σε δύο διαστάσεις. Μαθηματικά είναι μια συνάρτηση

$$f : \mathbb{R}^3 \to \mathbb{R}^2$$

η οποία μετασχηματίζει τις τρισδιάστατες συντεταγμένες του τυχαίου σημείου $\mathbf{P} = (X, Y, Z)$ στις $\mathbf{p} = (x, y)$. Η πιο γνωστή μέθοδος προβολής είναι η προοπτική και με αυτήν θα ασχοληθούμε στην παρούσα εργασία.

Η προοπτική προβολή συνίσταται στην «1-1» αντιστοίχηση (mapping)των σημείων του χώρου στο πέτασμα της κάμερας αν μιλάμε για ένα τεχνητό οπτικό μέσο ή στον χιτώνα του ματιού αν εξετάζουμε το οπτικό μέσο του ανθρώπου. Η αντιστοίχηση υλοποιείται μέσω του μοντέλου «κάμερας μικρής οπής» (pinhole model).

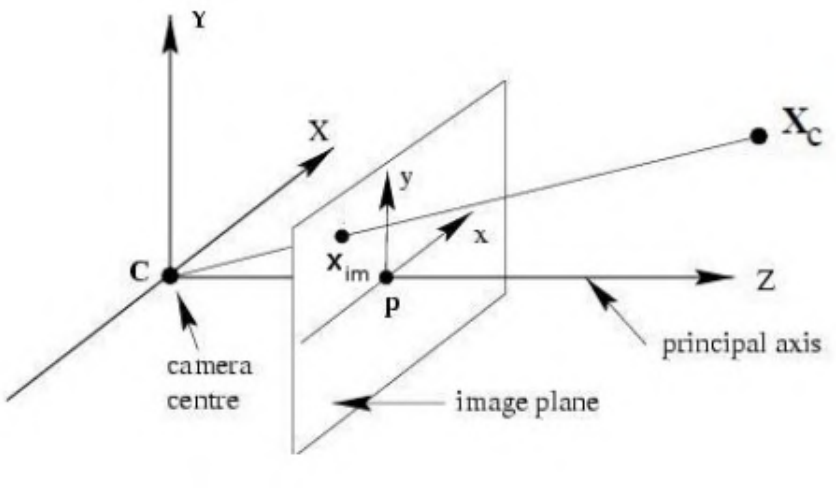
Η f_{pr} περιγράφεται ποιοτικά από την παρακάτω μη γραμμική συνάρτηση:[1]

$$\mathbf{p} = f_{pr}(\mathbf{P}) : (x, y) = \left(f\frac{X}{Z}, f\frac{Y}{Z} \right) \tag{2.1}$$

η οποία εκφράζεται και σε μορφή πίνακα με την χρήση ομογενών συντεταγμένων:

$$Z \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2.2}$$

[1] Πληρέστερη μοντελοποίηση της προοπτικής προβολής δίνεται στο παράρτημα Α′.1



Σχημα 2.1: Γεωμετρική απεικόνιση της προοπτικής προβολής σημείου.

Σχημα 2.2: Η ανάκτηση του βάθους είναι μια δύσκολη υπόθεση

Η προοπτική προβολή είναι *μη* αντιστρέψιμη πράξη:

$$\nexists f_{pr}^{-1} : \{\mathbb{R}^2 \to \mathbb{R}^3\} : \mathbf{P} = f_{pr}^{-1}(\mathbf{p}) \Leftrightarrow (X, Y, Z) = f_{pr}^{-1}(x, y)$$

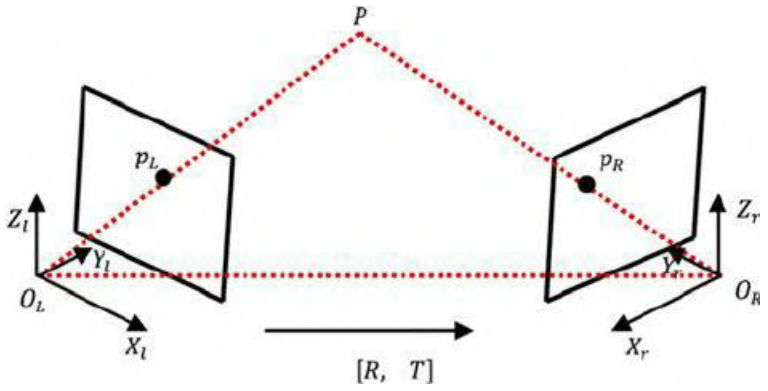
Πρακτικά αυτό σημαίνει ότι για κάθε σημείο $\mathbf{p}$ του πετάσματος της κάμερας είναι αδύνατος ο προσδιορισμός του σημείου $\mathbf{P}$ του τρισδιάστατου χώρου από το οποίο προήλθε. Γνωρίζουμε ότι οι υποψήφιες θέσεις του σημείου $\mathbf{P}$ κινούνται στην ευθεία που ενώνει το οπτικό κέντρο O της κάμερας με το σημείο $\mathbf{p}$. Διανυσματικά η ευθεία αυτή περιγράφεται από την έκφραση:[2]

$$\mathbf{l} = (X, Y, Z) = t \cdot \left(\frac{x}{f}, \frac{y}{f}, 1\right), t \in [f, +\infty) \tag{2.3}$$

Ερμηνεύοντας το αποτέλεσμα αντίστροφα, οποιοδήποτε σημείο $\mathbf{P}$ που ανήκει στην ευθεία $\mathbf{l}$ θα προβληθεί στο ίδιο ακριβώς σημείο $\mathbf{p}$ του πετάσματος.

Όπως παραστατικά αποτυπώνεται στην εικόνα 2.2 και μαθηματικά στην εξίσωση 2.3, δεν αρκεί μια εικόνα για την ακριβή ανάκτηση της αρχικής τρισδιάστατης πληροφορίας.

[2]Πληρέστερη έκφραση της ευθείας δίνεται στο παράρτημα Α΄.2



Σχημα 2.3: Τριγωνοποίηση

2.2 Γεωμετρία πολλαπλών προβολών

Στην γεωμετρία πολλαπλών προβολών διαθέτουμε $n \geq 2$ λήψεις της ίδιας τρισδιάστατης σκηνής. Θα μας απασχολήσει η περίπτωση των $n = 2$ λήψεων.

2.2.1 Προσδιορισμός τρισδιάστατης θέσης σημείου από δύο λήψεις

Υποθέτουμε ότι:

1. Ως σύστημα αναφοράς έχουμε ορίσει το σύστημα συντεταγμένων της μία εκ των δύο λήψεων. Την λήψη αυτή την ονομάζουμε λήψη αναφοράς και στην παρούσα εργασία επιλέγουμε να είναι η αριστερή.

2. Γνωρίζουμε την ακριβή θέση και τον προσανατολισμό της έτερης (δεξιάς) λήψης στο χώρο. Συγκεκριμένα για τον πλήρη προσδιορισμό μιας λήψης ως προς μια άλλη χρειάζονται ένας πίνακα περιστροφής $R \in \mathbb{R}^{3\times3} : \| R \| = 1$ και ένας πίνακας μετατόπισης $T \in \mathbb{R}^3$. Αυτοί οι δύο πίνακες μαζί συνθέτουν έναν μετασχηματισμό affine, που τον συμβολίζουμε ως $g = (R, T)$. Εάν γνωρίζουμε τις συντεταγμένες ενός τυχαίου σημείου P ως προς ένα ορισμένο σύστημα συντεταγμένων (στην περίπτωσή μας της αριστερής κάμερας), μπορούμε μέσω του μετασχηματισμού affine g να βρούμε τις συντεταγμένες του P ως το σύστημα συντεταγμένων της δεξιάς κάμερας.

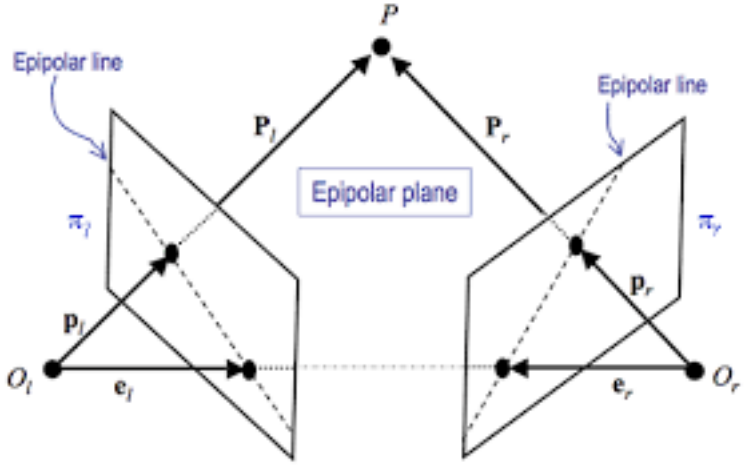
3. Γνωρίζουμε τις ακριβείς θέσεις $\mathbf{p_l}, \mathbf{p_r} \in \mathbb{R}^2$ στις οποίες έχει προβληθεί το *ίδιο* σημείο του τρισδιάστατου χώρου $\mathbf{P}$.

Με χρήση της σχέσης 2.3, προκύπτουν οι τρισδιάστατες *ομοεπίπεδες* ευθείες l_1 και l_2 που αναλογούν στα σημεία $\mathbf{p_l}$ και $\mathbf{p_r}$. Το σημείο τομής των δύο ευθειών είναι η αρχική θέση $\mathbf{P}$ των σημείων $\mathbf{p_l}, \mathbf{p_r}$. Η διαδικασία αυτή, που φαίνεται στην εικόνα 2.3, ονομάζεται τριγωνοποίηση (triangulation).

2.2.2 Το πρόβλημα της αντιστοίχησης

Η μέθοδος της τριγωνοποίησης προϋποθέτει την γνώση των σημείων $\mathbf{p_l}$ και $\mathbf{p_r}$[3]. Επομένως για την ανάκτηση της τρισδιάστατης θέσης ενός τυχαίου σημείου $\mathbf{p_l}$ απαιτείται η

[3]στο εξής θα αναφέρονται ως «αντίστοιχα σημεία»



Σχημα 2.4: Επιπολικό επίπεδο και επιπολική ευθεία

ανεύρεση του αντίστοιχου σημείου του $\mathbf{p_r}$ στην έτερη λήψη. Αυτή η αναζήτηση, που ονομάζεται «πρόβλημα αντιστοίχησης» (correspodence problem), είναι το κυρίαρχο πρόβλημα προς επίλυση σε κάθε εφαρμογή ανακατασκευής τρισδιάστατης πληροφορίας.

Η αναζήτηση του «αντίστοιχου σημείου» δεν γίνεται κατά τυχαίο τρόπο στο σύνολο της έτερης λήψης. Αντίθετα η γεωμετρία των δύο προβολών περιορίζει την αναζήτησή του κατά μήκος μιας συγκεκριμένης ευθείας που βρίσκεται στο πέτασμα της έτερης λήψης. Η ευθεία αυτή ονομάζεται «επιπολική ευθεία».

Αρχικά ορίζεται ο πίνακας essential matrix ως:

$$E = [T]R \in \mathbb{R}^{3\times 3}$$

Ως $[T]$ συμβολίζουμε τον 3×3 πίνακα εξωτερικού γινομένου του διανύσματος T. Ο essential matrix υπολογίζεται άμεσα εφόσον γνωρίζουμε τον affine μετασχηματισμό $g = (R, T)$ και συμπυκνώνει την γεωμετρική συσχέτιση των σημείων $\mathbf{p_l}$ και $\mathbf{p_r}$ καθώς ικανοποιεί την σχέση:[4]

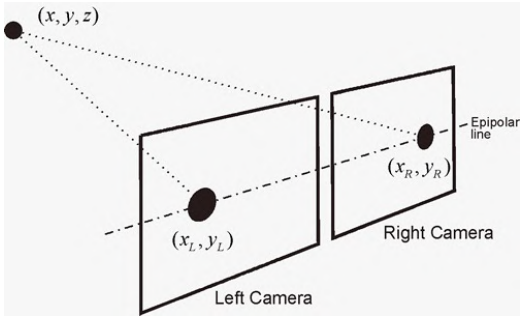
$$\mathbf{p_r}^T E \mathbf{p_l} = 0 \tag{2.4}$$

Η σχέση 2.4 ονομάζεται στερεοσκοπικός περιορισμός και από αυτήν προκύπτουν οι «επιπολικές ευθείες» κατά μήκος των οποίων βρίσκονται οι υποψήφιες θέσεις του «αντίστοιχου σημείου». Συγκεκριμένα εάν συμβολίσουμε με έναν πίνακα-γραμμή $l = [a\ b\ c]$ μια τυχαία ευθεία του επιπέδου $\mathbb{R}^2$[5] τότε ισχύει:

- το «αντίστοιχο σημείο» του $\mathbf{p_r}$ θα βρίσκεται πάνω στην επιπολική ευθεία $l_l = \mathbf{p_r}^T E$. Η ευθεία l_l ορίζεται ως προς το σύστημα αναφοράς της λήψης 1 και βρίσκεται πάνω στο επίπεδο του πετάσματος της κάμερας 1, δηλαδή το επίπεδο $z = f$.
- το «αντίστοιχο σημείο» του $\mathbf{p_l}$ θα βρίσκεται πάνω στην επιπολική ευθεία $l_r = \mathbf{p_l} E$. Η ευθεία l_r ορίζεται ως προς το σύστημα αναφοράς της λήψης 2 και βρίσκεται πάνω στο επίπεδο του πετάσματος της κάμερας 2, δηλαδή το επίπεδο $z = f$.

[4] Η απόδειξη της σχέσης 2.4 παρατίθεται αναλυτικά στο παράρτημα Α'.3

[5] Έτσι ώστε το εσωτερικό γινόμενο του διανύσματος ευθείας $\mathbf{l}$ και ενός σημείου $\mathbf{p}$ να κάνει μηδέν μόνο όταν το σημείο ανήκει στην ευθεία



(Α′) Στερεοσκοπική γεωμετρία

(Β′) Στερεοσκοπική διάταξη

2.2.3 Αντιστοίχηση στη στερεοσκοπική όραση

Στην στερεοσκοπική όραση, οι επιπολικές ευθείες είναι *οριζόντιες*, όπως φαίνεται στην εικόνα 2.5α′. Η προϋπόθεση αυτή ικανοποιείται είτε με φυσικό τρόπο, από την κατάλληλη τοποθέτηση στο χώρο των δύο λήψεων σε στερεοσκοπική διάταξη[6], είτε με την εφαρμογή κατάλληλων μετασχηματισμών σε ένα οποιοδήποτε ζεύγος εικόνων, διαδικασία που ονομάζεται ευθυγράμμιση[7] (rectification).

Επομένως σε ένα στερεοσκοπικό ζεύγος κάθε τυχαίο σημείο $\mathbf{p_l} = (x, y)$ της αριστερής λήψης:

- είτε θα δεν θα βρίσκεται **πουθενά** στο πέτασμα της δεξιάς (εκτός οπτικού πεδίου της)
- είτε θα βρίσκεται στο σημείο $\mathbf{p_r} = (x - d, y)$, όπου $x - d \geq 0 \Leftrightarrow d \leq x$ [8]

Η τιμή d, που εκφράζει την οριζόντια μετατόπιση του σημείου ανάμεσα στις δύο λήψεις, ονομάζεται παράλλαξη (disparity). Με γνωστά τα μεγέθη της στερεοσκοπικής διάταξης f, B και της παράλλαξης d, ο υπολογισμός του βάθους Z του τυχαίου σημείου p είναι άμεσος:

$$\left.\begin{aligned} x_1 &= -f\frac{X_1}{Z_1} \\ x_2 &= -f\frac{X_1 + B}{Z_1} \end{aligned}\right\} \Rightarrow Z_1 = \frac{fB}{x_1 - x_2} = \frac{fB}{d} \tag{2.5}$$

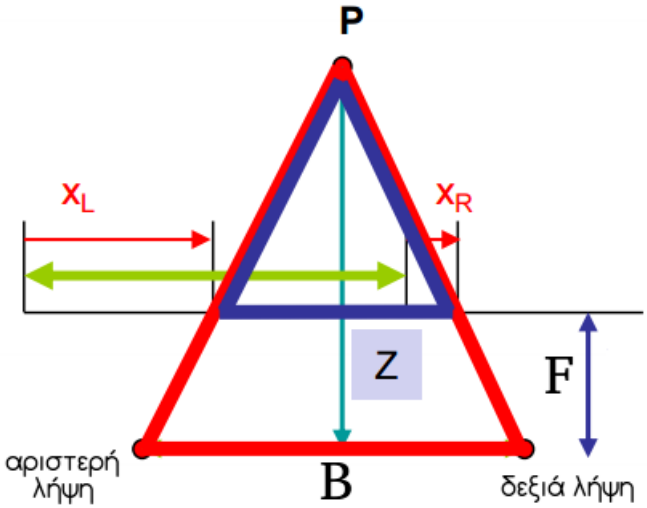
Γεωμετρικά η απόδειξη προκύπτει με κανόνα όμοιων τριγώνων φαίνεται στο σχήμα 2.6:

$$\frac{B}{Z} = \frac{(B + x_R) - x_L}{Z - f} \Rightarrow d = x_L - x_r = \frac{fB}{Z}$$

2.2.4 Πεδίο παραλλάξεων και πεδίο βάθους

Ο υπολογισμός της παράλλαξης d κάθε σημείου της εικόνας αναφοράς θα μας οδηγήσει σε έναν νέο πίνακα D όμοιων διαστάσεων με την αρχική εικόνα $D \in \mathbb{R}^{height \times width}$, όπου σε κάθε του θέση θα είναι αποθηκευμένη η πληροφορία παράλλαξης του αντίστοιχου

[6] Απόδειξη Α′.4
[7] Απόδειξη Α′.5
[8] η περίπτωση όπου $d > x$ αναλογεί στην πρώτη υποπερίπτωση (εκτός οπτικού πεδίου της κάμερας 2)



Σχημα 2.6: Γεωμετρική απόδειξη σχέσης βάθους και παράλλαξης.

σημείου, όπως φαίνεται στο σχήμα 2.7. Η παράλλαξη μετράται είτε σε *pixels* είτε σε *cm*. Επί της ουσίας οι δύο μονάδες μέτρησης είναι ισοδύναμες καθώς η οριζόντια πλευρά του *pixel* μετράται και αυτή σε *cm*. Στην παρούσα εργασία χρησιμοποιούμε ως μονάδα μέτρησης της παράλλαξης το *pixel*. Το πεδίο τιμών του χάρτη παραλλάξεων είναι το σύνολο $[0, width]$, διότι αν ένα σημείο αναλογεί σε παράλλαξη $> width$ τότε βρίσκεται εκτός οπτικού πεδίου της κάμερας 2.

Χάρτης βάθους (depth map) είναι ένας πίνακας $depth \in \mathbb{R}^{height \times width}$ που περιέχει την πληροφορία βάθους κάθε σημείου της εικόνας. Κατ' αντιστοιχία, επιλέγουμε ως μονάδα μέτρησης τα *pixels*. Το πεδίο τιμών του χάρτη παραλλάξεων είναι το σύνολο $[\frac{f \cdot b}{width}, \infty]$, διότι αν ένα σημείο αναλογεί σε βάθος $< \frac{f \cdot b}{width}$ τότε βρίσκεται εκτός οπτικού πεδίου της κάμερας 2.

Η μετάβαση από το πεδίο των παραλλάξεων στο πεδίο του βάθους είναι '1-1' και αντιστρέψιμη. Οι δύο αναπαραστάσεις είναι ισοδύναμες.

$$disparity \underset{g^{-1}}{\overset{g}{\longleftrightarrow}} depth$$

$$g : \mathbb{R}^{height \times width} \rightarrow \mathbb{R}^{height \times width} : depth = g(disparity) = \frac{fB}{disparity}$$

$$g^{-1} : \mathbb{R}^{height \times width} \rightarrow \mathbb{R}^{height \times width} : disparity = g^{-1}(depth) = \frac{fB}{depth}$$

2.3 Αρχές και περιορισμοί της στερεοσκοπικής αντιστοίχησης

Για την διεκπεραίωση της στερεοσκοπικής αντιστοίχησης βασιζόμαστε σε υποθέσεις που οδηγούν σε αντίστοιχους περιορισμούς, κάποιοι εκ των οποίων έχουν καθολική κι άλλοι μερική ισχύ.[32] Παραθέτουμε τους βασικότερους:

(α′) Αριστερή λήψη (β′) Δεξιά λήψη (γ′) χάρτης παράλλαξης

Σχημα 2.7: παράδειγμα στερεοσκοπικής λήψης

1. **Στερεοσκοπικός περιορισμός (stereo constraint):** Όπως αποδείχθηκε στη σχέση Α′.4, η αναζήτηση του αντίστοιχου σημείου περιορίζεται αυστηρά και μόνο κατά μήκος της οριζόντιας επιπολικής ευθείας.

2. **Περιορισμός συνέχειας/ασυνέχειας παράλλαξης (disparity continuity/discontinuity constraint):**

   - Κατά μήκος συνεχών επιφανειών, οι τιμές της παράλλαξης είναι συνεχείς. Ο περιορισμός αίρεται μόνο στην περίπτωση όπου μια συνεχής επιφάνεια δημιουργεί εσωτερικά «κρυμμένα σημεία».[9]
   - Κατά μήκος ασυνεχών επιφανειών, οι τιμές της παράλλαξης είναι ασυνεχείς. Ο περιορισμός αίρεται στην περίπτωση που δύο ασυνεχείς επιφάνειες τυχαίνει να βρίσκονται στο ίδιο βάθος.[10]

   Επομένως, ασυνέχειες στις τιμές της παράλλαξης συμβαίνουν είτε σε μεταβάσεις από μια επιφάνεια του χώρου σε μια άλλη, είτε αν δημιουργείται εντός μιας επιφάνειας εσωτερικό «κρυμμένο σημείο».

3. **Περιορισμός μοναδικότητας (uniqueness constraint):** Σε αδιαφανή αντικείμενα, κάθε σημείο της μίας λήψης έχει το πολύ ένα σημείο αντιστοίχησης στην έτερη λήψη. Πιο συγκεκριμένα, ένα και μοναδικό, εάν είναι ορατό από την έτερη λήψη, και κανένα εάν δεν είναι ορατό.[11] Έστω $\mathbf{p_1} \in I^L$, τότε

$$\mathbf{p_1} \xrightarrow{corresponds} \begin{cases} \emptyset, \text{ απόκρυψη} \\ \mathbf{p_2} \in I^R, \text{ αλλιώς} \end{cases}$$

   Στις περιοχές που είναι ορατές και από τις δύο λήψεις η παραπάνω σχέση γίνεται «1-1». Μπορούμε δηλαδή να αντιστοιχήσουμε μια αντιστρέψιμη συνάρτηση g μετάβασης ανάμεσα στα αντίστοιχα σημεία.

$$p_1 \underset{g^{-1}}{\overset{g}{\rightleftharpoons}} p_2$$

   Παραστατική απεικόνιση του περιορισμού μοναδικότητας παρουσιάζεται στο σχήμα 2.8 Ο περιορισμός παραβιάζεται μόνο σε σπάνιες περιπτώσεις διαφανών αντικειμένων, όπου δύο σημεία του τρισδιάστατου χώρου μπορεί να αποτυπώνονται ταυτόχρονα στο ίδιο σημείο του ενός πετάσματος και σε δύο διακριτά σημεία του έτερου. 2.9

[9] «Κρυμμένα σημεία» ονομάζουμε τα σημεία του χώρου που δεν είναι ορατά από μια λήψη.
[10] Όπως για παράδειγμα μια επιγραφή σε έναν τοίχο.
[11] η περίπτωση αυτή ονομάζεται απόκρυψη.

4. **Περιορισμός διάταξης παραλλάξεων (ordering constraint):** Το σύνολο των σημείων που απαρτίζουν το είδωλο της επιφάνειας ενός αδιαφανούς αντικειμένου της τρισδιάστατης σκηνής, είναι διατεταγμένα κατά τον ίδιο τρόπο στις δύο λήψεις. Πρακτικά, ένα σημείο της επιφάνειες που αποτυπώθηκε πιο αριστερά από ένα άλλο στο ένα πέτασμα, δεν μπορεί να αποτυπωθεί αντίστροφα (πιο δεξιά) στο άλλο πέτασμα. 2.10α′ Ο ισχυρισμός καταρρέει αν μεταβούμε από την επιφάνεια ενός αντικειμένου σε αυτή ενός άλλου. 2.10β′ Την περιοχή του χώρου εντός της οποίας αίρεται ο περιορισμός διάταξης, την αποκαλούμε «απαγορευμένη ζώνη» (forbidden zone). 2.10γ′

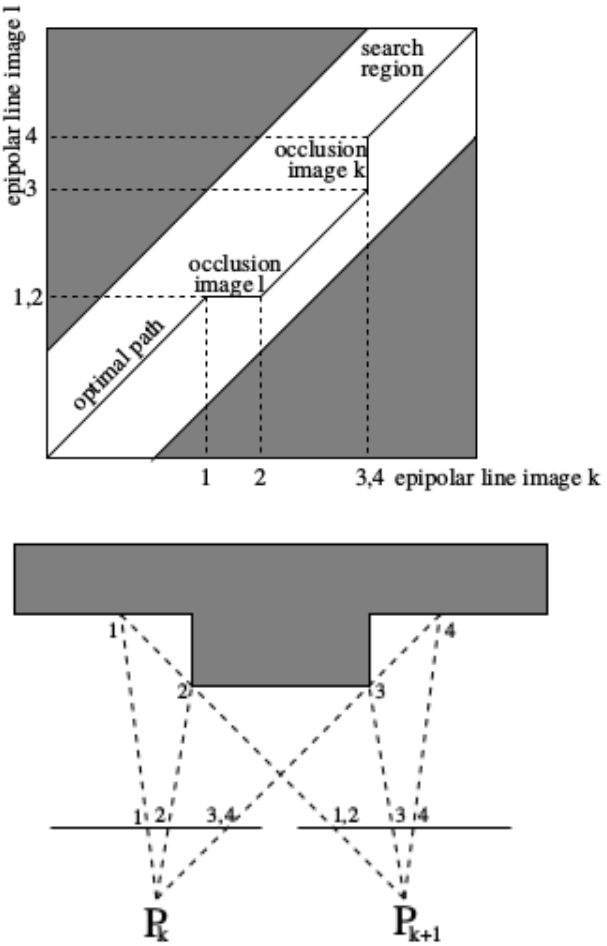
5. **Σταθερότητα φωτισμού (color constancy):** Ο φωτισμός και ο χρωματισμός κάθε σημείου μιας σκηνής παραμένει αμετάβλητος σε κάθε θέση παρατήρησης της σκηνής. Ο ισχυρισμός αυτός ισχύει για τις λαμπερτιανές επιφάνειες που προκαλούν διάχυση, αναιρείται όμως όταν η σκηνή περιλαμβάνει μη-λαμπερτιανές (λείες ή διαφανείς επιφάνειες) που προκαλούν φαινόμενα κατοπτρικής ανάκλασης και διάθλασης.

6. **Όριο μέγιστης παράλλαξης (disparity limit):** Η μέγιστη δυνατή παράλλαξη του τυχαίου σημείου $p_1 = (x, y)$ της αριστερής κάμερας είναι η $d = x$, τιμή μεταβλητή ανάλογα με την τετμημένη του μελετούμενου σημείου. Σε κάθε περίπτωση η πιθανή παράλλαξη δεν μπορεί να υπερβαίνει το πλάτος της εικόνας $d \in [0, width]$. Πολλές φορές θέτουμε ένα, σχετικά αυθαίρετο, αυστηρότερο όριο ως μέγιστη παράλλαξη d_{max}, όταν δεν μας ενδιαφέρει ο ακριβής υπολογισμός του βάθους αντικειμένων που βρίσκονται πιο κοντά από το όριο αυτό. Το επίπεδο $Z = Z_{min}$ θέτει έναν κόφτη πέρα του οποίου όλα τα αντικείμενα χαρακτηρίζονται ως «πολύ κοντινά». Η σύμβαση του μέγιστου ορίου παράλλαξης έχει το σημαντικό πλεονέκτημα της μείωσης της υπολογιστικής πολυπλοκότητας.

Θα μπορούσαμε να ομαδοποιήσουμε τους περιορισμούς 2, 3, 4 και 5 σε μια πιο ενιαία και ασαφή περιγραφή, αυτήν της «ομοιότητας γειτονιάς» (patches similarity). Η «ομοιότητα γειτονιάς», εγκολπώνοντας τους παραπάνω περιορισμούς, υποθέτει ότι μια δεδομένη περιοχή της τρισδιάστατης σκηνής θα έχει *παρόμοια* προβολή (σχήμα, χρώμα, μορφή, υφή) στα πετάσματα των δύο καμερών. Με δεδομένη αυτή την υπόθεση, θα αναζητήσουμε ακολούθως μετρικές ομοιότητας ώστε να αξιολογήσουμε την ομοιότητα περιοχών γύρω από τα σημεία ενδιαφέροντος και να «ταιριάξουμε» τα πλέον όμοια, δημιουργώντας τον ζητούμενο πίνακα παράλλαξης.

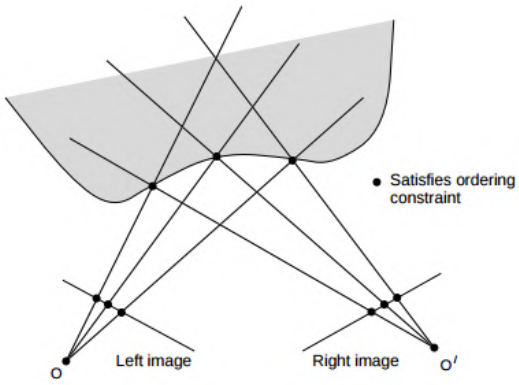
2.4 Μελέτη φαινομένων που αμφισβητούν τους περιορισμούς της στερεοσκοπικής αντιστοίχησης

Η υπόθεση της «ομοιότητας γειτονιάς» δεν έχει καθολική ισχύ. Εξ′ ορισμού η προοπτική προβολή είναι μια πράξη μετασχηματισμού που αλλοιώνει το σχήμα του προβαλλόμενου αντικειμένου. Οι διαφορετικές θέσεις λήψης δημιουργούν διαφορετικά προβαλλόμενα είδωλα. Για αυτόν τον λόγο διαχωρίζουμε τα προβλήματα της στερεοσκοπικής αντιστοίχησης σε δύο μεγάλες κατηγορίες, στα προβλήματα μικρής και μεγάλης απόστασης βάσης (small and wide baseline stereo rig problems). Η βάση αναφέρεται στην οριζόντια απόσταση B (baseline) των δύο λήψεων. Στα προβλήματα μεγάλης απόστασης βάσης τα δύο είδωλα του ίδιου τρισδιάστατου αντικειμένου αποκλίνουν έντονα σε σχήμα, χρώμα και μορφή. Η απόκλιση οξύνεται κατ′ αναλογία της αύξησης του μεγέθους B. 2.11 Στα προβλήματα μικρής απόστασης βάσης, που μας απασχολούν στην παρούσα εργασία, το

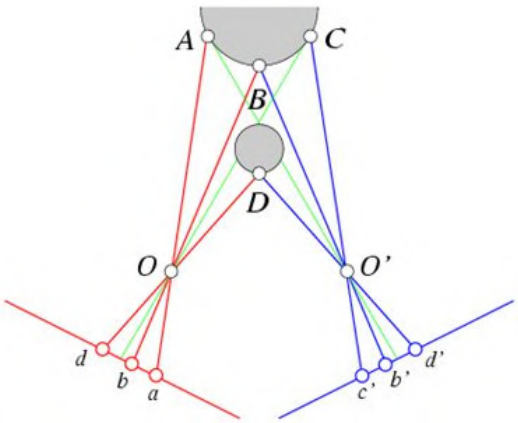


Σχημα 2.8: Περιορισμός μοναδικότητας. Στο γράφημα ο οριζόντιος άξονας είναι η επιπολική γραμμή της αριστερής λήψης, ενώ ο κάθετος της δεξιάς. Παρατηρούμε ότι στις περιοχές που δεν υφίσταται απόκρυψη, δηλαδή μέχρι το σημείο 1, ανάμεσα στα σημεία 2,3 και μετά το σημείο 4, η σχέση που συνδέει τα αντίστοιχα σημεία είναι «1-1». Αντιθέτως στα σημεία που υπάρχει απόκρυψη για κάποια από τις δύο λήψεις, δηλαδή στις περιοχές ανάμεσα στα σημεία 1,2 και 3,4, η σχέση δεν είναι αντιστρέψιμη.

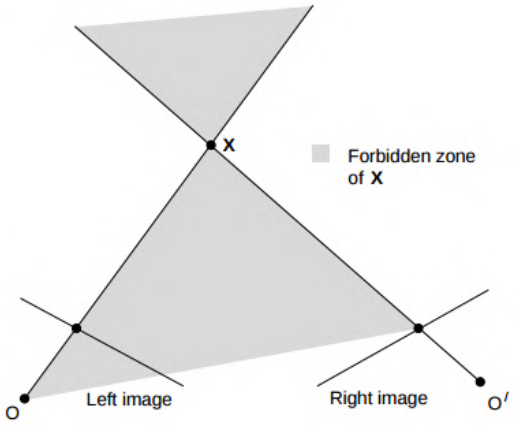
Σχημα 2.9: Τα διάφανα αντικείμενα παραβιάζουν τον περιορισμό μοναδικότητας.



(Α′) Περιορισμός διάταξης σε συνεχείς αδιαφανείς επιφάνειες



(Β′) Ο περιορισμός διάταξης αίρεται σε ασυνεχείς επιφάνειες



(Γ′) Απαγορευμένη περιοχή: οποιοδήποτε σημείο εντός της σκιασμένης περιοχής αναιρεί τον περιορισμό διάταξης παραλλάξεων

Σχημα 2.10: Περιορισμός διάταξης παραλλάξεων

φαινόμενο της σχηματικής, χρωματικής και μορφολογικής αλλοίωσης είναι αρκετά μειωμένο, όπως παρατηρούμε και στις εικόνες 2.12 2.7, χωρίς βέβαια να εξαφανίζεται πλήρως 2.13. Επομένως είναι αρκετά βάσιμο να στηριχτούμε στην «ομοιότητα γειτονιάς» σε ένα ικανοποιητικό υποσύνολο των περιοχών της εικόνας. Σε συγκεκριμένες θέσεις παρατηρούνται φαινόμενα που αναιρούν ευθέως την παραπάνω υπόθεση, και κατ΄ επέκταση τις τέσσερις βασικές υποθέσεις που έχει στηριχτεί. Παρακάτω παραθέτουμε μια συνοπτική ανάλυση των φαινομένων αυτών:

1. **Αποκρύψεις (Occlusions):** Η έστω και μικρή μετατόπιση στη γωνία θέασης δημιουργεί περιοχές του 3D χώρου που είναι ορατές μόνο από τη μία εκ των δύο λήψεων, όπως παρατηρούμε στο σχήμα 2.14. Οι περιοχές αυτές εντοπίζονται συνήθως σε ασυνέχειες βάθους, δηλαδή σε σημεία που μεταβαίνουμε από ένα αντικείμενο σε ένα άλλο και παραβιάζουν ολοκληρωτικά την υπόθεση της «ομοιότητας γειτονιάς» 2.15. Στις περιοχές αυτές πρέπει να δοθούν τιμές παράλλαξης εντός του διαστήματος τιμών $[D_L, D_R]$, όπου D_L, D_R οι τιμές παράλλαξης των σημείων που βρίσκονται εκατέρωθεν της κρυμμένης περιοχής. Βέλτιστα η μεταβολή των τιμών παράλλαξης εντός της κρυμμένης περιοχής πρέπει να ακολουθεί μια γραμμική μεταβολή από το D_L στο D_R, δηλαδή $d = D_L + (D_R - D_L)\dfrac{j - j_L}{j_R - j_L}$, όπου j συμβολίζει την οριζόντια θέση και D την τιμή της παράλλαξης.

2. **Φαινόμενο αυξομείωσης αποστάσεων και εμβαδών (Foreshortening effect):** Η προοπτική προβολή δεν κρατάει αναλλοίωτες τις αποστάσεις και τα εμβαδά. Το φαινόμενο της αυξομείωσης δημιουργεί διπλό πρόβλημα. Αφενός παραβιάζει την υπόθεση «ομοιότητας γειτονιάς», αφετέρου καθιστά και την αντιστοίχηση μη αντιστρέψιμη καθώς $n\,pixels$ της μιας λήψης αντιστοιχούνται σε $m \neq n\,pixels$ της έτερης, παρ΄ ότι το αντικείμενο είναι *απόλυτα ορατό και από τις δύο λήψεις*. 2.16

3. **Αλλοιώσεις φωτισμού**: Ο όρος φωτισμός περιγράφει την διαδικασία υπολογισμού της έντασης της φωτεινής ακτινοβολίας που προσλαμβάνει ο θεατής, που στην περίπτωσή μας είναι οι στερεοσκοπικές κάμερες. Η φωτεινή ακτινοβολία, δηλαδή το εγκάρσιο κύμα του ορατού φωτός που προσπίπτει στο πέτασμα της κάμερας, μπορεί να έχει προέλθει από τέσσερα φυσικά φαινόμενα: αυτοφωτισμό, ανάκλαση, διάθλαση και διάχυση. Εκ των τεσσάρων αυτών φαινομένων, η κατοπτρική ανάκλαση και η διάθλαση δημιουργούν φωτισμό που διαφέρει ανάλογα με την θέση του θεατή. Το φαινόμενο αυτό ονομάζεται μη λαμπερτιανό φαινόμενο φωτισμού (non-lambertian lighting effect)και οι επιφάνειες που το προκαλούν ονομάζονται μη λαμπερτιανές (non-lambertian) επιφάνειες. Παραθέτουμε επιλεκτικά κάποιες περιπτώσεις:

   (α΄) **Κατοπτρικές ανακλάσεις:** Σε αυτήν την περίπτωση μια μη λαμπερτιανή επιφάνεια (π.χ. καθρέπτης) αποτυπώνει στο πέτασμα της κάμερας το είδωλο ενός τρίτου αντικείμενου το οποίο κατοπτρίζει. Το είδωλο του τρίτου αντικειμένου, που τελικά αποτυπώνεται στον φακό, υφίσταται μεγάλες αλλοιώσεις ακόμα και σε μικρές μεταβολές της θέσης του θεατή. 2.17 Το παραπάνω πολύ δύσκολα αντιμετωπίσιμο φαινόμενο είναι σχετικά σπάνιο. Μια πιο συνηθισμένη εκδοχή του είναι ο εστιασμένος κατοπτρισμός μιας φωτεινής πηγής, που δημιουργεί αλλοιωμένη φωτεινότητα στα προβαλλόμενα είδωλα. 2.18.

   (β΄) **Φωτομετρικές αποκλίσεις (photometric variations):** Συνήθως δεν προκαλεί ολική αλλοίωση στα δύο είδωλα, αλλά μια απόκλιση κυρίως στη φωτεινότητα και στο χρώμα. Αν το φαινόμενο είναι πολύ έντονο, η μεγάλη απόκλιση στη διαφορά φωτεινότητας μπορεί να κάνει ακατάληπτο το σχήμα

(Α΄) Αριστερή λήψη (Β΄) Δεξιά λήψη

Σχημα 2.11: Στερεοσκοπική λήψη μεγάλης απόστασης βάσης

(Α΄) Αριστερή λήψη (Β΄) Δεξιά λήψη

Σχημα 2.12: Στερεοσκοπική λήψη μικρής απόστασης βάσης

του ειδώλου 2.19. Έντονες φωτομετρικές αποκλίσεις μεταξύ των δύο λήψεων προκαλούν συχνά οι σκιάσεις.

Υπάρχουν επίσης περιπτώσεις όπου ενώ η «ομοιότητα γειτονιάς» τηρείται, δημιουργούνται ιδιαίτερα μοτίβα που δημιουργούν σύγχυση σε μια απλή μετρική ομοιότητας:

1. **Επαναλαμβανόμενα μοτίβα, δομές και υφές(repetitive patterns, structures and textures):** αν το φωτογραφιζόμενο τοπίο εμφανίζει επαναλαμβανόμενα μοτίβα (όπως βιβλία σε μια βιβλιοθήκη, ρίγες μιας ζέβρας ή μια σκακιέρα) ή επαναλαμβανόμενες υφές (όπως μια εικόνα από πολλά τριαντάφυλλα) τότε έντονη «ομοιότητα γειτονιάς» εμφανίζεται σε περισσότερα από ένα σημεία στην έτερη εικόνα, προκαλώντας αδυναμία επιλογής της καταλληλότερης περιοχής για αντιστοίχηση. Μια μετρική ομοιότητας γειτονιών εμφανίζει σε αυτήν την περίπτωση πολλά τοπικά μέγιστα. 2.20

2. **Μεγάλες ομοιόμορφες περιοχές (uniform regions):** αν το εικονιζόμενο τοπίο χαρακτηρίζεται από μεγάλες ομοιόμορφες περιοχές, όπως για παράδειγμα ο ουρανός, ο τοίχος μιας πολυκατοικίας και πολλά άλλα, η «ομοιότητα γειτονιάς» παραμένει παρόμοια (κατά περιπτώσεις και τελείως ίδια) σε μεγάλο εύρος διαφορετικών παραλλάξεων. Η μετρική ομοιότητας μένει για πολλές διαδοχικές τιμές παράλλαξης κοντά στο ολικό μέγιστο.

Τέλος, υπάρχουν προβλήματα που δημιουργούνται λόγω θορύβου που προστίθεται κατά την λήψη της φωτογραφίας από το στερεοσκοπικό ζεύγος, όπως για παράδειγμα στην διαφορετική εστίαση όπου δημιουργείται θόλωμα (blur)κατά δυαδικό τρόπο σε κάθε λήψη δυσκολεύοντας την ανίχνευση της ομοιότητας. 2.21

Τα παραπάνω φαινόμενα αμφισβητούν, περισσότερο ή λιγότερο, τοπικά ή ολικά, τις αρχικές υποθέσεις στις οποίες βασίζονται οι μεθοδολογίες επίλυσης του προβλήματος της

(α′) Αριστερή λήψη (β′) Δεξιά λήψη

(γ′) Αριστερή λήψη (δ′) Δεξιά λήψη

Σχημα 2.13: Αλλοίωση των απεικονιζόμενων σχημάτων λόγω προοπτικής προβολής. Πηγή: [4]

στερεοσκοπικής αντιστοίχησης απαιτώντας ειδική αντιμετώπιση.

