Γρήγορη Αναζήτηση Εικόνων με βάση τα Οπτικά Χαρακτηριστικά και τη Γεωμετρία

Ιωάννης Καλαντίδης

January 19, 2009

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Γρήγορη Αναζήτηση Εικόνων με βάση τα Οπτικά Χαρακτηριστικά και τη Γεωμετρία

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ιωάννη Δ. Καλαντίδη

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2009

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Γρήγορη Αναζήτηση Εικόνων με βάση τα Οπτικά Χαρακτηριστικά και τη Γεωμετρία

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ιωάννη Δ. Καλαντίδη

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την ??η Ιανουαρίου 2009.

.....	.....	.....
Στέφανος Κόλλιας	Σταφυλοπάτης Ανδρέας-Γεώργιος	Τσανάκας Παναγιώτης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2009

.....

Ιωάννης Δ. Καλαντίδης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα τελευταία χρόνια έχουμε μια ραγδαία αύξηση του αριθμού των εικόνων που συναντάμε σε ψηφιακή μορφή. Σήμερα υπάρχουν διαθέσιμες στο διαδίκτυο τεράστιες συλλογές, ποικίλες ως προς το οπτικό και σημασιολογικό τους περιεχόμενο, που συνεχίζουν να μεγαλώνουν συνεχώς. Η σωστή αρχειοθέτηση όλου αυτού του όγκου πληροφοριών είναι πλέον πράξη αναγκαία, ώστε να μπορούμε να χρησιμοποιήσουμε γρήγορα και σωστά όλη αυτή την πληροφορία. Λειτουργίες όπως αναζήτηση και ανάκτηση εικόνων από ψηφιακές συλλογές γίνονται καθημερινή δραστηριότητα πολλών ανθρώπων. Στα πλαίσια της παρούσας διπλωματικής παρουσιάζονται τεχνικές αναζήτησης εικόνων με βάση αυτομάτως εξαγόμενα από αυτές οπτικά χαρακτηριστικά και αναπτύσσεται μία διαδικτυακή εφαρμογή για την παρουσίαση των αποτελεσμάτων. Οι τεχνικές αυτές ανήκουν σε ένα ευρύ φάσμα. Άλλοτε χρησιμοποιούνται τοπικά οπτικά χαρακτηριστικά, άλλοτε χαρακτηριστικά συνολικά για ολόκληρη την εικόνα, άλλοτε δημιουργείται οπτικός θησαυρός με συσταδοποίηση και άλλοτε εξετάζεται και η γεωμετρία των αποτελεσμάτων. Αποτελέσματα παρουσιάζονται για όλες τις παραπάνω τεχνικές, σε αρκετές ετερόκλητες συλλογές εικόνων για να ελεγχθεί η γενικότητα της εφαρμογής τους.

Λέξεις Κλειδιά

ανάκτηση εικόνων με βάση το οπτικό περιεχόμενο, οπτικό λεξικό, MPEG-7, οπτικός θησαυρός, οπτικές περιγραφές, σημασιολογικό κενό, ανάκτηση εικόνων με βάση χαρακτηριστικά από περιοχές τους, ανάκτηση εικόνων με βάση σημείων ενδιαφέροντος, περιγραφείς SURF, συσταδοποίηση k-means, RANSAC, Geo-tags.

Abstract

Over the recent years, the amount of digital images available online has increased rapidly. These huge multimedia collections contain diverse data and cover almost every aspect of life in terms of visual and semantic content. Proper indexing and analysis of such data is an essential process, in order to be able to retrieve its useful visual information. Searching through image libraries has become an everyday process, the same way as Google text-based search. In this diploma thesis, techniques for content-based image retrieval are presented and evaluated, and a web-based image search platform is created. Various techniques are applied, using either global or local features, such as the proliferous MPEG-7 and SURF descriptors, extracted locally, from points of interest or segmented regions. A bag-of-words model is used for indexing and geometric constraints are also taken into account. These techniques are evaluated over many common datasets, in order to test the universality of their use, towards a web-scale image retrieval system.

Keywords

content-based image retrieval, bag of words, MPEG-7, visual thesaurus, visual descriptors, semantic gap, region based retrieval, interest point based retrieval, SURF descriptors, k-means, k-d trees, Ukbench dataset, Zurich buildings dataset, Oxford Buildings dataset, RANSAC, Geo-tags,

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε κατά το ακαδημαϊκό έτος 2007- 2008 στο Εργαστήριο Ψηφιακής Επεξεργασίας Εικόνας, Βίντεο και Πολυμέσων (IVML) του Εθνικού Μετσόβιου Πολυτεχνείου. Θα ήθελα να ευχαριστήσω τον επιβλέποντα Καθηγητή κ. Στέφανο Κόλλια για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου την εργασία αυτή και για τη δυνατότητα που μου έδωσε να ασχοληθώ με το συγκεκριμένο ενδιαφέρον θέμα. Θα ήθελα επίσης να ευχαριστήσω τον Ερευνητή Δρ Ιωάννη Αβρίθη καθώς και τους Υποψήφιους Διδάκτορες Ευάγγελο Σπύρου, Καψάλα Πέτρο και Γεώργιο Τόλια για την καταλυτική συμβολη τους στη συγγραφή αυτής της εργασίας με την καθοδήγηση, τις πολύτιμες συμβουλές και υποδείξεις τους.

Περιεχόμενα

Περιεχόμενα Πινάκων

Περιεχόμενα Σχημάτων

Κεφάλαιο 1

Ανάλυση συστημάτων ανάκτησης εικόνων με βάση το οπτικό τους περιεχόμενο

1.1 Εισαγωγή

Αν δώσεις σε ένα παιδί τριών χρονών μερικές φωτογραφίες και του πεις να σου ξεχωρίσει αυτές που έχουν ανθρώπους από αυτές που δεν περιέχουν ανθρώπους είναι σχεδόν σίγουρο ότι θα τα καταφέρει περίφημα. Η ίδια εργασία όμως δεν είναι τόσο εύκολη ακόμη και για έναν εξελιγμένο ηλεκτρονικό υπολογιστή. Η σημασιολογική ανάλυση του οπτικού περιεχομένου μιας εικόνας μπορεί να είναι μια προαιώνια βιολογική διεργασία του ανθρώπινου εγκεφάλου που εκτελείται συνεχώς, για τον ηλεκτρονικό υπολογιστή είναι όμως μια πρόκληση των τελευταίων δεκαετιών.

Τα τελευταία χρόνια έχουμε μια ραγδαία αύξηση του αριθμού των εικόνων που συναντάμε σε ψηφιακή μορφή. Η διάδοση της ψηφιακής φωτογραφικής μηχανής επέτρεψε στον καθένα να φωτογραφίζει γρήγορα και εύκολα τον κόσμο που τον περιβάλλει. Με ελάχιστο κόστος μπορεί πια να ανεβάσει τις συλλογές φωτογραφιών του στο διαδίκτυο και έτσι να τις μοιράζεται με φίλους και αγνώστους. Υπάρχουν σήμερα διαθέσιμες σε ηλεκτρονική μορφή στο διαδίκτυο τεράστιες συλλογές, ποικίλες ως προς το οπτικό και σημασιολογικό περιεχόμενο, που συνεχίζουν να μεγαλώνουν συνεχώς.

Έτσι, η σωστή αρχειοθέτηση όλου αυτού του όγκου πληροφοριών είναι πλέον πράξη αναγκαία, ώστε να μπορούμε να χρησιμοποιήσουμε γρήγορα και σωστά όλη αυτή την πληροφορία. Λειτουργίες όπως αναζήτηση και ανάκτηση εικόνων από ψηφιακές συλλογές γίνονται καθημερινή δραστηριότητα πολλών ανθρώπων.

Η ανάκτηση εικόνων είναι μια ενεργή περιοχή έρευνας από την δεκαετία του '70, οπότε και γεννήθηκε με μίξη ερευνητών από δύο κυρίως ερευνητικές κοινότητες. Την κοινότητα οργάνωσης βάσεων δεδομένων και την κοινότητα της όρασης υπολογιστών. Οι δύο κλάδοι εξετάζουν την ανάκτηση εικόνων από διαφορετικές σκοπιές, η πρώτη με τεχνικές βασισμένες σε κείμενο και η δεύτερη με βάση τα οπτικά χαρακτηριστικά των εικόνων. Οι τεχνικές βασισμένες σε κείμενο αρχίζουν να αναπτύσσονται προς τα τέλη της δεκαετίας του '70. Ένα τότε δημοφιλές πλαίσιο ανάκτησης προϋπέθετε τον σχολιασμό του περιεχομένου κάθε εικόνας από τον άνθρωπο, και έπειτα η ανάκτηση εκτελούνταν από ένα σύστημα βάσης δεδομένων. Για καιρό υπήρξε έρευνα σε αυτόν τον τομέα, δύο όμως

βασικές δυσκολίες κατέστησαν τις μεθόδους αυτές αναποτελεσματικές. Πρώτον, οι συλλογές με εικόνες γινόνταν όλο και μεγαλύτερες και ήταν πλέον αδύνατον να σχολιαστούν όλες οι εικόνες. Το δεύτερο και αρκετά ουσιώδες πρόβλημα που παρατηρήθηκε, ήταν η υποκειμενικότητα της αντίληψης του περιεχομένου των εικόνων από άνθρωπο σε άνθρωπο.

Έτσι, στις αρχές της δεκαετίας του '90, προτάθηκε η αυτόματη εξαγωγή χαρακτηριστικών των εικόνων, με βάση το οπτικό περιεχόμενό τους. Από τότε έχουν αναπτυχθεί πάρα πολλές τεχνολογίες και τεχνικές εξαγωγής χαρακτηριστικών που οδήγησαν σε πολλά, ερευνητικά και εμπορικά, συστήματα ανάκτησης εικόνων με βάση το οπτικό περιεχόμενο.

Στον τομέα έρευνας της *«Ανάκτησης Εικόνων με βάση το οπτικό τους περιεχόμενο» (Content-based Image Retrieval - CbIR)* μπορούμε να εντάξουμε σήμερα οποιαδήποτε τεχνολογία βοηθάει στην οργάνωση εικόνων και στηρίζεται στο οπτικό περιεχόμενο της εικόνας, δηλαδή από ένα κριτήριο σύγκρισης περιοχών μέχρι μια εξελιγμένη εφαρμογή δεικτοδότησης και ένα ολοκληρωμένο σύστημα ανάκτησης.

Οι εφαρμογές της ανάκτησης εικόνων είναι πολλαπλές και πολλές από αυτές αναφέρονται στη συνέχεια της διπλωματικής αυτής εργασίας στα εδάφια για τις διάφορες τεχνικές. Επιγραμματικά μπορούμε να αναφέρουμε μερικά βασικά πεδία εφαρμογών:

- Ανάκτηση παρόμοιων εικόνων από συλλογές.
- Ανάκτηση μετεωρολογικών εικόνων για την πρόβλεψη – παρατήρηση καιρικών φαινομένων.
- Αναγνώριση περιοχών σε δορυφορικές εικόνες.
- Ταξινόμηση και ανάλυση ιατρικών εικόνων ώστε να μπορούν να εξαχθούν συμπεράσματα για την ύπαρξη η μη ανωμαλιών.
- Αναζήτηση έργων τέχνης από αντίστοιχες συλλογές – δημιουργία ψηφιακών μουσειακών συλλογών.
- Σημασιολογική εύρεση εικόνων στο διαδίκτυο.

Η παρούσα διπλωματική εργασία ασχολείται μόνο με την ανάκτηση εικόνων και όχι με την ανάκτηση βίντεο. Η ανάκτηση βίντεο μπορεί από πολλούς να θεωρηθεί ένα πιο ευρύ θέμα έρευνας, καθώς το βίντεο αποτελεί ουσιαστικά ακολουθία εικόνων, όμως πρακτικά η ύπαρξη διαδοχικών καρέ καταλήγει σε πολλές περιπτώσεις να αποτελεί παράγοντα που βοηθάει την ανάλυση. Στο βίντεο, τα διαδοχικά καρέ μπορούν να φανερώσουν πιο εύκολα τα αντικείμενα που περιέχονται, μιας και τα σημεία από τα οποία αποτελούνται κινούνται όλα μαζί.

Στην ανασκόπησή του για την ανάκτηση εικόνων με βάση το οπτικό τους περιεχόμενο ο Smeulders [Smeulders et al., 2000] εισάγει δύο «κενά» από τα οποία προκύπτουν πολλά προβλήματα:

Αισθητικό Κενό (Sensory Gap) είναι το κενό ανάμεσα στο πραγματικό αντικείμενο στον κόσμο και στην πληροφορία που μας δίνει μια (υπολογιστικά εξαγόμενη) περιγραφή του από μια αποτύπωση της σκηνής.

Λόγω της ύπαρξης αυτού του «κενού» η περιγραφή ενός αντικειμένου μέσα από μια εικόνα είναι πρακτικά αδύνατη, καθώς υπάρχει άγνοια της κατάστασης του αντικειμένου. Έχουμε έλλειψη πληροφορίας, γιατί βλέπουμε απλά μια προβολή ενός τρισδιάστατου αντικειμένου. Οι διδιάστατες προβολές

δύο διαφορετικών αντικειμένων μπορεί να είναι ίδιες. Συνεπώς, αν δεν υπάρχει επιπρόσθετη γνώση, μπορούμε μονάχα να υποθέσουμε ότι αντιπροσωπεύουν το ίδιο αντικείμενο.

Προς τις αρχές του 21ου αιώνα οι ερευνητές συνειδητοποίησαν πια ότι τα χαρακτηριστικά χαμηλού επιπέδου της εικόνας (όπως για παράδειγμα το χρώμα, το σχήμα και η υφή) δεν έρχονται σε άμεση αντιστοιχία με έννοιες της καθημερινής μας ζωής. Μπορεί να οριστεί, λοιπόν, και το:

Σημασιολογικό Κενό (Semantic Gap) ως την έλλειψη αντιστοιχίας ανάμεσα στην πληροφορία που μπορεί κάποιος να εξάγει από τα οπτικά δεδομένα και την ερμηνεία που έχουν τα ίδια δεδομένα για κάποιον χρήστη κάποια δεδομένη στιγμή.

Τα πρώτα συστήματα ανάκτησης εικόνων με βάση το οπτικό τους περιεχόμενο δεν ήταν τόσο φιλικά προς τον χρήστη και η εξαγωγή σωστών αποτελεσμάτων απαιτούσε βαθύτερη γνώση των χαρακτηριστικών, πράγμα το οποίο δυσχέραινε την αναζήτηση για κάποιον απλό χρήστη. Χρειάστηκε λοιπόν μια στροφή στον σχεδιασμό των συστημάτων ώστε να βάλουν πλέον τον χρήστη στο κέντρο και να εξελίσσονται με γνώμονα την ευκολία του. Για να συμβεί αυτό θα πρέπει τα συστήματα επόμενης γενιάς να καταλαβαίνουν το σημασιολογικό περιεχόμενο του ερωτήματος που θέτει ο χρήστης και όχι μόνο χαρακτηριστικά χαμηλού επιπέδου. Το πρόβλημα αυτό καλείται *γεφύρωση του σημασιολογικού κενού*.

Ένας από τους τρόπους αντιμετώπισης αυτού μπορεί να είναι η απλή μετάφραση - αντιστοιχία των εύκολα εξαγόμενων χαμηλού επιπέδου χαρακτηριστικών σε έννοιες που έχουν νόημα για τον χρήστη. Για παράδειγμα μπορεί εύκολα κάποιος να δώσει την έννοια «ξύλο» σε μια περιοχή εικόνας με την χαρακτηριστική όψη του ξύλου, και να την περιγράψει με χαρακτηριστικά χαμηλού επιπέδου όπως για παράδειγμα το χρώμα της και την υφή της. Έτσι, κάθε αποτέλεσμα ανάκτησης που προσεγγίζει τα ίδια χαρακτηριστικά θα μπορούσε επίσης να χαρακτηριστεί «ξύλο».

Παραδείγματα συστημάτων προς αυτή την κατεύθυνση υπάρχουν πολλά. Στον τομέα της αναγνώρισης προσώπων για παράδειγμα, υπάρχει η δουλειά του Rowley [Rowley et al., 1998] από το 1996. Ένα από τα πρώτα σύστηματα ανάκτησης εικόνων με βάση το οπτικό περιεχόμενο το οποίο καταπιάστηκε και με το πρόβλημα του σημασιολογικού κενού ήταν το ImageScape [Buijs & Lew, 1999] στο οποίο σύστημα μπορούσε ο χρήστης να θέτει εύκολα ερωτήματα, τα οποία περιέχουν κάποιες οπτικές έννοιες όπως *ουρανός, δέντρα, νερό* κλπ.. Το σύστημα χρησιμοποιούσε αρχές της θεωρίας πληροφορίας για να καθορίσει τα καλύτερα χαρακτηριστικά, ώστε να ελαχιστοποιηθεί η αβεβαιότητα στην ταξινόμηση.

Καθοριστικό βήμα προς το γεφύρωμα του σημασιολογικού κενού στα συστήματα ανάκτησης εικόνων με βάση το οπτικό τους περιεχόμενο, έγινε με την υιοθέτηση της *ανάδρασης σχετικότητας*. Έτσι, με διάφορες μεθόδους, ο χρήστης κατευθύνει το σύστημα προσαρμόζοντας τα ερωτήματα του όλο και προς τον στόχο της αναζήτησής του. Η ανάδραση σχετικότητας μελετάται εκτενέστερα στην ενότητα 1.9.

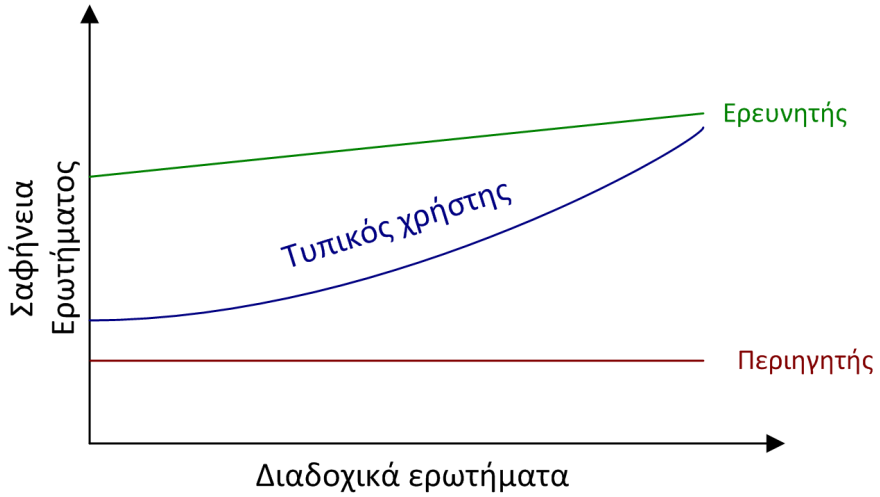
1.2 Επιδιώξεις και προθέσεις του χρήστη

Ο Datta [Datta et al., 2008] χωρίζει τους χρήστες, ανάλογα με τις προθέσεις τους και την σαφήνεια του σκοπού τους στις εξής κατηγορίες:

περιηγητής (browser): Ο χρήστης που δεν έχει καθαρό στόχο και απλά περιηγείται σε συλλογές εικόνων. Μια συνεδρία του χρήστη θα περιλαμβάνει διαδοχικά, χωρίς συνοχή μεταξύ τους ερωτήματα. Μπορεί να ψάχνει αρχικά για μηχανές και έπειτα για ηλιοβασιλέματα στη θάλασσα.

τυπικός χρήστης (surfer): Ο χρήστης ο οποίος έχει ενδιάμεση σαφήνεια ως προς τον σκοπό της αναζήτησης. Τα αρχικά ερωτήματα μπορεί να είναι εξερευνητικά, τα επόμενα όμως δείχνουν με μεγαλύτερη σαφήνεια τον στόχο της αναζήτησης.

ερευνητής (searcher): Ο χρήστης ο οποίος είναι ξεκάθαρος ως προς το τι ψάχνει στο σύστημα. Συνήθως μια συνεδρία του χρήστη θα είναι μικρή σε διάρκεια, με υψηλή συνοχή και οδηγεί τις περισσότερες φορές σε κάποια τελικά αποτελέσματα.



Σχήμα 1.1: Κατηγορία χρήστη και σαφήνεια

Η σαφήνεια του σκοπού παίζει, σύμφωνα με τον Datta, σημαντικό ρόλο στο τι περιμένει ένας χρήσης από ένα σύστημα ανάκτησης και μπορεί να δράσει σαν κατευθυντήρια γραμμή για τον σχεδιασμό του συστήματος. Ένας περιηγητής δίνει σημασία στην ευκολία της χρήσης και ίσως στην εμφάνιση του συστήματος. Συνήθως έχει και άνεση στον χρόνο, με αποτέλεσμα εικόνες «εκπλήξεις» από τυχαίες ή μη αναζητήσεις να είναι ευπρόσδεκτες. Ο τυπικός χρήστης από την άλλη, θα προτιμούσε ένα εύχρηστο περιβάλλον το οποίο τον διευκολύνει στο να διατυπώσει με σαφήνεια το ερώτημά του. Το όμορφο και εύχρηστο γραφικό περιβάλλον δεν είναι απαραίτητο για τον ερευνητή, ο οποίος προτιμά πιο ολοκληρωμένα και εύστοχα αποτελέσματα ίσως και τη δυνατότητα να έχει πλήρη έλεγχο της διαδικασίας ανάκτησης.

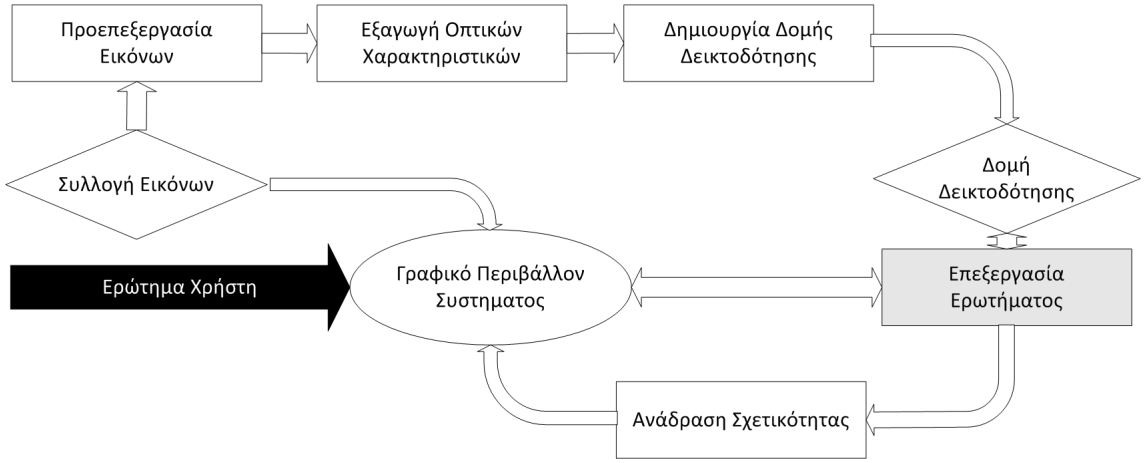
Το σχήμα 1.1 παρουσιάζει τη σαφήνεια του ερωτήματος που θέτει ο χρήστης σε σχέση με το σε ποιά κατηγορία ανήκει.

Είναι λίγες οι μελέτες που να ασχολούνται με τα συστήματα ανάκτησης από την μεριά του χρήστη βασισμένα σε πραγματικά δεδομένα. Σε μια από αυτές [Conescu & Christel, 2005] οι χρήστες χωρίζονται σε αρχάριους και προχωρημένους και μελετώνται τα μοτίβα αλληλεπίδρασης μέσω ενός

συστήματος ανάκτησης βίντεο. Στο [Armitage & Enser, 1997] γίνεται μια ανάλυση των αναγκών του χρήστη όσον αφορά στην ανάκτηση οπτικής πληροφορίας.

1.3 Δομή ενός συστήματος ανάκτησης

Στο Σχήμα 1.2 φαίνεται μια γενική διαγραμματική προσέγγιση ενός συστήματος ανάκτησης εικόνων με βάση το οπτικό τους περιεχόμενο και οι κύριες οντότητες που το αποτελούν.



Σχήμα 1.2: Η δομή ενός συστήματος ανάκτησης εικόνων με βάση το οπτικό τους περιεχόμενο

Οι εικόνες της συλλογής μπορεί αρχικά να υποστούν κάποιου είδους προεπεξεργασία, ώστε να μπορέσουμε στη συνέχεια να εξάγάγουμε σωστότερα οπτικά χαρακτηριστικά. Σε αυτό το στάδιο μπορεί να περιλαμβάνεται κάποιο φιλτράρισμα ή εξομάλυνση με σκοπό την αποθορυβοποίηση ή απλοποίηση των εικόνων. Σε αυτό το στάδιο μπορεί επίσης να γίνει κατάτμηση των εικόνων με σκοπό την εξαγωγή χαρακτηριστικών κατά περιοχές.

