Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Γεωμετρική Άθροιση Διανυσμάτων Περιγραφής για Ανάκτηση και Κατηγοριοποίηση Εικόνων

Διπλωματικη Εργασια

του

ΑΘΑΝΑΣΙΟΥ Γ. ΠΑΠΑΔΟΠΟΥΛΟΥ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εργαστηριο Επεξεργασιας Εικονας Βιντεο και Πολυμεσων
Αθήνα, Μάιος 2014

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Επεξεργασίας Εικόνας Βίντεο και Πολυμέσων

Γεωμετρική Άθροιση Διανυσμάτων Περιγραφής για Ανάκτηση και Κατηγοριοποίηση Εικόνων

Διπλωματικη Εργασια

του

ΑΘΑΝΑΣΙΟΥ Γ. ΠΑΠΑΔΟΠΟΥΛΟΥ

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Μαίου 2014.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Στέφανος Κόλλιας	Ανδρέας-Γεώργιος Σταφυλοπάτης	Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Λέκτορας Ε.Μ.Π.

Αθήνα, Μάιος 2014

(Υπογραφή)

.....

Αθανασιος Γ. Παπαδοπουλος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Επεξεργασίας Εικόνας Βίντεο και Πολυμέσων

Ευχαριστίες

Για την ολοκλήρωση αυτής της διπλωματικής εργασίας θα ήθελα αρχικά να ευχαριστήσω τον υπεύθυνο καθηγητή κ. Στέφανο Κόλλια για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου τη συγκεκριμένη εργασία. Θα ήθελα επίσης να ευχαριστήσω ιδιαιτέρως το Δρ. Γιάννη Αβρίθη για την πολύτιμη καθοδήγησή του καθ όλη τη διάρκεια της ερευνητικής μου προσπάθειας. Ευχαριστώ επίσης το Δρ. Κώστα Ραπαντζίκο για την σημαντική του βοήθεια στο ξεκίνημα της εργασίας μου, καθώς και όλα τα μέλη του Εργαστηρίου Επεξεργασίας Εικόνας, Βίντεο και Πολυμέσων του Εθνικού Μετσοβίου Πολυτεχνείου. Ένα ξεχωριστό ευχαριστώ οφείλω στην οικογένειά μου για τη συνεχή και αμέριστη στήριξη που μου προσέφερε. Τέλος, θα ήθελα να ευχαριστήσω το φίλο μου Παναγιώτη Γιαννούλη για την παραχώρηση των απαραίτητων υπολογιστικών πόρων προκειμένου να ολοκληρωθεί η πειραματική διαδικασία.

Περίληψη

Σε αυτή τη διπλωματική εργασία προτείνουμε μία νέα μέθοδο διανυσματικής αναπαράστασης εικόνων η οποία βασίζεται στην άθροιση διανυσμάτων περιγραφής, αξιοποιώντας παράλληλα τη χωρική τους πληροφορία. Η μέθοδος που αναπτύξαμε ονομάζεται Spatial Pyramid with Vectors of Locally Aggregated Descriptors (SP–VLAD) και σχεδιάστηκε με γνώμονα την εφαρμογή της στο πρόβλημα της ανάκτησης εικόνων με τρόπο που να επιτυγχάνεται υψηλή ακρίβεια και αποδοτικότητα, με χαμηλές απαιτήσεις μνήμης. Η μέθοδος SP–VLAD στηρίζεται στις ιδέες των μεθόδων Spatial Pyramid Matching (SPM) και Vector of Locally Aggregated Descriptors (VLAD). Συγκεκριμένα, συνδυάζει τη δομή της χωρικής πυραμίδας με τα διανύσματα VLAD. Η αξιολόγηση της μεθόδου SP–VLAD οδήγησε σε αισθητά υψηλότερες επιδόσεις σε σχέση με τις μεθόδους SPM και VLAD. Γι΄ αυτό το λόγο εφαρμόστηκε και στο πρόβλημα της κατηγοριοποίησης εικόνων, όπου ξεπέρασε το βαθμό κατηγορίοποιησης των μεθόδων SPM και VLAD στις βάσεις εικόνων στις οποίες δοκιμάστηκε. Οι εξαιρετικές επιδόσεις της μεθόδου SP–VLAD επετεύχθησαν με πολύ χαμηλές απαιτήσεις μνήμης μετά την μείωση της διάστασης των τελικών διανυσμάτων περιγραφής στις 128 και 64 διαστάσεις μέσω της μεθόδου PCA. Στο πλαίσιο της ερευνητικής μας δραστηριότητας δημιουργήθηκε η νέα βάση εικόνων Flowers 15 με σκοπό την εφαρμογή της μεθόδου SP–VLAD σε εικόνες ανθοφόρων φυτών.

Λέξεις Κλειδιά

ανάκτηση εικόνων, κατηγοριοποίηση εικόνων, ανιχνευτής, διάνυσμα περιγραφής, οπτικό λεξικό, διάνυσμα περιγραφής, χωρική πυραμίδα, άθροιση διανυσμάτων

Abstract

We propose a new method for the vector representation of an image which is based on aggregation of image descriptors while exploiting their spatial information. Our method, Spatial Pyramid with Vectors of Locally Aggregated Descriptors (SP–VLAD), was designed for the problem of image retrieval in order to achieve high accuracy and efficiency, with low memory requirements. SP–VLAD is based on the ideas of two other methods, Spatial Pyramid Matching (SPM) and Vector of Locally Aggregated Descriptors (VLAD). Specifically, it combines the idea of spatial pyramid with the VLAD descriptor vectors. The SP–VLAD method achieves high accuracy and significantly outperforms SPM and VLAD methods. The promising results led us to apply our method to the problem of image classification as well; exceeding the classification rate of SPM and VLAD methods on all databases which were used. The excellent results of our method were achieved with low memory usage after the dimension reduction of the descriptor vectors with the PCA method which resulted to vectors of 64 or 128 dimensions. We also created the dataset Flowers 15 for the needs of our research in order to be able to test the SP–VLAD method upon images of flowering plants.

Keywords

image retrieval, image classification, detector, descriptor, visual vocabulary, spatial pyramid, aggregation of vectors

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Στο παρόν εισαγωγικό κεφάλαιο παρουσιάζεται ο στόχος, το κίνητρο και η συνεισφορά της ερευνητικής δραστηριότητας που συντελέστηκε στο πλαίσιο αυτής της διπλωματικής εργασίας, και παρατίθεται το απαραίτητο θεωρητικό υπόβαθρο γύρω από το πρόβλημα της ανάκτησης εικόνων.

1.1 Στόχος και Κίνητρο

Στόχος. Στόχος της παρούσας διπλωματικής εργασίας είναι να προτείνουμε μία νέα μέθοδο διανυσματικής αναπαράστασης εικόνων την οποία σχεδιάσαμε και υλοποιήσαμε με γνώμονα την εφαρμογή της στο πρόβλημα της ανάκτησης εικόνων, επιδιώκοντας να συνδυάζει υψηλή ακρίβεια και αποδοτικότητα, με χαμηλές απαιτήσεις μνήμης. Η μέθοδος που αναπτύξαμε ονομάζεται Spatial Pyramid with Vectors of Locally Aggregated Descriptors (SP–VLAD) και αξιοποιεί τη χωρική πληροφορία που εμπεριέχεται στις εικόνες ενώ παράλληλα ωφελείται από τη χρησιμοποίηση τεχνικών άθροισης διανυσμάτων περιγραφής.

Κίνητρο. Το κίνητρο για την ανάπτυξη μίας νέας μεθόδου διανυσματικής αναπαράστασης εικόνων η οποία συνδυάζει υψηλή διακριτική ικανότητα με μειωμένες απαιτήσεις σε υπολογιστικούς πόρους και μνήμη, βασίζεται στην διαρκώς αυξανόμενη δημιουργία και χρήση παντός είδους εικόνων. Στις μέρες μας είναι εξαιρετικά διαδεδομένη η χρήση ψηφιακών καμερών και αυτό συνεπάγεται τη δημιουργία αντίστοιχα μεγάλου αριθμού συλλογών από εικόνες. Ενδεικτικά αναφέρουμε πως το Facebook, που αποτελεί τη δημοφιλέστερη ιστοσελίδα κοινωνικής δικτύωσης, δέχεται 7 petabytes από καινούριες εικόνες κάθε μήνα και 300 εκατομμύρια εικόνες κάθε ημέρα [12]. Είναι δύσκολο λοιπόν κάποιος να διανοηθεί τον αριθμό και την ποικιλία των εικόνων οι οποίες είναι διαθέσιμες σε ολόκληρο το διαδίκτυο, το οποίο μάλιστα αποτελεί μονάχα μία από τις πολυάριθμες πηγές εικόνων. Συνεπώς, είναι δεδομένη η ανάγκη οργάνωσης των εικόνων προκειμένου να είναι εφικτή η αναζήτηση και ανάκτησή τους μέσα σε έναν τόσο μεγάλο όγκο πληροφορίας. Αυτό το ρόλο καλούνται να διαδραματίσουν τα συστήματα ανάκτησης εικόνων και για αυτόν ακριβώς το λόγο κρίνεται επιβεβλημένη η συνεχής βελτίωσή τους μέσω μεθόδων όπως η SP–VLAD.

1.2 Ανάκτηση Εικόνων

Η ανάκτηση εικόνων αποτελεί ενεργό επιστημονικό πεδίο από τα τέλη της δεκαετίας του 1970 και στοχεύει στην αποτελεσματική αναζήτηση ψηφιακών εικόνων βάσει ενός κριτηρίου ομοιότητας. Έστω ότι διαθέτουμε λοιπόν μία βάση εικόνων. Δοθείσης μίας εικόνας−ερώτημα (query image) η οποία περιλαμβάνει ένα συγκεκριμένο οπτικό περιεχόμενο, ο στόχος είναι να παρουσιαστεί στο χρήστη ένα υποσύνολο της βάσης με όμοιες οπτικά εικόνες ως προς την εικόνα−ερώτημα. Για παράδειγμα, εάν η εικόνα−ερώτημα αναπαριστά κάποιο τοπίο, επιδιώκουμε να ανακτηθούν εικόνες με το ίδιο τοπίο. Στην εξέλιξη του πεδίου αυτού συνέβαλλαν κυρίως δύο επιστημονικές κοινότητες, αυτή της διαχείρισης βάσεων δεδομένων και αυτή της όρασης υπολογιστών. Κάθε μία προσέγγισε το πρόβλημα της ανάκτησης εικόνων από διαφορετική οπτική γωνία δημιουργώντας δύο μεγάλες ερευνητικές περιοχές [35]:

- **Ανάκτηση εικόνων βασισμένη στο κείμενο (text-based):** Οι εικόνες πρώτα υποσημειώνονται με μεταδεδομένα κειμένου (π.χ. λέξεις κλειδιά) και μετά χρησιμοποιούνται τεχνικές συστημάτων βάσεων δεδομένων για την ανάκτησή τους.
- **Ανάκτηση εικόνων βασισμένη στο οπτικό περιεχόμενο (content-based):** Οι εικόνες αναπαρίστανται και συγκρίνονται με βάση το οπτικό τους περιεχόμενο με χρήση μεθόδων της όρασης υπολογιστών.

Η ανάκτηση εικόνων με βάση το κείμενο ήταν η πρώτη απόπειρα στο συγκεκριμένο χώρο τη δεκαετία του 1970 και βασίστηκε στην περιγραφή των εικόνων με υποσημειώσεις κειμένου. Σχετικό παράδειγμα αποτελεί ο τρόπος με τον οποίο εμφανίζει εικόνες η μηχανή αναζήτησης Google όταν δοθούν ως είσοδος κάποιες λέξεις κλειδιά. Τα αποτελέσματα είναι αρκετά ακριβή τις περισσότερες φορές, ωστόσο το κείμενο που περιγράφει τις εικόνες είναι συνήθως προϊόν ανθρώπινης εργασίας και αυτό έχει ως συνέπεια να παρουσιάζονται δύο βασικές δυσκολίες στη χρήση της συγκεκριμένης μεθόδου. Η πρώτη αφορά το μέγεθος των βάσεων δεδομένων και η δεύτερη την αξιοπιστία των μεταδεδομένων. Βάσεις δεδομένων μεγάλου μεγέθους απαιτούν την αφιέρωση αντίστοιχα μεγάλου χρόνου για την υποσημείωση των εικόνων. Δεδομένου μάλιστα πως οι ανάγκες για συλλογές εικόνων πολύ μεγάλης κλίμακας (δισεκατομμύρια εικόνες) έχουν αυξηθεί, αλλά και το γεγονός πως νέες εικόνες είθισται να προστίθενται σε υπάρχουσες βάσεις, η όλη διαδικασία της υποσημείωσης των εικόνων κρίνεται απαγορευτικά απαιτητική σε πολλές περιπτώσεις. Η δεύτερη και μάλιστα πιο σημαντική δυσκολία αναφορικά με την αξιοπιστία των μεταδεδομένων οφείλεται στην υποκειμενικότητα της ανθρώπινης αντίληψης. Την ίδια εικόνα, διαφορετικοί άνθρωποι μπορούν να την αντιληφθούν διαφορετικά και κατά συνέπεια να την υποσημειώσουν με διαφορετικό τρόπο. Ο ανθρώπινος παράγοντας λοιπόν αναπόφευκτα εισάγει το στοιχείο της υποκειμενικότητας το οποίο μπορεί να οδηγήσει σε μη αναστρέψιμα λάθη.

Στην επίλυση της προαναφερθείσας δεύτερης δυσκολίας επικεντρώνεται κατά κύριο λόγο η ανάκτηση εικόνων βασισμένη στο περιεχόμενο, και αυτός άλλωστε είναι ο κύριος λόγος που έκανε την εμφάνισή της στα τέλη της δεκαετίας του 1990. Οι εικόνες πλέον δεν αναπαρίστανται από λέξεις κλειδιά αλλά από τα οπτικά τους χαρακτηριστικά, όπως είναι το χρώμα, το σχήμα

και η υφή των αντικειμένων που περιλαμβάνονται σε αυτές. Η ποικιλία τόσο των οπτικών χαρακτηριστικών όσο και των τρόπων περιγραφής των εικόνων έχει οδηγήσει στην ύπαρξη πολυάριθμων τεχνικών με πολλά περιθώρια βελτίωσης, αλλά και στη δυνατότητα εμφάνισης πολλών ακόμα νέων τεχνικών. Ο κοινός παρονομαστής όλων των συστημάτων ανάκτησης εικόνων με βάση το περιεχόμενο είναι η προσπάθεια γεφύρωσης δύο, όπως συνηθίζεται να αποκαλούνται, κενών [8, 38]:

- **Αισθητήριο κενό (sensory gap):** Πρόκειται για τη διαφορά μεταξύ ενός αντικειμένου στον πραγματικό κόσμο και της πληροφορίας που περιλαμβάνεται σε μία περιγραφή του η οποία προέρχεται από την καταγραφή του αντικειμένου αυτού με οποιοδήποτε μέσο.

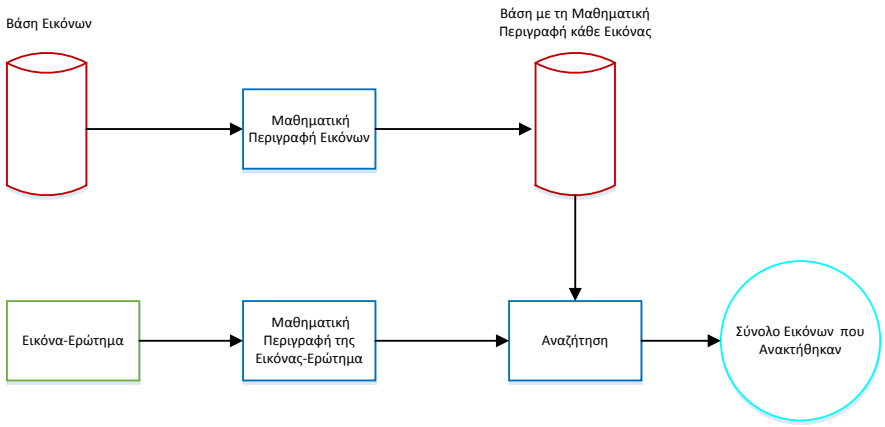
- **Σημασιολογικό κενό (semantic gap):** Πρόκειται για τη διαφορά μεταξύ της πληροφορίας που μπορεί κάποιος να εξάγει από τα οπτικά δεδομένα και της ερμηνείας που μπορεί να δώσει στα ίδια δεδομένα ένας οποιοδήποτε χρήστης κάτω από συγκριμένες συνθήκες.

Αναφορικά με το αισθητήριο κενό, ως παράδειγμα μπορούμε να αναφέρουμε τις φωτογραφίες που λαμβάνονται μέσω φωτογραφικών μηχανών. Συγκεκριμένα, μία φωτογραφία αποδίδει αντικείμενα του τριδιάστατου χώρου σε διδιάστατες εικόνες με συνέπεια να υπάρχει απώλεια πληροφορίας. Σε αυτό παίζει καθοριστικό ρόλο η γωνία λήψης, η φωτεινότητα του χώρου, η μερική ή ολική απόκρυψη αντικειμένων από άλλα αντικείμενα κτλ. Δηλαδή οι ιδιότητες ενός αντικειμένου όπως αναπαρίσταται σε μία εικόνα και οι ιδιότητες του ίδιου του αντικειμένου στον πραγματικό κόσμο ενδέχεται να διαφέρουν σημαντικά.

Από την άλλη, το σημασιολογικό κενό βασίζεται στο γεγονός πως ο χρήστης αναζητά σημασιολογική ομοιότητα μεταξύ εικόνων με βάση τις εμπειρίες που έχει από την ημέρα της γέννησής του, ενώ ένα σύστημα μπορεί να παρέχει μόνο ομοιότητα βασισμένη σε ηλεκτρονική επεξεργασία δεδομένων. Σε αυτό μπορούμε να προσθέσουμε και τη θεωρία της γενετικής εξέλιξης της ανθρώπινης όρασης ανά τους αιώνες ώστε να γίνει απολύτως κατανοητό πως είναι αναπόφευκτο να υπάρχει απόκλιση μεταξύ της ερμηνείας που δέχονται οι εικόνες από τους ανθρώπους και τις μηχανές. Άλλωστε, όπως έχουμε ήδη αναφέρει, υποκειμενικότητα υφίσταται ακόμη και στην ερμηνεία εικόνων μεταξύ διαφορετικών ανθρώπων, πόσο μάλλον όταν πρόκειται για μηχανές και ανθρώπους.

Το αισθητήριο κενό λοιπόν οφείλεται στους περιορισμούς που εισάγουν τα μέσα καταγραφής του πραγματικού κόσμου, ενώ το σημασιολογικό κενό οφείλεται στην εν γενεί δυσκολία της μαθηματικής απόδοσης του οπτικού περιεχομένου των εικόνων με τον ίδιο ακριβώς τρόπο που θα το ερμήνευε ένας άνθρωπος. Όπως προαναφέραμε, στόχος κάθε συστήματος είναι να εξαλείψει τα δύο αυτά κενά, ωστόσο με δεδομένο πως αυτό δεν είναι εφικτό σε απόλυτο βαθμό, ανάλογα με την εφαρμογή του συστήματος ενδέχεται να δοθεί μεγαλύτερη ή μικρότερη έμφαση στα προβλήματα που εκφράζουν τα δύο κενά.

Η ερευνητική μας εργασία αφορά την ανάκτηση εικόνων με βάση το περιεχόμενο και συνεπώς δεν θα δοθεί περαιτέρω έμφαση στην ανάκτηση εικόνων με βάση το κείμενο. Η διαδικασία της ανάκτησης εικόνων που βασίζεται στο οπτικό περιεχόμενο απεικονίζεται στο σχήμα 1.1.



Σχήμα 1.1: Δομή συστήματος ανάκτησης εικόνων βασισμένο στο περιεχόμενο.

Όπως μπορούμε να δούμε τα πάντα ξεκινάνε από την ύπαρξη μίας βάσης εικόνων και μίας εικόνας−ερώτημα με βάση την οποία θέλουμε να αναζητήσουμε όμοιες οπτικά εικόνες στη βάση. Προκειμένου να μπορέσουμε να συγκρίνουμε τις εικόνες μεταξύ τους οφείλουμε να τις αναπαραστήσουμε μαθηματικά. Δημιουργούμε λοιπόν μία νέα βάση δεδομένων όπου κάθε εικόνα περιλαμβάνεται με τη μαθηματική της αναπαράσταση. Έχοντας περιγράψει μαθηματικά και την εικόνα−ερώτημα, είμαστε σε θέση να τη συγκρίνουμε με τις μαθηματικές αναπαραστάσεις των εικόνων της βάσης ώστε να καταλήξουμε στο σύνολο των εικόνων που παρουσιάζουν τη μεγαλύτερη ομοιότητα με αυτήν. Ανακτώνται και παρουσιάζονται στο χρήστη είτε οι εικόνες ενός υποσυνόλου της βάσης οι οποίες εμφανίζουν την υψηλότερη ομοιότητα με την εικόνα−ερώτημα, είτε όλες οι εικόνες σε φθίνουσα σειρά ομοιότητας. Όπως γίνεται κατανοητό, για την επιτυχή πραγματοποίηση της παραπάνω διαδικασίας οφείλουμε να μελετήσουμε κάθε στάδιο ξεχωριστά.

1.2.1 Βάσεις Εικόνων

Οι βάσεις εικόνων που χρησιμοποιούνται στην ανάκτηση εικόνων αποτελούν οργανωμένες συλλογές από εικόνες. Ο τρόπος οργάνωσης και οι ιδιότητες των εικόνων ποικίλουν, κατά συνέπεια τα κριτήρια για την επιλογή ή τη δημιουργία της βάσης εικόνων που θα χρησιμοποιηθεί προσδιορίζονται από την εκάστοτε εφαρμογή.

Ο πιο εύκολος τρόπος να διαχωρίσουμε βάσεις εικόνων είναι με κριτήριο το μέγεθος τους. Μία βάση μπορεί να έχει από εκατοντάδες μέχρι δισεκατομμύρια εικόνες. Το μέγεθος της βάσης που θα χρησιμοποιηθεί σε μία εφαρμογή παίζει καθοριστικό ρόλο στο σχεδιασμό ενός συστήματος. Αρκεί να αναφέρουμε πως σε μία μικρή βάση μπορεί να πραγματοποιηθεί εξαντλητική αναζήτηση μεταξύ των εικόνων, ενώ αντίθετα, όσο αυξάνεται το μέγεθος της βάσης οι απαιτήσεις χρόνου και μνήμης επιβάλλουν τη χρήση προσεγγιστικών και βελτιστοποιημένων

μεθόδων. Ένα άλλο παράδειγμα αφορά συστήματα τα οποία χρειάζεται πρώτα να εκπαιδευτούν και έπειτα να δοκιμαστούν. Το στάδιο της εκπαίδευσης συνήθως απαιτεί τη χρήση ενός ελάχιστου αριθμού εικόνων που σαφώς πρέπει να καλύπτεται από τη βάση που είναι διαθέσιμη.

Αναφορικά με τον τρόπο οργάνωσης μίας βάσης, σημαντικό ρόλο παίζουν τα μεταδεδομένα που συνοδεύουν τις εικόνες. Στις περισσότερες ευρέως διαδεδομένες βάσεις παρέχονται μεταδεδομένα που βοηθούν την αξιολόγηση των συστημάτων ανάκτησης εικόνων που τις χρησιμοποιούν. Η όλη διαδικασία της δημιουργίας των μεταδεδομένων είναι γνωστή διεθνώς με τον όρο ''annotation''. Για παράδειγμα, στη βάση εικόνων Caltech 101 [10] που περιέχει εικόνες αντικειμένων, για κάθε εικόνα έχει υποσημειωθεί το περίγραμμα του αντικειμένου που περιλαμβάνει. Έτσι, ένα σύστημα αναγνώρισης αντικειμένων που χρησιμοποιεί την συγκεκριμένη βάση μπορεί άμεσα να αξιολογήσει τα αποτελέσματά του. Στις περισσότερες βάσεις είθισται να παρέχονται μεταδεδομένα που αφορούν το πεδίο ενδιαφέροντος κάθε εικόνας.

Η χρήση του παραδείγματος της βάσης Caltech 101 μας διευκολύνει στο να αναφερθούμε σε έναν ακόμα τρόπο οργάνωσης των εικόνων μίας βάσης ο οποίος αφορά τον χωρισμό των εικόνων ανά κατηγορία. Στην Caltech 101 που προαναφέραμε, περιλαμβάνονται 101 διαφορετικές κατηγορίες αντικειμένων και παρέχεται πρόσβαση σε κάθε μία ξεχωριστά. Αυτός ο τρόπος οργάνωσης είναι ιδιαίτερα χρήσιμος τόσο για την προσπέλαση των εικόνων όσο και για την αξιολόγηση της ανάκτησης καθώς είμαστε σε θέση να γνωρίζουμε εάν οι εικόνες που ανακτήθηκαν ανήκουν στην επιθυμητή κατηγορία.

Ένα εξαιρετικά σημαντικό σημείο διαφοροποίησης μεταξύ βάσεων εικόνων αποτελεί το εύρος του πεδίου αναζήτησης των εικόνων που περιέχουν. Συγκεκριμένα υπάρχουν δύο κατηγορίες [38]:

- **Στενό πεδίο εικόνων:** Τα οπτικά χαρακτηριστικά των εικόνων έχουν περιορισμένη και προβλέψιμη διακύμανση και μπορούν να οριστούν με σχετική ευκολία. Παράδειγμα αποτελεί ένα σύνολο με εικόνες που περιλαμβάνουν μπροστινές όψεις ανθρώπινων προσώπων με καθαρό φόντο. Σε αυτή την περίπτωση, ενώ τα χαρακτηριστικά του προσώπου διαφορετικών ανθρώπων διαφέρουν σημαντικά, υπάρχει μεγάλη συνάφεια όσων αφορά τα γεωμετρικά χαρακτηριστικά, όπως επίσης υπάρχει μικρή διακύμανση στο χρώμα.

- **Ευρύ πεδίο εικόνων:** Τα οπτικά χαρακτηριστικά των εικόνων έχουν απεριόριστη και απρόβλεπτη διακύμανση ακόμα και όταν πρόκειται για εικόνες με το ίδιο σημασιολογικό περιεχόμενο. Παράδειγμα αποτελούν εικόνες ανθρώπινων προσώπων που έχουν ληφθεί σε τυχαίες καθημερινές στιγμές, καθώς κάτι τέτοιο συνεπάγεται τυχαίο φόντο, απρόβλεπτη γωνία λήψης, διακύμανση στο φωτισμό κτλ.

Ο διαχωρισμός αυτός φανερώνει ένα αρκετά σημαντικό στοιχείο των βάσεων εικόνων που δεν είναι άλλο από τα οπτικά χαρακτηριστικά των εικόνων. Οι εικόνες λοιπόν ενδέχεται να αναπαριστούν είτε φυσικά τοπία είτε αντικείμενα, επίσης ενδέχεται τα αντικείμενα αυτά να είναι έμψυχα ή άψυχα. Μία ακόμα πηγή διαφοροποίησης του περιεχομένου των εικόνων εντοπίζεται στη γωνία λήψης, στη φωτεινότητα, στο εάν απεικονίζεται ολόκληρο ή μέρος του αντικειμένου που επιθυμούμε, όπως και στο εάν υπάρχει παρεμβολή άλλων αντικειμένων. Έπειτα, υπάρχουν περιπτώσεις που οι εικόνες τυγχάνουν περαιτέρω επεξεργασίας με αποτέλεσμα να αλλάζουν οι

διαστάσεις τους, να περιστρέφονται και γενικότερα να επιβάλλονται ποικίλα φίλτρα και γεωμετρικοί μετασχηματισμοί. Σημαντικό ρόλο στα παραπάνω διαδραματίζει και ο τρόπος λήψης μίας εικόνας. Διαφορετικό εργαλείο λήψης μπορεί να οδηγήσει εικόνες του ίδιου αντικειμένου να έχουν διαφορετική ανάλυση, διαφορετική φωτεινότητα κτλ.

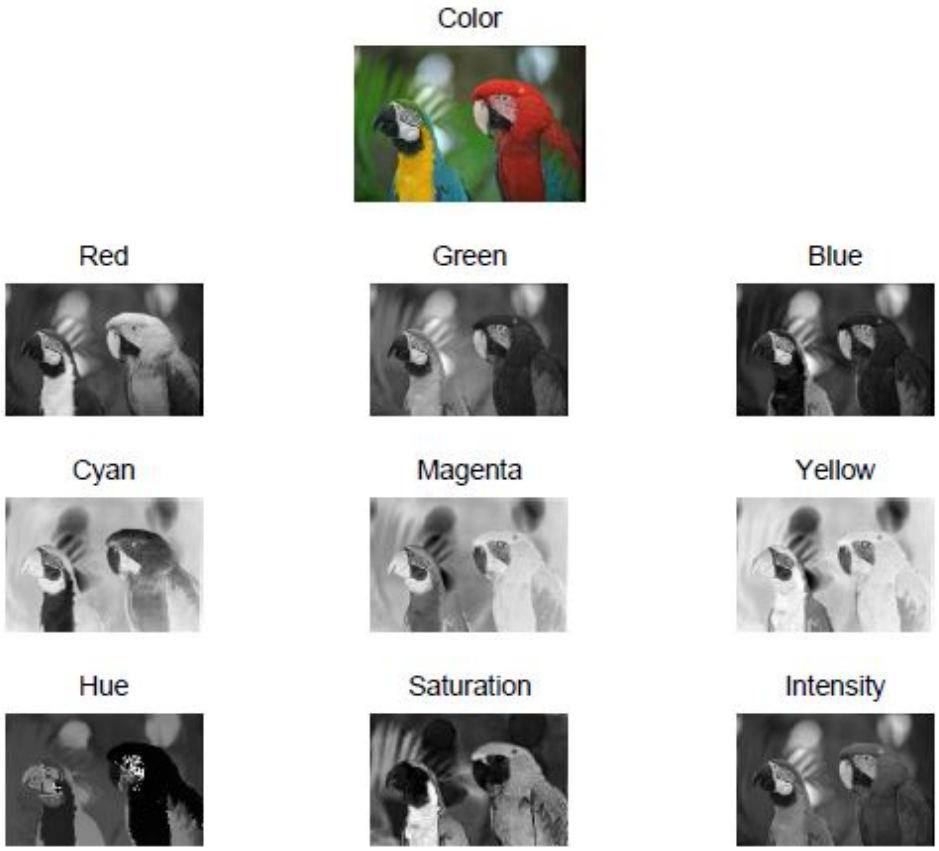
Οφείλουμε επίσης να προσθέσουμε τη σημασία που έχουν τα τεχνικά χαρακτηριστικά των εικόνων και ειδικότερα το μέγεθός τους. Το μέγεθος μίας εικόνας έχει αντίκτυπο στο πόσο πλούσιο είναι το περιεχόμενό της μέσω της διαθέσιμης ανάλυσης, αλλά κατά κύριο λόγο σχετίζεται με τους απαιτούμενους πόρους επεξεργασίας και αποθήκευσής της. Μικρότερες εικόνες συνήθως αντιστοιχούν σε μικρότερη ανάλυση και απαιτούν λιγότερη μνήμη για να αποθηκευτούν, λιγότερους υπολογιστικούς πόρους για την επεξεργασία τους και φυσικά η επεξεργασία αυτή γίνεται και σε μικρότερο χρόνο. Οι βάσεις εικόνων λοιπόν, συνήθως περιλαμβάνουν εικόνες ενός συγκεκριμένου μεγέθους ή τουλάχιστον ενός συγκεκριμένου εύρους διαστάσεων. Επιστρέφοντας στο παράδειγμα της βάσης Clatech 101, όλες οι εικόνες είναι διαστάσεων περίπου 300×300 εικονοστοιχεία.

Ένα ακόμα χαρακτηριστικό των βάσεων δεδομένων είναι ο τρόπος αποθήκευσης των εικόνων. Από αυτή την οπτική γωνία υπάρχουν εκατοντάδες διαφορετικά είδη εικόνων, και το κάθε ένα εξυπηρετεί καλύτερα μια διαφορετική χρήση. Τα πιο διαδεδομένα είναι τα αρχεία εικόνων jpg, png, και bmp που περιγράφουν κάθε εικονοστοιχείο χωριστά. Παράλληλα, υπάρχουν τα διανυσματικά αρχεία εικόνων όπως είναι τα cgm και svg που αποθηκεύουν μία γεωμετρική περιγραφή της εικόνας με αποτέλεσμα να καθίσταται εφικτή η προσαρμογή τους σε οθόνες οποιουδήποτε μεγέθους. Το ''format'', όπως αποκαλείται, των εικόνων ποικίλει καθώς ποικίλουν και οι εφαρμογές των εικόνων. Για παράδειγμα σε βιοϊατρικές εφαρμογές συναντώνται οι εικόνες mhd. Συνέπεια αυτού του πλήθους των διαφορετικών τρόπων αποθήκευσης είναι η ανάγκη ύπαρξης κατάλληλου λογισμικού για την ανάγνωση και αποθήκευση των εικόνων στην αντίστοιχη μορφή. Για αυτόν τον λόγο, βάσεις που περιέχουν εικόνες οι οποίες είναι αποθηκευμένες με εξεζητημένο τρόπο, συνοδεύονται από αρχεία κώδικα ή πακέτα λογισμικού που επιτρέπουν την ανάγνωση και αποθήκευση των εικόνων.

Συμπερασματικά λοιπόν, το πλήθος στα είδη των εικόνων δεν μπορεί να θεωρηθεί πεπερασμένο όπως δεν μπορούν να θεωρηθούν πεπερασμένες και οι αναπαραστάσεις του κόσμου γύρω μας. Το τι και πώς θα αναπαριστούν οι εικόνες μίας βάσης, όπως και το ποια θα είναι τα τεχνικά χαρακτηριστικά, ο τρόπος αποθήκευσης και ο τρόπος οργάνωσής τους μέσα στην ίδια τη βάση, υπόκεινται στην εκάστοτε εφαρμογή και τις ιδιαίτερες ανάγκες της.

1.2.2 Μαθηματική Αναπαράσταση

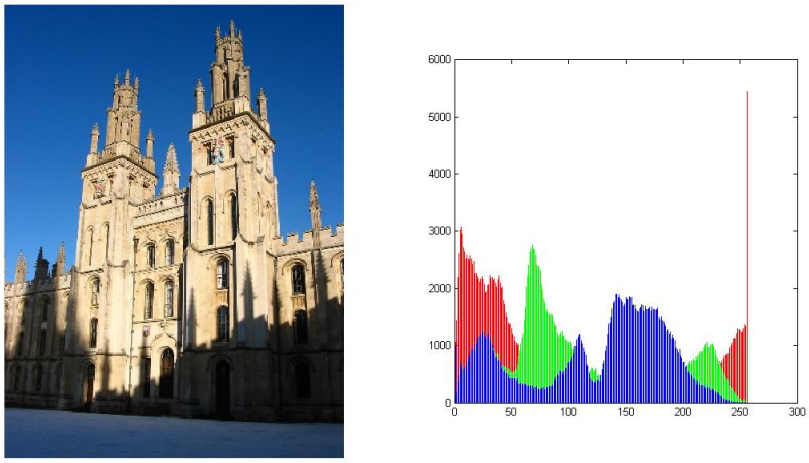
Προκειμένου να μπορέσουμε να πραγματοποιήσουμε οποιουδήποτε είδους ψηφιακή διεργασία με τη χρήση εικόνων οφείλουμε να τις αναπαραστήσουμε με τρόπο συμβατό προς τους ηλεκτρονικούς υπολογιστές και εύκολα αξιοποιήσιμο για το σκοπό που θέλουμε να τις χρησιμοποιήσουμε. Ο πιο απλός τρόπος μαθηματικής αναπαράστασης των εικόνων είναι μέσω πινάκων φωτεινότητας. Για παράδειγμα, μία ασπρόμαυρη εικόνα μπορεί να αναπαρασταθεί από έναν πίνακα διαστάσεων 200×300 κάθε στοιχείο του οποίου αντιστοιχεί σε μία τιμή φωτεινό-



Σχήμα 1.2: Παράδειγμα ανάλυσης εικόνας σε χρωματικά κανάλια ανάλογα με το χρωματικό χώρο που χρησιμοποιείται για την αναπαράστασή της μέσω του χρώματος [28].