2.5 Ανάλυση της στερεοσκοπικής αντιστοίχησης σε επιμέρους υποπροβλήματα

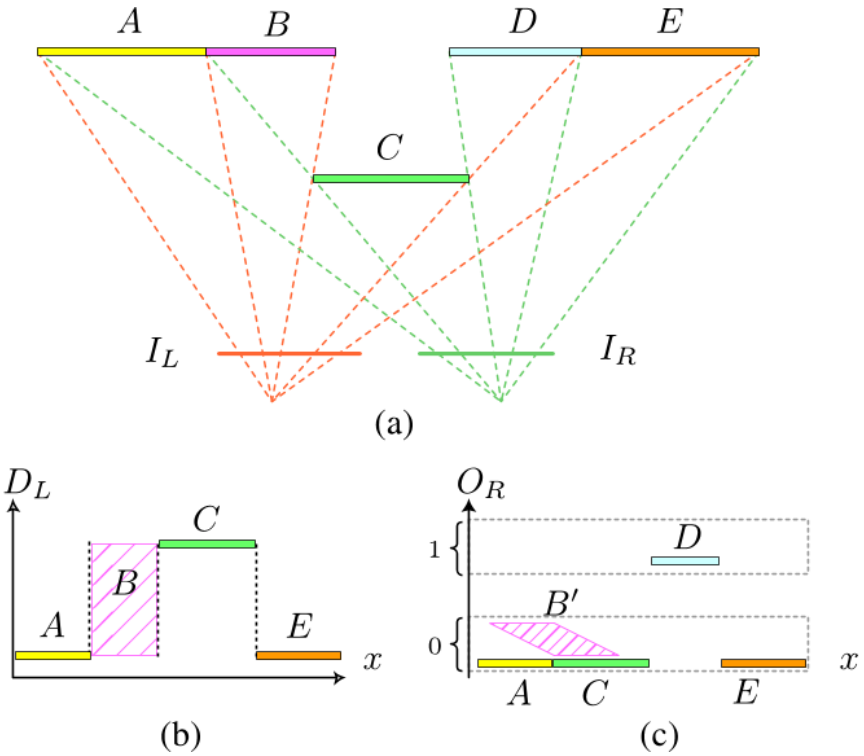
Οι Scharstein et Szeliski [35] επιμερίζουν τους αλγορίθμους επίλυσης προβλημάτων στερεοσκοπικής όρασης σε 4 βήματα:

1. Υπολογισμός κόστους αντιστοίχησης (matching cost computation)
2. Άθροιση κόστους (cost aggregation)
3. Υπολογισμός/βελτιστοποίηση χάρτη παράλλαξης (disparity computation/optimization)
4. Διόρθωση χάρτη παράλλαξης (disparity refinement)

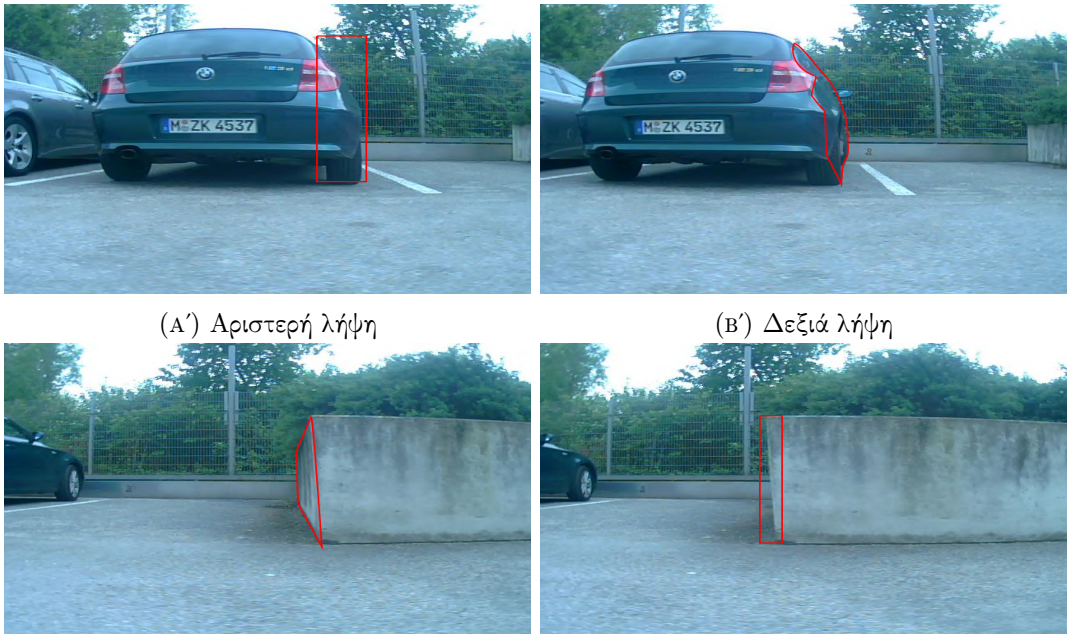
Οι Hirschmuller et Scharstein [13] ομαδοποίησαν περαιτέρω την παραπάνω κατηγοριοποίηση προτείνοντας δύο μόνο βήματα:

1. Αρχικοποίηση κόστους αντιστοίχισης (matching cost initialization)
2. Στερεοσκοπική μέθοδος (Stereo method)

Στην εργασία θα ακολουθήσουμε την ταξινόμηση των Hirschmuller et Scharstein.



Σχημα 2.14: Ανάλυση φαινομένου απόκρυψης. (α) Στερεοσκοπικό ζεύγος εικόνων I^L, I^R που αποτυπώνει μια σκηνή που περιλαμβάνει τα αντικείμενα A, B, C, D, E (β) Ο χάρτης παράλλαξης της αριστερής εικόνας. Παρατηρούμε ότι η παράλλαξη του αντικειμένου B είναι απροσδιόριστη λόγω απόκρυψης (δεν υπάρχει το είδωλό του στην έτερη λήψη). Υποχρεωτικά θα αντιστοιχηθεί είτε με το αντικείμενο A , είτε με το C , είτε θα επιμεριστεί σε δύο κομμάτια όπου αναλόγως θα αντιστοιχηθούν σε A και C . Γενικά, το αντικείμενο B θα λάβει τιμή παράλλαξης εντός του πεδίου $[D_A, D_C]$ (γ) Ο χάρτης απόκρυψης της δεξιάς εικόνας. Για την δεξιά εικόνα, το αντικείμενο D βρίσκεται σε απόκρυψη και όμοια με το (β) θα αντιστοιχηθεί είτε στο C είτε στο E .



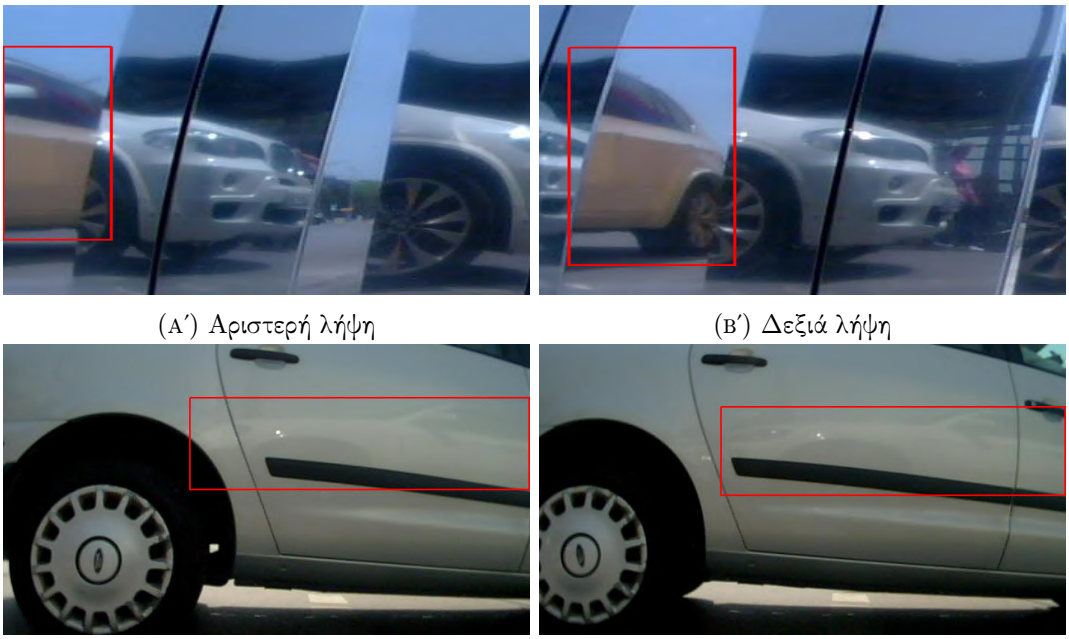
(α′) Αριστερή λήψη (β′) Δεξιά λήψη

(γ′) Αριστερή λήψη (δ′) Δεξιά λήψη

Σχημα 2.15: Φαινόμενο απόκρυψης. Πηγή: [4]

(α′) Αριστερή λήψη (β′) Δεξιά λήψη

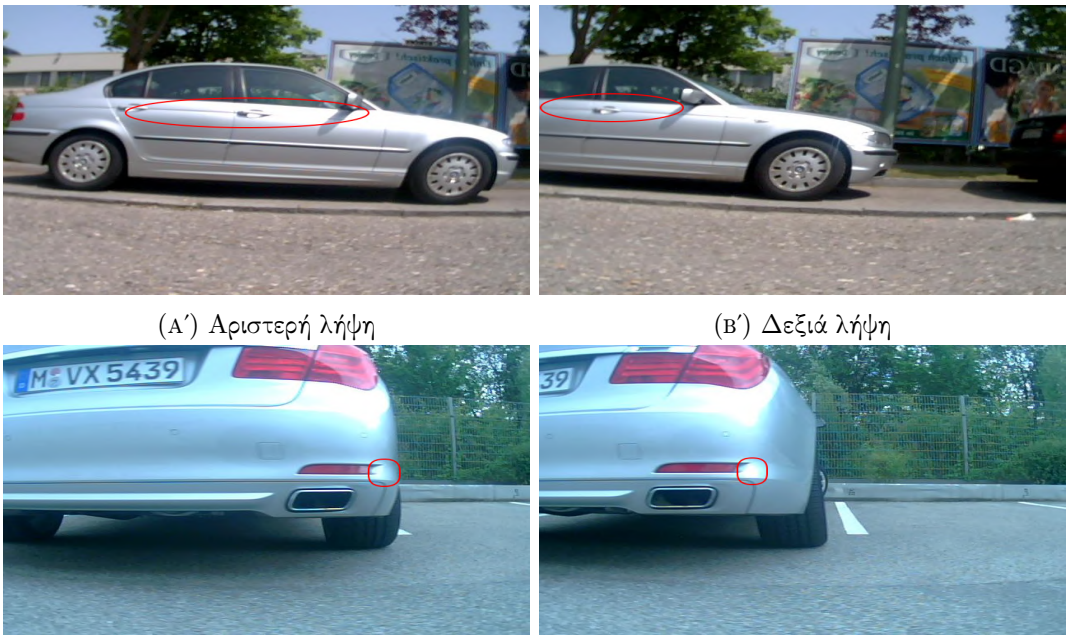
Σχημα 2.16: Φαινόμενο αυξομείωσης αποστάσεων και εμβαδών. Πηγή: [4]



(Α΄) Αριστερή λήψη (Β΄) Δεξιά λήψη

(Γ΄) Αριστερή λήψη (Δ΄) Δεξιά λήψη

Σχημα 2.17: Φαινόμενο κατοπτρικής ανάκλασης τρίτου αντικειμένου. Πηγή: [4]



(Α΄) Αριστερή λήψη (Β΄) Δεξιά λήψη

(Γ΄) Αριστερή λήψη (Δ΄) Δεξιά λήψη

Σχημα 2.18: Φαινόμενο ανάκλασης φωτεινής πηγής. Πηγή: [4]

(α′) Αριστερή λήψη (β′) Δεξιά λήψη

(γ′) Αριστερή λήψη (δ′) Δεξιά λήψη

Σχημα 2.19: Φαινόμενο φωτομετρικής απόκλισης. Πηγή: [4]

Σχημα 2.20: Επαναλαμβανόμενα μοτίβα και υφές

(Α′) Αριστερή λήψη (Β′) Δεξιά λήψη

Σχημα 2.21: Διαφορετική εστίαση σε κάθε λήψη

2.6 Αρχικοποίηση κόστους αντιστοίχισης

Στον παρών κεφάλαιο, όταν αναφερόμαστε σε σημείο $p = (x, y)$ της εικόνας εννοούμε το αντίστοιχο pixel. Επομένως, οι τιμές x, y είναι διακριτές με πεδίο τιμών $[0, \ldots, width-1]$ και $[0, \ldots, height-1]$ αντίστοιχα. Κατά τον περιορισμό 6 ορίζουμε ένα μέγιστο όριο στις υποψήφιες τιμές παράλλαξης που ερευνούμε και το ονομάζουμε `max_disparity`. Στόχος του παρόντος βήματος είναι για κάθε θέση p της εικόνας αναφοράς και για πιθανή παράλλαξη d να αναθέσουμε μια τιμή. Η τιμή αυτή θα προέλθει από την σύγκριση της γειτονιάς του p με την αντίστοιχη γειτονιά του κάθε υποψήφιου σημείου $q = (x - d, y)$ στην έτερη λήψη. Αν η τιμή εκφράζει ομοιότητα, όσο μεγαλύτερη τόσο πιο όμοιες οι συγκρινόμενες γειτονιές, ενώ το αντίθετο ισχύει αν εκφράζει κόστος. Στην παρούσα εργασία επιλέγουμε η τιμή να εκφράζει κόστος. Συμβολίζουμε τη γειτονιά ενός σημείου p με το σύμβολο N_p, και την ορίζουμε ως ένα τετράγωνο χωρίο με κέντρο το σημείο p. Το μέγεθος της γειτονιάς, δηλαδή η πλευρά του τετραγώνου, είναι μια ελεύθερη παράμετρος προς πειραματισμό.

Επομένως κατά το βήμα αυτό δημιουργούμε έναν τρισδιάστατο πίνακα κόστους:

$$C(d, x, y) : \mathbb{Z}^3 \rightarrow \mathbb{R} : C(d, x, y) = cost(I^L(x, y), I^R(x - d, y))$$

Η συνάρτηση *cost* είναι μια μετρική ομοιότητας. Παρακάτω παρατίθενται οι πιο γνωστές μετρικές που έχουν εφαρμοστεί:

- «Άθροισμα απόλυτων διαφορών» (sum of absolute differences - SAD):[10]

$$C(d,p) = -\sum_{q \in N_p} |I^L(q) - I^R(q-d)|$$

Η πιο απλή μέθοδος, η οποία αξιοποιείται ως επίδοση βάσης. Βασίζεται στην παραδοχή ότι η φωτεινότητα μένει σταθερή κατά την προβολή μιας γειτονιάς σημείων του χώρου σε δυο διαφορετικές λήψεις.

- «Άθροισμα τετραγώνων διαφορών» (sum of square differences - SSD):[16]

$$C(d,p) = -\sum_{q \in N_p} \left(I^L(q) - I^R(q-d)\right)^2$$

Βασίζεται στις ίδιες παραδοχές με την μέθοδο SAD. Λόγω του τετραγώνου εμφανίζει εντονότερη πόλωση στα μεγάλα σφάλματα.

- «Κανονικοποιημένη ετεροσυσχέτιση» ή «ομοιότητα συνημιτόνου»:

$$C(\mathbf{p},d) = \frac{\sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} I^L(\mathbf{q}) I^R(\mathbf{q}-\mathbf{d})}{\sqrt{\sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} I^L(\mathbf{q})^2 \sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} I^R(\mathbf{q}-\mathbf{d})^2}}.$$

- Απόσταση Hamming σε μετασχηματισμό Census. Αρχικά εφαρμόζουμε στη γειτονιά $\mathcal{N}_\mathbf{p}$ τον μετασχηματισμό Census. Ο μετασχηματισμός αυτός συγκρίνει την φωτεινότητα κάθε σημείου της γειτονιάς $\mathcal{N}_\mathbf{p}$ με την φωτεινότητα του κεντρικού pixel, αποθηκεύοντας την ετικέτα 1 αν είναι φωτεινότερο το περιφερειακό σημείο και 0 σε αντίθετη περίπτωση:

$$\mathbf{c}_{i,j} = \begin{cases} 1 & \text{εάν } |I(i,j) - I_c| \geqslant 0 \\ 0 & \text{εάν } |I(i,j) - I_c| < 0 \end{cases}$$

Έπειτα μετατρέπουμε τον πίνακα $\mathbf{c}_{i,j}$ σε ένα δυαδικό διάνυσμα $census(\mathbf{p})$ n θέσεων, όσο και το μέγεθος της γειτονιάς. Τέλος εφαρμόζουμε την απόσταση Hamming στα διανύσματα $census$ των συγκρινόμενων περιοχών:

$$XNOR = census(I^L(\mathcal{N}_\mathbf{p}) \odot census(I^R(\mathcal{N}_{\mathbf{p}-\mathbf{d}})$$
$$C(\mathbf{p},d) = XNOR \cdot XNOR$$

Έχουν προταθεί πολλές ακόμη μέθοδοι αρχικοποίησης του πίνακα κόστους. Το βήμα αυτό είναι το σημαντικότερο στην αλληλουχία βημάτων που καταλήγει στον υπολογισμό του χάρτη παράλλαξης και γι΄ αυτό η βιβλιογραφία είναι μεγάλη. Στο επόμενο κεφάλαιο θα δούμε πως με τη χρήση μηχανικής μάθησης μέσω τεχνητού νευρωνικού δικτύου, καταφέρνουμε την αρχικοποίηση πίνακα κόστους πολύ μεγαλύτερης ακρίβειας από αυτόν που πετυχαίνουν όλες οι παραπάνω μέθοδοι.

2.7 Στερεοσκοπική Μέθοδος (stereo method)

Η αρχικοποίηση του όγκου κόστους C περιέχει αρκετές λανθασμένες εκτιμήσεις, τοποθετημένες στα σημεία που αναιρούνται οι περιορισμοί της στερεοσκοπικής αντιστοίχησης

(αποκρύψεις, φωτομετρικές αλλοιώσεις, ομοιόμορφες περιοχές κ.α.). Στην στερεοσκοπική μέθοδο, εφαρμόζουμε τεχνικές που βελτιώνουν τις αρχικές εκτιμήσεις, οδηγώντας σε πιο ακριβή τελικό χάρτη παράλλαξης.

Θα αξιοποιήσουμε κυρίως τις μεθόδους που χρησιμοποίησαν οι Mei et al. (2011) [28] κι αναπροσάρμοσαν οι Zbontar et LeCun (2016) [45].

2.7.1 Άθροιση κόστους σε περιοχή υποστήριξης

Βασιζόμενοι στην υπόθεση «ομοιότητας γειτονιάς» (η οποία καταρρέει τοπικά σε ασυνέχειες βάθους), μπορούμε να αναθέσουμε σε κάθε pixel τον μέσο όρο των αρχικοποιημένων τιμών κόστους ολόκληρης της γειτονιάς του. Υποθέτουμε δηλαδή ότι αν ένα σημείο $\mathbf{p}$ έχει παράλλαξη d τότε και τα γειτονικά του σημεία θα βρίσκονται σε ίδια ή κοντινή παράλλαξη, αρκεί να μη μεταβαίνουμε από ένα αντικείμενο σε ένα άλλο.

Το σχήμα και το μέγεθος της γειτονιάς, που στο εξής θα ονομάζουμε περιοχή υποστήριξης (support region) μπορούν να προσδιοριστούν με δύο τρόπους:

- Ως μια ορθογώνια περιοχή προαποφασισμένου μεγέθους.
- Ως μια προσαρμοσμένη περιοχή, διαφορετική για κάθε σημείο $\mathbf{p}$.

Ορθογώνια περιοχή

Η περιοχή υποστήριξης είναι ένα τετράγωνο χωρίο το μέγεθος του οποίου επιλέγεται πειραματικά. Η επιλογή:

- μικρής περιοχής υποστήριξης προκαλεί ασάφειες (πολλές διαφορετικές τιμές παράλλαξης με παρόμοιο κόστος), καθώς δεν συνδυάζει αρκετή πληροφορία. Έτσι προκαλείται ένας χάρτης παράλλαξης αρκετά «τραχύς» με πολλές εξωκείμενες τιμές. Α΄.3γ΄
- μεγάλης περιοχής υποστήριξης έχει τα αντίστροφα αποτελέσματα. Συνδυάζει αρκετή πληροφορία, δημιουργώντας λείο χάρτη παράλλαξης, αλλά αδυνατεί να εντοπίσει τις ακμές τις οποίες και θολώνει καθώς πάσχει σε περιοχές ασυνέχειας βάθους, όπου οι αποκρύψεις καταστρατηγούν την «ομοιότητα γειτονιάς». Α΄.3δ΄

Η μαθηματική πράξη που υλοποιεί την παραπάνω διεργασία είναι η χωρική συνέλιξη με χωρικό φίλτρο μέσου όρου:

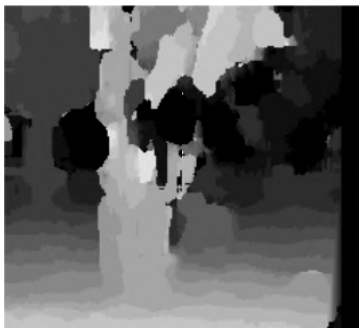
$$\texttt{for d} = 0 : \texttt{max\_disparity}$$
$$C_{agg}(d,:,:) = C_{init}(d,:,:) * h$$

όπου C_{init} ο αρχικός πίνακας κόστους, C_{agg} αυτός που προκύπτει μετά την συνέλιξη και h το φίλτρο μέσου όρου.

Η παραπάνω πράξη μπορεί μάλιστα να εφαρμοστεί κατ΄ επανάληψη, που ισοδυναμεί βαθμιαία με συνέλιξη με γκαουσιανό φίλτρο μεγαλύτερων διαστάσεων, όπως αποδεικνύεται στο παράρτημα Α΄.6.

(Α′) στερεοσκοπικό ζεύγος

(Β′) μικρό παράθυρο $w = 3$ πιξελς



(Γ′) μεγάλο παράθυρο $w = 20$ πιξελς

Σχημα 2.22: χάρτης παράλλαξης για μεταβλητό μέγεθος παραθύρου

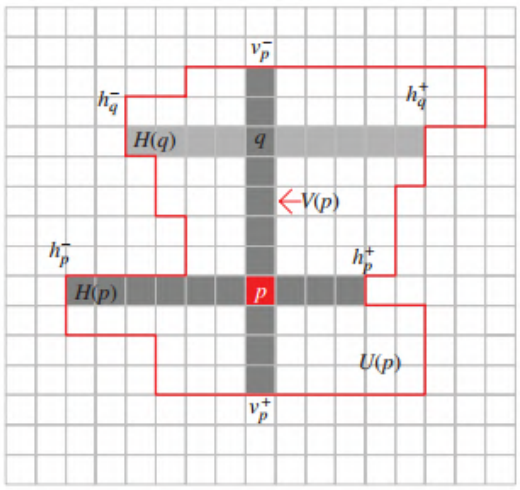
Προσαρμοσμένη περιοχή υποστήριξης

Ο ορισμός ενιαίας περιοχής υποστήριξης για όλα τα σημεία προκαλεί τα παραπάνω προβλήματα. Επιθυμούμε την δημιουργία περιοχής υποστήριξης διαφορετικού μεγέθους και σχήματος ανά σημείο. Στόχος είναι η περιοχή υποστήριξης να μην περιλαμβάνει μεταβάσεις από ένα αντικείμενο σε ένα άλλο. Έτσι οι μέσοι όροι θα συνδυάζουν πληροφορία από γειτονικά pixels που ανήκουν στο ίδιο αντικείμενο και άρα έχουν ίδια ή παρόμοια τιμή παράλλαξης.

Βασιζόμαστε στην υπόθεση ότι η μετάβαση από ένα αντικείμενο σε ένα άλλο αποτυπώνεται στο πέτασμα μέσω απότομης αλλαγής στη φωτεινότητα των pixels. [23] Ακολουθώντας την μεθοδολογία που πρότειναν οι Zhang et al. (2009) [46], αντιστοιχίζουμε σε κάθε pixel τέσσερις τιμές που συγκροτούν έναν σταυρό. Συγκεκριμένα, οι τέσσερις τιμές $[v_p^-, v_p^+, h_p^-, h_p^+]$ δηλώνουν πόσο εκτείνεται ο σταυρός στις τέσσερις κατευθύνσεις: πάνω, κάτω - κάθετος άξονας - και αριστερά, δεξιά - οριζόντιος άξονας.

Αποφασίζουμε την τιμή των $[v_p^-, v_p^+, h_p^-, h_p^+]$ που αντιστοιχούν στο pixel p ακολουθώντας δύο κριτήρια. Μετατοπιζόμαστε κατά μήκος της κάθε κατεύθυνσης όσο τηρούνται οι περιορισμοί:

- $|I(\mathbf{p}) - I(\mathbf{p}')| <$ `intensity_threshold`: η διαφορά στη φωτεινότητα μεταξύ του υπό εξέταση pixel $\mathbf{p}$ και του $\mathbf{p}'$ είναι μικρότερο ενός ορίου που θέτουμε ως παράμετρο. Όσο μικρότερη τιμή ορίου, τόσο αυξανόμενη ευαισθησία στην ανεύρεση συνόρου.
- $\|\mathbf{p} - \mathbf{p}'\| <$ `distance_threshold`: η μέγιστη τιμή προς κάθε κατεύθυνση περιορίζεται από το όριο `distance_threshold` επιβάλλοντας περιορισμό στο μέγιστο



Σχημα 2.23: παράδειγμα δημιουργίας περιοχής υποστήριξης τυχαίου pixel $\mathbf{p}$. Αφού, υπολογιστούν οι τιμές $[v_p^-, v_p^+, h_p^-, h_p^+]$ υπολογίζονται οι οριζόντιοι άξονες όλων των θέσεων $\mathbf{q}$ κατά μήκος του κάθετου άξονα

εμβαδόν της περιοχής υποστήριξης.

Αφού υπολογιστούν οι τέσσερις τιμές $[v_p^-, v_p^+, h_p^-, h_p^+] \forall \mathbf{p}$, η περιοχή υποστήριξης του $\mathbf{p}$ υπολογίζεται ως η ένωση των τιμών $[h_p^-, h_p^+]$ όλων των θέσεων $\mathbf{q}$ κατά μήκος του κάθετου άξονα του $\mathbf{p}$, όπως φαίνεται στο σχήμα 2.23

Η τελική επιλογή περιοχής υποστήριξης πρέπει να λαμβάνει υπ' όψη και την αντίστοιχη περιοχή στο αντίστοιχο pixel της έτερης λήψης. Επομένως, συμβολίζοντας $U^L(\mathbf{p})$ την περιοχή υποστήριξης στην αριστερή λήψη, $U^R(\mathbf{p})$ στην δεξιά, η τελική περιοχή υποστήριξης για την υπό εξέταση παράλλαξη d είναι η:

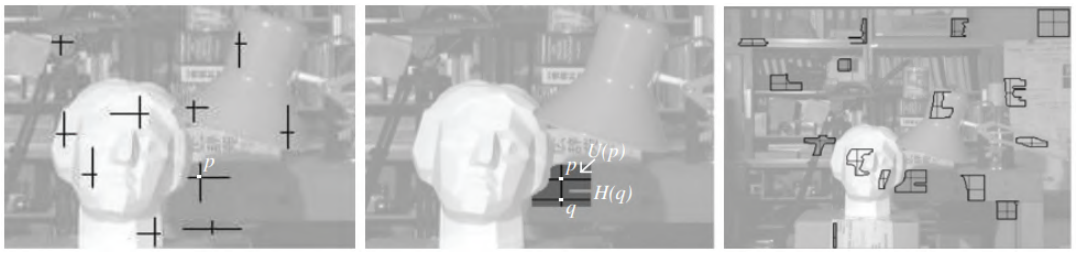
$$U_d(\mathbf{p}) = U^L(\mathbf{p}) \cup U^R(\mathbf{p} - \mathbf{d}) \Rightarrow$$
$$U_d(\mathbf{p}) = \{\mathbf{q} | \mathbf{q} \in U^L(\mathbf{p}), \mathbf{q} - \mathbf{d} \in U^R(\mathbf{p} - \mathbf{d})\}$$

Στο σχήμα 2.24, φαίνεται ότι η προσαρμοσμένη περιοχή υποστήριξης αποδίδει με ικανοποιητική ακρίβεια, καταφέρνοντας να περιορίσει τις περιοχές υποστήριξης εντός των ίδιων αντικειμένων και αποφεύγοντας την εμπλοκή εξωκείμενων τιμών στον υπολογισμό των μέσων όρων.

Η υπολογιστική πολυπλοκότητα του υπολογισμού της γειτονιάς για κάθε σημείο είναι $O(H \cdot W \cdot \texttt{distance\_threshold})$, ενώ για την άθροιση του κόστους $O(H \cdot W \cdot D \cdot (2 \cdot \texttt{distance\_threshold})^2)$.

2.7.2 Ημικαθολική αντιστοίχηση (semi-global matching)

Η μέθοδος άθροισης κόστους σε περιοχές υποστήριξης του προηγούμενο κεφαλαίου πετυχαίνει την τοπική εξομάλυνση των τιμών του κόστους. Οι νέες τιμές κόστους που υπολογίζει για κάθε pixel εξαρτώνται μόνο από τις τιμές της «γειτονιάς» του.



Σχημα 2.24: εφαρμογή μεθόδου υπολογισμού προσαρμόσιμων περιοχών υποστήριξης.

Η ημικαθολική αντιστοίχηση επιχειρεί να επιβάλει εξωτερικούς περιορισμούς που λαμβάνουν υπ΄ όψιν τις τιμές κόστους όλης της εικόνας καθολικά κι όχι ξεχωριστά κάθε γειτονιάς, όπως η προηγούμενη μέθοδος. Έτσι, επιδιώκει τον σχηματισμό ενός καθολικά ομαλού χάρτη παράλλαξης κι όχι ομαλού κατά γειτονιές μόνο. Όπως είναι λογικό, η πρόκληση που αντιμετωπίζει αφορά τις ασυνέχειες βάθους, τις οποίες πρέπει να εντοπίσει και να χειριστεί κατάλληλα, ώστε να μην προκύψει χάρτης παράλλαξης με θολωμένες (blur)τις άκρες των αντικειμένων.

Ορίζουμε μια συνάρτηση αναφοράς (συνήθως αποκαλείται μεταφορικά συνάρτηση ενέργειας, χωρίς να σχετίζεται άμεσα με την φυσική έννοια της ενέργειας) E που εξαρτάται από τον χάρτη παράλλαξης D και τον αρχικοποιημένο πίνακα κόστους C. Στο παρών βήμα θεωρούμε σταθερό τον πίνακα κόστους C και μεταβλητό τον χάρτη παράλλαξης D, επομένως συμβολίζουμε την συνάρτησή μας ως $E_C(D)$ και την ορίζουμε ως:

$$E_C(D) = \sum_{\mathbf{p}} \Bigg(C(\mathbf{p}, D(\mathbf{p})) + \sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} P_1 \cdot 1\{|D(\mathbf{p}) - D(\mathbf{q})| = 1\} + \sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} P_2 \cdot 1\{|D(\mathbf{p}) - D(\mathbf{q})| > 1\} \Bigg), \quad (2.6)$$

Η συνάρτηση $1\{\cdot\}$ ισούται με

$$1\{\text{statement}\} = \left\{ \begin{array}{ll} 1, & \text{if, statement: True} \\ 0, & \text{if, statement: False} \end{array} \right\}$$

Στην συνάρτηση 2.6 ο όρος:

- $\sum_{\mathbf{p}} C(\mathbf{p}, D(\mathbf{p}))$ «τιμωρεί» τις επιλογές παράλλαξης υψηλού κόστους
- $\sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} P_1 \cdot 1\{|D(\mathbf{p}) - D(\mathbf{q})| = 1\}$ «τιμωρεί» με την τιμή P_1 κάθε γειτονικό pixel με παράλλαξη που διαφέρει κατά 1
- $\sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} P_2 \cdot 1\{|D(\mathbf{p}) - D(\mathbf{q})| > 1\}$ «τιμωρεί» με την τιμή $P_2 > P_1$ κάθε γειτονικό pixel με παράλλαξη που διαφέρει περισσότερο από 1

Η τιμές P_1, P_2 δεν πρέπει να παραμείνουν αμετάβλητες σε όλο το εύρος της εικόνας. Μια τέτοια προσέγγιση θα «τιμωρούσε» την απότομη μεταβολή στην παράλλαξη γειτονικών pixel που βρίσκονται σε διαφορετικό επίπεδο, μη επιτρέποντας την ασυνέχεια σε ασυνεχείς επιφάνειες. Αντιθέτως, επιθυμούμε μεροληπτική συμπεριφορά που θα ευνοεί τις ασυνέχειες σε ασυνεχείς επιφάνειες και θα τις τιμωρεί σε συνεχείς, σεβόμενοι την

αντίστοιχη υπόθεση. Θα αξιοποιήσουμε την μέθοδο της ανίχνευσης ακμών από τις απότομες μεταβάσεις στη φωτεινότητα. Ορίζουμε:

$$diff_l = |I^L(\mathbf{p}) - I^L(\mathbf{q})| \quad , \quad \mathbf{p} \in I^L, \quad \mathbf{q} \in \mathcal{N}_\mathbf{p}$$

$$diff_r = |I^R(\mathbf{p} - \mathbf{d}) - I^R(\mathbf{q} - \mathbf{d})| \quad , \quad d : \texttt{disparity of p}$$

κι επιλέγουμε την τιμή των P_1, P_2 σύμφωνα με τους κανόνες:

$P_1 = \texttt{P1\_ref}$,	$P_2 = \texttt{P2\_ref}$	if $diff_l < \texttt{thres}, diff_r < \texttt{thres}$
$P_1 = \texttt{P1\_ref/big\_factor}$,	$P_2 = \texttt{P2\_ref/big\_factor}$	if $diff_l \geq \texttt{thres}, diff_r \geq \texttt{thres}$
$P_1 = \texttt{P1\_ref/small\_factor}$,	$P_2 = \texttt{P2\_ref/small\_factor}$	otherwise.

Οι υπερπαράμετροι της μεθοδολογίας είναι οι

- `P1_ref, P2_ref` που εκφράζουν την βασική τιμή αναφοράς για τις ασυνέχειες στην τιμή της παράλλαξης
- `thres` , όριο στην διαφορά της φωτεινότητας
- `small_factor` , διαιρεί την τιμή αναφοράς όταν μία εκ των τιμών $diff_l$, $diff_r$ ξεπεράσει το όριο διαφοράς φωτεινότητας
- `big_factor` , διαιρεί την τιμή αναφοράς όταν οι $diff_l$, $diff_r$ αμφότερες ξεπεράσουν το όριο διαφοράς φωτεινότητας

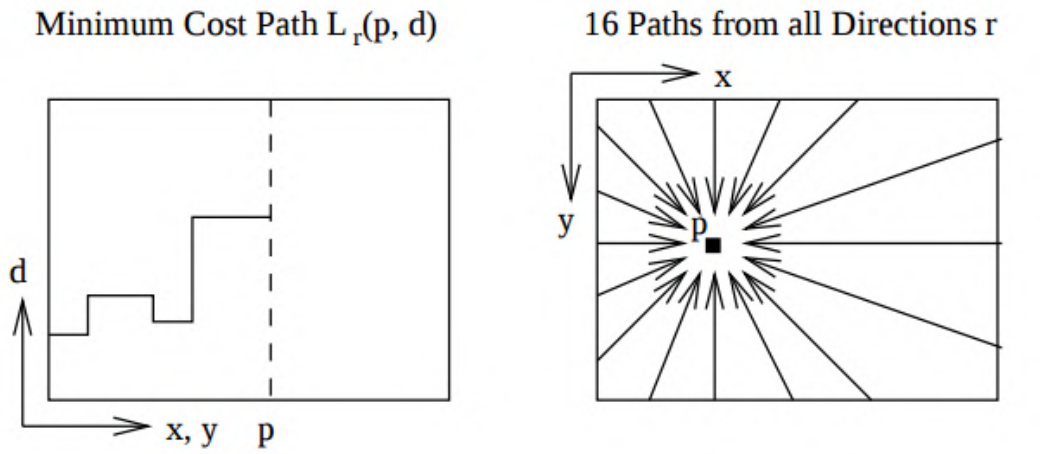
Επίλυση του προβλήματος ημικαθολικής αντιστοίχησης

Το πρόβλημα ελαχιστοποίησης της συνάρτησης $E_C(D)$ εμφανίζει δυσκολία επίλυσης για δύο λόγους:

- Η συνάρτηση $E_C(D)$ δεν είναι συνεχής ώστε να προσεγγίσουμε το ελάχιστό της μέσω μεθοδολογιών πρώτης παραγώγου
- Οι πιθανές τιμές D είναι $(m \times n)^{\text{max\_disparity}}$ που δημιουργεί μη αντιμετωπίσιμη υπολογιστική πολυπλοκότητα. Για παράδειγμα, μια εικόνα πολύ μικρής ανάλυσης 50×100pixels με μέγιστη παράλλαξη μόλις τα 10pixels δημιουργεί χώρο $(50 \times 100)^{10} \approx 10^{37}$ πιθανών τιμών.

Είναι υπολογιστικά αδύνατη η επίλυση του προβλήματος προς όλες τις κατευθύνσεις ταυτόχρονα. Αναγκαστικά, συμβιβαζόμαστε στην επίλυσή του προς μία κατεύθυνση $\mathbf{r}$ τη φορά με τη βοήθεια δυναμικού προγραμματισμού. Φυσικά, αυτός ο συμβιβασμός μας εγγυάται την επιβολή των εξωτερικών περιορισμών ομαλότητας **μόνο** στις κατευθύνσεις που θα τον εφαρμόσουμε.

Ο Hirschmüller (2008) [12] πρότεινε την ελαχιστοποίηση της ενέργειας σε 16 κατευθύνσεις, όπως φαίνονται στην εικόνα 2.25, και τον τελικό υπολογισμό του μέσου όρου αυτών.



Σχημα 2.25: οι 16 κατευθύνσεις που προτάθηκαν από τον Hirschmüller (2008) [12]

Επίλυση του προβλήματος ημικαθολικής αντιστοίχησης σε μία κατεύθυνση

Η επίλυση του προβλήματος σε μια προαποφασισμένη κατεύθυνση μπορεί να υλοποιηθεί με χρήση δυναμικού προγραμματισμού. Οι νέες τιμές του πίνακα κόστους $C_{\mathbf{r}}(\mathbf{p}, d)$ υπολογίζονται σύμφωνα με την αναδρομική σχέση:

$$C_{\mathbf{r}}(\mathbf{p}, d) = C(\mathbf{p}, d) - \min_k C_r(\mathbf{p} - \mathbf{r}, k) + \min\Big\{ C_r(\mathbf{p} - \mathbf{r}, d), C_r(\mathbf{p} - \mathbf{r}, d-1) + P_1, \\ C_r(\mathbf{p} - \mathbf{r}, d+1) + P_1, \min_k C_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, k) + P_2 \Big\}$$

Με αυτήν την μέθοδο η υπολογιστική πολυπλοκότητα του αλγορίθμου είναι $O(D \cdot H \cdot W)$ ανά κατεύθυνση.

2.7.3 Υπολογισμός χάρτη παράλλαξης (disparity map computation)

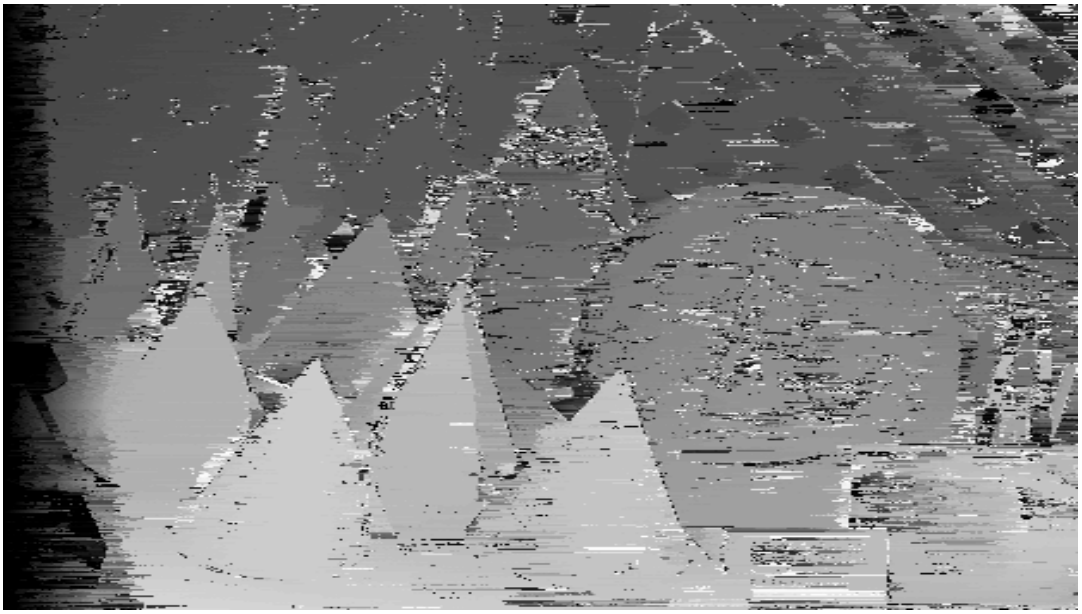
Μετά την επιβολή των παραπάνω περιορισμών ομαλότητας, ο χάρτης παράλλαξης υπολογίζεται σε κάθε θέση $\mathbf{p}$ ως η παράλλαξη με το ελάχιστο κόστος αντιστοίχησης (winner takes it all strategy):

$$D(\mathbf{p}) = \operatorname{argmin}_d C(\mathbf{p}, d).$$

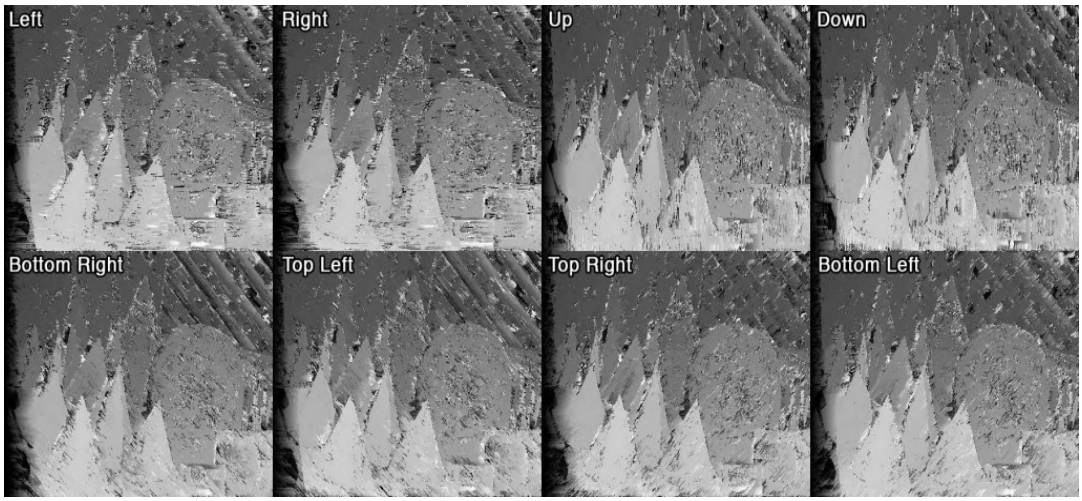
2.7.4 Εντοπισμός εξωκείμενων τιμών στον χάρτη παράλλαξης (outlier values detection in disparity map)

Θεωρούμε δεδομένη μια πρώτη εκτίμηση του χάρτη παράλλαξης D_{init}, όπως προέκυψε από το προηγούμενο βήμα. Μας ενδιαφέρει να ερευνήσουμε αν τηρούνται οι περιορισμοί

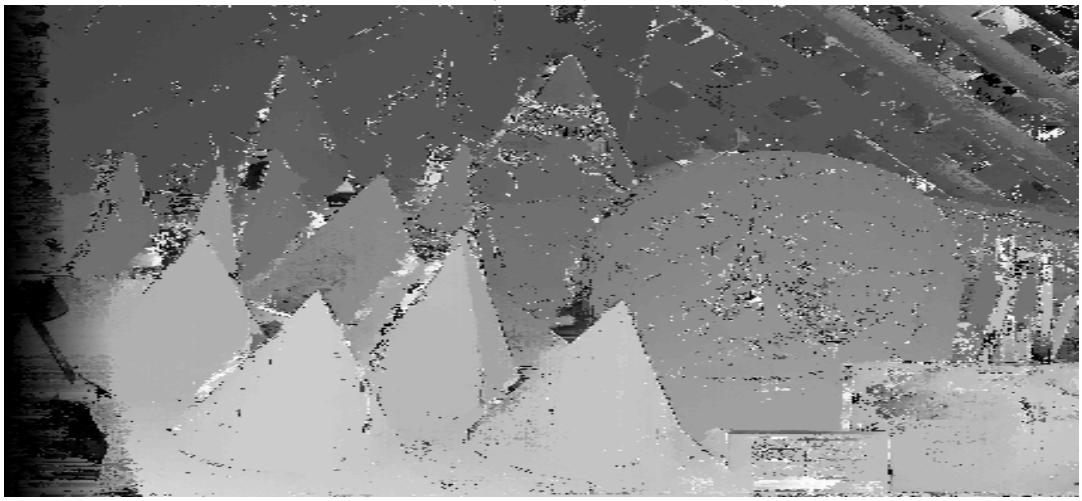
(Α′) Στερεοσκοπικό ζεύγος από το αρχείο Middlebury [38]



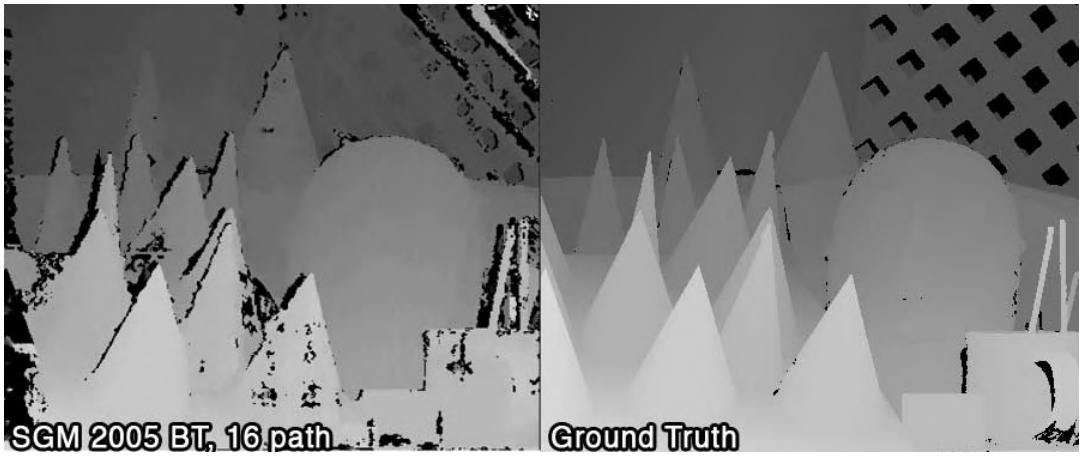
(Β′) Εφαρμογή του αλγορίθμου ημικαθολικής αντιστοίχησης μόνο κατά την οριζόντια κατεύθυνση $\boldsymbol{\rho} = (1, 0)$. Παρατηρούμε ότι η ανυπαρξία κάποιου περιορισμού ομαλότητας προς οποιαδήποτε άλλη κατεύθυνση δημιουργεί τα επονομαζόμενα φαινόμενα ραβδώσεων (streaking effects). Πηγή: [30]



(α′) Εφαρμογή του αλγορίθμου στις 8 βασικές κατευθύνσεις: 4 διευθύνσεις (οριζόντια, κάθετη, πρώτη και δεύτερη διαγώνιος) και 2 φορές ανά διεύθυνση. Το φαινόμενο ραβδώσεων εμφανίζεται πάντα σε διαφορετική κατεύθυνση.