Επόμενο στάδιο είναι η εξαγωγή των οπτικών χαρακτηριστικών είτε από ολόκληρη την εικόνα είτε από περιοχές της είτε από κάποια σημεία ενδιαφέροντος. Τεχνικές που αφορούν στην εξαγωγή χαρακτηριστικών και την έξυπνη αναπαράσταση των εικόνων θα αναφερθούν εκτενώς στήν ενότητα 1.4.

Από τα δεδομένα που εξήχθησαν με την προηγούμενη διαδικασία, παράγεται για κάθε εικόνα μια «υπογραφή», δηλαδή περιγραφή του περιεχομένου της με βάση τα παραπάνω χαρακτηριστικά. Οι περιγραφές αυτές όλων των εικόνων πρέπει να αποθηκευτούν αποδοτικά, σε μια βάση δεδομένων για παράδειγμα, ώστε να μπορεί έπειτα το σύστημα να χρησιμοποιήσει όλη αυτή την πληροφορία. Επιπλέον, μπορεί να γίνει κάποια ομαδοποίηση - συσταδοποίηση των περιγραφών από την οποία να προκύψει ένα «λεξικό» διάφορων οπτικών χαρακτηριστικών με βάση τα οποία θα χαρακτηριστεί κάθε εικόνα. Τεχνικές δεικτοδότησης αναφέρονται στην ενότητα 1.6 και τεχνικές με την χρήση οπτικού θυσαυρού στην ενότητα 1.7.

Ο πυρήνας του συστήματος ανάκτησης ενεργοποιήται με την είσοδο ενός ερωτήματος από τον χρήστη. Εκεί γίνεται το ταίριασμα των χαρακτηριστικών του ερωτήματος με τα χαρακτηριστικά των

εικόνων της συλλογής και ως έξοδο έχουμε κάποιες εικόνες ταξινομημένες με αύξουσα σειρά ως προς κάποιο κριτήριο απόστασης από την εικόνα - ερώτημα. Τεχνικές ταιριάσματος εικόνων αναφέρονται στην ενότητα 1.5. Οι διάφοροι τύποι ερωτήματος αναφέρονται στην ενότητα 1.8.

Σε πολλά σύγχρονα συστήματα δίνεται η δυνατότητα στον χρήστη να κρίνει τα αποτελέσματα του ερωτήματος θετικά ή αρνητικά και με αυτόν τον τρόπο έχει το σύστημα ένα είδος ανάδρασης. Η ανάδραση σχετικότητας μελετάται στην ενότητα 1.9.

1.4 Εξαγωγή οπτικών χαρακτηριστικών και αναπαράσταση εικόνων

Η εξαγωγή οπτικών χαρακτηριστικών είναι το θεμέλιο της ανάκτησης εικόνων με βάση το οπτικό τους περιεχόμενο. Με αυτήν τη διαδικασία κωδικοποιούνται στοιχεία του περιεχομένου των εικόνων με βάση τα οποία θα αναπαρίσταται η εικόνα στο σύστημα. Θα εξαχθεί κατά κάποιο τρόπο μια *οπτική «υπογραφή»* (visual signature) της εικόνας.

Τα χαρακτηριστικά μπορούν να εξαχθούν είτε από ολόκληρη την εικόνα (global features) είτε από ομάδες pixel – περιοχές της εικόνας (region features). Τα πλέον χρησιμοποιούμενα χαρακτηριστικά είναι αυτά που κωδικοποιούν *το χρώμα, την υφή, το σχήμα* ή την περιοχή γύρω από κάποια *σημεία ενδιαφέροντος* της εικόνας. Οι κατηγορίες αυτές θα περιγραφούν περιληπτικά παρακάτω.

Στην περίπτωση που τα χαρακτηριστικά εξάγονται από ολόκληρη την εικόνα, επιδιώκεται να κωδικοποιηθούν συνολικά χαρακτηριστικά της. Με αυτόν τον τρόπο η εικόνα αναπαρίσταται ολόκληρη με ένα πολυδιάστατο διάνυσμα, που συνήθως έχει διάσταση της τάξης του 10^2. Τα χαρακτηριστικά που εξάγονται από ολόκληρη την εικόνα χρησιμοποιούνται όλο και λιγότερο τα τελευταία χρόνια καθώς δεν είναι αποτελεσματικά σε εικόνες με πυκνό οπτικό περιεχόμενο. Με γενικά χαρακτηριστικά μπορεί να κωδικοποιηθεί αποτελεσματικά μια εικόνα ηλιοβασιλέματος, αλλά όχι μια εικόνα από τους δρόμους του Πεκίνου σε ώρα αιχμής (Σχήμα 1.3). Η αριστερή εικόνα αναπαρίσταται αποδοτικά με περιγραφείς εξαγόμενους από ολόκληρη την εικόνα (για παράδειγμα με κάποιον περιγραφέα χρώματος) καθώς ολόκληρο το σημασιολογικό της περιεχόμενο μπορεί να περιγραφεί αποδοτικά μονάχα με την έννοια «ηλιοβασίλεμα». Για την εικόνα στα δεξιά θα χρειαστούν πιο εξελιγμένες τεχνικές εξαγωγής για να κωδικοποιηθεί αξιοποιήσιμα το περιεχόμενο της.


Σχήμα 1.3: Εικόνες με ποικιλία στην πυκνότητα του οπτικού τους περιεχομένου

Στην περίπτωση εξαγωγής χαρακτηριστικών από περιοχές της εικόνας απαιτείται κάποια προεργασία εξαγωγής των ορίων των περιοχών. Συνήθεις μέθοδοι είναι η κατάτμησή της όπως και η εξαγωγή περιοχών γύρω από κάποια σημεία υψηλού ενδιαφέροντος της εικόνας.

1.4.1 Χρώμα

Το χρώμα είναι ένα από τα πιο διαδεδομένα οπτικά χαρακτηριστικά εικόνων, με την έρευνα στον τομέα αυτό να ανθεί στα πρώιμα στάδια ανάπτυξης των συστημάτων ανάκτησης εικόνων [Wang et al., 1997]. Μεγάλη έμφαση δόθηκε τότε στην εξερεύνηση διαφόρων *χώρων χρώματος* (color spaces) που φαίνεται να προσεγγίζουν καλύτερα τον τρόπο με τον οποίο το ανθρώπινο μάτι αντιλαμβάνεται το χρώμα σε σχέση με τον κλασσικό RGB χώρο (ένα παράδειγμα είναι ο χώρος LUV). Μελετήθηκε επίσης το πώς μεταβάλλεται το χρώμα με αλλαγές στον φωτισμό, την θέση της κάμερας και την κλίση των επιφανειών και έχουν γίνει προσπάθειες να αναπτυχθούν μοντέλα με ανοχή σε αυτές τις μεταβολές [Sebe & Lew, 2001]. Το μοντέλο της «αντίθετης αναπαράστασης χρωμάτων», για παράδειγμα [Swain & Ballard, 1991], κατάφερε να απομονώσει τη μεταβολή της φωτεινότητας (brightness) σε έναν άξονα, με αποτέλεσμα να υπάρχει η δυνατότητα να απαλειφθεί αυτή η συνιστώσα.

Το ιστόγραμμα χρωμάτων είναι μια πάγια τεχνική που χρησιμοποιήθηκε σε πολλά συστήματα ανάκτησης της προηγούμενης δεκαετίας, όπως για παράδειγμα το σύστημα QBIC [Flickner et al., 1995] της IBM. Στατιστικά το ιστόγραμμα χρωμάτων υποδηλώνει την από κοινού πιθανότητα των εντάσεων των τριών χρωματικών καναλιών. Μια εναλλακτική του ιστογράμματος αυτού είναι η χρήση των *«στιγμών χρώματος»* (Color moments) [Stricker & Orengo, 1995a].

Με το πέρασμα των χρόνων η έρευνα εστιάστηκε στην εξαγωγή μιας όσο το δυνατόν καλύτερης «περίληψης» των χρωμάτων της εικόνας, λαμβάνοντας υπόψη και την χωρική τους κατανομή. Το πρότυπο MPEG-7 ([Sikora, 2001], [Chang et al.,]) περιλαμβάνει σύγχρονους περιγραφείς χρώματος [Manjunath et al., 2001], περιγραφείς που θα αναλυθούν σε επόμενο κεφάλαιο καθώς χρησιμοποιήθηκαν και στο σύστημα ανάκτησης που υλοποιήθηκε.

1.4.2 Υφή

Με τον όρο *υφή* αναφερόμαστε σε οπτικά μοτίβα που έχουν ομογενείς ιδιότητες, οι οποίες δεν είναι αποτέλεσμα της παρουσίας απλά ενός χρώματος ή έντασης. Με τα χαρακτηριστικά υφής γίνεται προσπάθεια να κωδικοποιηθεί αυτή η πληροφορία της υφής που υπάρχει σε διάφορα υλικά ή αντικείμενα της εικόνας. Για παράδειγμα, το σώμα της ζέβρας ή της τίγρης, οι ξύλινες επιφάνειες, οι τούβλινοι τοίχοι, τα κεντήματα αλλά ακόμα και ο ουρανός, τα σύννεφα και οι φυλλωσιές, έχουν χαρακτηριστική όψη και επαναλαμβανόμενα μοτίβα, άλλοτε τραχιά και άλλοτε πιο απαλά. Ενώ το χρώμα είναι χαρακτηριστικό ενός μονάχα pixel, η υφή είναι ένα χαρακτηριστικό που αναφέρεται σε μια γειτονιά – περιοχή pixel. Δεν έχει νόημα να μιλάμε για υφή σε ένα σημείο μιας ξύλινης επιφάνειας αλλά σε μια περιοχή μέσα στην επιφάνεια αυτή.

Σε αρκετές περιπτώσεις, όπως δορυφορικές [Li & Castelli, 1997] ή ιατρικές εικόνες, τα χαρακτηριστικά υφής είναι ιδιαίτερα χρήσιμα καθώς οι περιοχές ομοιόμορφης υφής σε αυτές τις περιπτώσεις, έχουν άμεση σχέση με σημασιολογικές έννοιες υψηλού επιπέδου. Αποτελεσματικά αποδείχτηκαν τα χαρακτηριστικά υφής και σε εικόνες από κείμενα [Cullen et al., 1997].

Συστηματική έρευνα γύρω από χαρακτηριστικά υφής άρχισε με τον Haralick από την δεκαετία

του '70, που πρότεινε την αναπαράστασή της μέσω του πίνακα συνύπαρξης (co-occurrence matrix representation) ([Haralick et al., 1973], [Aksoy & Haralick, 1998]). Στα τέλη της ίδιας δεκαετίας ο Tamura μελέτησε και αυτός την υφή αλλά από άλλη σκοπιά, εμπνευσμένος από ψυχολογικές μελέτες για την αντίληψη της υφής από το ανθρώπινο μάτι [H. Tamura & Yamawaki, 1978]. Στο σύστημα QBIC [Flickner et al., 1995] της IBM χρησιμοποιείται ένα εξελιγμένο μοντέλο περιγραφής υφής βασισμένο στην δουλειά του Tamura.

Στην δεκαετία του '90, χρισιμοποιήθηκε ιδιαίτερα ο *μετασχηματισμός κυματιδίων* (wavelet transform) ως μέθοδος περιγραφής της υφής, με τους Smith και Chang [Smith & Chang, 1994] να έχουν καλά αποτελέσματα. Οι Wolf και Bishop υποστηρίζουν ότι τα φίλτρα Gabor είναι τα πλέον κατάλληλα για την εξαγωγή χαρακτηριστικών υφής για ανάκτηση εικόνων, καθώς δεν εξαρτώνται από την κλίμακα, όπως η συχνότητα, η κατεύθυνση και η φάση – παράμετροι που χρησιμοποιούνταν παλιότερα για εξαγωγή υφής [Wolf et al., 2000]. Στο ίδιο συμπέρασμα φτάνουν και οι Ma και Manjunath [Ma & Manjunath, 1995], συγκρίνοντας διάφορες μεθόδους εξαγωγής χαρακτηριστικών από κυματίδια.

Πιο σύγχρονοι περιγραφείς υφής, που προσφέρουν ανεξαρτησία στην περιστροφή και οι οποίοι χρησιμοποιήθηκαν και στην υλοποίηση του συστήματος, περιέχονται στο MPEG-7 [Manjunath et al., 2001], και θα αναλυθούν σε επόμενο κεφάλαιο.

1.4.3 Σχήμα

Η εξαγωγή του σχήματος ενός αντικειμένου είναι ταυτόσημη με την κατάτμηση μιας εικόνας στα αντικείμενα που περιλαμβάνει. Παρόλα αυτά, οι περισσότεροι αλγόριθμοι κατάτμησης δεν είναι αποδοτικοί στην ανάκτηση εικόνων, καθώς η κατάτμηση μεγάλων συλλογών εικόνων είναι ιδιαίτερα χρονοβόρα διαδικασία.

Στα συστήματα ανάκτησης εικόνων απαιτείται μια εύρωστη περιγραφή αναπαράστασης που θα έχει όσο το δυνατόν ανοχή και σε παραμορφώσεις του σήματος. Καθώς κάποια λάθη στην ανάκτηση είναι αποδεκτά, τα μέτρα ακρίβειας μπορούν να χαλαρωθούν με αποτέλεσμα να έχουμε μεγαλύτερη ευρωστία και μικρότερη υπολογιστική πολυπλοκότητα. Μέθοδοι με ανοχή στις παραμορφώσεις μπορούν να χρησιμοποιηθούν αποδοτικά και σε ερωτήματα με σκίτσο από τον χρήστη [Bimbo & Pala, 1997].

Γενικά, οι αναπαραστάσεις σχήματος χωρίζονται σε εκείνες που περιγράφουν μόνο το περίγραμμα και σε εκείνες που περιγράφουν όλη την περιοχή που καταλαμβάνει το σχήμα [Rui et al., 1999]. Σε αυτές τις δύο κατηγορίες χωρίζει και τους περιγραφείς του MPEG-7 ο Bober [Bober, 2001] και χρησιμοποιεί στην περίπτωση της περιγραφής περιγράμματος την αναπαράσταση στον *χώρο κλίμακας της καμπυλότητας* (*curvature Scale Space – CSS*) [Abbasi et al., 1999]. Μια ανάλυση των τεχνολογιών ταιριάσματος σχημάτων από την σκοπιά της υπολογιστικής γεωμετρίας βρίσκεται στο [Veltkamp & Hagedoorn, 2001].

1.4.4 Σημεία Ενδιαφέροντος

Πολλές φορές σε μια εικόνα δεν μας ενδιαφέρει το γενικό της περιβάλλον, αλλά ο εντοπισμός συγκεκριμένων αντικειμένων μέσα σε αυτή. Πρέπει λοιπόν να εστιάσουμε την προσοχή μας σε σημεία της εικόνας που μπορεί να περιγράφουν μέρη από κάποια αντικείμενα. Ως σημεία ενδιαφέροντος θεωρούμε τα σημεία, στα οποία η γύρω περιοχή έχει πλούσιο οπτικό περιεχόμενο. Αυτά μπορεί

να είναι σημεία στην περιφέρεια, στις γωνίες ή και στο εσωτερικό ενός αντικειμένου. Συνήθως σε ομογενείς επιφάνειες δεν έχουμε σημεία ενδιαφέροντος. Η σπουδαιότητα αυτών των σημείων έγκειται στο ότι περιγράφουν αποτελεσματικά σημαντικές περιοχές τις εικόνας με λίγες σχετικά τιμές, με αποτέλεσμα να είναι ιδιαίτερα διαδεδομένα στην ανάκτηση αντικειμένων. Για να μην υπάρχει εξάρτηση από την κλίμακα (μέγεθος) η διαδικασία ανίχνευσης σημείων ενδιαφέροντος εκτελείται σε πολλαπλές κλίμακες (multi-scale analysis). Δημιουργείται για αυτό τον σκοπό ένας χώρος κλίμακας με διαοχικές συνελίξεις της εικόνας με γκαουσιανούς πυρήνες αυξανόμενης τυπικής απόκλισης σ και σε κάθε επίπεδο εκτελείται η ανίχνευση των σημείων ενδιαφέροντος. Ανάλυση χώρων κλίμακας και εφαρμογές τους στην όραση υπολογιστών δίνει ο Lindeberg [Lindeberg, 1994].

Συνήθως μαζί με τους τοπικούς περιγραφείς λαμβάνεται υπόψη και η χωρική κατανομή των σημείων. Σε μερικές περιπτώσεις τα σημεία ομαδοποιούνται με σκοπό να αναπαραστήσουν τμήματα ενός αντικειμένου, και έπειτα λαμβάνεται υπόψη η χωρική κατανομή των τμημάτων.

Αφού έχουν βρεθεί τα σημεία ενδιαφέροντος, επόμενο βήμα είναι να βρεθεί μια εύρωστη και αποτελεσματική αναπαράσταση της γύρω περιοχής. Ο πιο απλός περιγραφέας θα ήταν ένα διάνυσμα με τις εντάσεις της φωτεινότητας των τριγύρω σημείων, με την ετεροσυσχέτηση σαν μέτρο ομοιότητας δύο τέτοιων περιοχών. Μια τέτοια αναπαράσταση όμως δεν είναι ιδιαίτερα αποτελεσματική καθώς απαιτεί αρκετό χώρο αποθήκευσης, και δεν έχει ανοχή σε φαινόμενα όπως περιστροφή ή αλλαγή κλίμακας.

Από το 1997 η Cordelia Schmid πρότεινε το πλαίσιο ενός συστήματος ανάκτησης εικόνων με βάση τοπικούς περιγραφείς γύρω από σημεία ενδιαφέροντος [Schmid & Mohr, 1997]. Για την περιγραφή των περιοχών χρησιμοποιεί το *local jet* – δηλαδή συνέλιξη με διάφορες γκαουσιανές παραγώγους και εκτελεί τη διαδικασία εξαγωγής με γκαουσιανές διαφορετικής διασποράς για να επιτύχει ανεξαρτησία στην κλίμακα.

Σημαντικό βήμα στην περιγραφή τοπικών χαρακτηριστικών έγινε το 2004 από τον Lowe, ο οποίος πρότεινε τους περιγραφείς *SIFT* (*Scale Invariant Feature Transform*) [Lowe, 2004]. Ο Lowe χρησιμοποιεί ως προσέγγιση της λαπλασιανής της γκαουσιανής (Laplacian-of-Gausian – LoG), για να επιταχύνει τη διαδικασία, τη διαφορά των γκαουσιανών (Difference-of-Gaussian – DoG) και εκτελεί και αυτός τη διαδικασία σε αυξανόμενες κλίμακες για να επιτύχει ανεξαρτησία από την κλίμακα. Σαν περιγραφέα της περιοχής γύρω από τα σημεία, προτείνει τη χρήση ιστογραμμάτων των παραγώγων, κρατώντας παράλληλα τη θέση και την κατεύθυνσή τους. Με έναν κβαντισμό των κατευθύνσεων και θέσεων των παραγώγων που εκτελείται, επιτυγχάνεται μια ανοχή σε μικρές περιστροφές και γεωμετρικές παραμορφώσεις. Εκτεταμένη αξιολόγηση και σύγκριση τοπικών περιγραφέων γίνεται από τους Mikolajczyk και Schmid [Mikolajczyk & Schmid, 2005].

Επειδή η εξαγωγή περιγραφέων όπως τα *SIFT* είναι ιδιαίτερα χρονοβόρα διαδικασία, η ενσωμάτωση τέτοιου είδους τοπικών περιγραφέων σε συστήματα ανάκτησης εικόνων θα τα επιβάρυνε αρκετά. Έτσι, έχουν γίνει προσπάθειες τα τελευταία χρόνια να αναπτυχθούν περιγραφείς που μοιράζονται την ίδια φιλοσοφία με τα SIFT, είναι όμως πιο γρήγορα υπολογίσιμοι. Σε αυτή την κατηγορία ανήκουν οι περιγραφείς *SURF* (*Speeded-Up Robust Features*) [Bay et al., 2008] οι οποίοι θα αναλυθούν σε επόμενο κεφάλαιο, μιας και είναι οι περιγραφείς που χρησιμοποιήθηκαν σε μια από τις τεχνικές ανάκτησης που περιγράφονται στην παρούσα δηπλωματική.

1.5 Τεχνικές ταιριάσματος εικόνων

Όσο σημαντικό είναι να έχουμε εύρωστους και διακριτέους περιγραφείς των οπτικών χαρακτηριστικών της εικόνας, άλλο τόσο σημαντικό είναι να έχουμε και έναν αποτελεσματικό τρόπο για να μετράμε την ομοιότητα μεταξύ δύο εικόνων.

Εκτενής ανάλυση πάνω στην ομοιότητα χαρακτηριστικών κάνει ο Jolion όπου και παρουσιάζει τις πιο διαδεδομένες μεθόδους για τον υπολογισμό της ομοιότητας [Jolion, 2001]. Η ομοιότητα μεταξύ δύο *διανυσμάτων χαρακτηριστικών* (*Feature Vectors*), που μπορεί να περιέχουν χαρακτηριστικά από μια εικόνα, μια περιοχή της εικόνας ή μια περιοχή γύρω από κάποιο σημείο ενδιαφέροντος, μπορεί να εκφραστεί ως:

$$S_{q,db} = d(F_q, F_{db}) \tag{1.1}$$

όπου F_q και F_{db} είναι δύο διανύσματα χαρακτηριστικών, το πρώτο από την εικόνα ερωτήματος του χρήστη και το δεύτερο από μία από τις εικόνες της βάσης δεδομενων και $d(i, j)$ κάποια συνάρτηση απόστασης.

Σε περιπτώσεις που ο χώρος των χαρακτηριστικών είναι διανυσματικός, αποτελεσματικό κριτήριο ομοιότητας είναι η *Ευκλείδεια απόσταση*, ή *L2* όπως αλλιώς είναι γνωστή. Έρευνες έδειξαν ότι δεν υπάρχει ένα γενικά αποτελεσματικότερο κριτήριο - μέτρο απόστασης, αλλά η επιλογή πρέπει να γίνει σε σχέση με την συλλογή εικόνων που διαθέτουμε και την φύση των διανυσμάτων των περιγραφέων που θέλουμε να συγκρίνουμε [Sebe et al., 2000].

Για να αποφανθούμε για την απόσταση μεταξύ ιστογραμμάτων έχουν προταθεί διάφορες μέθοδοι που συγκρίνουν τις σχετικές κορυφές των ιστογραμμάτων όπως η απόσταση *Minkowski*. Προτάθηκε επίσης η χρήση αθροιστικών ιστογραμμάτων καθώς υποστηρίζεται ότι είναι περισσότερο εύρωστα στον θόρυβο [Stricker & Orengo, 1995b].

Ο πλέον συνήθης τρόπος χειρισμού των ιστογραμμάτων είναι η μετατροπή τους σε διάνυσμα. Έτσι, ένα ιστόγραμμα με k κορυφές, αναπαρίσταται με ένα k-διάστατο διάνυσμα: $[f_1, f_2, \ldots, f_k]$, με f_i την τιμή της i-οστής κορυφής. Ένα μειονέκτημα αυτής της μεθόδου, όμως, είναι ότι με αυτόν τον τρόπο αγνοείται η πληροφορία της θέσης από όπου εξήχθησαν οι κορυφές του ιστογράμματος [Rubner et al., 1998]. Προτάθηκε, λοιπόν, το 1998 από τους Rubner και Tomasi η χρήση της *Απόστασης των Περιπατητών στην Γη* (*Earth Movers Distance - EMD*), για την χρήση της οποίας αναπαρίστανται τα ιστογράμματα σαν διανύσματα που περιέχουν ζεύγη τιμών και χωρικής πληροφορίας: $[(z_1, f_1), (z_2, f_2), \ldots, (z_k, f_k)]$, όπου f_i είναι η τιμή της i-οστής κορυφής και z_i η θέση – το κέντρο της κορυφής [Rubner et al., 2000].

Άλλη χρησιμοποιούμενη απόσταση είναι η απόσταση *Mahalanobis*, η οποία χρησιμοποιήθηκε κυρίως σε συστήματα ανάκτησης κατά περιοχές όπως το *Blobworld* για να συγκρίνει πιο σύνθετες δομές (*Blobs*) που περιέχουν χαρακτηριστικά χρώματος, υφής και χωρικής κατανομής μαζί [Carson et al., 1997] [Carson et al., 2002].

1.6 Τεχνικές δεικτοδότησης

Στα πρώιμα συστήματα ανάκτησης οι εξαγόμενες από τις εικόνες πληροφορίες σώζονταν απλά σε αρχεία ή αποτελούσαν εγγραφές σε κάποια βάση δεδομένων. Η απόδοση και των δύο μεθόδων, από την σκοπιά της υπολογιστικής αποτελεσματικότητας, έπεφτε με τον καιρό καθώς το μέγεθος

των συλλογών αύξανε γεωμετρικά. Επίσης, τα διανύσματα χαρακτηριστικών που εξάγονται από τις εικόνες της συλλογής έχουν συνήθως μεγάλη διάσταση. Έτσι, κυρίως σε μεγάλες συλλογές, είναι πλέον απαραίτητη η χρήση κάποιας έξυπνης μορφής δεικτοδότησης ή και κάποιας μορφής συσταδοποίησης – ομαδοποίησης της πληροφορίας.