τητας. Οι τιμές φωτεινότητας αποδίδουν τις αποχρώσεις του γκρι έχοντας ως ελάχιστη τιμή το μαύρο και ως μέγιστη το λευκό. Ωστόσο, μία τέτοια αναπαράσταση η οποία βασίζεται σε τιμές φωτεινότητας δεν προσεγγίζει την απόκριση του οπτικού συστήματος του ανθρώπου και ούτε περιγράφει ικανοποιητικά το σημασιολογικό περιεχόμενο μιας εικόνας. Συνεπώς αρκεί για την πραγματοποίηση περιορισμένων και απλών διεργασιών όπως είναι η προβολή μίας εικόνας.

Οπτικά χαρακτηριστικά. Όταν επιθυμούμε να αναπαραστήσουμε εικόνες για την χρησιμοποίησή τους στην ανάκτηση εικόνων επιχειρούμε να τις περιγράψουμε βασισμένοι σε αντιπροσωπευτικά οπτικά χαρακτηριστικά. Η έννοια του ‘‘αντιπροσωπευτικού’’ αποδίδει τη διακριτική ικανότητα που προσδίδει ένα χαρακτηριστικό στην μαθηματική περιγραφή μίας εικόνας και είναι σαφώς σχετική, καθώς εξαρτάται από το πρόβλημα που έχουμε να αντιμετωπίσουμε. Τα δημοφιλέστερα χαρακτηριστικά που χρησιμοποιούνται σε συστήματα ανάκτησης εικόνων είναι το χρώμα, το σχήμα, η υφή και τα σημεία ενδιαφέροντος. Για κάθε ένα υπάρχουν ποικίλοι τρόποι περιγραφής καθώς και συνδυαστικές μέθοδοι. Επίσης, πέραν των προαναφερθέντων χαρακτηριστικών υπάρχουν αρκετά ακόμα τα οποία σχετίζονται κατά βάση με την εκάστοτε εφαρμογή. Αξίζει να σημειωθεί πως σε αυτή την κατεύθυνση έχει συμβάλει τα μέγιστα η

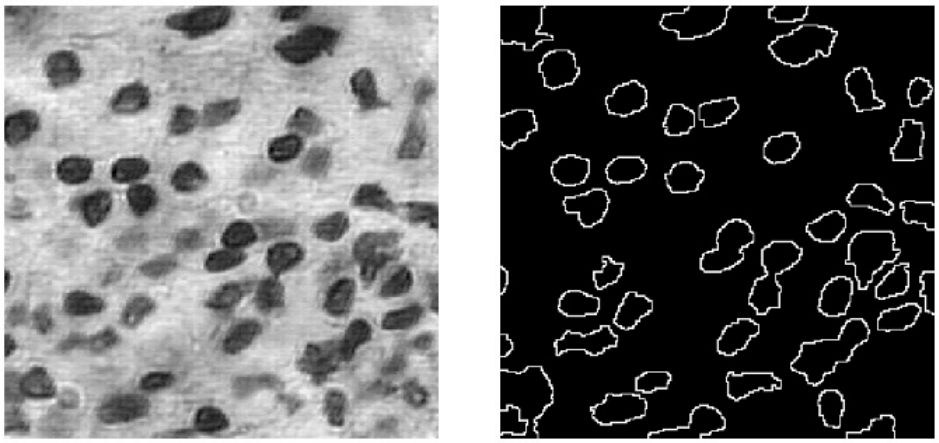


Σχήμα 1.3: Παράδειγμα περιγραφής εικόνας μέσω ιστογράμματος χρώματος.

επιστημονική περιοχή της αναγνώρισης προτύπων.

Το χρώμα αποτελεί ένα από τα πιο διαδεδομένα χαρακτηριστικά και ο πιο γνωστός τρόπος αναπαράστασής του στην τεχνολογία ψηφιακών εικόνων είναι μέσω του χρωματικού χώρου RGB. Ο όρος RGB αποτελεί ακρωνύμιο του Red-Green-Blue και σύμφωνα με τον χώρο αυτό κάθε χρώμα μπορεί να παραχθεί από τη μίξη τριών θεμελιωδών χρωμάτων, του κόκκινου, του πράσινου και του μπλε. Συνεπώς, το χρώμα του κάθε εικονοστοιχείου ορίζεται ως μία τριάδα τιμών που αντιπροσωπεύουν το ποσοστό μίξης του κάθε ενός από τα τρία προαναφερθέντα πρωταρχικά χρώματα. Ωστόσο, υπάρχουν αρκετοί ακόμα χρωματικοί χώροι, όπως για παράδειγμα οι HSI (Hue-Intensity-Saturation) και CMY (Cyan-Magenta-Yellow). Σχετικό παράδειγμα παρατίθεται στο σχήμα 1.2. Η ύπαρξη διαφορετικών χρωματικών χώρων συνδέεται με τις διαφορετικές ανάγκες που παρουσιάζει κάθε εφαρμογή. Για παράδειγμα ο χρωματικός χώρος HSI αντιστοιχεί καλύτερα στην ανθρώπινη αντίληψη της τοπολογίας του χώρου των χρωμάτων, δηλαδή συμπίπτει σε μεγάλο βαθμό με τον τρόπο που αντιλαμβάνεται την όραση ο ανθρώπινος οφθαλμός μέσω της απόχρωσης, της έντασης και του κορεσμού. Ανάλογα με τη χρήση λοιπόν, πρέπει να επιλεγεί ο καταλληλότερος χρωματικός χώρος. Οι συνηθέστεροι τρόποι μαθηματικής περιγραφής μίας εικόνας μέσω του χρώματος είναι με χρήση ιστογραμμάτων και στατιστικών ιδιοτήτων. Ένα ιστόγραμμα περιέχει τη διακύμανση των τιμών φωτεινότητας μίας εικόνας. Ωστόσο, υπάρχουν και ιστογράμματα που εμπεριέχουν και χωρική πληροφορία αναφορικά με το πού εμφανίζεται κάθε χρώμα. Ένα απλό παράδειγμα ιστογράμματος χρώματος απεικονίζεται στο σχήμα 1.3. Οι στατιστικές ιδιότητες αναφέρονται στον υπολογισμό μίας κατανομής για τις τιμές χρώματος μίας εικόνας αλλά και στον υπολογισμό μεγεθών όπως ροπές πρώτης, δεύτερης και τρίτης τάξης [39]. Ένα ακόμα παράδειγμα περιγραφής εικόνων μέσω χαρακτηριστικών χρώματος αποτελούν οι MPEG-7 Color Descriptors [22].

Το σχήμα ως χαρακτηριστικό περιγραφής εικόνων παρουσιάζει μεγάλη ποικιλία καθώς



Σχήμα 1.4: Παράδειγμα κατάτμησης εικόνας μέσω του αλγορίθμου watershed.

υπάρχει πληθώρα κατηγοριών σχημάτων τόσο στη φύση όσο και στο χώρο των μαθηματικών. Ο τρόπος αναπαράστασης του σχήματος των αντικειμένων μπορεί να χωριστεί σε δύο κατηγορίες. Η πρώτη αναφέρεται στη χρησιμοποίηση του περιγράμματος και η δεύτερη στη χρησιμοποίηση της εσωτερικής περιοχής ενός σχήματος. Το περίγραμμα μπορεί να περιγραφεί μέσα από fourier μετασχηματισμό της καμπύλης του περιγράμματος, ενώ το εσωτερικό ενός σχήματος μέσω ροπών και απλών ιδεών αναλυτικής γεωμετρίας. Το σχήμα ενός αντικειμένου μπορεί επίσης να περιγραφεί μέσω προβολών όπως η οριζόντια, η κάθετη και η διαγώνια προβολή. Σημαντικό ρόλο διαδραματίζουν και οι έννοιες της καμπυλότητας αλλά και αυτή του σκελετού ενός σχήματος [28]. Ιδιαίτερη αναφορά αξίζει στη διαδικασία της κατάτμησης εικόνων η οποία οδηγεί στο διαχωρισμό των αντικειμένων μίας εικόνας από το φόντο. Συγκεκριμένα, η εικόνα χωρίζεται σε μικρότερα τμήματα μέσω ενός κριτηρίου ομοιογένειας. Υπάρχει μεγάλος αριθμός αλγορίθμων κατάτμησης, ενδεικτικά αναφέρουμε τους αλγορίθμους n-cut [36] και watershed [14]. Σε κάθε περίπτωση, η διαδικασία της κατάτμησης μπορεί να αποτελέσει προπαρασκευαστικό στάδιο ώστε να προσδιοριστεί το σχήμα των αντικειμένων που περιλαμβάνονται στις εικόνες. Ένα παράδειγμα κατάτμησης φαίνεται στο σχήμα 1.4.

Η υφή αναφέρεται σε επαναλαμβανόμενα οπτικά μοτίβα επιφανειών μέσα σε μία εικόνα. Τα μοτίβα αυτά παρουσιάζουν ιδιότητες ομοιογένειας οι οποίες όμως δεν βασίζονται στην ύπαρξη ενός μόνο χρώματος ή φωτεινότητας, αλλά εμπεριέχουν σημαντικές πληροφορίες αναφορικά με τη δομή και την οργάνωση των επιφανειών και της σχέσης τους με τον περιβάλλοντα χώρο. Στο σχήμα 1.5 φαίνεται η διαφορετική υφή ενός ξύλινου τοίχου και ενός με τούβλα. Όπως γίνεται κατανοητό, η υφή ενός αντικειμένου αποτελεί σημαντική πληροφορία η οποία όμως είναι δύσκολο τόσο να προσδιοριστεί όσο και να περιγραφεί μαθηματικά. Προκειμένου να περιγραφεί η υφή υπάρχουν τριών ειδών μέθοδοι, στατιστικές, γεωμετρικές και ενέργειας [28]. Ενδεικτικά αναφέρουμε πως στις στατιστικές μεθόδους γίνεται χρήση των Gibbs και Markov Random Fields, στις γεωμετρικές εξέχουσα θέση έχουν τα fractals, και στις μεθόδους ενέργειας δεσπόζουν τα φίλτρα Gabor. Σημαντικό ρόλο επίσης διαδραματίζουν μετασχηματισμοί

Σχήμα 1.5: Παράδειγμα διαφορετικής υφής μεταξύ εικόνων.

όπως ο μετασχηματισμός κυματιδίων και ο διακριτός μετασχηματισμός συνημιτόνου. Επιπλέον, αναφορικά με τα χαρακτηριστικά υφής σε ένα συγκεκριμένο σημείο, πρέπει να σημειώσουμε πως έχουν σημασία μόνο συναρτήσει των γειτονικών εικονοστοιχείων του σημείου αυτού. Η έννοια της υφής λοιπόν αναφέρεται σε γειτονιές από εικονοστοιχεία, και το μέγεθος της γειτονιάς τωνεικονοστοιχείων που λαμβάνονται υπόψιν για την περιγραφή των χαρακτηριστικών λογίζεται ως η κλίμακα (scale) του εκάστοτε χαρακτηριστικού.

Τα σημεία ενδιαφέροντος (keypoints) αντιστοιχούν σε τοπικά χαρακτηριστικά όπως γωνίες, ακμές κτλ. Ένα παράδειγμα ανιχνευτή γωνιών φαίνεται στο σχήμα 1.6. Η σημασία των χαρακτηριστικών αυτών εντοπίζεται στο γεγονός πως μπορούν να αναπαραστήσουν σημαντικές περιοχές μίας εικόνας με αποδοτικό τρόπο. Τέτοιου είδους χαρακτηριστικά θα αποτελέσουν το κύριο πεδίο ενδιαφέροντος στο επόμενο κεφάλαιο λόγω της χρησιμοποίησής τους από τη μέθοδό μας.

Ανίχνευση και περιγραφή χαρακτηριστικών. Ο τρόπος ανίχνευσης και περιγραφής των τοπικών χαρακτηριστικών μίας εικόνας που θα μας απασχολήσει και στη συνέχεια, πραγματοποιείται σε δύο στάδια:

- **Ανίχνευση χαρακτηριστικών (feature detection):** Ο στόχος είναι η εύρεση αντιπροσωπευτικών χαρακτηριστικών σε μία εικόνα. Ο αλγόριθμος για την πραγματοποίηση αυτής της διαδικασίας ονομάζεται ανιχνευτής (detector) και μας επιτρέπει να γνωρίζουμε εάν κάθε ειχνοστοιχείο μίας εικόνας αντιστοιχεί σε σημείο ενδιαφέροντος.

- **Περιγραφή χαρακτηριστικών (feature description):** Ο στόχος είναι να αποδοθεί σε κάθε σημείο ενδιαφέροντος ένα διάνυσμα περιγραφής (descriptor) το οποίο να το περιγράφει.

Ένας ανιχνευτής λοιπόν εντοπίζει σημεία ενδιαφέροντος τα οποία αντιστοιχούν σε συγκεκριμένα τοπικά χαρακτηριστικά και στη συνέχεια εξάγονται διανύσματα περιγραφής. Υπάρχουν ποικίλοι τρόποι για την πραγματοποίηση τόσο της ανίχνευσης όσο και της περιγραφής των οπτικών χαρακτηριστικών μίας εικόνας. Ενδεικτικά μπορούμε να αναφέρουμε τους ανιχνευτές SIFT [27], SURF [3] και Hessian affine [30], καθώς και τις τεχνικές υπολογισμού διανυσμάτων περιγραφής SIFT [27], SURF [3] και DAISY [41].

Σχήμα 1.6: Παράδειγμα εφαρμογής ανιχνευτή γωνιών.

Ένας διαφορετικός τρόπος ανίχνευσης χαρακτηριστικών είναι η πυκνή δειγματοληψία (dense sampling). Συγκεκριμένα, δειγματοληπτείται ο χώρος των εικονοστοιχείων μέσω ενός ομοιογενούς πλέγματος και στη συνέχεια για κάθε δείγμα που αντιστοιχεί σε ένα εικονοστοιχείο, υπολογίζεται ένα διάνυσμα περιγραφής. Η μέθοδος αυτή εφαρμόζεται σε εικόνες πλούσιες σε πληροφορία από τις οποίες θέλουμε να εξάγουμε χαρακτηριστικά στο σύνολό τους.

Ένας εξαιρετικά σημαντικός παράγοντας για την επιλογή του κατάλληλου τρόπου ανίχνευσης και περιγραφής χαρακτηριστικών αποτελεί η έννοια της ανεξαρτησίας (invariance). Η ανεξαρτησία αναφέρεται στην ανάγκη ανίχνευσης όμοιων εικόνων ανεξάρτητα από τον τρόπο που αναπαρίσταται το περιεχόμενό τους. Για παράδειγμα, στο σχήμα 1.7 απεικονίζονται εικόνες του ίδιου περιεχομένου, στις οποίες όμως έχουν εφαρμοστεί αφινικοί μετασχηματισμοί. Εάν η αρχική εικόνα αποτελούσε εικόνα−ερώτημα, μέσω της διαδικασίας της ανάκτησης θα θέλαμε να μας επιστραφούν όλες οι υπόλοιπες. Δηλαδή θα θέλαμε τα χαρακτηριστικά που θα εξαχθούν από τις εικόνες να μην επηρεαστούν από τους μετασχηματισμούς που έχουν υποστεί. Γίνεται κατανοητό λοιπόν πως ανάλογα με την εφαρμογή, τα χαρακτηριστικά που ανιχνεύονται και εξάγονται από τις εικόνες είναι επιθυμητό να είναι ανεξάρτητα ως προς κάποιες ιδιότητες. Πιο συνηθισμένες τέτοιες ιδιότητες είναι η θέση του χώρου ενδιαφέροντος της εικόνας, το μέγεθός του, η γωνία περιστροφής του ως προς κάποιον άξονα, οι διακυμάνσεις της φωτεινότητας ακόμα και οι διαφοροποιήσεις στη γωνία λήψης. Πρέπει να σημειώσουμε πως η ανεξαρτησία σε μεγάλο εύρος ιδιοτήτων ενδέχεται να είναι απευκταία καθώς σε ορισμένες περιπτώσεις συνεπάγεται χαμηλότερη διακριτική ικανότητα. Συνεπώς, η επιλογή των ιδιοτήτων ως προς τις οποίες εξασφαλίζεται ανεξαρτησία πρέπει να γίνεται προσεκτικά και σύμφωνα πάντα με τις απαιτήσεις της εφαρμογής.

Είναι πολλές οι περιπτώσεις στις οποίες δεν αρκεί η ανίχνευση και η περιγραφή των χα-

Σχήμα 1.7: Παράδειγμα αφινικών μετασχηματισμών εικόνας [28].

ρακτηριστικών προκειμένου να αποδώσουμε υψηλή διακριτική ικανότητα στη μαθηματική περιγραφή ενός συνόλου από εικόνες. Για αυτόν ακριβώς το λόγο, συχνά πραγματοποιείται ένα επιπλέον στάδιο επεξεργασίας στο οποίο χρησιμοποιούνται μέθοδοι οι οποίες λαμβάνουν ως είσοδο τα διανύσματα περιγραφής και προβαίνουν σε περαιτέρω επεξεργασία προκειμένου να αποδοθεί η τελική μαθηματική περιγραφή των εικόνων. Ιδιαίτερο ενδιαφέρουν παρουσιάζουν τέτοιες μέθοδοι οι οποίες αποδίδουν μαθηματικά τις εικόνες μέσω ενός μοναδικού διανύσματος. Η μέθοδός μας, SP–VLAD, ανήκει σε αυτήν την κατηγορία μεθόδων, όπως και άλλες εξαιρετικά σημαντικές μέθοδοι τις οποίες θα παρουσιάσουμε λεπτομερώς σε επόμενα κεφάλαια.

1.2.3 Αναζήτηση

Έχοντας επιλέξει τα κατάλληλα χαρακτηριστικά και έχοντας προβεί στην μαθηματική αναπαράσταση των εικόνων μίας βάσης, στοχεύουμε στην πραγματοποίηση αποδοτικής αναζήτησης μεταξύ των εικόνων της βάσης ώστε να επιλεγούν αυτές που θα ανακτηθούν. Οι τρεις βασικές κατηγορίες αναζήτησης που συναντώνται στην ανάκτηση εικόνων είναι οι ακόλουθες [38]:

- **Αναζήτηση βάσει σχέσης:** Δεν αναζητούνται ξεκάθαρα κάποια ή κάποιες εικόνες, αντίθετα απλώς επιδιώκεται ένα ενδιαφέρον αποτέλεσμα. Η αναζήτηση επαναλαμβάνεται περισσότερες από μία φορές και κάθε φορά το αποτέλεσμα αξιολογείται ώστε να βελτιωθεί στην επόμενη επανάληψη [23].

- **Αναζήτηση βάσει στόχου:** Αναζητούνται μία ή περισσότερες συγκεκριμένες εικόνες. Παράδειγμα θα μπορούσε να αποτελεί η αναζήτηση ενός συγκεκριμένου έργου τέχνης σε μία συλλογή εικόνων με πίνακες ζωγραφικής.

- **Αναζήτηση βάσει κατηγορίας:** Αναζητούνται εικόνες που ανήκουν σε μία συγκεκριμένη σημασιολογική κλάση. Για παράδειγμα, σε μία συλλογή εικόνων με έπιπλα ενδέχεται να αναζητούμε εικόνες οι οποίες απεικονίζουν καρέκλες.

Στην ερευνητική μας εργασία επικεντρωνόμαστε στην αναζήτηση βάσει κατηγορίας. Η ποιότητα μίας τέτοιας αναζήτησης προσδιορίζεται κυρίως από τρεις παράγοντες: την ακρίβεια, την αποδοτικότητα και τις απαιτήσεις χρησιμοποίησης μνήμης. Συγκεκριμένα, η ακρίβεια αναφέρεται στην σημασιολογική ομοιότητα των ανακτηθέντων εικόνων, δηλαδή στο κατά πόσο οι εικόνες που ανακτήθηκαν ήταν αυτές που θα θέλαμε ως χρήστες. Η αποδοτικότητα αναφέρεται στο βαθμό που χρησιμοποιήθηκαν ικανοποιητικά οι υπολογιστικοί πόροι, κάτι που συνδέεται αφενός με την απαιτούμενη υπολογιστική ισχύ και αφετέρου με τον χρόνο απόκρισης, δηλαδή το πόσο γρήγορα ολοκληρώνεται η αναζήτηση. Τέλος, οι απαιτήσεις μνήμης προκύπτουν τόσο από το μέγεθος της μαθηματικής αναπαράστασης των εικόνων όσο και από τον τρόπο αναζήτησης λόγω των απαιτήσεων μνήμης που παρουσιάζονται στα ενδιάμεσα στάδια της χρησιμοποιούμενης μεθόδου.

Μεταξύ της αποδοτικότητας και των απαιτήσεων μνήμης υπάρχει στενή σχέση, καθώς η αποδοτικότητα μπορεί να προσεγγιστεί από το ποσό της μνήμης που πρέπει να προσπελαστεί. Ωστόσο, και οι τρεις προαναφερθέντες παράγοντες συνδέονται μεταξύ τους και οποιαδήποτε αλλαγή στον τρόπο αναζήτησης ενδέχεται να επηρεάζει και τους τρεις.

Καθοριστικό ρόλο στη διαδικασία της αναζήτησης διαδραματίζει ο τρόπος μαθηματικής αναπαράστασης των εικόνων. Στην πλειονότητα των εφαρμογών οι εικόνες αναπαρίστανται μαθηματικά μέσω διανυσμάτων. Δηλαδή μία εικόνα αναπαρίσταται από ένα ή περισσότερα διανύσματα. Κατά αυτόν τον τρόπο γίνεται η αναπαράσταση και από τη μέθοδό μας και για αυτό το λόγο θα επικεντρωθούμε σε αυτήν την περίπτωση.

Γνωρίζοντας τον τρόπο μαθηματικής αναπαράστασης των εικόνων, οφείλουμε να προσδιορίσουμε ένα τρόπο σύγκρισής τους. Αυτό επιτυγχάνεται μέσω της χρήσης μίας συνάρτησης ομοιότητας. Στον ορισμό μίας συνάρτησης ομοιότητας σημαντικό ρόλο παίζουν παράγοντες όπως η ευκολία υπολογισμού της συνάρτησης, η ικανότητα συσχετισμού της μαθηματικής με τη σημασιολογική ομοιότητα, η ανοχή στην ύπαρξη θορύβου, η γραμμικότητα κ.α. Για παράδειγμα, στην περίπτωση που χρησιμοποιούνται διανύσματα για την περιγραφή των εικόνων, η ομοιότητα μπορεί να υπολογιστεί είτε μέσω της νόρμας L_1 που αναπαριστά την απόσταση Manhattan, είτε της νόρμας L_2 που υπολογίζει την ευκλείδεια απόσταση. Εάν τα διανύσματα αναπαριστούν ιστογράμματα είθισται η χρήση συναρτήσεων ομοιότητας όπως η histogram intersection. Σε περιπτώσεις που συγκρίνουμε σύνολα διανυσμάτων μπορεί να γίνει χρήση της απόστασης Hausdorff. Έτσι, οι διάφορες μέθοδοι σύγκρισης των μαθηματικών αναπαραστάσεων των εικόνων διαφέρουν ως προς τη μορφή της εισόδου, τη συνάρτηση ομοιότητας και την πολυπλοκότητα υπολογισμού.

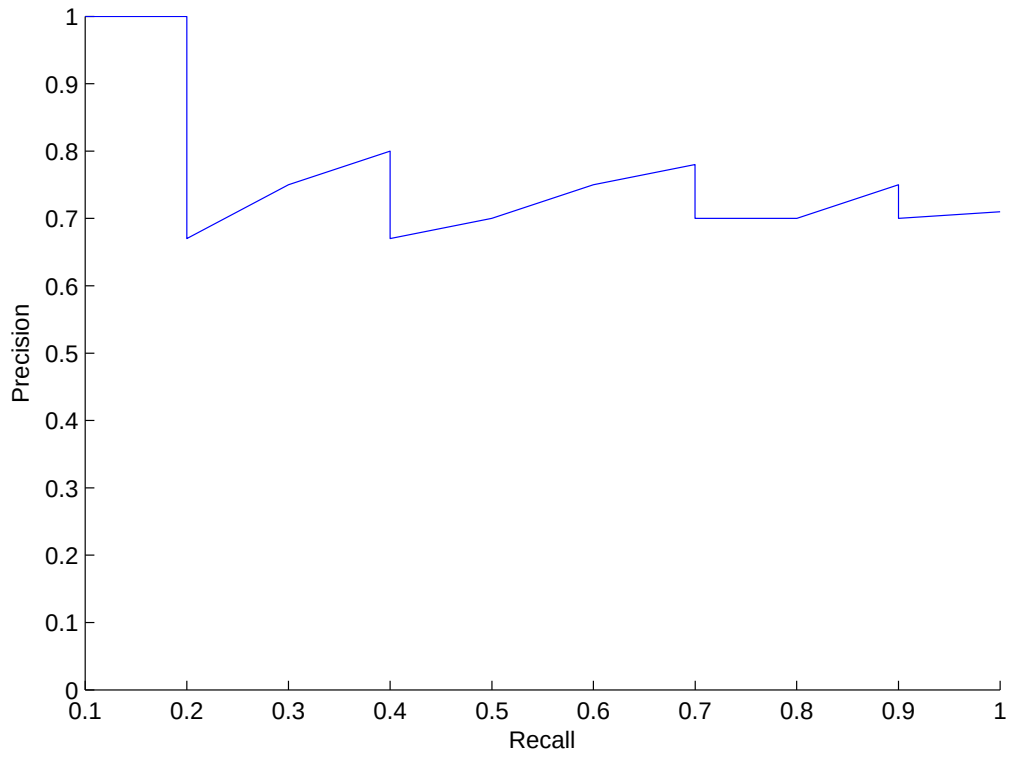
Κρίσιμο ρόλο για την ποιότητα μίας αναζήτησης παίζει το μέγεθος των διανυσμάτων που χρησιμοποιούνται καθώς επηρεάζει καθοριστικά τόσο το χρόνο επεξεργασίας τους όσο και τον απαιτούμενο χώρο αποθήκευσης. Είθισται λοιπόν να γίνεται χρήση μεθόδων μείωσης του μεγέθους των διανυσμάτων αυτών. Διανύσματα χιλιάδων διαστάσεων μπορούν να αντικατασταθούν από διανύσματα εκατοντάδων ή και δεκάδων διαστάσεων, δίχως να χάνουν μεγάλο μέρος της διακριτικής τους ικανότητας ή ακόμα και να την αυξάνουν. Μία από τις δημοφιλέστερες τεχνικές μείωσης του μεγέθους διανυσμάτων είναι η Principal Component Analysis (PCA) η οποία παρουσιάζεται σε επόμενο κεφάλαιο.

Πέραν όμως από τα προαναφερθέντα χαρακτηριστικά μίας αναζήτησης, άκρως σημαντικός είναι και ο τρόπος που οργανώνεται. Ο πιο απλός αλλά και λιγότερο αποδοτικός τρόπος είναι η εξαντλητική αναζήτηση μεταξύ των εικόνων της βάσης. Σε αυτή την περίπτωση, έχοντας ορίσει μία συνάρτηση ομοιότητας, συγκρίνουμε μέσω της συνάρτησης αυτής όλες τις εικόνες της βάσης με την εικόνα−ερώτημα. Έτσι, βρίσκουμε με απόλυτη ακρίβεια τις εικόνες που έχουν τη μεγαλύτερη ομοιότητα με την εικόνα που μας ενδιαφέρει. Ωστόσο, η εξαντλητική αναζήτηση δεν είναι εφικτή σε αρκετές περιπτώσεις λόγω του χρόνου αλλά και της μνήμης που απαιτείται καθώς, όπως έχουμε αναφέρει, το μέγεθος των βάσεων δεδομένων ενδέχεται να εκτείνεται σε εκατομμύρια εικόνες.

Προκειμένου να επιλυθεί το συγκεκριμένο πρόβλημα εφαρμόζονται προσεγγιστικές μέθοδοι οι οποίες οργανώνουν την αναζήτηση με τέτοιο τρόπο ώστε να μειώνεται είτε το κόστος είτε το πλήθος των συγκρίσεων ή και τα δύο. Σε αυτές τις περιπτώσεις παρουσιάζεται μία μικρή απώλεια ακρίβειας αλλά μεγάλη αύξηση της αποδοτικότητας και ενδεχομένως μείωση της απαιτούμενης μνήμης. Τέτοιες μέθοδοι οργανώνονται συνήθως με δένδρα αναζήτησης (search trees) ή πίνακες κατακερματισμού (hash tables). Συχνή είναι επίσης η χρήση κβαντιστών (quantizers) οι οποίοι υλοποιούνται με χρήση τεχνικών συσταδοποίησης (clustering), και έχουν ως στόχο την κωδικοποίηση των διανυσμάτων περιγραφής ώστε να διευκολύνονται οι συγκρίσεις και να απαιτείται λιγότερη μνήμη.

Στην περίπτωση που κάθε εικόνα αναπαρίσταται από ένα σύνολο διανυσμάτων περιγραφής, η αναζήτηση μπορεί να πραγματοποιηθεί με τη χρήση ενός αρχείου που ονομάζεται inverted file, το οποίο αποτελεί ιδέα η οποία έχει προέλθει από τον τρόπο σχεδιασμού των συστημάτων ανάκτησης κειμένου [37].

Ο τρόπος που μπορεί να πραγματοποιηθεί μία αναζήτηση γίνεται κατανοητό πως παρουσιάζει μεγάλη ποικιλία και οποιαδήποτε επιλογή απαιτεί πρώτα τον προσδιορισμό των απαιτήσεων σε ακρίβεια, αποδοτικότητα και χρησιμοποίηση μνήμης. Για παράδειγμα, συστήματα πραγματικού χρόνου απαιτούν μεγάλη αποδοτικότητα ώστε να παρουσιάζουν υψηλή ταχύτητα απόκρισης, και συνεπώς ενδέχεται να έχουν υψηλές απαιτήσεις μνήμης ώστε η αναζήτηση να γίνεται οργανωμένα με όσο το δυνατόν πιο αποδοτικές δομές, παράλληλα όμως πιθανόν να έχουν μειωμένη ακρίβεια λόγω πιθανής υψηλής μείωσης του αριθμού των διαστάσεων των διανυσμάτων.



Σχήμα 1.8: Παράδειγμα καμπύλης precision-recall.

1.2.4 Αξιολόγηση

Προκειμένου η ερευνητική προσπάθεια να οδηγηθεί προς την σωστή κατεύθυνση υπάρχει ανάγκη αξιολόγησης της ανάκτησης των εικόνων. Η αξιολόγηση αυτή μπορεί να επιτευχθεί με τον υπολογισμό διαφόρων δεικτών. Δύο δείκτες που χρησιμοποιούνται κατά κύριο λόγο είναι οι precision και recall. Προκειμένου να ορίσουμε τους δύο αυτούς δείκτες θα εισάγουμε την ακόλουθη ορολογία. Έστω ότι ανακτώνται n συνολικά εικόνες και υπάρχουν m εικόνες που ανήκουν στην ίδια κατηγορία με την εικόνα−ερώτημα μίας αναζήτησης. Οι εικόνες που ανήκουν στην ίδια κατηγορία με την εικόνα−ερώτημα και έχουν ανακτηθεί, ορίζονται ως σωστές θετικές (true positive) εικόνες. Οι εικόνες που ανήκουν στην ίδια κατηγορία με την εικόνα−ερώτημα αλλά δεν ανήκουν στις n εικόνες που έχουν ανακτηθεί ονομάζονται λανθασμένες αρνητικές (false negative). Οι εικόνες που έχουν ανακτηθεί αλλά δεν ανήκουν στην ίδια κατηγορία με την εικόνα−ερώτημα ονομάζονται λανθασμένες θετικές (false positive). Έστω ότι σε μία αναζήτηση υπάρχουν tp σωστές θετικές εικόνες, fn λανθασμένες αρνητικές και fp λανθασμένες θετικές εικόνες. Ο δείκτης precision ορίζεται ως εξής:

$$precision = \frac{tp}{tp + fp} = \frac{tp}{n} \tag{1.1}$$

Ο δείκτης precision λοιπόν υπολογίζει το λόγο των εικόνων που σωστά ανακτήθηκαν ως προς το σύνολο των ανακτηθέντων εικόνων. Ο δείκτης recall υπολογίζει το λόγο των εικόνων που ανακτήθηκαν σωστά ως προς το σύνολο των εικόνων που ανήκουν στην ίδια κατηγορία με

την εικόνα−ερώτημα, δηλαδή όλες εκείνες τις εικόνες που θα ήταν σωστό να ανακτηθούν. Ορίζεται από τον ακόλουθο τύπο:

$$recall = \frac{tp}{tp + fn} = \frac{tp}{m} \tag{1.2}$$

Οι δύο αυτοί δείκτες δέχονται τιμές στο διάστημα $[0, 1]$ και συνδέονται με αντίστροφη σχέση ως προς το πλήθος n των εικόνων που ανακτώνται. Καθώς ο αριθμός n αυξάνεται, το precision σταδιακά μειώνεται χωρίς να αποκλείονται προσωρινές αυξήσεις. Αντίθετα, η αύξηση του n οδηγεί σε σταδιακή αύξηση του recall. Το ζητούμενο είναι το precision να παραμένει υψηλό και το recall να αυξάνεται γρήγορα. Ένας συνήθης τρόπος οπτικοποίησης των δύο δεικτών είναι μέσω της σχεδίασής τους σε ένα κοινό διάγραμμα ως προς τον αριθμό n. Ένα τέτοιο διάγραμμα παρατίθεται στο σχήμα 1.8. Τα διαγράμματα σαν και αυτό ονομάζονται καμπύλες precision–recall.

Αδυναμία των δεικτών precision και recall είναι το γεγονός ότι πολλές φορές συναντάται μεγάλη δυσκολία στο να διαχωριστούν οι εικόνες σε δύο σύνολα, σε αυτό με τις εικόνες που είναι όμοιες με την εικόνα−ερώτημα και σε αυτό με τις εικόνες που δεν είναι. Όπως έχουμε αναφέρει και σε προηγούμενη ενότητα αυτό οφείλεται στην υποκειμενικότητα της κρίσης της σημασιολογικής ομοιότητας δύο εικόνων. Επίσης, το γεγονός ότι διαχωρίζονται δυαδικά οι εικόνες σε σωστά και λανθασμένα ανακτηθείσες έχει ως αποτέλεσμα να μην συνυπολογίζεται το γεγονός ότι οι εικόνες που κακώς ανακτώνται ενδέχεται να διαφέρουν τελείως με την εικόνα−ερώτημα αλλά ενδέχεται και να διαφέρουν οριακά. Στην δεύτερη περίπτωση, το σφάλμα ανάκτησης θα μπορούσε να θεωρηθεί πιο ‘‘ήπιο’’ και κατά συνέπεια να συνυπολογιστεί με μικρότερη βαρύτητα.