(β′) Χάρτης παράλλαξης βασισμένος στον μέσο όρο των τιμών που προκύπτουν από τις 8 βασικές κατευθύνσεις. Παρατηρούμε σαφή μείωση του φαινομένου των ραβδώσεων.



(γ′) αριστερά: Χάρτης παράλλαξης βασισμένος στον μέσο όρο των τιμών που προκύπτουν από τις 16 βασικές κατευθύνσεις, όπως ακριβώς προτάθηκε από τον Hirschmüller (2008) [12]. Δεξιά: οι πραγματικές τιμές του χάρτη παράλλαξης μετρημένες εργαστηριακά. Παρατηρούμε την έντονη ομοιότητα στο μεταξύ των δύο εικόνων.

Σχημα 2.27: Παραδείγματα επίλυσης του προβλήματος της ημικαθολικής αντιστοίχησης σε συγκεκριμένες κατευθύνσεις. Πηγή: [30]

που διατυπώθηκαν στο κεφάλαιο 2.3.

Σύμφωνα με τον περιορισμό μοναδικότητας 3, υπάρχει μια «1-1» αντιστρέψιμη σχέση που συνδέει τα σημεία των προβαλλόμενων ειδώλων μιας περιοχής ορατής και από τις δύο λήψεις. Ο περιορισμός αναιρείται σε περίπτωση αποκρύψεων 1.

Συμβολίζουμε ως D^L τον χάρτη παράλλαξης που προκύπτει θεωρώντας την αριστερή εικόνα I^L ως εικόνα αναφοράς και D^R τον αντίστοιχο με αναφορά στην εικόνα I^R. Μελετάμε αν υπάρχει συμφωνία στις τιμές παράλλαξης που προβλέπουν καταλήγοντας σε τρία πιθανά ενδεχόμενα για κάθε $\mathbf{\pi}$:

- **ορθή παράλλαξη**, αν

$$|D^L(\mathbf{p}) - D^R(\mathbf{p} - \mathbf{d})| \leqslant 1$$

  διότι οι προβλέψεις ταυτίζονται.

- **απόκρυψη**, αν

$$|d - D^R(\mathbf{p} - \mathbf{d})| > 1 \quad \forall d : \{\mathbf{p} - \mathbf{d} \geqslant 0\}$$

  Διατρέχουμε την οριζόντια επιπολική ευθεία στην δεξιά εικόνα κι ελέγχουμε ότι κανένα άλλο pixel **p'** δεν αντιστοιχίζεται με το pixel **p** . Τότε υποθέτουμε ότι υπάρχει φαινόμενο απόκρυψης. 2.28

- **αναντιστοιχία για οποιονδήποτε άλλο λόγο**, αν

$$\exists \quad d : \{\mathbf{p} - \mathbf{d} \geqslant 0\} \quad \text{τέτοιο ώστε :} \quad |d - D^R(\mathbf{p} - \mathbf{d})| \leqslant 1$$

  Αν αντίθετα με την προηγούμενη περίπτωση υπάρχει pixel **p'** που αντιστοιχίζεται με το pixel **p** θεωρούμε ότι για οποιονδήποτε από τους λόγους του κεφαλαίου 2.4, έχει παραβιαστεί η ομοιότητα γειτονιάς και έχουμε αστοχία πρόβλεψης.

Διόρθωση τιμών παράλλαξης σε pixels με σήμανση «απόκρυψη»

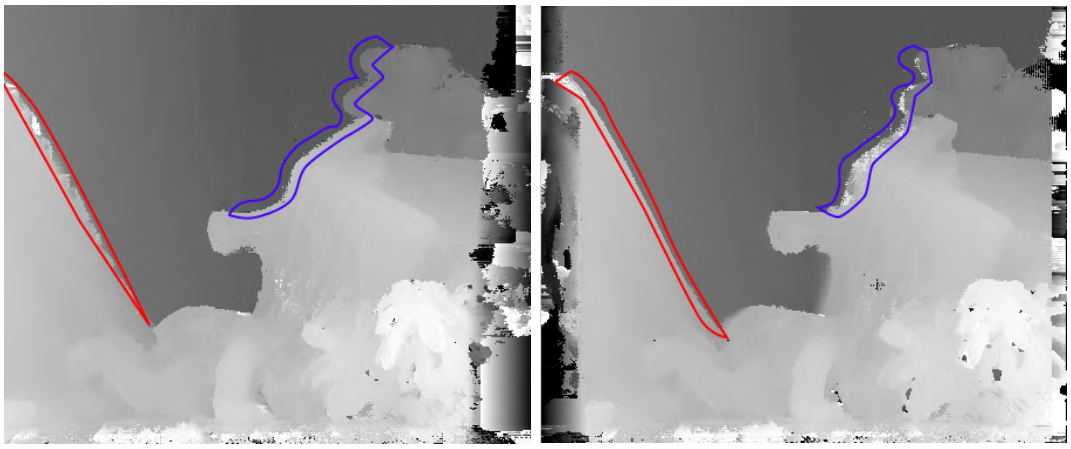
Για τα pixels με σήμανση «απόκρυψη», θέλουμε η παράλλαξή τους να υπολογιστεί με βάση γειτονικά pixels ταυτοποιημένα ως «ορθή παράλλαξη».

Οι Mei et. al (2009) [28] προτείνουν τη μέθοδο iterative region voting (επαναληπτική ψηφοφορία περιοχής υποστήριξης). Σ' αυτήν την μέθοδο δημιουργείται ένα ιστόγραμμα H_p με $d_{max}+1$ στάθμες, από τις τιμές παράλλαξης των pixels στην περιοχή υποστήριξης του $\mathbf{p}$ που έχουν σημανθεί ως «ορθή παράλλαξη». Συμβολίζουμε με d_p^* την υψηλότερη στάθμη. Αν ο αριθμός των ψηφισάντων S_p (pixels με σήμανση «ορθή παράλλαξη») υπερβαίνει ένα κατώφλι τ_S και η αναλογία $\dfrac{H_p(d_p^*)}{S_p}$ είναι μεγαλύτερη ενός άλλου ορίου τ_H, τότε η εκτίμηση d_p^* θεωρείται ασφαλής και ανατίθεται στο pixel $\mathbf{p}$:

$$S_p > \tau_s, \quad \frac{H_p(d_p^*)}{S_p} > \tau_H$$

Η παραπάνω μέθοδος μπορεί να χρησιμοποιηθεί επαναληπτικά.

Οι Zbontar et LeCun (2016) [45] πρότειναν μια πολύ πιο απλοποιημένη μέθοδο κατά την οποία αναζητούμε το πιο κοντινό pixel με σήμανση «ορθή παράλλαξη» κατά μήκος της



(Α′) εικόνα 1: χάρτης παράλλαξης D^L (Β′) εικόνα 2: χάρτης παράλλαξης D^R

Σχημα 2.28: Παράδειγμα αναντιστοιχίας στους χάρτες παράλλαξης D^L, D^R εστιασμένη σε περιοχές απόκρυψης που προκαλούνται λόγω μετάβασης από ένα αντικείμενο σε ένα άλλο

επιπολικής ευθείας κι αναθέτουμε την τιμή της παράλλαξής του στο «κρυμμένο» pixel.

Η μέθοδος Zbontar et LeCun είναι πιο συνεπής απέναντι στην στερεοσκοπική γεωμετρία και κυρίως στον περιορισμό μοναδικότητας 3, ενώ η μέθοδος Mei et. al εμφανίζει μεγαλύτερη ευστάθεια αξιοποιώντας πληροφορία πολλαπλών pixels.

Διόρθωση τιμών παράλλαξης σε pixels με σήμανση «αναντιστοιχία»

Τα pixels με σήμανση «αναντιστοιχία» περιέχουν άστοχη τιμή παράλλαξης. Η αστοχία μπορεί να έχει προέλθει από οποιαδήποτε παραβίαση των περιορισμών την στερεοσκοπικής αντιστοίχησης, όπως παρουσιάστηκαν στο κεφάλαιο 2.3, εκτός του περιορισμού μοναδικότητας που προκαλείται από το φαινόμενο αποκρύψεων. Σε κάθε περίπτωση εφόσον εξαιρούμε το ενδεχόμενο απόκρυψης, ισχύουν οι περιορισμοί συνέχειας/ασυνέχειας 2 και διάταξης παραλλάξεων 4. Βασιζόμενοι σε αυτές επιθυμούμε την πρόβλεψη της τιμής της παράλλαξης από τα γειτονικά pixels.

Υλοποιούμε την εξής μεθοδολογία. Κινούμενοι αποκλειστικά εντός της περιοχής υποστήριξης $U_d(\mathbf{p})$ (ώστε να εξασφαλίζουμε πλοήγηση εντός του ίδιου αντικειμένου), εκμεταλλευόμαστε την πληροφορία όλων των "pixels" με σήμανση «ορθή παράλλαξη». Συμβολίζουμε $d_{U_d}(\mathbf{p})$ το σύνολο των τιμών παράλλαξης των pixels, εντός της περιοχής υποστήριξης. Ως παράλλαξη του pixel $\mathbf{p}$ αναθέτουμε την διάμεσο του συνόλου $d_{U_d}(\mathbf{p})$.

2.7.5 Βελτιστοποίηση με ακρίβεια υποπίξελ

Έως τώρα ο χάρτης παράλλαξης περιέχει ακέραιες τιμές. Η παράλλαξη είναι από τη φύση της πραγματικός αριθμός. Προκειμένου να μετατρέψουμε τις ακέραιες τιμές σε πραγματικές, τοποθετούμε μια τετραγωνική καμπύλη (quadratic curve) ανάμεσα στα γειτονικά κόστη ώστε να υπολογίσουμε έναν νέο χάρτη παράλλαξης:

$$D_{\text{SE}}(\mathbf{p}) = D_{\text{INT}}(\mathbf{p}) - \frac{C(\mathbf{p}, d+1) - C(\mathbf{p}, d-1)}{2(C(\mathbf{p}, d+1) - 2C(\mathbf{p}, d) + C(\mathbf{p}, d-1))},$$

2.8 Αξιολόγηση χάρτη παράλλαξης (Disparity map evaluation)

Ο χάρτης παράλλαξης οπτικοποιείται μέσω εικόνας όπου ο χρωματισμός κάθε σημείου ορίζεται από την αντίστοιχη τιμή της παράλλαξης, όπως φαίνεται στην εικόνα 2.29γ′. Υπάρχουν δύο βασικές μετρικές αξιολόγησης της ακρίβειας του υπολογισμένου χάρτη παράλλαξης, το «απόλυτο σφάλμα πρόβλεψης» και το «απόλυτο σφάλμα πρόβλεψης με ανώφλι».

2.8.1 Απόλυτο σφάλμα πρόβλεψης (Absolute prediction error)

Έστω $\mathbf{D}^L_{\mathbf{ground\_truth}} \in \mathbb{R}^{M\times N}$ ο πραγματικός χάρτης παράλλαξης, ο οποίος διαθέτει πληροφορία για ένα υποσύνολο των σημείων της εικόνας. Έστω $\mathbb{A} \subseteq \mathbb{C} = \{(x, y) : x \in [0, M], y \in [0, N]\}$ αυτό το υποσύνολο και $\mathbb{B} = [\{(x, y) : x \in [0, M], y \in [0, N]\} - \mathbb{A}]$ τον δυαδικό του. Συμβολίζουμε με $\mathbf{D}^L_{\mathbf{predicted}} \in \mathbb{R}^{M\times N}$ τον χάρτη παράλλαξης της ίδιας εικόνας που έχει προέλθει από υπολογισμό.

Ο πίνακας «απόλυτου σφάλματος» ορίζεται ως:

$$\mathbf{AD}(\mathbf{p}) = \begin{cases} |\mathbf{D}^L_{\mathbf{ground\_truth}}(\mathbf{p}) - \mathbf{D}^L_{\mathbf{predicted}}(\mathbf{p})|, & \mathbf{p} \in \mathbb{A} \\ \mathbf{None}, & \mathbf{p} \in \mathbb{B} \end{cases}$$

Ο πίνακας «απόλυτου σφάλματος» έχει μονάδα μέτρησης pixel και οπτικοποιείται είτε μέσω ιστογράμματος των τιμών του 2.29ε′, είτε με εικόνα όπου ο χρωματισμός κάθε θέσης εξαρτάται από το αντίστοιχο «απόλυτο σφάλμα» 2.29ζ′.

2.8.2 Απόλυτο σφάλμα πρόβλεψης με ανώφλι (Absolute prediction error with threshold)

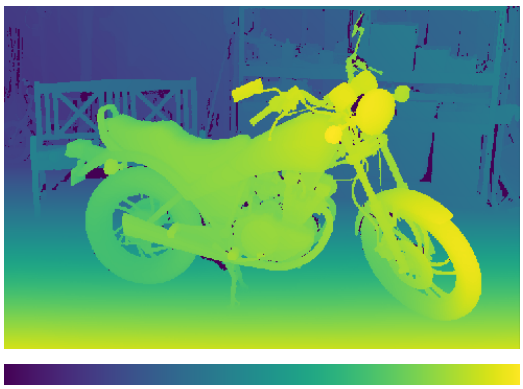
Ο πίνακας «απόλυτου σφάλματος με ανώφλι» $\mathbf{AD}_{\mathbf{threshold}}$ αντιστοιχίζει κάθε σημείο $\mathbf{p}$ της εικόνας σε τρεις διακριτές τιμές με τον ακόλουθο τρόπο:

$$\mathbf{AD}_{\mathbf{threshold}}(\mathbf{p}) = \begin{cases} 1, & \mathbf{p} \in \mathbb{A} \text{ και } |\mathbf{D}^L_{\mathbf{ground\_truth}}(\mathbf{p}) - \mathbf{D}^L_{\mathbf{predicted}}(\mathbf{p})| > \texttt{threshold} \\ 0.5, & \mathbf{p} \in \mathbb{A} \text{ και } |\mathbf{D}^L_{\mathbf{ground\_truth}}(\mathbf{p}) - \mathbf{D}^L_{\mathbf{predicted}}(\mathbf{p})| \leqslant \texttt{threshold} \\ 0, & \mathbf{p} \in \mathbb{B} \end{cases}$$

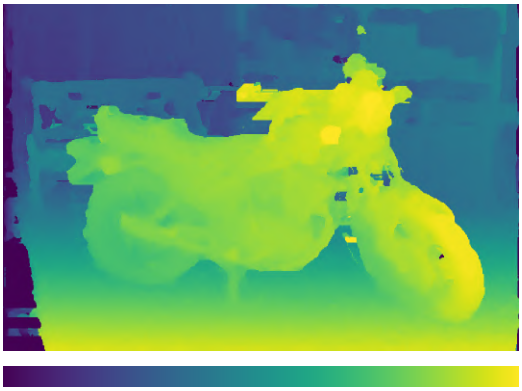
Η τιμή του ορίου `threshold` ορίζει την ανοχή στο σφάλμα πρόβλεψης. Ο πίνακας «απόλυτου σφάλματος με ανώφλι» οπτικοποιείται ως μια εικόνα τριών διακριτών χρωματισμών. 2.29η′

(Α′) $\mathbf{I}^L$

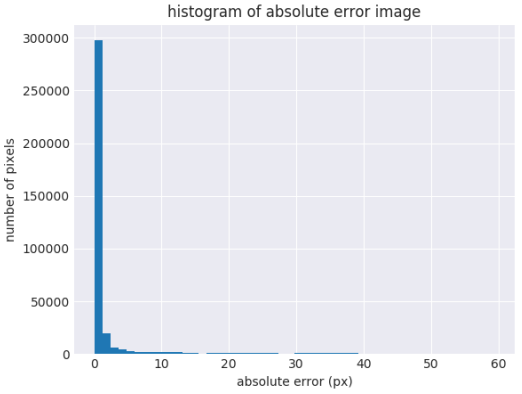
(Β′) $\mathbf{I}^R$



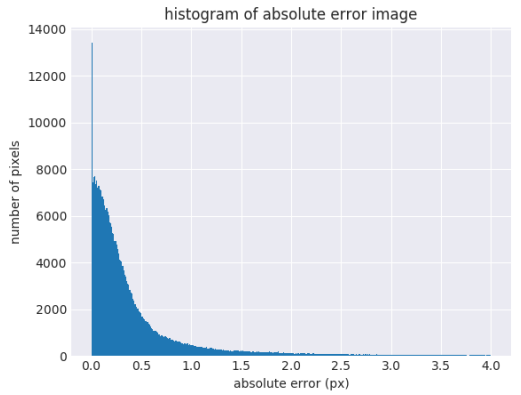
(Γ′) $\mathbf{D}^L_{\mathbf{ground\_truth}}$



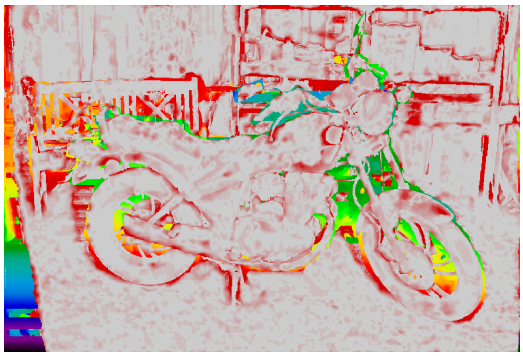
(Δ′) $\mathbf{D}^L_{\mathbf{predicted}}$



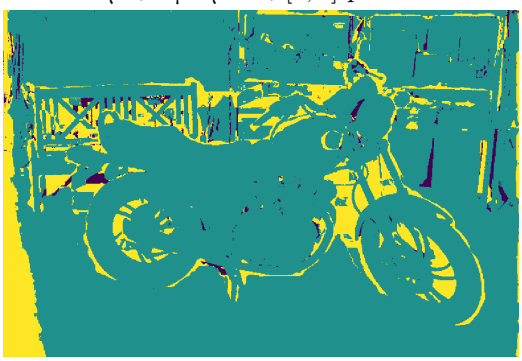
(Ε′) Ιστόγραμμα πίνακα απόλυτου σφάλματος **AD**



(Ϛ′) Ιστόγραμμα πίνακα **AD** εστιασμένο στις τιμές σφάλματος $[0, 4]$ pixels.



(Ζ′) $\mathbf{AD} = |\mathbf{D}^L_{\mathbf{ground\_truth}} - \mathbf{D}^L_{\mathbf{predicted}}|$



(Η′) $\mathbf{AD}_{\mathbf{threshold}}$

Σχημα 2.29: Παράδειγμα οπτικοποίησης όλων των μεθόδων αξιολόγησης του υπολογισμένου χάρτη παράλλαξης. Στο συγκεκριμένο παράδειγμα οι χρωματικές κλίμακες αναπαριστούν τιμές παράλλαξης ή σφάλματος εντός του διαστήματος $[0, 63]$ pixels . Το μέσο απόλυτο σφάλμα είναι $2.058px$ και το ποσοστό απόλυτου σφάλματος $> 3px$ είναι 9.891%

2.8.3 Προσδιορισμός ακρίβειας χάρτη παράλλαξης με μία τιμή

Επιχειρούμε να συμπυκνώσουμε την πληροφορία των πινάκων «απόλυτου σφάλματος» και «απόλυτου σφάλματος με ανώφλι» σε μία μοναδική τιμή η οποία θα αποτελεί τον δείκτη ποιότητας του υπολογισμένου χάρτη παράλλαξης. Η επικρατέστερη μεθοδολογία χρησιμοποιεί την μέση τιμή του αντίστοιχου πίνακα.

- το μέσο απόλυτο σφάλμα $\mathbf{AD_{average}}$, είναι η μέση τιμή του πίνακα $\mathbf{AD}$ υπολογισμένη επί του συνόλου $\mathbb{A}$:

$$\mathbf{AD_{average}} = \frac{1}{\text{πληθυσμός}(\mathbb{A})} \sum_{\mathbf{p}\in\mathbb{A}} |\mathbf{D}^L_{\mathbf{ground\_truth}}(\mathbf{p}) - \mathbf{D}^L_{\mathbf{predicted}}(\mathbf{p})|$$

- το ποσοστό σφάλματος (error percentage), είναι το ποσοστό σημείων του πίνακα $\mathbf{AD_{threshold}}$ με τιμή 1 επί του συνόλου $\mathbb{A}$:

$$\mathbf{error\_percentage} = \left[\frac{1}{\text{πληθυσμός}(\mathbb{A})} \sum_{\mathbf{p}\in\mathbb{A}} 1\{\mathbf{AD_{threshold}}(\mathbf{p}) = 1\} \right] \times 100\%$$

2.8.4 Παρατηρήσεις

- Το ποσοστό των σημείων που ανήκουν στο σύνολο $\mathbb{A}$ ως προς τα σημεία που ανήκουν στο σύνολο $\mathbb{C}$, ονομάζεται πυκνότητα (density) του πραγματικού χάρτη παράλλαξης $\mathbf{D_{ground\_truth}}$. Η πυκνότητα χαρακτηρίζει την συλλογή δεδομένων, για παράδειγμα στις συλλογές KITTI 2012 και KITTI 2015 η πυκνότητα είναι περίπου 40% ενώ στην συλλογή Middlebury είναι $> 95\%$. Στις συνθετικές συλλογές δεδομένων, η πυκνότητα είναι συνήθως 100%.

- Οι παραπάνω μετρικές εφαρμόζονται με δύο παραλλαγές, είτε σύνολο της εικόνας, είτε μόνο στα σημεία της που απεικονίζονται και στις δύο λήψεις. Αν θεωρήσουμε εικόνα αναφοράς την αριστερή(δεξιά) εικόνα $\mathbf{I}^{L(R)}$, ένα ποσοστό των σημείων της που βρίσκονται στο αριστερό(δεξιό) άκρο της δεν απεικονίζονται καθόλου στην δεξιά(αριστερή) εικόνα. Για τα σημεία αυτά δεν υπάρχει κανένα στοιχείο στερεοσκοπικής φύσης για τον προσδιορισμό του βάθους τους, το οποίο συνήθως υπολογίζεται με παρεμβολή των τιμών των γειτονικών στοιχείων. Όπως φαίνεται και στην εικόνα 2.29η΄, η αστοχία στην πρόβλεψη αυτών των σημείων είναι έντονη έως καθολική. Αν η αξιολόγηση εφαρμόζεται στο σύνολο της εικόνας, τα σημεία αυτά συνυπολογίζονται ενώ στην έτερη περίπτωση εξαιρούνται.

- Το «απόλυτο σφάλμα με ανώφλι» έχει περισσότερη αξία όταν δεν μας ενδιαφέρει η απόλυτη ακρίβεια, αλλά το να μην υπάρχουν έντονα άστοχες προβλέψεις. Για παράδειγμα, αν στόχος της μεθοδολογίας είναι η πλοήγηση στο χώρο, δεν μας προβληματίζει μια αστοχία πρόβλεψης βάθους της τάξης του μισού μέτρου για ένα αντικείμενο που βρίσκεται στα 10 μέτρα ή των 0.01 μέτρων για ένα αντικείμενο στα 50 εκατοστά από το πέτασμα της κάμερας. Αντιθέτως προβληματίζει μια έντονη απόκλιση. Σε αυτές τι περιπτώσεις η μετρική του απόλυτου σφάλματος με ανώφλι αποτελεί την καταλληλότερη αξιολόγηση.

- Οι τιμές του «απόλυτου σφάλματος» έχουν μονάδα μέτρησης pixel επομένως η αντιστοιχία σε πραγματική απόσταση mm συναρτάται των διαστάσεων του pixel. Για παράδειγμα, σε δύο στερεοσκοπικά ζεύγη που έχουν ληφθεί υπό τις ίδιες ακριβώς παραμέτρους στερεοσκοπικής διάταξης (B,f) με διαφορετικές αναλύσεις $720p$ και $2 \times 720p$, απόλυτο σφάλμα 1 px για την πρώτη εικόνα αναλογεί σε σφάλμα 2 px στην δεύτερη. Το ίδιο ισχύει και αν εφαρμόσουμε τεχνικές αλλαγής μεγέθους της αρχικής εικόνας, όπως υποδειγματοληψία, μια τεχνική που συνηθίζεται στα συνελικτικά νευρωνικά δίκτυα όπου η μνήμη της κάρτας γραφικών είναι περιορισμένη. Για παράδειγμα, τα στερεοσκοπικά ζεύγη της συλλογής Middlebury 2014 έχουν ανάλυση περίπου 2000×3000 px. Η προώθησή τους στο συνελικτικό νευρωνικό δίκτυο που παρουσιάζεται στην εργασία, απαιτεί την υποδειγματοληψία τους στο $1/2$ για μια κάρτα γραφικών με μνήμη ≈ 12 **GB** και στο $1/4$ για μνήμη ≈ 6 **GB**. Τα τελικά αποτελέσματα «απόλυτου σφάλματος» θα πρέπει να πολλαπλασιαστούν με τους συντελεστές 2 και 4 αντίστοιχα, ώστε να ισοδυναμούν με το «απόλυτο σφάλμα» στην μέγιστη ανάλυση.
- Ο υπολογισμός του πραγματικού σφάλματος στον υπολογισμό του βάθους, με δεδομένο το «απόλυτο σφάλμα πρόβλεψης» εξαρτάται από τρία μεγέθη:
  - την οριζόντια απόσταση B των δύο λήψεων της στερεοσκοπικής διάταξης
  - την εστίαση f στις δύο λήψεις
  - την πραγματική απόσταση κατά τον άξονα zz' (z_{real})του εικονιζόμενου σημείου από τις δύο λήψεις

  και υπολογίζεται από τον τύπο:

$$\left| \frac{1}{z_{real}} - \frac{1}{z_{predicted}} \right| = \frac{|d_{real} - d_{predicted}|}{Bf}$$

  Παίρνουμε ως παράδειγμα το στερεοσκοπικό ζεύγος 2.29α′ το οποίο έχει ληφθεί με παραμέτρους $B = 193mm$ και $f = 3997.68px$. Αν ένα αντικείμενο βρίσκεται στα $z_{real} = 10m$ απόσταση από το οπτικό κέντρο έχει πραγματική τιμή παράλλαξης $d_{real} = 77,155px$. Αν η πρόβλεψη μας είναι κατά $3px$ μεγαλύτερη, δηλαδή $d_{predicted} = 80,155px$, σημαίνει απόλυτο σφάλμα βάθους $0,374m$, δηλαδή πρόβλεψη βάθους $z_{predicted} = 9.626m$. Αν το αντικείμενο βρισκόταν στα $z_{real} = 5m$, θα είχε $d_{real} = 154.31px$, και το ίδιο απόλυτο σφάλμα με πριν $3px$ θα σήμαινε απόλυτο σφάλμα βάθους $0,095m$, περίπου 4 φορές μικρότερο σε σχέση με πριν.

  Με αφορμή την παρατήρηση της εξάρτησης του σφάλματος βάθους από την πραγματική απόσταση του αντικειμένου από το οπτικό κέντρο z_{real}, μπορούμε να δημιουργήσουμε μια νέα κατηγορία μετρικών που μετρούν το «απόλυτο σφάλμα πρόβλεψης» και «απόλυτο σφάλμα πρόβλεψης με ανώφλι» στο επίπεδο του πραγματικού βάθους κατά τον άξονα zz' αντί της τιμής της παράλλαξης.

Κεφάλαιο 3

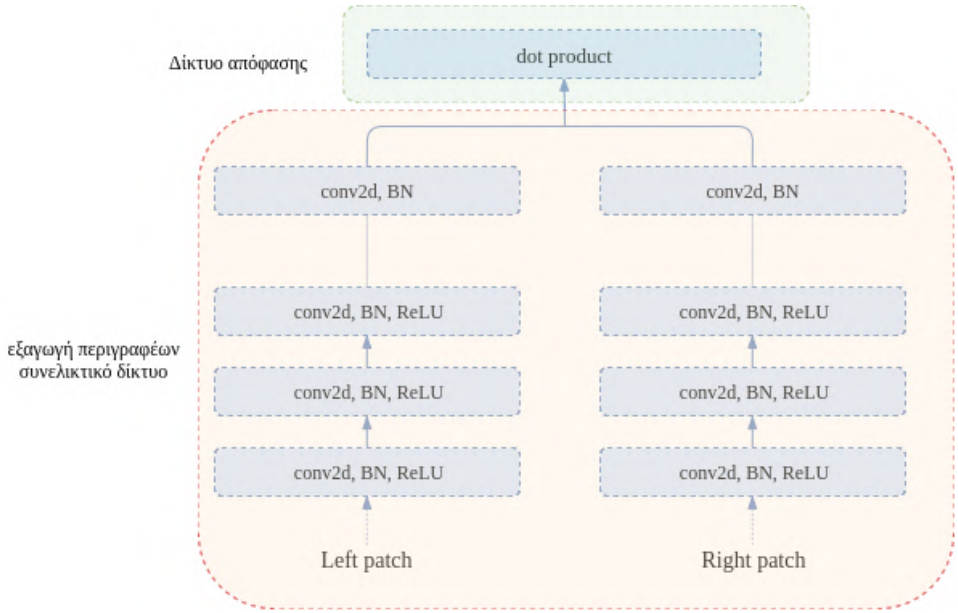
Θεωρητική ανάλυση τεχνητού νευρωνικού δικτύου

3.1 Χρήση νευρωνικών δικτύων στη στερεοσκοπική όραση

Η αρχικοποίηση του πίνακα κόστους με διαφορετικές τεχνικές μας βοηθάει να εξάγουμε ένα χρήσιμο συμπέρασμα: η απόδοση της μετρικής ομοιότητας μπορεί να βελτιωθεί αν η σύγκριση δεν βασιστεί στις αρχικές τιμές της φωτεινότητας των υπό σύγκριση χωρίων, αλλά σε ένα πιο αξιόπιστο περιγραφέα της γειτονιάς. Ως αξιόπιστο περιγράφουμε έναν περιγραφέα που είναι όσο το δυνατόν λιγότερο ευάλωτος στα φαινόμενα που προκαλούν αλλοίωση της «ομοιότητας γειτονιάς», όπως αυτά αναλύθηκαν στο προηγούμενο κεφάλαιο 2.4. Για παράδειγμα, ο μετασχηματισμός census [43] δημιουργεί τοπικό περιγραφέα που είναι ανεπηρέαστος από τις φωτομετρικές αποκλίσεις. Ταυτόχρονα όμως έχει το μειονέκτημα να δημιουργεί παρόμοιο περιγραφέα από τελείως διαφορετικά είδωλα που τυχαίνει να δημιουργούν γειτονιές με παρόμοια σχέση φωτεινότητας περιφέρειας και κεντρικού pixel. Η παρατήρηση ότι κάθε μέθοδος έχει διαφορετικά πλεονεκτήματα και αδυναμίες προτρέπει τον συνδυασμό μεθόδων στον υπολογισμό του τελικού κόστους για πιο αξιόπιστα αποτελέσματα, όπως επιτυχημένα υλοποιεί η μέθοδος AD-census [28].

Το πρόβλημα της εξαγωγής του πιο αξιόπιστου τοπικού περιγραφέα είναι ιδιαίτερα σύνθετο. Το εύρος των πιθανών επιλογών είναι χαοτικά μεγάλο και η βέλτιστη λύση είναι αδύνατο να προβλεφθεί από έναν προγραμματιστή. Αυτή η παρατήρηση μας οδηγεί στην αντιμετώπιση του προβλήματος με μεθόδους μηχανικής μάθησης (machine learning) και πιο συγκεκριμένα με τη χρήση συνελικτικών νευρωνικών δικτύων (convolutional neural networks). Έτσι δίνεται η δυνατότητα δόμησης ενός αλγορίθμου που θα δημιουργήσει μόνος του τον βέλτιστο περιγραφέα, μαθαίνοντάς τον από τα δεδομένα. Προϋπόθεση για να συμβεί αυτό είναι η ύπαρξη ικανοποιητικά μεγάλων συλλογών δεδομένων με στερεοσκοπικά ζεύγη που θα περιέχουν την πραγματική πληροφορία παράλλαξης. Τέτοιες συλλογές έχουν δημιουργηθεί και είναι διαθέσιμες τα τελευταία χρόνια.

Η αρχικοποίηση του κόστους αντιστοίχησης με χρήση συνελικτικών νευρωνικών δικτύων έχει δοκιμαστεί κυρίως τα τελευταία τρία χρόνια, έχοντας αποδώσει εξαιρετικά αποτελέσματα. Οι Zagoruyko, Komodakis [44] πρότειναν τρεις διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων για την σύγκριση τετράγωνων περιοχών εικόνας. Β′.1 Οι αρχιτεκτονικές τους εφαρμόστηκαν για την επίλυση προβλημάτων στερεοσκοπικής όρασης μεγάλης απόστασης βάσης (wide baseline).Οι Zbontar, Lecun [45] Β′.2 πρότειναν επίσης δύο αρχιτεκτονικές για την αρχικοποίηση του κόστους σε πρόβλημα μικρής απόστασης βάσης



Σχημα 3.1: Αρχιτεκτονική νευρωνικού δικτύου.

(small baseline).Οι Luo et. al [24] Β′.3, από την αρχιτεκτονική των οποίων έχει εμπνευστεί κι η παρούσα εργασία, αντιμετώπισαν το πρόβλημα της αρχικοποίησης του κόστους ως πρόβλημα ταξινόμησης πολλαπλών κατηγοριών. Τέλος, οι Alex Kendall et al [18] Β′.5 και οι Gydaris, Komodakis [8] Β′.4 αντιμετώπισαν το πρόβλημα του υπολογισμού του χάρτη παράλλαξης με χρήση συνελικτικών νευρωνικών δικτύων από την αρχή ως το τέλος. Οπτική αναπαράσταση των γνωστότερων αρχιτεκτονικών που χρησιμοποίηθηκαν για την αρχικοποίηση του πίνακα κόστους παρατίθεται στο παράρτημα Β′.1.

3.2 Αρχιτεκτονική νευρωνικού δικτύου

Προσεγγίζουμε το πρόβλημα της αρχικοποίησης του πίνακα κόστους, ως πρόβλημα ταξινόμησης πολλαπλών κατηγοριών. Κάθε pixel $\mathbf{p}$ της λήψης αναφοράς αντιστοιχίζεται (ταξινομείται) σε μία τιμή (κατηγορία) του συνόλου $\{0, 1, 2, \ldots, \texttt{max\_disparity}\}$. Το νευρωνικό δίκτυο χωρίζεται σε δύο μέρη, το συνελικτικό δίκτυο το οποίο αναλαμβάνει την εξαγωγή του τοπικού περιγραφέα και το δίκτυο απόφασης στο οποίο υπολογίζεται η τιμή του κόστους ομοιότητας σε κάθε θέση, όπως φαίνεται στο σχήμα 3.1.

3.2.1 Εξαγωγή τοπικών περιγραφέων - Συνελικτικό νευρωνικό δίκτυο

Στόχος της εξαγωγής τοπικού περιγραφέα είναι η αντιστοίχηση του τετράγωνου χωρίου πέριξ του σημείου ενδιαφέροντος $\mathbf{p}$, δηλαδή της γειτονιάς N_p, σε ένα διάνυσμα το οποίο συμβολίζουμε $\mathbf{I_{descriptor}}(\mathbf{p})$. Το μέγεθος της πλευράς της τετράγωνης γειτονιάς `patch_size` και το μέγεθος του τελικού διανύσματος `f_maps` είναι ελεύθερες παράμετροι προς επιλογή. Επομένως:

$$\mathbf{I_{descriptor}}(\mathbf{p}) = f_{desc}(N_p)$$

$$N_p \in \mathbb{R}^{\texttt{patch\_size}^2}, \mathbf{I_{descriptor}}(\mathbf{p}) \in \mathbb{R}^{\texttt{f\_maps}}$$

Κάθε μπλοκ του συνελικτικού νευρωνικού δικτύου απαρτίζεται από τα εξής επίπεδα:

- **Δισδιάστατη συνέλιξη**: Συνέλιξη με `f_maps` φίλτρα διαστάσεων [`kernel_size`× `kernel_size`×`f_maps`][1]. Η συνέλιξη εφαρμόζεται χωρίς zero padding στον πίνακα εισόδου. Συμβολίζουμε την πράξη ως:

$$y_{conv2d} = \mathbf{conv2d}(x, h)$$

- **Πόλωση:** Στο αποτέλεσμα της συνέλιξης του σήματος εισόδου με το κάθε φίλτρο προστίθεται ένας όρος (πόλωση), μετατρέποντας την συνολική πράξη σε μετασχηματισμό affine.

$$y_b = y_{conv2d} + b$$

- **Κανονικοποίηση δέσμης (batch normalization):** [14] Η κανονικοποίηση δέσμης αφήνει αμετάβλητες τις διαστάσεις του εισερχόμενου πίνακα, επηρεάζοντας μόνο την εσωτερική του στατιστική. Αναλυτική περιγραφή της μεθόδου και των σκοπών που επιτελεί δίνεται στο παράρτημα Β′.2.

$$y_{BN} = BN(y_b)$$

- **Συνάρτηση γραμμικού ανορθωτή (Rectified Linear Unit - ReLU):**Εισάγει την απαραίτητη μη γραμμικότητα μέσω της πράξης $\mathrm{ReLU}(x) = max(0, x)$σε κάθε τιμή του πίνακα εισόδου.

Η παραπάνω δομή επιπέδων επαναλαμβάνεται διαδοχικά `num_conv_layers` φορές. Έπειτα από κάθε μπλοκ, οι χωρικές διαστάσεις του χωρίου μειώνονται κατά $\frac{\texttt{kernel\_size} - 1}{2}$. Στο τέλος ολόκληρου του συνελικτικού νευρωνικού δικτύου, ο πίνακας που θα προκύψει θα έχει διαστάσεις $[1 \times 1 \times \texttt{f\_maps}]$ κι ουσιαστικά θα είναι ο τοπικός περιγραφέας του χωρίου που δόθηκε ως είσοδος στο δίκτυο.

3.2.2 Δίκτυο απόφασης

Στόχος του δικτύου απόφασης είναι η εκτίμηση της ομοιότητας των δύο περιγραφέων, που αναλογούν στα δύο υπό σύγκριση χωρία. Η εκτίμηση αυτή είναι το αποτέλεσμα μια συνάρτησης f_{comp}, που θα δέχεται ως είσοδο τους δύο τοπικούς περιγραφείς $\mathbf{I}^L_{descriptor}(\mathbf{p})$, $\mathbf{I}^R_{descriptor}(\mathbf{q})$ και θα επιστρέφει μια εκτίμηση ομοιότητας:

$$f_{comp} : \mathbb{R}^{2\times \texttt{f\_maps}} \rightarrow \mathbb{R}$$
$$s = f_{comp}(\mathbf{I}^L_{descriptor}\mathbf{p}, \mathbf{I}^R_{descriptor}\mathbf{q})$$

Η επιλογή της συνάρτησης αυτής μπορεί να γίνει με δύο τρόπους:

- Με την εφαρμογή μιας προαποφασισμένης συνάρτησης f_{comp} όπως η μέση απόλυτη διαφορά, η μέση τετραγωνική διαφορά, το εσωτερικό γινόμενο ή η ομοιότητα συνημιτόνου.

[1] Εκτός από το πρώτο μπλοκ, στο οποίο το φίλτρο συνέλιξης έχει διάσταση `kernel_size`×`kernel_size`× 1, καθώς δέχεται ως είσοδο grayscale τετράγωνο χωρίο διάστασης `patch_size` × `patch_size` × 1

- Με την χρήση μηχανικής μάθησης για την εκμάθηση της βέλτιστης συνάρτησης f_{comp} από τα δεδομένα. Στην περίπτωση, μια ενδεδειγμένη λύση είναι η εκπαίδευση ενός «πλήρως συνδεδεμένου» (fully connected) τεχνητού νευρωνικού δικτύου.

Ανάλυση της χρήσης τεχνητού νευρωνικού δικτύου ως δίκτυο απόφασης

- Η εκπαίδευση ενός νευρωνικού δικτύου για την εκτίμηση της ομοιότητας αποτελεί την βέλτιστη επιλογή με κριτήριο την ακρίβεια. Η επιλογή αυτή δίνει τη δυνατότητα στο δίκτυο να «μάθει» από τα δεδομένα την βέλτιστη συνάρτηση f_{comp} που θα περατώνει τον σκοπό της εκτίμησης ομοιότητας των δύο διανυσμάτων εισόδου.
- Από πλευράς ταχύτητας, η βέλτιστη επιλογή είναι η χρήση μιας προαποφασισμένης συνάρτησης, όπως για παράδειγμα το εσωτερικό γινόμενο. Η σύγκριση εκτελείται σειριακά $\texttt{max\_disparity} + 1$ φορές για κάθε σημείο της εικόνας, επομένως αν ο χρόνος περάτωσης της συνάρτησης σύγκρισης είναι t_f, ο συνολικός χρόνος θα είναι $(\texttt{max\_disparity} + 1) \times t_f$. Η χρήση μιας απαιτητικής υπολογιστικά συνάρτησης με πολύ μεγάλο t_f, όπως το «απόλυτα συνδεδεμένο» νευρωνικό δίκτυο αυξάνει πολύ έντονα το χρόνο εκτέλεσης. Αντιθέτως, μια συνάρτηση όπως το εσωτερικό γινόμενο, αφενός λόγω πολύ μικρού t_f κρατάει τον συνολικό χρόνο σε χαμηλά επίπεδα ακόμη και για μεγάλο `max_disparity` αφετέρου μπορεί να υλοποιηθεί παράλληλα σε κάρτα γραφικών.
- Η εκπαίδευση του συνελικτικού δικτύου και του δικτύου απόφασης γίνεται ενιαία. Αυτό προϋποθέτει τη χρήση μιας παραγωγίσιμης συνάρτησης f_{comp}, ώστε να μπορεί να εφαρμοστεί πάνω της ο κανόνας της οπισθοδιάδοσης (back propagation).

Παρατηρώντας ότι τα αποτελέσματα είναι ικανοποιητικά με χρήση μιας πράξης εσωτερικού γινομένου στα δύο διανύσματα, με ταυτόχρονη κατακόρυφη μείωση της υπολογιστικής πολυπλοκότητας το προτιμούμε σε σχέση με την εκπαίδευση ενός «απόλυτα συνδεδεμένου» νευρωνικού δικτύου.

3.3 Εκπαίδευση νευρωνικού δικτύου

3.3.1 Δημιουργία σετ εκπαίδευσης νευρωνικού δικτύου

Στο παράρτημα Β′.3 παραθέτουμε εκτενή ανάλυση των χαρακτηριστικών που διέπουν τις υπάρχουσες στερεοσκοπικές συλλογές.

Η δημιουργία του σετ εκπαίδευσης βασίζεται στις εικόνες της στερεοσκοπικής συλλογής KITTI με την ακόλουθη προεπεξεργασία σε κάθε εικόνα:[2]

- Μετατροπή τους σε εικόνες μηδενικής μέσης τιμής. Στην αρχική τους μορφή όλες οι εικόνες είναι πίνακες ακεραίων αριθμών $I \in \mathbb{Z} : \{0 \leqslant I \leqslant 255\}$. Εφαρμόζουμε την πράξη:

$$\mu = \frac{1}{M \times N} \cdot \sum_{\mathbf{p}} I(\mathbf{p})$$

[2] όλες οι εικόνες στις οποίες αναφερόμαστε είναι grayscale

$$I^{\mathbf{z\_m}} = I - \mu$$

Ο νέος πίνακας έχει σύνολο τιμών το σύνολο των πραγματικών αριθμών $I^{\mathbf{z\_m}} \in \mathbb{R}$.