Πολλές φορές προτού εφαρμοστεί κάποια μέθοδος δεικτοδότησης, μπορεί να δοκιμαστεί κάποια τεχνική για να ελαττωθεί η διάσταση των διανυσμάτων, όπως για παράδειγμα κάποια μορφή της *Ανάλυσης Πρωτευόντων Συνιστωσών* (*Principal Component Analysis - PCA*) [Ng & Sedighian, 1996]. Έπειτα, τα ευρετήρια που χρησιμοποιούνται πλέον για τον υπολογισμό ομοιότητας περιέχουν δομές δέντρων για να επιτύχουν λογαριθμική πολυπλοκότητα. Δομές όπως τα *k-d δέντρα, τα k-d δέντρα προτεραιότητας, τα R-δέντρα* χρησιμοποιήθηκαν και χρησιμοποιούνται σε διάφορες φάσης της διαδικασίας ανάκτησης [Egas et al., 1999] [Scott & Shyu, 2003]. Στο σύστημα που υλοποιήσαμε, χρησιμοποιήθηκαν k-d δέντρα για τον ταχύτερο προσδιορισμό του *κοντινότερου γείτονα* (Nearest Neighbour) μεταξύ δύο πολυδιάστατων διανυσμάτων χαρακτηριστικών.

Άλλες μέθοδοι που προτάθηκαν περιλαμβάνουν τον κβαντισμό των διανυσμάτων (vector quantization) [Ye & Xu, 2003] ή την χρήση δύο ειδών «υπογραφών» για κάθε εικόνα, με το ένα είδος να περιγράφει τα αντικείμενα που περιέχει η εικόνα και το άλλο είδος να περιγράφει την χωρική κατανομή στην εικόνα [El-Kwae & Kabuka, 2000].

Τέλος πολλές είναι οι τεχνικές που χρησιμοποιούν κάποιας μορφής *hashing*, δηλαδή προβολή - απεικόνηση των διανυσμάτων των εικόνων σε άλλους χώρους μικρότερης διάστασης [Greene et al., 1994], [Chum e

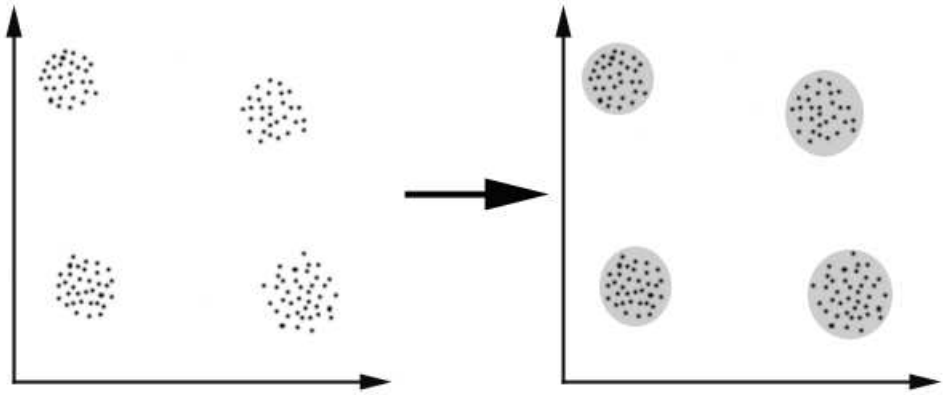
1.7 Τεχνικές με χρήση οπτικού θησαυρού

Σε περιπτώσεις όπου εξάγουμε χαρακτηριστικά από περιοχές εικόνων ή σημεία ενδιαφέροντος και όχι από ολόκληρη την εικόνα, ο όγκος της εξαγόμενης πληροφορίας αυξάνεται πολύ. Για παράδειγμα από 1000 μόνο εικόνες μπορεί να πάρουμε 60000 σημεία ενδιαφέροντος συνολικά, με το κάθε ένα να αναπαρίσταται από ένα διάνυσμα χαρακτηριστικών. Έτσι όταν το μέγεθος της συλλογής αυξάνει στην τάξη του 10^5, διαδικασίες που εκτελούνται κατ' επανάληψη στην ανάκτηση όπως η αναζήτηση και εξαγωγή απόστασης, ακόμα και με έναν αποδοτικό τρόπο δεικτοδότησης είναι ιδιαίτερα χρονοβόρες. Ως λύση αυτού του προβλήματος έχει προταθεί η συσταδοποίηση (*clustering*) των διανυσμάτων σε ομάδες (*clusters*). Κάθε ομάδα αντιπροσωπεύεται συνήθως από το κέντρο της κάθε συστάδας.

Μέθοδοι συσταδοποίησης υπάρχουν πολλές με πλέον διαδεδομένη την *k-means*, η λειτουργία της οποίας, μιας και χρησιμοποιήθηκε και στο σύστημά μας, θα αναλυθεί σε επόμενο κεφάλαιο. Χαρακτηριστικό της μεθόδου αυτής είναι ότι ο αριθμός των συστάδων πρέπει να καθοριστεί εκ των προτέρων. Άλλες μέθοδοι που έχουν χρησιμοποιηθεί είναι οι *kernel mapping, hierarchical clustering* και *metric learning*. Οι μέθοδοι συσταδοποίησης συνεχίζουν να εξελίσσονται μέχρι σήμερα και να βρίσκουν εφαρμογή σε συστήματα ανάκτησης εικόνων και αντικειμένων [Philbin et al., 2007].

Στην περίπτωση ανάκτησης αντικειμένων, απαιτείται και κατηγοριοποίηση περιοχών των εικόνων ανάλογα με τα αντικείμενα που περιέχουν. Αυτό προϋποθέτει να έχει προηγηθεί πριν τη χρήση του συστήματος μια φάση εκπαίδευσης. Πολλές μέθοδοι συσταδοποίησης και εκπαίδευσης αναλύονται στο βιβλίο *Elements of statistical learning* [Hastie et al., 2001].

Στην περίπτωση που εξάγονται οπτικά χαρακτηριστικά χρώματος και υφής από περιοχές της εικόνας (που συνήθως προέρχονται από κατάτμηση) και εκτελεστεί κάποιος αλγόριθμος συσταδοποίησης στα διανύσματα χαρακτηριστικών, θα πάρουμε κάποιους «τύπους περιοχών», κάθε έναν από τους



Σχήμα 1.4: Συσταδοποίηση διδιάστατων σημείων.

οποίους θα χαρακτηρίζει το κέντρο της συστάδας. Μπορεί, έτσι, να αναπαραστηθεί η κάθε εικόνα με ένα διάνυσμα που θα περιέχει τις ελάχιστες αποστάσεις των περιοχών της από τον κάθε τύπο περιοχής [Spyrou et al., 2008]. Η διαδικασία κατασκευής του οπτικού θησαυρού και της αναπαράστασης των εικόνων με αυτή την μέθοδο θα αναλυθεί σε παρακάτω κεφάλαιο, μιας και αποτελεί μέρος του συστήματος ανάκτησης εικόνων βασισμένο σε περιοχές που υλοποιήθηκε στα πλαίσια αυτής της διπλωματικής εργασίας.

Στην περίπτωση που τα διανύσματα χαρακτηριστικών προέρχονται από περιοχές γύρω από σημεία ενδιαφέροντος της εικόνας, με την εφαρμογή κάποιου είδους συσταδοποίησης μπορεί να παραχθεί ένα *οπτικό «λεξικο»* μικρών περιοχών (*visual vocabulary*). Κάθε στοιχείο του λεξικού, δηλαδή κάθε μια από τις παραγόμενες συστάδες, θα εκφράζει, σε περίπτωση που χρησιμοποιηθούν περιγραφείς τύπου *SIFT*, μια περιοχή ως προς τις κατευθύνσεις των παραγώγων των φωτεινοτήτων μέσα σε αυτή.

Στο σύστημα ανάκτησης αντικειμένων μέσα από καρέ ταινίων *Video Google*, οι Sivic και Zisserman χρησιμοποιούν τον αλγόριθμο *k-means* για συσταδοποίηση και δημιουργία *οπτικού «λεξικού»*. Στη συνέχεια για την αναπαράσταση των καρέ, επιστρατεύουν τεχνικές *συχνότητας εμφάνισης όρων* που χρησιμοποιούνται για την ανάκτηση κειμένου από μηχανές αναζήτησης παίρνοντας σαν «λέξεις» τα στοιχεία του *οπτικού «λεξικού»*. Τέλος, χρησιμοποιούν μια τεχνική χωρικής συνεκτικότητας σύμφωνα με την οποία ένα ταίριασμα σημείων είναι σωστό αν στην γύρω περιοχή υπάρχουν κι άλλα στοιχεία που ταιριάζουν [Sivic & Zisserman, 2003].

Στην τεχνική ανάκτησης με σημεία ενδιαφέροντος που υλοποιήθηκε στα πλαίσια αυτής της διπλωματικής εργασίας, κατασκευάστηκε *οπτικό λεξικό* και έπειτα αναπαραστάθηκε η κάθε εικόνα με ένα ιστόγραμμα «ψήφων» που δείχνει πόσες φορές «ψηφίστηκε» κάποια από τις *οπτικές λέξεις* (*Visual Words*) ως ο κοντινότερος γείτονας στα σημεία ενδιαφέροντος της εικόνας. Αναλυτική περιγραφή της όλης διαδικασίας γίνεται σε επόμενο κεφάλαιο.

Μια ιδιαίτερα πρόσφατη τάση είναι να μην αντιστοιχείται κάθε σημείο μιας εικόνας σε μία μονάχα οπτική λέξη. Πρόσφατα προτάθηκαν μεταξύ άλλων τρεις τεχνικές στις οποίες η αντιστοίχηση

σημείων και οπτικών λέξεων δεν γίνεται ένα προς ένα ντετερμινιστικά, αλλά είτε μέσω πυρήνων γκαουσιανών [Gemert et al., 2008], είτε με επιπρόσθετη πληροφορία για την σχετική θέση σημείου-οπτικής λεξης [Jegou et al., 2008a], είτε με την αντιστοίχηση παραπάνω από μια οπτικών λέξεων σε κάθε σημείο [Philbin et al., 2008].

1.8 Τύποι ερωτήματος και οπτικοποίηση των αποτελεσμάτων

Στον χρήστη ενός συστήματος ανάκτησης εικόνων, μπορούν να παρέχονται διάφοροι τρόποι με τους οποίους αυτός θα εκφράσει το ερώτημά του ώστε να πάρει τα επιθυμητά αποτελέσματα. Είναι πολύ χρήσιμο να μπορεί ο χρήστης να διαλέξει μεταξύ διαφόρων μορφών θέσης του ερωτήματος, για να μπορέσει να βρει τον καλύτερο ανάλογα με την περίσταση. Μερικές μορφές ερωτήματος είναι:

- *Ερώτημα με κείμενο*: Ο χρήστης θέτει το ερώτημά του πληκτρολογώντας κάποια λέξη ή λέξεις που σχετίζονται με το περιεχόμενο της εικόνας. Αυτή η μέθοδος είναι η πλέον διαδεδομένη σήμερα, με μηχανές αναζήτησης όπως τα Yahoo και Google να το χρησιμοποιούν. Βέβαια, στην περίπτωση αυτών των μηχανών αναζήτησης, η αναζήτηση δεν γίνεται με οπτικά χαρακτηριστικά, αλλά με μετα-δεδομένα τριγύρω από την εικόνα, όπως για παράδειγμα από το κείμενο που την συνοδεύει. Για να γίνει αναζήτηση με οπτικά χαρακτηριστικά από ερώτημα με κείμενο, θα πρέπει να προϋπάρχει γνώση που συνδέει τα χαμηλού επιπέδου οπτικά χαρακτηριστικά με τις υψηλού επιπέδου έννοιες .

- *Ερώτημα με εικόνα*: Ο χρήστης είτε ανεβάζει μια δικιά του εικόνα είτε διαλέγει κάποια από την συλλογή και ψάχνει για παρόμοιες. Το σύστημα πρέπει να εξάγει οπτικά χαρακτηριστικά από την εικόνα του χρήστη και να τα συγκρίνει με τα αντίστοιχα χαρακτηριστικά των εικόνων της συλλογής. Αυτή είναι η πλέον διαδεδομένη μορφή ερωτήματος στην ανάκτηση εικόνων με βάση το οπτικό τους περιεχόμενο. /item *Ερώτημα με περιοχή μιας εικόνας* Ο χρήστης είτε ανεβάζει μια δικιά του εικόνα είτε διαλέγει κάποια από την συλλογή και μέσα σε αυτήν επιλέγει την περιοχή που τον ενδιαφέρει. Προϋπόθεση είναι να υπάρχει κάποιο είδος κατάτμησης των εικόνων ώστε να μπορούν να απομονωθούν οι διάφοροι τύποι περιοχών αποδοτικά. Αυτού του είδους τα ερωτήματα φιλοξενούν τα συστήματα ανάκτησης αντικειμένων, καθώς η περιοχή ενδιαφέροντος μπορεί να είναι ένα ζώο ή αντικείμενο, το οποίο πρέπει να απομονωθεί από την υπόλοιπη – μη ενδιαφέρουσα – εικόνα.

- *Ερώτημα με γράφημα-σκίτσο από τον χρήστη*: Στο γραφικό περιβάλλον του συστήματος μπορεί να υπάρχει μια περιοχή σχεδίασης, εντός της οποίας ο χρήστης σχεδιάζει απλά και χοντροκομμένα το επιθυμητό ερώτημα. Το σύστημα έπειτα, δεχόμενο το σκίτσο ως εικόνα, με μεθόδους ανάλυσης σχήματος εξάγει οπτικά χαρακτηριστικά και εκτελεί την αναζήτηση.

- *Ερώτημα με σύνθεση τύπων περιοχών*: Ο χρήστης θα μπορεί να φτιάξει ένα «προσχέδιο» της δομής των εικόνων που ψάχνει ορίζοντας ποιους τύπους περιοχών θα θέλει να υπάρχουν και ίσως και σε ποιο μέρος της εικόνας. Για παράδειγμα θα μπορεί να ζητήσει από το σύστημα εικόνες με δενδρόφυτη βουνοπλαγιά στα αριστερά και τον ήλιο να δύει στο πάνω δεξιά κομμάτι της εικόνας.

- *Σύνθετο ερώτημα*: Μπορεί να περιλαμβάνει συνδυασμό των παραπάνω μεθόδων, κυρίως μείξη εικόνας/σκίτσου με μια λίστα επιθυμητών σημασιολογικών εννοιών. Για παράδειγμα μια εικόνα που ανεβάζει ο χρήστης και απεικονίζει ένα αυτοκίνητο μπροστά από τον πύργο του Άιφελ, θα μπορεί να συνοδευτεί από τις λέξεις car, citroen και c2 για να μπορέσει ίσως να καταλάβει το σύστημα ότι ο χρήστης δεν ενδιαφέρεται για εικόνες που περιέχουν το γνωστό σιδερένιο μνημείο του Παρισιού, αλλά θέλει εικόνες που περιέχουν το αυτοκίνητο που δείχνει η εικόνα. Η έννοια του σύνθετου ερωτήματος θα επεκταθεί στην επόμενη ενότητα με την εισαγωγή της ανάδρασης σχετικότητας.

Υπάρχουν αμέτρητες μορφές ερωτήματος, μορφές που προσαρμόζονται στα συστήματα τα οποία τις χρησιμοποιούν. Σε πιο καινούριες δημοσιεύσεις εισάγονται νέες μορφές ερωτήματος. Για παράδειγμα προτείνεται από τον Assfalg η χρήση ερωτημάτων που περιέχουν *τρισδιάστατα μοντέλα*, καθώς υποστηρίζει ότι τα διδιάστατα ερωτήματα δεν περιγράφουν σωστά την χωρική κατανομή των αντικειμένων στην εικόνα [Assfalg et al., 2002]. Σε μια άλλη ενδιαφέρουσα εφαρμογή, ο Kaster προτείνει την *χρήση χειρονομιών και ομιλίας* σαν ερωτήματα και σαν ανάδραση για ανάκτηση εικόνων [Käster et al., 2003].

Η *οπτικοποίηση* των αποτελεσμάτων της αναζήτησης είναι ένας από τους πιο σημαντικούς παράγοντες που μπορούν να οδηγήσουν στην αποδοχή και δημοτικότητα ενός συστήματος ανάκτησης. Ο Datta [Datta et al., 2008] κατηγοριοποιεί διάφορους τρόπου οπτικοποίησης ως εξής:

- Ταξινόμηση αποτελεσμάτων με βάση την σχετικότητα ως προς το ερώτημα.
- Χρονολογική ταξινόμηση των αποτελεσμάτων.
- Συσταδοποιημένη οπτικοποίηση των αποτελεσμάτων.
- Ιεραρχικά δομημένη οπτικοποίηση.
- Σύνθετη οπτικοποίηση, με συνδιασμό των παραπάνω.

Η ταξινόμηση αποτελεσμάτων με βάση την σχετικότητα ως προς το ερώτημα είναι ο πιο συνηθισμένος τρόπος εμφάνισης των αποτελεσμάτων, σύμφωνα με τον οποίο τα αποτελέσματα ταξινομούνται με βάση ένα αυξάνον μέτρο απόστασης από την εικόνα - ερώτημα. Η χρονολογική ταξινόμηση των αποτελεσμάτων είναι αποδοτική σε εφαρμογές οργάνωσης προσωπικών συλλογών εικόνων. Η συσταδοποιημένη οπτικοποίηση των αποτελεσμάτων εφαρμόστηκε από τον Chen στο σύστημα *CLUE* [Chen et al., 2005]. Η ιεραρχικά δομημένη οπτικοποίηση εμφανίζει τα αποτελέσματα σε κάποια ιεραρχική δομή όπως για παράδειγμα με μορφή δέντρου.

Τρισδιάστατη απεικόνιση των αποτελεσμάτων προτείνεται από το σύστημα *3D Mars* όπου οι εικόνες εμφανίζονται σε τρισδιάστατο πλέγμα σαν σε εφαρμογή εικονικής πραγματικότητας [Nakazato & Huang, 2001]. Ιδιαίτερα σημαντικό για την δημιουργία του γραφικού περιβάλλοντος ενός συστήματος ανάπτυξης είναι να μελετηθεί και να κατανοηθεί το πώς ο χρήστης συνήθως διαχειρίζεται τις εικόνες που έχει και πως διατυπώνει και περιγράφει αυτό που θέλει να πάρει ως αποτέλεσμα. Έρευνα πάνω στην ανθρώπινη συμπεριφορά ως προς την διαχείρηση του οπτικού υλικού έχουν διεξάγει οι Rodden και Wood [Rodden & Wood, 2003] και ο Cunningham αναλύει το πώς ο άνθρωπος περιγράφει την ανάγκη του για οπτική πληροφορία [Cunningham et al., 2004] [Cunningham & Masoodian, 2006].

Περιορισμούς στην τρόπο θέσης των ερωτημάτων θέτουν φορητές συσκευές, όπως κινητά τηλέφωνα ή PDA. Τα τελευταία χρόνια κινητά τερματικά γίνονται όλο και πιο συχνά χρήστες απομακρυσμένων συστημάτων ανάκτησης πολυμέσων. Σε τέτοια συστήματα όμως, λειτουργίες όπως η κύλιση και γενικότερα η περιήγηση στο περιβάλλον δυσχεραίνεται από το μικρό μέγεθος της οθόνης. Έρευνα για την οπτικοποίηση σε τέτοιου είδους συσκευές υπάρχει με προσεγγίσεις βασισμένες στην προσοχή του χρήστη [Chen et al., 2003], και στην προσωποποίηση των σελίδων [Bertini et al., 2005].

1.9 Ανάδραση σχετικότητας

Από την δημιουργία των πρώτων συστημάτων ανάκτησης δύο κύρια προβλήματα παρατηρήθηκαν: Το *σημασιολογικό κενό*, η έλλειψη δηλαδή αντιστοιχίας ανάμεσα στην εξαγόμενη οπτική πληροφορία και την ερμηνεία της πληροφορίας αυτής στον χρήστη, και η *υποκειμενικότητα της ανθρώπινης αντίληψης* ως προς το οπτικό περιεχόμενο. Η υποκειμενικότητα υπάρχει σε πολλαπλά επίπεδα, για παράδειγμα κάποιος χρήστης μπορεί να ενδιαφέρεται για το χρώμα μιας εικόνας την μια φορά, ενώ κάτω από άλλες συνθήκες να ενδιαφέρεται για την υφή.

Γι' αυτό τον λόγο, από τα μια-και-έξω ερωτήματα των πρώιμων συστημάτων ανάκτησης, η επόμενη γενιά προσπάθησε να ενσωματώσει μια συνεχή ανάδραση από τον χρήστη, τοποθετώντας τον ουσιαστικά στο κέντρο, για να μάθει περισσότερα για τις προθέσεις - στόχους του (*human in the loop*). Η διαδραστική διαδικασία του να κρίνει ο χρήστης τα αποτελέσματα της κάθε αναζήτησης και να αναπροσαρμόζει το ερώτημα με βάση αυτά, ονομάζεται *ανάδραση σχετικότητας (Relevance Feedback)*, σε αντιστοιχία και με παλιότερες παρόμοιες προσεγγίσεις για κείμενο. Η ανάδραση σχετικότητας μπορεί να θεωρηθεί μια ειδική περίπτωση *αναδυόμενης σημασιολογίας (emergent semantics)* [Lew et al., 2006].

Στην πιο απλή μορφή της, η μέθοδος ανάδρασης σχετικότητας του Rocchio, το 1971, πρότεινε την μετακίνηση του *σημείου ερωτήματος* (query point) πιο κοντά σε σχετικά αποτελέσματα και πιο μακριά από τα μη σχετικά [Salton, 1971]. Βέβαια πλέον το ερώτημα δεν μπορεί να παρασταθεί με ένα μόνο σημείο και από το 1971 πάρα πολλές μέθοδοι προσαρμοσμένες στην ανάκτηση εικόνων με βάση το οπτικό τους περιεχόμενο έχουν προταθεί.

Ο Sclaroff χρησιμοποιεί στο σύστημά *ImageRover* αποστάσεις Minkovski για την ανάδραση σχετικότητας, με τον χρήστη να τικάρει τις σχετικές εικόνες από τα αποτελέσματα [Sclaroff et al., 2001], ενώ ο Chen χρησιμοποιεί *SVM* μιας κλάσης για την εκμάθηση της ανάδρασης [Chen et al., 2001]. Οι Tieu και Viola χρησιμοποιούν τον αλγόριθμο εκμάθησης *AdaBoost* για την ανάδραση σχετικότητας με καλά αποτελέσματα καθώς ο AdaBoost λειτουργεί καλά και με περιορισμένα σετ εκμάθησης [Tieu & Viola, 2000].

Οι Rui και Huang προτείνουν ένα πολυμεσικό μοντέλο των αντικειμένων, που περιέχει διάφορα χαρακτηριστικά και πολλαπλές αναπαραστάσεις, κάθε μια από τις οποίες συνοδεύεται με κάποιον συντελεστή - βάρος. Με τα διαδοχικά ερωτήματα και σύμφωνα με τις αποφάσεις του χρήστη, τα βάρη μεταβάλλονται, μεταβάλλοντας ταυτόχρονα και τα αποτελέσματα της αναζήτησης [Rui et al., 1997b].

Την απόδοση διαφόρων τεχνικών ανάδρασης σχετικότητας σε περιβάλλοντα με μικρές οθόνες (κινητά τηλέφωνα, PDA) μελετάει ο Vinay, εξετάζοντας το κατά πόσο η ανάδραση σχετικότητας μαζί με εναλλακτικές στρατηγικές οπτικοποίησης, μπορούν να χρησιμοποιηθούν για να μειώσουν τον αριθμό των εγγράφων που πρέπει να εξετάσει ο χρήστης της κινητής συσκευής μέχρι να φτάσει στο στόχο της αναζήτησής του. Κατασκευάζει γι' αυτό το λόγο μια δομή δέντρου με όλες τις

πιθανές ενέργειες χρήστη - συστήματος, ώστε να προσδιορίσει μια τιμή μεγίστου στην απόδοση του συστήματος [Vinay et al., 2005].

1.10 Σύγχρονα συστήματα ανάκτησης και εφαρμογές

Την πληθώρα συστημάτων ανάκτησης εικόνων με βάση το οπτικό τους περιεχόμενο που αναπτύχθηκαν μέχρι την προηγούμενη πενταετία, ερευνούν εξονυχιστικά αρκετές σχετικές βιβλιογραφικές μελέτες. Η πιο ολοκληρωμένη για συστήματα μέχρι το 2002 είναι η επισκόπηση των Veltkamp και Tanase από το πανεπιστήμιο της Ουτρέχτης [Veltkamp & Tanase, 2002]. Σε αυτή περιγράφονται αναλυτικά 46 συστήματα ανάκτησης. Τα περισσότερα από αυτά δημιουργήθηκαν ερευνητικά αλλά αρκετά από αυτά είχαν και εμπορική εφαρμογή.