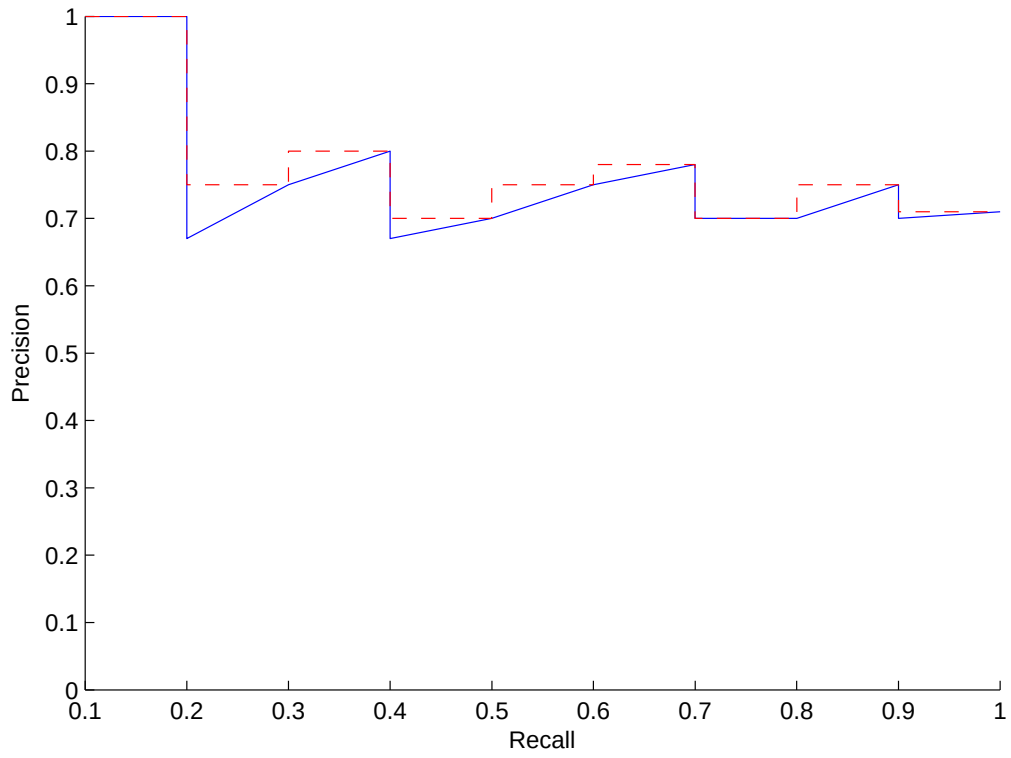
Ένας ακόμα πολύ χρήσιμος δείκτης στην ανάκτηση εικόνων είναι ο mean average precision (mAP). Για τον ορισμό του απαιτείται πρώτα να προσδιορίσουμε το μέγεθος average precision (AP). Το average precision υπολογίζεται για μία εικόνα−ερώτημα και αποτελεί το άθροισμα των τιμών του precision για όλες τις τιμές του recall. Συγκεκριμένα δίνεται από τον ακόλουθο τύπο:

$$average\ precision = \int_0^1 p(r)dr \tag{1.3}$$

όπου $p(r)$ είναι το precision για την τιμή r του recall. Η μέγιστη τιμή του δείκτη αυτού είναι 1 και μπορεί να υπολογιστεί και μέσω του εμβαδού κάτω από την καμπύλη precision recall. Στην πράξη, ο δείκτης average precision υπολογίζεται από την ακόλουθη ποσότητα:

$$\sum_{i=1}^{N} P(i)dr(i) \tag{1.4}$$

όπου N είναι το συνολικό πλήθος των εικόνων της βάσης, $P(i)$ είναι το precision για πλήθος i ανακτηθέντων εικόνων, και $dr(i)$ είναι η διαφορά του recall μεταξύ των τιμών $i - 1$ και i του πλήθους των ανακτηθέντων εικόνων. Η τιμή της παραπάνω ποσότητας προσεγγίζει την τιμή του average precision και σε συνέχεια του προηγούμενου παραδείγματος, μπορεί να υπολογιστεί γεωμετρικά μέσω του εμβαδού κάτω από τη διακεκομμένη καμπύλη στο σχήμα 1.9.



Σχήμα 1.9: Προσεγγιστικός γεωμετρικός υπολογισμός του average precision από την καμπύλη precision-recall.

Εν τέλει, ο δείκτης mean average precision υπολογίζεται ως η μέση τιμή των average precision για έναν αριθμό εικόνων−ερωτημάτων. Η μέγιστη δυνατή τιμή του mAP, όπως και του AP, είναι 1. Εάν λοιπόν έχουμε k εικόνες−ερωτήματα, και η συνάρτηση $AP(x)$ επιστρέφει την τιμή του average precision για την εικόνα−ερώτημα x, προκύπτει ο παρακάτω τύπος:

$$mean\ average\ precision = \frac{\sum_{i=1}^{k} AP(i)}{k} \tag{1.5}$$

Η επιλογή του δείκτη που θα χρησιμοποιηθεί εξαρτάται από την εφαρμογή που έχουμε και κυρίως από το μέτρο σύγκρισης που διαθέτουμε. Η σύγκριση συστημάτων μπορεί να επιτευχθεί μόνο μέσω της χρήσης των ίδιων δεικτών αξιολόγησης.

1.3 Συνεισφορά

Η πολυτιμότερη συνεισφορά της παρούσας διπλωματικής εργασίας είναι η δημιουργία μίας καινούριας μεθόδου διανυσματικής αναπαράστασης εικόνων, της Spatial Pyramid with Vectors of Locally Aggregated Descriptors (SP–VLAD). Η συγκεκριμένη μέθοδος βασίζεται στις ιδέες της χωρικής πυραμίδας και της άθροισης διανυσμάτων περιγραφής, όπως αυτές αξιοποιούνται στις μεθόδους Spatial Pyramid Matching (SPM) [25] και Vector of Locally Aggregated Descriptors (VLAD) [20] αντίστοιχα. Η μέθοδος SP–VLAD λοιπόν, επιχειρεί

να συνδυάσει τα θετικά στοιχεία των μεθόδων SPM και VLAD εξασφαλίζοντας ένα τελικό διάνυσμα αναπαράστασης της κάθε εικόνας το οποίο είναι πιο πλούσιο σε πληροφορία, προκειμένου να επιτυγχάνεται υψηλότερη ακρίβεια, ενώ παράλληλα το υπολογιστικό κόστος να παραμένει χαμηλό. Το ίδιο ισχύει και για τις απαιτήσεις μνήμης, καθώς τα διανύσματα διατηρούν εξαιρετική διακριτική ικανότητα ακόμα και μετά τη μείωση τους στις 128 ή και τις 64 διαστάσεις μέσω της μεθόδου PCA.

Η μέθοδός μας SP–VLAD, παρά το γεγονός πως σχεδιάστηκε για την εφαρμογή της στην ανάκτηση εικόνων, οι ιδιαιτέρως ικανοποιητικές επιδόσεις της μας οδήγησαν στην εφαρμογή της και στο πρόβλημα της κατηγοριοποίησης εικόνων. Και στα δύο προαναφερθέντα προβλήματα, της ανάκτησης και της κατηγοριοποίησης, η μέθοδος SP–VLAD οδήγησε σε αισθητά υψηλότερα αποτελέσματα σε σχέση με τις μεθόδους SPM και VLAD. Οι βάσεις που χρησιμοποιήθηκαν για την αξιολόγηση των μεθόδων είναι οι ευρέως διαδεδομένες στην επιστημονική κοινότητα INRIA Holidays [18] και Caltech 101 [11], καθώς και η βάση ανθοφόρων φυτών Flowers 15 την οποία δημιουργήσαμε με εικόνες από το διαδικτυακό χώρο flickr (www.flickr.com). Η συνεισφορά λοιπόν της ερευνητικής μας δραστηριότητας συνοψίζεται στα ακόλουθα σημεία:

- Σχεδιάσαμε εξ ολοκλήρου τη μέθοδο SP–VLAD και την υλοποιήσαμε στη γλώσσα προγραμματισμού C++ με χρήση των βιβλιοθηκών OpenCV [6] και Boost [5].
- Υλοποιήσαμε τις μεθόδους SPM και VLAD στη γλώσσα προγραμματισμού C++.
- Δημιουργήσαμε τη βάση εικόνων Flowers 15 με εικόνες ανθοφόρων φυτών από το flickr (www.flickr.com). Η βάση αυτή περιλαμβάνει 450 εικόνες από 15 διαφορετικά γένη φυτών όπως αυτά ταξινομούνται βάσει του συστήματος Angiosperm Phylogeny Group (APG) III [40].
- Εφαρμόσαμε πειραματικά στο πρόβλημα της ανάκτησης εικόνων τις μεθόδους SP–VLAD, VLAD και SPM χρησιμοποιώντας τις βάσεις εικόνων INRIA Holidays, Caltech 101 και Flowers 15. Η μέθοδος SP–VLAD πέτυχε υψηλότερο δείκτη mAP και στις 3 βάσεις εικόνων.
- Εφαρμόσαμε πειραματικά στο πρόβλημα της κατηγοριοποίησης εικόνων τις μεθόδους SP–VLAD, VLAD και SPM χρησιμοποιώντας τις βάσεις εικόνων INRIA Holidays, Caltech 101 και Flowers 15. Η μέθοδος SP–VLAD πέτυχε υψηλότερο βαθμό κατηγοριοποίησης και στις 3 βάσεις εικόνων.

1.4 Δομή Διπλωματικής Εργασίας

Στο κεφάλαιο 2 περιγράφεται ο τρόπος ανίχνευσης και περιγραφής των χαρακτηριστικών των εικόνων. Στο κεφάλαιο 3 παρουσιάζεται η μέθοδός μας SP–VLAD ενώ αναλύονται και οι μέθοδοι αναπαράστασης εικόνων στις οποίες βασίστηκε η δημιουργία της. Στο κεφάλαιο 4 επεξηγείται η πειραματική διαδικασία και παρουσιάζονται τα τελικά αποτελέσματα. Τέλος, στο

κεφάλαιο 5 συνοψίζονται τα συμπεράσματα της ερευνητικής μας εργασίας και παρατίθενται κάποιες σκέψεις αναφορικά με προοπτικές βελτίωσης και επέκτασής της.

Κεφάλαιο 2

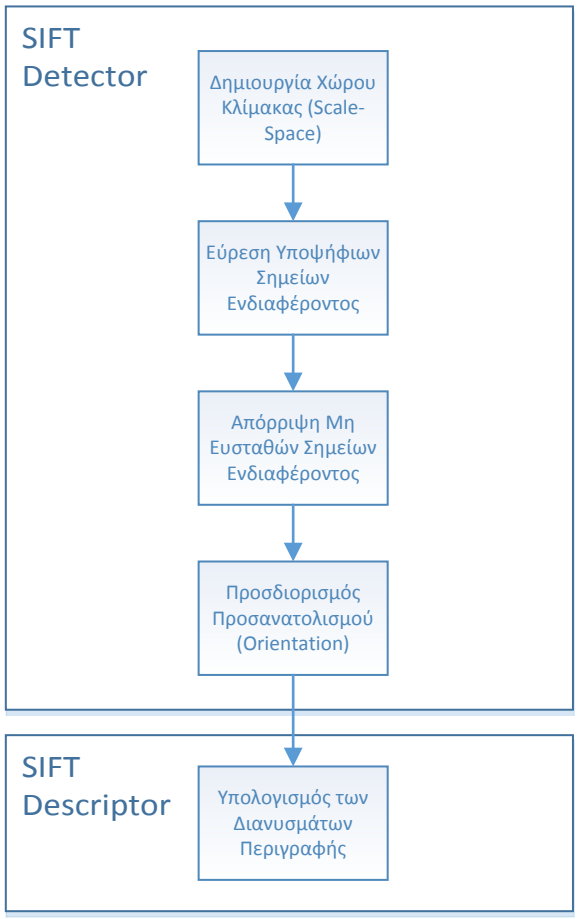
Ανίχνευση και Περιγραφή Χαρακτηριστικών

Στο κεφάλαιο αυτό περιγράφεται ο τρόπος με τον οποίο πραγματοποιείται η ανίχνευση και η περιγραφή των χαρακτηριστικών προκειμένου να χρησιμοποιηθούν από τη μέθοδό μας, SP–VLAD. Συγκεκριμένα, η προσοχή εστιάζεται στη μέθοδο Scale-Invariant Feature Transform (SIFT) [27], και στη μέθοδο της πυκνής δειγματοληψίας.

2.1 Μέθοδος Scale-Invariant Feature Transform (SIFT)

Η μέθοδος SIFT χρησιμοποιείται για την εξαγωγή ανεξάρτητων (invariant) τοπικών χαρακτηριστικών υψηλής διακριτικής ικανότητας. Η επιτυχία της μεθόδου βασίζεται στην ανεξαρτησία των χαρακτηριστικών ως προς την κλίμακα (scale) της εικόνας, και στην περιγραφή των χαρακτηριστικών μέσω τοπικών παραγώγων 1ου βαθμού, εξασφαλίζοντας υψηλή διακριτική ικανότητα με σχετικά χαμηλές απαιτήσεις σε υπολογιστικούς πόρους και μέγεθος μνήμης. Πιο συγκεκριμένα, τα χαρακτηριστικά που εξάγονται είναι ανεξάρτητα της κλίμακας (scale) και της περιστροφής (rotation) της εικόνας, καθώς και μερικώς ανεξάρτητα αναφορικά με τις αλλαγές στο φωτισμό και στη γωνία λήψης από 3D κάμερα. Επιπλέον, η μέθοδος SIFT παρέχει τη δυνατότητα εξαγωγής μεγάλου αριθμού χαρακτηριστικών, επιτρέποντας την επιτυχή αναγνώριση αντικειμένων ακόμα και σε περιπτώσεις στις οποίες η περιοχή ενδιαφέροντος της εικόνας επικαλύπτεται από άλλα αντικείμενα.

Η διαδικασία ανίχνευσης και περιγραφής των χαρακτηριστικών παρουσιάζεται μέσω του σχήματος 2.1. Στο πρώτο υπολογιστικό στάδιο δημιουργείται ένας χώρος κλίμακας. Συγκεκριμένα, δημιουργείται μία αλληλουχία εικόνων διαφορετικής κλίμακας ξεκινώντας από την αρχική εικόνα και εφαρμόζοντας διαδοχικά φίλτρα Gauss διαφορετικής διακύμανσης. Στη συνέχεια, σχηματίζεται η συνάρτηση Difference of Gaussian (DoG) και τα ακρότατά της αποτελούν τα υποψήφια σημεία ενδιαφέροντος, τα οποία αντιστοιχούν στα επιθυμητά χαρακτηριστικά. Τα τελικά σημεία ενδιαφέροντος προκύπτουν με την απόρριψη των μη ευσταθών υποψήφιων σημείων. Μη ευσταθή θεωρούνται τα σημεία που είτε βρίσκονται κοντά σε ακμές είτε παρουσιάζουν χαμηλή αντίθεση φωτεινότητας. Στο επόμενο στάδιο υπολογίζεται ένας ή



Σχήμα 2.1: Διάγραμμα ροής για τον υπολογισμό των SIFT detector και descriptor.

και περισσότεροι προσανατολισμοί για κάθε σημείο. Για τον προσδιορισμό των προσανατολισμών χρησιμοποιούνται οι κατευθύνσεις των τοπικών παραγώγων γύρω από το εκάστοτε σημείο ενδιαφέροντος. Για κάθε επόμενο υπολογισμό, τα δεδομένα των εικόνων μετασχηματίζονται με βάση την κλίμακα, τον προσανατολισμό και τη θέση των σημείων ενδιαφέροντος, επιτυγχάνοντας έτσι ανεξαρτησία ως προς αυτές τις παραμέτρους. Τα τελικά διανύσματα περιγραφής αποτελούνται από ιστογράμματα των τοπικών παραγώγων, υπολογισμένα στην περιοχή γύρω από κάθε σημείο ενδιαφέροντος στην αντίστοιχη κλίμακα.

Τα στάδια υπολογισμού της μεθόδου SIFT όπως παρουσιάζονται στο σχήμα 2.1 αναλύονται λεπτομερώς στις ενότητες που ακολουθούν.

2.1.1 Δημιουργία Χώρου Κλίμακας (Scale-Space)

Τα πραγματικά αντικείμενα, σε αντίθεση με μαθηματικές έννοιες όπως το σημείο ή η ευθεία, ενδέχεται να διαφέρουν σημαντικά ανάλογα με την κλίμακα στην οποία τα παρατηρούμε. Για παράδειγμα, οι απεικονίσεις ενός χάρτη αλλάζουν σημαντικά διαφοροποιώντας την κλίμακα. Κυμαίνονται από ολόκληρη τη γη μέχρι μία μόνο πόλη. Δηλαδή ένας άτλας θα μπορούσε να θεωρηθεί μία αναπαράσταση του κόσμου γύρω μας σε πολλαπλές κλίμακες. Η έννοια της κλίμακας λοιπόν διαδραματίζει σημαντικό ρόλο στην προσπάθεια περιγραφής των αντικειμένων. Σε ένα σύστημα ανάκτησης εικόνων όπως το δικό μας δεν είμαστε σε θέση να γνωρίζουμε από πριν την κλίμακα στην οποία απεικονίζονται οι δομές ενδιαφέροντος κάθε εικόνας. Επιπροσθέτως, στην ίδια εικόνα είναι πιθανό να περιλαμβάνονται αντικείμενα διαφορετικής κλίμακας.

Η ασάφεια που εμπεριέχεται στην έννοια της κλίμακας είναι εγγενής στις εικόνες και παρόλο που δεν μπορεί να εξαλειφθεί, είναι εφικτή η κατάλληλη διαχείρισή της. Αυτό επιτυγχάνεται ανιχνεύοντας σημεία ενδιαφέροντος σε όλες τις πιθανές κλίμακες, εξασφαλίζοντας έτσι ανεξαρτησία ως προς την κλίμακα για τα χαρακτηριστικά που εντοπίζονται. Με αυτή τη λογική προκύπτει η ανάγκη δημιουργίας ενός χώρου κλίμακας [43]. Η βασική ιδέα είναι να ενσωματωθεί η εκάστοτε αρχική εικόνα σε μία οικογένεια νέων εικόνων μίας παραμέτρου. Κάθε εικόνα της οικογένειας αυτής προκύπτει μέσω εξομάλυνσης της αρχικής εικόνας σε διαφορετικό βαθμό. Η παράμετρος αφορά το βαθμό στον οποίο έχει εξομαλυνθεί κάθε εικόνα. Διαφορετικός βαθμός εξομάλυνσης αντιστοιχεί σε διαφορετική κλίμακα. Η σταδιακή εξομάλυνση των εικόνων αντιστοιχεί σε σταδιακή μείωση της παρουσίας των λεπτομερειών που εμφανίζονται στις κλίμακες υψηλής ακρίβειας. Προκειμένου να πραγματοποιηθεί η εξομάλυνση των εικόνων, δεδομένων ορισμένων παραδοχών, ο μόνος αποδεκτός τρόπος είναι μέσω της συνέλιξής τους με συναρτήσεις Gauss μεταβλητής διακύμανσης [24, 26]. Συνεπώς, ο χώρος κλίμακας μίας εικόνας αποτελεί μία συνάρτηση μίας μεταβλητής. Έστω, $I(x, y) : \Re^N \longrightarrow \Re$ η εικόνα εισόδου, και $G(x, y, \sigma) : \Re^N \times \Re_+ \backslash \{0\} \longrightarrow \Re$ είναι η συνάρτηση Gauss δύο διαστάσεων. Η $G(x, y, \sigma)$ δίνεται από τον ακόλουθο τύπο:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{2.1}$$

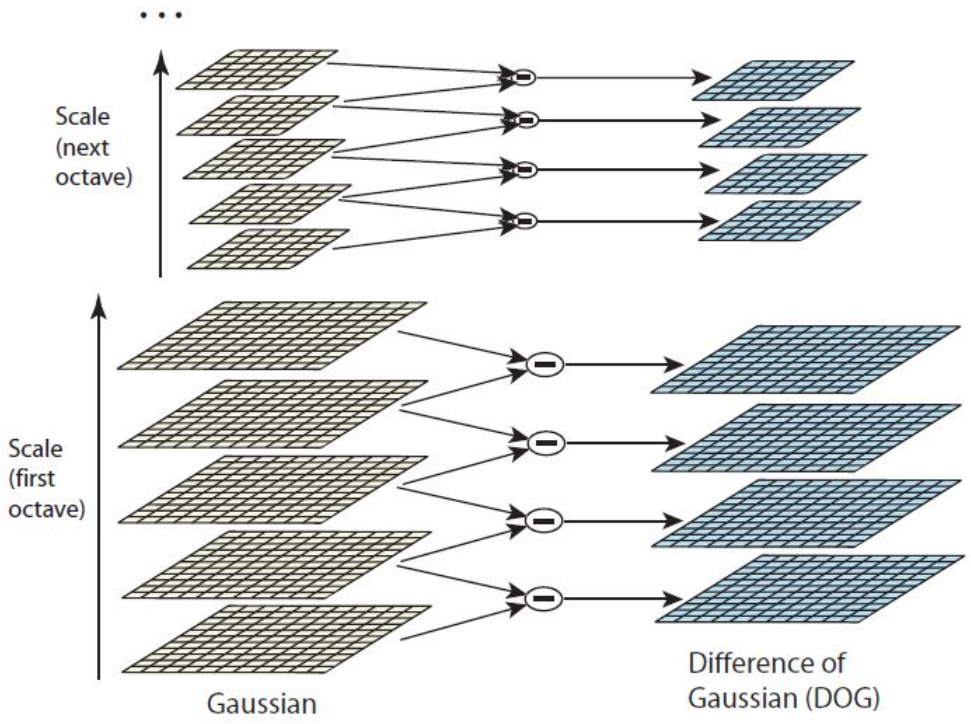
Ο χώρος κλίμακας $L(x, y, \sigma) : \Re^N \times \Re_+ \longrightarrow \Re$ ορίζεται από την παρακάτω σχέση:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2.2}$$

όπου με ‘‘$*$’’ συμβολίζεται η διαδικασία της συνέλιξης, $\sigma \in \Re_+$ είναι η παράμετρος κλίμακας, και $L(x, y, 0) = I(x, y)$.

Διαισθητικά, το αποτέλεσμα της συνέλιξης μίας εικόνας με τη συνάρτηση Gauss τυπικής απόκλισης σ είναι μία εικόνα όπου οι δομές αντικειμένων μικρότερου μεγέθους από σ έχουν εξαλειφθεί.

Όπως προαναφέρθηκε, η μόνη κατάλληλη συνάρτηση για τη δημιουργία του χώρου κλίμακας είναι η συνάρτηση Gauss. Αυτό ισχύει έχοντας ως αξίωμα πως η συνάρτηση του χώρου κλίμακας, $L(x, y, \sigma)$, πρέπει να είναι αιτιατή, ισοτροπική και ομοιογενής. Κατά αυτό τον τρόπο



Σχήμα 2.2: Σχηματική αναπαράσταση του χώρου κλίμακας και της συνάρτησης Difference of Gaussian [27].

εξασφάλιζεται η μη δημιουργία καινούριων ακρότατων όσο αυξάνεται η παράμετρος κλίμακας σ, και επίσης εξασφαλίζεται πως όλα τα σημεία των εικόνων σε κάθε κλίμακα αντιμετωπίζονται με τον ίδιο τρόπο.

2.1.2 Εύρεση Υποψήφιων Σημείων Ενδιαφέροντος

Τα χαρακτηριστικά που θέλουμε να εντοπίσουμε στις εικόνες αφορούν πλούσια τοπική πληροφορία οποιουδήποτε είδους. Δηλαδή η περιοχή της εικόνας γύρω από ένα σημείο ενδιαφέροντος οφείλει να είναι πλούσια σε πληροφορία αναφορικά με τις τοπικές δομές της εικόνας. Τέτοιες μπορούν να θεωρηθούν για παράδειγμα οι ακμές και οι γωνίες των αντικειμένων ή ακόμα και η υφή τους. Στόχος είναι η πληροφορία αυτή να ενισχύει όσο το δυνατόν περισσότερο τη διάκριση μεταξύ διαφορετικών χαρακτηριστικών και φυσικά την ομοιότητα μεταξύ χαρακτηριστικών της ίδιας κατηγορίας.

Προκειμένου να εντοπίσουμε τα επιθυμητά σημεία ενδιαφέροντος στρεφόμαστε στην ανίχνευση περιοχών του χώρου κλίμακας με έντονη διακύμανση στη φωτεινότητα. Αυτό επιτυγχάνεται μέσω της εύρεσης των ακρότατων της συνάρτησης Difference of Gaussian (DoG). Η συνάρτηση DoG προκύπτει από την αφαίρεση περιοχών του χώρου κλίμακας μεταξύ των οποίων η μεταβλητή κλίμακας σ διαφέρει κατά ένα σταθερό παράγοντα k. Έστω $DG(x, y, \sigma) : \Re^N \times \Re_+ \longrightarrow \Re$ είναι η συνάρτηση DoG. Εάν διατηρήσουμε το συμβολισμό που

είχε εισαχθεί πρωτύτερα στο παρόν κεφάλαιο, καταλήγουμε πως:

$$DG(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \tag{2.3}$$

Ο τρόπος δημιουργίας της συνάρτησης Difference of Gaussian αναπαρίσταται στο σχήμα 2.2. Η αρχική εικόνα συνελίσσεται σταδιακά με φίλτρα Gauss αυξανόμενης διακύμανσης σχηματίζοντας ένα σύνολο εξομαλυμένων εικόνων που διαχωρίζονται από τον παράγοντα k στο χώρο κλίμακας. Το συγκεκριμένο σύνολο εικόνων απεικονίζεται στο αριστερό τμήμα του σχήματος 2.2. Στη συνέχεια, διαδοχικές εικόνες αφαιρούνται μεταξύ τους παράγοντας τη συνάρτηση Difference of Gaussian, η οποία απεικονίζεται στο δεξί μέρος του σχήματος 2.2.

Για τη μείωση του υπολογιστικού κόστους, οι εξομαλυμένες εικόνες διαχωρίζονται σε οκτάβες. Συγκεκριμένα, κάθε φορά που η παράμετρος σ διπλασιάζεται, η εκάστοτε εικόνα υποδειγματοληπτείται διατηρώντας ένα ανά δύο εικονοστοιχεία. Κατά αυτό τον τρόπο, το υπολογιστικό κόστος για κάθε επόμενη οκτάβα μειώνεται σημαντικά, ενώ η ακρίβεια της δειγματοληψίας ως προς την παράμετρο κλίμακας σ παραμένει σταθερή για κάθε οκτάβα. Κάθε οκτάβα διαχωρίζεται σε σταθερό αριθμό διαστημάτων s. Συνεπάγεται λοιπόν πως $k = 2^{1/s}$. Επίσης, επιβάλλεται ο σχηματισμός $s + 3$ εικόνων για κάθε οκτάβα προκειμένου η συνάρτηση DoG να είναι πλήρης.

Ο λόγος που επιλέγεται η χρησιμοποίηση των ακρότατων της συνάρτησης DoG για την εύρεση των υποψήφιων σημείων ενδιαφέροντος είναι διπλός. Αρχικά, σημαντικό πλεονέκτημα για τη χρήση της συνάρτησης DoG αποτελεί η χαμηλή πολυπλοκότητα υπολογισμού της. Συγκεκριμένα, ο χώρος κλίμακας υπολογίζεται ούτως ή άλλως και συνεπώς η παραγωγή της συνάρτησης DoG απαιτεί μόνο την πραγματοποίηση αφαιρέσεων μεταξύ εικόνων. Ο δεύτερος λόγος που επιλέγεται η συνάρτηση DoG για την εύρεση των υποψήφιων σημείων ενδιαφέροντος αφορά το γεγονός πως αποτελεί ικανοποιητική προσέγγιση της κανονικοποιημένης συνάρτησης Laplacian of Gaussian (LoG) [26]. Η σχέση μεταξύ των συναρτήσεων DoG και LoG παρουσιάζει ιδιαίτερη σημασία διότι έχει αποδειχθεί πειραματικά πως τα ακρότατα της κανονικοποιημένης συνάρτησης Laplacian of Gaussian παρουσιάζουν την πιο ικανοποιητική ευστάθεια σε σχέση με τα σημεία που προκύπτουν από συναρτήσεις όπως η Hessian ή ο ανιχνευτής γωνιών Harris [29].

Έστω $LG(x, y, \sigma) : \Re^N \times \Re_+ \backslash \{0\} \longrightarrow \Re$ είναι η κανονικοποιημένη συνάρτηση LoG. Η κανονικοποίηση της LoG επιτελείται μέσω στάθμισης με τον παράγοντα σ^2. Η $LG(x, y, \sigma)$ δίνετε από τον ακόλουθο τύπο:

$$LG(x, y, \sigma) = \sigma^2 \nabla^2 G * I(x, y) \tag{2.4}$$

Η προσεγγιστική σχέση μεταξύ της συνάρτησης Difference of Gaussian και της κανονικοποιημένης Laplacian of Gaussian γίνεται κατανοητή μέσω της εξίσωσης διάχυσης θερμότητας (heat diffusion equation). Η συνάρτηση Gauss αποτελεί τη συνάρτηση Green για την εξίσωση διάχυσης θερμότητας και συνεπώς την ικανοποιεί:

$$\frac{\partial G}{\partial \sigma} = \sigma \nabla^2 G \tag{2.5}$$

Από την παραπάνω σχέση προκύπτει:

$$\sigma\nabla^2 G = \frac{\partial G}{\partial \sigma} \approx \frac{G(x, y, k\sigma) - G(x, y, \sigma)}{k\sigma - \sigma} \iff$$

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k-1)\sigma^2\nabla^2 G \tag{2.6}$$

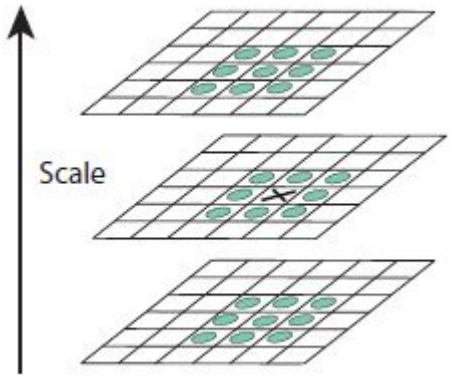
Συνεπάγεται λοιπόν πως η συνάρτηση DoG μπορεί πράγματι να χρησιμοποιηθεί ως προσέγγιση της LoG για τον προσδιορισμό των επιθυμητών υποψήφιων σημείων ενδιαφέροντος. Ο όρος $(k-1)$ είναι σταθερός για όλες τις κλίμακες και συνεπώς δεν επηρεάζει τον υπολογισμό των ακρότατων της $DG(x, y, \sigma)$. Το σφάλμα προσέγγισης τείνει στο 0 όταν το k τείνει στο 1. Αξίζει να σημειώσουμε πως η κανονικοποίηση της Laplacian of Gaussian με τον παράγοντα σ^2 γίνεται για να εξασφαλιστεί πραγματική ανεξαρτησία των σημείων ενδιαφέροντος ως προς την κλίμακα.

Έχοντας λοιπόν σχηματίσει τη συνάρτηση Difference of Gaussian, επιθυμούμε την εύρεση των ακρότατων σημείων της. Ο τρόπος εύρεσής τους απεικονίζεται στο σχήμα 2.3. Συγκεκριμένα, κάθε σημείο συγκρίνεται με τα οχτώ γειτονικά του σημεία στην κλίμακα που εξετάζουμε, αλλά και με τους εννέα γείτονές του τόσο στην επόμενη όσο και στην προηγούμενη κλίμακα. Στο σχήμα 2.3, το εικονοστοιχείο που εξετάζουμε είναι σημειωμένο με X και τα 26 εικονοστοιχεία με τα οποία πραγματοποιούνται οι συγκρίσεις στις διαδοχικές κλίμακες σημειώνονται με τους έγχρωμους κύκλους. Σημεία ενδιαφέροντος θεωρούνται όσα σημεία είναι μεγαλύτερα ή μικρότερα από όλους τους 26 γείτονές τους.

Η μέθοδος αυτή παρουσιάζει χαμηλό υπολογιστικό κόστος. Πιο συγκεκριμένα, το χαμηλό υπολογιστικό κόστος συνδέεται αφενός με το γεγονός πως η πράξη της σύγκρισης πραγματοποιείται εξαιρετικά αποδοτικά, και αφετέρου με το γεγονός πως για τα περισσότερα σημεία αρκούν πολύ λιγότερες από 26 συγκρίσεις προκειμένου να τα απορρίψουμε. Αρκεί να βρεθεί ένας γείτονας με μικρότερη τιμή και ένας με μεγαλύτερη, πράγμα που συνήθως επιτυγχάνεται μετά από λίγες μόνο συγκρίσεις.

Αυτό που απομένει να προσδιοριστεί είναι η δειγματοληψία στον χώρο της εικόνας και της κλίμακας. Συγκεκριμένα, η δειγματοληψία στο χώρο κλίμακας αναφέρεται στον αριθμό των διαστημάτων s κάθε οκτάβας, και η δειγματοληψία στο χώρο της εικόνας προσδιορίζεται από τον βαθμό εξομάλυνσης της αρχικής εικόνας κάθε οκτάβας. Δυστυχώς δεν υπάρχουν τιμές για τους προαναφερθέντες ρυθμούς δειγματοληψίας που θα επέτρεπαν την πλήρη ανίχνευση όλων των επιθυμητών σημείων ενδιαφέροντος. Συνεπάγεται λοιπόν πως πρέπει να υπάρξει ισορροπία μεταξύ πληρότητας και αποδοτικής χρήσης των υπολογιστικών πόρων. Η πληρότητα μέσω της αύξησης των ρυθμών δειγματοληψίας που προαναφέρθηκαν είναι επιθυμητή, ωστόσο ακόμα και σε αυτήν την περίπτωση, ελλοχεύει ο κίνδυνος ανίχνευσης ασταθών σημείων. Τα σημεία που βρίσκονται αρκετά κοντά χωρικά παρουσιάζουν έλλειμμα ευστάθειας ως προς μικρές μεταβολές που ενδέχεται να εμφανιστούν μεταξύ των εικόνων. Συνεπώς, οι υψηλοί ρυθμοί δειγματοληψίας, με δεδομένο και το αυξανόμενο υπολογιστικό κόστος, ενδέχεται να παρουσιάζουν μη ικανοποιητική ανταποδοτικότητα.

Συνυπολογίζοντας όλα τα προαναφερθέντα, το συμπέρασμα είναι πως η DoG διαθέτει μεγάλο πλήθος τοπικών ακρότατων και είναι αδύνατο να ανιχνευτούν όλα, ανεξάρτητα από το πόσο



Σχήμα 2.3: Σχηματική αναπαράσταση του τρόπου εύρεσης των ακρότατων της συνάρτησης Difference of Gaussian τα οποία αποτελούν τα υποψήφια σημεία ενδιαφέροντος [27].

πυκνή είναι η δειγματοληψία στο χώρο τόσο της εικόνας όσο και της κλίμακας. Η προσοχή λοιπόν στρέφεται στην εύρεση των πιο ευσταθών υποψήφιων σημείων. Αυτό πραγματοποιείται επιλέγοντας σχετικά μικρό αριθμό διαστημάτων s για κάθε οκτάβα και εξομαλύνοντας την αρχική εικόνα κάθε οκτάβας με ένα φίλτρο Gauss σταθερής διακύμανσης. Ο αριθμός s και ο βαθμός εξομάλυνσης προσδιορίζονται πειραματικά.

Σχετικά με τον αριθμό των διαστημάτων ανά οκτάβα, επιλέγεται συνήθως μία τιμή μεταξύ $s = 3$ και $s = 5$ [27]. Αυξάνοντας περαιτέρω το πλήθος των διαστημάτων, αυξάνεται και το πλήθος των υποψήφιων σημείων ενδιαφέροντος που εντοπίζονται. Η αύξηση του πλήθους των σημείων είναι επιθυμητή, αλλά προστίθενται πολλά μη ευσταθή σημεία, γεγονός που σίγουρα είναι απευκταίο. Επίσης, μεγαλύτερο s αντιστοιχεί σε μεγαλύτερο υπολογιστικό κόστος.

Το φίλτρο Gauss που χρησιμοποιείται για την εξομάλυνση της πρώτης εικόνας κάθε οκτάβας ορίζεται στις περισσότερες εφαρμογές με τυπική απόκλιση $\sigma = 1.6$. Επιπλέον, προκειμένου να αυξηθεί ο αριθμός των ευσταθών σημείων που ανιχνεύονται, η αρχική εικόνα διπλασιάζεται σε μέγεθος πριν τη χρησιμοποίησή της για τη δημιουργία της πρώτης οκτάβας. Ο διπλασιασμός του μεγέθους της αρχικής εικόνας πραγματοποιείται μέσω γραμμικής παρεμβολής λόγω της δυνατότητας αποδοτικής υλοποίησης της συγκεκριμένης μεθόδου. Γίνεται η υπόθεση πως η αρχική εικόνα είναι ήδη εξομαλυμένη τουλάχιστον κατά $\sigma = 0.5$ ώστε να αποφευχθεί το φαινόμενο της αναδίπλωσης (aliasing). Έπεται πως μετά τον διπλασιασμό της διαθέτει $\sigma = 1.0$ σε σχέση με τις νέες αποστάσεις μεταξύ των εικονοστοιχείων της. Συνεπώς, για την αρχική εικόνα της πρώτης οκτάβας απαιτείται μικρή επιπρόσθετη εξομάλυνση. Ο διπλασιασμός αυτός της αρχικής εικόνας σε συνδυασμό με την εξομάλυνση των αρχικών εικόνων κάθε οκτάβας παρατηρήθηκε πειραματικά πως τετραπλασιάζει τον αριθμό των ευσταθών σημείων που ανιχνεύονται [27].