- Κανονικοποίηση (normalization) μέσω της μετατροπής τους σε εικόνες μοναδιαίας διακύμανσης. Εφαρμόζουμε την πράξη:

$$\sigma^2 = \frac{1}{M \times N} \cdot \sum_{\mathbf{p}} I_{\mathbf{z\_m}}(\mathbf{p})$$

$$I^{\mathbf{unit\_var}} = \frac{I^{\mathbf{zm}}}{\sigma}$$

Επισημάνσεις:

- Η μετατροπή σε εικόνες μηδενικής μέσης τιμής και μοναδιαίας διακύμανσης γίνεται στο σύνολο της εικόνας κι όχι στο κάθε τετράγωνον χωρίο (patch) που θα εξαχθεί κατά τη δημιουργία του σετ εκπαίδευσης. Αυτό συμβαίνει διότι κατά την εκτέλεση του αλγορίθμου θα προωθούμε ολόκληρη την εικόνα στο δίκτυο (όχι κάθε χωρίο ξεχωριστά) και θέλουμε το δίκτυο να εκπαιδευτεί σε δεδομένα ίδιας στατιστικής με αυτά που θα αντιμετωπίσει κατά τον έλεγχο (testing).
- Οι υπολογισμοί της μέσης τιμής και της τυπικής απόκλισης γίνεται σε κάθε εικόνα ξεχωριστά. Δεν ακολουθούν τον ορισμό της κανονικοποίησης που ορίζει την κανονικοποίηση της κάθε διάστασης (κάθε ξεχωριστό pixel στην προκειμένη) κατά μήκος όλου του σετ εκπαίδευσης. Αυτή η εναλλακτική μορφή κανονικοποίησης είναι η πιο διαδεδομένη μέθοδος προεπεξεργασίας όταν η συλλογή περιλαμβάνει εικόνες.

Συμβολίζουμε με

$$< \mathcal{P}^L_{n\times n}(\mathbf{p}), \mathcal{P}^R_{(\texttt{max\_disparity}+n)\times n}(\mathbf{q}), \text{label} >$$

κάθε εγγραφή του σετ εκπαίδευσης, όπου:

- $\mathcal{P}^L_{n\times n}(\mathbf{p})$, είναι το τετράγωνο χωρίο διάστασης $n \times n$ της αριστερής εικόνας με κέντρο το σημείο $\mathbf{p}$
- $\mathcal{P}^R_{(\texttt{max\_disparity}+n)\times n}(\mathbf{q})$, είναι το παραλληλόγραμμο χωρίο διάστασης $\texttt{max\_disparity}+n) \times n$ της δεξιάς λήψης. Ουσιαστικά περιέχει τα τετράγωνα χωρία που ορίζονται γύρω από όλες τις υποψήφιες θέσεις παράλλαξης $\mathbf{p} - \mathbf{d}$.
- label, συμβολίζουμε ένα διάνυσμα $\texttt{max\_disparity} + 1$ θέσεων το οποίο λαμβάνει τιμές με τον ακόλουθο κανόνα:

$$label[i] = \begin{cases} 0.5 & \text{εάν } i = true\_disparity \\ 0.2 & \text{εάν } |i - true\_disparity| = 1 \\ 0.05 & \text{εάν } |i - true\_disparity| = 2 \\ 0 & \text{αλλιώς} \end{cases}$$

Ο λόγος που επιλέγουμε αυτήν την κατανομή στο διάνυσμα label , αντί του συνηθισμένου one-hot encoding οφείλεται στο ότι επιθυμούμε το δίκτυο να υπολογίζει μικρό κόστος

Σχημα 3.2: Αριστερή εικόνα I^L στερεοσκοπικού ζεύγους της συλλογής KITTI (2012)

(Α′) $\mathcal{P}^L_{150\times150}(163, 731)$ (Β′) $\mathcal{P}^R_{(150+150)\times150}(163, 701)$

Σχημα 3.3: παράδειγμα δημιουργίας εγγραφής στο σετ εκπαίδευσης ταξινόμησης πολλαπλών κατηγοριών. Η επιλογή μεγέθους ορθογώνιου χωρίου 150×150 δεν είναι ρεαλιστική, αλλά έγινε για να είναι εύληπτη η εικόνα. Στην πραγματικότητα το μέγεθος του χωρίου δεν υπερέβη ποτέ τα $n = 40$pixels στους πειραματισμούς μας.

ομοιότητας όχι μόνο στην σωστή παράλλαξη, αλλά και στις γειτονικές περιοχές εντός του ορίου ± 2 θέσεων.

Αναλυτική περιγραφή με αλγοριθμικές λεπτομέρειες της δημιουργίας των παραδειγμάτων εκπαίδευσης δίνεται στο παράρτημα Β′.4.

Ένα παράδειγμα εγγραφής του σετ εκπαίδευσης φαίνεται στο σχήμα 3.2. Δημιουργούνται περίπου $8 \cdot 10^4 \times \frac{\text{εγγραφές}}{\text{εικόνα}}$ οπότε το συνολικό σετ εκπαίδευσης περιέχει περίπου 32×10^6 εγγραφές.

3.3.2 Υπολογισμός συνάρτησης κόστους προς ελαχιστοποίηση

Συμβολίζουμε ως Θ το σύνολο των εκπαιδεύσιμων παραμέτρων του νευρωνικού δικτύου. Παρακάτω αναλύουμε πως δομείται η συνάρτηση κόστους $J_X(\theta)$ την οποία θα ελαχιστοποιήσουμε.

Κάθε εγγραφή του σετ εκπαίδευσης συμβολίζεται ως

$$< \mathcal{P}^L_{n\times n}(\mathbf{p}), \mathcal{P}^R_{(\texttt{max\_disparity}+n)\times n}(\mathbf{q}), \text{label} >$$

Block	Περιγραφή Επιπέδων	$\mathcal{P}^L_{n\times n}(\mathbf{p})$	$\mathcal{P}^R_{(\texttt{max\_disparity}+n)\times n}(\mathbf{q})$
Local descriptors extraction (Εξαγωγή περιγραφέων)			
	Siamese network (Σιαμαίο δίκτυο)		
Είσοδος		$n \times n \times 1$	$(d+n) \times n \times 1$
1	conv2d, F=64	$(n-x) \times (n-x) \times F$	$(d+n-x) \times (n-x) \times F$
2	conv2d, F=64	$(n-2x) \times (n-2x) \times F$	$(d+n-2x) \times (n-2x) \times F$
i	conv2d, F=64	$(n-ix) \times (n-ix) \times F$	$(d+n-ix) \times (n-ix) \times F$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1/x$	conv2d, F=64	$1 \times 1 \times F$	$(d+1) \times 1 \times F$
Έξοδος		F	$(d+1) \times F$

TABLE 3.1: Περίληψη επιπέδων δικτύου κι αντίστοιχων διαστάσεων πινάκων κατά την εκπαίδευση. Ως d συμβολίζεται η μέγιστη παράλλαξη `max_disparity`, ως x την μείωση των διαστάσεων του πίνακα εισόδου κατά τη συνέλιξη $\frac{\texttt{kernel\_size}-1}{2}$, ως n τη διάσταση του χωρίου εισόδου `patch_size`.

Το συνελικτικό νευρωνικό δίκτυο αποτελείται από δύο όμοιους κλάδους (αρχιτεκτονική σιαμαίων δικτύων). Ο ένας κλάδος δέχεται ως είσοδο το τετράγωνο χωρίο διάστασης [`patch_size` × `patch_size` × 1] που αναλογεί στο $\mathcal{P}^L_{n\times n}(\mathbf{p})$ και ο άλλος κλάδος το παραλληλόγραμμο διάστασης [(`patch_size`+`max_disparity`) × `patch_size` × 1] που αναλογεί στο $\mathcal{P}^R_{(\texttt{max\_disparity}+n)\times n}(\mathbf{q})$. Στην έξοδο του συνελικτικού δικτύου λαμβάνουμε:

- τον περιγραφέα $\mathbf{I}^L_{descriptor}(\mathbf{p})$ του σημείου $\mathbf{p}$ της αριστερής λήψης. Το διάνυσμα αυτό αναπαρίσταται από έναν πίνακα διάστασης `f_maps`.
- τους περιγραφείς $\mathbf{I}^R_{descriptor}(\mathbf{p}-\mathbf{d})$ όλων των υποψήφιων «αντίστοιχων σημείων» της δεξιάς λήψης. Το διάνυσμα αυτό αναπαρίσταται από έναν πίνακα διάστασης (`max_disparity` + 1) × `f_maps`.

Η διαδοχή των βημάτων και των διαστάσεων των πινάκων φαίνεται συμπυκνωμένα στον πίνακα 3.1.

Εφαρμόζουμε την πράξη του εσωτερικού γινομένου (`max_disparity` + 1) φορές, υπολογίζοντας έτσι το διάνυσμα $\mathbf{score} \in \mathbb{R}^{\texttt{max\_disparity}+1}$. Το διάνυσμα **score** περιέχει το σκορ ομοιότητας του χωρίου αναφοράς με το αντίστοιχο χωρίο της έτερης λήψης, σε κάθε πιθανή θέση παράλλαξης.

Η μετατροπή των τιμών του διανύσματος $\mathbf{score} \in \mathbb{R}^{\texttt{max\_disparity}+1}$ σε πιθανότητες γίνεται μέσω της συνάρτησης softmax:

$$p_d = \frac{e^{\mathbf{score}_d}}{\sum_{d=0}^{\texttt{max\_disparity}} e^{\mathbf{score}_d}}, \quad \forall d \in [0, \texttt{max\_disparity}+1]$$

Μέσω αυτής της πράξης δημιουργείται ένα διάνυσμα πιθανοτήτων $\mathbf{poss} \in \mathbb{R}^{\texttt{max\_disparity}+1}$, με τιμές εντός του διανύσματος $[0,1]$ το άθροισμα των οποίων ισούται με την μονάδα $\sum_{d=0}^{\texttt{max\_disparity}} p_d = 1$.

Ορίζουμε συνάρτηση εντροπίας ως:

$$H = \sum_{d=0}^{\texttt{max\_disparity}} \text{label(d)} log(\mathbf{poss}(d))$$

Αναπαριστούμε την συνάρτηση κόστους του κάθε παραδείγματος εκπαίδευσης ως H_j. Σε κάθε βήμα εκπαίδευσης το συνολικό κόστος υπολογίζεται στο σύνολο της δέσμης εκπαίδευσης:

$$L = \sum_{j=0}^{\texttt{batch\_size}} H_j$$

κι είναι μια τιμή που εξαρτάται από:

- τις εκπαιδεύσιμες παραμέτρους του δικτύου, έστω ότι τις συμβολίζουμε με Θ
- τα παραδείγματα της δέσμης εκπαίδευσης με τις αντίστοιχες ετικέτες τους (labels), έστω ότι τα συμβολίζουμε ως X

Υπολογίζουμε σφάλμα γενίκευσης:

$$R = \lambda \sum_i \Theta_i^2$$

Η ολική συνάρτηση κόστους είναι το άθροισμα της εντροπίας υπολογισμένη στην δέσμη παραδειγμάτων X και του σφάλματος γενίκευσης:

$$J_X(\Theta) = L + R$$

Θεωρούμε δεδομένα τα παραδείγματα της δέσμης εκπαίδευσης κι αντιμετωπίζουμε ως μεταβαλλόμενο μέγεθος μόνο τις παραμέτρους του δικτύου(βάρη των φίλτρων, πολώσεις, συντελεστές γ, β της κανονικοποίησης δέσμης). Η συνάρτηση κόστους επομένως είναι μια συνάρτηση:

$$J_X(\Theta) \ : \mathbb{R}^n \to \mathbb{R}$$

Στόχος μας είναι να βρούμε τις τιμές εκείνες των παραμέτρων Θ που φτάνουν «κοντά» στην ελάχιστη τιμή της J:

$$|J_X^{optimal}(\Theta) - J_X^{min}(\Theta)| < \varepsilon \tag{3.1}$$

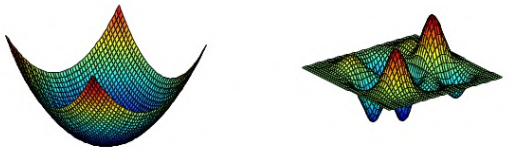
Στο παράρτημα Β΄.5 αναλύεται γιατί στοχεύουμε στον υπολογισμό ενός διανύσματος Θ στην γειτονιά του Θ_{min}.

3.3.3 Αρχικοποίηση των εκπαιδεύσιμων παραμέτρων του δικτύου

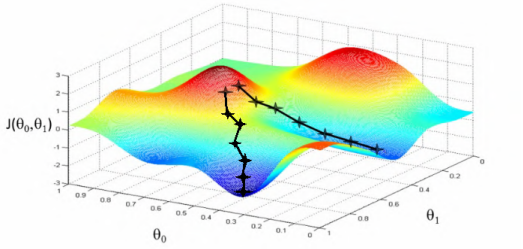
Η συνάρτηση $J_X(\Theta)$ είναι:

- συνεχής
- μη-κυρτή (non-convex)3.4α΄
- μη κοίλη (non-concave)

Η δεύτερη και τρίτη ιδιότητα δυσκολεύουν την προσπάθεια εύρεσης του ολικού ελάχιστου. Δεν υπάρχει κανένα εχέγγυο ότι ακολουθώντας την διεύθυνση της κλίσης της



(Α′) Παράδειγμα κυρτής (αριστερά) και μη-κυρτής (δεξιά) συνάρτησης.



(Β′) Διαφορετική αρχικοποίηση οδηγεί τον αλγόριθμο «απότομης καθόδου» σε διαφορετικό τελικό αποτέλεσμα.

συνάρτησης κόστους θα οδηγηθούμε στο ολικό ελάχιστο. Αντίθετα είναι σχεδόν βέβαιο ότι η μέθοδός μας θα «παγιδευτεί» σε κάποιο από τα τοπικά ελάχιστα. Αυτό ταυτόχρονα σημαίνει ότι για διαφορετικές αρχικές τιμές παραμέτρων το τελικό αποτέλεσμα θα είναι πιθανότατα διαφορετικό, όπως φαίνεται στην εικόνα 3.4β′.

Η πειραματικά επιβεβαιωμένη παρατήρηση που διευκολύνει την επίλυση του προβλήματος εντοπίζει ότι η συνάρτηση $J_X(\Theta)$ εμφανίζει σχετικά μικρή απόκλιση ανάμεσα στα τοπικά και στο ολικό της ελάχιστο. Η προσέγγιση ενός τοπικού ελάχιστου δεν θα δώσει ιδιαίτερα χειρότερο αποτέλεσμα από την προσέγγιση του ολικού ελαχίστου. Η παρατήρηση αυτή (μη αποδεδειγμένη μαθηματικά) δίνει αρκετές πιθανότητες η εκπαίδευση του δικτύου να σημειώσει πρόοδο και να καταλήξει σε ένα ικανοποιητικό αποτέλεσμα, εκκινώντας από αρκετά διαφορετικά σημεία. Παρ΄ όλα αυτά, η αρχικοποίηση των παραμέτρων κατέχει σημαντικό ρόλο στην εξέλιξη της εκπαίδευσης του δικτύου.

Αρχικοποιούμε το δίκτυο με τις ακόλουθες κατανομές:

- οι τιμές των φίλτρων του νευρωνικού δικτύου (βάρη) αρχικοποιούνται με την κατανομή που πρότειναν οι He et. al [11]. Η κατανομή αυτή είναι μια κανονική γκαουσιανή την οποία κλιμακώνουμε με την τιμή $^2/_n$, όπου n ο αριθμός των νευρώνων του προηγούμενου επιπέδου. Έτσι εξασφαλίζουμε ότι η κατανομή έχει διακύμανση $^2/_n$ κι η διακύμανση της κατανομής στην έξοδο του συγκεκριμένου επιπέδου θα είναι μοναδιαία.
- οι πολώσεις αρχικοποιούνται με μηδενικές τιμές
- οι κλιμακώσεις γ των επιπέδων κανονικοποίησης δέσμης με μονάδα
- οι μετατοπίσεις β των επιπέδων κανονικοποίησης δέσμης με μηδενικές τιμές

Οι εκπαιδεύσιμες(μεταβλητές) παράμετροι του δικτύου φαίνονται αναλυτικά στον πίνακα 3.2.

3.3.4 Βελτιστοποίηση των εκπαιδεύσιμων παραμέτρων του δικτύου

Έστω $\Theta_0 \in \mathbb{R}^n$ το διάνυσμα που περιέχει τις αρχικές τιμές των εκπαιδεύσιμων παραμέτρων του δικτύου. Κατά την εκπαίδευση, στοχεύουμε να οδηγηθούμε στο διάνυσμα $\Theta^{trained} \in \mathbb{R}^n$ για το οποίο ισχύει:

$$|J(\Theta^{trained}) - J_{min}| < \varepsilon$$

Συνελικτικό νευρωνικό δίκτυο - Εξαγωγή τοπικού περιγραφέα		
	Επίπεδο	Εκπαιδεύσιμες παράμετροι
	block 1	
1	**conv2d**	$\texttt{patch\_size}^2 \times 1 \times \texttt{f\_maps}$
2	**biases**	$\texttt{f\_maps}$
3	**BN**	$2 \times \texttt{f\_maps}$
4	**ReLU**	0
	block 2	
1	**conv2d**	$\texttt{patch\_size}^2 \times \texttt{f\_maps}^2$
2	**biases**	$\texttt{f\_maps}$
3	**BN**	$2 \times \texttt{f\_maps}$
4	**ReLU**	0
⋮	⋮	⋮
	block n_blocks	
1	**conv2d**	$\texttt{patch\_size}^2 \times \texttt{f\_maps}^2$
2	**biases**	$\texttt{f\_maps}$
3	**BN**	$2 \times \texttt{f\_maps}$
	Δίκτυο απόφασης	
1	**dot product**	0
	Σύνολο	
	$(\texttt{patch\_size}^2 \times \texttt{f\_maps}^2 + 3\texttt{f\_maps})\texttt{n\_blocks}$	

TABLE 3.2: Περίληψη εκπαιδεύσιμων παραμέτρων του νευρωνικού δικτύου.

Αξιοποιούμε ότι η συνάρτηση J είναι συνεχής και το διάνυσμα κλίσης της $\nabla J(\Theta)$ είναι άμεσα υπολογίσιμο μέσω του κανόνα της αλυσίδας (οπισθοδιάδοση).

Έστω:

- $\Theta_i \in \mathbb{R}^n$: το διάνυσμα όλων των n εκπαιδεύσιμων παραμέτρων του δικτύου κατά το i-οστό βήμα του αλγορίθμου
- η: η βαθμωτή ποσότητα που δείχνει την ταχύτητα ανανέωσης των παραμέτρων
- $J_{X_i}(\Theta_i)$: η συνάρτηση κόστους υπολογισμένη σε όλες τις εγγραφές της δέσμης εκπαίδευσης του βήματος i

Ο αλγόριθμος ADAM [19] που χρησιμοποιούμε στην εργασία συνδυάζει την προσέγγιση των μεθόδων RMSprop και momentum.[3] Ο ADAM αναπροσαρμόζει τις τιμές του διανύσματος Θ βηματικά κατά την ακόλουθη λογική:

$$\mathbf{m} = \beta_1 \cdot \mathbf{m} + (1 - \beta_1)\nabla J_{X_i}(\Theta_i)$$
$$\mathbf{v} = \beta_2 \cdot \mathbf{v} + (1 - \beta_2)(\nabla J_{X_i}(\Theta_i))^2$$
$$\Theta_{i+1} = \Theta_i - \eta \frac{\mathbf{m}}{\sqrt{\mathbf{v} + \epsilon}}$$

[3] Οι μέθοδοι αυτοί αναλύονται στο παράρτημα B′.6.

Οι πράξεις $(\nabla J_{X_i}(\Theta_i))^2$ και $\frac{\mathbf{m}}{\sqrt{\mathbf{v}+\epsilon}}$ περιγράφουν πολλαπλασιασμό (τετράγωνο) και διαίρεση μεταξύ διανυσμάτων στοιχείο προς στοιχείο.[4]

Για να αντιμετωπιστεί το φαινόμενο οι πίνακες $\mathbf{m}$ και $\mathbf{v}$, που αρχικοποιούνται με μηδενικές τιμές, να «αργούν» να αυξήσουν τις τιμές τους, ο αλγόριθμος ADAM εισάγει δύο βοηθητικά βήματα:

$$\mathbf{m} = \beta_1 \cdot \mathbf{m} + (1-\beta_1)\nabla J_{X_i}(\Theta_i)$$
$$\mathbf{m_1} = \frac{\mathbf{m}}{1-\beta_1^i}$$
$$\mathbf{v} = \beta_2 \cdot \mathbf{v} + (1-\beta_2)(\nabla J_{X_i}(\Theta_i))^2$$
$$\mathbf{v_1} = \frac{\mathbf{v}}{1-\beta_2^i}$$
$$\Theta_{i+1} = \Theta_i - \eta\frac{\mathbf{m_1}}{\sqrt{\mathbf{v_1}+\epsilon}}$$

Τυπικές τιμές των παραμέτρων είναι οι $\eta = 0.001, \beta_1 = 0.9$, $\beta_2 = 0.999$ και $\epsilon = 10^{-8}$.

3.4 Χρήση αλγορίθμου κατά την εκτέλεση

Αφού ολοκληρωθεί η εκπαίδευση, εφαρμόζουμε κάποιες τεχνικές αλλαγές στο δίκτυο ώστε να είναι έτοιμο να δεχθεί ως είσοδο ολόκληρες εικόνες. Κατά την εκτέλεση του αλγορίθμου επιθυμούμε το εκπαιδευμένο δίκτυο να υπολογίσει τον τοπικό περιγραφέα $\mathbf{I}_{descriptor}(\mathbf{p})$ για κάθε θέση $\mathbf{p}$ της αριστερής και δεξιάς λήψης.

Θα ήταν εξαιρετικά ασύμφορο να αντιμετωπίζαμε την κάθε εικόνα του στερεοσκοπικού ζεύγους ως ένα σύνολο $height \times width$ επικαλυπτόμενων χωρίων τα οποία θα προωθούσαμε ξεχωριστά. Αυτή την προσέγγιση ακολουθήσαμε κατά την εκπαίδευση με σκοπό να αυξήσουμε το σύνολο των παραδειγμάτων (από κάθε στερεοσκοπικό ζεύγος δημιουργούσαμε περίπου 10^5 παραδείγματα). Κατά την εκτέλεση, κρατάμε την ίδια δομή στο δίκτυο, εφαρμόζοντας τις εξής τροποποιήσεις:

- Η πράξη **conv2d** εφαρμόζεται με κατάλληλο zero padding στο σήμα εισόδου, ώστε οι χωρικές διαστάσεις σε είσοδο και έξοδο να μένουν αναλλοίωτες.
- Η κανονικοποίηση δέσμης (BN) κανονικοποιεί το σήμα εισόδου με βάση μια αντιπροσωπευτική μέση τιμή μ και τυπική απόκλιση σ που έχει αποθηκεύσει από την διαδικασία της εκπαίδευσης. Οι τιμές αυτές έχουν υπολογιστεί με την τεχνική του κινούμενο μέσου όρου.

Με αυτές τις τροποποιήσεις καταφέρνουμε να υπολογίσουμε σε ένα βήμα προώθησης (forward pass) τους τοπικούς περιγραφείς όλων των σημείων της εικόνας, αποφεύγοντας εξαντλητικούς επανυπολογισμούς επικαλυπτόμενων χωρίων. Τα μόνα σημεία της εικόνας που υφίστανται αλλοιώσεις είναι τα σημεία της περιφέρειας στα οποία οι συνελίξεις επηρεάζονται από το zero padding. Τα σημεία αυτά θα είχαν ούτως ή άλλως άστοχα αποτελέσματα αφού δεν ανήκουν στο training set και έχουν διαφορετική κατανομή τιμών.

[4] Η έξοδός τους είναι ίδιας διάστασης με την είσοδο.

Θα μπορούσαμε να τα αντιμετωπίσουμε με ειδική μέθοδο αλλά δεν μας απασχολούν ιδιαίτερα καθώς αποτελούν ελάχιστο ποσοστό του συνόλου της εικόνας. Χαρακτηριστικά σε μια εικόνα $720p$ και σε ένα δίκτυο εκπαιδευμένο σε `patch_size` $=$ `19` τα σημεία της περιφέρειας αποτελούν το 0.08% του συνόλου της εικόνας.

Κεφάλαιο 4

Υλοποίηση - Πειραματικό Μέρος

4.1 Υπολογιστικό Σύστημα

Τα πειράματα υλοποιήθηκαν σε σταθερό υπολογιστή με τα εξής χαρακτηριστικά:

- **Λειτουργικό σύστημα:** Linux Ubuntu 16.04
- **Επεξεργαστής:** 4x Intel® Core™ i5-4690 CPU @ 3.50 GHz (6M Cache)
- **Μνήμη RAM:** DDR3 16GB
- **Κάρτα γραφικών:** NVIDIA GeForce GTX 1060 (memory: 6GB, clock: 1.5 GHz, bandwidth: $192\frac{GB}{s}$, CUDA cores:1280)

Όλα τα πειράματα υλοποιήθηκαν σε περιβάλλον Python. Σε συγκεκριμένες περιπτώσεις για την επιτάχυνση των αλγορίθμων αξιοποιήθηκε ο στατικός μεταγλωτιστής Cython. Χρησιμοποιήθηκαν επίσης οι βιβλιοθήκες:

- tensorflow
- numpy
- matplotlib
- opencv

4.2 Εκπαίδευση νευρωνικού δικτύου

Εκπαιδεύουμε το νευρωνικό δίκτυο σε εγγραφές που εξάγουμε αποκλειστικά από το σετ δεδομένων KITTI 2012 που περιέχει συνολικά 194 εικόνες αφού το διαχωρίσουμε σε σετ εκπαίδευσης και σετ επικύρωσης, με συντελεστή αναλογίας $\texttt{pcg\_tr} = 0.6$. Επομένως, χρησιμοποιούνται $\textbf{int}(\texttt{pcg\_tr} \times 194) = 116$ εικόνες για την εκπαίδευση και $\textbf{int}((1 - \texttt{pcg\_tr}) \times 194) = 78$ για την επικύρωση. Κάθε δέσμη (batch) εκπαίδευσης περιέχει $\texttt{batch\_size} = 128$ εγγραφές που αναλογούν σε τυχαίες θέσεις εντός της εικόνας επιδιώκοντας την μικρότερη δυνατή συσχέτιση των παραδειγμάτων εντός της ίδιας δέσμης. Ορίζουμε ανώτατο όριο batches ανά εικόνα $\texttt{batches\_limit} = 40$, για να εξασφαλίσουμε μικρή συσχέτιση μεταξύ τους. Με αυτήν την μεθοδολογία:

- από κάθε εικόνα εξάγονται $\texttt{batches\_limit} \times \texttt{batch\_size} = 5120$ παραδείγματα.

- κάθε εποχή εκπαίδευσης (training epoch) περιέχει

$$\textbf{int}(195 \times \texttt{pcg\_tr}) \times \texttt{batches\_limit} = 4640 \text{ δέσμες}$$

που αναλογούν σε

$$\textbf{int}(195 \times \texttt{pcg\_tr}) \times \texttt{batches\_limit} \times \texttt{batch\_size} = 593920 \text{ παραδείγματα.}$$

Κατά την εκπαίδευση, σε κάθε μπλοκ conv2d + BN + ReLU προσθέτουμε ένα επίπεδο dropout [40] με ποσοστό αποκοπής νευρώνων 40%.

Χρησιμοποιούμε τον βελτιστοποιητή (optimizer) ADAM [19] με αρχικό βαθμό εκμάθησης (learning rate) 0.01. Μειώνουμε τον βαθμό εκμάθησης κατά τον παράγοντα 5 κάθε 6000 επαναλήψεις εκπαίδευσης, από την επανάληψη 20000 και μετά.

Παρακολουθούμε τη μεταβολή του κόστους εντροπίας 4.1 σε κάθε βήμα εκπαίδευσης και ελέγχουμε το απόλυτο σφάλμα πρόβλεψης και το απόλυτο σφάλμα πρόβλεψης με ανώφλι κάθε 1000 βήματα εκπαίδευσης[1] 4.2. Παρατηρούμε ότι οι καμπύλες σφάλματος και απόκλισης μεταξύ του σετ εκπαίδευσης και του σετ επικύρωσης εμφανίζουν πολύ μεγάλη ομοιότητα, σχεδόν ταύτιση, επιβεβαιώνοντας ότι το δίκτυο δεν έχει υποστεί υπερπροσαρμογή (overfitting). Από τις καμπύλες παρατηρούμε ότι η εκπαίδευση τελματώνεται μετά από περίπου 20000 επαναλήψεις. Επιλέγουμε να κρατήσουμε τις τιμές όλων των εκπαιδεύσιμων παραμέτρων κατά το βήμα 25000.[2] Η εκπαίδευση διήρκεσε συνολικά περίπου 7 ώρες.

4.3 Αρχικοποίηση κόστους αντιστοίχησης

Σε κάθε μια από τις παρακάτω ενότητες, υπολογίζουμε τον πίνακα κόστους C με τους υπό δοκιμή αλγορίθμους και ακολούθως υπολογίζουμε τον χάρτη παράλλαξης, εφαρμόζοντας την πράξη $D(\mathbf{p}) = \textbf{arg\_min}_d(\mathbf{p})$. Επιλέγουμε τυχαία το στερεοσκοπικό ζεύγος 4 της συλλογής KITTI 2012 για την οπτικοποίηση των αποτελεσμάτων.

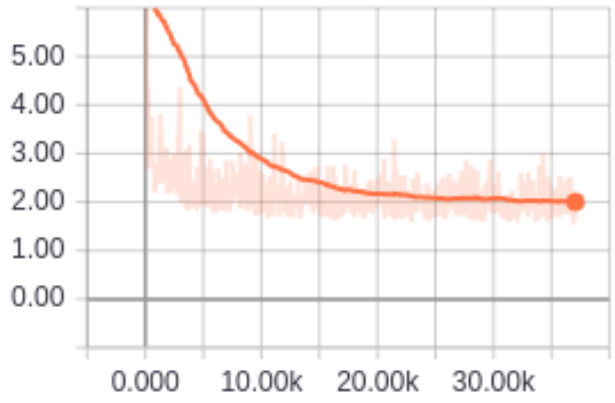
4.3.1 Συμβατικές μέθοδοι

Άθροισμα απόλυτων διαφορών

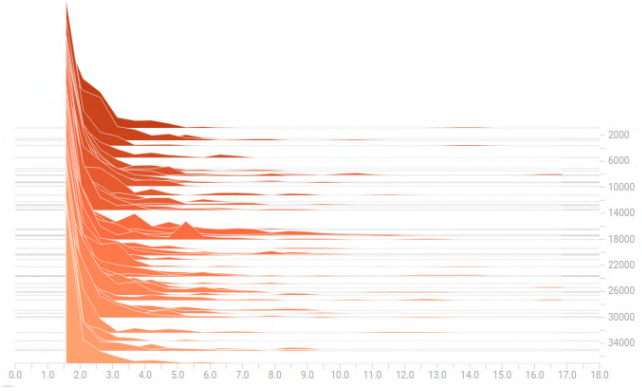
Όπως είδαμε στο κεφάλαιο 2 είναι δύσκολη η επιλογή περιοχής άθροισης κατάλληλου μεγέθους. Όσο μεγαλώνει, τόσο ο χάρτης εμφανίζεται πιο λείος αλλά με θολωμένες τις ακμές, ενώ αντίθετα όσο μικραίνει τόσο αυξάνεται η λεπτομέρεια στις ακμές με κόστος πολλές εξωκείμενες τιμές. Στο σχήμα 4.4 απεικονίζεται το παραπάνω φαινόμενο. Η καλύτερη επίδοση στο παράδειγμά εμφανίζεται για μέγεθος παραθύρου $\texttt{window\_size} = 21\ px$ με ποσοστό απόλυτου σφάλματος 16.7%.

[1] Για λόγους ταχύτητας, τα απόλυτα σφάλματα υπολογίζονται στις μισές εικόνες εκπαίδευσης κι επικύρωσης της συλλογής.

[2] Κατά την εκπαίδευση αποθηκεύονται όλες οι εκπαιδεύσιμες παράμετροι του δικτύου κάθε 1000 επαναλήψεις.

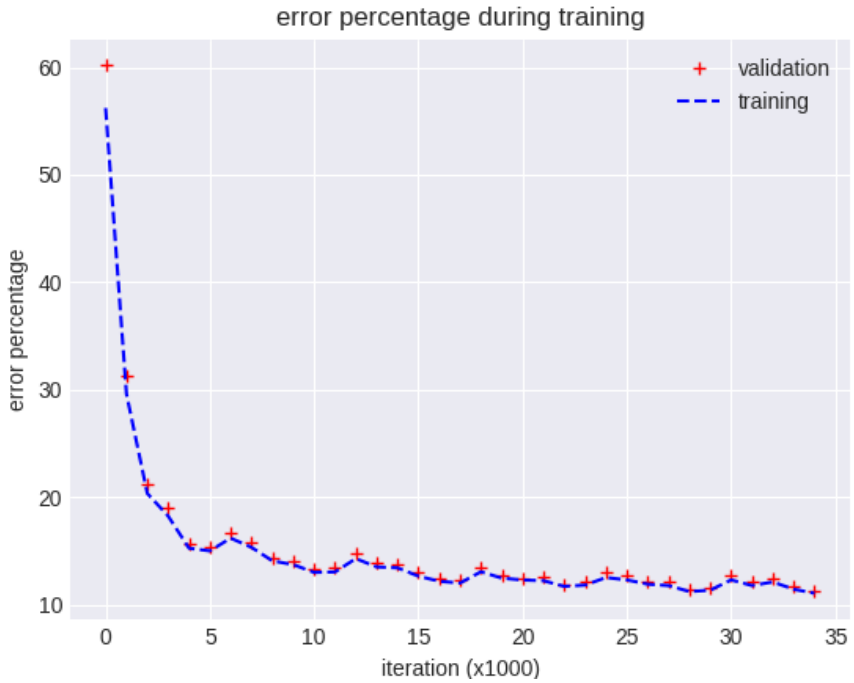


(α′) Μέσο κόστος εντροπίας ανά δέσμη κατά την διάρκεια της εκπαίδευσης.



(β′) Ιστόγραμμα κόστους εντροπίας για κάθε παράδειγμα της δέσμης εκπαίδευσης κατά την διάρκεια όλης της εκπαίδευσης.

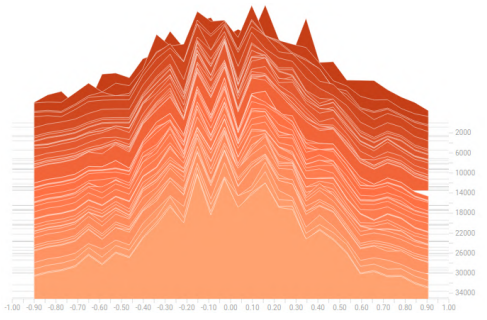
Σχημα 4.1: Η μείωση του κόστους εντροπίας είναι ικανοποιητική μέχρι το βήμα 20000. Από εκεί και πέρα η βελτίωση του δικτύου είναι ισχνή.



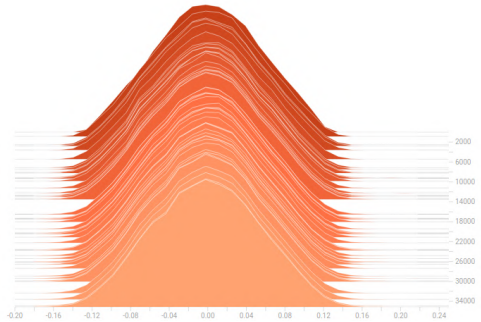
(Α΄) Εξέλιξη μέσου απόλυτου σφάλματος με κατώφλι κατά την διάρκεια της εκπαίδευσης.

(Β΄) Εξέλιξη μέση απόλυτου σφάλματος κατά την διάρκεια της εκπαίδευσης.

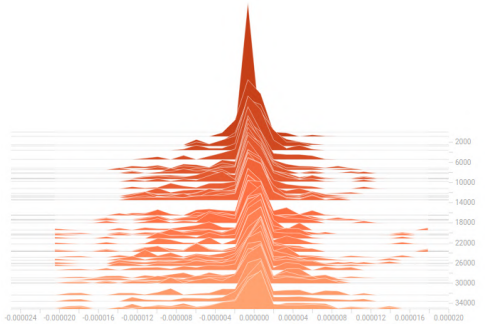
Σχημα 4.2: Παρατηρούμε και στα δύο γραφήματα οι καμπύλες που αναλογούν στο σετ εκπαίδευσης και επικύρωσης σχεδόν ταυτίζονται, φαινόμενο που αποτελεί έντονη ένδειξη ότι το δίκτυο δεν έχει υποστεί υπερπροσαρμογή. Τα γραφήματα επιβεβαιώνουν ότι η βελτίωση μετά το βήμα εκπαίδευσης 20000 είναι ανεπαίσθητη.



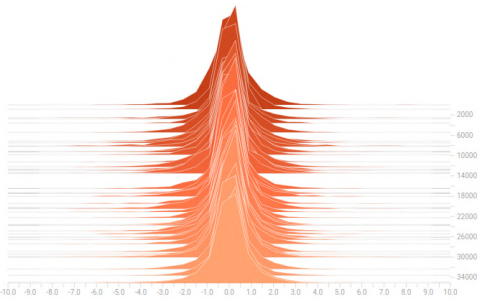
(Α′) Ιστόγραμμα βαρών του πρώτου επιπέδου.



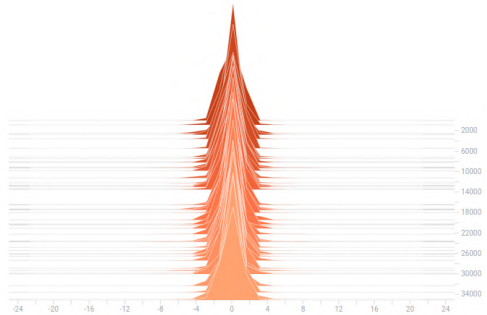
(Β′) Ιστόγραμμα βαρών του τέταρτου επιπέδου.



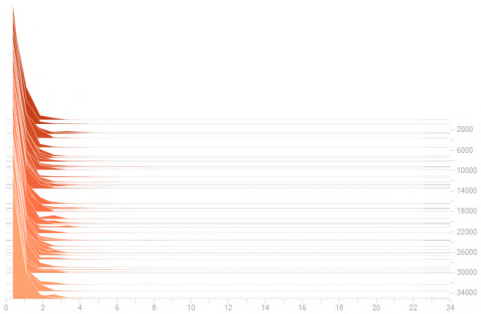
(Γ′) Ιστόγραμμα πολώσεων πρώτου επιπέδου.



(Δ′) Ιστόγραμμα εξόδου συνέλιξης πρώτου επιπέδου.



(Ε′) Ιστόγραμμα εξόδου επιπέδου batch normalization.

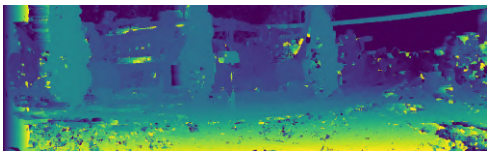


(Ϛ′) Ιστόγραμμα εξόδου συνάρτησης ενεργοποίησης ReLU.

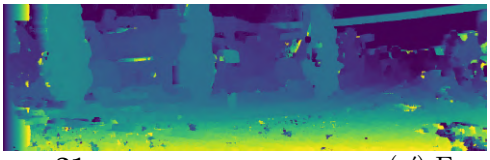
Σχημα 4.3: Παρατηρήσεις: (i) Η κατανομή των βαρών δεν είναι όμοια σε όλα τα επίπεδα. Όπως φαίνεται στην εικόνα (Α′) τα βάρη του πρώτου επιπέδου δεν ακολουθούν απόλυτα κανονική κατανομή και κυμαίνονται στο διάστημα $[-0.8, 0.8]$, ενώ αυτά του τέταρτου επιπέδου ακολουθούν απόλυτα κανονική και κυμαίνονται στο διάστημα $[-0.15, 0.15]$. (ii) Οι πολώσεις του πρώτου επιπέδου (σχήμα (Β′», όπως και των υπόλοιπων επιπέδων είναι τάξης μεγέθους 10^{-5}. Ουσιαστικά το δίκτυο δεν χρειάζεται τον αφινικό μετασχηματισμό $y = (conv2d(x, h)) + b$ και τον περιορίζει στον γραμμικό $y = (conv2d(x, h))$. Το φαινόμενο αυτό παρουσιάζεται συχνά όταν ακολούθως χρησιμοποιείται επίπεδο κανονικοποίησης δέσμης.(iii) Στα σχήματα (Δ′),(Ε′),(ΣΤ′) οι τιμές βρίσκονται εντός των διαστημάτων $[-7, 7]$, $[-7, 7]$ και $[0, 7]$ αντίστοιχα.

(Α′) I^L (Β′) I^R (Γ′) **ground_trouth**

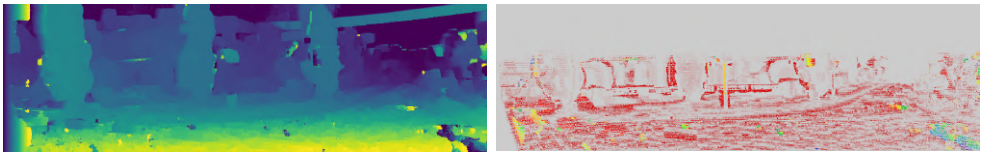
(Δ′) **window = 9px**



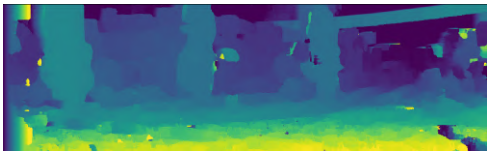
(Ε′) **window = 15px**



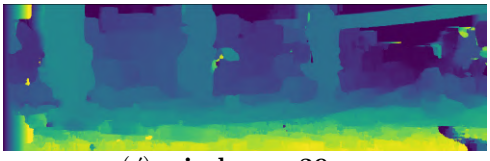
(Ϛ′) **window = 21px** (Ζ′) Error: 16.7%



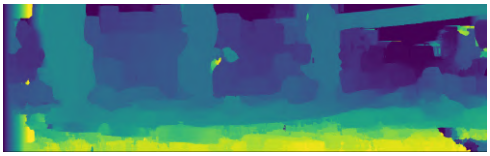
(Η′) **window = 27px**



(Θ′) **window = 33px**



(Ι′) **window = 39px**



Σχημα 4.4: Άθροισμα απόλυτων διαφορών σε παράθυρα διαφορετικού μεγέθους.

Μετασχηματισμός Census

Ο μετασχηματισμός census [43] δημιουργεί τοπικό περιγραφέα που είναι ανεπηρέαστος από τις φωτομετρικές αποκλίσεις. Ταυτόχρονα όμως έχει το μειονέκτημα να δημιουργεί παρόμοιο περιγραφέα από τελείως διαφορετικά είδωλα που τυχαίνει να δημιουργούν γειτονιές με παρόμοια σχέση φωτεινότητας περιφέρειας και κεντρικού pixel. Στο σχήμα 4.5 αποτυπώνεται αυτό το φαινόμενο καθώς οι υπολογισμένοι χάρτες παράλλαξης εμφανίζουν έντονες ασυνέχειες. Βέλτιστη επίδοση εμφανίζει για μέγεθος παραθύρου `window_size` $= 27\,px$ με ποσοστό απόλυτου σφάλματος 32.27%.

Μέθοδος AD-Census

Ο συνδυασμός του «αθροίσματος απόλυτων διαφορών» και μετασχηματισμού Census δημιουργεί την μέθοδο AD-Census. [28]. Στο σχήμα 4.6 φαίνεται ο χάρτης παράλλαξης για διαφορετικές τιμές παραθύρου. Οι υπολογισμοί έγιναν με τιμές παραμέτρων `l_AD` $=$ 10, `l_Census` $= 30$. Βέλτιστη επίδοση πετυχαίνεται για επιλογή παραθύρου μεγέθους `window_size` $= 21\,px$ με ποσοστό απόλυτου σφάλματος 19.84%.

Σύγκριση μεθόδων

Τα ποσοστά σφάλματος και τα απόλυτα σφάλματα των τριών παραπάνω μεθόδων για διαφορές τιμές παραθύρων άθροισης φαίνονται στα γραφήματα του σχήματος 4.7. Οι μετρήσεις δεν είναι αρκετά ευσταθείς καθώς έγιναν σε ένα μόνο στερεοσκοπικό ζεύγος εικόνων. Η μέθοδος AD-Census αν και εμφανίζει το δεύτερο καλύτερο αποτέλεσμα μετά την μέθοδο «αθροίσματος απόλυτων διαφορών» είναι η πιο αξιόπιστη σε συνδυασμό με τα βήματα της στερεοσκοπικής μεθόδου, γι΄ αυτό την επιλέγουμε για την σύγκριση με τα αποτελέσματα του νευρωνικού δικτύου.