Στα πιο πρόσφατα έτη, όχι μόνο δεν έχουν βγει στη δημοσιότητα πολλά συστήματα ανάκτησης βασισμένα στο οπτικό περιεχόμενό των εικόνων αλλά και η συντριπτική πλειοψηφία των παλαιότερων συστημάτων έχει σταματήσει να εξελίσσεται.

Ένα σύστημα ανάκτησης το οποίο άντεξε στον χρόνο και υπάρχει ακόμη διαθέσιμο στο ίντερνετ είναι το *Ikona*[1], δημιουργημένο από το γαλλικό ισντιτούτο Inria. Σε αυτό χρησιμοποιούνται για την σύγκριση οπτικά χαρακτηριστικά χρώματος, σχήματος και υφής. Υπάρχει επίσης η δυνατότητα ανάκτησης με χαρακτηριστικά εξαγόμενα από περιοχές της εικόνας, ανάκτηση με ανάδραση σχετικότητας, ανάκτηση με τοπικά χαρακτηριστικά των εικόνων καθώς και ανάκτηση τρισδιάστατων μοντέλων.

Το πιο ενδιαφέροντα συστήματα που εξελίσσονται σήμερα προέρχονται από την, ιδιαίτερα ενεργή στον χώρο, ομάδα του πανεπιστημίου της Οξφόρδης. Σε πλήρη εφαρμογή υπάρχουν δύο συστήματα, το ένα από τα οποία είναι το *Video Google*[2] στο οποίο ο χρήστης διαλέγει ένα αντικείμενο από κάποιο χαρακτηριστικό καρέ μιας ταινιας και αναζητώνται άλλα καρέ της ταινίας στα οποία αυτό περιέχεται. Ένα ακόμα σύστημα που μπορεί να δοκιμάσει κανείς στο σαιτ της ομάδας είναι το σύστημα που εντοπίζει κάποια κτίρια της Οξφόρδης μεσα σε μια τεράστια συλλογή φωτογραφιών[3]. Το σύστημα αυτό βασίζεται στο προηγούμενο αλλά έχουν γίνει μεγάλες βελτιώσεις πάνω στον έλεγχο χωρικής συνεκτικότητας, την δεικτοδότηση και την ταχύτητα σε σχέση με το Video Google [Chum et al., 2007] [Philbin et al., 2007].

Ένα άλλο σύγχρονο σύστημα ανάκτησης είναι το *SIMPLYcity* [4] (αρχικά για το Semantics-sensitive Integrated Matching for Picture LIbraries) ένα σύστημα ανάκτησης το οποίο είναι, όπως λέει και το όνομα του, ευαίσθητο σε σημασιολογικές έννοιες [Wang et al., 2001]. Συνδυάζει δηλαδή μεταδεδομένα της εικόνας με οπτικά χαρακτηριστικά για την εξαγωγή του αποτελέσματος. Το σύστημα αυτό συνδέεται άμεσα και με το σύστημα *Alipr*[5], το οποία συνδιάζει επίσης μεταδεδομένα των εικόνων και δίνει στον χρήστη την επιλογή ανάκτησης «σχετικών» (related) ή «παρόμοιων» (similar) εικόνων. Στην πρώτη περίπτωση, η αναζήτηση γίνεται χρησιμοποιώντας ταμπέλες (με την μορφή κειμένου) των εικόνων ενώ στην δεύτερη με οπτικά χαρακτηριστικά. Το όλο σύστημα στοχεύει

[1] http://www-rocq.inria.fr/cgi-bin/imedia/circario.cgi/demos
[2] http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html
[3] http://www.robots.ox.ac.uk/~vgg/research/oxbuildings/index.html
[4] http://wang14.ist.psu.edu/cgi-bin/zwang/regionsearch_show.cgi
[5] http://alipr.com/

στην αυτόματη γλωσσική δεικτοδότηση-ταξινόμηση των εικόνων με μεταδεδομένα και επιδιώκει μια στατιστική προσέγγιση προς αυτή την κατεύθυνση [Li & Wang, 2003]. Τα αρχικά Alipr άλλωστε σημαίνουν Automatic Linguistic Indexing and Picture Retrieval. Ένα ακόμα σύστημα που συνδιάζει παρόμοιες τεχνολογίες (και δανείζεται και το ίδιο γραφικό περιβάλλον) είναι το CLUE[6].

Από την κολοσσιαία ανάπτυξη των φωτογραφικών συλλογών που υπάρχουν στο ίντερνετ εμπνεύστηκε η ομάδα του Torralba και σχεδίασε ένα σύστημα ανάκτησης το οποίο ήθελε να περιέχει όσο το δυνατόν περισσότερες εικόνες. Ανέπτυξε λοιπόν το σύστημα *80 Million Tiny Images*[7] το οποίο περιέχει μια τεράστια συλλογή από πολύ μικρές εικόνες (32×32 εικονοστοιχείων) [Torralba et al., 2007].

Ένα ακόμα σύστημα ανάκτηση που μπορεί να βρει κανείς online στο ίντερνετ είναι το *Accio!*[8]. Σε αυτό, ο χρήστης δίνει στο σύστημα κάποιες εικόνες που περιέχουν το αντικείμενο προς ανάκτηση και κάποιες που δεν το περιέχουν, και με την χρησιμοποίηση ενός είδος κατάτμησης από σημεία ενδιαφέροντος και ενός αλγορίθμου μάθησης από πολλαπλές εικόνες ταξινομούνται οι εικόνες και επιστρέφονται στον χρήστη οι κοντινότερες.

Οι *ιατρικές εικόνες* είναι ένας τομέας στον οποίον η χρήση συστημάτων ανάκτησης με βάση το οπτικό περιεχόμενο είναι ιδιαίτερα διαδεδομένη. Τα αποτελέσματα στον τομέα αυτόν είναι εντυπωσιακά και ιατρικά κέντρα σε όλο τον κόσμο χρησιμοποιούν καθημερινά την τεχνολογία αυτή για αναγνώριση παθήσεων από ακτινογραφίες, μαγνητικές τομογραφίες ή εγκεφαλογραφήματα [Tagare et al., 1997]. Τα συστήματα αυτά δεν διαφέρουν σε πολλά σημεία από τα συστήματα ανάκτησης φωτογραφιών γενικού περιεχομένου, αλλά αποτελούν στις περισσότερες περιπτώσεις επεκτάσεις τους. Το σύστημα που προτείνουν οι Lehmann et al. από το πανεπιστήμιου του Άαχεν [Lehmann et al., 2000] [Lehmann et al., 2004], είναι μια επέκταση του γνωστού συστήματος Blobworld [Carson et al., 2002].

Επισκόπηση των συστημάτων ανάκτησης με βάση το οπτικό περιεχόμενο των εικόνων τα οποία έχουν ιατρικές εφαρμογές, έγινε το 2004 από τον Muller [Müller et al., 2004]. Σε αυτήν παρουσιάζονται τεχνικές και εφαρμογές συστημάτων ανάκτησης με γνώμονα τις ιατρικές εικόνες, δίνεται σχετική βιβλιογραφία και προτείνονται σενάρια για μελλοντική εξέλιξη του κλάδου.

[6] http://wang14.ist.psu.edu/cgi-bin/yixin/spectralsearch_show.cgi

[7] http://people.csail.mit.edu/torralba/tinyimages/

[8] http://www.cs.wustl.edu/accio/

Κεφάλαιο 2

Αναζήτηση εικόνων με βάση οπτικά χαρακτηριστικά από ολόκληρη την εικόνα

2.1 Εισαγωγή

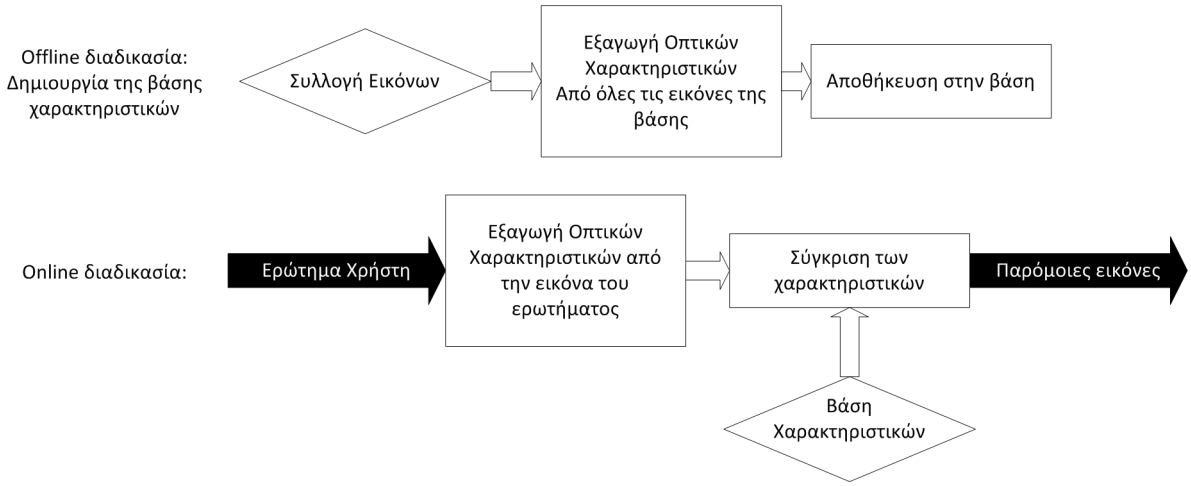
Σε αυτό το κεφάλαιο παρουσιάζεται η πιο απλή τεχνική ανάκτησης εικόνων με βάση το οπτικό τους περιεχόμενο. Σε αυτή την περίπτωση χαρακτηριστικά εξάγονται από ολόκληρη την εικόνα (*global features*) και σύμφωνα με αυτά συγκρίνεται η εικόνα του ερωτήματος με τις υπόλοιπες εικόνες της βάσης. Την προηγούμενη δεκαετία δημιουργήθηκαν πολλά παρόμοια συστήματα αναζήτησης, εμπορικά και μη, βασισμένα σε οπτικά χαρακτηριστικά χρώματος και υφής εξαγόμενα από ολόκληρη την εικόνα.

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, οι περιγραφείς που εξάγονται κατά αυτόν τον τρόπο μπορούν να περιγράψουν αποτελεσματικά εικόνες με αραιό οπτικό - σημασιολογικό περιεχόμενο, καθώς για να εντοπιστεί μια έννοια με χρήση αυτών των περιγραφέων, πρέπει να κατέχει ένα σημαντικό ποσοστό της εικόνας. Σε εικόνες με πυκνό οπτικό περιεχόμενο, οι περιγραφείς αυτοί, στην προσπάθεια τους να αναπαραστήσουν ολόκληρη την εικόνα, «μπερδεύονται» και χάνουν την διακριτικότητα τους.

Κύριο πλεονέκτημα αυτής της μεθόδου είναι ότι μπορεί να κατηγοριοποιήσει αποτελεσματικά εικόνες οι οποίες περιέχουν στο σύνολο ή σε μεγάλο ποσοστό της επιφάνειας τους έννοιες με χαρακτηριστική υφή ή χρώμα. Εικόνες από ένα ηλιοβασίλεμα, για παράδειγμα μπορούν να ξεχωριστούν εύκολα με βάση χαρακτηριστικά χρώματος από εικόνες ενός λιβαδιού ή γενικά από εικόνες με φυσικά τοπία (βλέπε σχήματα 2.6, 2.7 και 2.8) Χαρακτηριστική υφή έχουν και οι εικόνες με καταρράκτες και με τα κατάλληλα χαρακτηριστικά υφής μπορούν να διαφοροποιηθούν από άλλες – παρόμοιες χρωματικά – εικόνες.

Δυσκολίες στην ανάκτηση μπορούν να παρατηρηθούν σε ανομοιόμορφες συλλογές εικόνων. Αν, για παράδειγμα, το ταίριασμα γίνει μονάχα με χρωματικά κριτήρια, οι δύο εικόνες του σχήματος 2.1 θα επιστραφούν ως παρόμοιες. Το ίδιο μπέρδεμα μπορεί να παρατηρηθεί σε άλλες περιπτώσεις εικόνων αν χρησιμοποιηθούν αποκλειστικά περιγραφείς υφής. Τα πειράματα έδειξαν ότι ένας συνδυασμός περιγραφέων υφής και χρώματος μπορεί να έχει πολύ καλά αποτελέσματα σε εικόνες από φυσικά τοπία.

Σχήμα 2.1: Οι δύο εικόνες μπορεί να επιστραφούν ως παρόμοιες αν χρησιμοποιηθούν αποκλειστικά χρωματικοί περιγραφείς.



Σχήμα 2.2: Σχηματικό διάγραμμα της τεχνικής αναζήτησης με χαρακτηριστικά εξαγόμενα από ολόκληρη την εικόνα.

Σύμφωνα με την απλή διαδικασία του σχήματος 2.2, οπτικά χαρακτηριστικά εξάγονται συνολικά από κάθε εικόνα της συλλογής. Τα χαρακτηριστικά αυτά σχηματίζουν ένα πολυδιάστατο διάνυσμα αναπαράστασής της και αποθηκεύονται στην βάση δεδομένων του συστήματος. Αυτή η διαδικασία εκτελείται μία φορά για την βάση δεδομένων όχι σε πραγματικό χρόνο (offline) και πρέπει να ξαναεκτελεστεί όταν επιθυμείται να εισαχθούν καινούριες εικόνες στην βάση. Όταν γίνει ένα *ερώτημα* (*query*) για αναζήτηση, τα οπτικά χαρακτηριστικά της εικόνας του ερωτήματος εξάγονται την στιγμή εκείνη (online) και συγκρίνονται με τα αντίστοιχα χαρακτηριστικά όλων των εικόνων από την βάση. Από την σύγκριση επιστρέφονται στον χρήστη οι κοντινότερες, σύμφωνα με κάποια απόσταση από την εικόνα του ερωτήματος, εικόνες ως παρόμοιες. Φυσικά τα οπτικά χαρακτηριστικά δεν χρειάζεται να εξαχθούν ξανά αν η εικόνα του ερωτήματος είναι μια από τις εικόνες της βάσης. Σε αυτή την περίπτωση τα χαρακτηριστικά λαμβάνονται από την βάση.

2.2 Οι οπτικοί περιγραφείς του MPEG7

Για την εξαγωγή των χαρακτηριστικών από τις εικόνες χρησιμοποιούμε οπτικούς περιγραφείς του προτύπου *MPEG-7* [Chang et al.,]. Το MPEG-7, γνωστό και ως Διεπαφή Περιγραφής Πολυμεσικού Περιεχομένου (Multimedia Content Description Interface), είναι το τελευταίο πρότυπο που ανέπτυξε το MPEG, μετά τα ιδιαίτερα διαδεδομένα πρότυπα MPEG-1, MPEG-2 και MPEG-4. Ενώ οι πρόγονοί του εστίαζαν στην κωδικοποίηση, συμπίεση και αναπαράσταση οπτικοακουστικού υλικού, το MPEG-7 εστιάζει στην περιγραφή του πολυμεσικού περιεχομένου [Sikora, 2001]. Αναφέρεται στο περιεχόμενο διαφορετικών κατηγοριών όπως εικόνα, βίντεο, ήχος, ομιλία, γραφικά καθώς και συνδυασμούς των κατηγοριών αυτών. Η ανάπτυξή του άρχισε το 1998 και έγινε διεθνές πρότυπο το 2001. Παρακάτω επεξηγούνται συνοπτικά οι τρεις περιγραφείς χρώματος και οι δύο περιγραφείς υφής του MPEG-7 που χρησιμοποιήθηκαν [Manjunath et al., 2001]. Για την εξαγωγή των περιγραφέων από τις εικόνες χρησιμοποιήθηκε και ενσωματώθηκε στο σύστημα το πρόγραμμα *VDE*[1] (*Visual Descriptor Extraction*), περισσότερες πληροφορίες για το οποίο υπάρχουν στο [Tolias, 2007].

2.2.1 Κλιμακωτός περιγραφέας χρώματος

Ο *κλιμακωτός περιγραφέας χρώματος* (*Scalable Color Descriptor*) είναι ένα σχήμα κωδικοποίησης βασισμένο στην μετατροπή Haar πάνω στις τιμές ενός ιστογράμματος χρώματος στον HSV χώρο. Το ιστόγραμμα κβαντίζεται στα 256 bins και στη συνέχεια, με άθροιση γειτονικών κορυφών μπορεί να μειωθεί περαιτέρω το μέγεθος του περιγραφέα. Η βασική μονάδα της μετατροπής Haar αποτελείται από μια λειτουργία αθροίσματος (βαθυπερατό φίλτρο) και μια λειτουργία διαφοράς (υψιπερατό φίλτρο). Η μορφή του διανύσματος που προκύπτει δίνεται από τη σχέση

$$SCD = [c_1, c_1, \ldots, c_N] \tag{2.1}$$

όπου N είναι το μέγεθος του ιστογράμματος.

2.2.2 Περιγραφέας δομής χρώματος

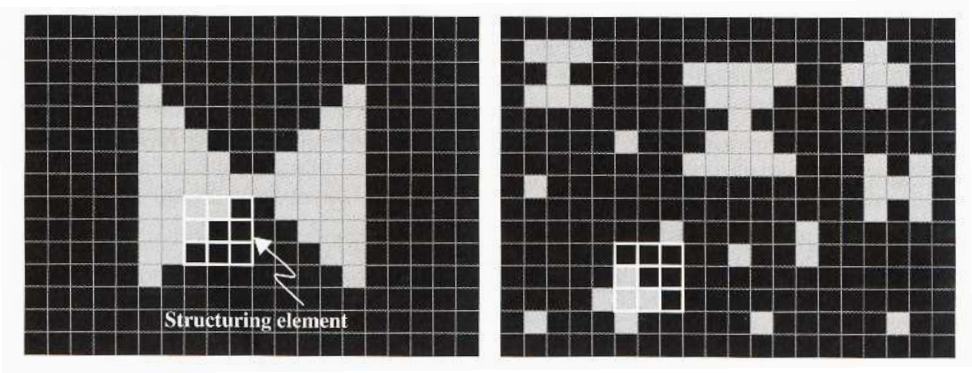
Ο *περιγραφέας δομής χρώματος* (*Color Structure Descriptor*) προσπαθεί να αποτυπώσει τα τοπικά χρωματικά χαρακτηριστικά της εικόνας. Για να το πετύχει αυτό, ένα παράθυρο διάστασης 8x8 pixels σαρώνει την εικόνα. Σε κάθε μετατόπιση του παραθύρου, καταμετράται ο αριθμός των φορών που το κάθε χρώμα περιλαμβάνεται μέσα στο δομικό στοιχείο, και δημιουργείται έτσι ένα ιστόγραμμα χρωμάτων. Οι δύο εικόνες του σχήματος 2.3 είναι παρόμοιες σαν ιστογράμματα (2 κορυφών), οι περιγραφείς δομής χρώματός τους όμως είναι εντελώς διαφορετικοί. Η μορφή του διανύσματος δίνεται από την σχέση

$$CSD = [h_1, h_1, \ldots, h_N] \tag{2.2}$$

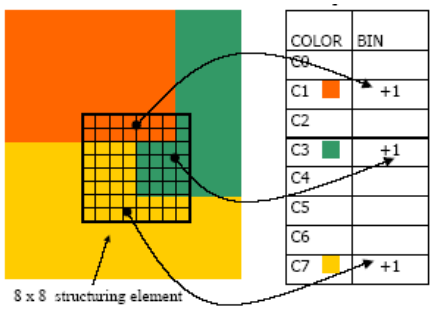
όπου N είναι το μέγεθος του ιστογράμματος, δηλαδή ο αριθμός των χρωμάτων. Για να είναι αποδοτικός ο περιγραφέας, το χρώμα κβαντίζεται σε 256 επίπεδα και έτσι οδηγούμαστε σε ενα διάνυσμα περιγραφής με 256 στοιχεία. Ο υπολογισμός των τιμών του περιγραφέα στην περίπτωση μιας

[1] http://image.ntua.gr/smag/tools/vde

έγχρωμης εικόνας (με 8 διαφορετικά χρώματα) φαίνεται στο σχήμα 2.4. Το δομικό στοιχείο είναι τετραγωνικό και έχει το μέγεθος 8×8 εικονοστοιχεία.



Σχήμα 2.3: Εικόνες 2 χρωματικών επιπέδων με διαφορετική δομή χρώματος (από το [Manjunath et al., 2001]).

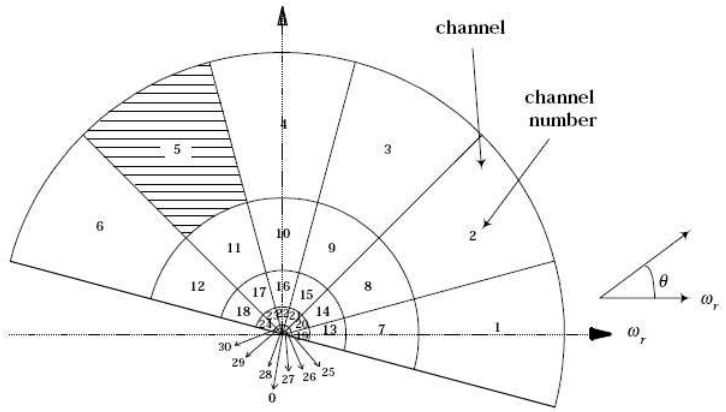


Σχήμα 2.4: Υπολογισμός περιγραφέα δομής χρώματος (από το [Manjunath et al., 2001]).

2.2.3 Περιγραφέας διάταξης χρώματος

Ο *περιγραφέας διάταξης χρώματος* (*Color Layout Descriptor*) αναπαριστά αποτελεσματικά την χωρική κατανομή του χρώματος στην εικόνα. Καθώς δίνει βάση στην χωρική δομή, είναι ιδιαίτερα χρήσιμος σε εφαρμογές ανάκτησης με βάση σκίτσο του χρήστη. Για την εξαγωγή του περιγραφέα, η εικόνα διαιρείται σε 64 μπλοκ, σε κάθε ένα από τα οποία εξάγεται ένα αντιπροσωπευτικό χρώμα. Στην μικροσκοπική εικόνα μεγέθους 8×8 που δημιουργείται από την παραπάνω διαδικασία εκτελείται διακριτός μετασχηματισμός συνημιτόνου (ΔΜΣ- Discrete Cosine Transform - DCT) ξεχωριστά σε κάθε ένα από τα τρία χρωματικά κανάλια. Εν τέλει, ο περιγραφέας εξάγεται από τους πρώτους N μη ομοιόμορφα κβαντισμένους συντελεστές (με σάρωση σε οδοντωτή τροχιά - zigzag scan) του μετασχηματισμού για το κάθε κανάλι. Ο περιγραφέας διάταξης χρώματος ορίζεται από την σχέση

$$CLD = \left[\left\{DY_{DC}, DY_{AC_i}\right\}, \left\{DCr_{DC}, DCr_{AC_j}\right\}, \left\{DCb_{DC}, DCb_{AC_k}\right\}\right] \tag{2.3}$$



Σχήμα 2.5: Ο διαμερισμός της συχνότητας σε 30 κανάλια για την εξαγωγή του περιγραφέα ομοιογενούς υφής (από το [Manjunath et al., 2001]).

όπου τα i, j, k δηλώνουν τον αριθμό των AC συντελεστών και μπορούν να πάρουν τις τιμές 3, 6, 10, 15, 21, 28, 64. Χρησιμοποιήθηκαν 12 συντελεστές συνολικά, 6 για την φωτεινότητα και 3 για κάθε χρωματικό κανάλι στον χρωματικό χώρο YCrCb.

2.2.4 Περιγραφέας ομοιογενούς υφής

Ο *περιγραφέας ομοιογενούς υφής* (*Homogeneous Texture Descriptor*) περιγράφει την κατευθυντικότητα, τραχύτητα και τακτικότητα διάφορων προτύπων/περιοχών της εικόνας. Βασίζεται στον υπολογισμό τοπικών χωρικών και συχνοτικών στατιστικών της υφής. Η εικόνα φιλτράρεται με φίλτρα προσανατολισμού και κλίμακας και υπολογίζεται η μέση τιμή και η τυπική απόκλιση (f_{DC}, f_{SD}) των αποτελεσμάτων στο πεδίο της συχνότητας. Το σχήμα 2.5 δείχνει το πώς χωρίζεται το επίπεδο των συχνοτήτων σε 30 κανάλια από τα φίλτρα. Η όλη διαδικασία επιταχύνεται σημαντικά αν οι παραπάνω τιμές υπολογιστούν στην περιοχή της συχνότητας και όχι στην χωρική περιοχή. Ο περιγραφέας ομοιογενούς υφής ορίζεται από την παρακάτω σχέση:

$$HTD = [f_{DC}, f_{SD}, e_1, e_2, \ldots, e_N, d_1, d_2, \ldots, d_N] \tag{2.4}$$

Όπου f_{DC} είναι η μέση και f_{SD} η τυπική απόκλιση των τιμών της εικόνας, N ο αριθμός των καναλιών στα οποία χωρίζει ο περιγραφέας τον χώρο συχνοτήτων και e_i και d_i η μέση ενέργεια και η απόκλιση ενέργειας του καθενός από αυτά. Το τελικό διάνυσμα του περιγραφέα έχει μήκος 62 στοιχείων.