2.1.3 Απόρριψη Μη Ευσταθών Σημείων Ενδιαφέροντος

Ένα κρίσιμο χαρακτηριστικό των σημείων ενδιαφέροντος είναι η ευστάθεια. Η υψηλή ευστάθεια επιτρέπει τον κατ΄ επανάληψη εντοπισμό των ίδιων σημείων ενδιαφέροντος ακόμα και μετά από την εφαρμογή τοπικών ή καθολικών μεταβολών στο χώρο της εικόνας. Για παράδειγμα, μία καθολική αλλαγή θα μπορούσε να θεωρηθεί η προσθήκη θορύβου στην εικόνα. Έχοντας λοιπόν εντοπίσει τα υποψήφια σημεία ενδιαφέροντος μέσω των ακρότατων της συνάρτησης DoG, ο στόχος είναι να απορριφθούν όσα σημεία κρίνονται μη ευσταθή. Δύο χαρακτηριστικά που ενδέχεται να εμφανίζουν τα μη ευσταθή σημεία είναι η μικρή αντίθεση φωτεινότητας και η λανθασμένη ένταξη τους κατά μήκος μίας ακμής. Για κάθε μία από τις δύο αυτές περιπτώσεις χρησιμοποιούνται συγκεκριμένες τεχνικές εντοπισμού των ασταθών σημείων, οδηγώντας βέβαια στην απόρριψή τους.

Η αντίθεση φωτεινότητας που εμφανίζει ένα σημείο προσδιορίζεται από την τιμή της συνάρτησης Difference of Gaussian στο σημείο αυτό. Αποδεκτές είναι οι τιμές της DoG που είναι είτε υψηλότερες είτε χαμηλότερες από ένα κατώφλι. Θεωρητικά, η θέση του κάθε υποψήφιου σημείου ενδιαφέροντος είναι γνωστή και συνεπώς ο υπολογισμός της τιμής της συνάρτησης DoG θα μπορούσε να θεωρηθεί ιδιαιτέρως απλός. Ωστόσο, προκειμένου να αυξηθεί η ακρίβεια του υπολογισμού, σε κάθε υποψήφιο σημείο ενδιαφέροντος εφαρμόζεται ένα λεπτομερές 3D μοντέλο για τον ακριβή προσδιορισμό της θέσης και της κλίμακάς του [7]. Συγκεκριμένα, γίνεται χρήση του αναπτύγματος Taylor της συνάρτησης Difference of Gaussian μέχρι τον τετραγωνικό όρο. Το σημείο γύρω από το οποίο υπολογίζεται το ανάπτυγμα Taylor είναι το εκάστοτε υποψήφιο σημείο ενδιαφέροντος. Έχουμε λοιπόν την ακόλουθη σχέση:

$$DG(\mathrm{x}) = DG + \frac{\partial DG^T}{\partial \mathrm{x}}\mathrm{x} + \frac{1}{2}\mathrm{x}^{\mathrm{T}}\frac{\partial^2 DG}{\partial \mathrm{x}^2}\mathrm{x} \tag{2.7}$$

όπου DG η συνάρτηση DoG, και $\mathrm{x} = (x, y, \sigma)^T$ η απόκλιση από τη θέση του σημείου ενδιαφέροντος που μελετάμε. Η συνάρτηση DoG και οι παράγωγοί της υπολογίζονται στο σημείο ενδιαφέροντος το οποίο εξετάζουμε. Προκειμένου να προσδιορίσουμε την τιμή της μεταβολής x, υπολογίζουμε το τοπικό ακρότατο $\widehat{\mathrm{x}}$ της συνάρτησης DG, απαιτώντας η πρώτη παράγωγος του αναπτύγματος Taylor, το οποίο παρουσιάστηκε στη σχέση 2.7, να είναι ίση με το μηδέν. Το τοπικό ακρότατο $\widehat{\mathrm{x}}$ λοιπόν, θα δίνεται από την ακόλουθη σχέση:

$$\widehat{\mathrm{x}} = -\frac{\partial^2 DG^{-1}}{\partial \mathrm{x}^2}\frac{\partial DG}{\partial \mathrm{x}} \tag{2.8}$$

Η σχέση 2.8 αποτελεί ένα γραμμικό σύστημα 3×3 το οποίο μπορεί να επιλυθεί με πολύ μικρό υπολογιστικό κόστος. Έχοντας υπολογίσει το $\widehat{\mathrm{x}}$, εξετάζουμε την τιμή του. Συγκεκριμένα, εάν το $\widehat{\mathrm{x}}$ είναι μεγαλύτερο του 0.5 σε οποιαδήποτε διάσταση, τότε συμπεραίνουμε πως το ακρότατο της DoG βρίσκεται πιο κοντά σε άλλο σημείο και όχι στο σημείο ενδιαφέροντος το οποίο εξετάζουμε. Συνεπώς, το υποψήφιο σημείο ενδιαφέροντος αντικαθίσταται από το κοντινότερό του με βάση την τιμή του $\widehat{\mathrm{x}}$ και η όλη διαδικασία της παρεμβολής επαναλαμβάνεται μέχρι το $\widehat{\mathrm{x}}$ να είναι μικρότερο ή ίσο του 0.5 προς όλες τις διαστάσεις. Μόλις εξασφαλιστεί αυτό, η τιμή $\widehat{\mathrm{x}}$ προστίθεται στη θέση του υποψήφιου σημείου ενδιαφέροντος ολοκληρώνοντας έτσι τον

υψηλής ακρίβειας προσδιορισμό της θέσης του. Καθίσταται πλέον δυνατός ο υπολογισμός της τιμής της DoG στο σημείο που εξετάζουμε αντικαθιστώντας τη σχέση 2.8 στη 2.7:

$$DG(\hat{\mathrm{x}}) = DG + \frac{1}{2}\frac{\partial DG^T}{\partial \mathrm{x}}\hat{\mathrm{x}} \tag{2.9}$$

Κάθε υποψήφιο σημείο ενδιαφέροντος για το οποίο ισχύει $|DG(\hat{\mathrm{x}})| > h$ απορρίπτεται. Με δεδομένο πως οι τιμές φωτεινότητας των εικονοστοιχείων ανήκουν στο διάστημα $[0, 1]$, στο κατώφλι h αποδίδονται τιμές από το ίδιο διάστημα. Ενδεικτικές συνήθεις επιλογές είναι 0.03 και 0.04. Υψηλότερη τιμή του κατωφλίου συνεπάγεται απόρριψη μεγαλύτερου αριθμού σημείων.

Η δεύτερη κατηγορία μη ευσταθών σημείων ενδιαφέροντος που φιλοδοξούμε να απορρίψουμε συγκεντρώνει όλα εκείνα τα σημεία που ενώ παρουσιάζουν υψηλή αντίθεση φωτεινότητας επειδή βρίσκονται κατά μήκος μίας ακμής, η θέση τους κρίνεται επισφαλής. Για παράδειγμα, η εμφάνιση θορύβου ενδέχεται να τα αποσταθεροποιήσει, και να τεθούν εκτός ακμής. Οι ακμές είναι από τις πιο θεμελιώδεις δομές που παρουσιάζουν έντονη διακύμανση στη φωτεινότητα και η σωστή ανίχνευση των σημείων τους είναι επιβεβλημένη. Για την απόρριψη των σημείων αυτής της κατηγορίας γίνεται χρήση της έννοιας της κύριας καμπυλότητας (principal curvature). Συγκεκριμένα, ένα ακρότατο της DoG το οποίο δεν είναι καλώς ορισμένο, αναμένεται να παρουσιάζει υψηλή κύρια καμπυλότητα κατά μήκος της ακμής στην οποία εντάσσεται, και χαμηλή τιμή κύριας καμπυλότητας ως προς την κάθετη κατεύθυνση στην ακμή. Ένα επιπλέον χαρακτηριστικό για τις καμπυλότητες που προαναφέρθηκαν είναι ότι πρέπει να έχουν το ίδιο πρόσημο. Συνεπώς, υπολογίζουμε το λόγο της υψηλής καμπυλότητας ως προς την χαμηλή, ώστε να αποφανθούμε για την ευστάθεια του αντίστοιχου σημείου ενδιαφέροντος. Όταν η τιμή του λόγου αυτού ξεπερνάει ένα κατώφλι, το υποψήφιο σημείο ενδιαφέροντος απορρίπτεται. Προκειμένου να απλοποιήσουμε τους υπολογισμούς, εκμεταλλευόμαστε το γεγονός πως οι καμπυλότητες της DG είναι ανάλογες των ιδιοτιμών του ακόλουθου 2×2 Hessian πίνακα:

$$\mathrm{H} = \begin{bmatrix} DG_{xx} & DG_{xy} \\ DG_{xy} & DG_{yy} \end{bmatrix} \tag{2.10}$$

όπου H ο Hessian πίνακας. Αρκεί λοιπόν ο υπολογισμός του λόγου r των ιδιοτιμών του H [15]. Έστω, α η ιδιοτιμή του H με το μεγαλύτερο μέτρο, και β η ιδιοτιμή με το μικρότερο. Για τον λόγο r θα ισχύει $\alpha = r\beta$. Προκείμενου να αποφύγουμε περιττούς υπολογισμούς, θεωρούμε τις ακόλουθες ποσότητες:

$$\mathrm{Tr}(\mathrm{H}) = DG_{xx} + DG_{yy} = \alpha + \beta \tag{2.11}$$

$$\mathrm{Det}(\mathrm{H}) = DG_{xx}DG_{yy} - (DG_{xy})^2 = \alpha\beta \tag{2.12}$$

$$\frac{\mathrm{Tr}(\mathrm{H})^2}{\mathrm{Det}(\mathrm{H})} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r} \tag{2.13}$$

όπου Tr(H) είναι το ίχνος του Hessian πίνακα και Det(H) η ορίζουσά του. Όπως μπορεί εύκολα να διαπιστωθεί από τη σχέση 2.13, η ποσότητα $\frac{(r+1)^2}{r}$ παρουσιάζει ελάχιστη τιμή όταν

$\alpha = \beta$ και είναι αύξουσα ως προς το r. Συνεπώς, αρκεί να θέσουμε ένα κατώφλι για τον λόγο r και να αξιολογήσουμε την ακόλουθη σχέση:

$$\frac{\mathrm{Tr(H)}^2}{\mathrm{Det(H)}} < \frac{(r+1)^2}{r} \tag{2.14}$$

Ο έλεγχος της σχέσης αυτής έχει εξαιρετικά χαμηλό υπολογιστικό κόστος και τα υποψήφια σημεία ενδιαφέροντος τα οποία δεν την επιβεβαιώνουν, απορρίπτονται. Μία ενδεικτική τιμή για το κατώφλι είναι $r = 10$. Κατά αυτόν τον τρόπο απορρίπτονται όσα υποψήφια σημεία ενδιαφέροντος έχουν λόγο μεταξύ των δύο principal curvatures μεγαλύτερο του 10. Αυξάνοντας την τιμή του r απορρίπτονται λιγότερα σημεία ενδιαφέροντος. Επιπροσθέτως, αξίζει να σημειώσουμε πως εάν η ορίζουσα, η οποία υπολογίζεται από τη σχέση 2.12, έχει αρνητικό πρόσημο, τότε οι κύριες καμπυλότητες του υποψήφιου σημείου ενδιαφέροντος έχουν αντίθετα πρόσημα και επομένως το σημείο αυτό απορρίπτεται αυτόματα.

Οφείλουμε να παρατηρήσουμε πως για την αποτίμηση των σχέσεων 2.8 και 2.10, οι παράγωγοι της DG υπολογίζονται προσεγγιστικά μέσω διαφορών μεταξύ γειτονικών σημείων.

2.1.4 Προσδιορισμός Προσανατολισμού (Orientation)

Ο προσανατολισμός ενός σημείου ενδιαφέροντος ορίζεται από την επικρατέστερη κατεύθυνση των τοπικών παραγώγων των σημείων που ανήκουν στη γειτονιά του. Η έννοια της επικρατέστερης κατεύθυνσης είναι ποσοτική και σχετίζεται με το πλήθος των σημείων της προαναφερθείσας γειτονίας που διαθέτουν παραγώγους με κατευθύνσεις σε ένα συγκεκριμένο εύρος μοιρών. Ο λόγος που υπολογίζουμε τον προσανατολισμό των σημείων ενδιαφέροντος είναι για να μπορούν να εκφραστούν ως προς αυτόν τα διανύσματα περιγραφής και κατά συνέπεια να εξασφαλίζεται ανεξαρτησία αναφορικά με τον περιστροφή της εικόνας.

Προκειμένου λοιπόν να υπολογίσουμε τον προσανατολισμό ενός σημείου ενδιαφέροντος, αρχικά υπολογίζουμε το μέτρο και την κατεύθυνση των τοπικών παραγώγων όλων των σημείων που ανήκουν στην εικόνα του χώρου κλίμακας η οποία έχει την ίδια τιμή της παραμέτρου κλίμακας σ με την κλίμακα του σημείου ενδιαφέροντος. Συγκεκριμένα:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{2.15}$$

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))) \tag{2.16}$$

όπου $m(x,y)$ είναι το μέτρο και $\theta(x,y)$ ο προσανατολισμός της τοπικής παραγώγου σε κάθε σημείο της εξομαλυμένης εικόνας $L(x,y,\sigma)$ του χώρου κλίμακας. Όπως είναι εμφανές, οι υπολογισμοί βασίζονται σε διαφορές μεταξύ των pixels. Σημειώνουμε πως η μεταβλητή σ της εξομαλυμένης εικόνας $L(x,y,\sigma)$ έχει την ίδια τιμή με την κλίμακα του σημείου ενδιαφέροντος προκειμένου να εξασφαλιστεί ανεξαρτησία ως προς την κλίμακα.

Έχοντας ολοκληρώσει τους παραπάνω υπολογισμούς, σχηματίζουμε ένα ιστόγραμμα με βάση τις κατευθύνσεις των τοπικών παραγώγων. Συγκεκριμένα, το ιστόγραμμα χωρίζεται σε 36 διαστήματα καλύπτοντας το εύρος τον 360 μοιρών. Προκειμένου να ορίσουμε τη γειτονία του σημείου ενδιαφέροντος η οποία λαμβάνεται υπόψιν στο σχηματισμό του ιστογράμματος,

κάθε δείγμα που προστίθεται στο ιστόγραμμα πρώτα σταθμίζεται από ένα κυκλικό παράθυρο Gauss με τυπική απόκλιση 1.5 φορές μεγαλύτερη από την κλίμακα του σημείου ενδιαφέροντος. Επίσης, κάθε δείγμα του ιστογράμματος σταθμίζεται με το μέτρο της παραγώγου. Η μέγιστη συχνότητα του ιστογράμματος αποτελεί την επικρατέστερη κατεύθυνση. Επιπλέον, για οποιαδήποτε άλλη συχνότητα η οποία έχει τιμή που ισούται ή υπερβαίνει το 80% της μέγιστης συχνότητας, δημιουργείται ξεχωριστό σημείο ενδιαφέροντος. Όλα αυτά τα σημεία ενδιαφέροντος έχουν την ίδια θέση και κλίμακα και διαφέρουν μόνο ως προς τον προσανατολισμό. Τέλος, να σημειώσουμε πως η ακριβής τιμή της κάθε συχνότητας που εξετάζουμε προσεγγίζεται μέσω παρεμβολής. Συγκεκριμένα, μία παραβολή εφαρμόζεται μεταξύ των τριών κοντινότερων τιμών του ιστογράμματος στην εκάστοτε συχνότητα.

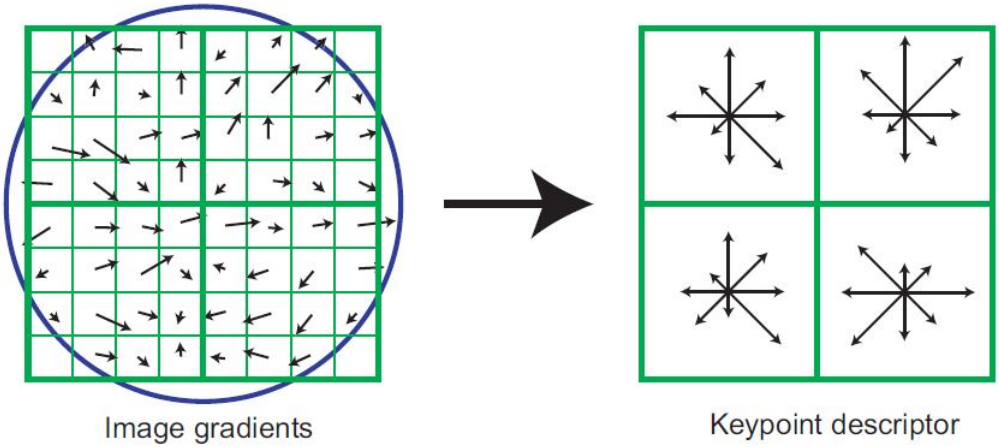
2.1.5 Υπολογισμός των Διανυσμάτων Περιγραφής

Η περιγραφή των σημείων ενδιαφέροντος βασίζεται στην τοπική πληροφορία της εικόνας και πραγματοποιείται με τρόπο που εξασφαλίζεται αφενός η υψηλή διακριτική ικανότητα των διανυσμάτων περιγραφής και αφετέρου η ανεξαρτησία ως προς την κλίμακα, την περιστροφή και τις μεταβολές στο φωτισμό.

Ο τρόπος υπολογισμού των διανυσμάτων περιγραφής αναπαρίσταται στο σχήμα 2.4 και θα χρησιμοποιηθεί σαν σημείο αναφοράς στη συνέχεια. Η περιγραφή ενός σημείου ενδιαφέροντος βασίζεται στην έννοια της γειτονιάς, δηλαδή των σημείων που βρίσκονται γύρω του. Για τον προσδιορισμό της γειτονιάς αυτής, θεωρούμε την εικόνα του χώρου κλίμακας η οποία έχει την ίδια τιμή κλίμακας με το σημείο ενδιαφέροντος που εξετάζουμε. Για κάθε εικονοστοιχείο της γειτονίας του σημείου ενδιαφέροντος υπολογίζεται το μέτρο και η κατεύθυνση της παραγώγου σύμφωνα με τις σχέσεις 2.15 και 2.16. Το μέτρο και η κατεύθυνση της παραγώγου σε κάθε σημείο αναπαρίστανται με ένα βέλος κατάλληλου μήκους και προσανατολισμού όπως μπορεί να παρατηρηθεί στο αριστερό μέρος της εικόνας 2.4. Προκειμένου να εξασφαλιστεί η ανεξαρτησία ως προς την περιστροφή της εικόνας, οι παράγωγοι των δειγμάτων της γειτονιάς και οι συντεταγμένες τους περιστρέφονται σύμφωνα με τον προσανατολισμό του σημείου ενδιαφέροντος.

Το επόμενο βήμα που απαιτείται είναι ο διαχωρισμός της γειτονιάς του κάθε σημείου ενδιαφέροντος σε υποπεριοχές. Συγκεκριμένα, σχηματίζονται υποπεριοχές μεγέθους 4×4 εικονοστοιχεία. Όπως φαίνεται στο αριστερό τμήμα του σχήματος 2.4, η γειτονιά έχει μέγεθος 8×8 εικονοστοιχεία και αποτελείται από τέσσερις υποπεριοχές μεγέθους 4×4 εικονοστοιχεία. Το μέγεθος της γειτονιάς επηρεάζει άμεσα το μέγεθος του τελικού διανύσματος και την πολυπλοκότητα υπολογισμού.

Στη συνέχεια, από κάθε υποπεριοχή δειγμάτων σχηματίζεται ένα ιστόγραμμα προσανατολισμών. Τα ιστογράμματα αυτά είναι χωρισμένα σε r διαστήματα καλύπτοντας το εύρος των 360 μοιρών. Τα μέτρα των παραγώγων των δειγμάτων της υποπεριοχής που εξετάζουμε, ανάλογα με τον προσανατολισμό τους, προστίθενται στο αντίστοιχο διάστημα του ιστογράμματος. Το αποτέλεσμα της παραπάνω διαδικασίας είναι ο σχηματισμός των ιστογραμμάτων προσανατολισμών τα οποία συνθέτουν τον πίνακα περιγραφής ο οποίος αποτυπώνεται γραφικά



Σχήμα 2.4: Σχηματική αναπαράσταση του τρόπου υπολογισμού των διανυσμάτων περιγραφής [27].

στο δεξιό μέρος του σχήματος 2.4. Όπως μπορούμε να διαπιστώσουμε, κάθε ιστόγραμμα είναι χωρισμένο σε οκτώ διαστήματα και το μήκος κάθε βέλους αντιστοιχεί στη συχνότητα του αντίστοιχου διαστήματος του ιστογράμματος προσανατολισμών.

Πρέπει να σημειωθεί πως προτού σχηματιστούν τα ιστογράμματα των προσανατολισμών, το μέτρο της παραγώγου του κάθε δείγματος μίας γειτονιάς σταθμίζεται μέσω ενός φίλτρου Gauss με τυπική απόκλιση σ ίση με 1.5 φορές το πλάτος της γειτονιάς. Το φίλτρο Gauss αναπαρίσταται στο αριστερό μέρος του σχήματος 2.4 από τον μεγάλο μπλε κύκλο. Ο στόχος της στάθμισης είναι να αποφευχθούν έντονες διακυμάνσεις στις τιμές των ιστογραμμάτων εξαιτίας μικρών μετατοπίσεων της γειτονιάς ενός σημείου ενδιαφέροντος. Επίσης, δίνεται μικρότερη έμφαση στα δείγματα που απέχουν περισσότερο από το κέντρο της γειτονιάς καθώς τα δείγματα αυτά συγκεντρώνουν υψηλότερη πιθανότητα να έχουν συμπεριληφθεί λανθασμένα στη γειτονιά που εξετάζουμε.

Μία βασική αρχή πάνω στην οποία στηρίζεται η δημιουργία των διανυσμάτων περιγραφής είναι η ύπαρξη της δυνατότητας μετατόπισης των τοπικών δομών της εικόνας χωρίς οι τιμές των διανυσμάτων να διαφοροποιούνται. Οι τοπικές δομές της εικόνας εκφράζονται μέσω των εικονοστοιχείων που ανήκουν στη γειτονιά του εκάστοτε σημείου ενδιαφέροντος και μας ενδιαφέρει δύο όμοιες γειτονιές οι οποίες διαφέρουν μόνο στις θέσεις μερικών εικονοστοιχείων να περιγράφονται με τον ίδιο τρόπο. Συγκεκριμένα, ένα εικονοστοιχείο θα μπορούσε να μετατοπιστεί μέχρι και τέσσερις θέσεις παραμένοντας στην ίδια υποπεριοχή της γειτονιάς και συνεπώς να συνεισφέρει στο ίδιο ιστόγραμμα προσανατολισμών.

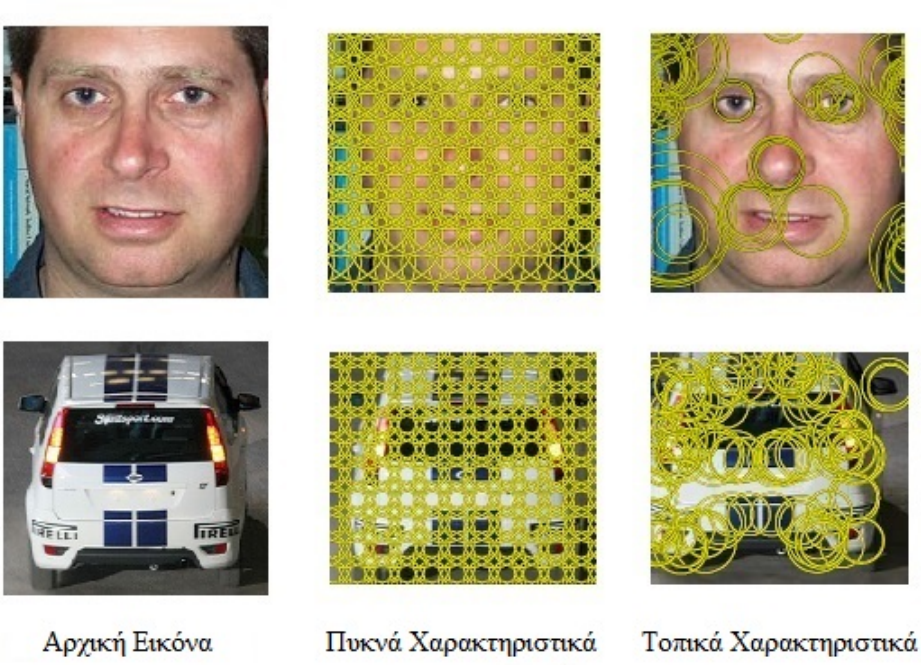
Με βάση την ίδια λογική, είναι σημαντικό να αντιμετωπιστούν οι περιπτώσεις που οι τιμές των διανυσμάτων περιγραφής διαφοροποιούνται απρόβλεπτα είτε εξαιτίας της μετατόπισης ενός δείγματος από μία υποπεριοχή σε άλλη, είτε εξαιτίας της μετατόπισης του προσανατολισμού ενός δείγματος από ένα διάστημα του ιστογράμματος σε ένα άλλο. Πραγματοποιείται λοιπόν

τριγραμμική παρεμβολή (trilinear interpolation) ώστε να κατανεμηθεί η τιμή του μέτρου της παραγώγου του κάθε δείγματος στα διπλανά διαστήματα του ιστογράμματος προσανατολισμών. Συγκεκριμένα, προτού ένα δείγμα ενταχθεί σε ένα ιστόγραμμα, το μέτρο της παραγώγου του σταθμίζεται με τον παράγοντα $(1 - d)$ σε κάθε διάσταση, όπου d είναι η απόσταση του δείγματος από την κεντρική τιμή του διαστήματος στο οποίο εντάσσεται.

Το τελικό διάνυσμα περιγραφής περιλαμβάνει τις τιμές όλων των ιστογραμμάτων προσανατολισμών. Το μόνο που απομένει είναι η κατάλληλη προσαρμογή του ώστε να εξασφαλιστεί ανεξαρτησία ως προς τις μεταβολές του φωτισμού. Οι μεταβολές αυτές ενδέχεται να είναι είτε γραμμικές είτε μη γραμμικές. Στην πρώτη περίπτωση εντάσσονται οι αλλαγές στην αντίθεση της εικόνας οι οποίες προκύπτουν από πολλαπλασιασμό της τιμής της φωτεινότητας του κάθε εικονοστοιχείου με μία σταθερά. Με την ίδια σταθερά θα είναι πολλαπλασιασμένο και το μέτρο των παραγώγων των εικονοστοιχείων, και συνεπώς εάν τα διανύσματα περιγραφής κανονικοποιηθούν ώστε να έχουν μοναδιαίο μέτρο, οι μεταβολές στην αντίθεση μπορούν να ακυρωθούν. Μία ακόμα γραμμική μεταβολή αποτελεί η ομοιόμορφη αλλαγή στη φωτεινότητα. Στην περίπτωση αυτή, μία σταθερά προστίθεται στην τιμή του κάθε εικονοστοιχείου και συνεπώς οι τοπικές παράγωγοι παραμένουν ανεπηρέαστες καθώς υπολογίζονται μέσω διαφορών μεταξύ των τιμών των εικονοστοιχείων όπως υποδηλώνεται και από τη σχέση 2.15. Οι μη γραμμικές μεταβολές καλύπτουν ένα πολύ μεγάλο εύρος αλλαγών στο φωτισμό, όπως για παράδειγμα η μεταβολή στην φωτεινότητα ανάλογα με τον προσανατολισμό της 3D επιφάνειας ενός αντικειμένου. Τέτοιες αλλαγές συνήθως επιδρούν στο μέτρο των παραγώγων των σημείων με διαφορετικό τρόπο, αλλά δεν επηρεάζουν τον προσανατολισμό των παραγώγων. Συνεπώς, επιδιώκεται να μετριαστεί η απρόβλεπτη συνέπεια της υπερβολικής αύξησης του μέτρου μίας παραγώγου και να δοθεί έμφαση στον τρόπο που κατανέμονται οι κατευθύνσεις των παραγώγων. Αυτό επιτυγχάνεται περιορίζοντας την μέγιστη επιτρεπτή τιμή του μέτρου των παραγώγων στο 0.2. Έπειτα, τα διανύσματα περιγραφής κανονικοποιούνται και πάλι ώστε να έχουν μοναδιαίο μέτρο. Η τιμή κατωφλίου 0.2 έχει προσδιοριστεί πειραματικά [27].

Η πολυπλοκότητα υπολογισμού των διανυσμάτων περιγραφής προσδιορίζεται από δύο παράγοντες, τον αριθμό των διαστημάτων r κάθε ιστογράμματος και τη διάσταση n των $n \times n$ πινάκων ιστογραμμάτων. Το μέγεθος του τελικού διανύσματος περιγραφής είναι rn^2. Ενδεχόμενη αύξηση στο μέγεθος του διανύσματος συνεπάγεται υψηλότερη διακριτική ικανότητα μεταξύ μεγάλου πλήθους χαρακτηριστικών, αλλά παράλληλα οδηγεί και σε εντονότερη ευαισθησία σε περιπτώσεις που τα αντικείμενα παρουσιάζουν παραμόρφωση ή μερική επικάλυψη με άλλα αντικείμενα της εικόνας. Πειραματικά έχει επιλεχθεί $r = 8$ και $n = 4$ [27]. Συνεπώς, σε κάθε σημείο ενδιαφέροντος αντιστοιχίζεται μία γειτονιά των 16×16 εικονοστοιχείων από την οποία προκύπτει ένας 4×4 πίνακας ιστογραμμάτων προσανατολισμών και εν τέλει ένα διάνυσμα περιγραφής 128 διαστάσεων.

Τέλος, αξίζει να σημειώσουμε πως το μέτρο και η κατεύθυνση κάθε σημείου υπολογίζεται εκ των προτέρων για όλες τις εικόνες του χώρου κλίμακας, και κατά συνέπεια δεν επιβαρύνεται με τις πράξεις αυτές η πολυπλοκότητα του αλγορίθμου που υπολογίζει τα διανύσματα περιγραφής.

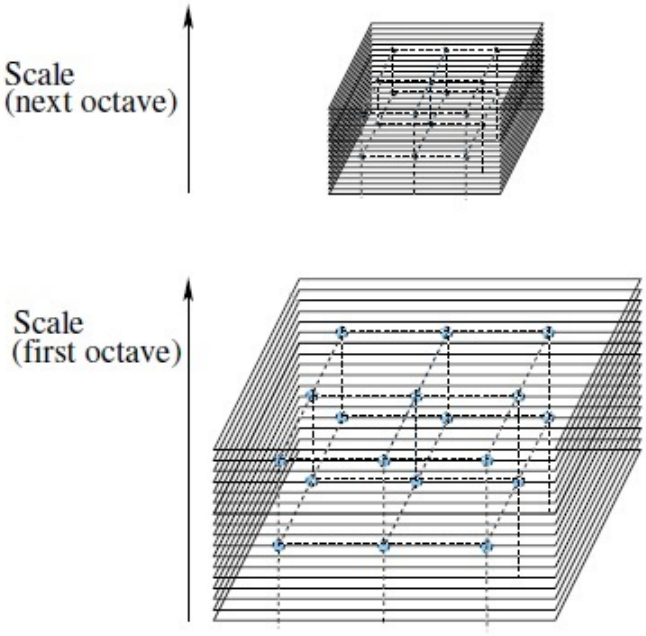


Σχήμα 2.5: Παράδειγμα εφαρμογής της πυκνής δειγματοληψίας και του ανιχνευτή Hessian Laplace ο οποίος χρησιμοποιείται για την ανίχνευση των τοπικών χαρακτηριστικών [42].

2.2 Πυκνή Δειγματοληψία

Η πυκνή δειγματοληψία αποτελεί μέθοδο ανίχνευσης χαρακτηριστικών και συνεπώς απαιτείται να συμπληρωθεί με τη χρήση μίας μεθόδου υπολογισμού διανυσμάτων περιγραφής. Προκειμένου να προσδιοριστούν τα σημεία ενδιαφέροντος σε μία εικόνα μέσω της πυκνής δειγματοληψίας, ένα ομοιόμορφο πλέγμα τοποθετείται στην εικόνα, υποδεικνύοντας τα εικονοστοιχεία που θα αποτελέσουν τα σημεία ενδιαφέροντος μέσω των κόμβων του πλέγματος. Γίνεται κατανοητό λοιπόν πως η βασική διαφορά της πυκνής δειγματοληψίας και των τοπικών χαρακτηριστικών, όπως αυτά ανιχνεύονται από έναν ανιχνευτή όπως ο SIFT, έγκειται στο γεγονός πως τα σημεία ενδιαφέροντος τοποθετούνται σε σταθερές θέσεις οι οποίες προσδιορίζονται ανεξάρτητα από το είδος και την πληροφορία της εικόνας. Για αυτόν ακριβώς το λόγο, η πυκνή δειγματοληψία βασίζεται στην υπόθεση πως η εικόνα είναι πλούσια σε πληροφορία στο σύνολό της και συνεπώς δεν αρκεί να ανιχνευτούν τοπικά σημεία αλλά είναι προτιμότερο να δειγματοληπτηθεί ομοιόμορφα ο χώρος της εικόνας.

Ένα παράδειγμα πυκνής δειγματοληψίας σε σύγκριση με μεθόδους ανίχνευσης τοπικών χαρακτηριστικών παρουσιάζεται στο σχήμα 2.5. Τα χαρακτηριστικά αναπαρίστανται από κίτρινους κύκλους, το μέγεθος των οποίων αντιστοιχεί στην κλίμακα των χαρακτηριστικών. Η κλίμακα ενός σημείου ενδιαφέροντος προσδιορίζει το μέγεθος της γειτονιάς του και χρησιμοποιείται τόσο για την εξαγωγή πληροφορίας όπως ο προσανατολισμός του σημείου, όσο και για το σχηματισμό των διανυσμάτων περιγραφής. Όπως φαίνεται στο σχήμα 2.5, στην πυκνή



Σχήμα 2.6: Τρόπος εφαρμογής της πυκνής δειγματοληψίας σε πολλαπλές κλίμακες μέσω του χώρου κλίμακας της εικόνας [42].

δειγματοληψία όλα τα σημεία έχουν την ίδια κλίμακα, ωστόσο δύναται να εφαρμοστεί και σε πολλαπλές κλίμακες. Η προσέγγιση αυτή απεικονίζεται στο σχήμα 2.6, όπου ένα ομοιόμορφο πλέγμα τοποθετείται στο χώρο κλίμακας.

Η πυκνή δειγματοληψία λοιπόν χρησιμοποιείται στις περιπτώσεις που είναι επιθυμητή η κάλυψη όλου του εύρους της πληροφορίας που εμπεριέχεται σε μία εικόνα. Συγκεκριμένα, στην περιγραφή μίας εικόνας περιλαμβάνονται ισότιμα και περιοχές με χαμηλή αντίθεση φωτεινότητας, οι οποίες θα είχαν αγνοηθεί από έναν ανιχνευτή τοπικών χαρακτηριστικών. Η λογική είναι πως ακόμα και αν είναι δύσκολη η ταύτιση τέτοιων περιοχών μεταξύ διαφορετικών εικόνων, περιλαμβάνουν πληροφορία η οποία είναι ενδεικτική του περιεχομένου της εικόνας και συνεπώς μπορούν να ωφελήσουν στην ανάκτηση εικόνων. Επιπλέον, μέσω της πυκνής δειγματοληψίας τα σημεία ενδιαφέροντος συνδέονται χωρικά μέσω της διάταξης του ομοιόμορφου πλέγματος που χρησιμοποιείται. Κατά αυτόν τον τρόπο, παρουσιάζεται η δυνατότητα αξιοποίησης της πληροφορίας που συνεπάγεται η χωρική διάταξη των σημείων με πιο δομημένο τρόπο σε σύγκριση με τα τοπικά χαρακτηριστικά, τα οποία ενδέχεται να βρίσκονται οσοδήποτε μακριά μεταξύ τους. Ένα ακόμα θετικό χαρακτηριστικό της πυκνής δειγματοληψίας είναι ότι παράγεται υψηλός, σταθερός και προκαθορισμένος αριθμός σημείων ενδιαφέροντος. Ο αριθμός αυτός εξαρτάται από τις διαστάσεις της εκάστοτε εικόνας και την απόσταση μεταξύ των κόμβων του ομοιόμορφου πλέγματος.

Στα μειονεκτήματα της πυκνής δειγματοληψίας περιλαμβάνονται οι υψηλές απαιτήσεις σε υπολογιστικούς πόρους. Ο απαιτούμενος χρόνος για την εξαγωγή των χαρακτηριστικών συν-

δέεται άμεσα με την πυκνότητα του πλέγματος, ωστόσο αναμένεται να είναι αισθητά μεγαλύτερος από τον χρόνο που απαιτείται για την εξαγωγή τοπικών χαρακτηριστικών. Παρομοίως, αυξημένες παρουσιάζονται και οι απαιτήσεις μνήμης λόγω του μεγάλου πλήθους χαρακτηριστικών. Σημαντική σημείωση αποτελεί το γεγονός ότι ενδέχεται αξιοσημείωτο μέρος του υπολογιστικού και αποθηκευτικού κόστους που προαναφέρθηκε να επενδύεται σε χαρακτηριστικά που δεν συνεισφέρουν ουσιαστικά στη διαδικασία ανάκτησης των εικόνων λόγω της πληροφορίας που εμπεριέχουν.