4.3.2 Νευρωνικό δίκτυο ταξινόμησης πολλαπλών κατηγοριών

Για να μειώσουμε την υπολογιστική πολυπλοκότητα επιλέγουμε να προωθήσουμε ολόκληρες τις εικόνες I^L, I^R του στερεοσκοπικού ζεύγους στους δύο κλάδους του σιαμαίου δικτύου, αντί να προωθούμε κάθε χωρίο τους ξεχωριστά. Στην έξοδό του παίρνουμε τους αντίστοιχους συνολικούς περιγραφείς της εικόνας:

$$I^L_{desc} = f_{siamese}(I^L), \quad I^R_{desc} = f_{siamese}(I^R), \text{ όπου: } I^L_{desc}, I^R_{desc} \in \mathbb{R}^{m \times n \times F}$$

Το κόστος ομοιότητας υπολογίζεται ως το αντίθετο του εσωτερικού γινομένου των τοπικών περιγραφέων των υπό σύγκριση χωρίων. Για λόγους ταχύτητας, υπολογίζουμε κάθε επίπεδο d του πίνακα κόστους C σε μια πράξη εσωτερικού γινομένου:

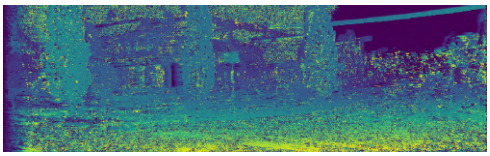
```
for d in range(max_disparity + 1):
        C_d = = −dot_product_F(I^L_desc[:, d :, :], I^R_desc[:, : −d, :])
        C[d, :, :] = C_d
```

Στον πίνακα 4.1 φαίνονται συμπυκνωμένα οι παραπάνω πράξεις με τους αντίστοιχους χρόνους που χρειάζονται για την εκτέλεσή τους. Στο γράφημα 4.8 φαίνονται οι χρόνοι εκτέλεσης για διαφορετικά μεγέθη εικόνων και μέγιστες παραλλάξεις.

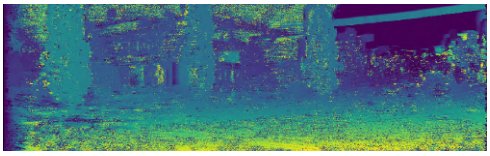
Το δίκτυο είναι εκπαιδευμένο σε `patch_size` $= 19$.

(Α′) I^L (Β′) I^R (Γ′) **ground_trouth**

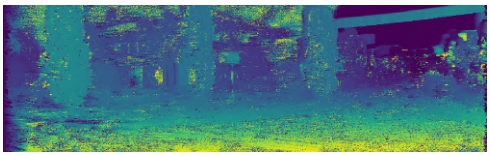
(Δ′) **window = 9px**



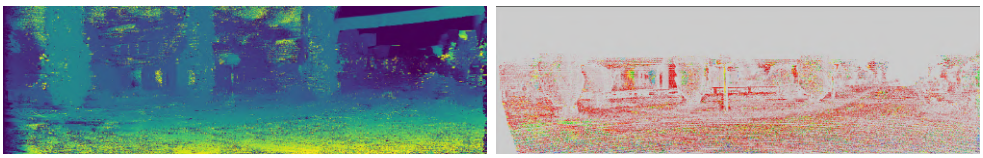
(Ε′) **window = 15px**



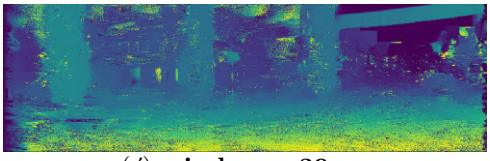
(Ϛ′) **window = 21px**



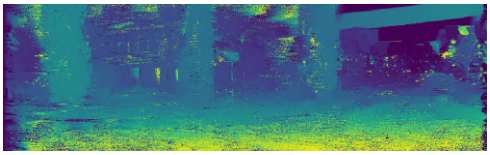
(Ζ′) **window = 27px** (Η′) Error: 32.27%



(Θ′) **window = 33px**



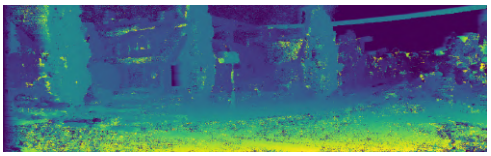
(Ι′) **window = 39px**



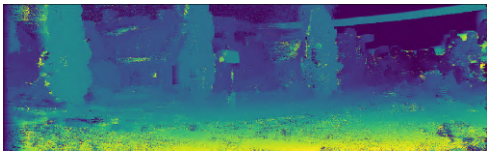
Σχημα 4.5: Μετασχηματισμός Census σε παράθυρα διαφορετικού μεγέθους.

(Α′) I^L (Β′) I^R (Γ′) **ground_trouth**

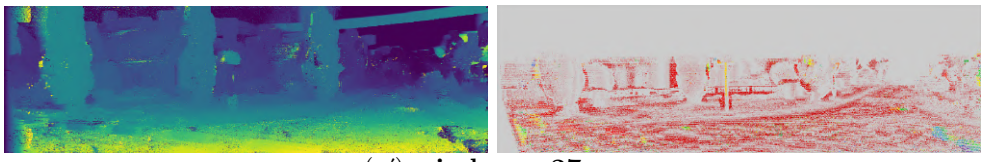
(Δ′) **window = 9px**



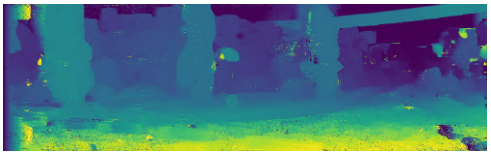
(Ε′) **window = 15px**



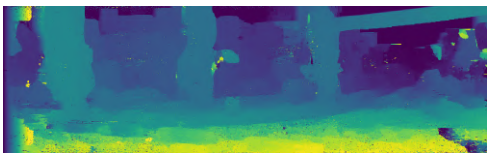
(Ϛ′) **window = 21px** (Ζ′) Error: 19.84%



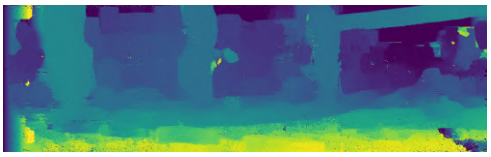
(Η′) **window = 27px**



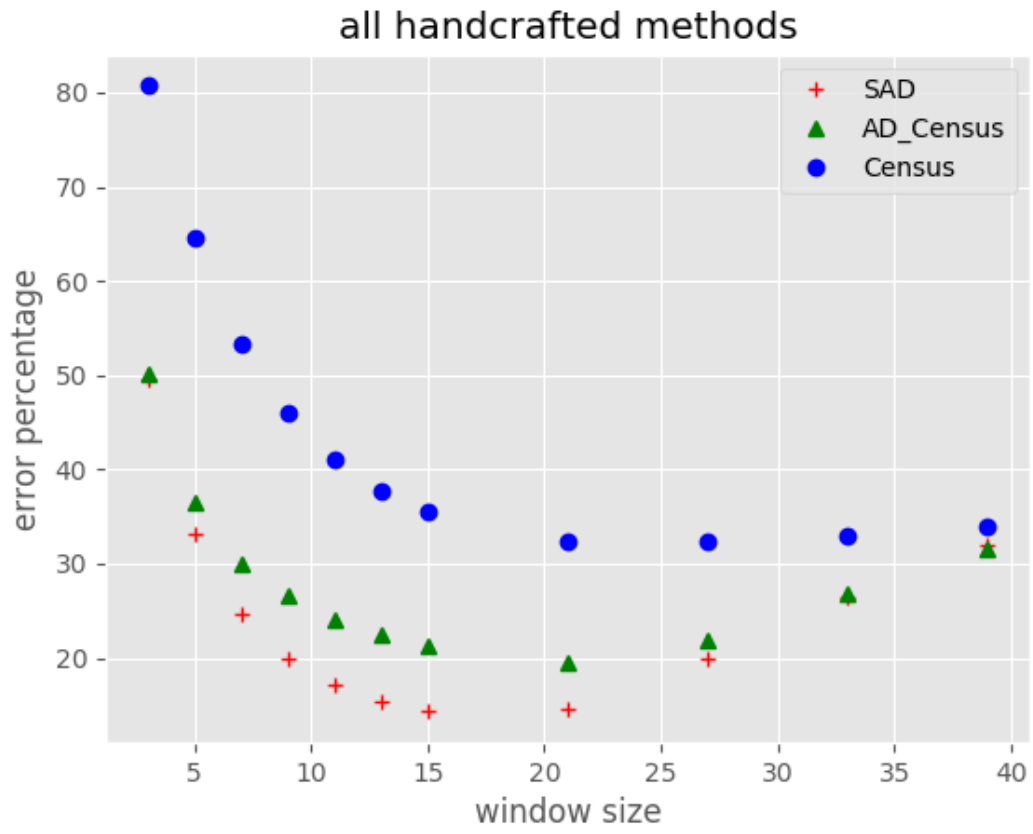
(Θ′) **window = 33px**



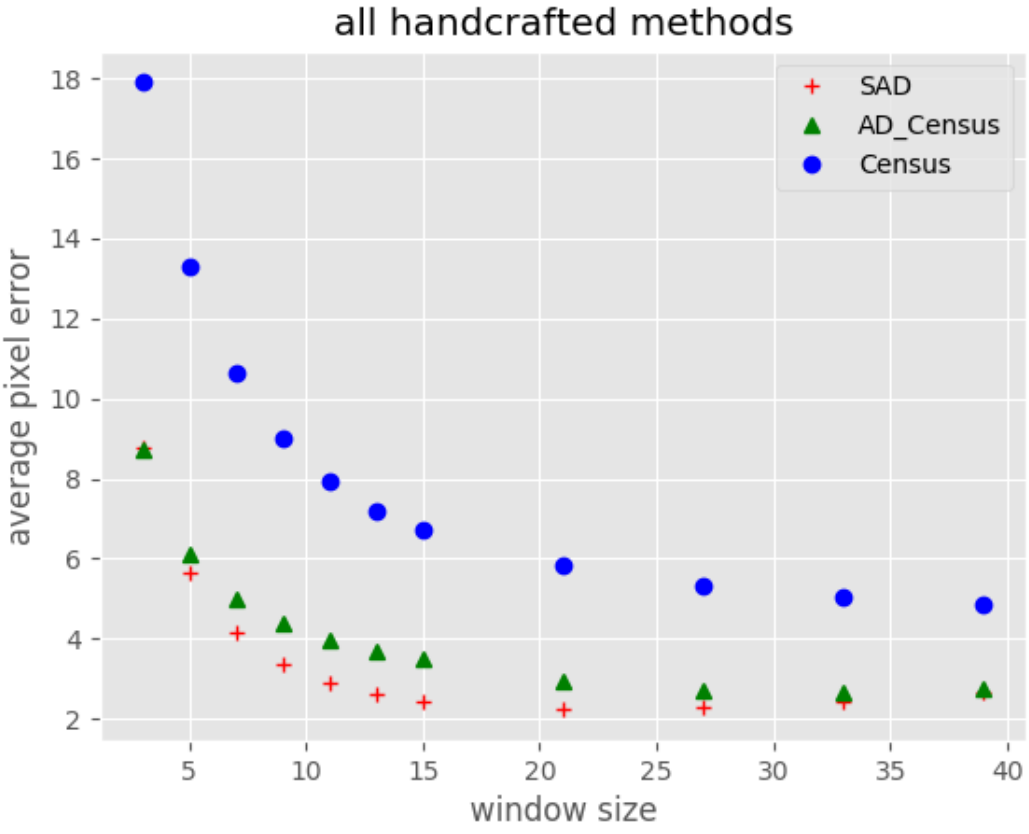
(Ι′) **window = 39px**



Σχημα 4.6: Μετασχηματισμός AD-Census σε παράθυρα διαφορετικού μεγέθους.



(α′) Μέσο απόλυτο σφάλμα με κατώφλι, για διαφορετικά μεγέθη παραθύρου.



(β′) Μέσο απόλυτο σφάλμα, για διαφορετικά μεγέθη παραθύρου.

Σχημα 4.7: Σύγκριση αποτελεσμάτων συμβατικών μεθόδων

	Περιγραφή Επιπέδων	Διαστάσεις	Χρόνος
Local descriptors extraction (Εξαγωγή περιγραφέων)			
Siamese network (Σιαμαίο δίκτυο)			
	Είσοδος I^L, I^R	H×W×1	
1	3×3 conv2d+BN+ReLU, F=64	H×W×F	0.06 sec
2	3×3 conv2d+BN+ReLU, F=64	H×W×F	0.06 sec
⋮	⋮	⋮	⋮
8	3×3 conv2d+BN+ReLU, F=64	H×W×F	0.06 sec
9	3×3 conv2d+BN (not ReLU), F=64	H×W×F	0.06 sec
	Έξοδος I^L_{desc}, I^R_{desc}	H×W×F	
Υπολογισμός κόστους ομοιότητας $\mathbf{C_d}$			
10	$\mathbf{C_d} = -\mathbf{dot\_product_F}(I^L_{desc}[:, d:, :], I^R_{desc}[:, :-d, :])$	1×H×W	0.95 sec

TABLE 4.1: Περίληψη επιπέδων δικτύου ταξινόμησης πολλαπλών επιπέδων. Όλα τα βήματα υπολογίζονται μια φορά, εκτός του βήματος 10 που υπολογίζεται ξεχωριστά για κάθε διαφορετικό επίπεδο d. Οι χρόνοι έχουν υπολογιστεί για μία μέση φωτογραφία με $H = 400, W = 1200,$ `max_disparity` $= 150$.

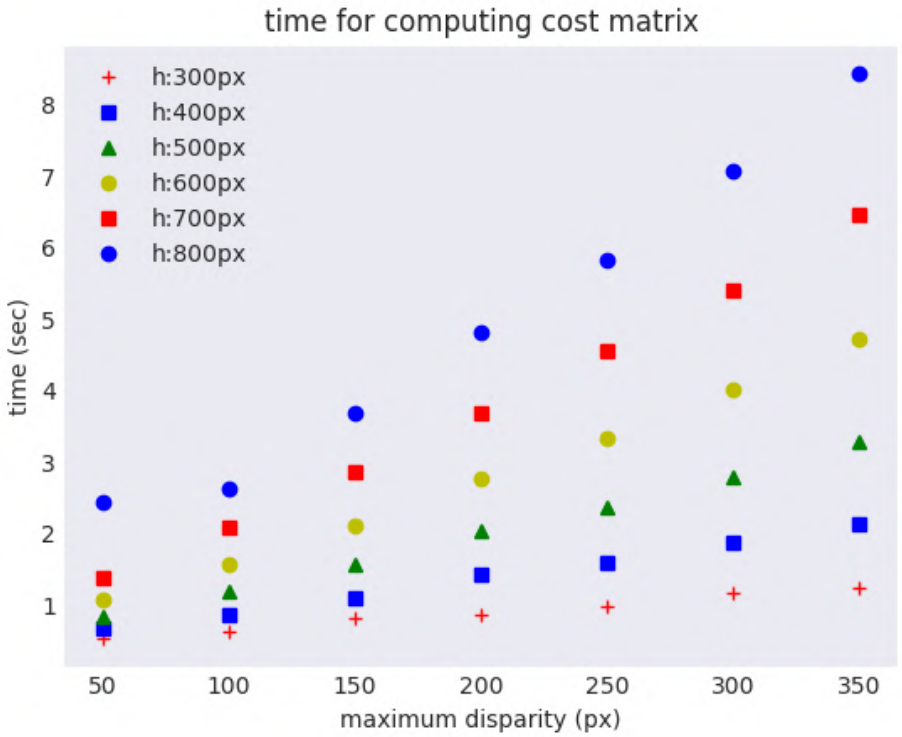
Στις εικόνες 4.9 φαίνεται ο χάρτης παράλλαξης που προέκυψε από τον πίνακα κόστους που αρχικοποίησε το νευρωνικό δίκτυο. Το νευρωνικό δίκτυο έχει εκπαιδευτεί να συγκρίνει γειτονιές μεγέθους `patch_size` $= 19\,px$. Το ποσοστό σφάλματος είναι 5.846% και το μέσο απόλυτο σφάλμα $1.132\,px$. Η βελτίωση σε σχέση με τις προηγούμενες μεθόδους είναι έντονη και γενικεύεται στο σύνολο των σετ δεδομένων.

4.4 Στερεοσκοπική Μέθοδος

Η μελέτη των βημάτων γίνεται παρουσιάζοντας τα αποτελέσματα πριν και μετά από το κάθε βήμα της στερεοσκοπικής μεθόδου. Θα δοκιμάζονται οι πίνακες κόστους και οι χάρτες παράλλαξης που έχουν προκύψει από την μέθοδο AD-Census και το νευρωνικό δίκτυο.

4.4.1 Άθροιση κόστους σε προσαρμόσιμη περιοχή υποστήριξης

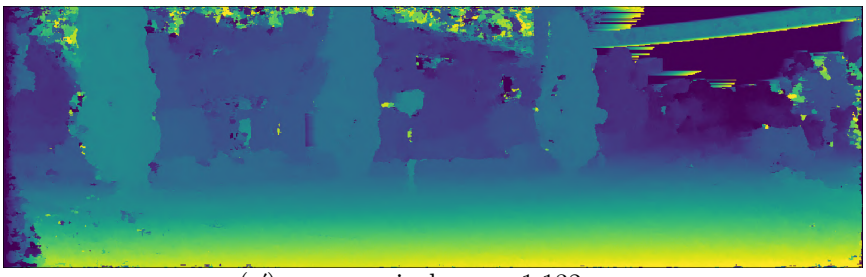
Οι υπολογισμοί υλοποιούνται με τιμές παραμέτρων `intensity_threshold` $= 0.13$ και `distance_threshold` $= 5$ που κάνουν τον αλγόριθμο αρκετά ευαίσθητο στην ανεύρεση ακμών και το εύρος των περιοχών υποστήριξης περιορισμένο. Έτσι έχουμε αρκετά εχέγγυα ότι δεν θα περάσουμε σχεδόν ποτέ από μια επιφάνεια σε μια άλλη, με κόστος οι περιοχές υποστήριξης να περιορίζονται σε πολύ μικρότερα εμβαδά απ΄ τη συνολική επιφάνεια του αντικειμένου πάνω στο οποίο υπολογίζονται. Αυτή η επιλογή «μικρού ρίσκου» δίνει τη δυνατότητα η επανάληψη του αλγορίθμου πολλές φορές να εξομαλύνει τον χάρτη παράλλαξης χωρίς να καταστρέφει (θολώνει) τις ακμές, όπως επιβεβαιώνεται και στα γραφήματα της εικόνας 4.10. Οφείλουμε να σχολιάσουμε ότι τα βήματα της άθροισης κόστους σε προσαρμόσιμη περιοχή υποστήριξης και της ημι-καθολικής αντιστοίχησης έχουν όμοιο στόχο: την εξομάλυνση του χάρτη παράλλαξης. Επομένως η εξαντλητική χρησιμοποίηση του ενός εργαλείου, για παράδειγμα την εφαρμογή του παρόντος βήματος επαναληπτικά `cbca_iter` > 10 φορές, μειώνει εξαιρετικά την επίδραση του έτερου. Παρατηρούμε ότι στην περίπτωση που τα κόστη έχουν αρχικοποιηθεί με νευρωνικό δίκτυο,



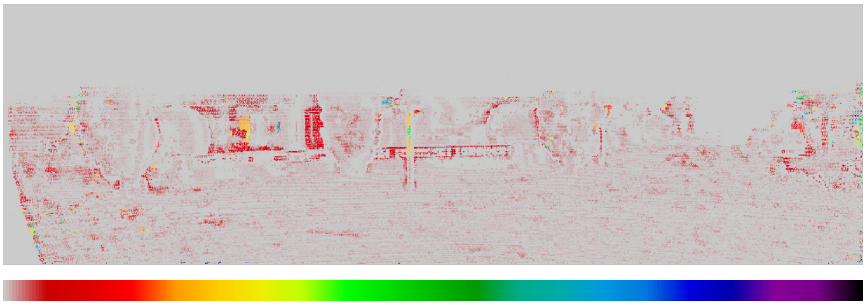
Σχημα 4.8: Χρόνος υπολογισμού του πίνακα κόστους από το νευρωνικό δίκτυο συναρτήσει του μεγέθους της εικόνας και της μέγιστης παράλλαξης. Το μέγεθος της εικόνας εισόδου είναι $h \times \frac{16}{9} h px$. Παρατηρούμε ότι οι χρόνοι δεν ξεπερνούν τα $9\,sec$ ακόμη και για εικόνα 800×1420 (μεγαλύτερη ανάλυση από το όριο του high-definition) και μέγιστη παράλλαξη τα $350px$. Για εικόνες μεγαλύτερου γινομένου $h \times \frac{16}{9} h \times$ `max_disparity` δεν επαρκούσε η μνήμη της κάρτας γραφικών.

(Α′) I^L (Β′) I^R (Γ′) **ground_trouth**

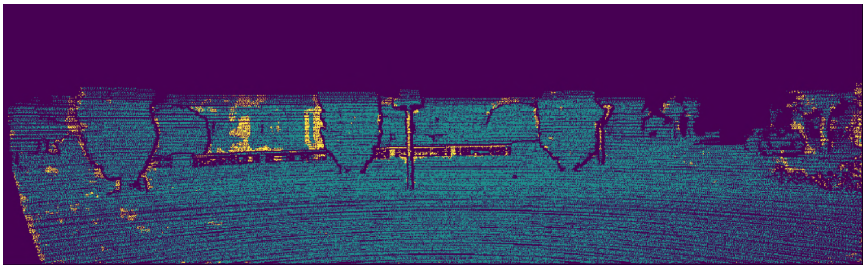
(Δ′) Χάρτης παράλλαξης νευρωνικού δικτύου



(Ε′) average pixel error: 1.132 px



(Ϛ′) Error: 5.846%



Σχημα 4.9: Χάρτης παράλλαξης υπολογισμένος με βάση τον πίνακα κόστους που αρχικοποιεί το νευρωνικό δίκτυο.

η επαναληπτική χρήση της «άθροισης κόστους» βελτιώνει το αποτέλεσμα διαρκώς μέχρι και τις 40 επαναλήψεις όπου το ποσοστό σφάλματος έχει πέσει στο 3.2%. Αντίθετα στην περίπτωση αρχικοποίησης με AD-Census η βελτίωση τελματώνει στις 25 επαναλήψεις με ποσοστό σφάλματος 7.8%. Πρέπει να επισημάνουμε ότι τα αποτελέσματα της «άθροισης κόστους» δεν είναι πάντα το ίδιο ευεργετικά όσο το συγκεκριμένο παράδειγμα. Γενικότερα, συνηθίζεται η βελτίωση του σφάλματος, σε συνδυασμό με το επόμενο βήμα της ημικαθολικής αντιστοίχησης, να σταματάει μετά τις 3 επαναλήψεις. Έτσι, συνήθως επιλέγουμε μια τιμή στο εύρος $[0, 5]$.

4.4.2 Ημι-καθολική αντιστοίχηση

Επιλέγουμε τιμές παραμέτρων:

- `P1_ref` $= 1.5$
- `P1_ref` $= 30$
- `thres` $= 0.13$
- `big_factor` $= 6$
- `small_factor` $= 3$

Επιλέγουμε το ίδιο όριο στη διαφορά φωτεινότητας `thres` με την μέθοδο «άθροισης κόστους». Η μέθοδος AD-Census μαζί με την ημικαθολική αντιστοίχηση εμφανίζουν σφάλμα 5.993%, κατά δύο ποσοστιαίες μονάδες μικρότερο σε σχέση με τον συνδυασμό AD-Census και cbca. Το νευρωνικό δίκτυο μαζί με την ημικαθολική αντιστοίχηση εμφανίζει σφάλμα 4.324%. Με την παρεμβολή και της μεθόδου cbca για δύο επαναλήψεις το σφάλμα μειώνεται στο 4.054%, ενώ η βέλτιστη επίδοσή του εμφανίζεται με παρεμβολή της μεθόδου cbca για 30 επαναλήψεις στο $2,97\%$.

4.4.3 Εντοπισμός εξωκείμενων τιμών στον χάρτη παράλλαξης

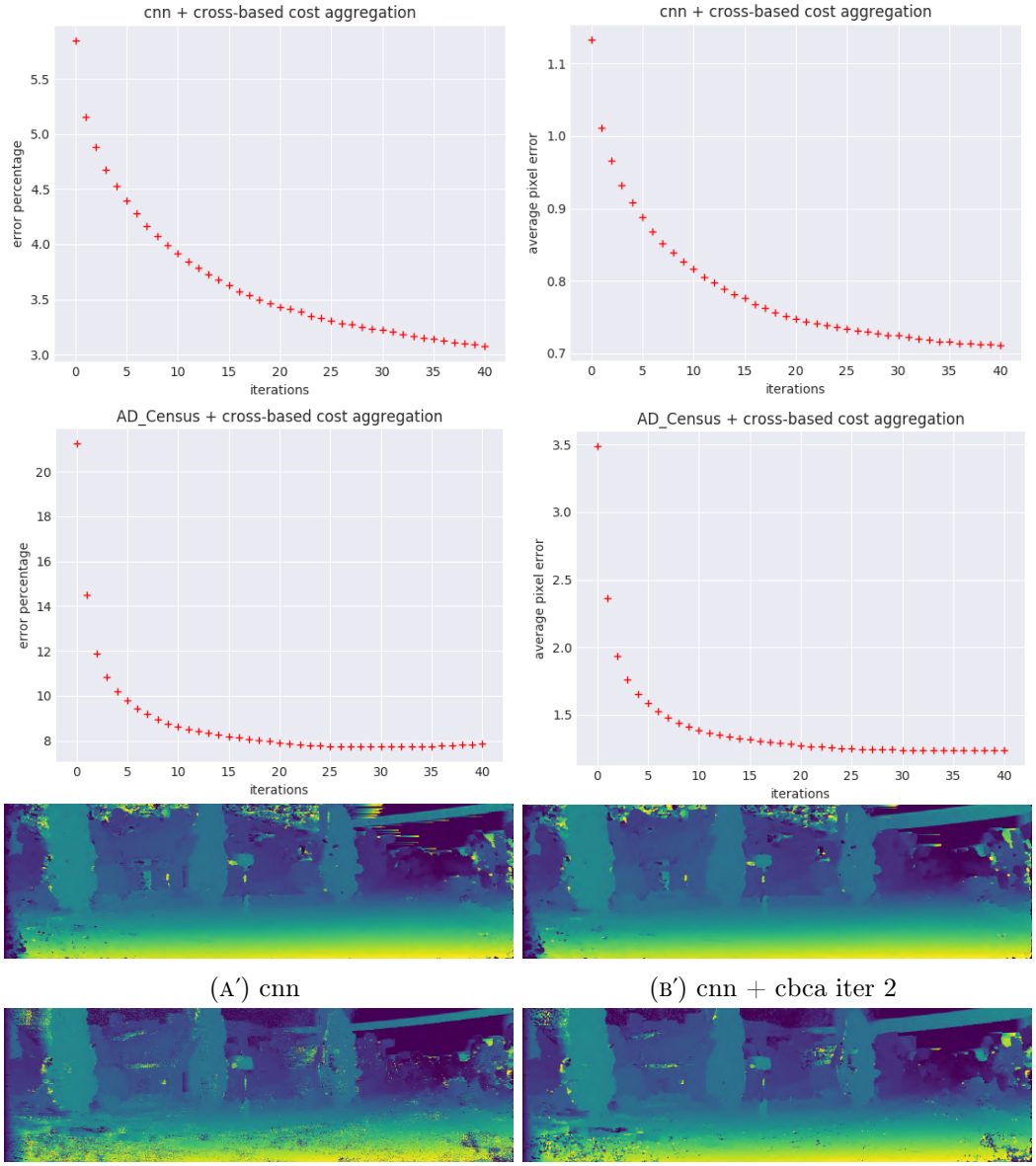
Εφαρμόζουμε τον αλγόριθμο εντοπισμού εξωκείμενων τιμών στις μέχρι τώρα βέλτιστες επιδόσεις. Στον συνδυασμό AD-Census + sgm η εφαρμογή του αλγορίθμου βελτιώνει το σφάλμα κατά 0.55% στο 5.44%, ενώ στον συνδυασμό AD-Census + cbca30 + sgm κατά 0.1% στο 2.87%.

4.4.4 Βελτιστοποίηση με ακρίβεια δεκαδικού pixel

Τέλος, η βελτιστοποίηση με ακρίβεια υποπίξελ βελτιώνει το συνολικό σφάλμα κατά 0.022% και 0.2%. Συνολικά, η μέθοδος βασισμένη στα αρχικά κόστη του αλγορίθμου AD-Census πετυχαίνει ποσοστό σφάλματος 5.422% ενώ η μέθοδος βασισμένη στο νευρωνικό δίκτυο 2.67%. Τα αντίστοιχα μέσα απόλυτα σφάλματα είναι $0.832\,px$ και $0.62\,px$ αντίστοιχα.

4.4.5 Χρόνος εκτέλεσης στερεοσκοπικής μεθόδου

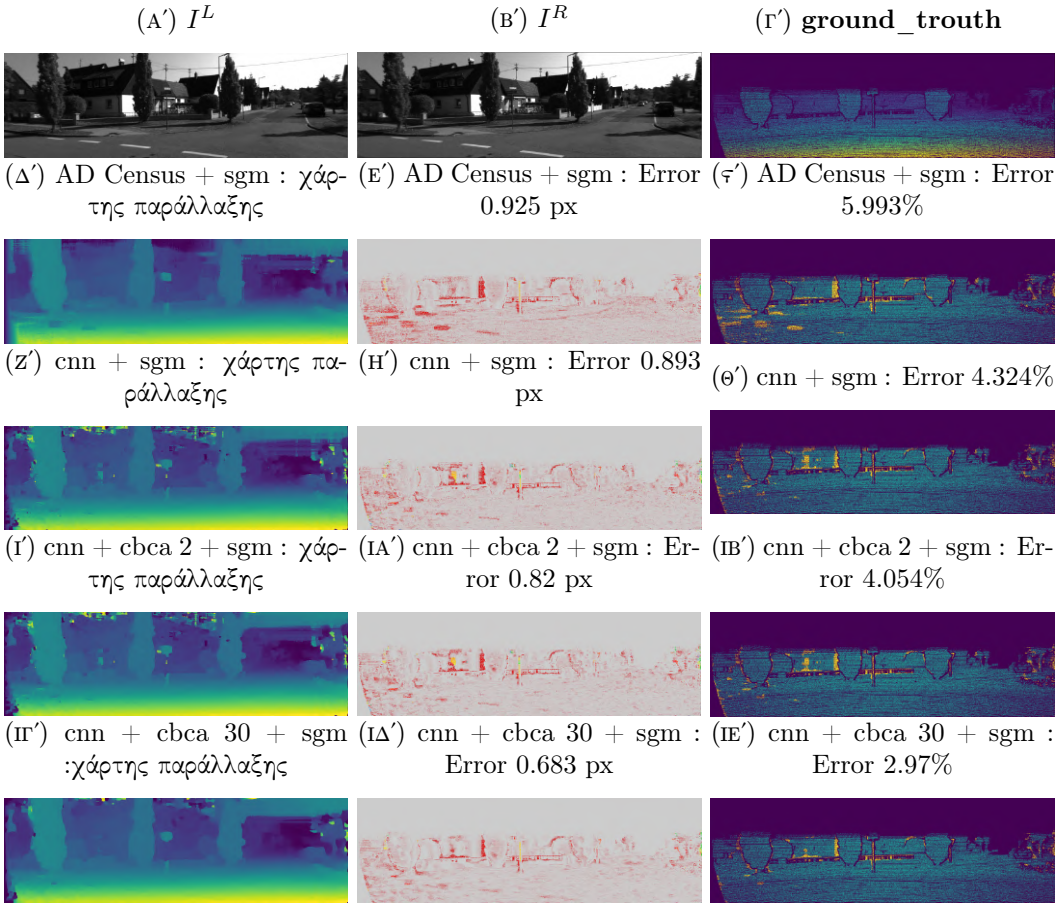
Οι μετρήσεις χρόνου εκτέλεσης όλων των βημάτων της στερεοσκοπικής μεθόδου απεικονίζονται στα γραφήματα της εικόνας 4.15. Όλοι οι αλγόριθμοι έχουν υλοποιηθεί με χρήση της Cython. Η υλοποίηση σε Python προκαλούσε τεράστιους χρόνους εκτέλεσης,



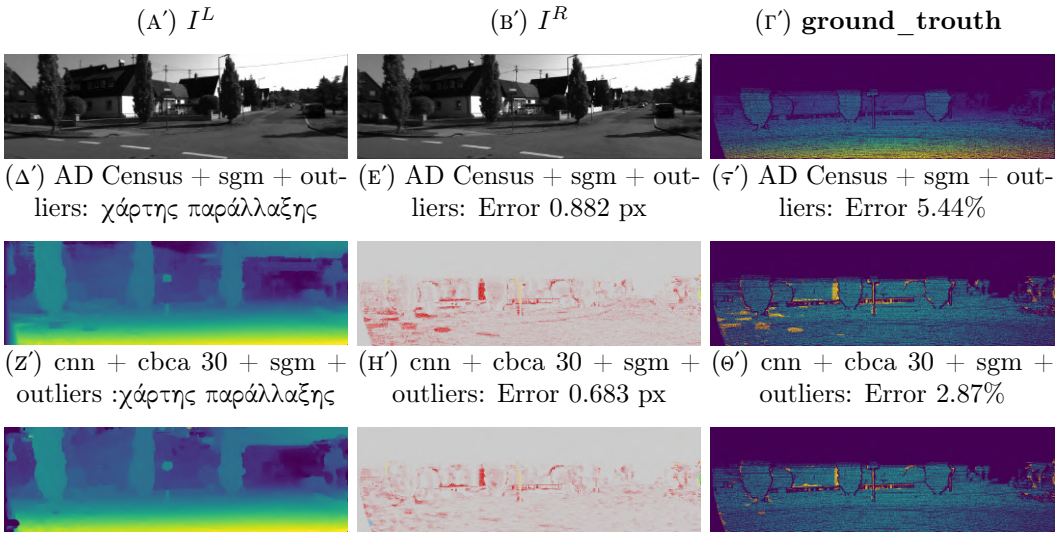
(Α′) cnn (Β′) cnn + cbca iter 2

(Γ′) AD-Census iter 0 (Δ′) AD-Census iter 2

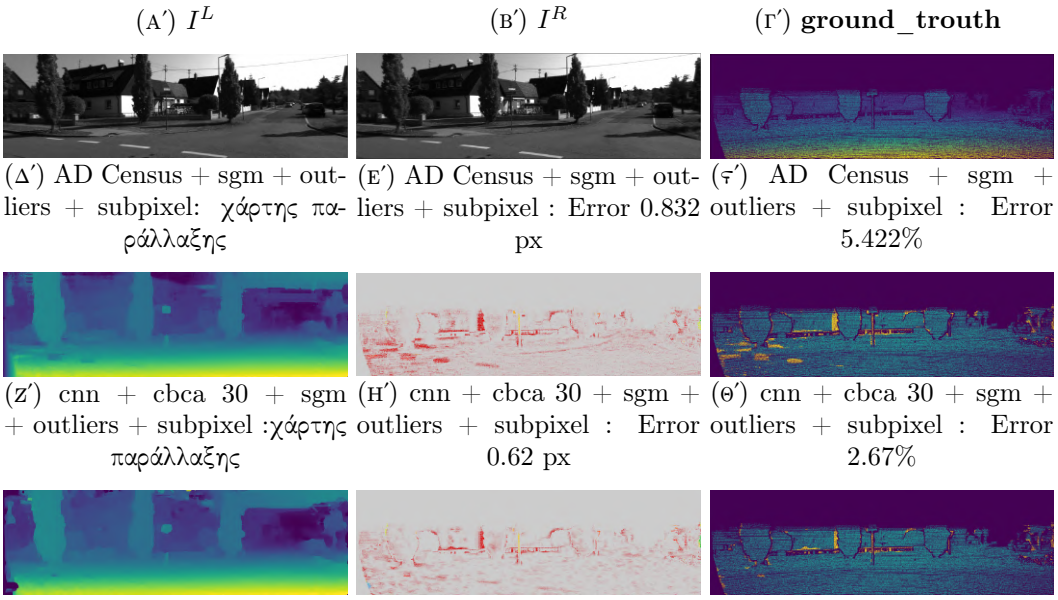
Σχημα 4.10: Τα αποτελέσματα της μεθόδου cross-based cost aggregation συναρτήσει των επαναλήψεων εφαρμογής της.



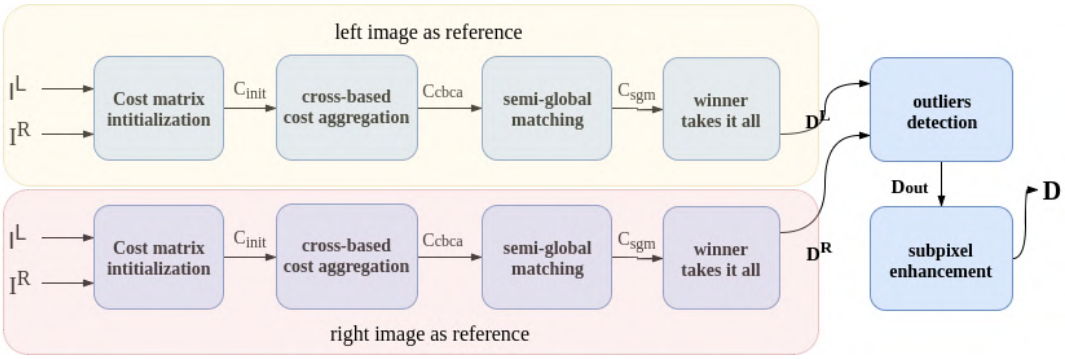
Σχημα 4.11: Παραδείγματα εφαρμογής semi global matching.



Σχημα 4.12: Παραδείγματα εφαρμογής outliers.



Σχημα 4.13: Παραδείγματα εφαρμογής subpixel enhancement.



Σχημα 4.14: Σχεδιάγραμμα ολόκληρης της μεθοδολογίας.

Πίνακας παραμέτρων	
Νευρωνικό δίκτυο	
patch_size	19×19
num_conv_layers	9
f_maps	64
kernel_size	3
pcg_tr	0.6
batches_limit	40
batches_size	128
learning_rate	0.01
Cross based cost aggregation	
intensity_threshold	0.13
distance_threshold	5
cbca_num_iterations_1	2
Semi global matching	
P1_ref	1.5
P2_ref	30
thres	0.13
small_factor	3
big_factor	6
Γενικές παράμετροι	
max_disparity	230

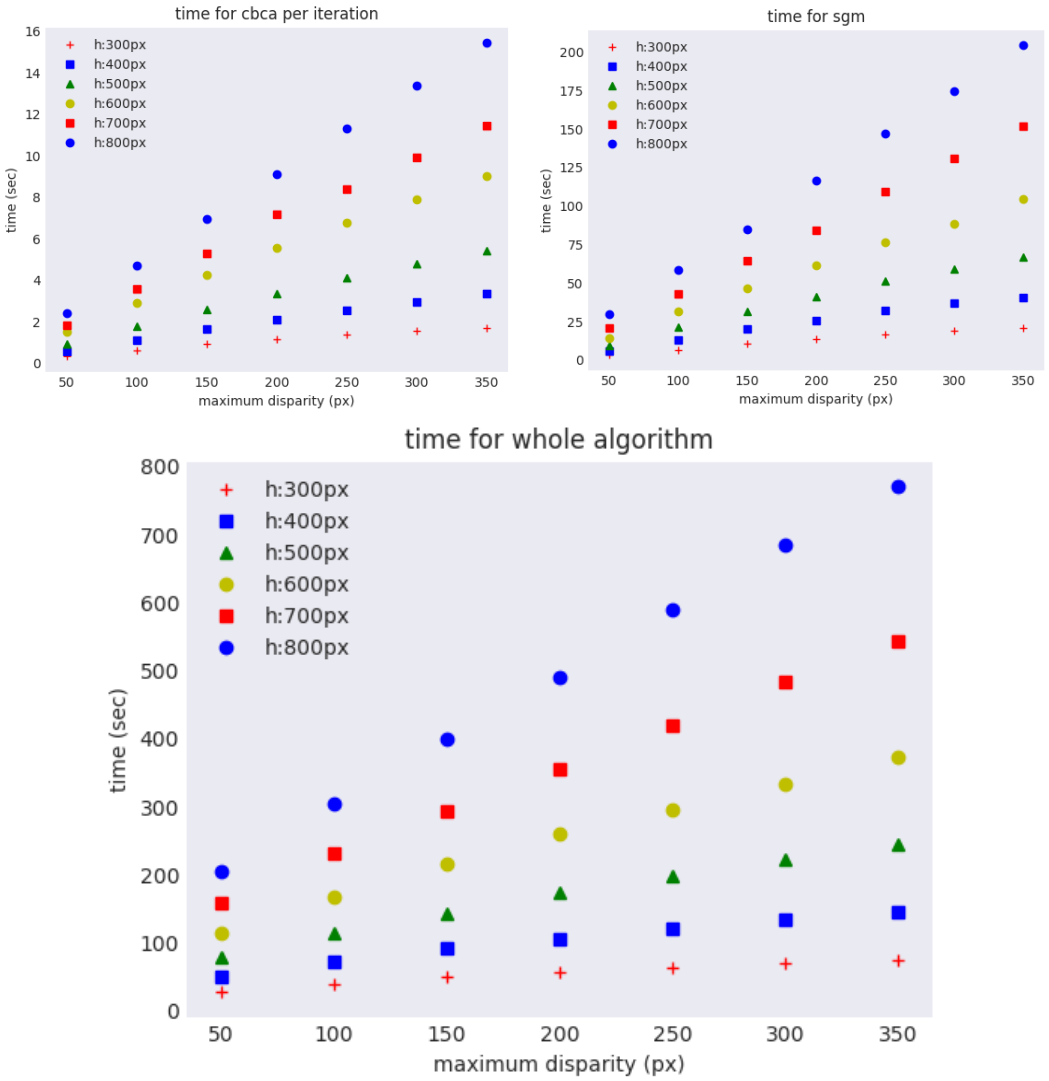
Πινακας 4.2: Όλες οι τιμές των παραμέτρων που επιλέχθηκαν.

με τεράστια καθυστέρηση στα σημεία των αλγορίθμων που ήταν αδύνατη η χρήση των έτοιμων vectorised πράξεων πάνω σε πίνακες του πακέτου numpy, όπως για παράδειγμα στα βήματα cross-based cost aggregation και semi-global matching. Σε αυτά τα βήματα, οι υπολογισμοί γίνονται ξεχωριστά για κάθε pixel και για κάθε επίπεδο παράλλαξης (εντός τριπλού βρόγχου επανάληψης `for`) λόγω διαφορετικής περιοχής υποστήριξης ανά pixel. Η επιτάχυνση που προκύπτει από την στατική δήλωση των μεταβλητών και την μεταγλώττιση (compiling) ολόκληρου του κώδικα πριν την εκτέλεσή του, μειώνει τον χρόνο εκτέλεσης των αλγορίθμων σε υποφερτά επίπεδα.

Σχεδόν όλα τα βήματα της στερεοσκοπικής μεθόδου είναι παραλληλοποιήσιμα και μπορούν να επιταχυνθούν εξαιρετικά, με κατάλληλη υλοποίηση στο περιβάλλον CUDA ώστε να εκτελούνται σε κάρτα γραφικών (GPU). Η συνολική επιτάχυνση μπορεί να φτάσει τις $x100$, με αποτέλεσμα η συνολική μεθοδολογία να περατώνεται σε χρόνους $< 1\mathbf{sec}$. Στην παρούσα εργασία μας ενδιέφερε περισσότερο η ποιοτική αξιολόγηση των αποτελεσμάτων σε σχέση με την επίτευξη μιας γρήγορης υλοποίησης γι΄ αυτό αρκεστήκαμε στην σειριακή εκδοχή των αλγορίθμων.

4.5 Συνολικά αποτελέσματα

Εκτελούμε όλη την μέθοδολογία στα σετ δεδομένων KITTI 2012 και KITTI 2015. Οι τιμές των παραμέτρων που χρησιμοποιούμε φαίνονται στον πίνακα 4.2.



Σχημα 4.15: Μελέτη χρόνου εκτέλεσης όλων των μεθόδων του στερεοσκοπικού αλγορίθμου συναρτήσει του μεγέθους της εικόνας και του ορίου της μέγιστης δυνατής παράλλαξης. Το μέγεθος της εικόνας είναι $\mathbf{H} \times \mathbf{W} = \mathbf{H} \times {}^{16}/_{9}\mathbf{H}$. Παρατηρούμε ότι το σκέλος του νευρωνικού δικτύου, αν και το πιο δαπανηρό από πλευράς υπολογισμών, καταναλώνει πολύ λίγο χρόνο, λόγω της καθόλα παραλληλοποιημένης εκτέλεσής του στην κάρτα γραφικών. Αντιθέτως, οι υπόλοιπες συναρτήσεις, λόγω σειριακής εκτέλεσης στον επεξεργαστή, καταναλώνουν αρκετά περισσότερο χρόνο.