2.2.5 Περιγραφέας Ιστογράμματος ακμών

Ο *περιγραφέας ιστογράμματος ακμών* (*Edge Histogram Descriptor*) είναι ένα διάνυσμα το οποίο αναπαριστά τη χωρική κατανομή των ακμών μέσα στην εικόνα. Η κατανομή των ακμών είναι μια καλή αναπαράσταση υφής, χρήσιμη για το ταίριασμα φυσικών εικόνων.

Για τον υπολογισμό του περιγραφέα, η εικόνα διαιρείται σε 4×4 υποεικόνες κι έπειτα η κάθε μία από αυτές διαιρείται επιπλέον σε μη επικαλυπτόμενα τετράγωνα μπλοκ, το μέγεθος των οποίων

καθορίζεται ανάλογα με την ανάλυση της εικόνας. Ο αριθμός των μπλοκ είναι σταθερός για όλες τις εικόνες και είναι ίσος με 1100. Οι ακμές χωρίζονται σε 5 κατηγορίες: οριζόντιες, κάθετες, διαγώνιες με προσανατολισμό 45°, διαγώνιες με προσανατολισμό 135° και μη κατευθυντικές ακμές.

Στην συνέχεια κάθε ένα από τα μπλοκ ταξινομείται σε μία από τις πέντε κατηγορίες ακμών ή σαν μπλοκ χωρίς ακμές. Για να γίνει η κατηγοριοποίηση εφαρμόζονται ανιχνευτές ακμών σε κάθε μπλοκ αφού αυτά θεωρηθούν σαν εικόνα 2×2. Μετά την κατηγοριοποίηση όλων των μπλοκ, υπολογίζονται τα ιστογράμματα ακμών με 5 κορυφές, μία για κάθε είδος ακμής, για τις 16 υποεικόνες. Έχουμε δηλαδή συνολικά $16 \times 5 = 80$ κορυφές, άρα ο περιγραφέας ιστογράμματος ακμών είναι ένα 80-διάστατο διάνυσμα.

2.3 Ταίριασμα περιγραφέων

Μετά την εξαγωγή των περιγραφέων από μια εικόνα, την εικόνα αυτήν αναπαριστούν πια στο σύστημα μας όχι οι τιμές φωτεινότητας των εικονοστοιχείων της, αλλά οι τιμές των περιγραφέων που εξήχθησαν. Αν υπολογίσουμε τα επιμέρους μεγέθη του κάθε περιργαφέα, πλέον η κάθε εικόνα αναπαρίσταται στο σύστημα μας με ένα διάνυσμα διάστασης 666, ανεξαρτήτως του μεγέθους και της ανάλυσης της (πίνακας 2.1). Πρέπει να σημειωθεί ότι για να εξαχθούν σωστά όλοι οι περιγραφείς από μια εικόνα, απαιτείται να έχει ελάχιστο μέγεθος 128×128 εικονοστοιχεία.

SC	CS	EH	HT	CL
256	256	80	62	12

Πίνακας 2.1: Το διάνυσμα οπτικών περιγραφέων μιας εικόνας. Οι συντομογραφίες των ονομάτων των περιγραφέων στον πίνακα 2.2.

Η δεικτοδότηση στην απλή αυτή μορφή συστήματος είναι εξίσου απλή, καθώς για κάθε εικόνα σώζεται στην βάση δεδομένων το διάνυσμα του σχήματος 2.1. Οι τιμές του κάθε περιγραφέα σώζονται σε διακριτά πεδία της βάσης για ταχύτητα, καθώς, όπως αναφέρεται παρακάτω, για την σύγκριση των εικόνων υπολογίζονται ξεχωριστά οι επιμέρους αποστάσεις για κάθε περιγραφέα.

Για τη σύγκριση δύο εικόνων, υπολογίζονται αρχικά οι επιμέρους αποστάσεις μεταξύ των αντίστοιχων περιγραφέων τους και έπειτα, μετά και από μια διαδικασία κανονικοποίησης, βγαίνει από αυτές τις επιμέρους αποστάσεις η μέση απόσταση, ανάλογα και με τα βάρη που έχουν καθοριστεί για τον κάθε περιγραφέα.

Το πρότυπο MPEG-7 δεν ορίζει ρητά κάποια συνάρτηση υπολογισμού απόστασης για όλους τους περιγραφείς. Πειράματα όμως έχουν δείξει ότι η ευρέως χρησιμοποιούμενη και ιδιαίτερα απλή L1 απόσταση μπορεί να εφαρμοστεί με επιτυχία για τη σύγκριση των περιγραφέων δομής χρώματος, ιστογράμματος ακμών, ομοιογενούς υφής και τον κλιμακωτό περιγραφέα χρώματος. Οι εξισώσεις δίνονται παρακάτω:

$$D(SCD_1, SCD_2) = \sum_{i=1}^{256} \mid c_{1i} - c_{2i} \mid \tag{2.5}$$

$$D(CSD_1, CSD_2) = \sum_{i=1}^{256} \mid h_{1i} - h_{2i} \mid \tag{2.6}$$

$$D(HTD_1, HTD_2) = \sum_{i=1}^{60} \mid h_{1i} - h_{2i} \mid \tag{2.7}$$

$$D(EHD_1, EHD_2) = \sum_{i=1}^{80} \mid e_{1i} - e_{2i} \mid \tag{2.8}$$

Για τον υπολογισμό της απόστασης μεταξύ δύο περιγραφέων διάταξης χρώματος, το MPEG-7 προτείνει την εξής συνάρτηση:

$$D(CLD_1, CLD_2) = \sqrt{\sum (w_{yi}(DY_i^1 - DY_i^2)^2 + w_{rj}(DCr_j^1 - DCr_j^2)^2 + w_{bk}(DCb_k^1 - DCb_k^2)^2)} \tag{2.9}$$

Όπου τα βάρη w_{yi} ,w_{rj} , w_{bk} είναι στην προκειμένη περίπτωση ίσα με 1.

Για τον υπολογισμό της συνολικής απόστασης μεταξύ των δύο εικόνων, υπολογίζεται ο μέσος όρος των παραπάνω αποστάσεων με βάρη, σύμφωνα με την σχέση:

$$D(I_1, I_2) = \frac{w_{CS} \cdot dist_{CS} + w_{EH} \cdot dist_{EH} + w_{HT} \cdot dist_{HT} + w_{CL} \cdot dist_{CL} + w_{SC} \cdot dist_{SC}}{w_{CS} + w_{EH} + w_{HT} + w_{CL} + w_{SC}} \tag{2.10}$$

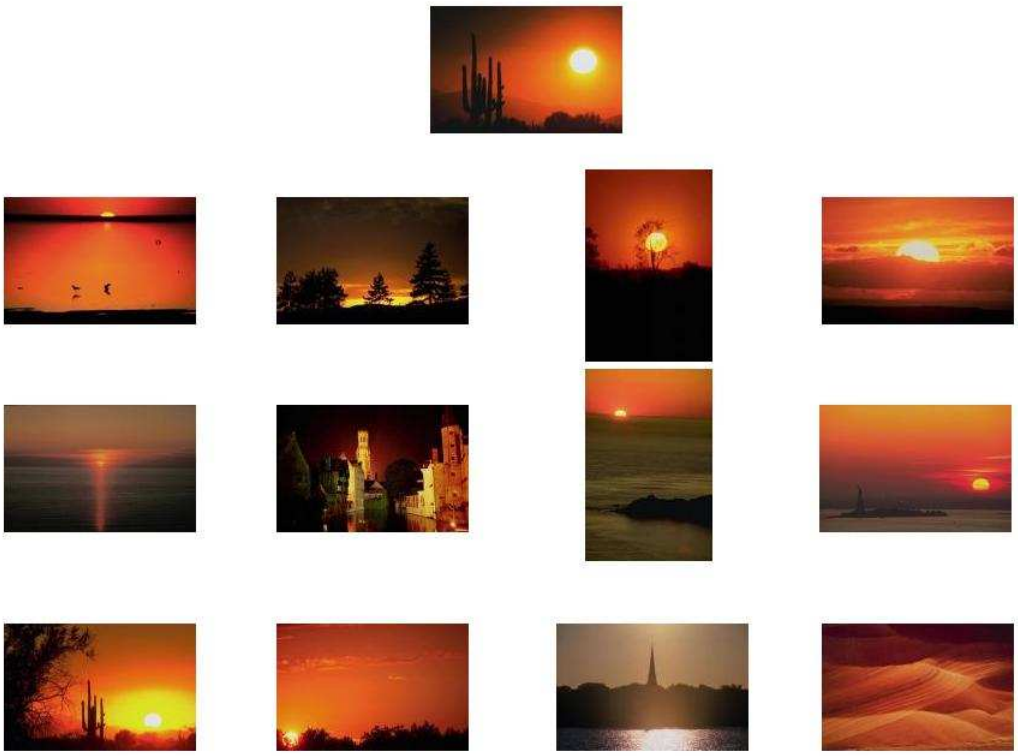
όπου $dist_X$ είναι η απόσταση των δύο εικόνων σύμφωνα με τον περιγραφέα X και w_X το αντίστοιχο βάρος του. Οι συντομογραφίες των περιγραφέων φαίνονται στον πίνακα 2.2.

Ονομασία περιγραφέα	**Συντομογραφία**
Περιγραφέας Δομής Χρώματος	**CS**
Κλιμακωτός Περιγραφέας Χρώματος	**SC**
Περιγραφέας Ιστογράμματος Ακμών	**EH**
Περιγραφέας Ομοιογενούς Υφής	**HT**
Περιγραφέας Διάταξης Χρώματος	**CL**

Πίνακας 2.2: Συντομογραφίες των περιγραφέων

Στην «απλή» αυτοματοποιημένη (με ένα κλικ) αναζήτηση στο σύστημα όλα τα βάρη έχουν τιμή 1, άρα όλοι οι περιγραφείς συμμετέχουν εξίσου στην τελική ανάκτηση. Το χρώμα έχει έτσι ένα προβάδισμα σε σχέση με την υφή καθώς χρησιμοποιούνται τρεις περιγραφείς χρώματος και δύο υφής.

Μερικά αποτελέσματα με αυτά τα βάρη φαίνονται στα σχήματα 2.6, 2.7 και 2.8. Στο πάνω μέρος είναι η εικόνα του ερωτήματος και από κάτω ταξινομημένες από τα αριστερά προς τα δεξιά και από πάνω προς τα κάτω οι κοντινότερες 12 εικόνες. Στα παραδείγματα φαίνεται καθαρά ότι οι ανακτώμενες ως κοντινότερες εικόνες μοιάζουν πολύ με την εικόνα του ερωτήματος ως προς το χρώμα και την υφή.

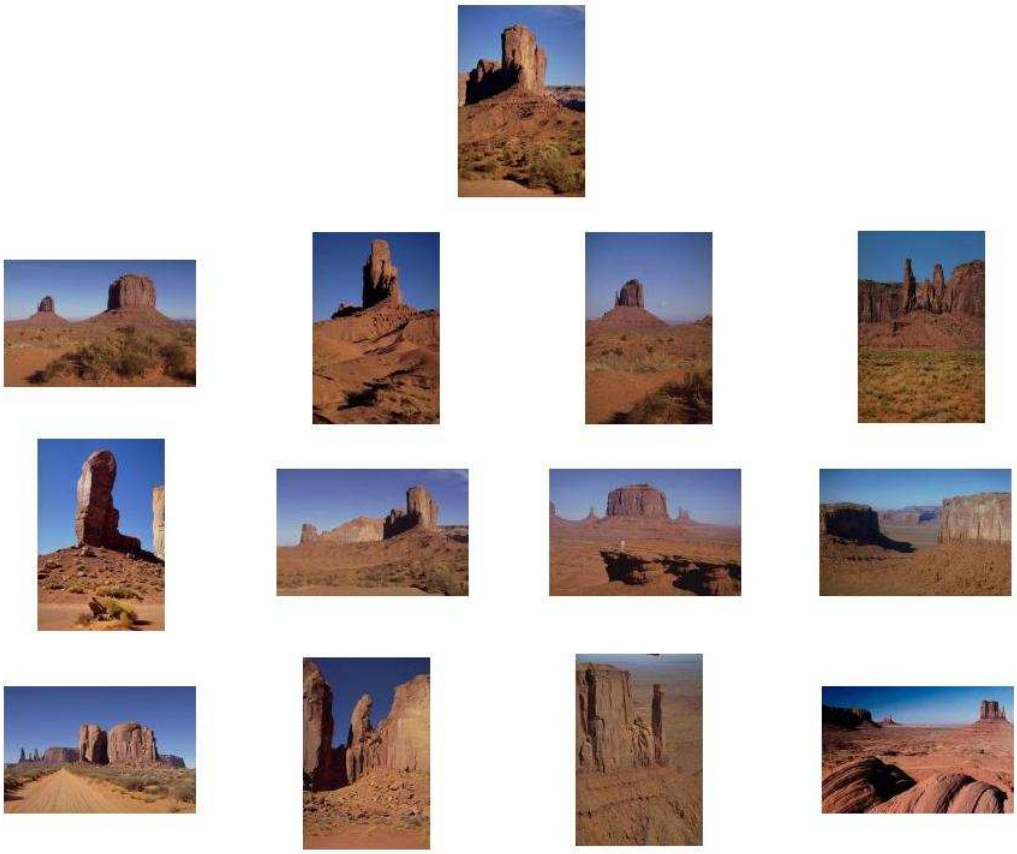


Σχήμα 2.6: Αποτελέσματα αναζήτησης με βάρη ίσα με 1 σε συλλογή με φυσικές εικόνες από το Corel. Ηλιοβασίλεμα.

2.3.1 Προηγμένες δυνατότητες αναζήτησης

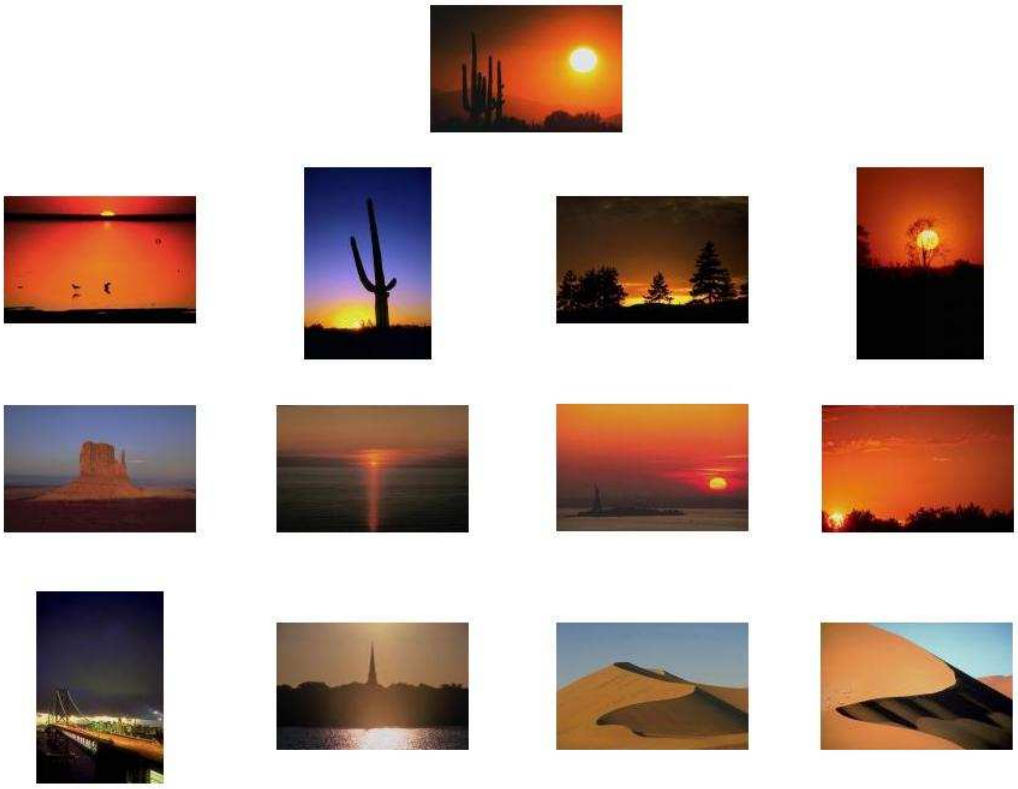
Στον χρήστη του συστήματος δίνεται η δυνατότητα να μεταβάλλει κατά βούληση τα βάρη των περιγραφέων. Έτσι ο προηγμένος χρήστης, μεταβάλλοντας τα βάρη, μπορεί να χρησιμοποιήσει τους κατάλληλους περιγραφείς ώστε να κατευθύνει την αναζήτηση προς τα χαρακτηριστικά που επιθυμεί.

Στο σχήμα 2.9 φαίνονται τα αποτελέσματα αναζήτησης της εικόνας του σχήματος 2.6, αλλά μονάχα με τους περιγραφείς υφής. Μπορεί να παρατηρήσει εύκολα κανείς ότι όλες οι «λάθος» ανακτήσεις μοιάζουν με την εικόνα του ερωτήματος ως προς την υφή.



Σχήμα 2.7: Αποτελέσματα αναζήτησης με βάρη ίσα με 1 σε συλλογή με φυσικές εικόνες από το Corel. Έρημος.

Σχήμα 2.8: Αποτελέσματα αναζήτησης με βάρη ίσα με 1 σε συλλογή με φυσικές εικόνες από το Corel. Βλάστηση.



Σχήμα 2.9: Αποτελέσματα αναζήτησης μόνο με περιγραφείς υφής σε συλλογή με φυσικές εικόνες από το Corel. Ηλιοβασίλεμα

Κεφάλαιο 3

Το μοντέλο bag-of-words και αναζήτηση με βάση οπτικά χαρακτηριστικά περιοχών

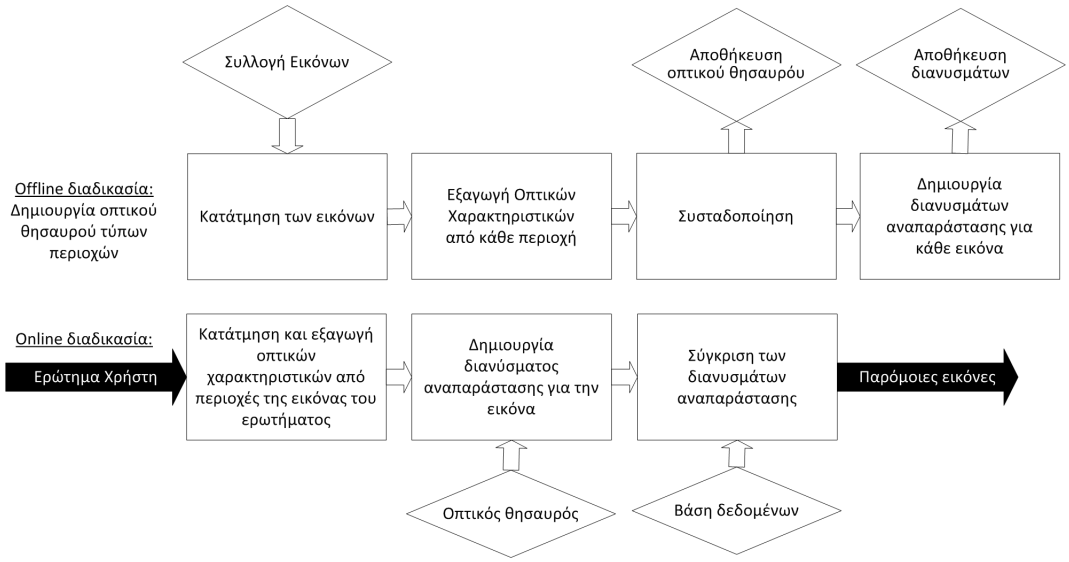
3.1 Εισαγωγή

Όταν το σημασιολογικό περιεχόμενο των εικόνων εμπεριέχει περισσότερες από μια έννοιες υψηλού επιπέδου, τότε μια περιγραφή εξαγόμενη από ολόκληρη την εικόνα περιγράφει όλες τις περιεχόμενες έννοιες μαζί. Καθίσταται λοιπόν σχεδόν αδύνατο το να διαχωριστούν πολλαπλές έννοιες αυτές. Αυτό έχει ως αποτέλεσμα, εικόνες που περιέχουν υποσύνολό των ζητούμενων εννοιών της εικόνας του ερωτήματος, να μην επιστρέφονται σαν παρόμοιες. Ερευνήθηκε προς αυτή την κατεύθυνση μια τεχνική αναζήτησης που βασίζεται σε χαρακτηριστικά περιοχών της εικόνας, σαν μια προσπάθεια να αποτυπωθούν καλύτερα και να είναι διαχωρίσιμες πολλαπλές έννοιες υψηλού επιπέδου.

Σε αυτό το κεφάλαιο παρουσιάζεται αυτή η πιο σύνθετη τεχνική ανάκτησης εικόνων. Σύμφωνα με αυτή, η κάθε εικόνα υπόκειται αρχικά σε μια διαδικασία κατάτμησης. Χωρίζεται σε περιοχές, από τις οποίες έπειτα εξάγονται τα οπτικά χαρακτηριστικά. Έπειτα, για να επιτευχθεί ταχύτητα και αποδοτικότητα, εκτελείται συσταδοποίηση όλων των οπτικών χαρακτηριστικών των περιοχών των εικόνων της βάσης, με αποτέλεσμα να εξάγονται κάποιες ομάδες περιοχών από διάφορες εικόνες, οι οποίες μοιάζουν ως προς τα οπτικά χαρακτηριστικά. Μπορούμε να πούμε ότι κάθε μια από αυτές σχηματίζει και έναν τύπο περιοχής.

Για κάθε εικόνα μετά την κατάτμηση, μπορεί να υπολογιστεί η απόσταση – με βάση κάποιο κριτήριο – των περιοχών της από κάθε τύπο περιοχής, συγκρίνοντας με τα οπτικά χαρακτηριστικά τους. Αν από κάθε περιοχή κρατηθεί η ελάχιστη από αυτές τις αποστάσεις, τότε αυτή θα εμφανίζεται στον τύπο περιοχής, στον οποίο η συγκεκριμένη περιοχή είναι πιο «κοντά» . Η εικόνα μπορεί πλέον να αναπαρίσταται από τους κοντινότερους τύπους περιοχών των περιοχών της καθώς και από τις αντίστοιχες αποστάσεις. Αυτά ορίζουν το *διάνυσμα αναπαράστασης* της εικόνας. Για την σύγκριση των εικόνων, μπορούν να συγκριθούν τα αντίστοιχα διανύσματα αναπαράστασης των εικόνων. Στον χρήστη επιστρέφονται ως παρόμοιες οι εικόνες τα διανύσματα αναπαράστασης των οποίων απέχουν λιγότερο από το αντίστοιχο διάνυσμα της εικόνας του ερωτήματος.

Η δομή του συστήματος φαίνεται στο σχήμα 3.1. Η διαδικασία δημιουργίας του οπτικού θησαυρού, δηλαδή του λεξικού τύπων περιοχών, καθώς και η εξαγωγή διανυσμάτων αναπαράστασης για



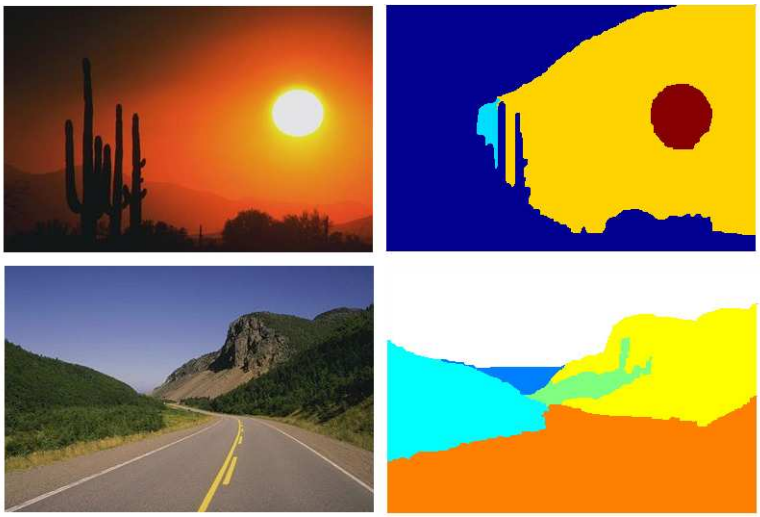
Σχήμα 3.1: Σχηματικό διάγραμμα του συστήματος αναζήτησης με βάση χαρακτηριστικά εξαγόμενα από περιοχές της εικόνας.

κάθε μια από τις εικόνες της βάσης εκτελείται μία φορά. Σε πραγματικό χρόνο ερωτήματος η διαδικασία κατάτμησης, εξαγωγής χαρακτηριστικών και διανύσματος αναπαράστασης εκτελείται μόνο για την εικόνα του ερωτήματος, εφόσον αυτή δεν ανήκει ήδη στη βάση. Αρκετά παρόμοια συστήματα αναζήτησης έχουν αναπτυχθεί μετά το 2000, αλλά η εξέλιξη τους έχει στις περισσότερες περιπτώσεις πλέον σταματήσει.