Σε κάθε περίπτωση, η πυκνή δειγματοληψία είναι μία απλή και ευρέως διαδεδομένη διαδικασία εξαγωγής σημείων ενδιαφέροντος η οποία ενδείκνυται σε εφαρμογές όπου ο χώρος ενδιαφέροντος των εικόνων εκτείνεται σχεδόν σε ολόκληρη την έκτασή τους. Υπενθυμίζουμε επίσης την ανάγκη συνδυασμού της με μία μέθοδο περιγραφής χαρακτηριστικών ώστε να ολοκληρωθεί η διαδικασία εξαγωγής των διανυσμάτων περιγραφής.

Κεφάλαιο 3

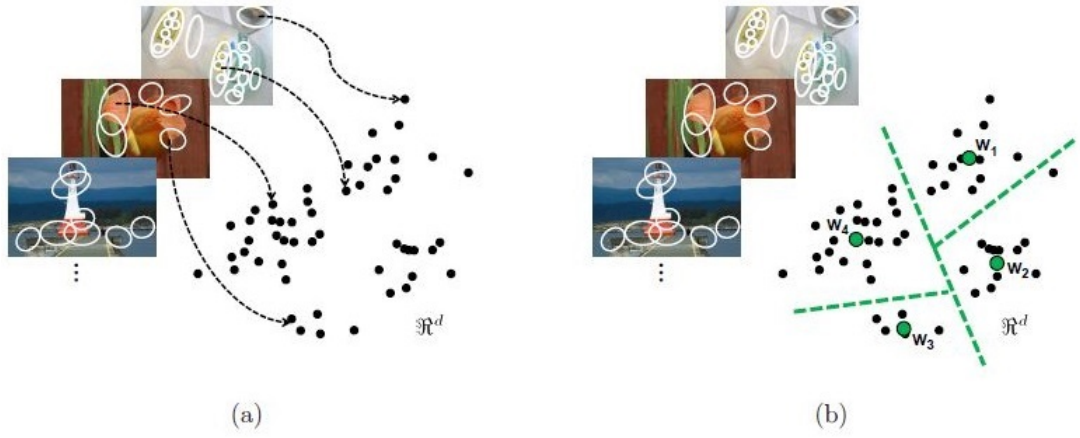
Γεωμετρική Άθροιση Διανυσμάτων Περιγραφής

Η μέθοδος διανυσματικής αναπαράστασης εικόνων Spatial Pyramid with Vectors of Locally Aggregated Descriptors (SP–VLAD), η οποία βασίζεται στην γεωμετρική άθροιση διανυσμάτων περιγραφής, αναπτύχθηκε εξ ολοκλήρου στο πλαίσιο της παρούσας διπλωματικής εργασίας και περιγράφεται αναλυτικά στο παρόν κεφάλαιο. Αρχικά παρουσιάζονται 3 μέθοδοι διανυσματικής αναπαράστασης εικόνων στις ιδέες των οποίων βασίστηκε η SP–VLAD. Οι μέθοδοι αυτοί είναι οι Bag of Words (BoW) [37], Spatial Pyramid Matching (SPM) [25] και Vector of Locally Aggregated Descriptors (VLAD) [20]. Στη συνέχεια περιγράφεται λεπτομερώς η μέθοδος SP–VLAD.

3.1 Μέθοδος Bag of Words (BoW)

Η ανάκτηση κειμένου (text retrieval) έχει γνωρίσει μεγάλη άνθιση και βρίσκεται σε ιδιαιτέρως προχωρημένο επίπεδο, ιδιαίτερα σε σχέση με την ανάκτηση εικόνων. Σε αυτό το πλαίσιο, έχει επιχειρηθεί η εφαρμογή ιδεών οι οποίες προέρχονται από συστήματα ανάκτησης κειμένου σε αυτά της ανάκτησης εικόνων. Μία τέτοια περίπτωση αποτελεί και η μέθοδος Bag of Words.

Πιο συγκεκριμένα, στα συστήματα ανάκτησης κειμένου συνήθως ακολουθείται μία σειρά από καθορισμένα βήματα. Αρχικά, τα αρχεία κειμένου χωρίζονται σε λέξεις. Στη συνέχεια, οι λέξεις χωρίζονται σε κατηγορίες ανάλογα με τη ρίζα τους και αναπαρίστανται με βάση αυτήν. Για παράδειγμα, οι λέξεις write, writing και writer αναπαρίστανται όλες από τη λέξη write που είναι η κοινή τους ρίζα. Λέξεις όπως είναι π.χ. τα άρθρα, αγνοούνται, καθώς εμφανίζονται με μεγάλη συχνότητα στα περισσότερα κείμενα και συνεπώς δεν αναμένεται να συνεισφέρουν στη διακριτική ικανότητα του συστήματος. Οι λέξεις που απομένουν κωδικοποιούνται κατάλληλα έτσι ώστε για κάθε κείμενο να σχηματιστεί ένα ιστόγραμμα του οποίου η κάθε τιμή προκύπτει ως η συχνότητα εμφάνισης της κάθε λέξης στο κείμενο αυτό. Κατ΄ επέκταση, ένα κείμενο αναπαριστάται από ένα διάνυσμα το οποίο σε κάθε του διάσταση περιλαμβάνει την αντίστοιχη τιμή του προαναφερθέντος ιστογράμματος. Τέλος, ενδέχεται στις τιμές του διανύσματος αυτού



Σχήμα 3.1: Διαδικασία δημιουργίας ενός οπτικού λεξικού [13].

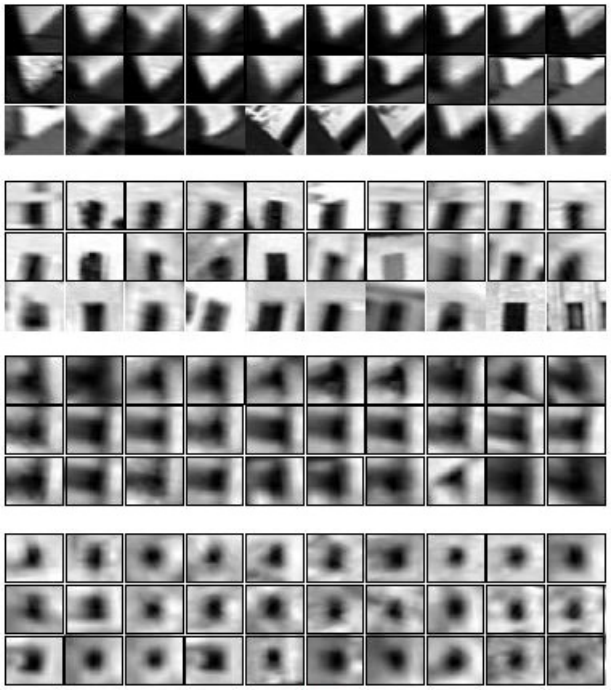
να αντιστοιχίζονται βάρη ανάλογα με την εφαρμογή στην οποία απευθύνεται το εκάστοτε σύστημα ανάκτησης.

Προκειμένου να πραγματοποιηθεί η ανάλογη διαδικασία στο χώρο των εικόνων απαιτείται η δημιουργία ενός είδους οπτικού λεξικού (visual vocabulary) το οποίο να αντικατοπτρίζει την οπτική πληροφορία των εικόνων. Υπό αυτή την έννοια, οι εικόνες αντιμετωπίζονται σαν ένα είδος αρχείου το οποίο αποτελείται από ένα σύνολο οπτικών λέξεων (visual words) και συνεπώς μπορεί να επιτευχθεί άμεσα η κατ΄ αναλογία εφαρμογή μεθόδων ανάκτησης κειμένου.

3.1.1 Οπτικό Λεξικό

Μία εικόνα μπορεί να θεωρηθεί ως ένα σύνολο διανυσμάτων περιγραφής τα οποία προκύπτουν από τη χρήση μίας μεθόδου ανίχνευσης και περιγραφής χαρακτηριστικών όπως είναι για παράδειγμα η SIFT. Ωστόσο, το μέγεθος των διανυσμάτων αυτών αλλά και η ποικιλία των τιμών που μπορεί να λάβει κάθε στοιχείο τους, συνήθως συνεπάγεται πως ο χώρος των χαρακτηριστικών (feature space) είναι ιδιαιτέρως πολλών διαστάσεων και συνεπώς δεν μας επιτρέπεται να θεωρήσουμε κάθε διάνυσμα περιγραφής ως μία ξεχωριστή οπτική λέξη. Αντίθετα, κατ΄ αναλογία με τη διαδικασία ομαδοποίησης των λέξεων μέσω της κοινής τους ρίζας, προκειμένου να εξαχθούν οι οπτικές λέξεις επιβάλλεται η ομαδοποίηση των διανυσμάτων περιγραφής. Αυτό επιτυγχάνεται μέσω της διακριτοποίησης του χώρου των χαρακτηριστικών έτσι ώστε κάθε κομμάτι του διακριτοποιημένου χώρου να εκφράζει μία οπτική λέξη. Κατά αυτόν τον τρόπο, κάθε διάνυσμα αντιστοιχίζεται σε μία οπτική λέξη ανάλογα με το κομμάτι του χώρου στο οποίο ανήκει.

Η διαδικασία της δημιουργίας του οπτικού λεξικού πραγματοποιείται σε δύο στάδια τα οποία αποτυπώνονται στο σχήμα 3.1. Στο πρώτο στάδιο, (a), εξάγονται τα διανύσματα περιγραφής από ένα σύνολο αντιπροσωπευτικών εικόνων της βάσης η οποία χρησιμοποιείται στη διαδικασία ανάκτησης. Οι λευκές ελλείψεις στις εικόνες αντιστοιχούν στις περιοχές που έχει εντοπίσει ο ανιχνευτής, και οι μαύρες κουκκίδες αντιπροσωπεύουν τα διανύσματα περιγραφής



Σχήμα 3.2: Περιοχές εικόνων από τις οποίες έχουν εξαχθεί διανύσματα περιγραφής τα οποία αντιπροσωπεύουν 4 διαφορετικές οπτικές λέξεις [37].

d διαστάσεων του κάθε σημείου ενδιαφέροντος στον χώρο $\Re^d$. Στο δεύτερο στάδιο, (b), δίνονται ως είσοδος σε έναν αλγόριθμο συσταδοποίησης τα διανύσματα περιγραφής και προκύπτουν k συστάδες, τα κέντρα των οποίων αποτυπώνονται στον χώρο $\Re^d$ μέσω των πράσινων κουκκίδων. Το κέντρο κάθε συστάδας αντιπροσωπεύει μία οπτική λέξη, οδηγώντας έτσι στη δημιουργία k οπτικών λέξεων. Ένας από τους δημοφιλέστερους αλγόριθμος συσταδοποίησης που χρησιμοποιούνται γι αυτόν τον σκοπό είναι ο k-means. Η διακριτοποίηση του χώρου $\Re^d$ αποτυπώνεται μέσω των διακεκομμένων πράσινων γραμμών του διαγράμματος Voronoi, το οποίο υποδηλώνει ότι τα διανύσματα περιγραφής αντιστοιχίζονται στην οπτική λέξη την οποία αντιπροσωπεύει το κέντρο στο οποίο βρίσκονται πιο κοντά. Οι αποστάσεις των διανυσμάτων που υπολογίζονται στον $\Re^d$ είναι συνήθως ευκλείδειες. Το αποτέλεσμα αυτής της διαδικασίας είναι η δημιουργία ενός οπτικού λεξικού το οποίο αποτελείται από k οπτικές λέξεις, οι οποίες εκφράζονται μέσω των διανυσμάτων των κέντρων των συστάδων τα οποία είναι d διαστάσεων, όπως και τα διανύσματα περιγραφής.

Σημαντικό στοιχείο αποτελεί το είδος της πληροφορίας που κωδικοποιούν οι οπτικές λέξεις που δημιουργούνται κατά αυτόν τον τρόπο. Η πληροφορία αυτή ποικίλει και εξαρτάται κυρίως από τέσσερα στοιχεία: τον τρόπο ανίχνευσης και περιγραφής των χαρακτηριστικών, το μέγεθος του λεξικού, τον αλγόριθμο συσταδοποίησης και το σύνολο των εικόνων το οποίο επιλέγεται για να εξαχθεί το λεξικό. Σε κάθε περίπτωση όμως, υπάρχει σημασιολογική ερμηνεία της πληροφορίας που φέρουν οι οπτικές λέξεις όπως κατ΄ αναλογία μπορούμε να αν-

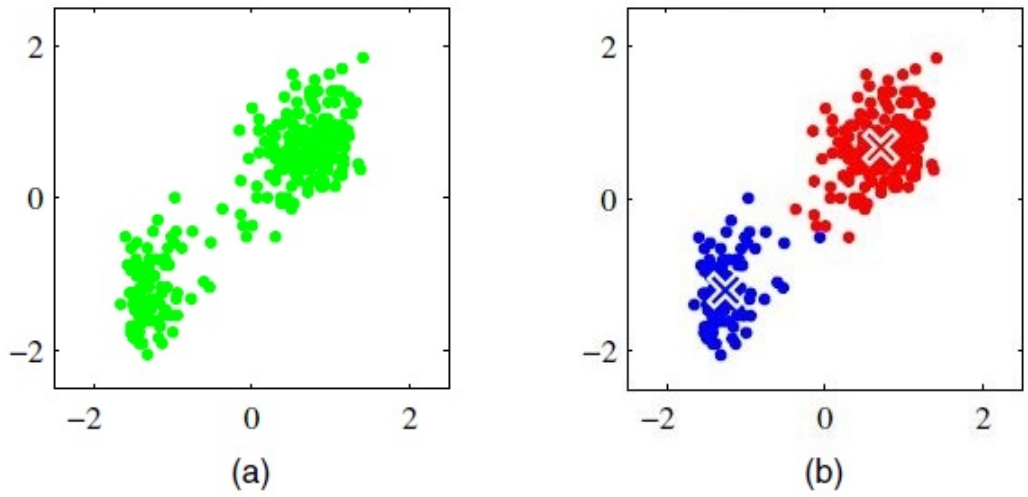
τιληφθούμε το νόημα των πραγματικών λέξεων ως συνθετικά στοιχεία ενός κειμένου. Αυτό αποτυπώνεται στο σχήμα 3.2 όπου παρουσιάζονται περιοχές εικόνων από τις οποίες προέρχονται διανύσματα περιγραφής τα οποία ανήκουν στις ίδιες συστάδες και συνεπώς αντιστοιχούν στις ίδιες οπτικές λέξεις.

Αξίζει να παρατηρήσουμε τη σημασία που έχει ο τρόπος επιλογής των εικόνων οι οποίες θα χρησιμοποιηθούν για να δημιουργηθεί το οπτικό λεξικό. Συγκεκριμένα, τα καλύτερα αποτελέσματα επιτυγχάνονται όταν χρησιμοποιούνται εικόνες οι οποίες προέρχονται από τη βάση δεδομένων η οποία θα χρησιμοποιηθεί μετέπειτα στη διαδικασία της ανάκτησης. Επίσης, στις περιπτώσεις που χρησιμοποιείται υποσύνολο των εικόνων της βάσης, είναι επιθυμητό το υποσύνολο αυτό να καλύπτει όσο το δυνατόν μεγαλύτερο εύρος της πληροφορίας που φέρουν οι εικόνες της βάσης έτσι ώστε να είναι όσο το δυνατόν πιο αντιπροσωπευτικό του συνόλου. Για παράδειγμα, σε μία βάση με εικόνες διαφορετικών αντικειμένων συνίσταται να επιλεχθούν εικόνες όλων των κατηγοριών. Ωστόσο, είναι συχνές οι περιπτώσεις που ένα οπτικό λεξικό παράγεται από ένα σύνολο εικόνων το οποίο δεν σχετίζεται με τη βάση δεδομένων στην οποία αναμένεται να εφαρμοστεί. Κατά αυτόν τον τρόπο ένα οπτικό λεξικό μπορεί να χρησιμοποιηθεί σε περισσότερες από μία διαφορετικές εφαρμογές παρά το γεγονός ότι αναμένονται μη βέλτιστα αποτελέσματα. Με αυτή τη λογική δημιουργούνται ''οικουμενικά'' οπτικά λεξικά που μπορούν να χρησιμοποιηθούν σε πλήθος εφαρμογών.

Συμπερασματικά, ένα πραγματικό λεξικό χρησιμοποιείται για τη σύνθεση προτάσεων που αποτελούν μονοδιάστατες δομές που έχει δημιουργήσει ο άνθρωπος, ενώ ένα οπτικό λεξικό επιχειρεί να περιγράψει εικόνες που αποτελούν διδιάστατες αναπαραστάσεις του τριδιάστατου πραγματικού κόσμου. Συνεπώς, αναλογιζόμενοι το σημαντικό ρόλο που διαδραματίζει το οπτικό λεξικό στην ακρίβεια της ανάκτησης, πρέπει να δημιουργείται με γνώμονα την εφαρμογή στην οποία θα εφαρμοστεί έτσι ώστε να επιτυγχάνονται τα καλύτερα δυνατά αποτελέσματα.

3.1.2 Τεχνική Συσταδοποίησης k–means

Υπάρχουν τρεις βασικές κατηγορίες τεχνικών συσταδοποίησης, αυτές που βασίζονται στις αποστάσεις των δεδομένων, αυτές που βελτιστοποιούν μία συνάρτηση ποιότητας συνολικά για τα δεδομένα, και αυτές που χρησιμοποιούν στατιστικά μοντέλα. Στη δεύτερη κατηγορία ανήκει ένας από τους δημοφιλέστερους αλγόριθμους συσταδοποίησης, ο k-means, ο οποίος χρησιμοποιείται όπως είδαμε για τη δημιουργία οπτικών λεξικών. Πρόκειται για προσεγγιστικό και επαναληπτικό αλγόριθμο στον οποίο δίνεται ως είσοδος ο επιθυμητός αριθμός συστάδων k. Ο αλγόριθμος ξεκινάει ορίζοντας τυχαία τα κέντρα των συστάδων στο χώρο των δεδομένων και κάθε του επανάληψη χωρίζεται σε δύο βήματα. Στο πρώτο βήμα κάθε στοιχείο του χώρου αντιστοιχίζεται στη συστάδα από το κέντρο της οποίας απέχει τη μικρότερη απόσταση. Στο δεύτερο βήμα επαναπροσδιορίζονται τα κέντρα των συστάδων υπολογίζοντας τη μέση τιμή των δεδομένων που έχουν αντιστοιχιστεί στην εκάστοτε συστάδα. Κατά συνέπεια, σε κάθε επανάληψη ορίζονται εκ νέου οι συστάδες έως ότου ο αλγόριθμος να συγκλίνει. Κατά αυτόν τον τρόπο, επιδιώκουμε να ελαχιστοποιήσουμε το συνολικό άθροισμα των αποστάσεων του κάθε δεδομένου από το κοντινότερο κέντρο. Κριτήριο σύγκλισης αποτελεί είτε ο αριθμός των



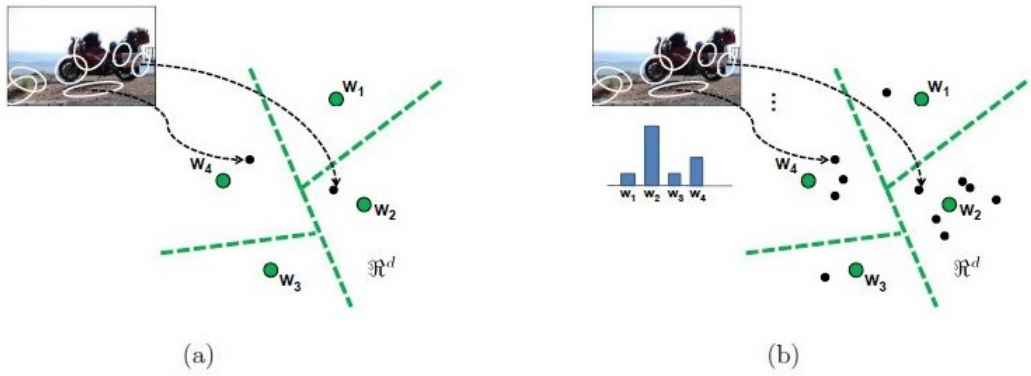
Σχήμα 3.3: Παράδειγμα εφαρμογής του αλγορίθμου k-means [4].

επαναλήψεων είτε η μέγιστη μεταβολή των κέντρων των συστάδων μεταξύ δύο διαδοχικών επαναλήψεων. Οι αποστάσεις συνήθως υπολογίζονται μέσω του τετραγώνου της ευκλείδειας απόστασης.

Ένα παράδειγμα εφαρμογής του αλγορίθμου k-means αποτυπώνεται στο σχήμα 3.3. Πρόκειται για μία απλή περίπτωση όπου επιλέγουμε $k = 2$. Τα δεδομένα αποδίδονται με πράσινα σημεία στο διδιάστατο ευκλείδειο χώρο (a). Οι συστάδας που προκύπτουν μετά την εφαρμογή του αλγορίθμου αποδίδονται με κόκκινα και μπλε σημεία αντίστοιχα (b). Τα κέντρα των 2 συστάδων αποδίδονται με τα σύμβολα "×" του αντίστοιχου χρώματος.

3.1.3 Διανυσματική Αναπαράσταση

Η μέθοδος BoW στοχεύει στο να αποτυπώσει την κατανομή των διανυσμάτων περιγραφής μίας εικόνας μέσω ενός ιστογράμματος, κάθε τιμή του οποίου υπολογίζεται από τον αριθμό των διανυσμάτων περιγραφής που αντιστοιχίζονται σε μία συγκεκριμένη οπτική λέξη. Σε αυτό το πλαίσιο, η διαδικασία διανυσματικής αναπαράστασης μίας εικόνας συντελείται σε δύο στάδια, τα οποία αποτυπώνονται στο σχήμα 3.4. Συγκεκριμένα, στο πρώτο στάδιο, (a), διαθέτοντας ένα οπτικό λεξικό και μία εικόνα από την οποία έχει εξαχθεί ένα σύνολο διανυσμάτων περιγραφής, κάθε διάνυσμα αντιστοιχίζεται σε μία οπτική λέξη, υπολογίζοντας την ελάχιστη απόσταση του διανύσματος αυτού από όλες τις οπτικές λέξεις. Μέσω αυτής της διαδικασίας, μία εικόνα η οποία αρχικά αποτελούσε ένα σύνολο διανυσμάτων περιγραφής, πλέον μπορεί να αποδοθεί ως ένα σύνολο από οπτικές λέξεις (bag of words). Στο δεύτερο στάδιο, (b), υπολογίζεται ένα ιστόγραμμα το οποίο αποτυπώνει τη συχνότητα εμφάνισης κάθε οπτικής λέξης στην εικόνα. Κατά αυτόν τον τρόπο, κάθε τέτοιο ιστόγραμμα περιλαμβάνει k τιμές, ίσες με το πλήθος των οπτικών λέξεων, και συνεπώς κάθε εικόνα μπορεί να αναπαρασταθεί από ένα διάνυσμα k διαστάσεων.



Σχήμα 3.4: Διαδικασία διανυσματικής αναπαράστασης εικόνων μέσω της μεθόδου Bag of Words [13].

Η διαδικασία της διανυσματικής αναπαράστασης ολοκληρώνεται κανονικοποιώντας το διάνυσμα του ιστογράμματος. Ο πιο διαδεδομένος τρόπος κανονικοποίησης αφορά την απόδοση βαρών σε κάθε τιμή του διανύσματος μέσω της τεχνικής ‘‘term frequency – inverse document frequency $(tf - idf)$’’ [37]. Η απόδοση βαρών κατά αυτόν τον τρόπο έχει ως εξής. Έστω ότι διαθέτουμε ένα λεξικό k οπτικών λέξεων και κάθε εικόνα i αναπαρίσταται από ένα διάνυσμα BoW, $V_i = (t_1, \ldots, t_j, \ldots, t_k)^T$. Τότε:

$$t_j = n_{ji} \cdot \frac{1}{n_i} \cdot log\frac{N}{n_j} \tag{3.1}$$

όπου n_{ji} είναι η συχνότητα εμφάνισης της οπτικής λέξης j στην εικόνα i, n_i είναι ο συνολικός αριθμός οπτικών λέξεων στην εικόνα i, n_j είναι η συχνότητα εμφάνισης της οπτικής λέξης j σε ολόκληρη τη βάση εικόνων, και N είναι ο συνολικός αριθμός εικόνων στη βάση. Ο όρος tf αντιστοιχεί στην ποσότητα n_{ji}/n_i και δίνει έμφαση στις οπτικές λέξεις που εμφανίζονται συχνά σε μία εικόνα. Η διαίρεση με τον όρο n_i εξασφαλίζει ανεξαρτησία ως προς τον αριθμό των οπτικών λέξεων που διαθέτει η κάθε εικόνα καθώς ο αριθμός αυτός δεν είναι απαραίτητα σταθερός. Ο όρος idf είναι ίσος με N/n_j και μειώνει την έμφαση που δίνεται σε οπτικές λέξεις οι οποίες εμφανίζονται με μεγάλη συχνότητα συνολικά στη βάση καθώς δεν προσδίδουν διακριτική ικανότητα στις εικόνες που τις περιλαμβάνουν. Αξίζει να σημειώσουμε πως η κανονικοποίηση μπορεί να πραγματοποιηθεί και μόνο με τη χρήση του όρου tf ή μόνο με τον πολλαπλασιασμό των τιμών του ιστογράμματος με τον όρο idf. Επίσης, η διαδικασία της κανονικοποίησης μπορεί να συντελεστεί και μέσω άλλων μεθόδων, όπως π.χ. με χρήση της νόρμας L_1 ή της νόρμας L_2, καθώς και μέσω συνδυασμού δύο ή και περισσότερων διαφορετικών μεθόδων.

Δύο εικόνες που αναπαρίστανται με τη μέθοδο BoW μπορούν να συγκριθούν με τη χρήση διαφόρων συναρτήσεων όπως είναι π.χ. το τετράγωνο της νόρμας L_2, η histogram intersection ή η πράξη του εσωτερικού γινομένου.

Ο αριθμός k των συστάδων ποικίλει ανάλογα με την εφαρμογή και ενδέχεται να εκτείνεται

από μερικές εκατοντάδες έως και αρκετές χιλιάδες. Σε κάθε περίπτωση όμως, η μέθοδος BoW παρέχει μία αρκετά απλή αναπαράσταση η οποία είναι και ιδιαιτέρως ''οικονομική'' σε επίπεδο μνήμης, καθώς για την περιγραφή μίας εικόνας δεν χρειάζεται ένα σύνολο μεγάλου πλήθους διανυσμάτων περιγραφής αλλά μόνο ένα αραιό ιστόγραμμα. Επίσης, όλες οι εικόνες αναπαρίστανται από ένα διάνυσμα σταθερού μεγέθους επιτρέποντας έτσι την ευκολότερη σύγκριση των διανυσμάτων αυτών, και επιπρόσθετα καθίσταται εφικτή η επέκταση της μεθόδου σε τεχνικές μηχανικής εκμάθησης (machine learning).

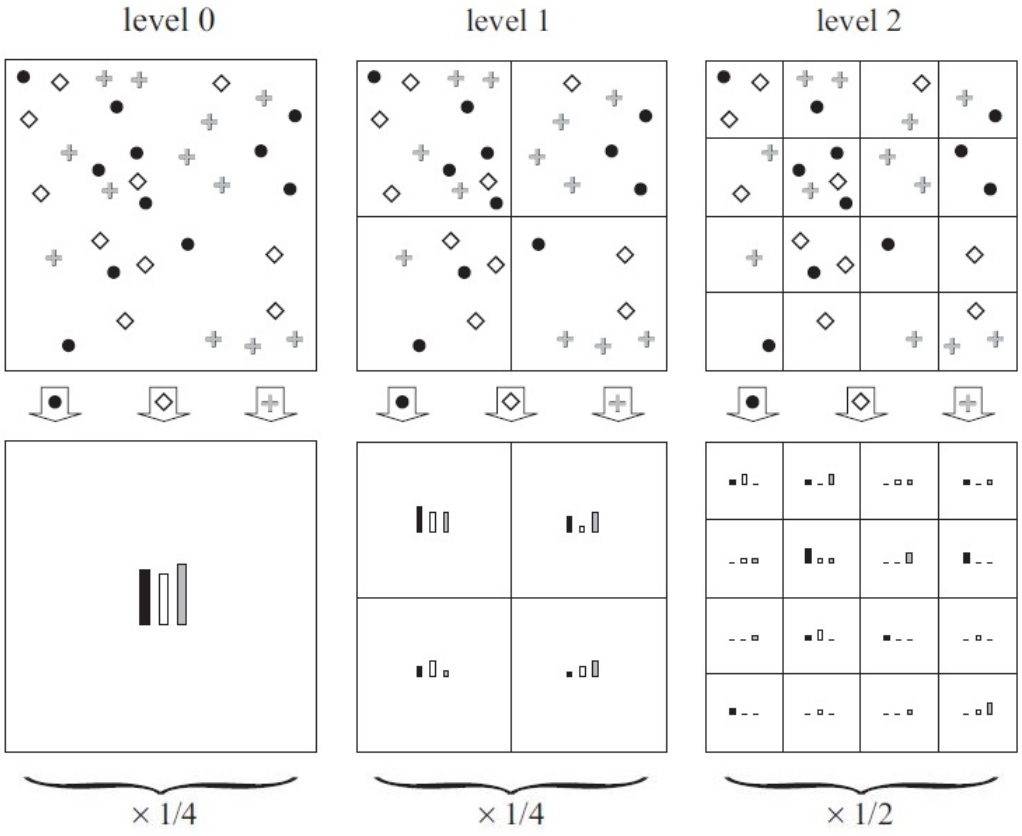
Η αναπαράσταση λοιπόν μίας εικόνας μέσω της μεθόδου BoW έχει αδιαμφισβήτητα πλεονεκτήματα και μπορεί να αποτελέσει τη βάση για πιο πολύπλοκες αναπαραστάσεις που θα επιφέρουν ακόμα υψηλότερη ακρίβεια σε ένα σύστημα ανάκτησης εικόνων.

3.2 Μέθοδος Spatial Pyramid Matching (SPM)

Η μέθοδος Spatial Pyramid Matching επεκτείνει τη μέθοδο BoW προς την κατεύθυνση της αξιοποίησης της χωρικής πληροφορίας που περιλαμβάνεται στις οπτικές λέξεις. Ο λόγος για κάτι τέτοιο βασίζεται στην ίσως σημαντικότερη διαφορά μεταξύ των οπτικών και των πραγματικών λέξεων, που είναι το γεγονός ότι οι οπτικές λέξεις διαθέτουν γεωμετρική δομή. Συγκεκριμένα, όταν πραγματοποιείται μία αναζήτηση σε ένα σύστημα ανάκτησης κειμένου μέσω ενός αριθμού λέξεων, οι λέξεις αυτές ενδέχεται να βρίσκονται σε οποιοδήποτε σημείο ενός κειμένου και μάλιστα χωρίς να είναι αυστηρή η σειρά μεταξύ τους. Αντίθετα, όταν μία αναζήτηση πραγματοποιείται μέσω οπτικών λέξεων, είναι αναπόφευκτο πως υπάρχει μία συσχέτιση μεταξύ τους η οποία αντιστοιχεί σε μία συγκεκριμένη αναπαράσταση του περιεχομένου της εικόνας που αντιπροσωπεύουν. Συνεπώς, είναι λογικό να επιχειρηθεί η ενσωμάτωση χωρικής πληροφορίας στην αναπαράσταση των εικόνων μέσω οπτικών λέξεων. Αυτό επιτυγχάνεται μέσω της λεγόμενης χωρικής πυραμίδας (spatial pyramid).

3.2.1 Χωρική Πυραμίδα

Προκειμένου να δημιουργηθεί η χωρική πυραμίδα, η εικόνα διαχωρίζεται διαδοχικά σε όλο και περισσότερες υποπεριοχές και για κάθε τέτοια υποπεριοχή υπολογίζεται το ιστόγραμμα της συχνότητας εμφάνισης των οπτικών λέξεων που περιέχει, όπως ακριβώς στη μέθοδο BoW. Η τεχνική της μεθόδου SPM οπτικοποιείται στο σχήμα 3.5. Αρχικά η εικόνα αποτελείται από ένα σύνολο διανυσμάτων περιγραφής καθένα από τα οποία έχει αντιστοιχιστεί σε μία οπτική λέξη. Αυτό είναι το επίπεδο 0 (level 0) στο σχήμα 3.5, όπου θεωρούμε πως υπάρχουν 3 είδη οπτικών λέξεων τα οποία συμβολίζονται μέσω κύκλων, ρόμβων και σταυρών. Το κάθε σύμβολο τοποθετείται στις συντεταγμένες του εικονοστοιχείου της εικόνας όπου ανιχνεύτηκε το χαρακτηριστικό που αντιπροσωπεύει. Στο επόμενο επίπεδο (level 1) δημιουργούνται 4 μη επικαλυπτόμενες υποπεριοχές στην εικόνα υποδιπλασιάζοντας τις αρχικές της συντεταγμένες. Στο τελευταίο επίπεδο (level 2) κάθε διάσταση των υποπεριοχών του προηγούμενου επιπέδου υποδιπλασιάζεται και πάλι, δημιουργώντας 16 μη επικαλυπτόμενες περιοχές. Γενικότερα, σε κάθε επίπεδο της πυραμίδας οι διαστάσεις των υποπεριοχών υποδιπλασιάζονται με βάση το προηγούμενο επίπεδο, με αποτέλεσμα ο αριθμός τους να αυξάνεται εκθετικά. Κάθε υποπεριοχή



Σχήμα 3.5: Παράδειγμα χωρικής πυραμίδας τριών επιπέδων.

αντιπροσωπεύεται από ένα διάνυσμα BoW σύμφωνα με τις οπτικές λέξεις που περιλαμβάνονται σε αυτήν. Τέλος, τα διανύσματα BoW κανονικοποιούνται, ολοκληρώνοντας έτσι τη δημιουργία της χωρικής πυραμίδας. Συγκεκριμένα, χρησιμοποιείται ο όρος tf, δηλαδή οι τιμές όλων των ιστογραμμάτων διαιρούνται με το συνολικό αριθμό των οπτικών λέξεων της εικόνας, εξασφαλίζοντας ουσιαστικά τη θεώρηση ίδιου αριθμού οπτικών λέξεων ανά εικόνα.

Προκειμένου να συγκριθούν δύο εικόνες που αναπαρίστανται από δύο χωρικές πυραμίδες ίδιου αριθμού επιπέδων, αρκεί να υπολογιστεί το σταθμισμένο άθροισμα του μέγιστου κοινού αριθμού οπτικών λέξεων ίδιου είδους που εμφανίζεται σε κάθε υποπεριοχή του κάθε επιπέδου. Αυτό επιτυγχάνεται μέσω της εφαρμογής της συνάρτησης histogram intersection η οποία λειτουργεί ως εξής. Έστω $H_X^{l,j}$ και $H_Y^{l,j}$ τα ιστογράμματα που αναπαριστούν διανυσματικά δύο αντίστοιχες υποπεριοχές j σε ένα συγκεκριμένο επίπεδο l των χωρικών πυραμίδων X και Y. Τότε, τα $H_X^{l,j}(i)$ και $H_Y^{l,j}(i)$ υποδηλώνουν τη συχνότητα εμφάνισης της οπτικής λέξης i στις υποπεριοχές j των επιπέδων l των πυραμίδων X και Y. Μέσω της συνάρτησης histogram intersection θα έχουμε:

$$I^j(H_X^{l,j}, H_Y^{l,j}) = \sum_{i=1}^{k} min(H_X^{l,j}(i), H_Y^{l,j}(i)) \tag{3.2}$$

Κατά αυτόν τον τρόπο υπολογίζεται ο μέγιστος κοινός αριθμός οπτικών λέξεων κάθε είδους

που υφίσταται σε μία υποπεριοχή δύο χωρικών πυραμίδων. Η εξίσωση (3.2) επεκτείνεται σε όλες τις υποπεριοχές ενός επιπέδου l με τον ακόλουθο τρόπο:

$$I^l = \sum_{j=1}^{N_l} I^j(H_X^{l,j}, H_Y^{l,j}) = \sum_{j=1}^{N_l} \sum_{i=1}^{k} min(H_X^{l,j}(i), H_Y^{l,j}(i)) \tag{3.3}$$

όπου N_l είναι ο αριθμός υποπεριοχών που εμφανίζονται σε ένα επίπεδο l. Πιο συγκεκριμένα, στο επίπεδο l μία εικόνα διαθέτει 2^l υποπεριοχές σε κάθε διάσταση, οδηγώντας στην ύπαρξη 4^l υποπεριοχών συνολικά στο επίπεδο. Όπως επιβεβαιώνεται και στο σχήμα 3.5, στο επίπεδο 0 ($l = 0$) έχουμε 1 υποπεριοχή, στο επίπεδο 1 ($l = 1$) έχουμε 4 και στο επίπεδο 2 ($l = 2$) έχουμε 16.