KITTI 2012 - Αποτελέσματα

	Training set	Validation set	total
Σύνολο εικόνων	116	78	194
AD-Census			
Μέσο σφάλμα %	–	–	29.397
Μέγιστο σφάλμα %	–	–	70.779
Ελάχιστο σφάλμα %	–	–	8.37
Μέσο σφάλμα απόστασης px	–	–	12.862
Μέγιστο σφάλμα απόστασης px	–	–	63.614
Ελάχιστο σφάλμα απόστασης px	–	–	2.301
AD-Census + stereo method			
Μέσο σφάλμα %	–	–	7.795
Μέγιστο σφάλμα %	–	–	37.77
Ελάχιστο σφάλμα %	–	–	0.759
Μέσο σφάλμα απόστασης px	–	–	1.856
Μέγιστο σφάλμα απόστασης px	–	–	21.659
Ελάχιστο σφάλμα απόστασης px	–	–	0.428
cnn			
Μέσο σφάλμα %	9.353	10.048	9.632
Μέγιστο σφάλμα %	33.209	39.376	39.376
Ελάχιστο σφάλμα %	1.256	1.288	1.256
Μέσο σφάλμα απόστασης px	3.634	3.828	3.712
Μέγιστο σφάλμα απόστασης px	19.216	13.808	19.216
Ελάχιστο σφάλμα απόστασης px	0.898	0.843	0.843
cnn + stereo method			
Μέσο σφάλμα %	5.648	5.992	5.787
Μέγιστο σφάλμα %	15.742	29.334	29.334
Ελάχιστο σφάλμα %	1.376	1.3	1.3
Μέσο σφάλμα απόστασης px	1.328	1.445	1.357
Μέγιστο σφάλμα απόστασης px	4.4986	6.326	6.326
Ελάχιστο σφάλμα απόστασης px	0.498	0.527	0.498

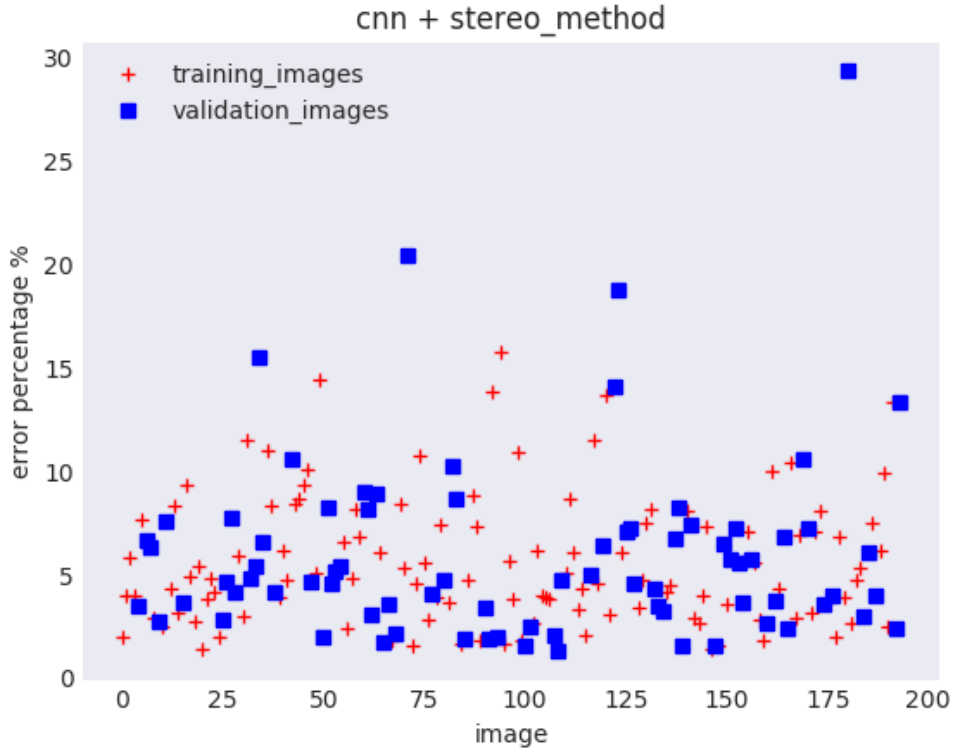
Table 4.3: Περίληψη αποτελεσμάτων στο σετ δεδομένων KITTI 2012.

4.5.1 KITTI 2012

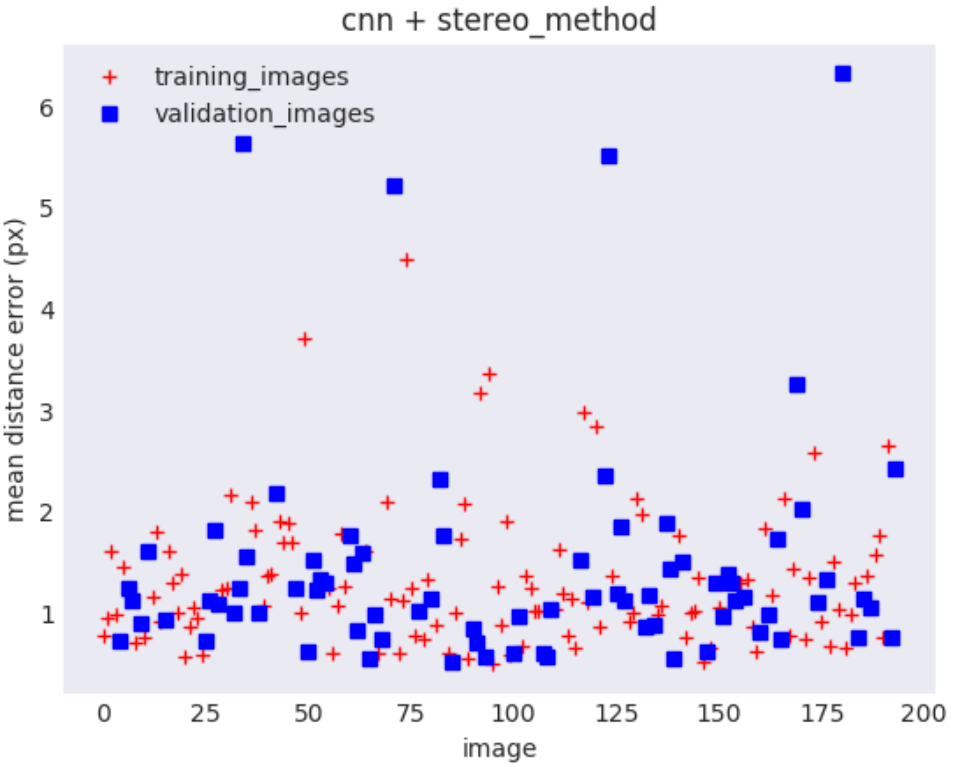
Τα συνολικά αποτελέσματα παρατίθενται στον πίνακα 4.3. Το μέσο σφάλμα είναι 5.648% και 5.992% στο σετ εκπαίδευσης και επικύρωσης αντίστοιχα. Η μικρή απόκλιση ανάμεσα στις δύο τιμές μας επιβεβαιώνει ότι έχουμε αποφύγει την υπερπροσαρμογή. Τα αντίστοιχα μέσα απόλυτα σφάλματα είναι 1.328**px** και 1.445**px** αντίστοιχα. Στα γραφήματα 4.16α′, 4.16β′, 4.17α′, 4.17β′, 4.18α′, 4.18β′, αποτυπώνονται αναλυτικά τα αποτελέσματα στο σετ εκπαίδευσης και επικύρωσης της συλλογής.

4.5.2 KITTI 2015

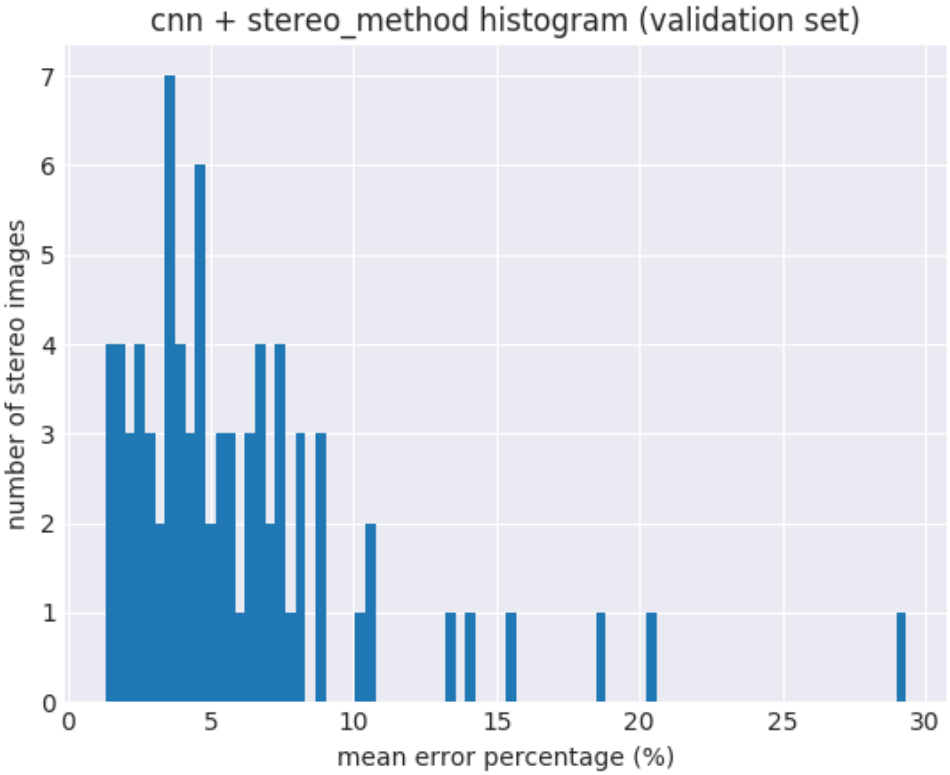
Τα συνολικά αποτελέσματα παρατίθενται στον πίνακα 4.4. Το μέσο σφάλμα είναι 6.545% και το μέσο απόλυτο σφάλμα 1.577**px**. Οι ιδιαίτερα καλές επιδόσεις επιβεβαιώνουν ότι η ικανότητα του νευρωνικού δικτύου να αρχικοποιεί τον πίνακα κόστους έχει γενική αξία,



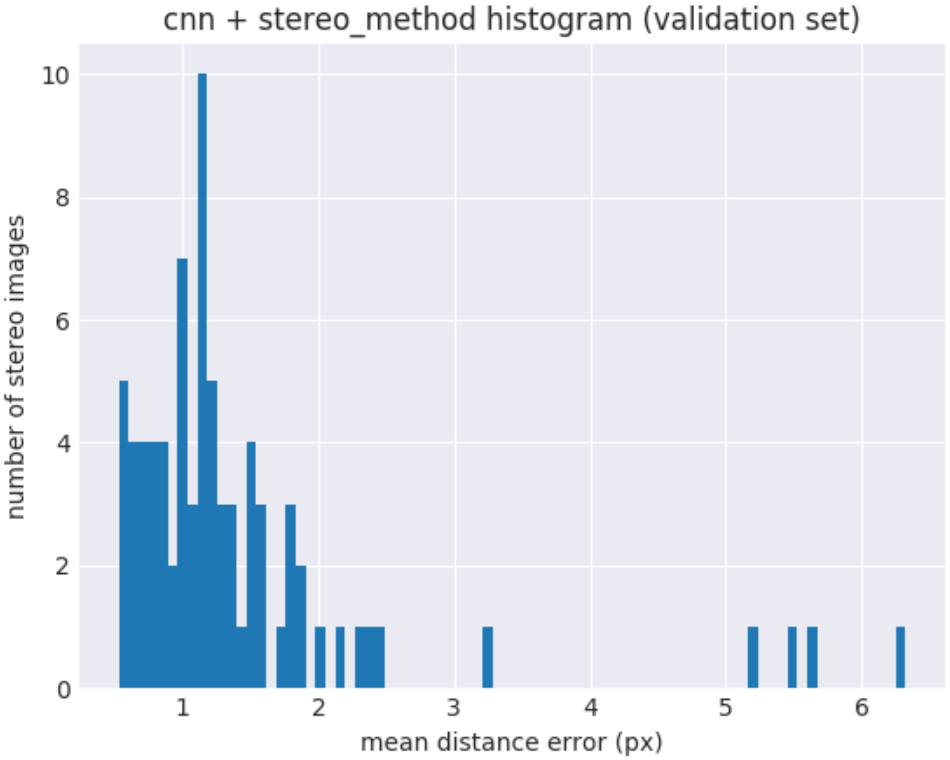
(α′) KITTI 2012: Ποσοστό σφάλματος ανά εικόνα.



(β′) KITTI 2012: Απόλυτο σφάλμα απόστασης ανά εικόνα.

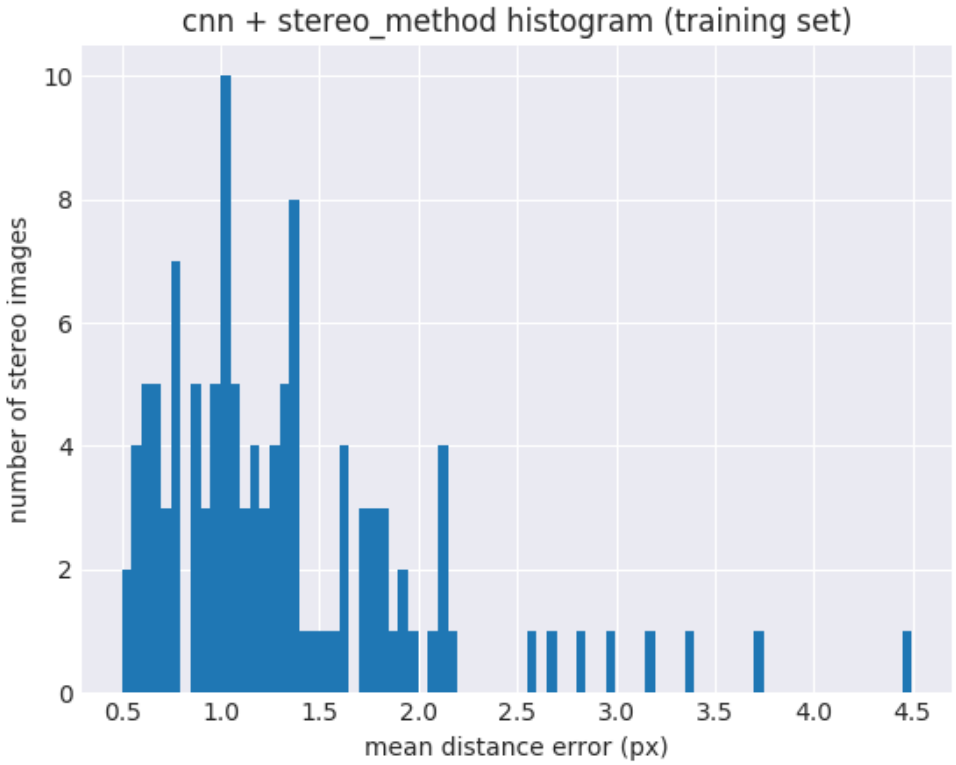


(Α′) KITTI 2012: Ιστόγραμμα ποσοστών σφάλματος στο σύνολο του σετ επικύρωσης.



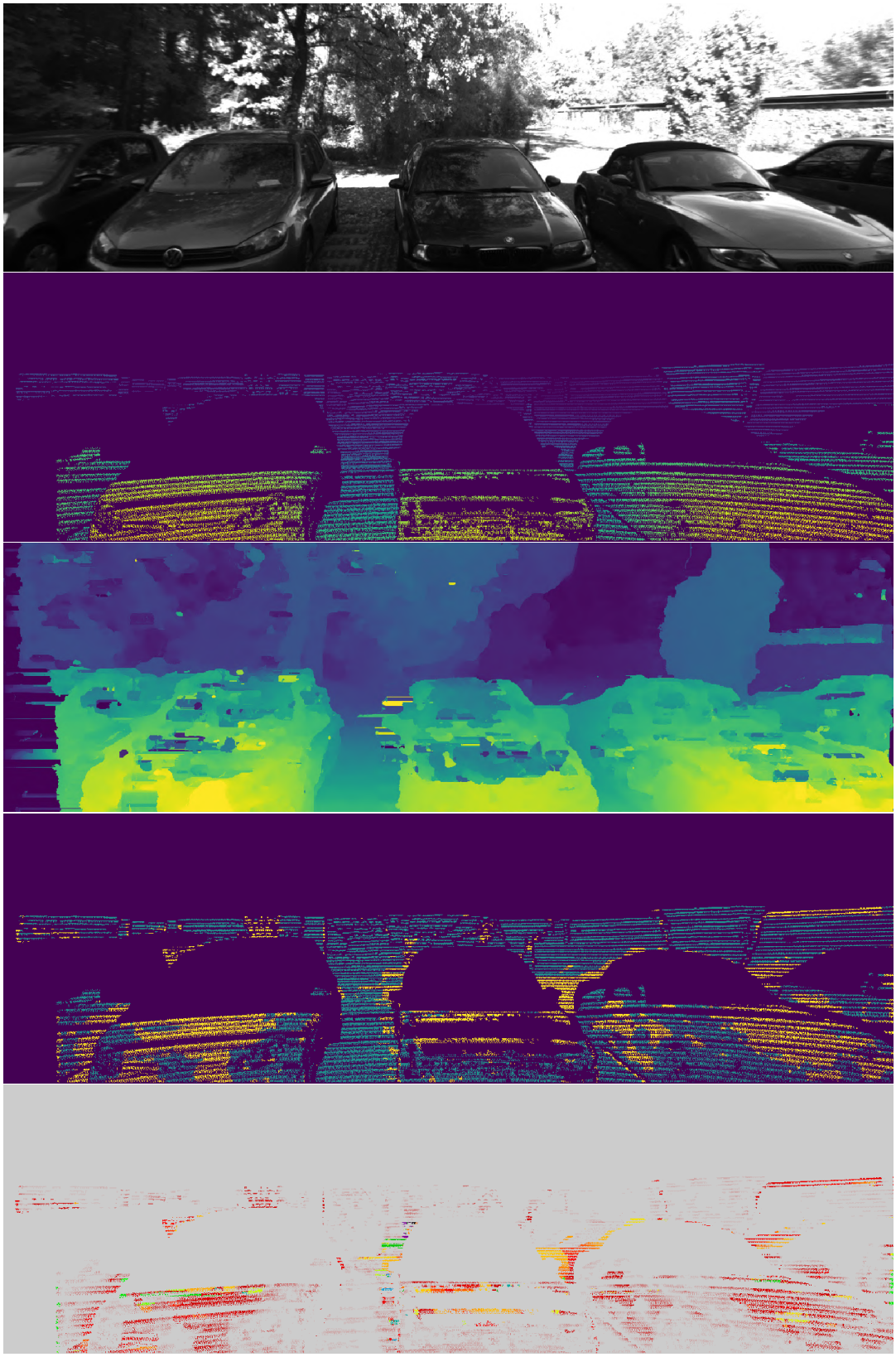
(Β′) KITTI 2012: Ιστόγραμμα απόλυτου σφάλματος στο σύνολο του σετ επικύρωσης.

(Α′) KITTI 2012: Ιστόγραμμα ποσοστών σφάλματος στο σύνολο του σετ εκπαίδευσης.



(Β′) KITTI 2012: Ιστόγραμμα απόλυτου σφάλματος στο σύνολο του σετ εκπαίδευσης.

Σχημα 4.19: Στερεοσκοπικό ζεύγος 108 του σετ επικύρωσης της συλλογής KITTI 2012. Καλύτερη επίδοση του αλγορίθμου με ποσοστό σφάλματος 1.3% και μέσο απόλυτο σφάλμα 1.445 px.



Σχημα 4.20: Στερεοσκοπικό ζεύγος 180 του σετ επικύρωσης της συλλογής KITTI 2012. Χειρότερη επίδοση του αλγορίθμου με ποσοστό σφάλματος 29.334% και μέσο απόλυτο σφάλμα 6.326 px.

KITTI 2015 - Αποτελέσματα	
Σύνολο εικόνων	200
AD-Census	
Μέσο σφάλμα %	27.064
Μέγιστο σφάλμα %	85.003
Ελάχιστο σφάλμα %	8.145
Μέσο σφάλμα απόστασης px	10.218
Μέγιστο σφάλμα απόστασης px	47.903
Ελάχιστο σφάλμα απόστασης px	3.284
AD-Census + stereo method	
Μέσο σφάλμα %	8.097
Μέγιστο σφάλμα %	53.295
Ελάχιστο σφάλμα %	0.755
Μέσο σφάλμα απόστασης px	1.546
Μέγιστο σφάλμα απόστασης px	7.56
Ελάχιστο σφάλμα απόστασης px	0.58
cnn	
Μέσο σφάλμα %	10.558
Μέγιστο σφάλμα %	81.191
Ελάχιστο σφάλμα %	1.846
Μέσο σφάλμα απόστασης px	2.338
Μέγιστο σφάλμα απόστασης px	16.57
Ελάχιστο σφάλμα απόστασης px	0.846
cnn + stereo method	
Μέσο σφάλμα %	6.545
Μέγιστο σφάλμα %	80.714
Ελάχιστο σφάλμα %	1.248
Μέσο σφάλμα απόστασης px	1.577
Μέγιστο σφάλμα απόστασης px	38.537
Ελάχιστο σφάλμα απόστασης px	0.663

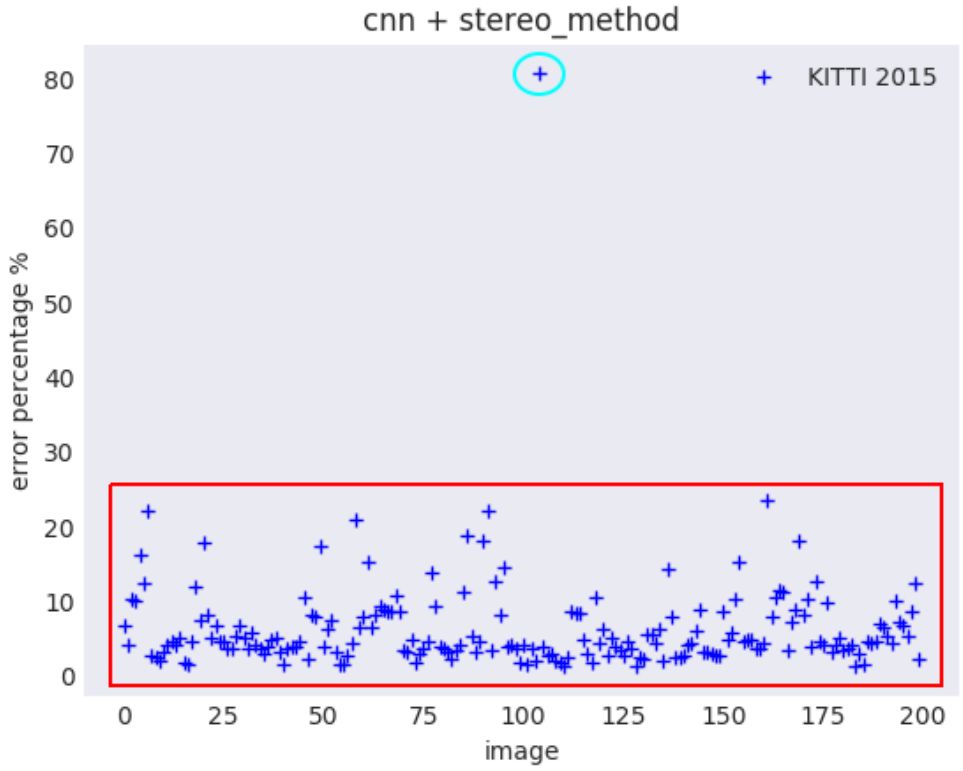
Table 4.4: Περίληψη αποτελεσμάτων στο σετ δεδομένων KITTI 2015.

καθώς αντιμετωπίζει επιτυχημένα σετ δεδομένων στο οποίο δεν έχει εκπαιδευτεί.[3] Στα γραφήματα 4.21α′, 4.21β′, 4.22α′, 4.22β′, 4.23α′, 4.23β′, 4.24α′, 4.24β′, αποτυπώνονται αναλυτικά τα αποτελέσματα στο σετ δεδομένων KITTI 2015.

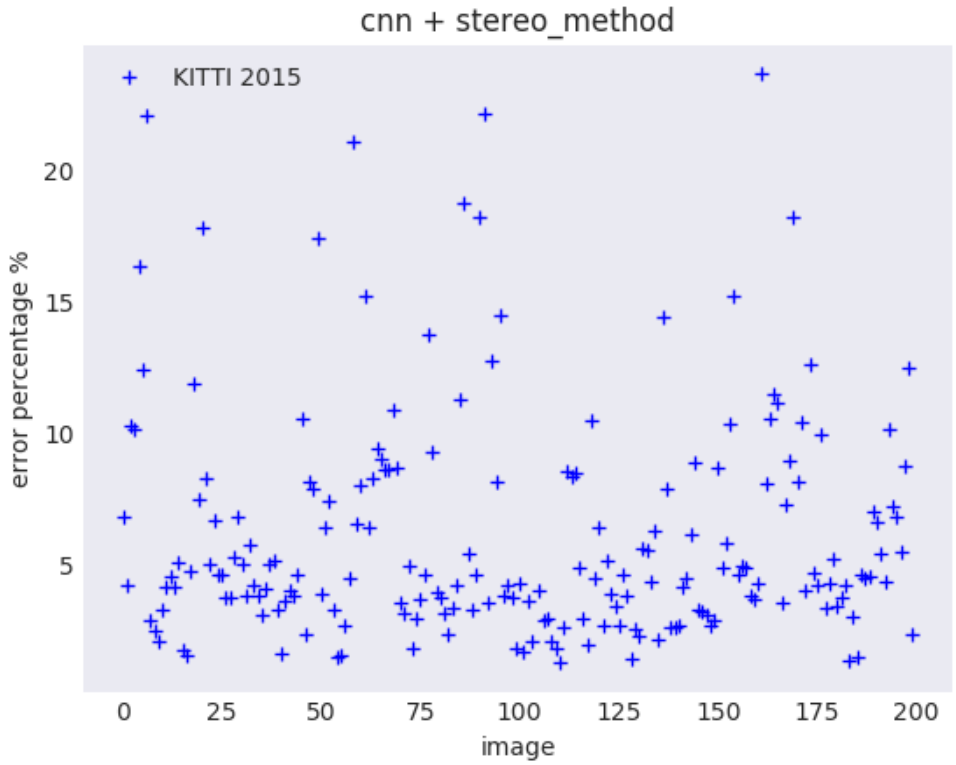
4.5.3 Middlebury

Η συλλογή Middlebury αποτελείται από τα σετ δεδομένων των χρονολογιών 2003, 2005, 2006 και 2014, με σύνολο παραδειγμάτων 2, 6, 21 και 22 αντίστοιχα. Διαφέρει έντονα σε σχέση με τις συλλογές KITTI στο είδος των εικόνων που περιέχει, όπως αναλύθηκε στο προηγούμενο κεφάλαιο. Η αξιόλογη επίδοση του αλγορίθμου και σε αυτή τη συλλογή επικυρώνει ότι το νευρωνικό δίκτυο μπορεί να ανταποκριθεί σε στερεοσκοπικά ζεύγη ιδιαίτερα διαφορετικών χαρακτηριστικών από αυτά στα οποία έχει εκπαιδευτεί. Το μέσο

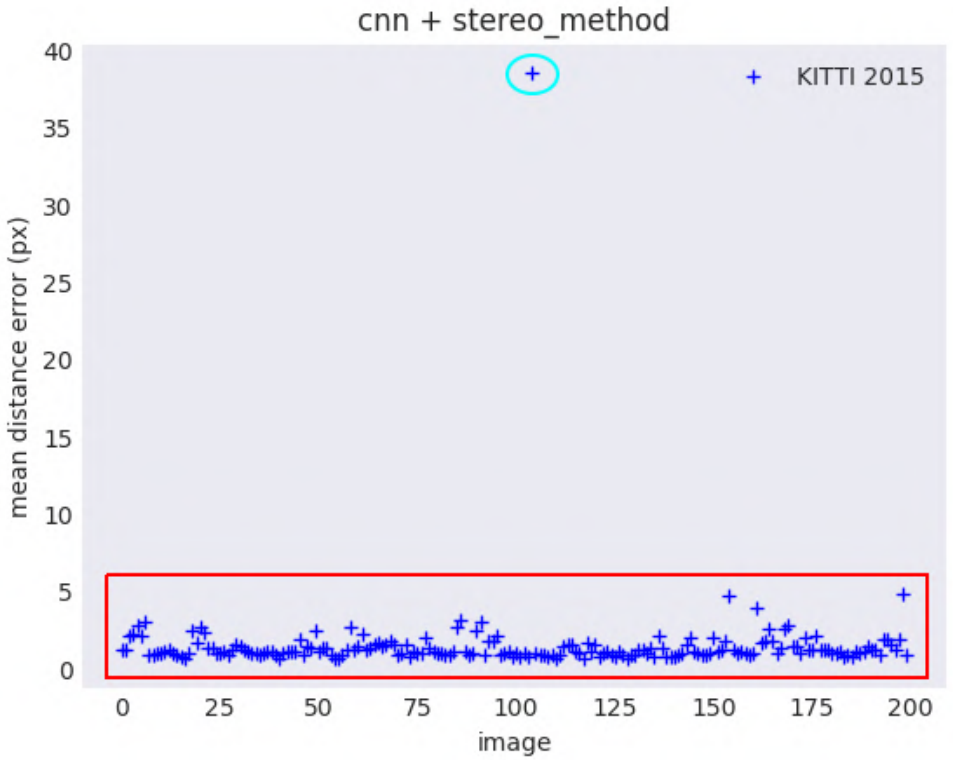
[3] Βέβαια η συλλογή δεδομένων KITTI 2015 περιέχει παρόμοιας κατηγορίας παραδείγματα με αυτά της KITTI 2012.



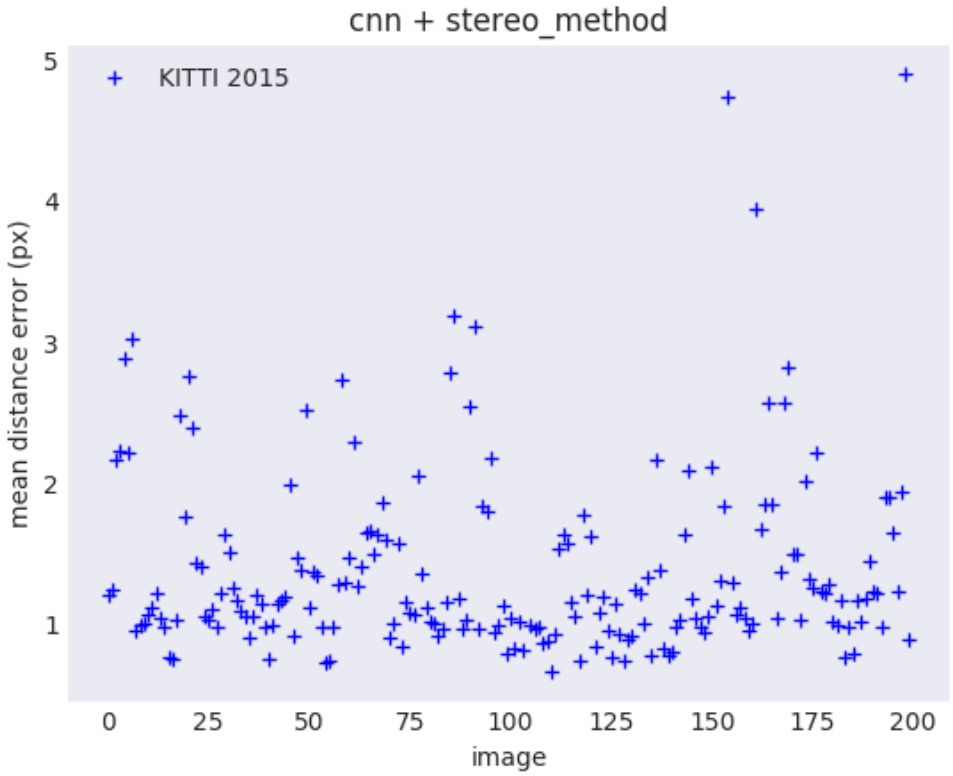
(Α′) KITTI 2015: Ποσοστό σφάλματος ανά εικόνα.



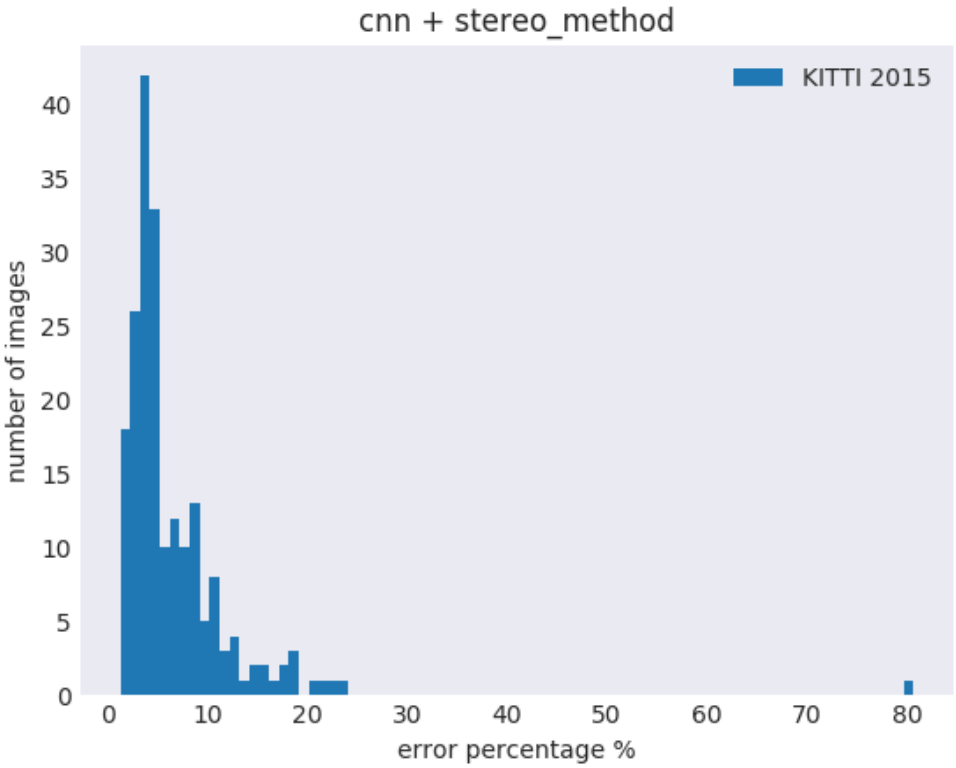
(Β′) KITTI 2015: Ποσοστό σφάλματος ανά εικόνα εστιασμένο στις τιμές του κόκκινου πλαισίου, εξαιρώντας την εξωκείμενη τιμή.



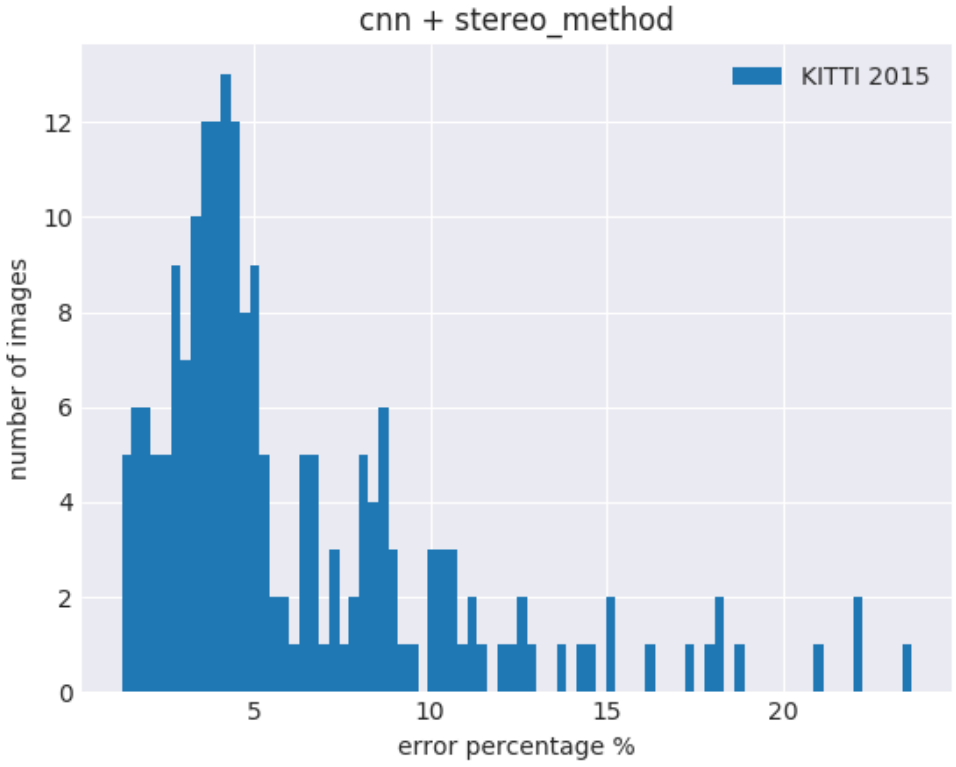
(Α′) KITTI 2015: Απόλυτο σφάλμα απόστασης ανά εικόνα.



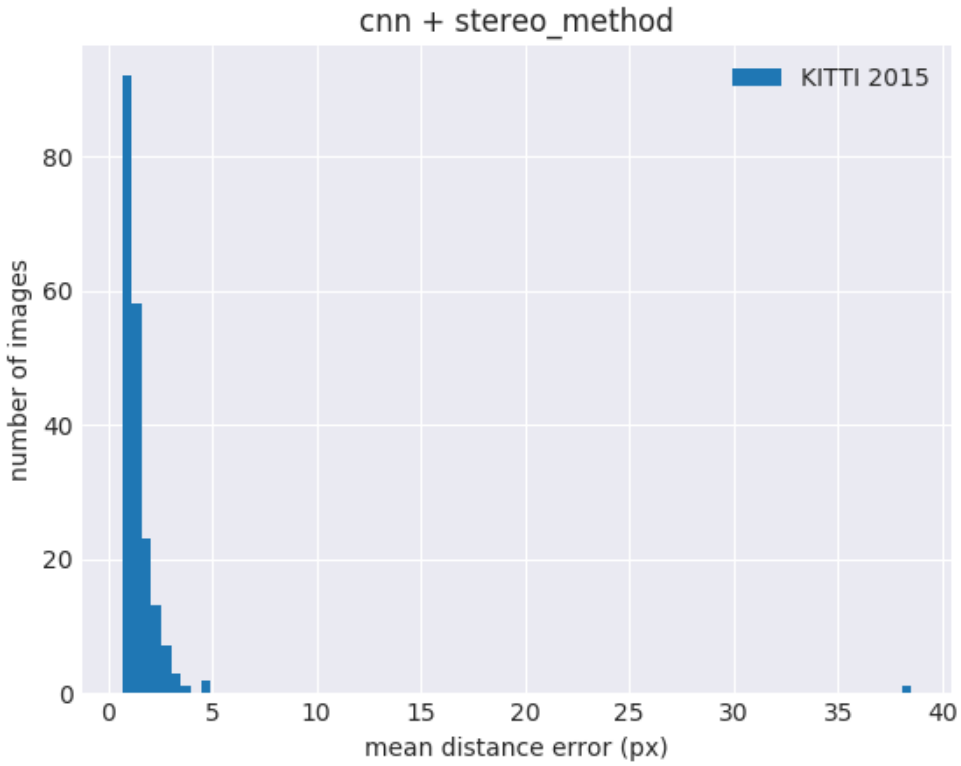
(Β′) KITTI 2015: Απόλυτο σφάλμα απόστασης ανά εικόνα, εστιασμένο στις τιμές του κόκκινου πλαισίου.



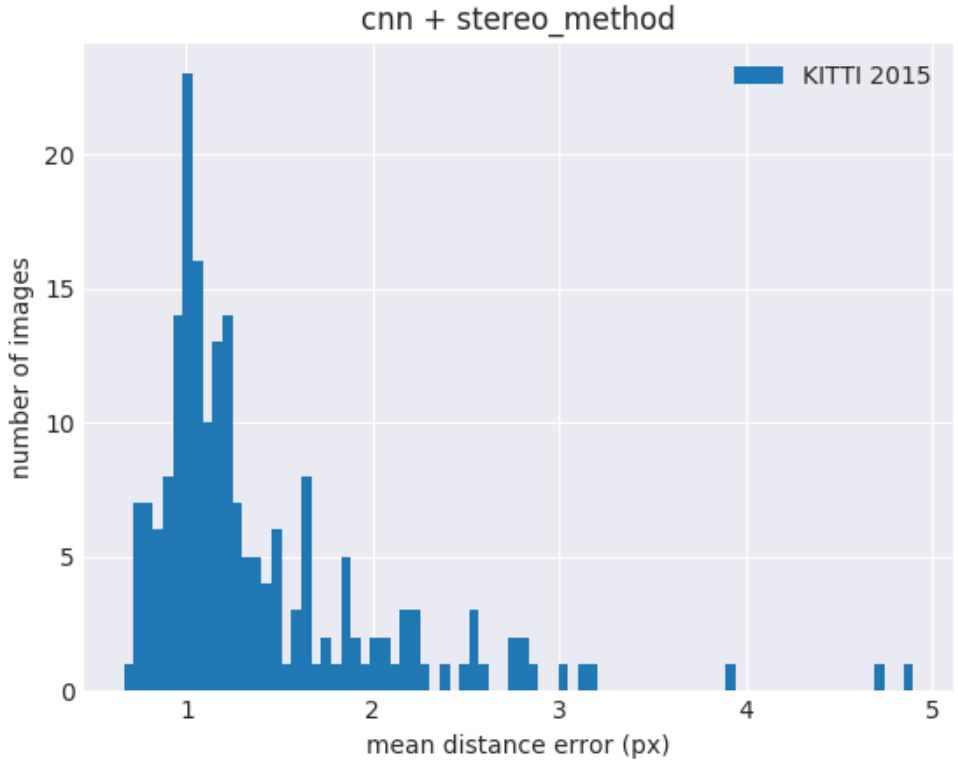
(Α′) KITTI 2015: Ιστόγραμμα ποσοστών σφάλματος στο σύνολο της συλλογής.



(Β′) KITTI 2015: Ιστόγραμμα ποσοστών σφάλματος στο σύνολο της συλλογής, εστιασμένο στις τιμές του κόκκινου πλαισίου.

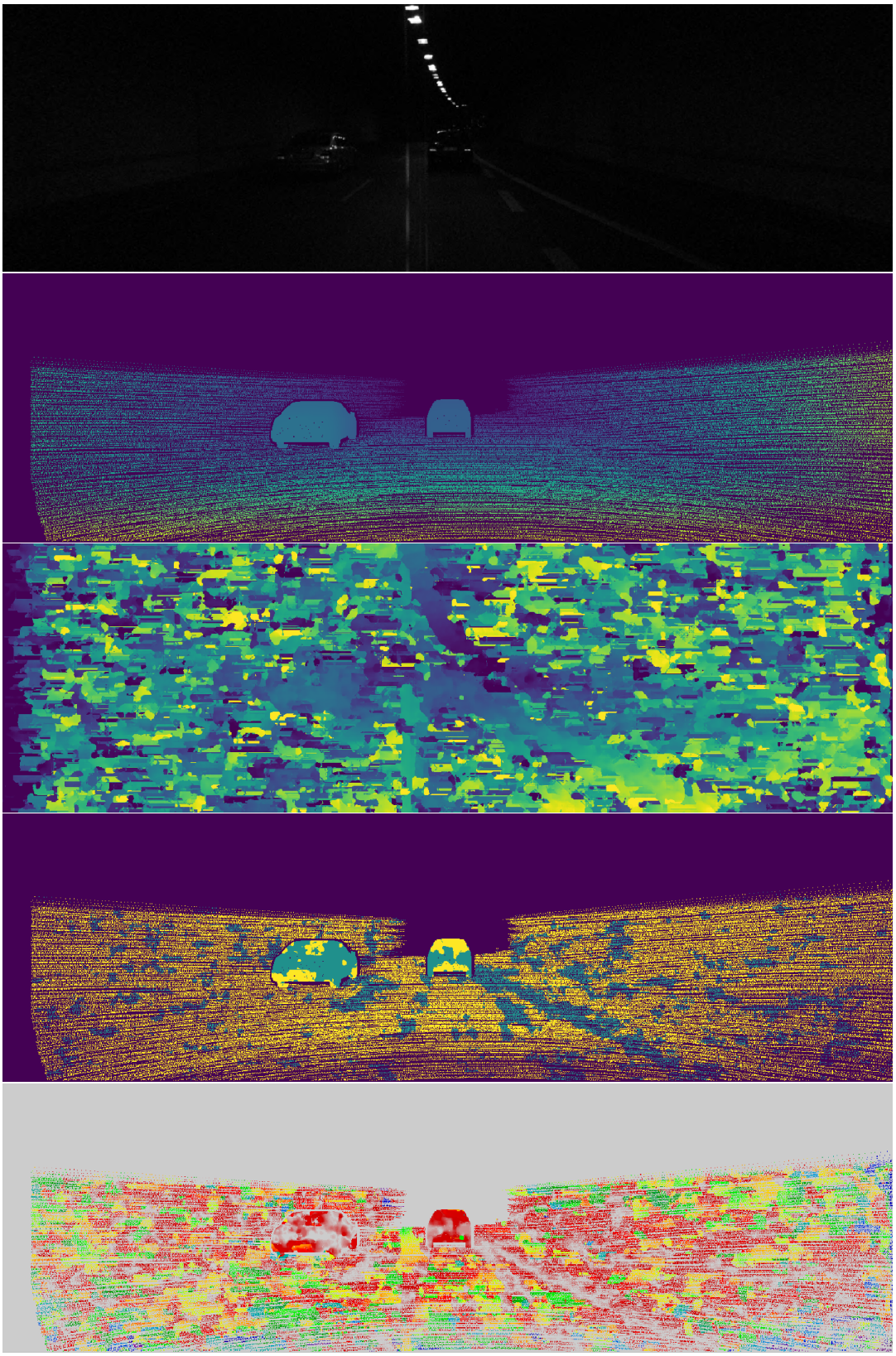


(Α′) KITTI 2015: Ιστόγραμμα απόλυτου σφάλματος στο σύνολο της συλλογής.