3.2 Κατάτμηση των εικόνων

Ως κατάτμηση μιας εικόνας ορίζεται η διαδικασία κατά την οποία μια εικόνα διαχωρίζεται σε ομοιόμορφες ως προς κάποιο κριτήριο περιοχές, οι οποίες είναι επιθυμητό να αντιστοιχούν σε αντικείμενα, τμήματα αντικειμένων ή γενικότερα έννοιες. Στην προκειμένη περίπτωση θα θέλαμε ιδανικά από την κατάτμηση να πάρουμε τις περιοχές των εικόνων που αντιστοιχούν στις σημασιολογικές έννοιες που περιέχονται στην εικόνα, ώστε να τις έχουμε διαχωρίσιμες. Κατά την ιδανική αυτή περίπτωση θα μπορούσαμε με τους οπτικούς περιγραφείς που αναφέρθηκαν στην παράγραφο 2.2 να εκφράσουμε επακριβώς κάθε υψηλού επιπέδου έννοια με χαμηλού επιπέδου χαρακτηριστικά χωρίς να ενυπάρχει «θόρυβος» από άλλες έννοιες στην περιοχή που αναπαριστούμε με τους οπτικούς περιγραφείς.

Η κατάτμηση των εικόνων αποτελεί την βάση της περαιτέρω ανάλυσης και η επιτυχία της καθορίζει και τη γενικότερη επιτυχία των επόμενων αποτελεσμάτων. Δυστυχώς η ιδανική κατάτμηση είναι πρακτικά αδύνατο να επιτευχθεί καθώς είναι και σε πολλές περιπτώσεις ουσιαστικά αδύνατο να οριστεί μια ιδανική κατάτμηση. Η επιλογή της μεθόδου και των παραμέτρων της κατάτμησης εξαρτάται πολύ από την συλλογή των εικόνων στις οποίες θα εφαρμοστεί.



Σχήμα 3.2: Εικόνες της συλλογής Corel (αριστερά) και η κατάτμηση τους (δεξιά).

Στην υλοποίηση που πραγματοποιήθηκε, χρησιμοποιήθηκε μια μέθοδος κατάτμησης που βασίζεται σε συντακτικά οπτικά χαρακτηριστικά [Adamek et al., 2005]. Η συγκεκριμένη μέθοδος επιλέχτηκε γιατί η κατάτμηση της οδηγεί συνήθως σε μεγάλες περιοχές, που ανταποκρίνονται στην χωρική δομή των περιεχόμενων στις φυσικές εικόνες εννοιών στις οποίες δοκιμάστηκε το σύστημα. Έννοιες όπως «βλάστηση», «χιόνι» και «έρημος» αναμένεται να καταλαμβάνουν ένα μεγάλο ποσοστό της εικόνας στην οποία εμπεριέχονται.

Στη σχετική δημοσίευση [Adamek et al., 2005], επεκτείνεται ο αλγόριθμος *Recursive Shortest Spanning Tree* (RSST) με ένα καινούριο μοντέλο χρώματος και *συντακτικά χαρακτηριστικά* [Bennstrom & Casas που χρησιμοποιούνται ως κριτήριο για την ένωση (merge) των περιοχών. Επίσης εισάγονται τρόποι ανάλυσης δομής της εικόνας με προσαρμογή της χωρικής πληροφορίας και του σχήματος των περιοχών. Παραδείγματα κατάτμησης εικόνων με την συγκεκριμένη μέθοδο φαίνονται στο σχήμα 3.2. Βλέπουμε ότι η συγκεκριμένη μέθοδος κατάτμησης οδηγεί συνήθως σε αρκετά μεγάλες περιοχές και αλλά πολλές φορές δημιουργούνται και μικρές περιοχές που ουσιαστικά αποτελούν θόρυβο για την ανάκτηση. Προσπαθώντας να μειωθεί το πλήθος αυτών των περιοχών, τέθηκε μέγιστο όριο κατάτμησης στις 8 περιοχές ανά εικόνα.

3.3 Εξαγωγή οπτικών χαρακτηριστικών

Για να εξάγουμε τα οπτικά χαρακτηριστικά των περιοχών των εικόνων χρησιμοποιούνται οι *περιγραφείς χρώματος και υφής* του *MPEG-7*, οι οποίοι περιγράφονται εκτενώς στην ενότητα 2.2. Όπως και στην περίπτωση εξαγωγής οπτικών χαρακτηριστικών από όλη την εικόνα, έτσι και εδώ στις περιοχές της, για την εξαγωγή των περιγραφέων χρησιμοποιείται το πρόγραμμα VDE[1] ανά

[1] `http://image.ntua.gr/smag/tools/vde`

περιοχή [Tolias, 2007]. Τελικά οι 5 περιγραφείς που εξάγονται, ενώνονται σε ένα *διάνυσμα χαρακτηριστικών* (*feature vector*) της περιοχής, που έχει την μορφή της σχέσης (3.1).

$$f_i = f(r_i) = [SCD(r_i), CSD(r_i), EHD(r_i), HTD(r_i), CLD(r_i)], \forall r_i \in I \tag{3.1}$$

όπου f_i το διάνυσμα χαρακτηριστικών που αντιστοιχεί στην περιοχή r_i της εικόνας I, $CLD(r_i)$ ο περιγραφέας διάταξης χρώματος για την περιοχή r_i, $DCD(r_i)$ ο περιγραφέας κύριων χρωμάτων για την περιοχή r_i κ.ο.κ.

Το διάνυσμα αυτό έχει μήκος 666 στοιχείων και αποτελεί την αναπαράσταση της περιοχής r_i στο σύστημα. Τα διανύσματα από όλες τις περιοχές όλων των εικόνων της συλλογής αποθηκεύονται προσωρινά στην βάση, ώστε να χρησιμοποιηθούν έπειτα για την δημιουργία της τελικής δεικτοδότησης των εικόνων του συστήματος.

Οι περιγραφείς που εξάγονται κατά αυτόν τον τρόπο μπορούν να περιγράψουν αποτελεσματικά τις περιοχές από τις οποίες εξάγονται, όταν αυτές έχουν αραιό οπτικό - σημασιολογικό περιεχόμενο. Αυτό αναμένεται να ισχύει για τις περιοχές των εικόνων της συλλογής, μιας και είναι περιοχές από κατάτμηση οι οποίες ιδανικά περιέχουν μια έννοια υψηλού επιπέδου της εικόνας. Πρακτικά, σε κάθε περιοχή δεν υπάρχει μονάχα μία έννοια, αλλά μπορούμε να θεωρήσουμε ότι υπάρχει μια κύρια έννοια η οποία κατέχει ένα σημαντικό ποσοστό της περιοχής, οπότε και πάλι η περιγραφή της συγκεκριμένης έννοιας από τους MPEG-7 περιγραφείς στην περιοχή αυτή, αναμένεται να είναι αποδοτικότερη από την εξαγωγή των περιγραφέων από ολόκληρη την εικόνα.

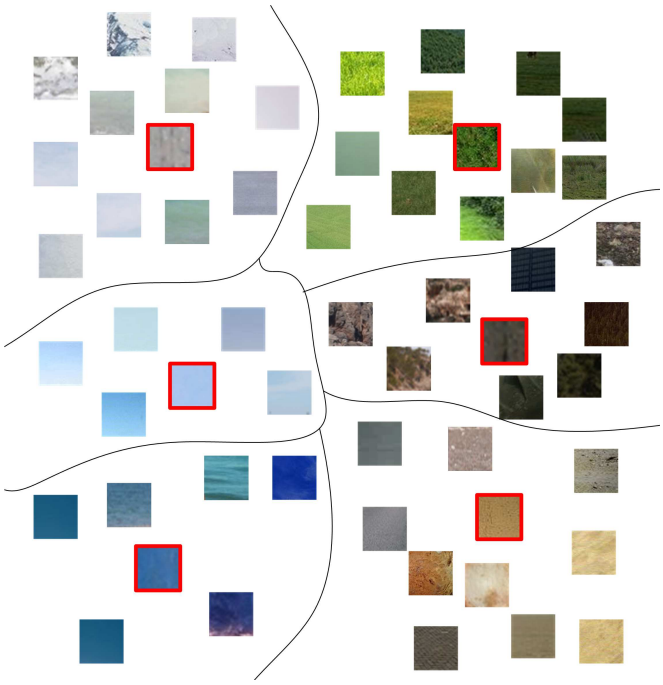
3.4 Συσταδοποίηση χαρακτηριστικών και δημιουργία οπτικού θησαυρού τύπων περιοχών

Στις περιοχές των εικόνων της εκάστοτε συλλογής περιέχονται έννοιες υψηλού επιπέδου οι οποίες υπάρχουν σε πολλές άλλες από τις εικόνες. Αυτές τις έννοιες είναι επιθυμητό το σύστημα να μπορεί να ξεχωρίζει και να εντοπίσει αποτελεσματικά, ώστε αν η εικόνα του ερωτήματος από τον χρήστη τις περιέχει, να περιέχονται και στις ανακτώμενες - παρόμοιες εικόνες.

Η ομαδοποίηση όλων αυτών των περιοχών από όλες τις εικόνες της συλλογής, των περιοχών δηλαδή των οποίων τα διανύσματα χαρακτηριστικών μοιάζουν (άρα εμπεριέχουν θεωρητικά και τις ίδιες έννοιες υψηλού επιπέδου) είναι λοιπόν μια επιβεβλημένη διαδικασία. Μια τέτοια ομαδοποίηση των περιοχών θα έδινε κάποιες συνήθεις «πατέντες» περιοχών, οι οποίες επαναλαμβάνονται μέσα στις εικόνες της συλλογής. Μπορούμε να πούμε ότι αυτά τα μοτίβα ορίζουν *τύπους περιοχών* (*region types*, σχήμα 3.3).

Εμφανίζεται λοιπόν το πρόβλημα της *ομαδοποίησης ή συσταδοποίησης* (*clustering*) των διανυσμάτων χαρακτηριστικών, που είναι ουσιαστικά πολυδιάστατα διανύσματα δεδομένων. Για την συσταδοποίηση δεδομένων υπάρχουν διαθέσιμες πολλές τεχνικές, στατιστικές και μη [Bishop, 2006]. Στην προκειμένη περίπτωση, λόγω και τις μεγάλης διάστασης των δεδομένων, προτιμάται ο ιδιαίτερα διαδεδομένος αλγόριθμος συσταδοποίησης *k-means*, που αναπτύσσεται στην υποενότητα 3.4.1.

Όταν προσδιοριστούν οι συστάδες αυτές, θα μπορεί κάθε διάνυσμα αναπαράστασης περιοχής να μπορεί να αντιστοιχηθεί στην κοντινότερη – με κάποιο μέτρο απόστασης – σε αυτό συστάδα.



Σχήμα 3.3: Ομαδοποίηση περιοχών σύμφωνα με τα όρια των τύπων περιοχών.

3.4.1 ο αλγόριθμος k-means

Η κατηγοριοποίηση δεδομένων σε k ομάδες μέσω υπολογιστή ονομάζεται συχνά και k-συσταδοποίηση (k-*clustering*) και σε αυτή την κατηγορία αλγορίθμων ανήκει και ο αλγόριθμος k-means που χρησιμοποιήθηκε. Ο αριθμός των συστάδων - ομάδων k είναι γνωστός εκ των προτέρων και το πρόβλημα είναι να κατηγοριοποιηθούν τα πολυδιάστατα διανύσματα χαρακτηριστικών σε αυτές τις k συστάδες.

Ο αλγόριθμος k-*means* υπολογίζει και αναθέτει κάθε δεδομένο στο κοντινότερό του κέντρο συστάδας και πρωτοεμφανίστηκε το 1967 [MacQueen,]. Το *κέντρο συστάδας* (*centroid*) ορίζεται ως ο μέσος όρος όλων των σημείων που ανήκουν στη συστάδα, δηλαδή ο αριθμητικός μέσος αυτών των σημείων σε κάθε διάσταση των δεδομένων προς συσταδοποίηση.

Ο αλγόριθμος αποτελείται από τα παρακάτω βήματα:

- Επιλογή του αριθμού των συστάδων, k.
- Τυχαία επιλογή k σημείων ως αρχικά κέντρα των συστάδων.
- Αντιστοίχηση κάθε σημείου στο κοντινότερο - σύμφωνα με κάποια απόσταση - κέντρο συστάδας.
- Επαναϋπολογισμός των κέντρων των συστάδων ως μέσο όρο των σημείων που αντιστοιχήθηκε προηγουμένως στην κάθε συστάδα.
- Επανάληψη των δύο παραπάνω βημάτων μέχρι να ικανοποιηθεί κάποιο κριτήριο τερματισμού.

Κριτήριο τερματισμού μπορεί να είναι η ελαχιστοποίηση κάποιου συναρτησιακού, όπως για παράδειγμα η μέγιστη απόσταση από το κέντρο της συστάδας για κάθε σημείο, το άθροισμα των μέσων αποστάσεων σε όλες τις συστάδες, το άθροισμα των τυπικών αποκλίσεων των συστάδων ή η συνολική απόσταση μεταξύ των σημείων μιας συστάδας και του κέντρου. Επίσης όταν σε κάποια από τις επαναλήψεις δεν υπάρχει πια καμία μετατόπιση σημείου, άρα και κέντρου συστάδας, τότε ο αλγόριθμος τερματίζεται.

Ως μέτρο απόστασης χρησιμοποιείται συνήθως η Ευκλείδεια απόσταση μεταξύ των N-διάστατων διανυσμάτων των σημείων. Για το σύστημά μας, μιας και τα δεδομένα προς συσταδοποίηση είναι διανύσματα χαρακτηριστικών με την γνωστή μορφή της εξίσωσης (3.1), η Ευκλείδεια απόσταση τροποποιήθηκε ώστε να υπολογίζεται σε κάθε περιγραφέα ξεχωριστά, σύμφωνα με την σχέση:

$$\begin{aligned} D(FV_1, FV_2) &= D(SCD_1, SCD_2) + D(CSD_1, CSD_2) + \\ &+ D(HTD_1, HTD_2) + D(EHD_1, EHD_2) + D(CLD_1, CLD_2) = \\ &= \sum_{i=1}^{256} fv_{1i} - fv_{2i} + \sum_{i=257}^{512} fv_{1i} - fv_{2i} + \\ &+ \sum_{i=513}^{575} fv_{1i} - fv_{2i} + \sum_{i=576}^{654} fv_{1i} - fv_{2i} + \sum_{i=655}^{666} fv_{1i} - fv_{2i} \end{aligned} \tag{3.2}$$

Το κριτήριο τερματισμού και η επιλογή του μέτρου απόστασης είναι οι παράγοντες που καθορίζουν το *σχήμα της συστάδας*. Επίσης είναι σημαντικό να τονιστεί ότι το αποτέλεσμα του αλγορίθμου επηρεάζεται από την τυχαία αρχική ανάθεση των κέντρων των συστάδων και για μικρό αριθμό συστάδων ενδέχεται να μην συγκλίνει πάντα στα ίδια κέντρα. Δυστυχώς, ο απλός αλγόριθμος k-means είναι αρκετά αργός για μεγάλο αριθμό συστάδων και πολυδιάστατα δεδομένα.

Ένα παράδειγμα συσταδοποίησης με τον αλγόριθμο k-means παρουσιάζεται στο σχήμα 3.5, διαγράμματα τα οποία δημιουργήθηκαν από την υλοποίηση [Pelleg & Moore, 1999]. Εδώ η διάσταση των δεδομένων είναι 2 και ο επιθυμητός αριθμός συστάδων 5. Τα αρχικά δεδομένα φαίνονται στο σχήμα 3.5(a), ενώ στο σχήμα 3.5(b) με κόκκινο σημειώνονται τα κέντρα των συστάδων τα οποία αρχικοποιήθηκαν τυχαία. Στο σχήμα 3.5(c) σημειώνονται τα όρια των συστάδων με αυτά τα τυχαία αρχικά κέντρα, γνωστά και ως *κελιά Βορονόι* (*Voronoi cells*). Τα σημεία που ανήκουν σε κάθε μια από τις συστάδες χρωματίζονται κατάλληλα (σχήμα 3.5(d)). Έπειτα τα κέντρα των συστάδων μετακινούνται στον μέσο όρο των σημείων που ανήκουν στην κάθε συστάδα (σχήμα 3.5(e)) και η διαδικασία επαναλαμβάνεται μέχρις ότου να μην έχουμε μετακινήσεις σημείων από την μία συστάδα σε άλλη. Η τελική συσταδοποίηση φαίνεται στο σχήμα 3.5(j).

Επεκτάσεις και βελτιώσεις του αλγορίθμου k-means, όπως είναι λογικό καθώς έχουν περάσει πάνω από 40 χρόνια από την εμφάνιση του, έχουν προταθεί πολλές [Philbin et al., 2007], [Kanungo et al., 2002], [Fang, 1996] και [Frossyniotis et al., 2004].

3.4.2 Δημιουργία Οπτικού θησαυρού

Με εφαρμογή του αλγόριθμου k-means στα διανύσματα χαρακτηριστικών όλων των περιοχών όλων των εικόνων της συλλογής, προκύπτουν κάποια κέντρα ομάδων, τα οποία αποτελούν μέσους

(a) Τύποι περιοχών από τον οπτικό θησαυρό, οι οποίοι περιλαμβάνουν έννοιες υψηλού επιπέδου.

(b) Τύποι περιοχών από τον οπτικό θησαυρό, οι οποίοι δεν περιλαμβάνουν έννοιες υψηλού επιπέδου.

Σχήμα 3.4: «Καλοί» και «κακοί» τύποι περιοχών από τον οπτικό θησαυρό.

όρους από διανύσματα χαρακτηριστικών εξαγμένα από παρόμοιες περιοχές. Το διάνυσμα του κάθε κέντρου, αποτελεί ένα διάνυσμα το οποίο προσεγγίζουν κάποιες περιοχές από εικόνες της συλλογής. Μπορούμε να πούμε ότι το κάθε κέντρο εκφράζει έναν *τύπο περιοχής* (*region type*), που μπορεί να είναι είτε υποσύνολο είτε υπερσύνολο μιας έννοιας υψηλού επιπέδου. Η έννοια ουρανός για παράδειγμα, μπορεί να είναι σε πολλούς διαφορετικούς (χρωματικά και ως προς την υφή) τύπους περιοχών, μιας και ο καθαρός ουρανός δεν είναι ίδιος με τον συννεφιασμένο, τον νυχτερινό ή του ηλιοβασιλέματος. Παρόμοια, σε έναν τύπο περιοχής είναι πιθανό να συνυπάρχουν παραπάνω από μια έννοιες υψηλού επιπέδου, κυρίως στην περίπτωση που δεν είναι διαχωρισμένες μετά την κατάτμηση. Παράδειγμα εννοιών που μπορεί να συνυπάρχουν σε μια συστάδα είναι ο ουρανός με την θάλασσα.

Για να μπορέσουμε να οπτικοποιήσουμε τους τύπους περιοχών, εκτός από το κέντρο της συστάδας ανατέθηκε στον κάθε έναν και το κοντινότερο σε αυτόν διάνυσμα χαρακτηριστικών, δηλαδή το διάνυσμα χαρακτηριστικών μιας συγκεκριμένης περιοχής μιας από τις εικόνες της συλλογής, το οποίο έχει την μικρότερη απόσταση από όλα στο κάθε κέντρο συστάδας. Παραδείγματα τέτοιων περιοχών, που χαρακτηρίζουν τις συστάδες, άρα και τύπους περιοχών φαίνονται στο σχήμα 3.4(a).

Δυστυχώς, μιας και η συσταδοποίηση γίνεται πάνω σε όλες τις περιοχές που εξάγει η κατάτμηση και όχι μόνο σε εκείνες που περιέχουν ωφέλιμες έννοιες, σχηματίζονται και αρκετοί τύποι περιοχών οι οποίοι δεν περιέχουν ουσιαστικά κάποια έννοια υψηλού επιπέδου και λειτουργούν αρνητικά στην διαδικασία της ανάκτησης. Παραδείγματα τέτοιων τύπων περιοχών φαίνονται στο σχήμα 3.4(b). Η ύπαρξη τέτοιων συστάδων είναι αναπόφευκτη σε μια διαδικασία μάθησης χωρίς επίβλεψη (*unsupervised learning*) όπως είναι η διαδικασία που εκτελείται για την δημιουργία του οπτικού θησαυρού.

Αξίζει να σημειωθεί ότι το σχήμα των περιοχών που φαίνονται στα σχήματα 3.4(a) και 3.4(b),

παρότι σε αρκετές περιπτώσεις είναι χαρακτηριστικό της έννοιας που περιγράφεται δεν κρατιέται και δεν περιλαμβάνεται στην πληροφορία του οπτικού θησαυρού.

3.5 Αναπαράσταση και δεικτοδότηση των εικόνων

Αφότου σχηματιστεί ο οπτικός θησαυρός περιοχών, κάθε εικόνα της συλλογής μπορεί να αναπαρασταθεί μέσω του θησαυρού αυτού, με ένα *διάνυσμα αναπαράστασης* (*Model Vector*). Το διάνυσμα αυτό για μια εικόνα, θα φανερώνει τους κοντινότερους σε αυτήν τύπους περιοχής.