Επιπρόσθετα, γίνεται η θεώρηση πως εικόνες οι οποίες εμφανίζουν κοινό αριθμό οπτικών λέξεων σε υποπεριοχές υψηλότερων επιπέδων είναι πιο πιθανό να είναι σημασιολογικά όμοιες και συνεπώς ανατίθενται βάρη αναλόγως. Συγκεκριμένα, το βάρος που ανατίθεται στο επίπεδο l είναι $1/2^{L-l}$, όπου $l = 0, \ldots, L$. Το βάρος αυτό είναι αντιστρόφως ανάλογο του μήκους των ακμών των υποπεριοχών του κάθε επιπέδου. Ωστόσο, το σύνολο κοινών οπτικών λέξεων σε ένα επίπεδο l δίνεται από τη σχέση (3.3) και περιλαμβάνει και όλες τις κοινές οπτικές λέξεις που βρίσκονται στο επίπεδο $l + 1$. Συνεπώς, ο αριθμός των κοινών οπτικών λέξεων που εντοπίζονται μόνο στο επίπεδο l δίνεται από την ποσότητα $I^l - I^{l+1}$ για $l = 0, \ldots, L - 1$.

Εν τέλει, η συνολική ομοιότητα δύο χωρικών πυραμίδων X και Y δίνεται από την ακόλουθη σχέση:

$$I = I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(I^l - I^{l+1}) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l \tag{3.4}$$

Από τη σχέση (3.4) γίνεται φανερός ο λόγος που τα βάρη που χρησιμοποιούνται στο σχήμα 3.5 είναι 1/4, 1/4 και 1/2 για τα επίπεδα 0, 1 και 2 αντίστοιχα.

3.2.2 Διανυσματική Αναπαράσταση

Η πληροφορία της χωρικής πυραμίδας συμπυκνώνεται σε ένα διάνυσμα τοποθετώντας διαδοχικά τα κανονικοποιημένα διανύσματα των ιστογραμμάτων όλων των υποπεριοχών. Η διάσταση ενός τέτοιου διανύσματος δίνεται από την ακόλουθη ποσότητα:

$$k \sum_{l=0}^{L} 4^l = k \frac{1}{3}(4^{L+1} - 1) \tag{3.5}$$

όπου k είναι το μέγεθος του λεξικού, και L ο αριθμός των επιπέδων της πυραμίδας. Πολλαπλασιάζοντας τις κανονικοποιημένες τιμές των ιστογραμμάτων κάθε επιπέδου με τα κατάλληλα βάρη, ολοκληρώνεται η αναπαράσταση μίας εικόνας μέσω της μεθόδου SPM. Η ενσωμάτωση των βαρών στα τελικά διανύσματα κατά αυτόν τον τρόπο, καθιστά εφικτή τη σύγκριση δύο εικόνων που έχουν περιγραφεί με τη μέθοδο SPM μέσω της απλής εφαρμογής της συνάρτησης histogram intersection, αποφέροντας ακριβώς το ίδιο αποτέλεσμα με την σχέση (3.5).

Το μέγεθος του λεξικού που επιλέγεται συνήθως είναι είτε $k = 200$ είτε $k = 400$ καθώς αποδεικνύεται πειραματικά [25] πως περαιτέρω αύξηση δεν επιφέρει αξιοσημείωτη βελτίωση

στην απόδοση, ειδικότερα σε σχέση με την παράλληλη αύξηση σε ανάγκες μνήμης καθώς αυξάνεται το μέγεθος των τελικών διανυσμάτων αναπαράστασης των εικόνων. Αντίστοιχα, ο αριθμός επιπέδων που είθισται να επιλέγεται είναι είτε $L = 2$ είτε $L = 3$. Στην περίπτωση που $L = 0$, το διάνυσμα της αναπαράστασης SPM ταυτίζεται με το διάνυσμα BoW της εικόνας.

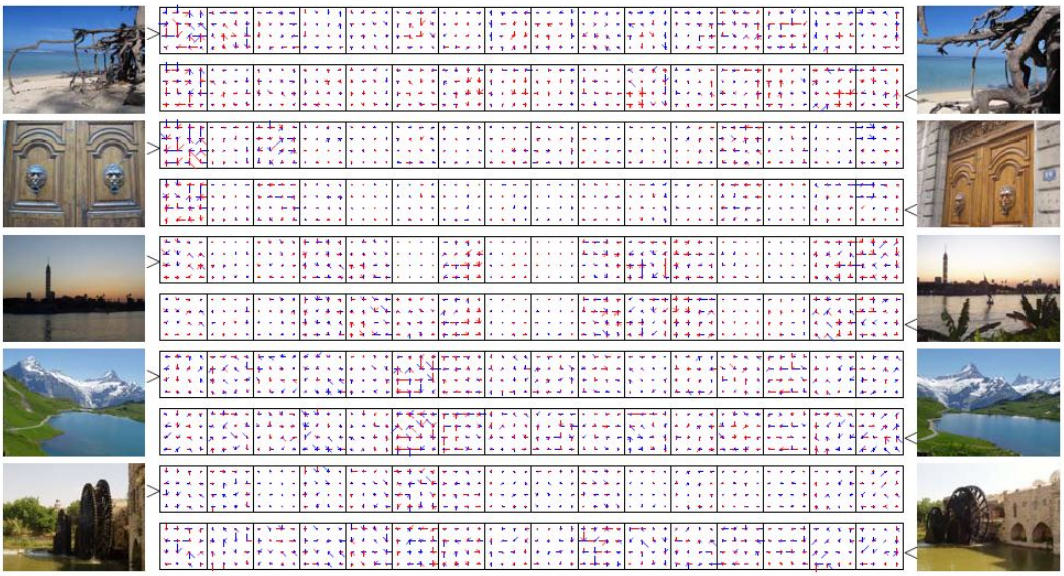
Ένα από τα πιο βασικά πλεονεκτήματα της μεθόδου SPM είναι πως εκμεταλλεύεται την πληροφορία πολλαπλών επιπέδων χωρίς απαραίτητα να επηρεάζεται από την ανομοιότητα που πιθανώς εμφανίζουν ορισμένες εικόνες σε μεμονωμένα επίπεδα. Στις περιπτώσεις βέβαια που μεταξύ όμοιων εικόνων υπάρχει ιδιαίτερα υψηλή διακύμανση στη δομή, η περιγραφή της μεθόδου SPM ενδέχεται να αποτυγχάνει. Ωστόσο, η γεωμετρική δομή που εισάγεται μέσω της χωρικής πυραμίδας παρέχει υψηλή διακριτική ικανότητα χωρίς να κρίνεται αναγκαία η ύπαρξη ενός μεγάλου οπτικού λεξικού όπως αυτών που χρησιμοποιούνται στη μέθοδο BoW, αποδεικνύοντας έτσι πως η προσθήκη χωρικής πληροφορίας προσφέρει μία περιγραφή πλουσιότερη σε πληροφορία σε σχέση με μία απλή αύξηση στο μέγεθος του λεξικού.

3.3 Μέθοδος Vector of Locally Aggregated Descriptors (VLAD)

Η μέθοδος VLAD χρησιμοποιείται για την διανυσματική αναπαράσταση εικόνων και βασίζεται στην ιδέα των διανυσμάτων περιγραφής και του οπτικού λεξικού όπως και η μέθοδος BoW. Η βασική της διαφορά ωστόσο από τη μέθοδο BoW είναι ότι η πληροφορία που αξιοποιεί δεν προέρχεται από τον αριθμό των διανυσμάτων περιγραφής που αντιστοιχούνται σε κάθε οπτική λέξη μέσω των συστάδων, αλλά από το βαθμό συγκέντρωσης των διανυσμάτων περιγραφής γύρω από τα κέντρα των συστάδων που αντιστοιχούνται. Περιγράφει δηλαδή την κατανομή που ακολουθούν τα διανύσματα της κάθε συστάδας ως προς το κέντρο της. Αυτό επιτυγχάνεται μέσω της άθροισης διανυσμάτων περιγραφής.

3.3.1 Άθροιση Διανυσμάτων Περιγραφής

Η κατανομή των διανυσμάτων περιγραφής γύρω από το κέντρο της συστάδας στην οποία ανήκουν αποτυπώνεται προσθέτοντας τις διαφορές όλων των διανυσμάτων από το κέντρο. Αυτή είναι η κεντρική ιδέα στην οποία βασίζεται η τεχνικής της άθροισης διανυσμάτων περιγραφής. Έστω λοιπόν ότι οι εικόνες που μας ενδιαφέρουν περιγράφονται από διανύσματα περιγραφής d διαστάσεων. Έστω επίσης ότι διαθέτουμε ένα οπτικό λεξικό C που διαθέτει k οπτικές λέξεις $C = \{c_1, \ldots, c_i, \ldots, c_k\}$, όπου οι οπτικές λέξεις αναπαριστώνται από τα κέντρα των συστάδων. Για παράδειγμα, η οπτική λέξη που αντιστοιχεί στη συστάδα i, αναπαρίσταται από το διάνυσμα d διαστάσεων c_i το οποίο αποτελεί το κέντρο της συστάδας αυτής. Κάθε διάνυσμα περιγραφής x αντιστοιχίζεται στην κοντινότερη οπτική λέξη c_i και συμβολίζεται με $NN(x) = c_i$. Οι αποστάσεις που υπολογίζονται είναι ευκλείδειες. Η κατανομή γύρω από το κέντρο c_i όλων των διανυσμάτων x για τα οποία ισχύει $NN(x) = c_i$, υπολογίζεται προσθέτοντας διαδοχικά τις διαφορές $x - c_i$. Συνεπώς, για κάθε συστάδα i μπορεί να σχηματιστεί



Σχήμα 3.6: Διανυσματική περιγραφή εικόνων μέσω της μεθόδου VLAD [20].

ένα διάνυσμα v_i διαστάσεων d, για το οποίο ισχύει:

$$v_{i,j} = \sum_{x:NN(x)=c_i} x_j - c_{i,j} \tag{3.6}$$

όπου τα $v_{i,j}$, x_j και $c_{i,j}$ συμβολίζουν την τιμή της j διάστασης των διανυσμάτων v_i, x και c_i αντίστοιχα. Τα διανύσματα $v_1, \ldots, v_k$ είναι το αποτέλεσμα της άθροισης των διανυσμάτων περιγραφής και εμπεριέχουν την πληροφορία της κατανομής των διανυσμάτων περιγραφής κάθε συστάδας ως προς το κέντρο της.

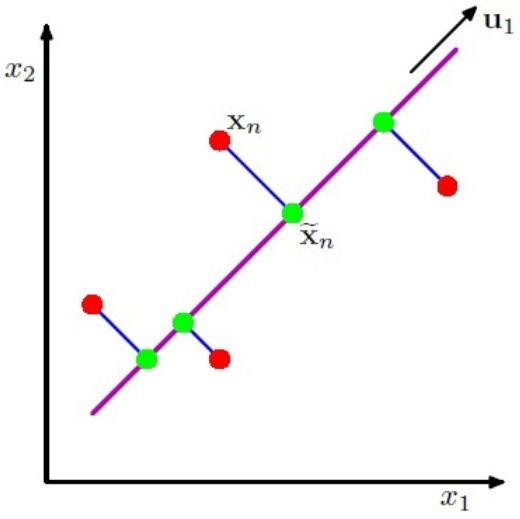
3.3.2 Διανυσματική Αναπαράσταση

Το διάνυσμα μαθηματικής αναπαράστασης μίας εικόνας μέσω της μεθόδου VLAD προκύπτει τοποθετώντας διαδοχικά τα διανύσματα $v_1, \ldots, v_k$. Κατά αυτόν τον τρόπο, προκύπτει ένα διάνυσμα $k \cdot d$ διαστάσεων.

Τα διανύσματα περιγραφής VLAD κανονικοποιούνται μέσω της μεθόδου Signed Square Rooting (SSR) [21]. Συγκεκριμένα, σε κάθε στοιχείο z ενός διανύσματος VLAD εφαρμόζεται η ακόλουθη συνάρτηση:

$$f(z) = sign(z)\sqrt{|z|} \tag{3.7}$$

Κατά αυτόν τον τρόπο επιχειρείται η αντιμετώπιση του φαινομένου των ‘‘bursty’’, όπως αποκαλούνται, χαρακτηριστικών, δηλαδή εκείνων των χαρακτηριστικών που παρουσιάζουν υψηλή συχνότητα εμφάνισης χωρίς αυτό να συνεπάγεται ουσιαστική συνεισφορά στη διακριτική ικανότητα των διανυσμάτων αναπαράστασης μίας εικόνας [19]. Στη συνέχεια, το διάνυσμα VLAD κανονικοποιείται μέσω της νόρμας L_2 ώστε να αποκτήσει μοναδιαίο μέτρο. Με αυτόν τον τρόπο ολοκληρώνεται η διανυσματική περιγραφή μίας εικόνας μέσω της μεθόδου VLAD.



Σχήμα 3.7: Παράδειγμα εφαρμογής της μεθόδου PCA ως ορθογώνια προβολή δεδομένων [4].

Να σημειώσουμε πως για την εξαγωγή των διανυσμάτων περιγραφής χρησιμοποιείται η μέθοδος SIFT και για την δημιουργία του λεξικού ο αλγόριθμος k-means.

Στο σχήμα 3.6 αποτυπώνονται τα διανύσματα VLAD για πέντε ζεύγη όμοιων σημασιολογικά εικόνων. Το μέγεθος του λεξικού είναι $k = 16$ και τα διανύσματα αναπαριστώνται όπως τα διανύσματα περιγραφής της μεθόδου SIFT. Συγκεκριμένα, χρησιμοποιείται ο 4×4 πίνακας ιστογραμμάτων προσανατολισμών για κάθε συστάδα. Με μπλε χρώμα αποδίδονται οι θετικές τιμές των διανυσμάτων VLAD και με κόκκινο οι αρνητικές. Είναι φανερό πως οι όμοιες οπτικά εικόνες παρουσιάζουν υψηλές τιμές στις ίδιες συστάδες καταδεικνύοντας την υψηλή διακριτική ικανότητα της μεθόδου.

Ένα από τα πιο θετικά χαρακτηριστικά της αναπαράστασης μίας εικόνας μέσω της μεθόδου VLAD είναι ότι δεν απαιτείται μεγάλο μέγεθος λεξικού προκειμένου να επιτευχθεί υψηλή διακριτική ικανότητα. Συγκεκριμένα, το μέγεθος του οπτικού λεξικού αρκεί να είναι μεταξύ $k = 16$ και $k = 256$. Επίσης, επιτρέπεται μεγάλη μείωση στη διάσταση των διανυσμάτων VLAD χωρίς να επηρεάζεται σημαντικά η διακριτική τους ικανότητα. Για αυτόν το σκοπό στα διανύσματα VLAD εφαρμόζεται η μέθοδος PCA στην οποία κάνουμε ειδική αναφορά στην επόμενη ενότητα. Οφείλουμε επίσης να σημειώσουμε πως μέσω της περιγραφής VLAD ευνοούνται προσεγγιστικές μέθοδοι αναζήτησης όπως η Asymmetric Distance Computation (ADC) [20]. Το στοιχείο αυτό σε συνδυασμό με το μικρό μέγεθος των διανυσμάτων VLAD που εξασφαλίζεται μέσω της εφαρμογής του PCA, καθιστούν την μέθοδο VLAD ικανή να χρησιμοποιηθεί σε εφαρμογές ανάκτησης εικόνων όπου οι βάσεις που χρησιμοποιούνται περιλαμβάνουν αριθμό εικόνων της τάξης του εκατομμυρίου.

3.3.3 Τεχνική Principal Component Analysis (PCA)

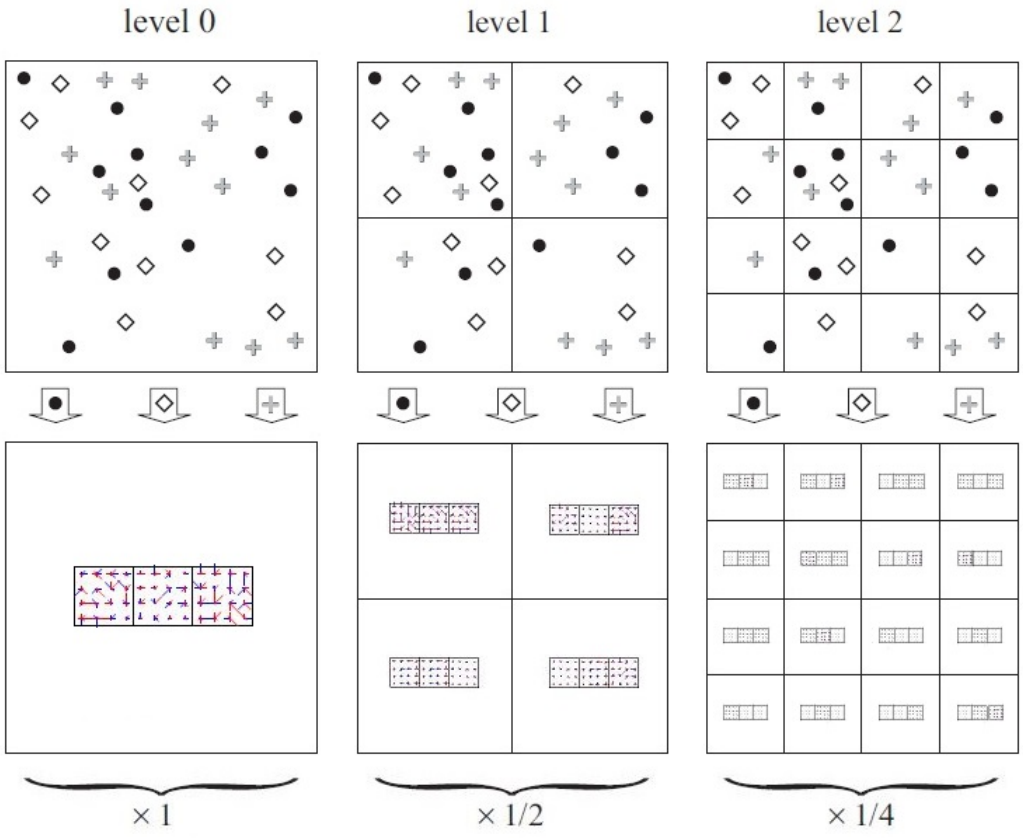
Η πιο γνωστή και διαδεδομένη μέθοδος για τη μείωση της διάστασης ενός συνόλου δεδομένων είναι η Principal Component Analysis (PCA) η οποία μπορεί να οριστεί με δύο τρόπους. Ο πρώτος τρόπος ορίζει τη μέθοδο PCA ως την ορθογώνια προβολή των δεδομένων σε ένα γραμμικό χώρο χαμηλότερης διάστασης, γνωστό ως κύριο υποχώρο (principal subspace), τέτοιο ώστε η διακύμανση των προβαλλόμενων δεδομένων να μεγιστοποιείται [16]. Ισοδύναμα, η μέθοδος PCA μπορεί να οριστεί και ως η γραμμική προβολή η οποία ελαχιστοποιεί το μέσο σφάλμα προβολής, το οποίο ορίζεται ως η μέση τετραγωνική απόσταση μεταξύ των δεδομένων και των προβολών τους [32]. Και στις δύο περιπτώσεις λοιπόν, το ζητούμενο είναι η προβολή των δεδομένων από έναν χώρο $\Re^d$ σε έναν χώρο $\Re^{d'}$, όπου $d' \leq d$. Τα ορθοκανονικά διανύσματα $\{u_1, \ldots, u_i, \ldots, u_{d'}\}$ τα οποία αποτελούν τα διανύσματα βάσης του νέου χώρου $\Re^{d'}$, αποδεικνύεται πως είναι τα ιδιοδιανύσματα που αντιστοιχούν στις d' μεγαλύτερες ιδιοτιμές του πίνακα συμμεταβλητότητας των δεδομένων [4]. Συγκεκριμένα, το ιδιοδιάνυσμα που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή ονομάζεται πρώτη κύρια συνιστώσα (principal component), το ιδιοδιάνυσμα που αντιστοιχεί στη δεύτερη μεγαλύτερη ιδιοτιμή ονομάζεται δεύτερη κύρια συνιστώσα κ.ο.κ. Κατά αυτόν τον τρόπο, τα ιδιοδιανύσματα συνθέτουν έναν πίνακα M, διαστάσεων $d' \times d$, και κάθε διάνυσμα δεδομένων x_n του $\Re^d$ αντιστοιχίζεται σε ένα διάνυσμα $\widetilde{x}_n$ του χώρου $\Re^{d'}$:

$$\widetilde{x}_n = Mx_n \tag{3.8}$$

Μέσω της μεθόδου PCA λοιπόν, ένα σύνολο διανυσμάτων δεδομένων προβάλλεται σε έναν νέο χώρο μικρότερης ή ίσης διάστασης, εξασφαλίζοντας πως η διακύμανση των προβαλλόμενων διανυσμάτων είναι μέγιστη, το μέσο συνολικό σφάλμα προβολής είναι ελάχιστο και οι διαστάσεις των προβαλλόμενων διανυσμάτων είναι ασυσχέτιστες. Η διαδικασία της ορθογώνιας προβολής απεικονίζεται στο σχήμα 3.7. Τα διδιάστατα δεδομένα x_n, που απεικονίζονται ως κόκκινα σημεία, προβάλλονται στον κύριο υποχώρο u_1 που είναι μικρότερης διάστασης. Τα προβαλλόμενα δεδομένα συμβολίζονται με $\widetilde{x}_n$ και αποτυπώνονται μέσω των σημείων πράσινου χρώματος.

3.4 Μέθοδος Spatial Pyramid with Vectors of Locally Aggregated Descriptors (SP–VLAD)

Η μέθοδος SP–VLAD προτείνεται για τη μαθηματική αναπαράσταση εικόνων και συνδυάζει τη χωρική πληροφορία των διανυσμάτων περιγραφής με την πληροφορία που εμπεριέχεται στον τρόπο κατανομής των διανυσμάτων αυτών γύρω από τα κέντρα των συστάδων στο χώρο των οπτικών λέξεων. Η αξία της χωρικής πληροφορίας πηγάζει από τη γεωμετρική δομή των οπτικών λέξεων μίας εικόνας και αξιοποιείται μέσω της τεχνικής της χωρικής πυραμίδας η οποία εισήχθη μέσω της μεθόδου SPM. Παράλληλα, η πληροφορία του τρόπου κατανομής των διανυσμάτων περιγραφής στον χώρο των οπτικών λέξεων εξασφαλίζεται μέσω της άθροισης των διανυσμάτων περιγραφής. Η τεχνική αυτή χρησιμοποιείται στα διανύσματα VLAD τα οποία παρέχουν μία εξαιρετικά συμπαγή περιγραφή η οποία χαρακτηρίζεται από χαμηλές



Σχήμα 3.8: Παράδειγμα χωρικής πυραμίδας τριών επιπέδων σύμφωνα με την τεχνική υπολογισμού της μεθόδου SP–VLAD.

απαιτήσεις μνήμης και υψηλή διακριτική ικανότητα. Η μέθοδος SP–VLAD λοιπόν, αθροίζει γεωμετρικά τα διανύσματα περιγραφής συνδυάζοντας τις ιδέες στις οποίες βασίζονται οι μέθοδοι SPM και VLAD, με στόχο να παρέχει μία περιγραφή ιδιαιτέρως πλούσια σε πληροφορία η οποία να υπολογίζεται αποδοτικά, να επιτυγχάνει υψηλή ακρίβεια και να διαθέτει χαμηλές απαιτήσεις μνήμης.

3.4.1 Τεχνική Υπολογισμού

Η μέθοδος SP–VLAD αθροίζει γεωμετρικά τα διανύσματα περιγραφής μίας εικόνας χρησιμοποιώντας την τεχνική της χωρικής πυραμίδας με τη βασική διαφορά ότι σε κάθε υποπεριοχή του κάθε επιπέδου της πυραμίδας δεν υπολογίζει ένα ιστόγραμμα με τις συχνότητες εμφάνισης των οπτικών λέξεων, αλλά ένα διάνυσμα VLAD. Δηλαδή η εικόνα αρχικά διαχωρίζεται σε L επίπεδα καθένα από τα οποία αποτελείται από 4^l μη επικαλυπτόμενες υποπεριοχές, όπου $l = 0, \ldots, L$. Στη συνέχεια, σε κάθε υποπεριοχή υπολογίζεται ένα διάνυσμα VLAD ανάλογα με τα διανύσματα περιγραφής που περιλαμβάνει. Τα διανύσματα VLAD κανονικοποιούνται αρχικά μέσω της μεθόδου SSR και έπειτα μέσω της νόρμας L_2. Η εξαγωγή των διανυσμάτων περιγραφής γίνεται με τη μέθοδο SIFT και η συσταδοποίηση των διανυσμάτων αυτών υλοποιεί-

ται με τον αλγόριθμο k-means χωρίς να υπάρχει κάποια δέσμευση για τη μη χρησιμοποίηση εναλλακτικών μεθόδων εξαγωγής χαρακτηριστικών και συσταδοποίησης.

Η τεχνική υπολογισμού καταδεικνύεται στο σχήμα 3.8. Παρατηρούμε πως υπάρχουν $L = 2$ επίπεδα και $k = 3$ οπτικές λέξεις οι οποίες συμβολίζονται μέσω κύκλων, ρόμβων και σταυρών. Για την αναπαράσταση των διανυσμάτων VLAD χρησιμοποιείται ο 4×4 πίνακας ιστογραμμάτων προσανατολισμών κατ΄ αντιστοιχία με τη μέθοδο SIFT.

Προκειμένου να συγκριθούν δύο εικόνες οι οποίες αναπαρίστανται με την τεχνική SP–VLAD υπολογίζεται το άθροισμα των αποτελεσμάτων της επιμέρους εφαρμογής μίας συνάρτησης αξιολόγησης στα διανύσματα VLAD των αντίστοιχων υποπεριοχών του κάθε επιπέδου. Η συνάρτηση αξιολόγησης f_d που επιλέγεται συνήθως είναι το τετράγωνο της ευκλείδειας απόστασης που για δύο διανύσματα x και y υπολογίζεται ως εξής:

$$f_d(x, y) = \|x - y\|^2 \tag{3.9}$$

Ωστόσο, όπως και στην περίπτωση της μεθόδου SPM, σε κάθε επίπεδο εφαρμόζονται διαφορετικά βάρη. Συγκεκριμένα, σε κάθε σύγκριση στο επίπεδο l εφαρμόζεται το βάρος:

$$2^{wl} \tag{3.10}$$

όπου w είναι η παράμετρος προσδιορισμού του τρόπου ανάθεσης των βαρών. Συγκεκριμένα, για $w > 0$ δίνεται μεγαλύτερη έμφαση σε υποπεριοχές υψηλότερων επιπέδων. Για παράδειγμα, εάν διαθέτουμε $L = 2$ και $w = 1$, για $l = 0$ το βάρος είναι 1, για $l = 1$ είναι 2 και για $l = 2$ είναι 4. Υπάρχουν όμως περιπτώσεις που τα υψηλότερα επίπεδα αποτυγχάνουν να αποδώσουν την πραγματική ομοιότητα δύο εικόνων καθώς στοχεύουν στην ανεύρεση γεωμετρικών ομοιοτήτων υψηλής λεπτομέρειας που είτε δεν υφίστανται είτε υπάρχουν σε περιορισμένο βαθμό. Σε αυτές τις περιπτώσεις εφαρμόζονται βάρη τα οποία στοχεύουν στο να δώσουν μεγαλύτερη έμφαση στα χαμηλότερα επίπεδα καθώς κρίνεται σκόπιμο τα υψηλότερα επίπεδα να συνεισφέρουν δευτερευόντως στη διακριτική ικανότητα των διανυσμάτων SP–VLAD. Αυτό επιτυγχάνεται επιλέγοντας $w < 0$. Εάν π.χ. διαθέτουμε $L = 2$ και θέσουμε $w = -1$, τότε από την σχέση (3.10) προκύπτει πως για $l = 0$ το βάρος είναι 1, για $l = 1$ είναι $1/2$ και για $l = 2$ είναι $1/4$. Αυτός ο τρόπος ανάθεσης βαρών αποτυπώνεται και στο σχήμα 3.8. Γενικότερα, αυξάνοντας τη θετική τιμή του w, η διαφορά μεταξύ των βαρών διαδοχικών επιπέδων γίνεται εντονότερη καθώς κινούμαστε σε υψηλότερα επίπεδα, και συνεπώς η διακριτική ικανότητα των διανυσμάτων SP–VLAD προσδιορίζεται σε όλο και μεγαλύτερο βαθμό από τα υψηλότερα επίπεδα. Αντιστρόφως, μειώνοντας την αρνητική τιμή του w, π.χ. επιλέγοντας $w = -2$, αποδίδεται ακόμα μεγαλύτερη έμφαση στα χαμηλότερα επίπεδα. Για $w = 0$ δεν αποδίδονται βάρη στα επίπεδα της πυραμίδας.

3.4.2 Διανυσματική Αναπαράσταση

Η διανυσματική αναπαράσταση μιας εικόνας με τη μέθοδο SP–VLAD επιτυγχάνεται τοποθετώντας διαδοχικά τα διανύσματα VLAD κάθε υποπεριοχής του κάθε επιπέδου. Το μέγεθος

του διανύσματος που προκύπτει δίνεται από τη σχέση:

$$d_{sp-vlad} = d_{vlad} \cdot N_{sp} = kd \cdot \sum_{l=0}^{L} 4^l = kd \cdot \frac{1}{3}(4^{L+1} - 1) \tag{3.11}$$

όπου d_{vlad} είναι το μέγεθος των διανυσμάτων VLAD, N_{sp} είναι ο αριθμός όλων των υποπεριοχών όλων των επιπέδων της πυραμίδας, d είναι το μέγεθος των διανυσμάτων περιγραφής, k είναι ο αριθμός των οπτικών λέξεων και L ο αριθμός των επιπέδων της πυραμίδας. Επίσης, εντάσσοντας τα βάρη στις τιμές των διανυσμάτων SP–VLAD, είναι εφικτή η σύγκριση δύο τέτοιων διανυσμάτων μέσω της απευθείας εφαρμογής της συνάρτησης της σχέσης 3.9. Αξίζει να παρατηρήσουμε πως στην περίπτωση που $L = 0$, τα διανύσματα SP–VLAD ταυτίζονται με της μεθόδου VLAD.

Όπως και στην περίπτωση της μεθόδου VLAD, το μέγεθος του οπτικού λεξικού επιλέγεται να είναι σχετικά μικρό, δηλαδή μεταξύ $k = 16$ και $k = 256$. Επίσης, ο αριθμός των επιπέδων συνήθως δεν ξεπερνά την τιμή $L = 3$. Ωστόσο, σε μια μέση περίπτωση όπου $k = 64$, $L = 2$ και $d = 128$ λόγω της χρήσης του SIFT descriptor, από τη σχέση 3.11 μπορούμε να υπολογίσουμε πως το μέγεθος ενός διανύσματος SP–VLAD θα είναι $d_{sp-vlad} = 172032$ διαστάσεων. Το μέγεθος αυτό είναι απαγορευτικό για την εφαρμογή της μεθόδου σε βάσεις εικόνων πολύ μεγάλου μεγέθους αλλά και γενικότερα για την επίτευξη γρήγορων αναζητήσεων. Είναι όμως εφικτή η μείωση της διάστασης των διανυσμάτων SP–VLAD σε πολύ μεγάλο βαθμό, χωρίς να επηρεάζεται σημαντικά η διακριτική τους ικανότητα. Για το σκοπό αυτό χρησιμοποιείται η μέθοδος PCA και επιτυγχάνεται μία ιδιαιτέρως οικονομική αναπαράσταση των εικόνων από άποψη μνήμης. Άλλωστε, όπως συμβαίνει και στη μέθοδο VLAD, τα διανύσματα περιγραφής δεν χρησιμοποιούνται μετά την εξαγωγή των τελικών διανυσμάτων SP–VLAD και συνεπώς δεν απαιτείται η αποθήκευση τους, ελευθερώνοντας έτσι σημαντικούς αποθηκευτικούς πόρους.

Γίνεται αντιληπτό πως η μέθοδος SP–VLAD διαθέτει αρκετές παραμέτρους, όπως το μέγεθος του οπτικού λεξικού, ο αριθμός των επιπέδων της χωρικής πυραμίδας αλλά και ο τρόπος ανάθεσης των βαρών κάθε επιπέδου. Συνεπώς είναι καθοριστικός ο τρόπος επιλογής των παραμέτρων αυτών προκειμένου να επιτευχθούν βέλτιστα αποτελέσματα για την εκάστοτε εφαρμογή.

Κεφάλαιο 4

Πειραματικά Αποτελέσματα

Στο παρόν κεφάλαιο παρουσιάζονται τα πειραματικά αποτελέσματα από την εφαρμογή της μεθόδου που προτείνουμε, SP–VLAD, σε 3 διαφορετικές βάσεις εικόνων. Οι δύο από αυτές είναι οι ευρέως διαδεδομένες βάσεις INRIA Holidays [18] και Caltech 101 [11], ενώ η τρίτη βάση εικόνων που χρησιμοποιήθηκε είναι η Flowers 15 την οποία δημιουργήσαμε από εικόνες λουλουδιών του flickr (www.flickr.com). Παράλληλα, σε αυτές τις τρεις βάσεις εικόνων εφαρμόστηκαν και οι μέθοδοι SPM και VLAD προκειμένου να υπάρξει σύγκριση των αποτελεσμάτων. Λόγω των ενθαρρυντικών αποτελεσμάτων επεκτείναμε την εφαρμογή της μεθόδου μας και στο πρόβλημα της κατηγοριοποίησης εικόνων (image classification) χρησιμοποιώντας τις βάσεις Caltech 101 και Flowers 15. Και σε αυτήν την περίπτωση οι μέθοδοι SPM και VLAD αποτέλεσαν το μέτρο σύγκρισης για την ποιότητα των αποτελεσμάτων.

4.1 Βάσεις Εικόνων

Η επιλογή των βάσεων που χρησιμοποιήθηκαν στην πειραματική διαδικασία έγινε με γνώμονα το είδος και το πλήθος των εικόνων που περιλαμβάνουν, αλλά και τον βαθμό χρησιμοποίησής τους από την ερευνητική κοινότητα. Παρακάτω παραθέτουμε αναλυτικά στοιχεία για την κάθε βάση εικόνων ξεχωριστά:

- **INRIA Holidays** [18]: Πρόκειται για ένα σύνολο 1491 εικόνων οι οποίες αποτελούν φωτογραφίες διακοπών. Οι εικόνες είναι υψηλής ανάλυσης, π.χ. 2448×3264 εικονοστοιχεία. Συνολικά υπάρχουν 500 διαφορετικές κατηγορίες εικόνων εκ των οποίων κάθε μία περιλαμβάνει μία διαφορετική σκηνή ή αντικείμενο. Για κάθε κατηγορία υπάρχουν από 2 μέχρι και 13 εικόνες. Δείγματα των εικόνων της βάσης παρουσιάζονται στο σχήμα 4.1. Η βάση αυτή επιλέχτηκε καθώς είχε χρησιμοποιηθεί ήδη για την αξιολόγηση της μεθόδου VLAD [20, 21] και αποτελεί σημείο αναφοράς για την ερευνητική κοινότητα στην ανάκτηση εικόνων.