(Β′) KITTI 2015: Ιστόγραμμα απόλυτου σφάλματος στο σύνολο της συλλογής, εστιασμένο στις τιμές του κόκκινου πλαισίου.

Σχημα 4.25: Στερεοσκοπικό ζεύγος 110 της συλλογής KITTI 2015. Καλύτερη επίδοση του αλγορίθμου με ποσοστό σφάλματος 1.248% και μέσο απόλυτο σφάλμα 1.577 px.



Σχημα 4.26: Στερεοσκοπικό ζεύγος 104 της συλλογής KITTI 2015. Χειρότερη επίδοση του αλγορίθμου με ποσοστό σφάλματος 80.714% και μέσο απόλυτο σφάλμα 38.537 px. Μας ξεφτίλισε.

Middlebury - Αποτελέσματα

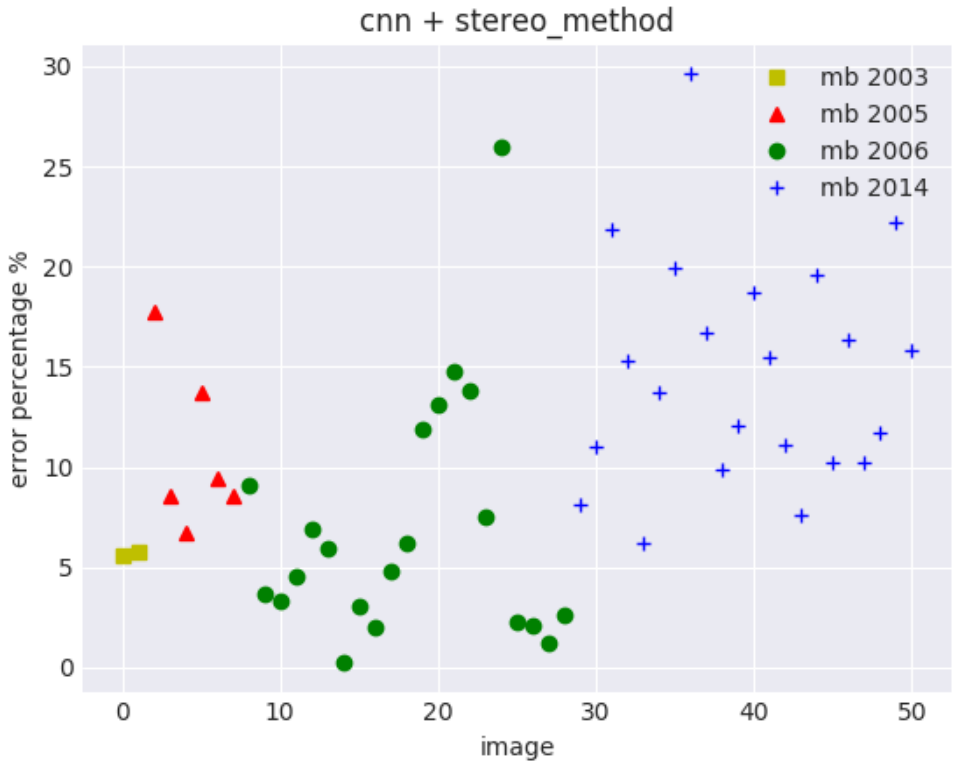
Dataset	2003	2005	2006	2014
Σύνολο εικόνων	2	6	21	22
AD-Census				
Μέσο σφάλμα %	16.578	21.32	18.591	30.894
Μέγιστο σφάλμα %	20.057	32.277	59.612	46.963
Ελάχιστο σφάλμα %	13.100	11.797	0.419	15.544
Μέσο σφάλμα απόστασης px	3.779	4.828	4, 776	7.144
Μέγιστο σφάλμα απόστασης px	2.888	7.38	17.782	20.791
Ελάχιστο σφάλμα απόστασης px	4.671	2.307	0.35	2.839
AD-Census + stereo method				
Μέσο σφάλμα %	7.693	11.983	7.41	18.134
Μέγιστο σφάλμα %	8.107	17.548	27.507	33.699
Ελάχιστο σφάλμα %	7.278	7.798	0.209	7.112
Μέσο σφάλμα απόστασης px	1.249	1.935	1.77	3.414
Μέγιστο σφάλμα απόστασης px	1.276	4.213	6.153	11.316
Ελάχιστο σφάλμα απόστασης px	1.221	0.989	0.204	1.12
cnn				
Μέσο σφάλμα %	10.030	17.012	14.313	22.912
Μέγιστο σφάλμα %	11.357	25.814	50.741	38.017
Ελάχιστο σφάλμα %	8.704	9.865	0.569	11.026
Μέσο σφάλμα απόστασης px	2.527	4.177	4.106	5.362
Μέγιστο σφάλμα απόστασης px	2.906	7.057	16.837	15.151
Ελάχιστο σφάλμα απόστασης px	2.148	2.315	0.481	2.168
cnn + stereo method				
Μέσο σφάλμα %	5.685	10.81	6.902	14.688
Μέγιστο σφάλμα %	5.754	17.733	25.978	29.615
Ελάχιστο σφάλμα %	5.616	6.761	0.215	6.164
Μέσο σφάλμα απόστασης px	1.197	2.144	1.881	2.967
Μέγιστο σφάλμα απόστασης px	1.213	4.424	7.952	9.989
Ελάχιστο σφάλμα απόστασης px	1.182	1.125	0.23	1.26

TABLE 4.5: Περίληψη αποτελεσμάτων στο σετ δεδομένων Middlebury.

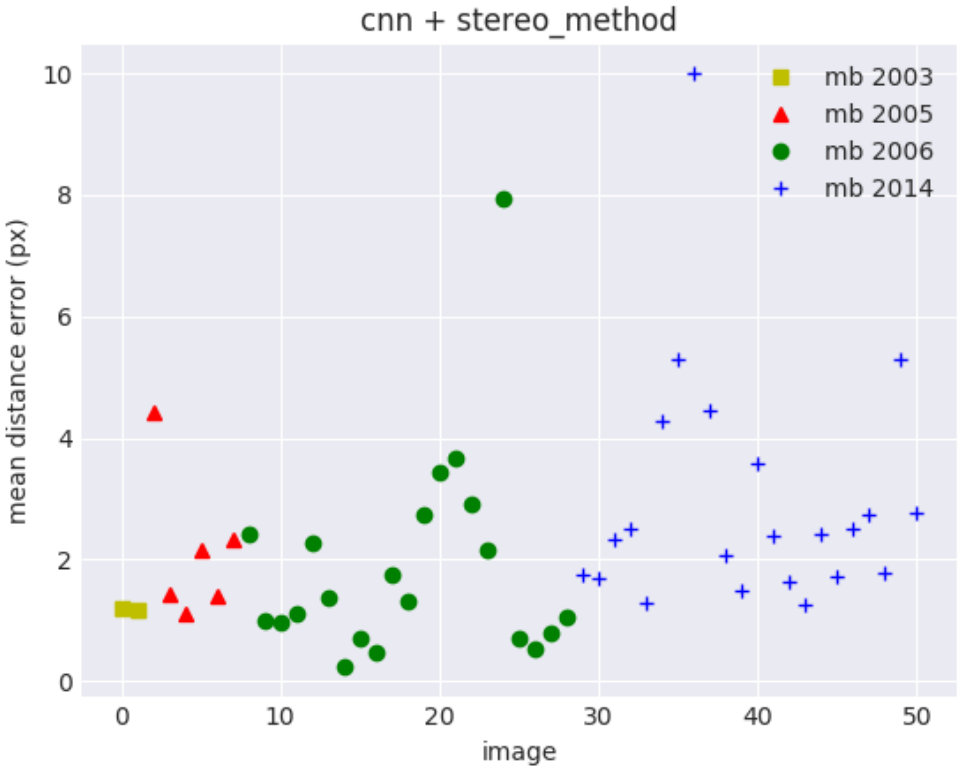
ποσοστό σφάλματος είναι 5.685%, 10.81%, 6.902%, 14.688% και το μέσο απόλυτο σφάλμα 1.197**px**, 2.144**px**, 1.881**px** και 2.967**px** αντίστοιχα. Στα γραφήματα 4.27α′, 4.27β′, 4.28α′ και 4.27α′ και στον πίνακα 4.5 παρουσιάζονται αναλυτικά τα αποτελέσματα.

4.6 Συμπεράσματα - Προτάσεις για το μέλλον

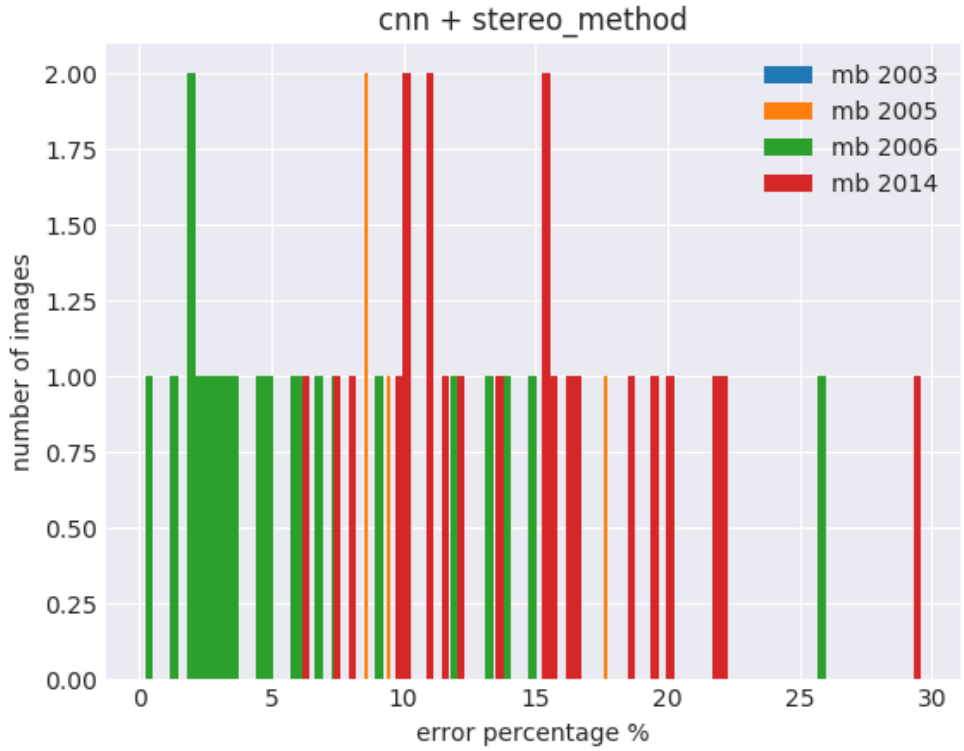
- Τα αποτελέσματα δικαιώνουν την αρχική πρόθεση της εργασίας να αποδείξει ότι μέσω μηχανικής μάθησης μπορεί να βελτιωθεί η ακρίβεια της σύγκρισης ομοιότητας. Συγκεκριμένα, το νευρωνικό δίκτυο που εκπαιδεύσαμε εμφανίζει αρκετά καλύτερα αποτελέσματα σε σχέση με την AD-Census η οποία είναι ίσως η πληρέστερη μέθοδος εκτίμησης του χάρτη παράλλαξης, χωρίς την χρήση μηχανικής μάθησης.
- Η βελτίωση δεν περιορίζεται σε παραδείγματα παρόμοια με εκείνα στα οποία το τεχνητό νευρωνικό δίκτυο έχει εκπαιδευτεί, αλλά επεκτείνεται και σε στερεοσκοπικά



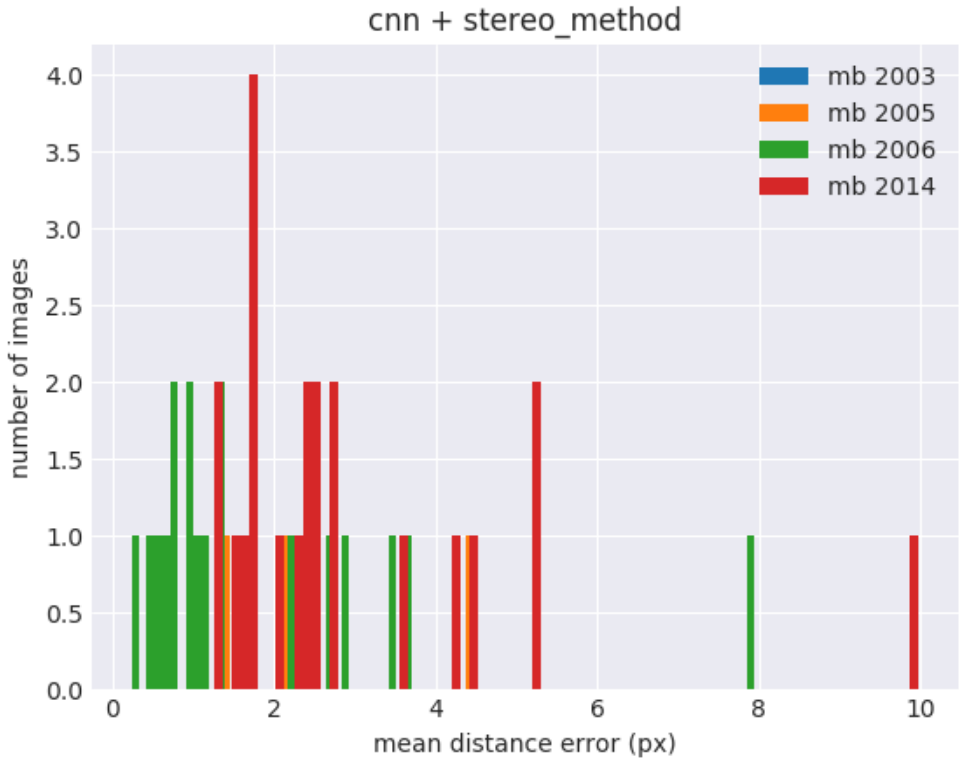
(Α′) Middlebury: Ποσοστό σφάλματος ανά εικόνα.



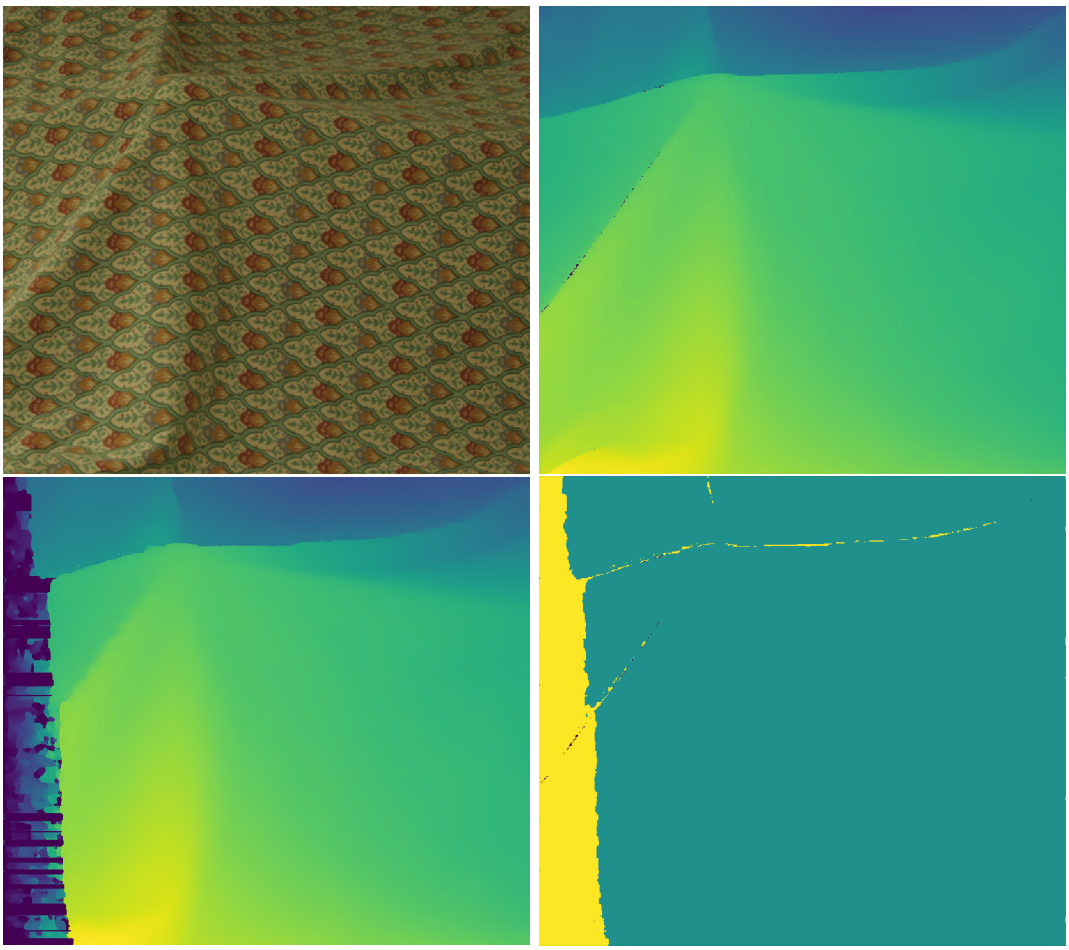
(Β′) Middlebury: Απόλυτο σφάλμα ανά εικόνα.



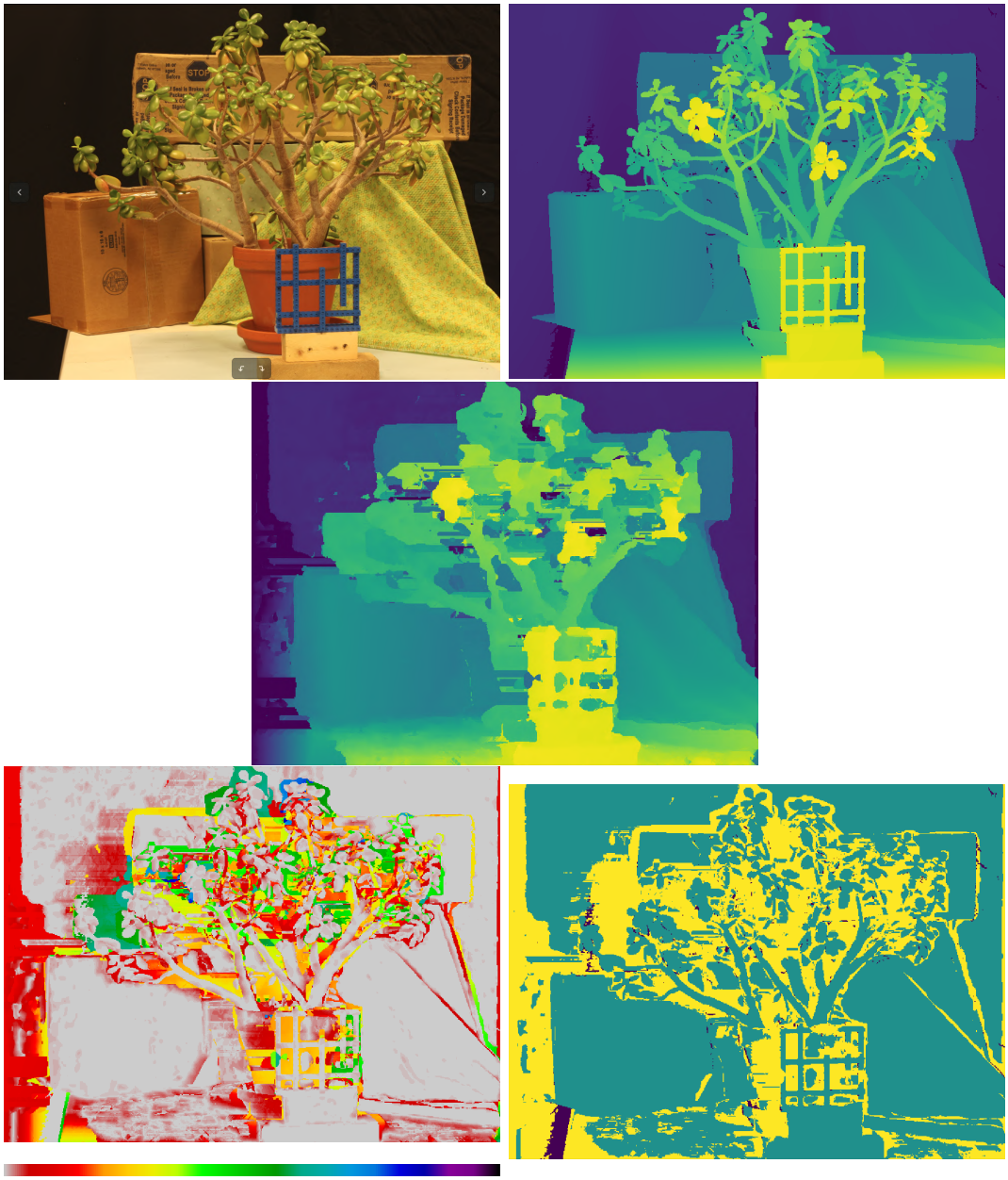
(Α′) Middlebury: Ιστόγραμμα ποσοστών σφάλματος στο σύνολο της συλλογής.



(Β′) Middlebury: Ιστόγραμμα απόλυτου σφάλματος στο σύνολο της συλλογής.



Σχημα 4.29: Στερεοσκοπικό ζεύγος 16 της συλλογής Middlebury 2006. Καλύτερη επίδοση του αλγορίθμου με ποσοστό σφάλματος 0.215% και μέσο απόλυτο σφάλμα 0.23 px.



Σχημα 4.30: Στερεοσκοπικό ζεύγος 7 της συλλογής Middlebury 2014. Χειρότερη επίδοση του αλγορίθμου με ποσοστό σφάλματος 29.615% και μέσο απόλυτο σφάλμα 9.989 px.

ζεύγη διαφορετικής φύσης. Έτσι επισφραγίζεται ότι το εκπαιδευμένο μοντέλο έχει «αντιληφθεί» έναν γενικό κανόνα σύγκρισης με πολύ μεγάλο πεδίο εφαρμογής.

- Το νευρωνικό δίκτυο προβλέπει αρχικές τιμές ομοιότητες ικανές να παράξουν ικανοποιητικό χάρτη παράλλαξης, **χωρίς κανένα επιπλέον βήμα επεξεργασίας.**
- Η ποιότητα των προβλέψεων του νευρωνικού δικτύου είναι ανάλογη του μεγέθους της συλλογής δεδομένων στην οποία έχει εκπαιδευτεί.
- Η μεγιστοποίηση της απόδοσής του σε στερεοσκοπικά ζεύγη ιδιαίτερων χαρακτηριστικών απαιτεί την κατάλληλη ρύθμιση των παραμέτρων (fine-tuning) του σε ένα μικρό σετ εκπαίδευσης ενδεικτικό των παραδειγμάτων που θα καλεστεί να αντιμετωπίσει.
- Πρέπει να αναπτυχθούν περισσότερες τεχνητές στερεοσκοπικές συλλογές, γενικής και ειδικής θεματολογίας. Σε αυτήν την προσπάθεια συνεισφέρει ιδιαίτερα η κατηγορία νευρωνικών δικτύων Generative Adversarial Networks (GAN). Η τεχνητή συλλογή [27] αποτελεί χαρακτηριστικό παράδειγμα. Μέσω των (GAN) μπορούμε να δημιουργήσουμε πολύ μεγάλες συλλογές τεχνητών δεδομένων από ένα πολύ μικρότερο αρχικό σύνολο πραγματικών παραδειγμάτων.
- Η στερεοσκοπική μέθοδος πρέπει να αντικατασταθεί κι αυτή από μεθόδους μηχανικής μάθησης και ταυτόχρονα να ενσωματωθεί σε μια ενιαία εκπαιδεύσιμη μέθοδο. Το πρόβλημα της στερεοσκοπικής όρασης όπως έχει διατυπωθεί ως σήμερα φέρει την ιδιαιτερότητα να υποδιαιρείται σε δύο υποπροβλήματα (αρχικοποίηση του πίνακα κόστους και στερεοσκοπική μέθοδο) τα οποία επιλύονται ανεξάρτητα το ένα από το άλλο κι έπειτα συνενώνονται για να αποτελέσουν μια ενιαία μέθοδο. Επομένως παρατηρείται το πρόβλημα δύο βέλτιστες λύσεις στα επιμέρους υποπροβλήματα να μην αποτελούν την βέλτιστη λύση του ενιαίου προβλήματος. Πρέπει επομένως να ερευνηθεί η δυνατότητα υλοποίησης της στερεοσκοπικής μεθόδου με αλγορίθμους μηχανικής μάθησης και της δημιουργία ενός ενιαίου μοντέλου.
- Η προτεινόμενη μέθοδος μπορεί να μορφοποιηθεί κατάλληλα ως πρόβλημα unsupervised ή semi-supervised learning. Αυτό μπορεί να γίνει μέσω της επαναπροβολής του ανακατασκευασμένου μοντέλου στις δύο κάμερες.
- Η προτεινόμενη μέθοδος πρέπει να επιταχυνθεί κατά τάξη μεγέθους $50x$ ώστε να μπορεί να εφαρμοστεί σε δεδομένα βίντεο.

Παράρτημα Α΄

Παράρτημα Κεφαλαίου 2

Α΄.1 Αναλυτική έκφραση προοπτικής προβολής

Στην πραγματικότητα παρεμβάλλονται κι άλλα φαινόμενα που χρήζουν μοντελοποίησης, ώστε η συνάρτηση f_{pr} να είναι ακριβής. Μια πιο πλήρης περιγραφή της αποδίδεται από την παρακάτω σχέση: [25]

$$Z\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha & -\alpha\cot\theta & s_x & 0 \\ 0 & \frac{\beta}{\sin\theta} & s_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} * \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

όπου:

- f, η εστιακή απόσταση της κάμερας (cm)
- k, αναλογία $\frac{pixels}{cm}$ κατά τον άξονα xx'
- l, αναλογία $\frac{pixels}{cm}$ κατά τον άξονα yy'[1]
- $\alpha = k * f$, η εστιακή απόσταση εκφρασμένη σε μονάδες οριζόντιας πλευράς pixel
- $\beta = l * f$, η εστιακή απόσταση εκφρασμένη σε μονάδες κάθετης πλευράς pixel
- $[s_x, s_y]^T$, μετατόπιση του κέντρου της εικόνας από το οπτικό κέντρο στο πάνω αριστερό άκρο $(pixel)$
- θ, απόκλιση της γωνίας των αξόνων της κάμερας από την τέλεια καθετότητα των $90°$ μοιρών (rad)

Το παραπάνω μοντέλο αγνοεί το φαινόμενο της ακτινικής παραμόρφωσης (radial distortion). Η ακτινική παραμόρφωση, όπως και η απόκλιση της γωνίας των αξόνων της κάμερας, δημιουργούν αποκλίσεις που διορθώνονται με κατάλληλη επεξεργασία (camera calibration). Προκειμένου να δώσουμε έμφαση στα ζητήματα της στερεοσκοπικής γεωμετρίας, οι μαθηματικές σχέσεις που χρησιμοποιούμε δεν περιλαμβάνουν αυτά τα φαινόμενα.

[1] η αναλογία είναι συνήθως διαφορετική κατά τους δύο άξονες, καθώς τα pixels είναι παραλληλόγραμμα

Α΄.2 Αναλυτική έκφραση ευθείας

Αν έχουμε μοντελοποιήσει την προοπτική προβολή μέσω της έκφρασης του κεφαλαίου Α΄.1, η ευθεία που ενώνει το οπτικό κέντρο O της κάμερας με το σημείο $\mathbf{p}$ δίνεται από την παρακάτω πιο αναλυτική διανυσματική έκφραση:

$$(X, Y, Z) = t \cdot \left(\frac{x - s_x}{a} + \frac{\cot\theta \sin\theta y s_y}{\beta}, \frac{(y - s_y)\sin\theta}{\beta}, 1 \right), t \in [f, +\infty)$$

Α΄.3 Απόδειξη στερεοσκοπικού περιορισμού

Υπόθεση:

1. Ως σύστημα αναφοράς έχουμε ορίσει το σύστημα συντεταγμένων της αριστερής λήψης.
2. Γνωρίζουμε τον affine μετασχηματισμό $g = (R, T)$ που προσδιορίζει την θέση και τον προσανατολισμό της δεξιάς λήψης.
3. Συμβολίζουμε με $\mathbf{p_1}, \mathbf{p_2}$ τις προβολές του **ίδιου** 3D σημείου στο σύστημα συντεταγμένων της αριστερής και δεξιάς λήψης αντίστοιχα

Τότε τα δύο αντίστοιχα σημεία τηρούν την σχέση:

$$\mathbf{p_2}^T E \mathbf{p_1} = 0, \text{όπου } E = [T]R$$

Ως $[T]$ συμβολίζουμε τον πίνακα εξωτερικού γινομένου του διανύσματος T.

Ο πίνακας E ονομάζεται essential matrix.

Για δεδομένο σημείο $\mathbf{p_1}$ αναζητούμε το «αντίστοιχο σημείο» $\mathbf{p_2}$. Η απόδειξη ισχύει και για το αντίστροφο:

$$
\begin{gathered}
\mathbf{P_2} = R\mathbf{P_1} + T \Leftrightarrow \\
f \cdot \begin{bmatrix} X_2/Z_2 \\ Y_2/Z_2 \\ 1 \end{bmatrix} = R \cdot f \begin{bmatrix} X_1/Z_1 \\ Y_1/Z_1 \\ 1 \end{bmatrix} + T \xRightarrow{[T]\cdot} \\
f \cdot [T] \begin{bmatrix} X_2/Z_2 \\ Y_2/Z_2 \\ 1 \end{bmatrix} = f \cdot [T] \cdot R \begin{bmatrix} X_1/Z_1 \\ Y_1/Z_1 \\ 1 \end{bmatrix} + [T]T \xLeftrightarrow{[T]T=0} \\
[T] \begin{bmatrix} X_2/Z_2 \\ Y_2/Z_2 \\ 1 \end{bmatrix} = [T] \cdot R \begin{bmatrix} X_1/Z_1 \\ Y_1/Z_1 \\ 1 \end{bmatrix} \xRightarrow{\mathbf{p_2^T}\cdot} \\
\begin{bmatrix} X_2/Z_2 & Y_2/Z_2 & 1 \end{bmatrix} [T] \begin{bmatrix} X_2/Z_2 \\ Y_2/Z_2 \\ 1 \end{bmatrix} = \begin{bmatrix} X_2/Z_2 & Y_2/Z_2 & 1 \end{bmatrix} [T] \cdot R \begin{bmatrix} X_1/Z_1 \\ Y_1/Z_1 \\ 1 \end{bmatrix} \xLeftrightarrow{\mathbf{p_2} \perp T \times \mathbf{p_2}} \\
0 = \begin{bmatrix} X_2/Z_2 & Y_2/Z_2 & 1 \end{bmatrix} [T] \cdot R \begin{bmatrix} X_1/Z_1 \\ Y_1/Z_1 \\ 1 \end{bmatrix} \Leftrightarrow \\
\mathbf{p_2}^T [T_x] R \mathbf{p_1} = 0 \xRightarrow{E=[T]R} \\
\mathbf{p_2}^T E \mathbf{p_1} = 0
\end{gathered}
\tag{Α΄.1}
$$

Α΄.4 Λήψη εικόνας από στερεοσκοπική διάταξη

Στη στερεοσκοπική διάταξη, η θέση της δεξιάς λήψης προκύπτει από αυτή της αριστερής, με μετατόπισης μόνο κατά τον άξονα xx' και χωρίς καμία περιστροφή. 2.5β΄ Ισχύουν επομένως οι εξής περιορισμοί:

$$
g_{stereo} = \left(R = I, T = \begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix} \right) \tag{Α΄.2}
$$

$$
E_{stereo} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -b \\ 0 & b & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -b \\ 0 & b & 0 \end{bmatrix} \tag{Α΄.3}
$$

$$
\begin{gathered}
p_1^T E p_2 = 0 \Leftrightarrow \\
\begin{bmatrix} x_1 & y_1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -b \\ 0 & b & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \\ 0 \end{bmatrix} = 0 \Rightarrow \\
\begin{bmatrix} x_1 & y_1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ -b \\ by_2 \end{bmatrix} = 0 \Rightarrow \\
-by_1 + by_2 = 0 \Rightarrow \\
y_1 = y_2
\end{gathered}
\tag{Α΄.4}
$$

Ο περιορισμός Α΄.4 ονομάζεται στερεοσκοπικός περιορισμός (stereo constraint) κι αποδεικνύει ότι η στερεοσκοπική διάταξη ικανοποιεί με φυσικό τρόπο τον ορισμό της στερεοσκοπικής όρασης.

Α΄.5 Ευθυγράμμιση (Rectification)

Δεν είναι μονόδρομος η χρήση στερεοσκοπικής διάταξης για να επιτευχθεί ο στερεοσκοπικός περιορισμός. Αντιθέτως, μπορεί να επιβληθεί μετά την λήψη των εικόνων με κατάλληλη επεξεργασία. Αν γνωρίζουμε τον affine μετασχηματισμό $g = (R, T)$, η μετατροπή του τυχαίου ζεύγους εικόνων σε στερεοσκοπικό είναι άμεση[5]. Αν επίσης αγνοούμε πλήρως την σχετική θέση των δύο λήψεων, μπορούμε πρώτα να υπολογίσουμε τον πίνακα $E = [T_x]R \in \mathbb{R}^{3\times3}$ μέσω κάποιας κατάλληλης τεχνικής (π.χ. αλγόριθμος 8 σημείων) κι ακολούθως να μετατρέψουμε το ζεύγος σε στερεοσκοπικό. Αμφότερα τα σενάρια είναι αντιμετωπίσιμα, με το πρώτο να έχει μεγαλύτερη ακρίβεια σε σχέση με το δεύτερο που συσσωρεύει κάποιο σφάλμα λόγω του εισαγωγικού βήματος του υπολογισμού του πίνακα E.

Η αναλυτική μεθοδολογία επίλυσης του προβλήματος ευθυγράμμισης (όπως και της εύρεσης του πίνακα E όταν είναι αναγκαίο) ξεφεύγει των σκοπών της συγκεκριμένης εργασίας. Παρακάτω δίνεται μια σύντομη περιγραφή των βημάτων της ευθυγράμμισης για την εποπτική κατανόηση της μεθοδολογίας: [41]

Υποθέτουμε ότι γνωρίζουμε affine μετασχηματισμό $g = (R, T)$. Τότε:

- Περιστρέφουμε την αριστερή λήψη ώστε η ευθεία βάσης (ευθεία που συνδέει το οπτικό κέντρο των δύο λήψεων) με το επίπεδο που ορίζει το πέτασμα της αριστερής κάμερας να γίνουν παράλληλα. Ο πίνακας που υλοποιεί αυτήν την περιστροφή ονομάζεται $R_{rect} \in \mathbb{R}^{3\times3} : \| R_{rect} \| = 1$.
- Η αντίστοιχη πίνακας για την δεξιά λήψη είναι ο $R_r = RR_{rect} \in \mathbb{R}^{3\times3}$
- Για κάθε σημείο της αριστερής λήψης $\mathbf{p}_L \in I^L$ υπολογίζουμε:

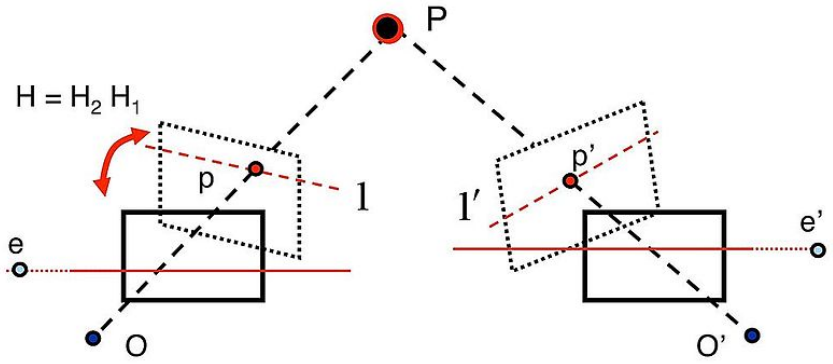
$$R_{rect}\mathbf{p}_L = [x', y', z'] \Rightarrow$$

$$\Rightarrow \mathbf{p}_L^{'} = [f \cdot \frac{x'}{z'}, f \cdot \frac{y'}{z'}, f]$$

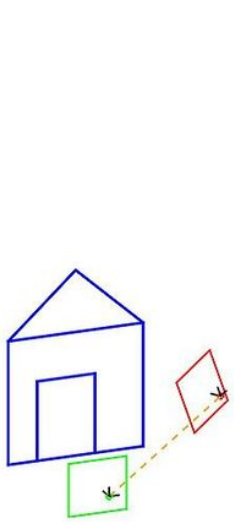
- Αντίστοιχα για την δεξιά λήψη, υπολογίζουμε τα σημεία $\mathbf{p}_R^{'}$ με χρήση του R_r αντί του R_{rect}

Όσα περιγράφηκαν παραπάνω σε επίπεδο σημείου $\mathbf{p}$ μπορούν να εφαρμοστούν συνολικά σε όλα τα σημεία του πετάσματος κι έτσι επιτυγχάνεται η δημιουργία ενός ζεύγους εικόνων που τηρεί τον στερεοσκοπικό περιορισμό. Πρέπει να παρατηρήσουμε ότι ενώ οι προ ευθυγράμμισης τιμές των σημείων είναι ακέραιοι αριθμοί (θέσεις pixel) οι επιστρεφόμενες τιμές είναι πραγματικοί αριθμοί. Επομένως, πρέπει να εφαρμοστεί ένα βήμα παρεμβολής τιμών σε ορθογώνιο χωρίο για τον υπολογισμό της τελικής ευθυγραμμισμένης εικόνας.

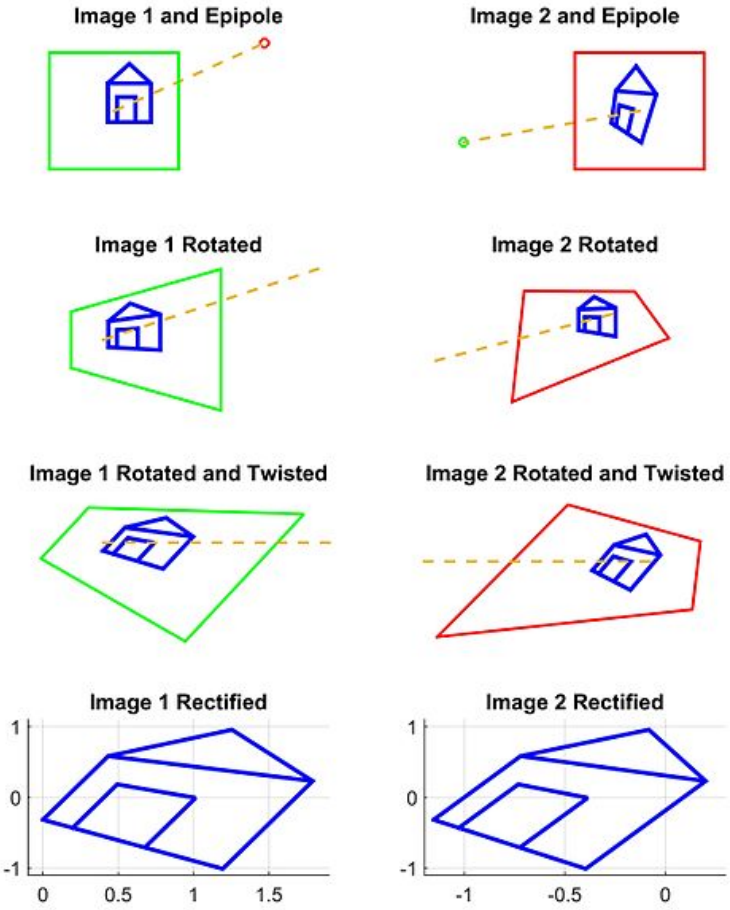
Η ευθυγράμμιση γενικεύει την αξία του προβλήματος της στερεοσκοπικής όρασης, αυξάνοντας το πεδίο εφαρμογής του από το στενό πλαίσιο των στερεοσκοπικών διατάξεων σε κάθε ζεύγος εικόνων, οποιασδήποτε γεωμετρίας.



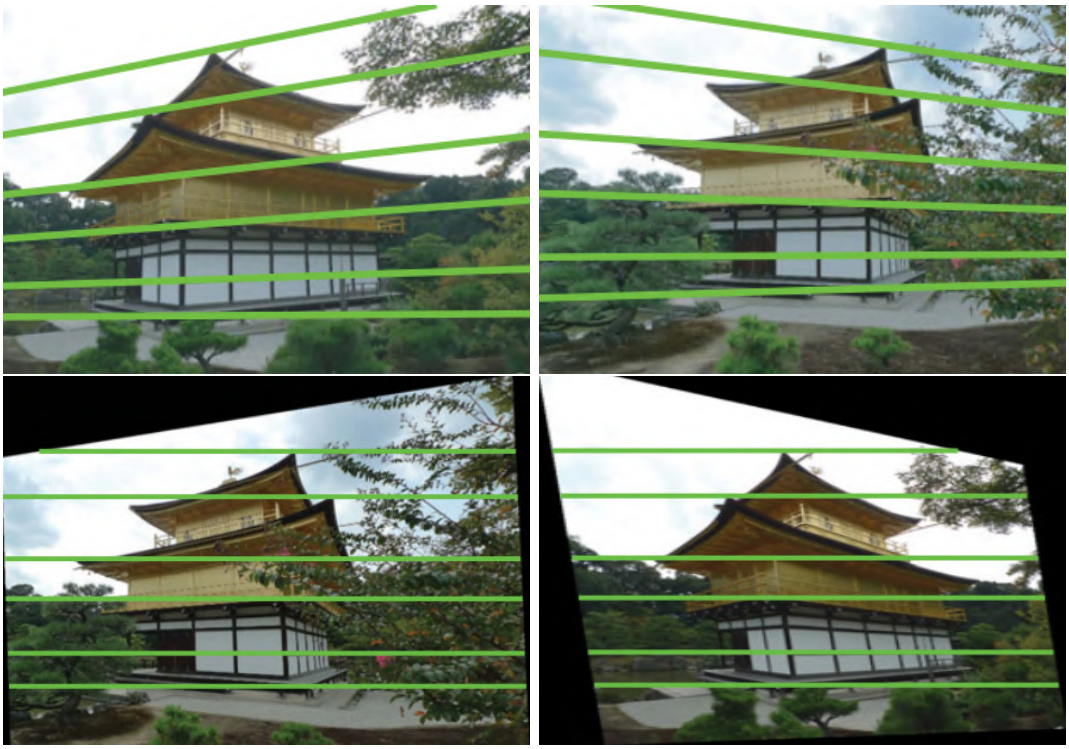
(Α΄) Γεωμετρική αποτύπωση της ευθυγράμμισης ενός ζεύγους εικόνων



(Β΄) Το αρχικό σχήμα που φωτογραφίζουν οι δύο κάμερες.



Σχημα Α΄.1: Αναλυτική παρουσίαση των μετασχηματισμών που ευθυγραμμίζουν το αρχικό ζεύγος εικόνων. Πηγή: [42]



Σχημα Α΄.2: Παράδειγμα ευθυγράμμισης από το βιβλίο Computer vision for visual effects. [33]. Επάνω: ζεύγος εικόνων τυχαίας γεωμετρίας. Κάτω: ζεύγος εικόνων που τηρεί τον στερεοσκοπικό περιορισμό, μετά από ευθυγράμμιση.