Το διάνυσμα αναπαράστασης θα αποτελείται από έναν αριθμό τύπων περιοχής και μια τιμή για κάθε έναν από αυτούς, η οποία φανερώνει το πόσο μοιάζει ο καθένας με κάποια από τις περιοχές της εικόνας. Για να εξαχθεί το διάνυσμα αυτό, εξάγονται πρώτα τα διανύσματα χαρακτηριστικών από κάθε περιοχή της εικόνας, τα οποία έχουν την μορφή της σχέσης (3.1). Έπειτα υπολογίζεται η απόσταση του κάθε διανύσματος χαρακτηριστικών από όλους τους τύπους περιοχής. Όπως αναφέρθηκε στην υποενότητα 3.4.1, τα κέντρα των συστάδων που εξάγονται από την διαδικασία συσταδοποίησης με τον αλγόριθμο k-means είναι ίδιας διάστασης με τα δεδομένα εισόδου, άρα έχουν και αυτά την διάσταση των διανυσμάτων χαρακτηριστικών. Μπορούμε λοιπόν χρησιμοποιώντας την σχέση (3.2) να υπολογίσουμε την απόσταση της κάθε περιοχής της εικόνας από τον κάθε τύπο περιοχής, βρίσκοντας την απόσταση των μεταξύ τους διανυσμάτων χαρακτηριστικών. Για κάθε περιοχή της εικόνας λοιπόν υπολογίζονται οι αποστάσεις μεταξύ αυτής κάθε τύπου περιοχής και εισάγονται στο διάνυσμα αναπαράστασης οι ν τύποι με τις κοντινότερες αποστάσεις, κάθε ένας σαν ζευγάρι μαζί με μια τιμή αντιστρόφως ανάλογη της απόστασης του από την περιοχή της εικόνας. Αν ένας από τους τύπους περιοχής βρίσκεται στους ν κοντινότερους για πάνω από μια από τις περιοχές της εικόνας τότε φυσικά εισάγεται μία μονάχα φορά, με τιμή ομοιότητας την μέγιστη από τις τιμές. Αν MV_i είναι το διάνυσμα της εικόνας i τότε θα έχει μορφή:

$$MV_i = \Big[RT_i(1), SV_i(1), RT_i(2), SV_i(2), \ldots, RT_i(j), SV_i(j), \ldots, RT_i(N_T), SV_i(N_T)\Big], \quad i = 1 \ldots N_{images} \tag{3.3}$$

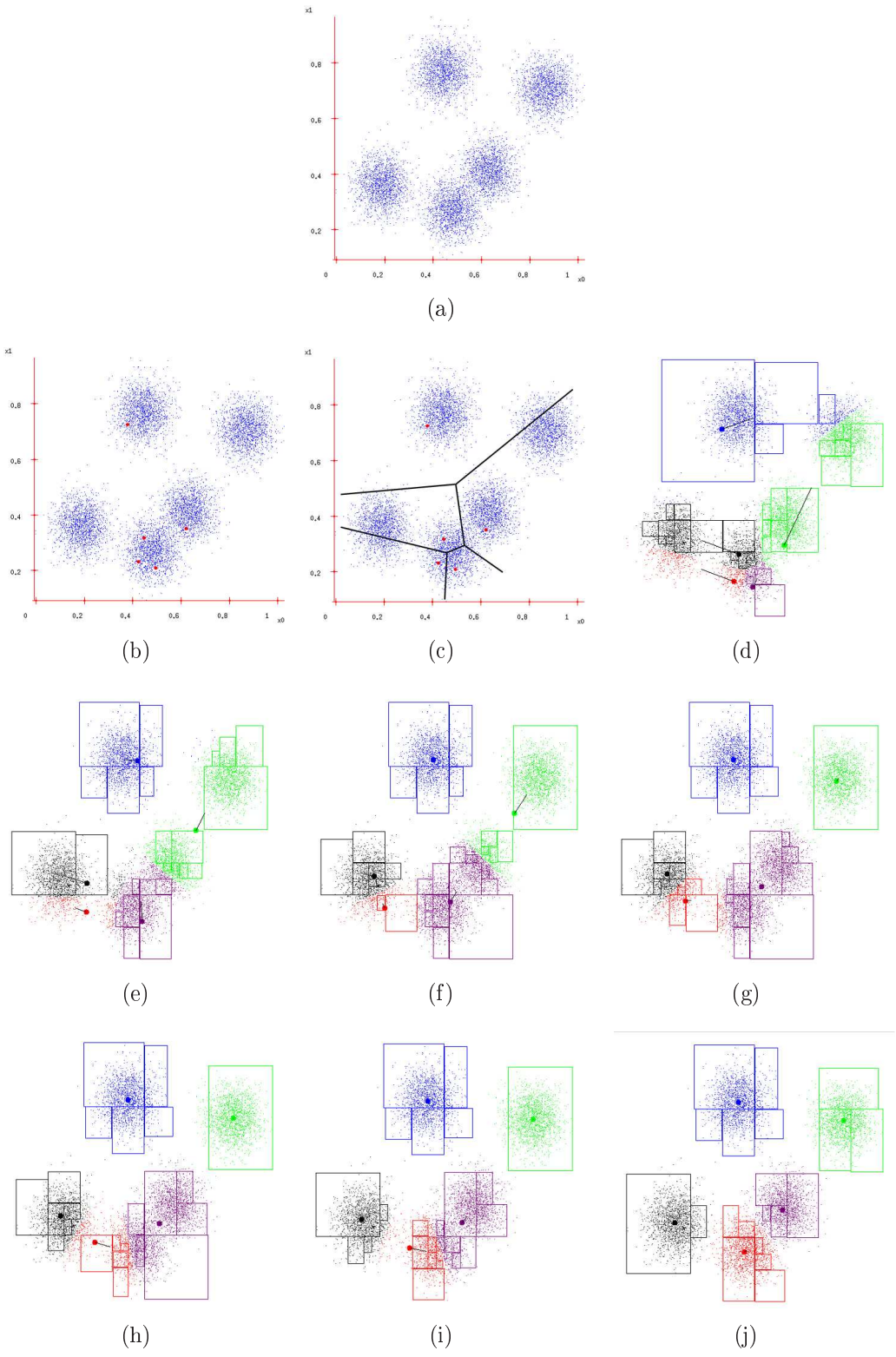
όπου N_{images} είναι ο αριθμός των εικόνων της συλλογής , $RT_i(j)$ ένας τύπος περιοχής «κοντινός» στην εικόνα i και $SV_i(j)$ είναι η τιμή ομοιότητας (*Similarity Value*) του, που δίνεται από τον τύπο:

$$SV_i(j) = \frac{\alpha}{d_{ij}} \tag{3.4}$$

Όπου d_{ij} είναι η ελάχιστη απόσταση του τύπου περιοχής $RT_i(j)$ από κάποια από τις περιοχές της εικόνας i. Το α εισάγεται για την κανονικοποίηση των τιμών ομοιότητας. Το d_{ij} θα δίνεται από την σχέση:

$$d_i(j) = \min_{r \in R_i)} \Big\{ D\Big(fv(rt_j), fv(r)\Big) \Big\}, \quad i = 1 \ldots N_{images}, \quad j = 1 \ldots N_T \tag{3.5}$$

όπου Reg_i είναι το σύνολο των περιοχών μετά την κατάτμηση της εικόνας i και $D\Big(fv(rt_k), fv(r)\Big)$ είναι η απόσταση μεταξύ του διανύσματος χαρακτηριστικών του τύπου περιοχής k (με $rt_k = RT_i(j)$) και του διανύσματος χαρακτηριστικών της περιοχής r της εικόνας i, που συμβολίζονται αντίστοιχα με $(fv(rt_k))$ και $(fv(r))$.



Σχήμα 3.5: Συσταδοποίηση δεδομένων σε $k = 5$ συστάδες με τον αλγόριθμο k-means (από το [Moore, b]).

Σχήμα 3.6: Οι 2 ($\nu = 2$) κοντινότεροι τύποι περιοχής για κάθε μία από τις περιοχές της εικόνας στα αριστερά.

Εύκολα παρατηρεί κανείς ότι το μήκος του διανύσματος αναπαράστασης που ισούται με $2 \times N_T$ είναι μεταβλητό. Αυτό συμβαίνει γιατί ο αριθμός των τύπων περιοχής που είναι κοντά στις περιοχές τις εικόνας εξαρτάται από τις περιοχές της κάθε εικόνας. Αν δεν εμφανιστεί κάποιος τύπος περιοχής ως ένας από τους ν κοντινότερους σε πάνω από μία περιοχή μιας εικόνας i τότε το N_T θα ισούται με

$$N_T = \nu * Reg_i \tag{3.6}$$

όπου ν είναι ο αριθμός των κοντινότερων τύπων περιοχής που κρατείται και Reg_i είναι το σύνολο των περιοχών μετά την κατάτμηση της εικόνας i. Στο σχήμα 3.6 φαίνονται οι 2 ($\nu = 2$) κοντινότεροι τύποι περιοχής για τις περιοχές μιας εικόνας.

Επίσης στον πίνακα 3.1 αποτυπώνεται σχηματικά ένα διάνυσμα αναπαράστασης:

ID εικόνας	Τύποι Περιοχής	Τιμές ομοιότητας
12345	12 43 66 1 32 4	0.155 0.03 0.4 0.2 0.17 0.045

Πίνακας 3.1: Τα πεδία του διανύσματος αναπαράστασης των εικόνων. Στο πεδίο Τύποι Περιοχής αποθηκεύεται μια λίστα από τους κοντινότερους στην εικόνα τύπους περιοχής και στο πεδίο Τιμές Ομοιότητας η λίστα με τις αντίστοιχες τιμές ομοιότητας.

Ο λόγος για τον οποίον επιλέχτηκε το να αποθηκεύονται οι ν κοντινότεροι σε κάθε περιοχή τύποι περιοχής είναι ότι παρατηρήθηκε ότι πολλές έννοιες υψηλού επιπέδου εκφράζονταν σε παραπάνω από έναν τύπους περιοχής. Έτσι αν για παράδειγμα δύο τύποι περιοχής εκφράζουν την άμμο μιας παραλίας και κρατιόταν μονάχα ο κοντινότερος τύπος περιοχής στο διάνυσμα αναπαράστασης, τότε οι περιοχές δύο εικόνων που εκφράζονται η μια από τον πρώτο και η άλλη από τον δεύτερο τύπο, λανθασμένα δεν θα είχαν καμία ομοιότητα. Μειώνεται με αυτόν τον τρόπο λοιπόν το σφάλμα της κβαντοποίησης.

Σημαντικό είναι να τονιστεί το ότι για την εξαγωγή των αποστάσεων μεταξύ δυο διανυσμάτων χαρακτηριστικών, στη σχέση (3.2) εισάγονται και βάρη κανονικοποίησης σε κάθε περιγραφέα και η

τελική απόσταση βγαίνει ως ο μέσος όρος με βάρη των επιμέρους αποστάσεων. Το βάρος για κάθε περιγραφέα ισούται με την μέγιστη δυνατή απόσταση μεταξύ δύο ίδιων περιγραφέων ανάμεσα σε όλες τις περιοχές των εικόνων της συλλογής.

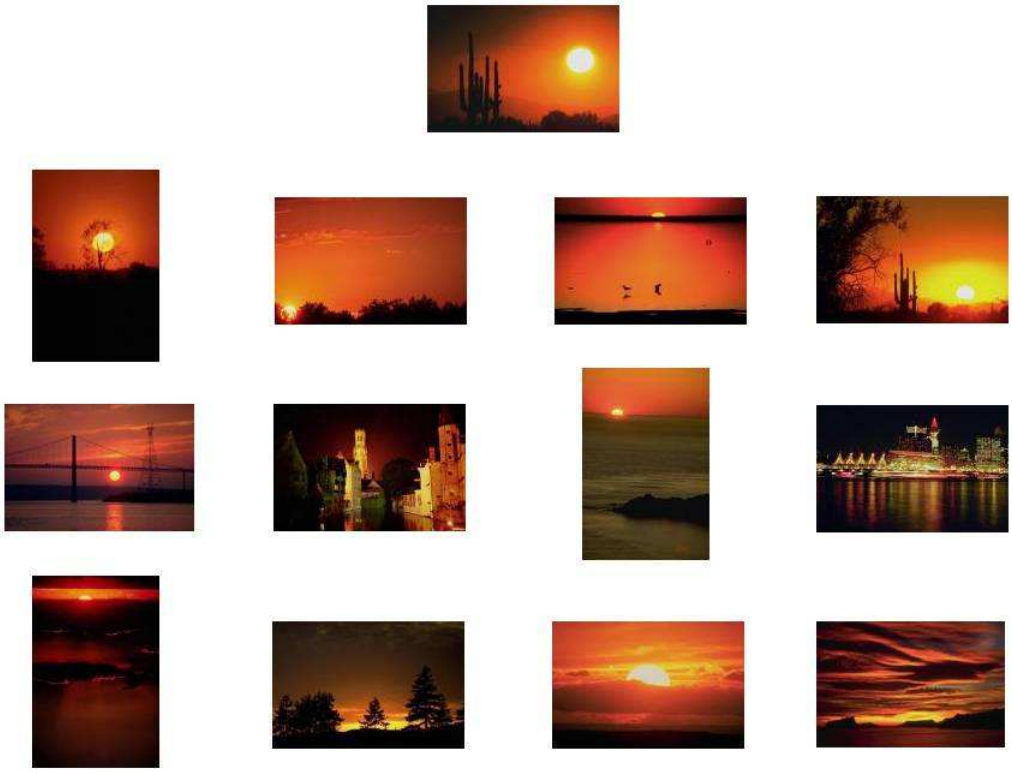
Στην βάση δεδομένων αποθηκεύεται το διάνυσμα αναπαράστασης για κάθε εικόνα και πλέον κάθε εικόνα αναπαρίσταται στη βάση με ένα $2 \times N_T$-διάστατο διάνυσμα τύπων περιοχής και τιμών ομοιότητας. Για μελλοντική προσθήκη νέων εικόνων ή για την εξαγωγή του διανύσματος αναπαράστασης από εικόνες που ανεβάζουν οι χρήστες, αποθηκεύονται επίσης στην βάση τα κέντρα των τύπων περιοχών.

3.6 Ταίριασμα των εικόνων

Για να υπολογιστεί η απόσταση μεταξύ δύο εικόνων αρκεί να υπολογιστεί η Ευκλείδεια απόσταση μεταξύ των διανυσμάτων αναπαράστασης των δύο αυτών εικόνων.

Όταν εισάγεται στο σύστημα ένα νέο ερώτημα ανάκτησης, τότε, αν η εικόνα είναι εικόνα της συλλογής, φορτώνεται από την βάση το διάνυσμα αναπαράστασης και υπολογίζεται η απόσταση του από όλα τα αντίστοιχα διανύσματα όλων των εικόνων της συλλογής. Οι αποστάσεις ταξινομούνται και στον χρήστη του συστήματος επιστρέφονται ως παρόμοιες οι εικόνες που έχουν την μικρότερη απόσταση.

Μερικά παραδείγματα αναζήτησης φαίνονται στα σχήματα 3.7 και 3.8. Στο πάνω μέρος είναι η εικόνα του ερωτήματος και από κάτω ταξινομημένες από τα αριστερά προς τα δεξιά και από πάνω προς τα κάτω οι κοντινότερες 12 εικόνες. Στο σχήμα 3.8 οι σωστές ανακτήσεις της έβδομης και όγδοης εικόνας δεν θα επιστρέφονταν με την αναζήτηση βασισμένη σε χαρακτηριστικά ολόκληρης της εικόνας. Η ανάκτηση μετά από κατάτμηση σε περιοχές είναι πιό ευαίσθητη σε εμπεριεχόμενες έννοιες υψηλού επιπέδου.



Σχήμα 3.7: Αποτελέσματα αναζήτησης με χαρακτηριστικά εξαγόμενα από περιοχές σε συλλογή με φυσικές εικόνες από το Corel. Ηλιοβασίλεμα.

Σχήμα 3.8: Αποτελέσματα αναζήτησης με χαρακτηριστικά εξαγόμενα από περιοχές σε συλλογή με φυσικές εικόνες από το Corel. Χιόνι.

Κεφάλαιο 4

Αναζήτηση με βάση τοπικά χαρακτηριστικά

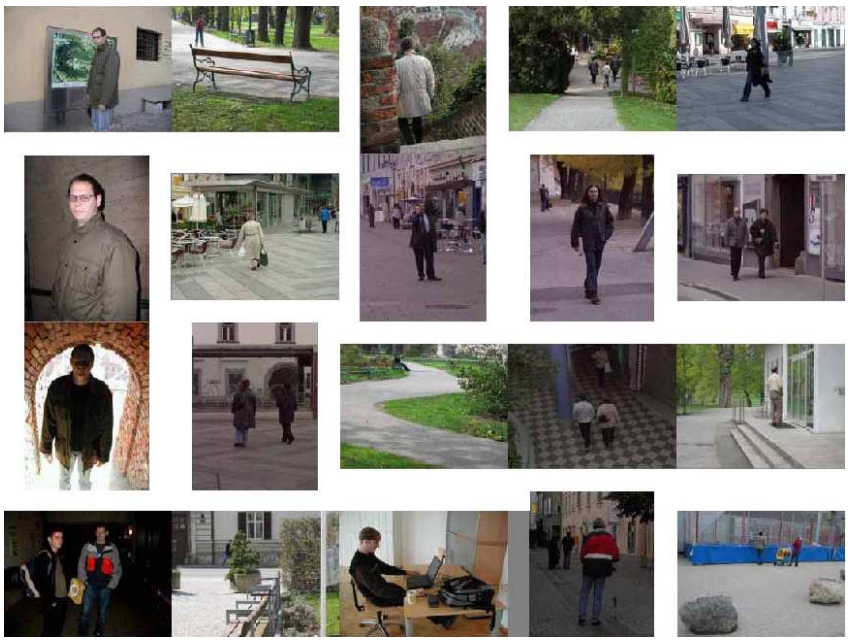
4.1 Εισαγωγή

Καθώς οι πολυμεσικές συλλογές γιγαντώνονται συνεχώς και το Περιεχόμενο τους εκτείνεται σχεδόν σε κάθε έκφανση της ζωής, η δυνατότητα ανάκτησης εικόνων με το ίδιο σημασιολογικό περιεχόμενο εύκολα, γρήγορα και με ακρίβεια θα ήταν ιδιαίτερα χρήσιμη. Μια τέτοια αναζήτηση έρχεται σε πλήρη αντιστοιχία με την αναζήτηση σελίδων μέσω μιας μηχανής αναζήτησης ιστού, όπως για παράδειγμα το Google.

Εύκολα καταλαβαίνει κανείς ότι περιγραφείς εξαγόμενοι από ολόκληρη την εικόνα δεν μπορούν σε τόσο μεγάλη κλίμακα να αναπαραστήσουν τον ποικιλόμορφο σημασιολογικό διάκοσμο που περιέχουν οι συλλογές. Το ίδιο ισχύει και για περιγραφείς από περιοχές κατάτμησης καθώς οι «τύποι περιοχών» δεν μπορούν παρά να χαρακτηρίσουν συγκεκριμένες περιοχές εικόνων. Με όλους τους προαναφερθέντες περιγραφείς, η ανάκτηση της έννοιας «άνθρωπος» από εικόνες μιας συλλογής όπως του σχήματος 4.1 είναι πρακτικά αδύνατη.

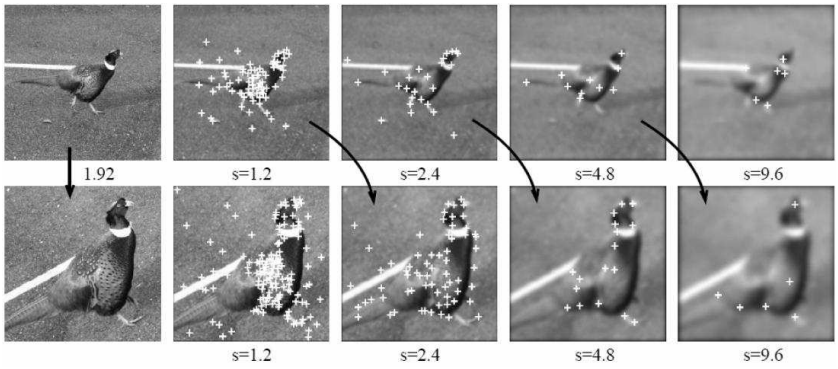
Για να μπορέσουν να αναπαρασταθούν σωστά οι παραπάνω εικόνες, ώστε Να μπορεί έπειτα εξαχθεί και να απομονωθεί η πληροφορία της ύπαρξης ανθρώπου, θα πρέπει να χρησιμοποιηθεί ένα μικρότερο δομικό στοιχείο περιγραφής από ότι ολόκληρη η εικόνα ή μεγάλες περιοχές της. Γι' αυτό το λόγο, τα τελευταία χρόνια χρησιμοποιούνται κατά κόρων περιγραφείς εξαγόμενοι από περιοχές γύρω από σημεία ενδιαφέροντος. Τα σημεία ενδιαφέροντος μιας εικόνας μπορεί να είναι ακμές ή γωνίες της εικόνας αλλά και σημεία έντονης υφής ή αλλαγής της υφής. Τα σημεία αυτά είναι εύρωστα σε μεταβολές κλίμακας όταν αυτά εξάγονται από πολλαπλές κλίμακες καθώς και σε περιστροφές (σχήματα 4.2 και 4.3). Η περιοχή γύρω από τα σημεία, από την οποία εξάγονται οπτικά χαρακτηριστικά, έχει μέγεθος ανάλογο με την κλίμακα στην οποία εντοπίζεται το σημείο.

Από κάθε μια από τις εικόνες της συλλογής εξάγονται οι περιγραφείς γύρω από τα σημεία ενδιαφέροντος και από αυτά δημιουργείται με συσταδοποίηση ένα *οπτικό λεξικό* σύμφωνα με τις «λέξεις» του οποίου θα αναπαρασταθούν τα σημεία όλων των εικόνων. Το λεξικό δημιουργείται γιατί η απευθείας εξαντλητική σύγκριση σημείο με σημείο ανάμεσα σε δύο εικόνες είναι μια ιδιαίτερα χρονοβόρα διαδικασία. Με την χρήση του οπτικού λεξικού η σύγκριση δύο εικόνων ανάγεται στην σύγκριση



Σχήμα 4.1: Ποικιλόμορφες εικόνες στις οποίες περιέχεται η έννοια «άνθρωπος».

Σχήμα 4.2: Σημεία ενδιαφέροντος σε εικόνες του ίδιου κτιρίου από διαφορετική οπτική γωνία.



Σχήμα 4.3: Σημεία ενδιαφέροντος εξαγόμενα από διάφορες κλίμακες με την μέθοδο Harris - Laplacian από δύο εικόνες του ίδιου αντικειμένου σε διαφορετική κλίμακα και οι αντιστοιχίες τους.

των διανυσμάτων αναπαράστασής τους.

4.2 Τα σημεία ενδιαφέροντος και οι περιγραφείς SURF

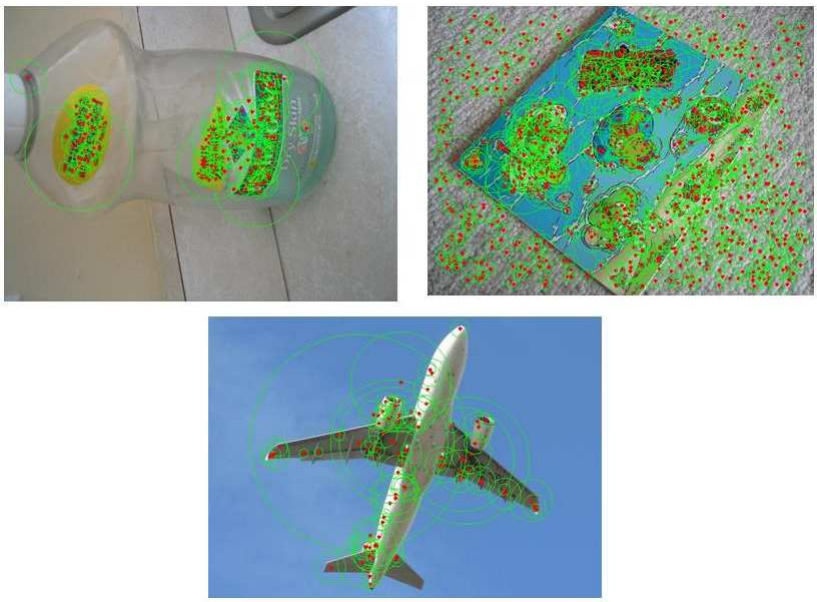
Για να μην υπάρχει εξάρτηση από την κλίμακα (μέγεθος), η διαδικασία ανίχνευσης σημείων ενδιαφέροντος εκτελείται σε πολλαπλές κλίμακες. Δημιουργείται για αυτό τον σκοπό ένας χώρος κλίμακας από διαδοχικές συνελίξεις της εικόνας με γκαουσιανούς πυρήνες αυξανόμενης τυπικής απόκλισης σ και σε κάθε επίπεδό της εκτελείται η εξαγωγή σημείων ενδιαφέροντος. Οι διαδοχικές συνελίξεις ουσιαστικά εξομαλύνουν την εικόνα όλο και πιο πολύ με αποτέλεσμα στις κλίμακες προς την κορυφή να έχουμε μια αρκετά απλοποιημένη εκδοχή της αρχικής εικόνας.

Πιο πρόσφατες τεχνικές προτείνουν τη δημιουργία χώρου κλίμακας με μορφολογικά φίλτρα αντί για συνέλιξη με γκουσιανές συναρτήσεις 4.3. Αυτό μπορεί να οδηγήσει σε σημεία προσδιορισμένα με μεγαλύτερη ακρίβεια, καθώς τα μορφολογικά φίλτρα διατηρούν σχεδόν αναλλοίωτες τις ακμές τις εικόνας ακόμα και σε υψηλά επίπεδα φιλτραρίσματος. Η εφαρμογή μορφολογικών χώρων κλίμακας στην εξαγωγή σημείων ενδιαφέροντος έγινε πολύ πρόσφατα με ενδιαφέροντα αποτελέσματα.

Την περιοχή γύρω από τα σημεία μπορούμε να την περιγράψουμε με διάφορους τρόπους, από πολύ απλούς που είναι οι τιμές της φωτεινότητας μέχρι ιδιαίτερα σύνθετους, όπως μια σειρά από ιστογράμματα των κατευθύνσεων των παραγώγων.

Η αποτελεσματικότητα των σημείων και των περιγραφέων μετριέται κυρίως με δύο κριτήρια: την επαναληψημότητα (repeatability) και το κατά πόσο είναι ευδιάκριτοι (distinctive) και χαρακτηριστικοί οι τοπικοί περιγραφείς. Επιθυμητό είναι επίσης να είναι τα σημεία που εξάγωνται ανεξάρτητα από αφινικές μεταβολές [Sapiro & Tannenbaum, 1993]. Όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο, αναλυτική σύγκριση και αξιολόγηση τοπικών περιγραφέων γίνεται από τους Mikolajczyk και Schmid [Mikolajczyk & Schmid, 2005].

Στην υλοποίηση μας χρησιμοποιήσαμε τα σημεία ενδιαφέροντος και τους περιγραφείς που πρότειναν οι Bay, Ess, Tuytelaars και Van Gool το 2006, τα *SURF* (*Speeded-Up Robust Features*) όπως



Σχήμα 4.4: Τρεις εικόνες και τα σημεία ενδιαφέροντός τους.

περιγράφονται στο [Bay et al., 2008]. Προτιμηθήκαν αυτοί οι περιγραφείς γιατί εκτός του ότι εξάγονται πολύ πιο γρήγορα από άλλους παρόμοιους περιγραφείς, όπως τα SIFT [Lowe, 2004], έχουν και καλύτερη απόδοση. Για την υλοποίηση χρησιμοποιήθηκε και ο κώδικας που δίνουν οι δημιουργοί για την εξαγωγή των σημείων και των τοπικών περιγραφέων.

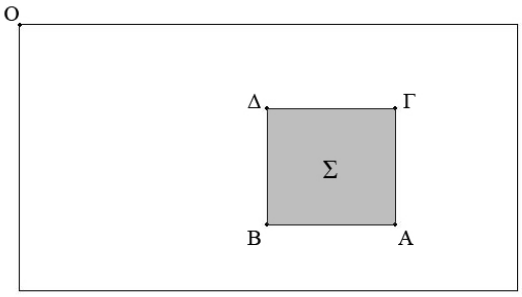
Στο σχήμα 4.4 φαίνονται τρείς εικόνες με τα σημεία ενδιαφέροντος αποτυπωμένα με κόκκινες τελείες. Η ακτίνα του πράσινου κύκλου φύρω από κάθε σημείο, είναι ανάλογη της κλίμακας στην οποία αυό βρέθηκε.