- **Caltech 101** [11]: Το συγκεκριμένο σύνολο εικόνων περιλαμβάνει 101 διαφορετικές κατηγορίες αντικειμένων. Ενδεικτικό δείγμα εικόνων της βάσης από διαφορετικές κατηγορίες αντικειμένων παρατίθεται στο σχήμα 4.2. Κάθε κατηγορία αποτελείται από 31

έως 800 εικόνες. Στα πειράματά μας χρησιμοποιήσαμε 100 από τις κατηγορίες εικόνων επιλέγοντας 10 εικόνες από την κάθε μία. Δημιουργήσαμε δηλαδή ένα υποσύνολο της βάσης Caltech 101 με συνολικά 1000 εικόνες. Όλες οι εικόνες είναι μέσης ανάλυσης, π.χ. 300×300 εικονοστοιχεία. Το συγκεκριμένο σύνολο εικόνων σχηματίστηκε λόγω του μεγάλου εύρους διαφορετικών κατηγοριών εικόνων που περιλαμβάνει, της ευρύτατης χρησιμοποίησης της βάσης Caltech 101 στο πρόβλημα της κατηγοριοποίησης και της ανάκτησης εικόνων, αλλά και λόγω της εφαρμογής της μεθόδου SPM σε αυτήν κατά το παρελθόν [25].

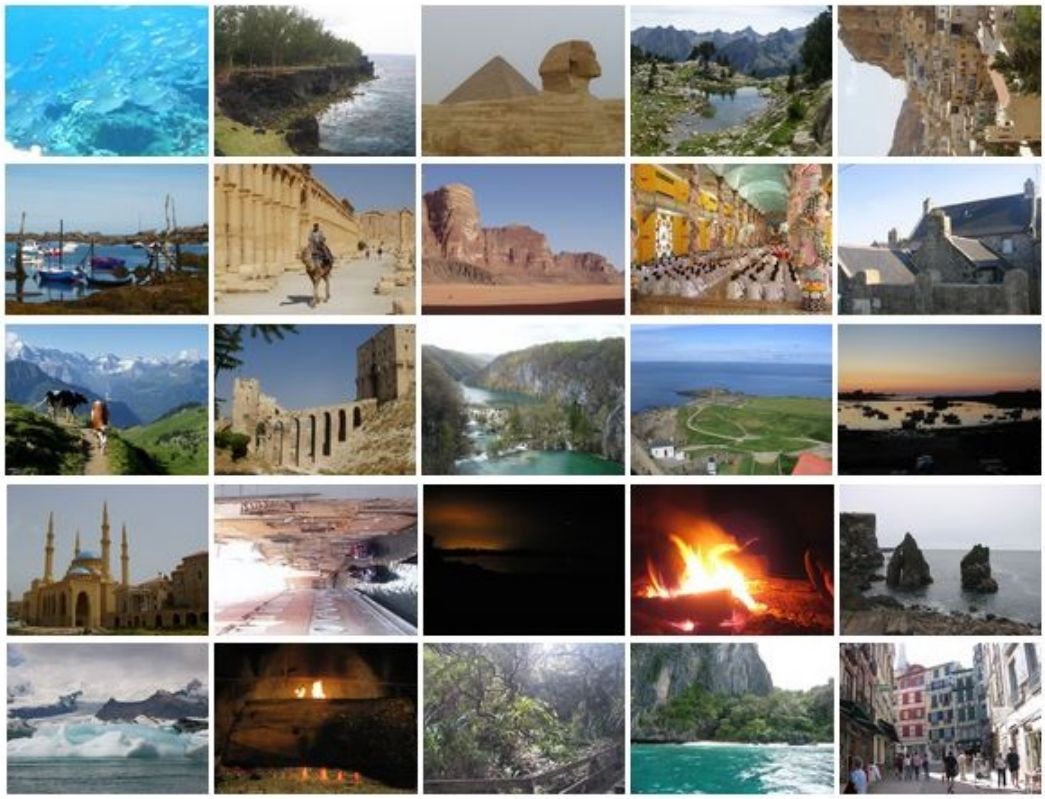
- **Flowers 15**: Το σύνολο εικόνων Flowers 15 αποτελείται από 15 διαφορετικά γένη ανθοφόρων φυτών. Ο διαχωρισμός των φυτών σε γένη έγινε με βάση το σύστημα Angiosperm Phylogeny Group (APG) III [40]. Τα γένη φυτών που περιλαμβάνονται είναι Abutilon, Acca, Albizia, Allamanda, Bauhinia, Caesalpinia, Callistemon, Catharanthus, Erythrina, Gladiolus, Lathyrus, Malva, Nerium, Passiflora και Rosa. Για κάθε ένα από αυτά τα 15 γένη ανθοφόρων φυτών υπάρχουν 30 εικόνες και ενδεικτικά παραδείγματα περιλαμβάνονται στο σχήμα 4.3. Οι εικόνες της βάσης είναι μέσης ανάλυσης, για παράδειγμα 500×500 εικονοστοιχεία, και προέρχονται από το flickr (www.flickr.com). Ο λόγος δημιουργίας αυτής της βάσης ήταν η επιθυμία εφαρμογής μεθόδων ανάκτησης και κατηγοριοποίησης σε εικόνες ανθοφόρων φυτών με προοπτική την ένταξή τους μελλοντικά στην μηχανή αναζήτησης εικόνων Floral [17].

4.2 Πειραματική Διαδικασία

Η πειραματική διαδικασία αποτυπώνεται στο διάγραμμα ροής του σχήματος 4.4. Όλα τα στάδια υλοποιήθηκαν στη γλώσσα προγραμματισμού C++ με χρήση των βιβλιοθηκών OpenCV [6] και Boost [5].

Ανεξάρτητα από το γεγονός ότι οι εικόνες και των τριών βάσεων είναι έγχρωμες, η επεξεργασία των εικόνων πραγματοποιήθηκε θεωρώντας αποχρώσεις του γκρι. Δεδομένης λοιπόν μίας εκ των τριών βάσεων εικόνων που διαθέτουμε, αρχικά εξάγονται τα διανύσματα περιγραφής από όλες τις εικόνες. Η ανίχνευση των χαρακτηριστικών γίνεται είτε μέσω της εφαρμογής του ανιχνευτή SIFT είτε μέσω πυκνής δειγματοληψίας. Όταν εφαρμόζεται ο ανιχνευτής SIFT, ο αριθμός s των διαστημάτων κάθε οκτάβας ορίζεται $s = 5$ και ο αριθμός των οκτάβων ορίζεται ακολούθως αυτόματα ανάλογα με την ανάλυση της εκάστοτε εικόνας. Η τυπική απόκλιση του φίλτρου Gauss το οποίο χρησιμοποιείται για την εξομάλυνση της αρχικής εικόνας στην πρώτη οκτάβα τίθεται $\sigma = 1.6$. Το κατώφλι για την απόρριψη μη ευσταθών σημείων με μικρή αντίθεση φωτεινότητας επιλέγεται $h = 0.04$ και το αντίστοιχο κατώφλι για τα μη ευσταθή σημεία των ακμών ορίζεται $r = 10$. Η πυκνή δειγματοληψία πραγματοποιείται μόνο σε ένα επίπεδο και στο ομοιόμορφο πλέγμα που τοποθετείται στην εκάστοτε εικόνα οι αποστάσεις μεταξύ των κόμβων είναι 5 εικονοστοιχεία. Ανεξάρτητα από τον τρόπο ανίχνευσης των χαρακτηριστικών, η περιγραφή τους πραγματοποιείται μέσω της μεθόδου SIFT.

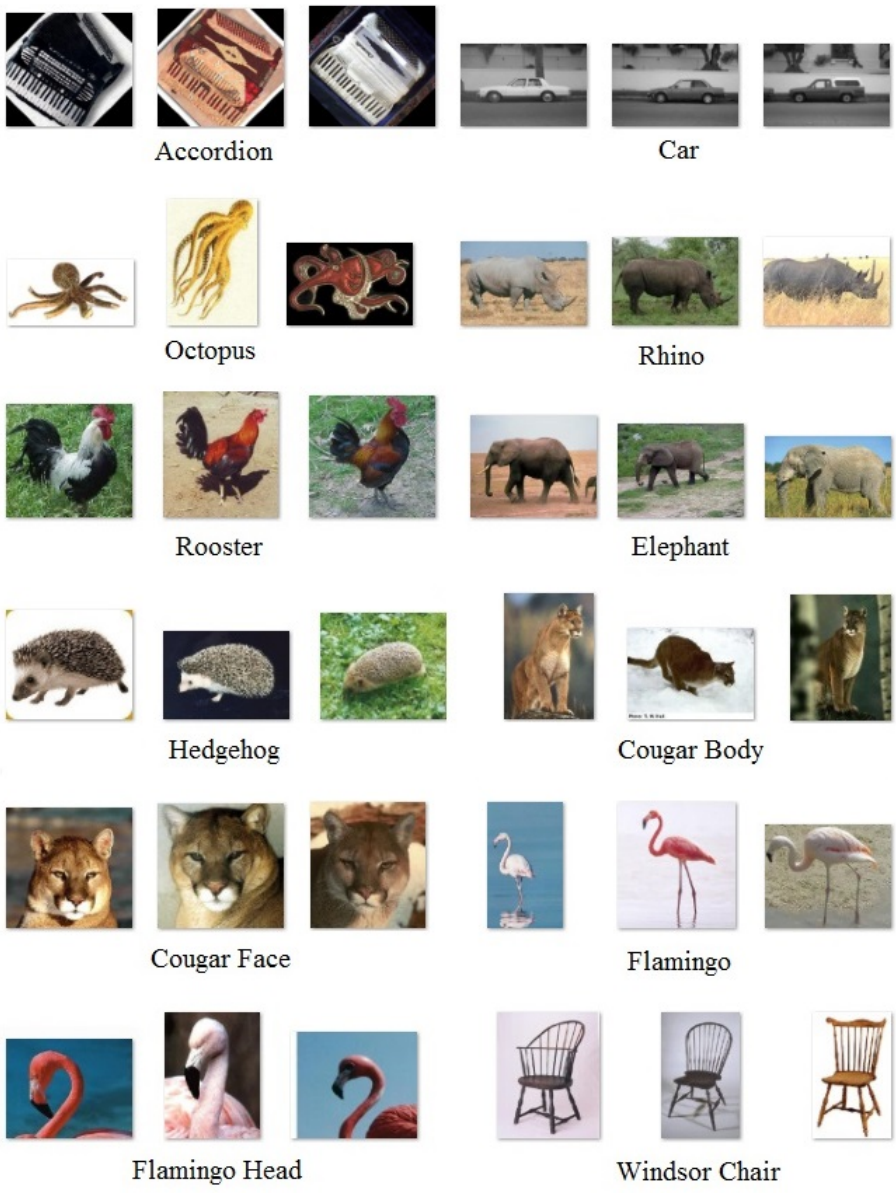
Στο επόμενο στάδιο δημιουργείται το οπτικό λεξικό. Συνολικά δημιουργούνται οπτικά



Σχήμα 4.1: Παραδείγματα εικόνων από τη βάση δεδομένων INRIA Holidays.

λεξικά 4 διαφορετικών μεγεθών, $k = 16, 64, 128, 200$. Ο αλγόριθμός που χρησιμοποιείται είναι ο k-means και ο αριθμός k αντιστοιχεί στον αριθμό των συστάδων και κατ΄ επέκταση των οπτικών λέξεων. Κάθε λεξικό δημιουργείται χρησιμοποιώντας όλο το πλήθος των διανυσμάτων περιγραφής στα οποία θα εφαρμοστεί. Εξάγονται διαφορετικά λεξικά για τις περιπτώσεις που η ανίχνευση των χαρακτηριστικών έχει πραγματοποιηθεί με τον ανιχνευτή SIFT και για αυτές που έχει χρησιμοποιηθεί πυκνή δειγματοληψία.

Η μαθηματική περιγραφή των εικόνων ολοκληρώνεται εφαρμόζοντας μία εκ των μεθόδων SP–VLAD, SPM και VLAD. Για τη μέθοδο SPM δημιουργούνται διανύσματα αναπαράστασης για 4 διαφορετικούς αριθμούς επιπέδων, συγκεκριμένα για $L = 0, 1, 2, 3$. Στην περίπτωση που $L = 0$, τα διανύσματα SPM ταυτίζονται με της μεθόδου BoW. Στη μέθοδο SP–VLAD επιλέγονται 3 διαφορετικοί αριθμοί επιπέδων, $L = 1, 2, 3$, και 3 διαφορετικοί τρόποι ανάθεσης βαρών, $w = 1, -1, -2$. Συγκεκριμένα, σε κάθε επίπεδο l, για $w = 1$ εφαρμόζονται τα βάρη 2^l που δίνουν μεγαλύτερη έμφαση στα ανώτερα επίπεδα, για $w = -1$ εφαρμόζονται τα βάρη $1/2^l$ τα οποία ενισχύουν τα χαμηλότερα επίπεδα, και για $w = -2$ εφαρμόζονται τα βάρη $1/2^{2l}$ τα οποία δίνουν ενισχυμένη έμφαση στα χαμηλότερα επίπεδα. Επίσης, για τις μεθόδους SP–VLAD και VLAD χρησιμοποιούνται οπτικά λεξικά μεγέθους $k = 16, 64, 128$, ενώ για τη μέθοδο SPM επιλέγεται $k = 200$. Για $k = 128$ και $L \geq 3$ η διάσταση των διανυσμάτων SP–VLAD ξεπερνάει το 1.000.000 και κρίνεται απαγορευτική η πραγματοποίηση οποιουδήποτε



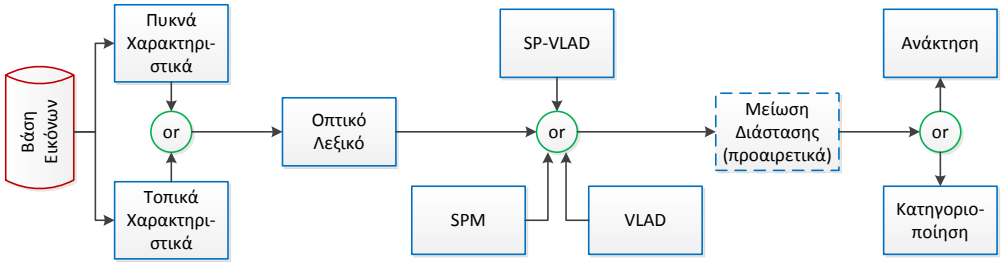
Σχήμα 4.2: Παραδείγματα εικόνων από τη βάση δεδομένων Caltech 101.

υπολογισμού με διανύσματα αυτού του μεγέθους. Συνεπώς, για $k = 128$, ο μέγιστος αριθμός επιπέδων που υπολογίζεται είναι $L = 2$. Σε όλες τις υπόλοιπες περιπτώσεις η μέγιστη τιμή επιπέδων που εφαρμόζεται είναι $L = 3$.

Η διάσταση των διανυσμάτων αναπαράστασης των εικόνων μειώνεται μέσω της μεθόδου PCA. Για την εκμάθηση της μεθόδου χρησιμοποιούνται τα διανύσματα κάθε πειράματος στο σύνολό τους. Μετά από αυτό το στάδιο, το διάνυσμα αναπαράστασης κάθε εικόνας είναι είτε 64 είτε 128 διαστάσεων. Το στάδιο αυτό είναι προαιρετικό υπό την έννοια ότι τα διανύσματα κάθε μεθόδου χρησιμοποιούνται και στην πλήρη τους διάσταση.

Ο τρόπος πραγματοποίησης της ανάκτησης αλλά και της κατηγοριοποίησης των εικόνων,

Σχήμα 4.3: Παραδείγματα εικόνων από τη βάση δεδομένων Flowers 15.



Σχήμα 4.4: Διαγραμματική αναπαράσταση της πειραματικής διαδικασίας.

καθώς και τα αντίστοιχα πειραματικά αποτελέσματα παρατίθενται στις δύο επόμενες ενότητες.

4.3 Ανάκτηση

Τρόπος Αξιολόγησης. Ο δείκτης που χρησιμοποιείται για την αξιολόγηση της ανάκτησης των εικόνων είναι ο mean average precision (mAP). Σε κάθε βάση επιλέγεται ένας αριθμός

Holidays (mAP)									
Μέθοδος	k	L	D ($\times 10^3$)	Τοπικά Χαρακτηριστικά			Πυκνά Χαρακτηριστικά		
				D	64	128	D	64	128
SPM	200	0	0.2	**0.449**	**0.403**	**0.417**	0.363	0.313	0.313
		1	1	0.443	0.389	0.410	**0.430**	**0.345**	0.349
		2	4	0.435	0.373	0.393	**0.430**	0.335	**0.376**
		3	17	0.406	0.318	0.350	0.374	0.281	0.311
VLAD	16	0	2	0.520	0.510	0.515	0.364	0.372	0.373
	64	0	8	0.543	**0.555**	0.572	0.383	0.396	0.400
	128	0	16	0.555	0.553	**0.589**	0.363	**0.401**	**0.404**
SP–VLAD $w = 1$	16	1	10	0.511	0.538	0.566	0.423	0.433	0.439
		2	43	0.090	0.553	0.553	0.407	0.443	0.457
		3	174	0.015	0.531	0.529	0.274	0.436	0.441
	64	1	41	0.334	0.566	0.609	0.414	0.449	0.460
		2	172	0.033	**0.570**	0.589	0.345	0.458	0.469
		3	696	0.008	0.525	0.540	0.181	0.453	0.466
	128	1	82	0.193	0.568	**0.613**	0.391	0.459	0.465
		2	344	0.021	0.558	0.600	0.296	**0.465**	**0.481**
SP–VLAD $w = -1$	16	1	10	0.532	0.540	0.566	0.427	0.433	0.438
		2	43	0.310	0.552	0.582	0.421	0.446	0.457
		3	174	0.072	0.563	0.591	0.318	0.444	0.454
	64	1	41	0.461	0.575	0.604	0.410	0.445	0.458
		2	172	0.103	**0.595**	0.621	0.374	0.461	0.467
		3	696	0.026	0.594	0.626	0.236	0.461	**0.482**
	128	1	82	0.337	0.578	0.612	0.393	0.454	0.461
		2	344	0.060	0.587	**0.629**	0.326	**0.465**	0.472
SP–VLAD $w = -2$	16	1	10	0.549	0.544	0.558	0.428	0.432	0.439
		2	43	0.518	0.548	0.572	0.428	0.439	0.454
		3	174	0.403	0.555	0.578	0.380	0.448	0.461
	64	1	41	0.539	0.573	0.602	0.414	0.442	0.454
		2	172	0.366	0.581	0.604	0.396	0.451	0.464
		3	696	0.169	0.578	0.614	0.308	**0.465**	**0.468**
	128	1	82	0.488	0.573	0.609	0.400	0.446	0.457
		2	344	0.222	**0.582**	**0.620**	0.354	0.456	0.470

Πίνακας 4.1: Τα αποτελέσματα της ανάκτησης εικόνων στη βάση δεδομένων INRIA Holidays.

από εικόνες−ερωτήματα (query images) οι οποίες χρησιμοποιούνται στη διαδικασία της ανάκτησης. Συγκεκριμένα, στη βάση INRIA Holidays επιλέγονται 500 εικόνες−ερωτήματα, μία από κάθε κατηγορία εικόνων που διαθέτει η βάση. Παρομοίως, από τη βάση Caltech

101 επιλέγονται 100 εικόνες−ερωτήματα. Κάθε μία προέρχεται από μία από τις 100 διαφορετικές κατηγορίες της βάσης. Με την ίδια λογική, από τη βάση Flowers 15 επιλέγονται 15 εικόνες−ερωτήματα, μία για κάθε γένος ανθοφόρων φυτών. Δοθείσης μίας εικόνας εισόδου λοιπόν, πραγματοποιείται εξαντλητική αναζήτηση μεταξύ όλων των εικόνων της βάσης και επιστρέφονται όλες οι εικόνες σε φθίνουσα σειρά ομοιότητας. Κατά αυτόν τον τρόπο μπορεί να υπολογιστεί ο δείκτης average precision (AP) και επαναλαμβάνοντας αυτή τη διαδικασία για όλες τις εικόνες−ερωτήματα μπορεί να υπολογιστεί ο δείκτης mAP. Η συνάρτηση ομοιότητας που επιλέγεται για τις μεθόδους SP–VLAD και VLAD είναι το τετράγωνο της ευκλείδειας απόστασης ενώ για τη μέθοδο SPM χρησιμοποιείται η συνάρτηση histogram intersection.

Αποτελέσματα. Στους πίνακες 4.1, 4.2 και 4.3 περιλαμβάνονται τα αποτελέσματα της ανάκτησης εικόνων για τις βάσεις INRIA Holidays, Caltech 101 και Flowers 15 αντίστοιχα. Σε κάθε πίνακα περιλαμβάνονται τα αποτελέσματα των μεθόδων SPM, VLAD και SP–VLAD. Ο αριθμός k αναφέρεται στο πλήθος των οπτικών λέξεων, το L στον αριθμό των επιπέδων της χωρικής πυραμίδας (η τιμή $L = 0$ αντιστοιχεί στις περιπτώσεις που δεν χρησιμοποιείται η πυραμίδα), το D στην πλήρη διάσταση των διανυσμάτων κάθε μεθόδου και το w στον τρόπο ανάθεσης βαρών. Παρουσιάζονται ξεχωριστά τα αποτελέσματα για τις περιπτώσεις που τα χαρακτηριστικά έχουν εξαχθεί με τη μέθοδο SIFT (τοπικά χαρακτηριστικά) και για αυτές που έχουν ανιχνευτεί με πυκνή δειγματοληψία (πυκνά χαρακτηριστικά). Σε κάθε περίπτωση, οι πειραματικοί υπολογισμοί πραγματοποιούνται χρησιμοποιώντας τα διανύσματα κάθε μεθόδου σε πλήρη διάσταση αλλά και μετά την εφαρμογή του PCA ώστε να έχουν μειωθεί στις 64 ή 128 διαστάσεις. Είναι σημαντικό να αναφέρουμε πως μειώνοντας το μέγεθος των διανυσμάτων σε ένα συγκεκριμένο αριθμό διαστάσεων καθίστανται συγκρίσιμα τα αποτελέσματα μεταξύ των διαφορετικών μεθόδων. Συνεπώς, για κάθε διαφορετική μέθοδο σημειώνεται με έντονους χαρακτήρες η μέγιστη τιμή που επιτυγχάνεται με μέγεθος διανυσμάτων 64 και 128 διαστάσεων αλλά και η συνολικά μέγιστη τιμή για τις περιπτώσεις που αυτή επιτυγχάνεται μέσω των διανυσμάτων πλήρους διάστασης της κάθε μεθόδου.

Αρχικά να σημειώσουμε πως και στις τρεις βάσεις εικόνων, τόσο συνολικά όσο και συγκεκριμένα για τις 64 και τις 128 διαστάσεις, η μέγιστη τιμή του mAP επιτυγχάνεται με τη μέθοδό μας, SP–VLAD. Αυτό ισχύει και για τα τοπικά αλλά και για τα πυκνά χαρακτηριστικά. Ωστόσο υπάρχουν πολλές επιπλέον παρατηρήσεις που μπορούμε να κάνουμε για την επίδραση των διαφορετικών παραμέτρων στην ακρίβεια της εκάστοτε ανάκτησης, εξάγοντας έτσι χρήσιμα συμπεράσματα για την συμπεριφορά της μεθόδου μας.

Η επίδραση της κάθε παραμέτρου εξαρτάται από το σύνολο εικόνων στο οποίο εφαρμόζεται η διαδικασία της ανάκτησης. Στον πίνακα 4.1 παρατηρούμε πως η μέθοδος SPM με τοπικά χαρακτηριστικά επιτυγχάνει τη μέγιστη τιμή της για $L = 0$, όπου τα διανύσματα SPM ταυτίζονται με της μεθόδου BoW. Η προσθήκη δηλαδή της χωρικής πυραμίδας μέσω της μεθόδου SPM δεν επιφέρει βελτίωση, αλλά μείωση της ακρίβειας. Η μέθοδος SPM με πυκνά χαρακτηριστικά παρουσιάζει μέγιστη τιμή για $L = 2$ αλλά με την προσθήκη ενός ακόμα επιπέδου, η ακρίβεια μειώνεται. Παρόμοια συμπεριφορά παρατηρούμε στη μέθοδο SP–VLAD. Για παράδειγμα, για τοπικά χαρακτηριστικά και στάθμιση με $w = 1$, η προσθήκη τρίτου επιπέδου έχει αρνητικό αντίκτυπο στην ακρίβεια της ανάκτησης, ανεξάρτητα από το μέγεθος του λεξι-

Caltech101 (mAP)									
Μέθοδος	k	L	D	Τοπικά Χαρακτηριστικά			Πυκνά Χαρακτηριστικά		
			($\times 10^3$)	D	64	128	D	64	128
SPM	200	0	0.2	0.178	0.154	0.152	0.201	0.183	0.182
		1	1	0.226	0.180	0.182	0.269	0.216	0.215
		2	4	0.282	**0.211**	**0.224**	0.318	0.257	0.261
		3	17	**0.309**	0.187	0.213	**0.336**	**0.271**	**0.279**
VLAD	16	0	2	0.222	0.228	0.229	0.229	0.240	0.238
	64	0	8	0.234	0.260	0.256	0.224	0.241	**0.239**
	128	0	16	0.252	**0.296**	**0.291**	0.216	**0.249**	0.238
SP–VLAD $w = 1$	16	1	10	0.241	0.291	0.292	0.288	0.300	0.304
		2	43	0.085	0.361	**0.365**	0.323	0.343	0.355
		3	174	0.041	**0.366**	0.363	0.317	0.340	0.358
	64	1	41	0.177	0.317	0.315	0.281	0.309	0.311
		2	172	0.05	0.349	0.352	0.315	0.345	0.356
		3	696	0.03	0.324	0.322	0.310	0.345	0.361
	128	1	82	0.137	0.329	0.338	0.271	0.311	0.317
		2	344	0.040	0.332	0.341	0.315	**0.350**	**0.362**
SP–VLAD $w = -1$	16	1	10	0.247	0.280	0.288	0.286	0.298	0.301
		2	43	0.165	0.336	0.346	0.319	0.340	0.345
		3	174	0.079	**0.363**	**0.370**	0.322	0.346	0.358
	64	1	41	0.211	0.303	0.301	0.280	0.304	0.308
		2	172	0.082	0.346	0.354	0.311	0.344	0.347
		3	696	0.048	0.355	0.365	0.316	**0.349**	**0.360**
	128	1	82	0.173	0.328	0.335	0.268	0.310	0.310
		2	344	0.062	0.349	0.365	0.311	0.344	0.354
SP–VLAD $w = -2$	16	1	10	0.248	0.262	0.269	0.282	0.295	0.298
		2	43	0.239	0.288	0.302	0.313	0.329	0.335
		3	174	0.190	0.303	0.319	0.319	0.334	0.345
	64	1	41	0.235	0.287	0.287	0.276	0.298	0.302
		2	172	0.157	0.313	0.323	0.304	0.332	0.336
		3	696	0.103	0.325	0.333	0.315	**0.341**	**0.347**
	128	1	82	0.220	0.318	0.327	0.265	0.305	0.303
		2	344	0.112	**0.341**	**0.349**	0.302	0.337	0.341

Πίνακας 4.2: Τα αποτελέσματα της ανάκτησης εικόνων στη βάση δεδομένων Caltech 101.

κού. Συμπεραίνουμε πως οι εικόνες του συνόλου INRIA Holidays δεν παρουσιάζουν μεγάλη χωρική ομοιότητα, όπως άλλωστε ήταν αναμενόμενο, καθώς πρόκειται για τυχαίες φωτογραφίες ή για εικόνες που προκύπτουν ως αφφινικοί μετασχηματισμοί άλλων εικόνων. Η έλλειψη

υψηλής γεωμετρικής συσχέτισης μεταξύ των εικόνων αυτής της βάσης επιβεβαιώνεται και από τις διαφορετικές τεχνικές που σταθμίζονται τα διανύσματα SP–VLAD όταν χρησιμοποιούνται τοπικά χαρακτηριστικά. Συγκεκριμένα, όταν η στάθμιση των διανυσμάτων γίνεται με $w = 1$ δίνοντας έμφαση στα υψηλότερα επίπεδα της πυραμίδας, η ακρίβεια που επιτυγχάνεται είναι αισθητά χαμηλότερη σε σχέση με τις αντίστοιχες περιπτώσεις που $w = -1$ ή $w = -2$. Αντίθετα, στη βάση Caltech 101 (πίνακας 4.2) η οποία διαθέτει εικόνες με υψηλή γεωμετρική ομοιότητα, η μέθοδος SPM παρουσιάζει μέγιστη ακρίβεια για $L = 3$. Επίσης, παρατηρώντας τις μέγιστες αποδόσεις της μεθόδου SP–VLAD, παρατηρούμε πως για $w = 1$ σε σχέση με την περίπτωση που $w = -1$, υφίσταται μικρότερο μέγιστο μόνο στην περίπτωση των τοπικών χαρακτηριστικών για 128 διαστάσεις. Επίσης, για $w = 1$ επιτυγχάνονται συνολικά καλύτερες μέγιστες αποδόσεις συγκριτικά με την περίπτωση που $w = -2$. Επιπλέον, η προσθήκη επιπέδων στη μέθοδο SP–VLAD βελτιώνει τα αποτελέσματα σχεδόν σε όλες τις περιπτώσεις, ανεξάρτητα από την τιμή των υπόλοιπων παραμέτρων. Αναφορικά με τη βάση Flowers 15 (πίνακας 4.3), παρατηρούμε πως η χωρική ομοιότητα μεταξύ των εικόνων είναι ιδιαιτέρως χαμηλή. Για τοπικά χαρακτηριστικά η μέθοδος SPM λειτουργεί καλύτερα για $L = 0, 1$ και η μέθοδος VLAD, η οποία δεν χρησιμοποιεί χωρική πληροφορία, παρουσιάζει υψηλότερες μέγιστες τιμές σε σχέση με την SP–VLAD ($w = 1$). Επίσης, η μέθοδος SP–VLAD με $w = -2$ παρουσιάζει συνολικά υψηλότερες επιδόσεις και από τη μέθοδο SP–VLAD με $w = 1$ αλλά και από την SP–VLAD με $w = -1$, αξιοποιώντας το γεγονός ότι δίνει την μικρότερη έμφαση στα υψηλά επίπεδα της χωρικής πυραμίδας. Στο ίδιο πλαίσιο, είναι ενδιαφέρον και το γεγονός ότι για τοπικά χαρακτηριστικά, οι μέγιστες τιμές που επιτυγχάνει η μέθοδος SP–VLAD, ανεξάρτητα από τον τρόπο στάθμισης, είναι με $L = 1$.

Στους πίνακες 4.1 και 4.3 είναι φανερό πως η χρήση των τοπικών χαρακτηριστικών σε σχέση με των πυκνών οδηγεί σε αισθητά καλύτερα αποτελέσματα όλες τις μεθόδους. Το γεγονός αυτό καταδεικνύει τη μη ύπαρξη χρήσιμης πληροφορίας σε όλη την έκταση των εικόνων των βάσεων INRIA Holidays και Flowers 15, καθώς σε αντίθετη περίπτωση τα πυκνά χαρακτηριστικά θα απέδιδαν υψηλότερα αποτελέσματα σε σχέση με τα τοπικά. Το ίδιο ισχύει και στη βάση Caltech 101 (πίνακας 4.2) ωστόσο η διαφορά στα αποτελέσματα μεταξύ των μεθόδων με τοπικά και πυκνά χαρακτηριστικά δεν είναι τόσο μεγάλη και μάλιστα σε ορισμένες περιπτώσεις, για τις ίδιες τιμές παραμέτρων, τα πυκνά χαρακτηριστικά εμφανίζουν καλύτερες επιδόσεις. Αυτό είναι λογικό για τη συγκεκριμένη βάση καθώς πολλές εικόνες έχουν έντονο υπόβαθρο το οποίο αποτυπώνεται μέσω των πυκνών χαρακτηριστικών ενισχύοντας έτσι τη διακριτική ικανότητα των διανυσμάτων αναπαράστασης κάθε εικόνας. Επίσης, υπάρχουν αρκετές εικόνες χωρίς καθόλου υπόβαθρο, γεγονός που επιτρέπει στα πυκνά χαρακτηριστικά να αποδώσουν λεπτομερώς τη χρήσιμη πληροφορία που περιλαμβάνεται στις περιοχές ενδιαφέροντος κάθε εικόνας.

Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός ότι η χρήση των διανυσμάτων SP–VLAD σε πλήρη διάσταση D με τοπικά χαρακτηριστικά οδηγεί και στις τρεις βάσεις δεδομένων σε αρκετά χαμηλές τιμές του δείκτη mAP, οι οποίες μάλιστα μειώνονται όσο προστίθενται περισσότερα επίπεδα L και όσο μεγαλώνει το μέγεθος του οπτικού λεξικού k. Ο λόγος που πιθανότατα συμβαίνει κάτι τέτοιο είναι ότι τα τοπικά χαρακτηριστικά που ανιχνεύονται είναι

Flowers15 (mAP)									
Μέθοδος	k	L	D	Τοπικά Χαρακτηριστικά			Πυκνά Χαρακτηριστικά		
			($\times 10^3$)	D	64	128	D	64	128
SPM	200	0	0.2	**0.268**	0.265	**0.266**	0.131	0.142	0.142
		1	1	0.259	**0.266**	0.264	0.154	0.169	0.169
		2	4	0.260	0.259	**0.266**	0.177	**0.189**	**0.186**
		3	17	0.261	0.245	0.250	0.181	0.180	0.184
VLAD	16	0	2	0.288	0.296	0.294	0.137	0.162	0.164
	64	0	8	0.301	0.308	0.303	0.158	0.187	0.189
	128	0	16	0.303	**0.320**	**0.306**	0.157	**0.193**	**0.194**
SP–VLAD $w = 1$	16	1	10	0.274	0.295	0.289	0.162	0.200	0.205
		2	43	0.129	0.275	0.260	0.183	0.218	0.222
		3	174	0.087	0.251	0.210	0.176	0.240	0.229
	64	1	41	0.265	**0.309**	**0.293**	0.169	0.218	0.216
		2	172	0.100	0.270	0.224	0.182	0.242	**0.241**
		3	696	0.083	0.234	0.208	0.175	**0.248**	0.233
	128	1	82	0.254	**0.309**	0.280	0.163	0.220	0.223
		2	344	0.094	0.251	0.220	0.179	0.239	0.236
SP–VLAD $w = -1$	16	1	10	0.281	0.298	0.291	0.162	0.199	0.205
		2	43	0.209	0.300	0.284	0.183	0.219	0.223
		3	174	0.112	0.293	0.270	0.181	0.239	0.230
	64	1	41	0.273	0.313	0.299	0.169	0.218	0.215
		2	172	0.140	0.300	0.272	0.181	0.239	0.241
		3	696	0.092	0.281	0.238	0.178	**0.251**	**0.242**
	128	1	82	0.270	**0.322**	**0.293**	0.162	0.220	0.223
		2	344	0.115	0.284	0.264	0.175	0.237	0.240
SP–VLAD $w = -2$	16	1	10	0.286	0.300	0.300	0.159	0.202	0.202
		2	43	0.277	0.305	0.292	0.181	0.213	0.222
		3	174	0.226	0.305	0.292	0.185	0.233	0.233
	64	1	41	0.292	0.309	**0.303**	0.169	0.213	0.217
		2	172	0.230	0.304	0.298	0.184	0.235	0.238
		3	696	0.149	0.304	0.287	0.181	**0.249**	**0.245**
	128	1	82	0.282	**0.325**	0.299	0.163	0.220	0.221
		2	344	0.176	0.316	0.288	0.177	0.237	0.241

Πίνακας 4.3: Τα αποτελέσματα της ανάκτησης εικόνων στη βάση δεδομένων Flowers 15.