Α΄.6 Επαναληπτική εφαρμογή φίλτρου μέσου όρου κατά την άθροιση κόστους σε ορθογώνια περιοχή

Η επαναληπτική εφαρμογή του φίλτρου μέσου όρου φαίνεται παρακάτω:

$$C^0_{agg} = C_{init}$$
$$C^k_{agg}(d,x,y) = C^{k-1}_{agg}(x,y,d) * h(x,y)$$

που ισοδυναμεί με συνέλιξη του αρχικού πίνακα κόστους με το φίλτρο:

$$h_1 = \underbrace{h(x,y) \ldots * h(x,y)}_{\text{k times}}$$
$$C^n_{agg}(d,x,y) = C_{init}(x,y,d) * h_1$$

Παρατηρούμε ότι η επαναληπτική συνέλιξη του φίλτρου με τον εαυτό του, αυξάνει σε κάθε βήμα τις διαστάσεις του. Πιο συγκεκριμένα, έστω παράθυρο h με αρχικές διαστάσεις $m \times n$, η επαναληπτική εφαρμογή του k φορές στον αρχικό πίνακα κόστους ισοδυναμεί,

με συνέλιξή του με ένα φίλτρο διαστάσεων

$$m_1 = m + (m-1) * (k-1)$$
$$n_1 = n + (n-1) * (k-1)$$

Η γραμμική αύξηση των διαστάσεων του ισοδύναμου φίλτρου κατά την αναδρομική άθροιση κόστους γειτονιάς φέρει τις ίδιες αδυναμίες και πλεονεκτήματα με την εξ΄ αρχής επιλογή μεγάλης περιοχής υποστήριξης. Υπάρχει βέβαια ποιοτική διαφορά μεταξύ τους καθώς η επαναληπτική συνέλιξη με τον εαυτό του τείνει να το μετατρέψει σε γκαουσιανό φίλτρο (gaussian filter), όπως παρατηρούμε στο σχήμα Α΄.3 Για παράδειγμα εάν εφαρμόσουμε ένα απλό 3×3 φίλτρο 2 και 3 και 4 φορές αντίστοιχα θα είναι σαν να είχαμε κάνει αρχική συνέλιξη με τα φίλτρα:

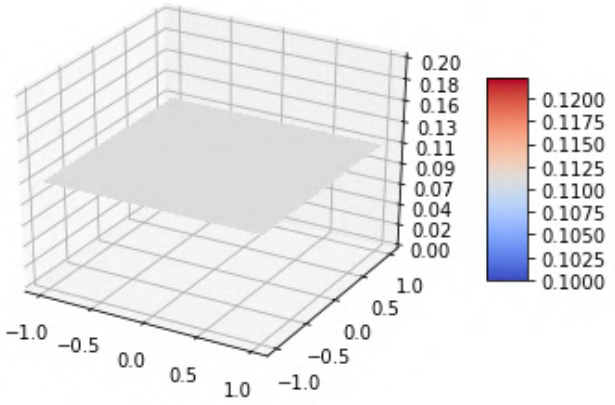
$$h = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$h_1 = \frac{1}{9} \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 4 & 6 & 4 & 2 \\ 3 & 6 & 9 & 6 & 3 \\ 2 & 4 & 6 & 4 & 2 \\ 1 & 2 & 3 & 2 & 1 \end{bmatrix}$$

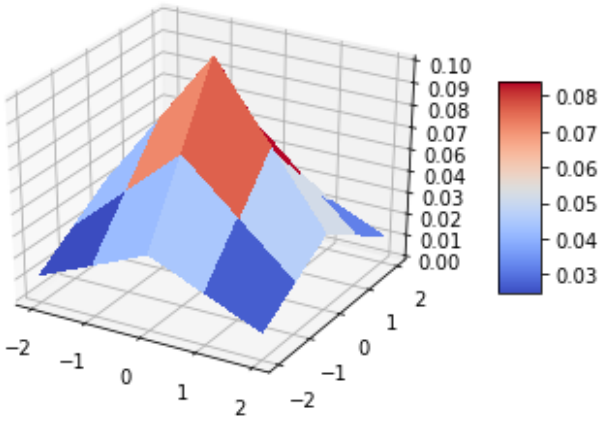
$$h_2 = \frac{1}{9^2} \begin{bmatrix} 1 & 3 & 6 & 7 & 6 & 3 & 1 \\ 3 & 9 & 18 & 21 & 18 & 9 & 3 \\ 6 & 18 & 36 & 42 & 36 & 18 & 6 \\ 7 & 21 & 42 & 49 & 42 & 21 & 7 \\ 6 & 18 & 36 & 42 & 36 & 18 & 6 \\ 3 & 9 & 18 & 21 & 18 & 9 & 3 \\ 1 & 3 & 6 & 7 & 6 & 3 & 1 \end{bmatrix}$$

$$h_3 = \frac{1}{9^3} \begin{bmatrix} 1 & 4 & 10 & 16 & 19 & 16 & 10 & 4 & 1 \\ 4 & 16 & 40 & 64 & 76 & 64 & 40 & 16 & 4 \\ 10 & 40 & 100 & 160 & 190 & 160 & 100 & 40 & 10 \\ 16 & 64 & 160 & 256 & 304 & 256 & 160 & 64 & 16 \\ 19 & 76 & 190 & 304 & 361 & 304 & 190 & 76 & 19 \\ 16 & 64 & 160 & 256 & 304 & 256 & 160 & 64 & 16 \\ 10 & 40 & 100 & 160 & 190 & 160 & 100 & 40 & 10 \\ 4 & 16 & 40 & 64 & 76 & 64 & 40 & 16 & 4 \\ 1 & 4 & 10 & 16 & 19 & 16 & 10 & 4 & 1 \end{bmatrix}$$

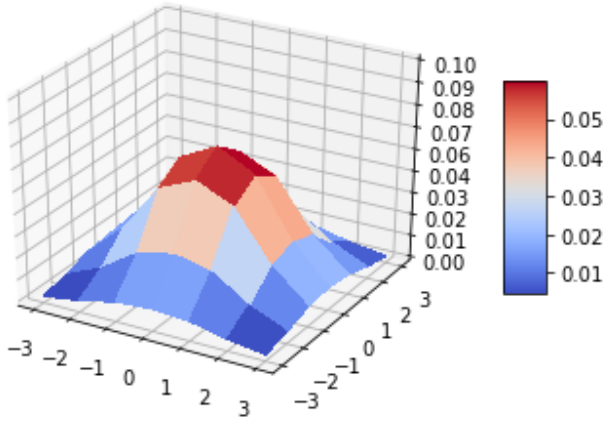
Η σταδιακή μετατροπή σε γκαουσιανό φίλτρο, πέρα από την αύξηση της περιοχής υποστήριξης, δίνει ιδιαίτερη βαρύτητα στα (pixels) κοντά στο (pixel) ενδιαφέροντος, αδιαφορώντας για το πιο απομακρυσμένα.



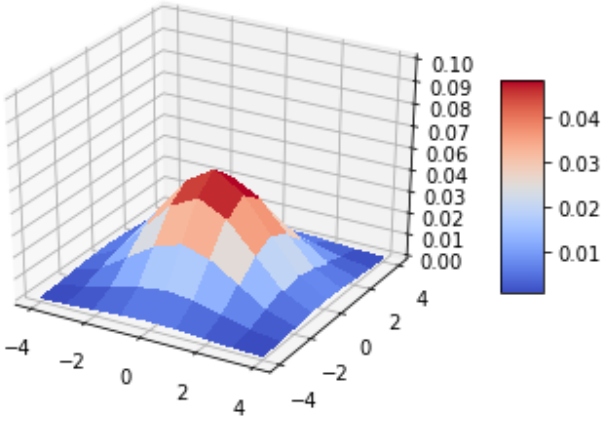
(Α΄) επανάληψη 1



(Β΄) επανάληψη 2



(Γ΄) επανάληψη 3



(Δ΄) επανάληψη 4

Σχημα Α΄.3: μεταβολή 3×3 φίλτρου μέσης τιμής κατά την συνέλιξη με τον εαυτό του

Παράρτημα Β′

Παράρτημα Κεφαλαίου 3

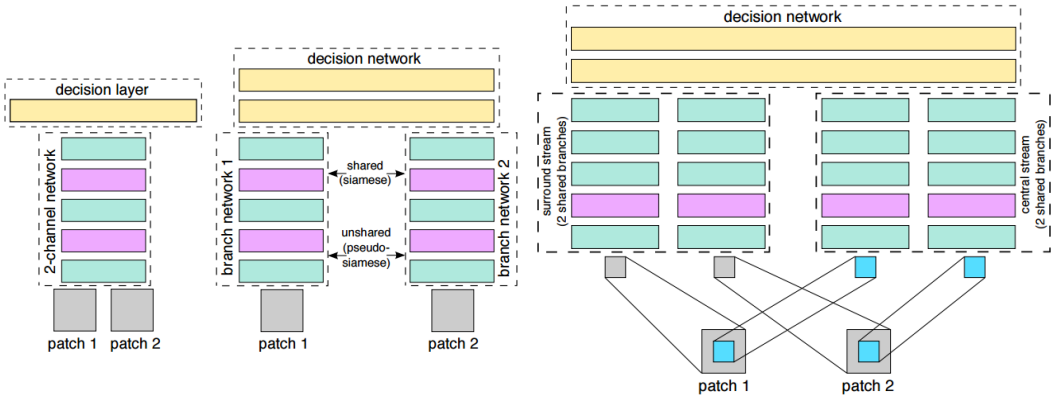
Β′.1 Γνωστές αρχιτεκτονικές νευρωνικών δικτύων για την αρχικοποίηση κόστους

Στις εικόνες Β′.1, Β′.2, Β′.3, Β′.4 και Β′.5 φαίνονται σχηματικά οι γνωστότερες αρχιτεκτονικές που έχουν προταθεί για την αρχικοποίηση του πίνακα κόστους.

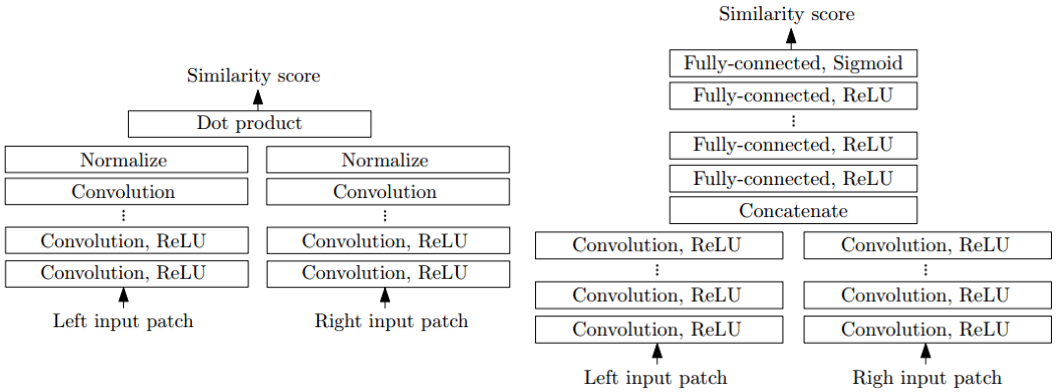
Β′.2 Επίπεδο κανονικοποίησης δέσμης

- Είσοδος: οι τρισδιάστατοι πίνακες $y_{\mathbf{conv2d}}$ όλης της δέσμης εκπαίδευσης.
- Έξοδος: οι τρισδιάστατοι πίνακες $y_{\mathbf{BN}}$ όλης της δέσμης εκπαίδευσης, κανονικοποιημένοι με βάση το κάθε ξεχωριστό feature map.
- Διαδικασία: Σε κάθε βήμα εκπαίδευσης, προωθούνται στο δίκτυο `batch_size` εγγραφές του σετ εκπαίδευσης. Κάθε εγγραφή, έχοντας περάσει από το επίπεδο της δισδιάστατης συνέλιξης αναπαρίσταται από έναν πίνακα y_{conv2d}. Ο τρισδιάστατος πίνακας y_{conv2d} απαρτίζεται από `f_maps` δισδιάστατους πίνακες $y_{conv2d,i}$. Δημιουργούμε `f_maps` ομάδες, μέλη της οποίας είναι κατ΄ αντιστοιχία, όλοι οι `batch_size` πίνακες $y_{conv2d,i}$. Τις ομάδες αυτές τις συμβολίζουμε ως X_j και κάθε ξεχωριστό στοιχείου τους ως X_j^i. Επί αυτών των ομάδων εφαρμόζεται η κανονικοποίηση. Συγκεκριμένα, η κανονικοποίηση δέσμης $X_{\mathbf{BN},j} = BN_{\gamma,\beta}(X_j)$ υλοποιείται με την ακόλουθη μεθοδολογία:

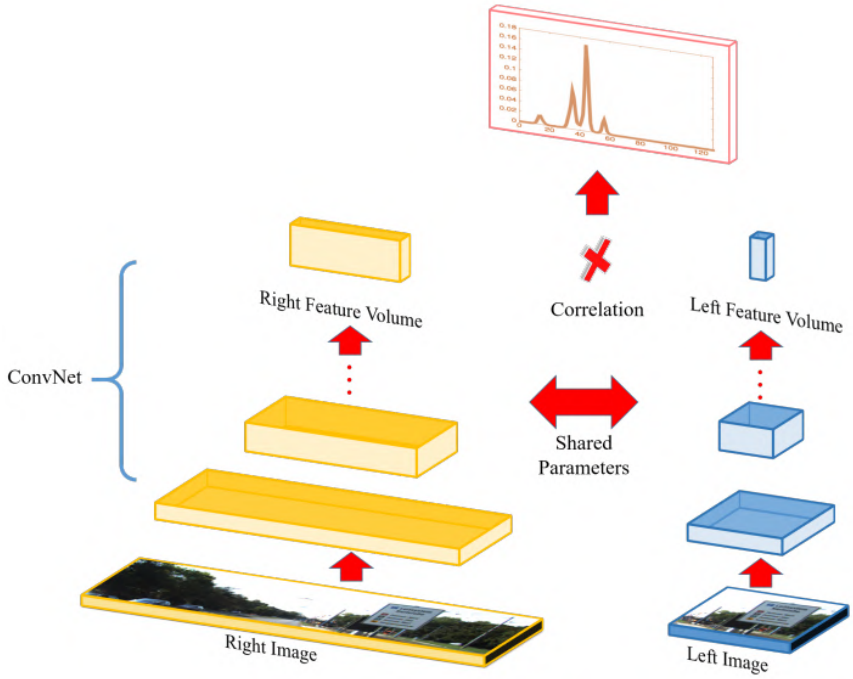
$$
\begin{aligned}
\mu_{X_j} &= {}^1\!/_m \sum_{i=1}^{m} X_j^i && \text{// μέσος όρος ομάδας} \\
\sigma^2_{X_j} &= {}^1\!/_m \sum_{i=1}^{m} (X_j^i - \mu_{X_j})^2 && \text{// διακύμανση ομάδας} \\
\hat{X}_j^i &= \frac{X_j^i - \mu_{X_j}}{\sqrt{\sigma^2_{X_j} + \epsilon}} && \text{// κανονικοποίηση} \\
\hat{X}_j^i &= \gamma \hat{X}_j^i + \beta && \text{// κλιμάκωση και μετατόπιση}
\end{aligned}
\tag{Β′.1}
$$



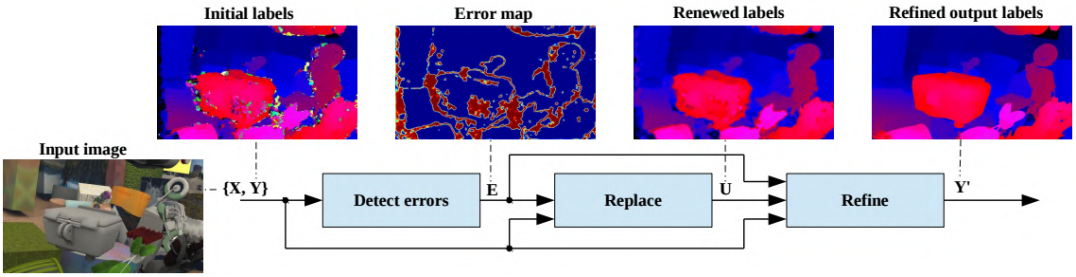
Σχημα Β΄.1: Αρχιτεκτονικές νευρωνικών δικτύων για την σύγκριση περιοχών εικόνας, όπως προτάθηκαν από τους Zagoruyko, Komodakis. [44]



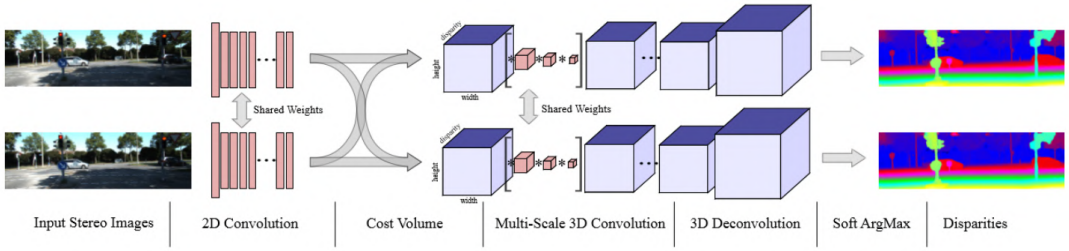
Σχημα Β΄.2: Αρχιτεκτονικές νευρωνικών δικτύων για την σύγκριση περιοχών εικόνας, όπως προτάθηκαν από τους Zbontar, Lecun. [45]



Σχημα Β΄.3: Αρχιτεκτονικές νευρωνικών δικτύων για την σύγκριση περιοχών εικόνας, όπως προτάθηκαν από τους Luo et al. [24]



Σχημα Β΄.4: Αρχιτεκτονικές νευρωνικών δικτύων για την σύγκριση περιοχών εικόνας, όπως προτάθηκαν από τους Gidaris et al. [8]



Σχημα Β΄.5: Αρχιτεκτονικές νευρωνικών δικτύων για την σύγκριση περιοχών εικόνας, όπως προτάθηκαν από τους Kendall et al. [8]

Οι μετασχηματισμένες τιμές $\hat{X}_j^i$ «αποστοιχίζονται» από τις ομάδες τους κι αναδιατάσσονται στη δομή που είχαν κατά την είσοδό τους στο επίπεδο. Έτσι συνολικά έχει επιτευχθεί η πράξη $y_{\mathbf{BN}} = BN(y_{\mathbf{conv2d}})$.

Β΄.3 Συλλογές στερεοσκοπικών δεδομένων με πληροφορία παράλλαξης

Οι συλλογές αυτές περιέχουν στερεοσκοπικά ζεύγη μαζί με την πληροφορία παράλλαξης, μετρημένη με κάποιο ειδικό εργαλείο όπως lidar ή laser. Η παράλλαξη συνήθως δεν είναι διαθέσιμη στο σύνολο των pixels του στερεοσκοπικού ζεύγους, αλλά σε ένα υποσύνολο αυτού. Το ποσοστό των διαθέσιμων τιμών επί του συνόλου της εικόνας ονομάζεται πυκνότητα στερεοσκοπικού ζεύγους (density of stereo pair):

$$\text{πυκνότητα} = \frac{\text{διαθέσιμες παραλλάξεις}}{\text{συνολικά pixels}} \times 100\%$$

Χαρακτηριστικό μέγεθος κάθε συλλογής είναι η μέση πυκνότητα των εγγραφών της

$$\text{μέση πυκνότητα} = \frac{1}{\text{σύνολο στερεοσκοπικών ζευγών}} \sum_i \text{πυκνότητα}_i$$

η οποία για προβλήματα πυκνής στερεοσκοπικής αντιστοίχησης (dense stereo matching) πρέπει να ξεπερνάει το 20%.

Οι πιο γνωστές στερεοσκοπικές συλλογές, που θα μας απασχολήσουν στην εργασία, είναι οι Middlebury stereo dataset, Kitti stereo benchmark και Synthetic stereo dataset.

Σχημα Β΄.6: Παραδείγματα εικόνων από την στερεοσκοπική συλλογή middlebury stereo dataset.

B.3.1 Middlebury stereo dataset

Το σετ δεδομένων Middlebury stereo dataset δημιουργήθηκε από το ομώνυμο πανεπιστήμιο των Ηνωμένων Πολιτειών το 2001 [**σχηαρστειν2002ταξονομψ**]. Έκτοτε η συλλογή έχει ανανεωθεί με νέες εκδόσεις τις χρονολογίες 2003 [36], 2005 [34], 2006 [13] και 20014 [38], περιλαμβάνοντας συνολικά περίπου 50 στερεοσκοπικά ζεύγη εικόνων. Οι διάφορες εκδόσεις εμφανίζουν μικρές διαφορές μεταξύ τους καθώς όλες χαρακτηρίζονται από τις παρακάτω ιδιότητες, όπως φαίνεται στην εικόνα Β΄.6:

- Οι φωτογραφίες έχουν ληφθεί σε εργαστηριακό περιβάλλον ελεγχόμενου φωτισμού
- Οι επιφάνειες των αντικειμένων που απαρτίζουν την σκηνή είναι λαμπερτιανές και συνήθως έχουν υφή
- οι αυξομειώσεις του βάθους είναι μικρές, άρα και το σύνολο των τιμών παράλλαξης. Προσεγγιστικά $d \in [10, 50]$ $pixels$
- Η πρόσοψη των αντικειμένων εμφανίζει πολύ μικρή κλίση με τον άξονα zz' της στερεοσκοπικής διάταξης

Η μέση πυκνότητα της συλλογής είναι περίπου 97%.

Οι δύο εικόνες των στερεοσκοπικών ζευγών έχουν ανάλυση 1988×2964 $pixels$, η απόσταση βάσης των δύο λήψεων είναι $B = 0.193m$, η εστίαση $f = 3997.68px$ κι η παράλλαξη κυμαίνεται στο διάστημα $[2, 265]px$.

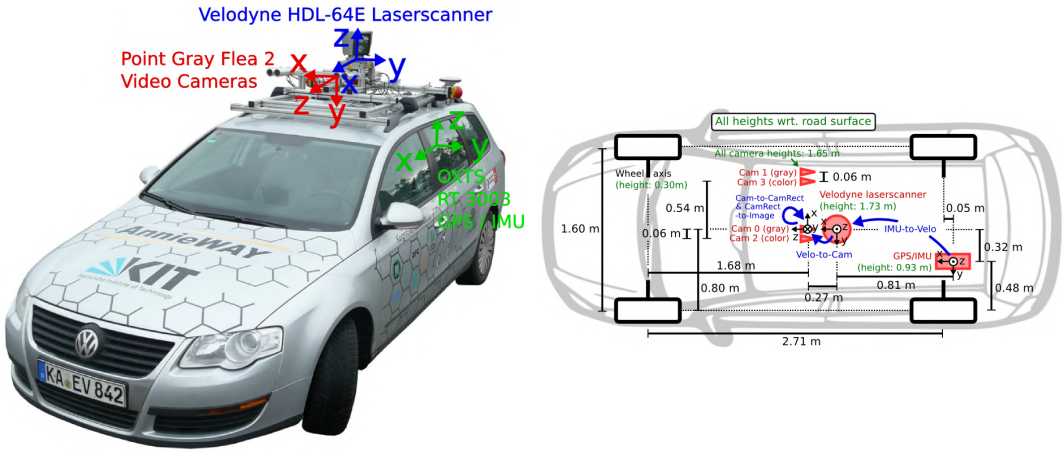
B.3.2 Kitti stereo benchmark

Η συλλογή δεδομένων Kitti stereo benchmark δημιουργήθηκε το 2012 [7] από το τεχνολογικό Ινστιτούτο της Καρλσρούης (Karlsruhe institute of Technology) σε συνεργασία με το τεχνολογικό Ινστιτούτο της Τογιότα στο Σικάγο (Toyota Technological Institute at Chicago). Ανανεώθηκε το 2015 [29] περιλαμβάνοντας πλέον συνολικά 400 στερεοσκοπικά ζεύγη εικόνων.

Οι εικόνες έχουν ληφθεί από την οροφή ενός αυτοκινήτου Β΄.7. Το πραγματικό βάθος των απεικονιζόμενων αντικειμένων έχει ληφθεί από στρεφόμενο σαρωτή λέιζερ (rotating laser scanner) τοποθετημένο πίσω από την αριστερή κάμερα.

Η γεωμετρία και στατιστική των εικόνων που περιλαμβάνει διαφέρει πλήρως από αυτή του Middlebury stereo dataset καθώς χαρακτηρίζεται από τις εξής ιδιότητες:

- Οι φωτογραφίες έχουν ληφθεί σε φυσικό περιβάλλον, στους δρόμους της Καρλσρούης μια ηλιόλουστη ημέρα. Περιλαμβάνουν κατά κύριο λόγο κινούμενα οχήματα, πεζοδρόμια, πεζούς και σπίτια εκατέρωθεν του δρόμου.



Σχημα Β΄.7: Το αυτοκίνητο που χρησιμοποιήθηκε για την συλλογή KITTI και η κάτοψή του.

- Οι επιφάνειες των αντικειμένων που απαρτίζουν την σκηνή δεν είναι στο σύνολό τους λαμπερτιανές. Ταυτόχρονα ο έντονος ήλιος δρα ως μια πολύ δυνατή πηγή φωτός με αποτέλεσμα να δημιουργούνται έντονα φαινόμενα κατοπτρικών ανακλάσεων.[1]
- Η έντονη πηγή φωτός δημιουργεί συχνά κορεσμό στον αισθητήρα *ccd* με αποτέλεσμα την αποτύπωση ειδώλων χωρίς υφή.
- οι αυξομειώσεις του βάθους, άρα και το σύνολο των τιμών παράλλαξης, είναι πολύ μεγάλο.
- Οι προσόψεις των αντικειμένων εμφανίζουν μεγάλες κλίση, σε σχέση με τον άξονα zz' της στερεοσκοπικής διάταξης δημιουργώντας εντονότερα φαινόμενα αυξομείωσης αποστάσεων και αποκρύψεων.

Η μέση πυκνότητα της συλλογής είναι περίπου 20%.

Οι δύο εικόνες των στερεοσκοπικών ζευγών έχουν ανάλυση $376 \times 1241 pixels$, η απόσταση βάσης των δύο λήψεων είναι $B = 0.54m$, η εστίαση $f = 707.09px$ κι η παράλλαξη κυμαίνεται στο διάστημα $[0, 230]px$.

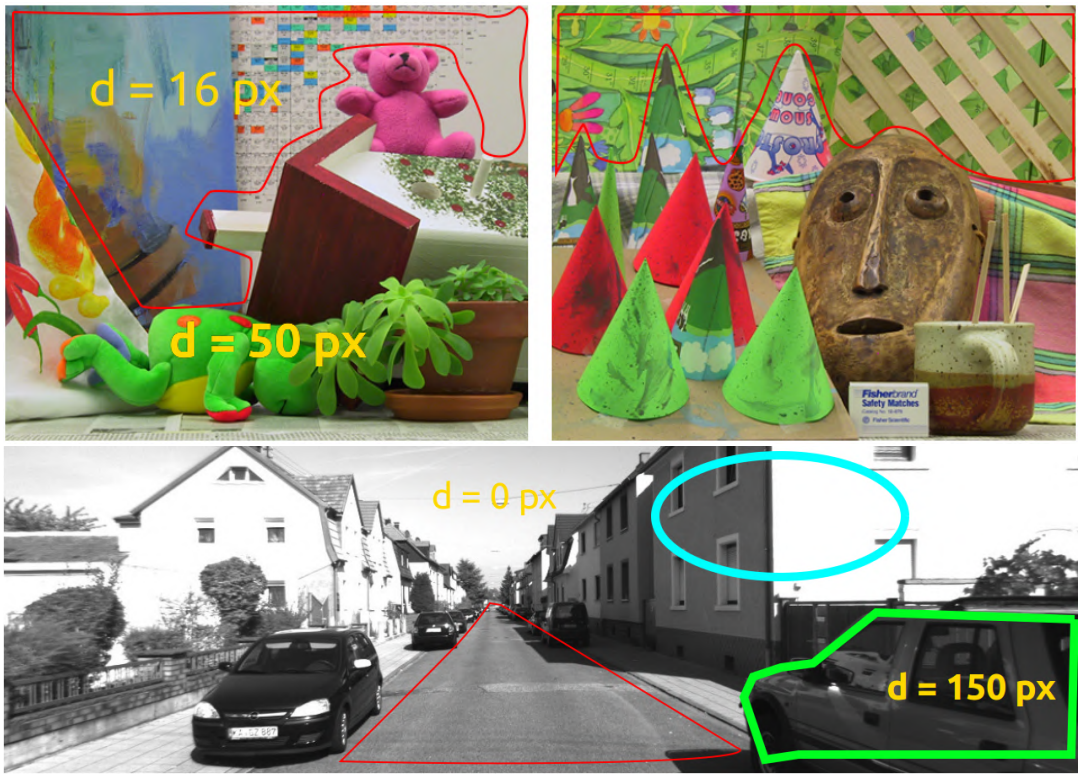
Στην εικόνα Β΄.9 αποτυπώνονται παραστατικά οι έντονες διαφορές ανάμεσα στις συλλογές KITTI και Middlebury.

Β.3.3 Synthetic stereo dataset

Το 2015, οι Mayer et al. [27] δημιούργησαν μια μεγάλη συνθετική συλλογή από στερεοσκοπικά ζεύγη. Η συλλογή αυτή προσεγγίζει σε ομοιότητα φυσικές εικόνες και ταυτόχρονα περιέχει πάρα πολλά στερεοσκοπικά ζεύγη (35.000) αποτελώντας βάση δεδομένων για την εκπαίδευση αλγορίθμων εκμάθησης μηχανής. Στην παρούσα εργασία δεν αξιοποιούμε την υπάρχουσα συλλογή. Παραδείγματα εικόνων της στερεοσκοπικής συλλογής φαίνονται στο σχήμα Β΄.10.

[1] Στην ανανέωση του 2015 τα τζάμια των αυτοκινήτων περιέχουν πληροφορία παράλλαξης και συμπεριλαμβάνονται στην αξιολόγηση, κάνοντας τη συλλογή πιο απαιτητική στον χειρισμό των κατοπτρικών επιφανειών

Σχημα Β΄.8: Παραδείγμα εικόνας από την στερεοσκοπική συλλογή kitti stereo benchmark.



Σχημα Β΄.9: **Κίτρινα χρώμα:** μέγιστη κι ελάχιστη τιμή παράλλαξης σε κάθε εικόνα. **Κόκκινο χωρίο:** διαφορά στην κλίση των εικονιζόμενων αντικειμένων. **Πράσινο χωρίο:** παράδειγμα κατοπτρικής ανάκλασης. **Γαλάζιο χωρίο:** παράδειγμα κορεσμού αισθητήρα λόγω έντονου φωτός, με αποτέλεσμα την απεικόνιση επιφάνειας χωρίς υφή

(α΄) Εικόνα από την συνθετική συλλογή

(β΄) Σύγκριση εικόνων δρόμου συνθετικής συλλογής (δεξιά) και συλλογής kitti stereo benchmark (αριστερά)

Σχημα Β΄.10: Συνθετική συλλογή στερεοσκοπικών εικόνων

Β΄.4 Αναλυτική περιγραφή της δημιουργίας του σετ εκπαίδευσης

Οι εγγραφές δημιουργούνται με την ακόλουθη μεθοδολογία. Ως εικόνα αναφοράς θεωρούμε την αριστερή εικόνα του στερεοσκοπικού ζεύγους. Σε κάθε σημείο $\mathbf{p} = (x, y) \in I^L$ όπου ισχύουν οι τρεις παρακάτω προϋποθέσεις:

- η τιμή της παράλλαξης είναι γνωστή
- $\texttt{max\_disparity} + \dfrac{n-1}{2} < x < \text{width} - \dfrac{n-1}{2}$,
- $\dfrac{n-1}{2} < y < \text{height} - \dfrac{n-1}{2}$

εφαρμόζουμε την εξής μεθοδολογία:

- Αποθηκεύουμε το τετράγωνο χωρίο διαστάσεων $n \times n$ pixels πέριξ του σημείου $\mathbf{p}$ ως $\mathcal{P}^L_{n\times n}(\mathbf{p})$.
- Αποθηκεύουμε το παραλληλόγραμμο χωρίο διαστάσεων $(\texttt{max\_disparity} + n) \times n$ ως $\mathcal{P}^R_{(\texttt{max\_disparity}+n)\times n}(\mathbf{q})$. Η θέση $\mathbf{q}$ υπολογίζεται ως

  $$\mathbf{q} = \texttt{int}(x - d, y)$$

  και το χωρίο εκτείνεται:
  - $\dfrac{n-1}{2}$ θέσεις προς τα πάνω, κάτω και δεξιά
  - $\texttt{max\_disparity} + \dfrac{n-1}{2}$ θέσεις προς τα αριστερά
- Αναθέτουμε ως label την τιμή της παράλλαξης του σημείου. Η ετικέτα label λαμβάνει τιμές στο διάστημα $[0, \texttt{max\_disparity}]$

Ουσιαστικά κάθε εγγραφή εκπαίδευση περιέχει το χωρίο αναφοράς $\mathcal{P}^L_{n\times n}(\mathbf{p})$ και όλα τα πιθανά χωρία που αυτό θα μπορούσε να αποτυπώνεται στην έτερη λήψη $\mathcal{P}^R_{(\texttt{max\_disparity}+n)\times n}(\mathbf{q})$.

Β΄.5 Επεξήγηση σχέσης 3.1

Ο λόγος που μας ενδιαφέρει η προσέγγιση της ελάχιστης τιμής, κι όχι η εύρεσή της, είναι ότι οι παράμετροι που οδηγούν στην ελάχιστη τιμή αυτή καθ΄ αυτή, είναι ευάλωτοι σε υπερπροσαρμογή (overfitting). Το ολικό ελάχιστο της συνάρτησης είναι έντονα επηρεασμένο από την τυχαία μορφή των συγκεκριμένων παραδειγμάτων X, ενώ η ευρύτερη περιοχή πέριξ αυτού αναπαριστά καλύτερα το γενικό μοτίβο που ακολουθούν τα παραδείγματα εκπαίδευσης και που τελικά θέλουμε το δίκτυό μας να «μάθει». Ο παραπάνω σχολιασμός θα είχε μεγαλύτερη αξία εάν εκπαιδεύαμε το δίκτυο πάνω σε **όλα** τα δεδομένα του σετ εκπαίδευσης μέσω του αλγορίθμου «απότομης καθόδου» (gradient descent). Η επιλογή μας να χρησιμοποιήσουμε την εναλλακτική μορφή του αλγορίθμου «στοχαστικής απότομης καθόδου μικρής δέσμης» (mini-batch stochastic gradient descent) μας απαλλάσσει από τον παραπάνω προβληματισμό, καθώς το δίκτυο αναπροσαρμόζει τις παραμέτρους του επιλύοντας διαρκώς διαφορετικά προβλήματα ελαχίστου, το καθένα βασισμένο σε ένα διαφορετικό υποσύνολο παραδειγμάτων της συλλογής εκπαίδευσης. Επομένως, είναι σχεδόν βέβαιο ότι με μια μικρή αναπροσαρμογή παραμέτρων στην κατεύθυνση του

εκάστοτε ελαχίστου, το δίκτυο δεν θα φτάσει ποτέ ούτως ή άλλως στο ολικό ελάχιστο, παρά μόνο θα το προσεγγίζει, ικανοποιώντας την συνθήκη 3.1.

Β΄.6 Περιγραφή μεθόδων τις οποίες συνδυάζει ο αλγόριθμος ADAM

Σε κάθε βήμα, οι παράμετροι W ανανεώνονται ως:

$$W_{i+1} = W_i - \texttt{learning\_rate} \times \bigtriangledown_{W_i} f(X_i, W_i)$$

Αυτή η απλή εκδοχή του αλγορίθμου «στοχαστικής απότομης καθόδου μικρής δέσμης» (mini-batch stochastic gradient descent) δυσκολεύεται σε περιοχές που η κλίση αποκλίνει από διάσταση σε διάσταση, δηλαδή το διάνυσμα $\bigtriangledown_{W_i} f(X_i, W_i)$ εμφανίζει μεγάλη διακύμανση. Τέτοια «φυσιολογία» συνηθίζουν να εμφανίζουν οι περιοχές κοντά στα τοπικά ελάχιστα, όπου ο αλγόριθμος «ταλαντώνεται» ανάμεσα στις δύο «πλαγιές» πλησιάζοντας πολύ διστακτικά το τοπικό ελάχιστο. Για τον λόγο αυτό εισάγεται στην ανανέωση των παραμέτρων ο όρος μομεντυμ v. Μπορούμε να παρομοιάσουμε αυτή την εκδοχή του αλγορίθμου με μία μπάλα που αφήνουμε να κατρακυλήσει στην πλαγιά ενός λόφου. Η πορεία που ακολουθεί δεν εξαρτάται μόνο από την κλίση του εδάφους σε κάθε σημείο, αλλά και από την ταχύτητα που έχει ήδη αναπτύξει. Έτσι πραγματοποιείται ταχύτερη σύγκλιση στο τοπικό ελάχιστο. Η συνηθισμένη τιμή της παραμέτρου `momentum_rate` κινείται κοντά στο 0.9.

$$v_{i+1} = \texttt{momentum\_rate} \times v_i + \texttt{learning\_rate} \times \bigtriangledown_{W_i} f(X_i, W_i)$$
$$W_{i+1} = W_i - v_{i+1}$$

Οι παραπάνω μέθοδοι αναπροσαρμόζουν τις τιμές των παραμέτρων κατά σταθερό `learning_rate`. Η εύρεση της κατάλληλης τιμής `learning_rate` απαιτεί χρονοβόρο πειραματισμό. Προς επίλυση αυτής της προβληματικής, έχουν αναπτυχθεί αλγόριθμοι που αυξομειώνουν εσωτερικά το `learning_rate`. Πιο συγκεκριμένα, ο αλγόριθμος RMSprop δημιουργεί έναν κινούμενο μέσο όρο που ρυθμίζει το `learning_rate` διαφορετικά για κάθε διάσταση σύμφωνα με τον ακόλουθο κανόνα: Συμβολίζουμε ως **record** το διάνυσμα ίδιων διαστάσεων με το διάνυσμα παραμέτρων W που κρατάει το «ιστορικό» των μεταβολών:

$$\textbf{record} = \texttt{decay\_rate} \cdot \textbf{record} + (1 - \texttt{decay\_rate}) \nabla_{W_i} f^2(X_i, W_i) \tag{Β΄.2}$$

$$W_{i+1} = \frac{-\texttt{learning\_rate} \cdot \nabla_{W_i} f(X_i, W_i)}{\sqrt{\textbf{record} + \epsilon}}$$

Bibliography

[1] Padmanabhan Anandan. "A computational framework and an algorithm for the measurement of visual motion". In: *International Journal of Computer Vision* 2.3 (1989), pp. 283–310.

[2] Stan Birchfield and Carlo Tomasi. "A pixel dissimilarity measure that is insensitive to image sampling". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.4 (1998), pp. 401–406.

[3] Yuri Boykov, Olga Veksler, and Ramin Zabih. "Fast approximate energy minimization via graph cuts". In: *IEEE Transactions on pattern analysis and machine intelligence* 23.11 (2001), pp. 1222–1239.

[4] Christian Unger for the lesson "Computer Aided Medical Procedures" at TUM. *Stereo Matching*. 2009. URL: http://campar.in.tum.de/twiki/pub/Chair/TeachingWs09Cv2/3D_CV2_WS_2009_Stereo.pdf.

[5] Olivier Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. 1993.

[6] Pedro F Felzenszwalb and Daniel P Huttenlocher. "Efficient belief propagation for early vision". In: *International journal of computer vision* 70.1 (2006), pp. 41–54.

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite". In: *Computer Vision and Pattern Recognition (CVPR),(Providence, USA)* (2012).

[8] Spyros Gidaris and Nikos Komodakis. "Detect, Replace, Refine: Deep Structured Prediction For Pixel Wise Labeling". In: *arXiv preprint arXiv:1612.04770* (2016).

[9] Ralf Haeusler, Rahul Nair, and Daniel Kondermann. "Ensemble Learning for Confidence Measures in Stereo Vision". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013.

[10] Marsha J Hannah. *Computer matching of areas in stereo images*. Tech. rep. STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1974.

[11] Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *The IEEE International Conference on Computer Vision (ICCV)*. 2015.

[12] Heiko Hirschmuller. "Stereo processing by semiglobal matching and mutual information". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30.2 (2008), pp. 328–341.

[13] Heiko Hirschmuller and Daniel Scharstein. "Evaluation of cost functions for stereo matching". In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE. 2007, pp. 1–8.

[14] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International Conference on Machine Learning*. 2015, pp. 448–456.

[15] Takeo Kanade and Masatoshi Okutomi. "A stereo matching algorithm with an adaptive window: Theory and experiment". In: *IEEE transactions on pattern analysis and machine intelligence* 16.9 (1994), pp. 920–932.

[16] Takeo Kanade et al. "Development of a video-rate stereo machine". In: *Journal of the Robotics Society of Japan* 15.2 (1997), pp. 261–267.

[17] Sing Bing Kang, Richard Szeliski, and Jinxiang Chai. "Handling occlusions in dense multi-view stereo". In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. Vol. 1. IEEE. 2001, pp. I–I.

[18] Alex Kendall et al. "End-to-End Learning of Geometry and Context for Deep Stereo Regression". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[19] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. 2014.

[20] Vladimir Kolmogorov and Ramin Zabih. "Computing visual correspondence with occlusions using graph cuts". In: *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. Vol. 2. IEEE. 2001, pp. 508–515.

[21] Dan Kong and Hai Tao. "A method for learning matching errors for stereo computation." In: *BMVC*. Vol. 1. 2004, p. 2.

[22] Yunpeng Li and Daniel P Huttenlocher. "Learning for stereo vision using the structured support vector machine". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, pp. 1–8.

[23] Jiangbo Lu, Gauthier Lafruit, and Francky Catthoor. "Anisotropic local high-confidence voting for accurate stereo correspondence". In: *Electronic Imaging 2008*. International Society for Optics and Photonics. 2008, 68120J–68120J.

[24] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun. "Efficient Deep Learning for Stereo Matching". In: *CVPR*. IEEE Computer Society, 2016, pp. 5695–5703.

[25] Yi Ma et al. *An invitation to 3-d vision: from images to geometric models*. Vol. 26. Springer Science & Business Media, 2012.

[26] Larry Matthies, Takeo Kanade, and Richard Szeliski. "Kalman filter-based algorithms for estimating depth from image sequences". In: *International Journal of Computer Vision* 3.3 (1989), pp. 209–238.

[27] Nikolaus Mayer et al. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4040–4048.

[28] Xing Mei et al. "On building an accurate stereo matching system on graphics hardware". In: 2011.

[29] Moritz Menze and Andreas Geiger. "Object Scene Flow for Autonomous Vehicles". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3061–3070.

[30] Zachary Moratto. *Semi-Global Matching*. September 4, 2013. URL: http://lunokhod.org/?p=1356.

[31] Min-Gyu Park and Kuk-Jin Yoon. "Leveraging stereo matching with learning-based confidence measures". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 101–109.

[32] Marc Pollefeys. "Visual 3D Modeling from Images." In: *VMV*. 2004, p. 3.

[33] R.J. Radke. *Computer Vision for Visual Effects*. Computer Vision for Visual Effects. Cambridge University Press, 2013. ISBN: 9780521766876. URL: https://books.google.gr/books?id=MgIVyAWf9VUC.

[34] Daniel Scharstein and Chris Pal. "Learning conditional random fields for stereo". In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE. 2007, pp. 1–8.

[35] Daniel Scharstein and Richard Szeliski. "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms". In: *International journal of computer vision* 47.1-3 (2002), pp. 7–42.

[36] Daniel Scharstein and Richard Szeliski. "High-accuracy stereo depth maps using structured light". In: *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. Vol. 1. IEEE. 2003, pp. I–I.

[37] Daniel Scharstein and Richard Szeliski. "Stereo matching with nonlinear diffusion". In: *International journal of computer vision* 28.2 (1998), pp. 155–174.

[38] Daniel Scharstein et al. "High-resolution stereo datasets with subpixel-accurate ground truth". In: *German Conference on Pattern Recognition*. Springer, Cham. 2014, pp. 31–42.

[39] Aristotle Spyropoulos, Nikos Komodakis, and Philippos Mordohai. "Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[40] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[41] Emanuele Trucco and Alessandro Verri. *Introductory techniques for 3-D computer vision*. 1998.

[42] Wikipedia, the free encyclopedia. *Image Rectification*. URL: https://en.wikipedia.org/wiki/File:2DRectificationBAG.jpg.

[43] Ramin Zabih and John Woodfill. "Non-parametric local transforms for computing visual correspondence". In: *European conference on computer vision*. Springer. 1994, pp. 151–158.

[44] Sergey Zagoruyko and Nikos Komodakis. "Learning to compare image patches via convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4353–4361.

[45] Jure Zbontar and Yann LeCun. "Stereo matching by training a convolutional neural network to compare image patches". In: *Journal of Machine Learning Research* 17.1-32 (2016), p. 2.

[46] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit. "Cross-based local stereo matching using orthogonal integral images". In: *IEEE transactions on circuits and systems for video technology* 19.7 (2009), pp. 1073–1079.

[47] Li Zhang and Steven M Seitz. "Estimating optimal parameters for MRF stereo from a single image pair". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.2 (2007), pp. 331–342.