περιορισμένα σε πλήθος και όταν χρησιμοποιούνται για τη δημιουργία διανυσμάτων εκατοντάδων χιλιάδων διαστάσεων, όπως αυτά της μεθόδου SP–VLAD, αποτυγχάνουν να αποδώσουν την κατανομή των οπτικών λέξεων με τρόπο που να προσδίδει στα διανύσματα ικανοποιητική

διακριτική ικανότητα. Συγκεκριμένα, όταν έχουμε $L \geq 2$, ο αριθμός των υποπεριοχών που δημιουργούνται είναι υψηλός και τα περιορισμένα διανύσματα περιγραφής αναμένεται να κατανέμονται με ανόμοιο τρόπο ακόμα και μεταξύ εικόνων υψηλής ομοιότητας, αφήνοντας αρκετές υποπεριοχές με μικρό αριθμό ή χωρίς καθόλου διανύσματα περιγραφής. Όπως είναι λογικό, το πρόβλημα αυτό εντείνεται όταν το μέγεθος του οπτικού λεξικού αυξάνεται καθώς καθίσταται ακόμα δυσκολότερο σε αντίστοιχες υποπεριοχές όμοιων εικόνων να υπάρχει ικανοποιητικός αριθμός διανυσμάτων που να αντιπροσωπεύει τις απαραίτητες οπτικές λέξεις. Παράλληλα, για $w = -1$ και κυρίως για $w = -2$ το πρόβλημα που μελετάμε μετριάζεται, υπογραμμίζοντας έτσι ότι έγκειται στα ανώτερα επίπεδα της πυραμίδας. Άλλωστε, το φαινόμενο αυτό δεν παρατηρείται στην περίπτωση των πυκνών χαρακτηριστικών τα οποία είναι πολλαπλάσια σε πλήθος σε σχέση με τα τοπικά και καλύπτουν ομοιόμορφα το χώρο των εικόνων. Ωστόσο, η εφαρμογή της μεθόδου PCA αλλάζει ολοκληρωτικά αυτή την εικόνα. Κατά αυτόν τον τρόπο αναδεικνύεται το γεγονός ότι τα διανύσματα SP–VLAD είναι πλούσια σε πληροφορία αλλά χρειάζεται καλύτερη αναδιοργάνωση της πληροφορίας αυτής, η οποία πραγματοποιείται μέσω της εφαρμογής του PCA. Η μέθοδος PCA παρόλο που μειώνει το μέγεθος των διανυσμάτων SP–VLAD και συνεπώς θα αναμέναμε να ελαττώνει και τη διακριτική τους ικανότητα, δημιουργεί διανύσματα των οποίων οι διαστάσεις είναι ασυσχέτιστες και διαθέτουν τη μέγιστη δυνατή διακύμανση. Σε αυτό το πλαίσιο λοιπόν, όταν υπάρχει πλεόνασμα πληροφορίας στα αρχικά διανύσματα SP–VLAD, είναι αναμενόμενο η εφαρμογή του PCA να βελτιώνει την ακρίβεια της ανάκτησης. Μάλιστα, σημαντικό στοιχείο αποτελεί το γεγονός πως, όπως έχουμε προαναφέρει, η εκμάθηση της μεθόδου PCA γίνεται με χρήση των διανυσμάτων SP–VLAD στα οποία στη συνέχεια θα εφαρμοστεί. Η εκμάθηση μέσω μέρους των διανυσμάτων αυτών ή τελείως ξένων διανυσμάτων SP–VLAD αναμένεται να οδηγήσει σε αισθητά χαμηλότερες τιμές τα αποτελέσματα. Παρατηρούμε άλλωστε πως και η μέθοδος VLAD ή ακόμα και η SPM εμφανίζουν σε αρκετές περιπτώσεις βελτιωμένα αποτελέσματα μετά την εφαρμογή του PCA.

Αναφορικά με το βαθμό στον οποίο είναι προτιμότερο να μειώσουμε το μέγεθος των διανυσμάτων SP–VLAD, σημαντικό ρόλο διαδραματίζει το μέγεθος της μνήμης που σκοπεύουμε να διαθέσουμε για τη διαδικασία της ανάκτησης. Παρατηρούμε ότι στις περισσότερες περιπτώσεις τα διανύσματα με 128 διαστάσεις δίνουν καλύτερα αποτέλεσμα από εκείνα με τις 64, αλλά συνήθως η διαφορά στην ακρίβεια δεν είναι μεγάλη, και συνεπώς μπορούμε να επιλέξουμε διανύσματα 64 διαστάσεων προκειμένου να εξοικονομήσουμε μνήμη εάν κάτι τέτοιο είναι επιθυμητό. Με την ίδια λογική, οι απαιτήσεις και η φύση κάθε εφαρμογής προσδιορίζουν τις τιμές και των υπολοίπων παραμέτρων, όπως είναι ο αριθμός των επιπέδων L, το μέγεθος του λεξικού k και η τεχνική ανάθεσης βαρών. Η αύξηση του μεγέθους του λεξικού έχει συνήθως θετικό αντίκτυπο στα αποτελέσματα καθώς επιτρέπει λεπτομερέστερη περιγραφή του οπτικού περιεχομένου των εικόνων, ωστόσο είναι σημαντικό να συνυπολογιστούν οι απαιτήσεις υπολογιστικών πόρων και μνήμης που προκύπτουν. Ο αριθμός των επιπέδων της χωρικής πυραμίδας που θα επιλεχθεί, εξαρτάται από το πόσο στενή γεωμετρική σχέση υπάρχει μεταξύ των εικόνων ίδιων κατηγοριών μίας βάσης. Η επιλογή της τεχνικής στάθμισης συνδέεται με το βαθμό στον οποίο συνεισφέρει η χωρική πληροφορία στη διαδικασία της ανάκτησης. Στην περίπτωση που θεωρούμε πως η χωρική πληροφορία έχει δευτερεύοντα ρόλο, όπως για παράδειγμα στη

βάση INRIA Holidays, είναι προτιμότερη η επιλογή $w = -1$ ή ακόμα και $w = -2$ παρά να παραληφθεί η χρήση της χωρικής πυραμίδας τελείως, καθώς τα αποτελέσματα βελτιώνονται σε σχέση με τη μέθοδο VLAD η οποία δεν χρησιμοποιεί καθόλου τη χωρική πληροφορία. Το γεγονός αυτό φανερώνει πως έστω και δευτερευόντως, η αξιοποίηση της χωρικής πληροφορίας αναμένεται να έχει ευεργετικά αποτελέσματα στην ανάκτηση εικόνων.

4.4 Κατηγοριοποίηση

Τρόπος Αξιολόγησης. Δοθείσης μίας εικόνας−ερώτημα και μίας βάσης με έναν αριθμό από διαφορετικές κατηγορίες εικόνων, η κατηγοριοποίηση στοχεύει στην σωστή επιλογή της κατηγορίας στην οποία ανήκει η εικόνα−ερώτημα. Η τεχνική που χρησιμοποιούμε για την πραγματοποίηση της κατηγοριοποίησης ονομάζεται k–Nearest Neighbors (k-NN). Σύμφωνα με τον αλγόριθμο k-NN, εντοπίζονται οι k εικόνες της βάσης οι οποίες παρουσιάζουν την υψηλότερη ομοιότητα με την εικόνα−ερώτημα και επιλέγεται η κατηγορία στην οποία ανήκει η πλειονότητα αυτών των k εικόνων. Συνεπώς, εάν η πλειονότητα των k εικόνων ανήκει στην ίδια κατηγορία με την εικόνα−ερώτημα, η εικόνα−ερώτημα κατηγοριοποιείται σωστά. Στη βάση Caltech 101 η οποία διαθέτει 100 κατηγορίες εικόνων με 10 εικόνες στην κάθε μία, ορίζεται $k = 10$. Στη βάση Flowers 15 που διαθέτει 30 εικόνες σε κάθε κατηγορία, ορίζεται $k = 30$. Η αξιολόγηση της κατηγοριοποίησης πραγματοποιείται υπολογίζοντας τον μέσο όρο των σωστών κατηγοριοποιήσεων για όλες τις εικόνες μίας κατηγορίας. Για παράδειγμα, στη βάση Caltech 101 θεωρούνται διαδοχικά ως εικόνες−ερωτήματα κάθε μία από τις 10 εικόνες μίας κατηγορίας και υπολογίζεται το ποσοστό όσων έχουν κατηγοριοποιηθεί σωστά. Επαναλαμβάνοντας αυτή τη διαδικασία και για τις 100 κατηγορίες εικόνων, μπορούμε να υπολογίσουμε τον μέσο όρο των ποσοστών κατηγοριοποίησης συνολικά για όλες τις κατηγορίες εικόνων, υπολογίζοντας έτσι το βαθμό κατηγοριοποίησης (classification rate). Με αυτόν τον τρόπο αξιολογείται η κατηγοριοποίηση των εικόνων από το σύστημα μας και τα σχετικά αποτελέσματα περιλαμβάνονται στους πίνακες 4.4 και 4.5 για τις βάσεις Caltech 101 και Flowers 15 αντίστοιχα. Οι πίνακες είναι δομημένοι ακριβώς όπως οι αντίστοιχοι πίνακες της ανάκτησης εικόνων. Επίσης, η συνάρτηση ομοιότητας που επιλέγεται στον αλγόριθμό k-NN για τις μεθόδους SP–VLAD και VLAD είναι το τετράγωνο της ευκλείδειας απόστασης ενώ για τη μέθοδο SPM χρησιμοποιείται η συνάρτηση histogram intersection.

Αποτελέσματα. Η μέθοδός μας, SP–VLAD, επιτυγχάνει και στις δύο βάσεις εικόνων αισθητά καλύτερες μέγιστες τιμές από τις υπόλοιπες μεθόδους, τόσο για διανύσματα μεγέθους 64 όσο και 128 διαστάσεων. Επιπλέον, και στις δύο βάσεις εικόνων, οι επιδόσεις όλων των μεθόδων είναι αδιαμφισβήτητα καλύτερες όταν χρησιμοποιούνται τοπικά αντί για πυκνά χαρακτηριστικά. Γενικότερα, η συμπεριφορά όλων των μεθόδων συναρτήσει των διαφόρων παραμέτρων είναι πανομοιότυπη με αυτήν που παρατηρήθηκε και αναλύθηκε στην ανάκτηση εικόνων. Ωστόσο, αναλύοντας την επίδοση της μεθόδου μας στις επιμέρους κατηγορίες εικόνων μπορούμε να εξάγουμε επιπρόσθετα πολύτιμα συμπεράσματα για τη συμπεριφορά της.

Στη βάση Caltech 101 η μέθοδος SP–VLAD παρατηρούμε πως παρουσιάζει τα υψηλότερα ποσοστά σωστής κατηγοριοποίησης σε κατηγορίες εικόνων οι οποίες διαθέτουν όμοια

Caltech101 (Classification Rate)									
Μέθοδος	k	L	D	Τοπικά Χαρακτηριστικά			Πυκνά Χαρακτηριστικά		
			($\times 10^3$)	D	64	128	D	64	128
SPM	200	0	0.2	0.378	0.403	0.399	0.442	0.403	0.400
		1	1	0.463	0.447	0.436	0.521	0.441	0.439
		2	4	0.561	**0.474**	**0.482**	0.542	**0.470**	0.479
		3	17	**0.618**	0.445	0.458	**0.564**	0.461	**0.489**
VLAD	16	0	2	0.510	0.503	0.508	0.465	0.458	**0.463**
	64	0	8	0.512	0.538	0.557	0.442	**0.481**	0.462
	128	0	16	0.512	**0.576**	**0.587**	0.444	0.474	0.460
SP–VLAD $w = 1$	16	1	10	0.507	0.569	0.569	0.533	0.535	0.549
		2	43	0.088	**0.670**	**0.693**	0.538	0.576	0.583
		3	174	0.021	0.646	0.658	0.539	**0.590**	0.625
	64	1	41	0.311	0.588	0.635	0.516	0.547	0.556
		2	172	0.038	0.610	0.664	0.527	0.587	0.589
		3	696	0.015	0.539	0.576	0.520	0.577	**0.627**
	128	1	82	0.217	0.615	0.651	0.510	0.546	0.544
		2	344	0.021	0.618	0.650	0.529	0.589	0.608
SP–VLAD $w = -1$	16	1	10	0.523	0.558	0.574	0.543	0.534	0.544
		2	43	0.310	0.622	0.664	0.544	0.567	0.578
		3	174	0.083	**0.652**	**0.697**	0.550	0.586	0.615
	64	1	41	0.420	0.587	0.620	0.524	0.547	0.549
		2	172	0.098	0.614	**0.697**	0.540	0.583	0.593
		3	696	0.034	0.605	0.658	0.541	**0.589**	**0.620**
	128	1	82	0.311	0.587	0.641	0.499	0.547	0.548
		2	344	0.070	0.607	0.672	0.518	0.574	0.578
SP–VLAD $w = -2$	16	1	10	0.528	0.539	0.568	0.529	0.529	0.549
		2	43	0.529	0.571	0.598	0.549	0.546	0.576
		3	174	0.381	0.580	0.629	0.552	0.568	0.585
	64	1	41	0.521	0.565	0.593	0.516	0.539	0.549
		2	172	0.295	0.592	0.632	0.545	**0.583**	0.592
		3	696	0.149	0.597	0.632	0.546	0.571	**0.606**
	128	1	82	0.451	0.579	0.630	0.496	0.540	0.533
		2	344	0.163	**0.605**	**0.657**	0.518	0.582	0.587

Πίνακας 4.4: Τα αποτελέσματα της κατηγοριοποίησης εικόνων στη βάση δεδομένων Caltech 101.

γεωμετρική διάταξη όπως οι εικόνες της κατηγορίας "Accordion" ή της κατηγορίας "Car". Ενδεικτικά, στις κατηγορίες που θα αναφέρουμε θα παραθέτουμε το ποσοστό κατηγοριοποί-

Flowers15 (Classification Rate)									
Μέθοδος	k	L	D	Τοπικά Χαρακτηριστικά			Πυκνά Χαρακτηριστικά		
			($\times 10^3$)	D	64	128	D	64	128
SPM	200	0	0.2	0.504	0.489	0.482	0.382	0.338	0.336
		1	1	0.513	0.496	**0.498**	0.400	0.367	0.369
		2	4	0.502	**0.500**	0.478	0.362	0.404	0.416
		3	17	**0.516**	0.491	0.464	0.302	**0.407**	**0.431**
VLAD	16	0	2	0.533	0.593	0.567	0.380	0.380	0.371
	64	0	8	0.540	0.638	0.627	0.380	0.391	**0.396**
	128	0	16	0.649	**0.680**	**0.680**	0.373	**0.393**	0.373
SP–VLAD $w = 1$	16	1	10	0.536	0.598	0.584	0.400	0.433	0.429
		2	43	0.364	0.653	0.642	0.327	0.500	0.498
		3	174	0.080	0.616	0.593	0.264	0.551	0.542
	64	1	41	0.567	0.662	0.607	0.389	0.469	0.438
		2	172	0.178	0.676	0.651	0.342	0.529	0.549
		3	696	0.069	0.571	0.578	0.336	**0.560**	**0.578**
	128	1	82	0.627	**0.689**	**0.669**	0.413	0.480	0.469
		2	344	0.147	0.684	0.651	0.347	0.527	0.567
SP–VLAD $w = -1$	16	1	10	0.536	0.627	0.602	0.389	0.433	0.431
		2	43	0.582	0.656	0.611	0.344	0.473	0.493
		3	174	0.273	0.669	0.653	0.298	0.527	0.547
	64	1	41	0.549	0.680	0.638	0.398	0.460	0.431
		2	172	0.424	0.709	0.656	0.398	0.516	0.549
		3	696	0.151	0.698	0.658	0.336	**0.564**	**0.562**
	128	1	82	0.627	**0.720**	0.665	0.420	0.469	0.456
		2	344	0.247	0.713	**0.698**	0.371	0.531	0.525
SP–VLAD $w = -2$	16	1	10	0.518	0.636	0.598	0.400	0.429	0.436
		2	43	0.573	0.649	0.596	0.407	0.471	0.476
		3	174	0.562	0.665	0.609	0.347	0.502	0.513
	64	1	41	0.549	0.651	0.600	0.413	0.467	0.456
		2	172	0.567	0.678	0.644	0.416	0.502	0.500
		3	696	0.453	0.671	0.653	0.389	**0.531**	**0.538**
	128	1	82	0.644	0.693	0.673	0.411	0.469	0.447
		2	344	0.518	**0.727**	**0.713**	0.400	0.500	0.496

Πίνακας 4.5: Τα αποτελέσματα της κατηγοριοποίησης εικόνων στη βάση δεδομένων Flowers 15.

ησης που επιτυγχάνει η μέθοδος SP–VLAD για κάθε μία από αυτές όταν οι παράμετροι είναι τέτοιες ώστε συνολικά να επιτυγχάνεται ο μέγιστος δυνατός βαθμός κατηγοριοποίησης. Η

τιμή αυτή του βαθμού κατηγοριοποίησης είναι 0.697 και εντοπίζεται στον πίνακα 4.4 για τοπικά χαρακτηριστικά, διανύσματα μεγέθους 128 διαστάσεων, $w = -1$, $k = 16$ και $L = 3$. Στις προαναφερθείσες κατηγορίες "Accordion" και "Car" λοιπόν, η μέθοδος SP–VLAD με τις παραμέτρους που προαναφέραμε, επιτυγχάνει ποσοστό κατηγοριοποίησης 100%. Τα χαμηλότερα ποσοστά σωστής κατηγοριοποίησης παρουσιάζονται σε κατηγορίες εικόνων οι οποίες εμφανίζουν είτε τυχαία γεωμετρική δομή όπως είναι η κατηγορία "Octopus" (0%) είτε έντονο υπόβαθρο όπως είναι για παράδειγμα εικόνες ζώων στο φυσικό τους περιβάλλον. Για παράδειγμα, τέτοιες είναι οι κατηγορίες "Rhino" (10%), "Rooster" (30%) και "Elephant" (20%). Αντίθετα, σε κατηγορίες ζώων με μονόχρωμο ή καθόλου υπόβαθρο, ακόμα κι αν παρουσιάζουν ιδιαίτερη υφή, επιτυγχάνονται αρκετά υψηλότερα ποσοστά, όπως για παράδειγμα στην κατηγορία "Hedgehog" (70%). Προς αυτή την κατεύθυνση ενδεικτική είναι η μεγάλη διαφορά που παρουσιάζεται μεταξύ των κατηγοριών "Cougar Body" (50%) και "Cougar Face" (100%) καθώς και των "Flamingo" (40%) και "Flamingo Head" (70%). Και στις δύο αυτές περιπτώσεις, όταν το Πούμα ή το Φλαμίνγκο αναπαρίστανται ολόκληρα εισάγεται πληροφορία από το υπόβαθρο η οποία δυσχεραίνει την κατηγοριοποίηση. Όταν όμως αναπαριστάται μόνο το κεφάλι των ζώων αυτών, τα αποτελέσματα είναι αισθητά καλύτερα. Εξαιρετικά είναι τα ποσοστά σε κατηγορίες όπως η "Windsor Chair" (100%) στην οποία οι εικόνες έχουν μονόχρωμο υπόβαθρο και τα αντικείμενα δεν έχουν έντονη υφή αλλά διαθέτουν λεπτές αναλογίες. Ενδεικτικά παραδείγματα εικόνων από όλες τις κατηγορίες που προαναφέραμε περιλαμβάνονται στο σχήμα 4.2.

Στη βάση Flowers 15 επιβεβαιώνεται το γεγονός ότι στα ανθοφόρα φυτά η χωρική πληροφορία δεν αποτελεί των πρωτεύοντα παράγοντα ενίσχυσης της ακρίβειας, καθώς η μέθοδος SPM περιλαμβάνει τα χαμηλότερα ποσοστά κατηγοριοποίησης ενώ τα υψηλότερα επιτυγχάνονται με την μέθοδο SP–VLAD για $w = -2$. Επίσης, τα πυκνά χαρακτηριστικά οδηγούν και πάλι σε αξιοσημείωτα χαμηλότερα ποσοστά καθιστώντας τη χρήση των τοπικών χαρακτηριστικών τη μόνη ενδεδειγμένη επιλογή. Παρατηρώντας την επίδοση της μεθόδου SP–VLAD στις επιμέρους κατηγορίες φυτών όταν χρησιμοποιούνται τοπικά χαρακτηριστικά, $w = -2$, αριθμός διαστάσεων 128, $k = 128$ και $L = 2$, συμπεραίνουμε πως η σύγχυση στην κατηγοριοποίηση υφίσταται μεταξύ φυτών που μοιάζουν οπτικά. Τέτοιες είναι για παράδειγμα οι κατηγορίες "Malva" (47%) και "Nerium" (37%). Αντίθετα, σε κατηγορίες φυτών όπως οι "Passiflora" (83%) και "Rosa" (83%) που έχουν ιδιαίτερα οπτικά χαρακτηριστικά, τα ποσοστά κατηγοριοποίησης είναι αρκετά υψηλότερα. Επομένως, για την κατηγοριοποίηση σε βάσεις με εικόνες ανθοφόρων φυτών μπορούμε να συμπεράνουμε πως μπορούν να επιτευχθούν αρκετά καλά αποτελέσματα όταν η συγγένεια των φυτών που επιχειρούμε να κατηγοροποιήσουμε δεν είναι ιδιαιτέρως κοντινή. Άλλωστε, και στη βάση Caltech 101 περιλαμβάνονται 3 κατηγορίες φυτών, οι "Sunflower" (100%), "Lotus" (90%) και "Water Lilly" (60%). Παρατηρούμε πως τα ποσοστά κατηγοριοποίησης είναι αρκετά υψηλά, γεγονός που φανερώνει ότι οι εικόνες φυτών μπορούν να κατηγοριοποιηθούν σωστά με μεγάλη επιτυχία σε βάσεις στις οποίες περιλαμβάνονται πολλές εικόνες τελείως διαφορετικών κατηγοριών.

Κεφάλαιο 5

Συμπεράσματα και Μελλοντικές Επεκτάσεις

Στόχος αυτού του κεφαλαίου είναι να συνοψίσουμε τα αποτελέσματα από την ερευνητική εργασία που πραγματοποιήθηκε στο πλαίσιο αυτής της διπλωματικής και να προτείνουμε μελλοντικά βήματα τα οποία μπορούν να συντελεστούν προκειμένου να εξελιχθεί και να δοκιμαστεί περαιτέρω η μέθοδός μας, SP–VLAD, τόσο αναφορικά με το πρόβλημα της ανάκτησης όσο και της κατηγοριοποίησης εικόνων.

5.1 Συμπεράσματα

Στο πλαίσιο αυτής της διπλωματικής εργασίας προτείναμε και υλοποιήσαμε τη μέθοδο SP–VLAD η οποία χρησιμοποιείται για την διανυσματική αναπαράσταση εικόνων, και εφαρμόστηκε ως δομικό στοιχείο σε συστήματα ανάκτησης αλλά και κατηγοριοποίησης εικόνων. Η έρευνά μας εφαρμόστηκε στις βάσεις εικόνων INRIA Holidays, Caltech 101 και Flowers 15 και συγκρίθηκε με τις μεθόδους SPM και VLAD. Τα συμπεράσματα που προέκυψαν είναι τα ακόλουθα:

- Στο πρόβλημα της ανάκτησης εικόνων η μέθοδος SP–VLAD επιτυγχάνει συνολικά καλύτερα αποτελέσματα σε όλες τις βάσεις εικόνων και από τη μέθοδο SPM και από τη μέθοδο VLAD. Το ίδιο ισχύει και για το πρόβλημα της κατηγοριοποίησης εικόνων. Η συνολικά ανώτερη επίδοση της μεθόδου SP–VLAD φανερώνει πως η πληροφορία που περιλαμβάνει είναι πλουσιότερη σε σχέση με των υπόλοιπων μεθόδων, οδηγώντας σε διανύσματα αναπαράστασης υψηλότερης διακριτικής ικανότητας.

- Αναφορικά με τις παραμέτρους της μεθόδου SP–VLAD, ο αριθμός των επιπέδων L της χωρικής πυραμίδας αντικατοπτρίζει το βαθμό στον οποίο υπάρχει γεωμετρική συσχέτιση μεταξύ των εικόνων. Το μέγεθος k του οπτικού λεξικού ορίζει το πόσο λεπτομερώς περιγράφεται το οπτικό περιεχόμενο των εικόνων. Η τεχνική ανάθεσης βαρών προσδιορίζει το πόσο κρίσιμη θεωρείται η χωρική πληροφορία προκειμένου να εξασφαλιστεί η υψηλότερη δυνατή διακριτική ικανότητα μεταξύ των διανυσμάτων SP–VLAD. Η επιλογή

της τιμής της κάθε παραμέτρου εξαρτάται από τη φύση των εικόνων στις οποίες θα εφαρμοστεί η μέθοδος και τις απαιτήσεις σε υπολογιστικούς πόρους και μέγεθος μνήμης της εκάστοτε εφαρμογής.

- Η ανίχνευση χαρακτηριστικών μέσω πυκνής δειγματοληψίας οδήγησε σε αισθητά χαμηλότερα πειραματικά αποτελέσματα τη μέθοδό μας σε σχέση με τα τοπικά χαρακτηριστικά και συνεπώς ενδείκνυται η χρησιμοποίηση ενός ανιχνευτή όπως ο SIFT.

- Η μέθοδός μας ανταποκρίνεται πολύ θετικά στη μείωση του μεγέθους των διανυσμάτων SP–VLAD μέσω της μεθόδου PCA, εξασφαλίζοντας έτσι πολύ πιο οικονομική αναπαράσταση αλλά και υψηλότερη ακρίβεια ανάκτησης και κατηγοριοποίησης καθώς απαλλάσσει τα αρχικά διανύσματα SP–VLAD από την πλεονάζουσα πληροφορία που εμπεριέχουν.

- Τα επίπεδα ακρίβειας που επιτυγχάνουν τα διανύσματα SP–VLAD σε πλήρη διάσταση είναι χαμηλότερα από των μεθόδων SPM και VLAD παρά το γεγονός ότι για τις ίδιες τιμές των παραμέτρων έχουν μεγαλύτερο μέγεθος. Μετά την μείωση της διάστασης των διανυσμάτων της κάθε μεθόδου, η μέθοδος SP–VLAD επιτυγχάνει συνολικά καλύτερα αποτελέσματα από τις μεθόδους SPM και VLAD τόσο για τις 128 όσο και για τις 64 διαστάσεις.

- Δημιουργήθηκε και χρησιμοποιήθηκε η βάση Flowers 15, με συνολικά 450 εικόνες από 15 διαφορετικά γένη ανθοφόρων φυτών, και προέκυψε πως η χωρική πληροφορία διαδραματίζει δευτερεύοντα ρόλο στην ανάκτηση και την κατηγοριοποίηση των εικόνων αυτών. Επιπροσθέτως, τα πυκνά χαρακτηριστικά δεν κατάφεραν να αποδώσουν με υψηλότερη ακρίβεια σε σχέση με τα τοπικά χαρακτηριστικά την πληροφορία που περιλαμβάνεται στις εικόνες ανθοφόρων φυτών.

5.2 Μελλοντικές Επεκτάσεις

Πρωτεύων μελλοντικός μας στόχος είναι η περαιτέρω βελτίωση των αποτελεσμάτων που επιτυγχάνει η μέθοδός μας, SP–VLAD, στα προβλήματα της ανάκτησης και της κατηγοριοποίησης εικόνων. Επιπροσθέτως, είναι επιθυμητή η δοκιμή της μεθόδου SP–VLAD σε μεγαλύτερο πλήθος εικόνων αλλά και εύρος κατηγοριών. Η ερευνητική μας εργασία λοιπόν, μπορεί να επεκταθεί μελλοντικά προς τις ακόλουθες κατευθύνσεις:

- Η διαδικασία της κατηγοριοποίησης των εικόνων μπορεί να πραγματοποιηθεί μέσω μηχανικής εκμάθησης (machine learning). Για παράδειγμα, η μέθοδος SPM έχει εφαρμοστεί στο πρόβλημα του classification με χρήση Support Vector Machines (SVM) επιτυγχάνοντας υψηλά ποσοστά αναγνώρισης στη βάση Caltech 101 [25].

- Στα διανύσματα VLAD που υπολογίζονται σε κάθε υποπεριοχή της χωρικής πυραμίδας μπορούμε να εφαρμόσουμε εναλλακτικές τεχνικές κανονικοποίησης όπως είναι η Intra–normalization [2].

- Δεδομένης της επιτυχημένης εφαρμογής των τοπικών χαρακτηριστικών μέσω του SIFT detector και descriptor, θα μπορούσαν να δοκιμαστούν και εναλλακτικές τεχνικές ανίχνευσης και περιγραφής χαρακτηριστικών όπως για παράδειγμα ο Hessian affine [30] και ο DAISY descriptor [41].

- Η μέθοδος SP–VLAD αναφορικά με το πρόβλημα της ανάκτησης αλλά και της κατηγοριοποίησης εικόνων, μπορεί να δοκιμαστεί σε επιπλέον βάσεις δεδομένων με διαφορετικές κατηγορίες εικόνων όπως είναι π.χ. η βάση κτιρίων Oxford Buildings [33] ή η University of Kentucky Benchmark [31].

- Δεδομένου του μικρού μεγέθους των διανυσμάτων SP–VLAD μετά την εφαρμογή του PCA, η μέθοδός μας θα μπορούσε να εφαρμοστεί σε μεγαλύτερες βάσεις δεδομένων με εκατομμύρια εικόνες χρησιμοποιώντας προσεγγιστικές μεθόδους αναζήτησης όπως έχει επιχειρηθεί για τη μέθοδο VLAD [20].

- Προκειμένου να επιτύχουμε μεγαλύτερη γενίκευση και ευκολότερη σύγκριση των αποτελεσμάτων μας, μπορούμε να εξάγουμε το οπτικό λεξικό και να πραγματοποιήσουμε την εκμάθηση της μεθόδου PCA χρησιμοποιώντας σύνολα εικόνων διαφορετικά από αυτά στα οποία πρόκειται μετέπειτα να εφαρμόσουμε τη μέθοδό μας.

Βιβλιογραφία

[1] Alexandre Alahi, Raphael Ortiz και Pierre Vandergheynst. Freak: Fast retina keypoint. Στο *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, σελίδες 510–517. IEEE, 2012.

[2] Relja Arandjelovic και Andrew Zisserman. All about vlad. Στο *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, σελίδες 1578–1585. IEEE, 2013.

[3] Herbert Bay, Tinne Tuytelaars και Luc Van Gool. Surf: Speeded up robust features. Στο *Computer Vision–ECCV 2006*, σελίδες 404–417. Springer, 2006.

[4] Christopher M Bishop και others. *Pattern recognition and machine learning*, τόμος 1. springer New York, 2006.

[5] C++ Boost. Libraries.(2012), 2012.

[6] G. Bradski. *Dr. Dobb's Journal of Software Tools*, 2000.

[7] Matthew Brown και David G Lowe. Invariant features from interest point groups. Στο *BMVC*, αριθμός s 1, 2002.

[8] Ritendra Datta, Dhiraj Joshi, Jia Li και James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[9] Richard O Duda, Peter E Hart και David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[10] L. Fei-Fei, R. Fergus και P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. Στο *Workshop on Generative-Model Based Vision*, 2004.

[11] Li Fei-Fei, Rob Fergus και Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[12] GigaOM. Facebook has 220 billion of your photos to put on ice, 2012.

[13] Kristen Grauman και Bastian Leibe. *Visual object recognition.* Αριθμός 11. Morgan & Claypool Publishers, 2011.

[14] Allan Hanbury. Image segmentation by region based and watershed algorithms. *Wiley Encyclopedia of Computer Science and Engineering*, 2008.

[15] Chris Harris και Mike Stephens. A combined corner and edge detector. Στο *Alvey vision conference*, τόμος 15, σελίδα 50. Manchester, UK, 1988.

[16] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[17] Image και Video Analysis Group (IVA). Flower visual image retrieval, 2008-2013.

[18] Herve Jegou, Matthijs Douze και Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. Στο *Computer Vision–ECCV 2008*, σελίδες 304–317. Springer, 2008.

[19] Hervé Jégou, Matthijs Douze και Cordelia Schmid. On the burstiness of visual elements. Στο *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, σελίδες 1169–1176. IEEE, 2009.

[20] Hervé Jégou, Matthijs Douze, Cordelia Schmid και Patrick Pérez. Aggregating local descriptors into a compact image representation. Στο *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, σελίδες 3304–3311. IEEE, 2010.

[21] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez και Cordelia Schmid. Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1704–1716, 2012.

[22] Eiji Kasutani και Akio Yamada. The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. Στο *Image Processing, 2001. Proceedings. 2001 International Conference on*, τόμος 1, σελίδες 674–677. IEEE, 2001.

[23] Toshikazu Kato, Takio Kurita, Nobuyuki Otsu και Kyoji Hirata. A sketch retrieval method for full color image database-query by visual example. Στο *Pattern Recognition, 1992. Vol. I. Conference A: Computer Vision and Applications, Proceedings., 11th IAPR International Conference on*, σελίδες 530–533. IEEE, 1992.

[24] Jan J Koenderink. The structure of images. *Biological cybernetics*, 50(5):363–370, 1984.

[25] Svetlana Lazebnik, Cordelia Schmid και Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. Στο *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, τόμος 2, σελίδες 2169–2178. IEEE, 2006.

[26] Tony Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994.

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[28] P. Maragos. *Image Analysis and Computer Vision*. 2013.

[29] Krystian Mikolajczyk και Cordelia Schmid. An affine invariant interest point detector. Στο *Computer Vision–ECCV 2002*, σελίδες 128–142. Springer, 2002.

[30] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir και Luc Van Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1-2):43–72, 2005.

[31] David Nister και Henrik Stewenius. Scalable recognition with a vocabulary tree. Στο *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, τόμος 2, σελίδες 2161–2168. IEEE, 2006.

[32] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[33] J. Philbin, O. Chum, M. Isard, J. Sivic και A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. Στο *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[34] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic και Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. Στο *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, σελίδες 1–8. IEEE, 2007.

[35] Yong Rui, Thomas S Huang και Shih Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation*, 10(1):39–62, 1999.

[36] Jianbo Shi και Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[37] Josef Sivic και Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. Στο *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, σελίδες 1470–1477. IEEE, 2003.

[38] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta και Ramesh Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[39] Markus A Stricker και Markus Orengo. Similarity of color images. Στο *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*, σελίδες 381–392. International Society for Optics and Photonics, 1995.

[40] t. f. e. Wikipedia. Apg iii system, 2013.

[41] E. Tola, V.Lepetit και P. Fua. A Fast Local Descriptor for Dense Matching. Στο *Proceedings of Computer Vision and Pattern Recognition*, Alaska, USA, 2008.

[42] Tinne Tuytelaars. Dense interest points. Στο *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, σελίδες 2281–2288. IEEE, 2010.

[43] Andrew Witkin. Scale-space filtering: A new approach to multi-scale description. Στο *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, τόμος 9, σελίδες 150–153. IEEE, 1984.

[44] Ian H Witten, Alistair Moffat και Timothy C Bell. *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann, 1999.

Κεφάλαιο 6

Ορολογία

Αγγλικός Όρος	Ελληνική Μετάφραση
accuracy	ακρίβεια
aggregating descriptors	άθροιση διανυσμάτων χαρακτηριστικών
aliasing	αναδίπλωσης
broad domain	ευρύ πεδίο
classification rate	βαθμός κατηγοριοποίησης
cluster	συστάδα
clustering	συσταδοποίηση
content-based retrieval	ανάκτηση βασισμένη στο περιεχόμενο
covariance matrix	πίνακας συμμεταβλητότητας
database	βάση δεδομένων
dense features	πυκνά χαρακτηριστικά
dense sampling	πυκνή δειγματοληψία
descriptor	διάνυσμα περιγραφής
detector	ανιχνευτής
determinant	ορίζουσα
efficiency	αποδοτικότητα
eigenvalue	ιδιοτιμή
eigenvector	ιδιοδιάνυσμα
feature vector	διάνυσμα χαρακτηριστικών
feature description	περιγραφή χαρακτηριστικών
feature detection	ανίχνευση χαρακτηριστικών
format	μορφότυπο
gradient	παράγωγος
hash table	πίνακας κατακερματισμού
heat diffusion equation	εξίσωση διάχυσης θερμότητας
Hessian matrix	Εσσιανος πίνακας
hue	απόχρωση

image classification	κατηγοριοποίηση εικόνων
intensity	ένταση
invariance	ανεξαρτησία
keypoint	σημείο ενδιαφέροντος
local features	τοπικά χαρακτηριστικά
machine learning	μηχανική εκμάθηση
metadata	μεταδεδομένα
narrow domain	στενό πεδίο
orientation	προσανατολισμός
pixel	εικονοστοιχείο
principal component	κύρια συνιστώσα
principal curvature	κύρια καμπυλότητα
principal subspace	κύριος υποχώρος
quantizer	κβαντιστής
query image	εικόνα-ερώτημα
rotation	περιστροφή
saturation	κορεσμός
scale	κλίμακα
scale space	χώρος κλίμακας
search tree	δένδρο αναζήτησης
semantic gap	σημασιολογικό κενό
sensory gap	αισθητήριο κενό
spatial pyramid	χωρική πυραμίδα
text-based retrieval	ανάκτηση βασισμένη στο κείμενο
texture	υφή
trace	ίχνος
trilinear interpolation	τριγραμμική παρεμβολή
visual vocabulary	οπτικό λεξικό
visual word	οπτική λέξη