4.2.1 Προσδιορισμός των σημείων ενδιαφέροντος

Για τον προσδιορισμό των σημείων ενδιαφέροντος η παραπάνω μέθοδος προτείνει μια ιδιαίτερα γρήγορη εκτίμηση της μήτρας Hessian. Η μήτρα Hessian (Hessian matrix) είναι ο τετραγωνικός πίνακας των μερικών παραγώγων δεύτερης τάξης μιας συνάρτησης. Για να εκτιμηθεί η μήτρα, γίνεται χρήση των *integral images*, με την χρήση των οποίων μειώνεται δραματικά ο χρόνος υπολογισμού της [Viola & Jones, 2001]. Η τιμή της Integral image $I_\Sigma(x)$ στην θέση $x = (x, y)$ περιέχει το άθροισμα όλων των pixel της αντίστοιχης εικόνας σε μια περιοχή ορθογωνίου, με άνω αριστερά άκρο την αρχή $(0, 0)$ και κάτω δεξιά άκρο το σημείο x, δηλαδή

$$I_\Sigma(x) = \sum_{i=0}^{i \le x} \sum_{j=0}^{j \le y} I(i, j) \tag{4.1}$$

Μετά τον υπολογισμό της integral image, απαιτούνται μονάχα τρεις αθροίσεις για να υπολογιστεί το άθροισμα των φωτεινοτήτων σε οποιαδήποτε ορθογώνια περιοχή της εικόνας (σχήμα 4.5). Ο



Σχήμα 4.5: Με χρήση των integral images το άθροισμα των φωτεινοτήτων στην περιοχή Σ υπολογίζεται με τρεις αθροίσεις και τέσσερις προσβάσεις στην μνήμη

υπολογισμός είναι ανεξάρτητος του μεγέθους της περιοχής στην οποία ζητάμε το άθροισμα των στοιχείων, γεγονός ιδιαίτερα σημαντικό καθώς όσο ανεβαίνει η κλίμακα στον χώρο κλίμακας θα έχουμε όλο και μεγαλύτερα μεγέθη φίλτρων.

Ο ανιχνευτής σημείων ενδιαφέροντος βασίζεται στη μήτρα Hessian για να βρει θέσεις με τοπικά μέγιστα. Η επιλογή της κλίμακας καθορίζεται από την ορίζουσα της μήτρας, όπως είχε προτείνει ο Lindeberg [Lindeberg, 1998]. Σε μια θέση $x = (x, y)$ της εικόνας I, η μήτρα Hessian στο σημείο αυτό και σε κλίμακα σ είναι:

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \tag{4.2}$$

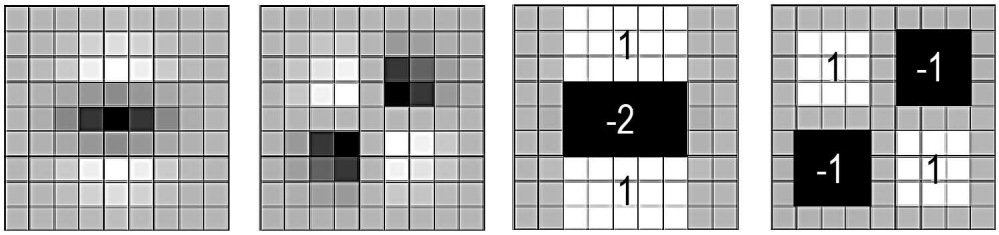
όπου $L_{xx}(x,\sigma)$ είανι η συνέλιξη της γκαουσιανής μερικής παραγώγου δευτέρας τάξης με την εικόνα I στο σημείο x. Οι γκαουσιανές θεωρούνται βέλτιστες για την ανάλυση σε κλίμακες, αλλά μιας και πρέπει να διακριτοποιηθούν και να κατωφλιωθούν, αποκτούν αδυναμία στην επαναληψημότητα τους σε περιστροφές γύρω από περιττά πολλαπλάσια του $\pi/4$. Αυτό συμβαίνει λόγω της τετραγωνικής φύσης του φίλτρου και ισχύει για όλους τους ανιχνευτές που χρησιμοποιούν την μήτρα Hessian. Για τον ίδιο λόγο, η μέγιστη επαναληψημότητα εμφανίζεται στα πολλαπλάσια των $\pi/2$ μοιρών. Για περεταίρω απλούστευση των υπολογισμών, οι μερικές παράγωγοι των γκαουσιανών προσεγγίζονται από φίλτρα-κουτιού (box filters) όπως φαίνεται στο σχήμα 4.6 χωρίς μείωση στην απόδοση.

Τα 9x9 φίλτρα του σχήματος είναι οι προσεγγίσεις της γκαουσιανής με $\sigma = 1.2$ και αποτελούν την ελάχιστη κλίμακα υπολογισμού. Αν συμβολίσουμε τα φίλτρα αυτά με D_{xx}, D_{yy} και D_{xy} τότε η ορίζουσα της μήτρας:

$$det(H_{\varepsilon\kappa\tau\iota\mu}) = D_{xx}D_{yy} - (wD_{xy})^2 \tag{4.3}$$

όπου το βάρος w χρησιμοποιείται για να διατηρηθεί η ενέργεια μεταξύ των γκαουσιανών πυρήνων και των εκτιμήσεων τους και παίρνει την τιμή 0.9. Οι αποκρίσεις των φίλτρων σε κάθε κλίμακα κανονικοποιούνται σε σχέση με την κλίμακα που βρίσκονται.

Η εκτιμώμενη τιμή της ορίζουσας της μήτρας Hessian αντιπροσωπεύει την απόκριση της εικόνας (*blob response*) στην θέση x. Οι αποκρίσεις από όλες τις κλίμακες αποθηκεύονται σε έναν χάρτη



Σχήμα 4.6: Από τα αριστερά: Οι διακριτοποιημένες γκαουσιανές μερικές παράγωγοι δευτέρας ταξης στην κατεύθυνση-y και στην κατεύθυνση-xy και οι δύο απλοποιήσεις - εκτιμήσεις τους. (από το [Bay et al., 2008]).

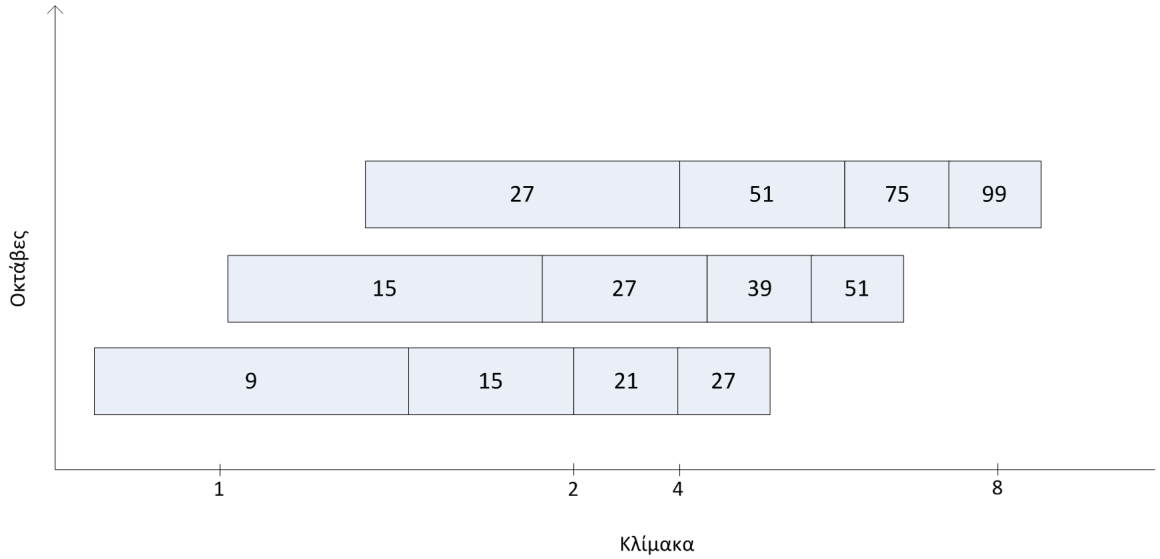
αποκρίσεων (blob response map) και από αυτόν έπειτα εξάγονται τοπικά μέγιστα (*local maxima*), τα οποία δηλώνουν σημείο ενδιαφέροντος.

Για να δημιουργηθεί ο χώρος κλίμακας απαιτείται το φιλτράρισμα της εικόνας για απλοποίηση και ταυτόχρονα η υποδειγματοληψία της, ώστε να ανεβαίνει η κλίμακα. Με αυτό τον τρόπο δημιουργούνται όλο και μικρότερες εικόνες, δομή που περιγράφεται συχνά ως πολυκλιμακωτή *πυραμίδα*. Ο Lowe αφαιρεί διαδοχικά επίπεδα της πυραμίδας για να πάρει μια εκτίμηση της διαφοράς των γκαουσιανών και να εκτιμήσει με αυτό τον τρόπο την απόκριση [Lowe, 2004].

Για τα SURF αντί να γίνεται υποδειγματοληψία της εικόνας σε κάθε επίπεδο, μεγαλώνει το μέγεθος του τετράγωνου φίλτρου με το οποίο αυτή απλοποιείται. Αυτό κάνει την εξαγωγή ταχύτερη, καθώς με τις integral images ο χρόνος εφαρμογής ενός φίλτρου στην εικόνα είναι ανεξάρτητος από το μέγεθός του φίλτρου. Έτσι η εικόνα φιλτράρεται αρχικά με τα 9x9 φίλτρα ως ελάχιστη κλίμακα και έπειτα διαδοχικά με φίλτρα μεγαλύτερου μεγέθους, αυξάνοντας κάθε φορά το παράθυρο ανά τουλάχιστον ένα pixel σε κάθε πλευρά της μάσκας που οδηγεί σε μια συνολική αύξηση 6 pixel σε κάθε επίπεδο για το φίλτρο. Για τη δεύτερη κλίμακα η εικόνα θα φιλτραριστεί με φίλτρο μεγέθους 15x15 pixel. Η διακριτή φύση των integral images επιβάλλει αυτόν τον χβαντισμό των κλιμάκων.

Ο χώρος κλίμακας διαιρείται σε *οκτάβες*. Μια οκτάβα αντιπροσωπεύει μια σειρά από φιλτραρίσματα που τελικά αντιστοιχούν σε μια αύξηση της κλίμακας περίπου κατά ένα παράγοντα 2. Σε κάθε επόμενη οκτάβα το μέγεθος παράθυρου του φίλτρου αυξάνεται με διπλάσια τιμή από ότι στην προηγούμενη οκτάβα. Αυτό σημαίνει ότι η δεύτερη οκτάβα, θα αρχίσει με φίλτρο μεγέθους 15x15 pixel (σε αυτό το μέγεθος έχει διπλασιαστεί περίπου το αρχικό μέγεθος της πρώτης οκτάβας που είναι 9x9) και το επόμενο φίλτρο θα έχει μέγεθος 27x27, υπέστει δηλαδή αύξηση 6x2=12 pixel ανα πλευρά. Το σχήμα 4.7 δείχνει τα μεγέθη των φίλτρων για τις τρεις πρώτες οκτάβες. Εύκολα παρατηρεί κανείς ότι οι οκτάβες επικαλύπτονται, πράγμα που συμβαίνει με σκοπό να καλυφθούν χωρίς κενά όλες οι πιθανές κλίμακες. Όπως είναι λογικό, ο αριθμός των εξαγόμενων σημείων ανά οκτάβα μειώνεται απότομα όσο ανεβαίνουν οι οκτάβες. Πρακτικά, μετά την τρίτη οκτάβα, τα εξαγόμενα σημεία είναι πολύ λίγα.

Για καθοριστεί η ακριβής θέση του σημείου ενδιαφέροντος πάνω στην εικόνα και ανεξάρτητα από την κλίμακα, εκτελείται ένας αποδοτικός αλγόριθμος για τον εντοπισμό τοπικών μέγιστων ανάμεσα σε κλίμακες (*non-maximum suppression*) σε μια 3x3x3(κλίμακες) γειτονιά όπως προτείνουν οι Neubeck και Van Gool [Neubeck & Van Gool, 2006]. Έπειτα, τα τοπικά μέγιστα από τις ορίζουσες



Σχήμα 4.7: Σχηματική παρουσίαση του μέγεθους των τετραγωνικών φίλτρων για τις τρεις πρώτες οκτάβες (από το [Bay et al., 2008]).

της μήτρας υπόκεινται σε παρεμβολή με την μέθοδο που είχε προτείνει ο Lowe [Lowe, 2004], διαδικασία ιδιαίτερα σημαντική γιατί σε αρκετές περιπτώσεις οι μεταβολές της κλίμακας σε διαδοχικά επίπεδα είναι αρκετά μεγάλες.

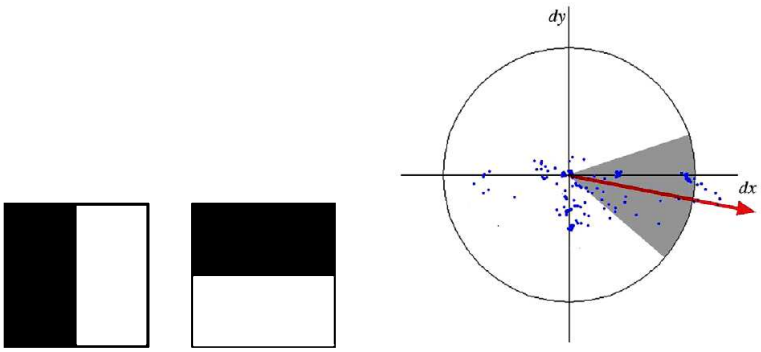
Σύμφωνα με τα παραπάνω, η μικρότερη δυνατή κλίμακα, μετά και την παρεμβολή, είναι για $\sigma = 1.6 = 1.2\frac{12}{9}$ που αντιστοιχεί σε φίλτρο 12x12 και η μεγαλύτερη για $\sigma = 3.2 = 1.2\frac{24}{9}$.

4.2.2 Εξαγωγή περιγραφέων

Ο περιγραφέας των SURF περιγράφει την κατανομή των εντάσεων στην περιοχή γύρω από το σημείο ενδιαφέροντος, παρόμοια με την πληροφορία για τις κατευθύνσεις των παραγώγων που εξάγεται από τα SIFT. Αντί για την κατανομή των παραγώγων όμως, στα SURF χρησιμοποιούνται οι κατανομές των κυματιδίων Haar πρώτης τάξης (*first order Haar wavelet responses*) στις κατευθύνσεις x και y. Η ταχύτητα εξαγωγής των περιγραφέων αυξάνεται πολύ με την χρήση των integral images και για να κρατηθεί μικρό το μέγεθος του περιγραφέα εν τέλει χρησιμοποιούνται 64 διαστάσεις.

Για να υπάρχει ανεξαρτησία στην περιστροφή, καθορίζεται σε κάθε σημείο ενδιαφέροντος ένας κύριος προσανατολισμός. Αρχικά υπολογίζονται οι αποκρίσεις των κυματιδίων Haar (τα wavelets που χρησιμοποιούνται φαίνονται στο σχήμα 4.8, αριστερά) στις κατευθύνσεις x και y σε μια ακτίνα $6s$ γύρω από το σημείο και με βήμα δειγματοληψίας s, με s να είναι η κλίμακα στην οποία βρέθηκε το σημείο. Το μέγεθος των κυματιδίων είναι επίσης εξαρτώμενο από την κλίμακα, καθώς έχουν μήκος πλευράς $4s$. Οι αποκρίσεις υπολογίζονται και αφού τους δοθούν βάρη με μια γκαουσιανή ($\sigma = 2$s) τις αναπαριστούμε σαν σημεία στον χώρο με την οριζόντια απόκριση σαν τετμημένη και την κάθετη σαν τεταγμένη. Ως κύριος προσανατολισμό ορίζουμε το μακρύτερο διάνυσμα που προκύπτει από την

άθροιση των αποκρίσεων σε ένα κυλιόμενο παράθυρο τόξου $\pi/3$ (σχήμα 4.8, δεξιά).



Σχήμα 4.8: Αριστερά: τα Haar Wavletes που χρησιμοποιούν τα SURF, δεξιά: εντοπισμός του κύριου προσανατολισμού από το κυλιόμενο παράθυρο (από το [Bay et al., 2008]).

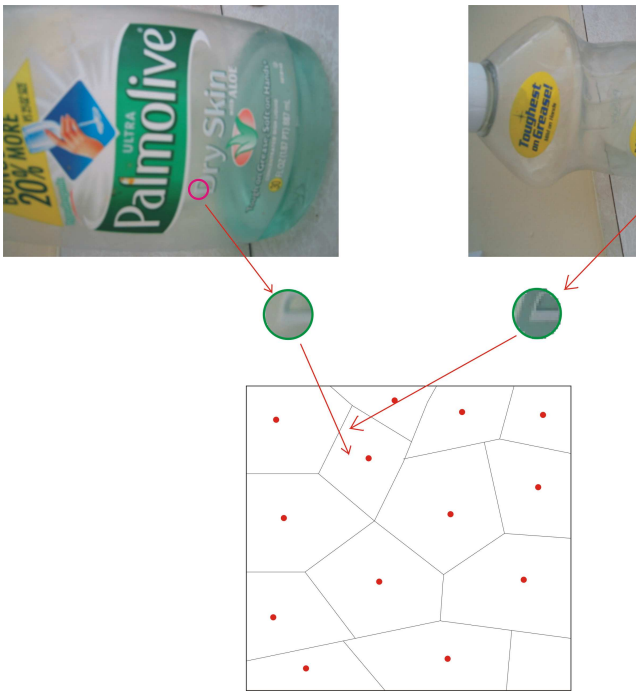
Μετά τον καθορισμό του κύριου προσανατολισμού, δημιουργούμε μια περιοχή πλάτους $20s$ γύρω από το σημείο ενδιαφέροντος, προσανατολισμένη κατάλληλα με τον προσανατολισμό του σημείου την οποία τεμαχίζουμε σε μικρότερες 4x4 τετράγωνες υποπεριοχές. Ο τεμαχισμός αυτός είναι απαραίτητος για να αποθηκεύεται επίσης χωρική πληροφορία. Σε κάθε υποπεριοχή υπολογίζονται οι αποκρίσεις των Haar κυματιδίων σε 5x5 ίσων αποστάσεων σημεία, και αθροίζονται ανά κατεύθυνση (οριζόντια και κάθετη σε σχέση πάντα με τον κύριο προσανατολισμό του σημείου) και αυτά τα αθροίσματα αποτελούν τις πρώτες τιμές του διανύσματος του περιγραφέα. Για να υπάρχει πληροφορία για την πολικότητα των μεταβολών των εντάσεων, αθροίζονται επίσης οι απόλυτες τιμές των αποκρίσεων στις δύο κατευθύνσεις.

Τελικά, κάθε υποπεριοχή χαρακτηρίζεται από τέσσερις τιμές: άθροισμα των οριζόντιων αποκρίσεων, άθροισμα των κάθετων αποκρίσεων, άθροισμα των απολύτων τιμών των οριζόντιων αποκρίσεων, άθροισμα των απολύτων τιμών των κάθετων αποκρίσεων. Αν ενώσουμε λοιπόν όλες τις 4x4 υποπεριοχές σε ένα διάνυσμα έχουμε 64 τιμές που χαρακτηρίζουν την περιοχή γύρω από το σημείο ενδιαφέροντος.

Για να υπάρχει ανεξαρτησία από την αντίθεση (contrast) μετατρέπουμε το διάνυσμα του περιγραφέα σε μοναδιαίο.

4.3 Δημιουργία οπτικού λεξικού και δεικτοδότηση

Το οπτικό λεξικό μπορεί να το σκεφτεί κανείς σε πλήρη αντιστοιχία με ένα συνηθισμένο γλωσσικό λεξικό. Σε αυτή την περίπτωση μια εικόνα μπορεί να αντιστοιχεί σε μια πρόταση, η καλύτερα σε μια παράγραφο κειμένου, το οποίο αναπαρίσταται στο χαρτί με κάποιες λέξεις. Δύο παράγραφοι πάνω στο ίδιο θέμα, για παράδειγμα το ποδόσφαιρο, θα περιέχουν αρκετές κοινές λέξεις όπως γήπεδο, ομάδα, προπονητής, γκολ και πολλές άλλες. Μπορούμε λοιπόν να συγκρίνουμε δύο παραγράφους ως προς το θέμα τους, ελέγχοντας την επικάλυψη όρων σε αυτές, χωρίς να χρειάζεται να ξέρουμε αυτό καθ' αυτό το θέμα. Στο σχήμα 4.9 φαίνονται δύο σημεία από δύο εικόνες τα οποία αναμένεται να



Σχήμα 4.9: Δύο σημεία απο δύο εικόνες τα οποία αντιστοιχίζονται στην ίδια οπτική λέξη.

αντιστοιχηθούν στην ίδια οπτική λέξη. Στο κάτω μέρος φαίνεται ο χώρος των οπτικών λέξεων που χωρίζεται από σε *κελιά Βορονόι*.

Βέβαια, είναι γεγονός ότι, την σημασιολογική σαφήνεια των γλωσσικών λέξεων δεν μπορεί να πλησιάσουν σε μεγάλο βαθμό οι οπτικές λέξεις, καθώς η πολυσημία της κάθε οπτικής λέξης καθιστά σε πολλές περιπτώσεις την αποσαφήνιση τους ακατόρθωτη.

4.3.1 Δημιουργία οπτικού λεξικού

Για την δημιουργία οπτικού λεξικού ακολουθείται η διαδικασία που παρουσιάστηκε και στην ενότητα 3.4. Εκτελείται δηλαδή *συσταδοποίηση* στα διανύσματα των 64-διάστατων SURF περιγραφέων που εξάγονται από έναν αριθμό εικόνων της συλλογής.

Η συσταδοποίηση γίνεται και πάλι με χρήση του αλγόριθμου *k-means*, ο οποίος περιγράφεται αναλυτικά στην ενότητα 3.4.1. Για μεγάλο αριθμό σημείων και συστάδων, η συσταδοποίηση είναι μια ιδιαίτερα χρονοβόρα διαδικασία. Είναι σύνηθες, ο αριθμός των «λέξεων» του οπικού λεξικού, άρα και ο αριθμός των συστάδων του k-means, να είναι ιδιαίτερα μεγάλος, πολλές φορές στην τάξη των δεκάδων χιλιάδων [Jegou et al., 2008b], [Philbin et al., 2007]. Με τον αριθμό των σημείων ανά εικόνα γύρω στα 500, η συσταδοποίηση σημείων από 10000 εικόνες σε 5000 σημεία μπορεί να πάρει αρκετές μέρες με τον βασικό αλγόριθμο k-means. Χωρίς βελτιώσεις του αλγορίθμου, τέτοια λεξικά είναι υπολογιστικά ανέφικτα.

Γι' αυτό το λόγο, η συσταδοποίηση δεν γίνεται σε όλες τις εικόνες της συλλογής, αλλά σε

υποσύνολο τους. Τα κέντρα των συστάδων που εξάγονται αποθηκεύονται στην βάση δεδομένων. Μπορούμε πλέον να αναπαραστήσουμε τις εικόνες της συλλογής όχι με τα σημεία τους, αλλά με βάση την απόσταση των σημείων αυτών αυτές τις οπτικές «λέξεις».

Αξίζει να σημειωθεί ότι από όσο μεγαλύτερη ποικιλία εικόνων εξάγεται το παραγόμενο από τη συσταδοποίηση λεξικό τόσο πιο καλά μπορούν να αναπαρασταθούν με αυτό εικόνες με διαφορετικά θεματικά περιεχόμενα. Το ίδιο συμβαίνει επίσης όσο αυξάνεται το μέγεθος του λεξικού.

4.3.2 Δημιουργία διανύσματος αναπαράστασης των εικόνων

Για να αναπαραστήσουμε τις εικόνες αξιοποιούμε το οπτικό λεξικό και σχηματίζουμε για κάθε εικόνα ένα *διάνυσμα αναπαράστασης* (*Model Vector*). Για κάθε περιγραφέα σημείου της εικόνας βρίσκουμε την κοντινότερη σε αυτόν οπτική λέξη. Έχουμε λοιπόν ένα ιστόγραμμα εμφάνισης των οπτικών λέξεων ως *κοντινότεροι γείτονες* (*nearest neighbors*) των σημείων.

Για την εύρεση της κοντινότερης σε κάθε σημείο οπτικής λέξης είναι, λόγω του μεγάλου αριθμού σημείων και οπτικών λέξεων, υπολογιστικά απαγορευτικό να ελεγχθούν εξαντλητικά όλες οι αποστάσεις. Χρησιμοποιούνται λοιπόν τα *k-d δέντρα* (*k-d trees*) τα οποία περιγράφονται στην επόμενη υποενότητα (4.3.3).

Το ιστόγραμμα εμφάνισης των οπτικών λέξεων ως κοντινότεροι γείτονες στα σημεία μια εικόνας κανονικοποιείται και οι μη μηδενικές τιμές του αποθηκεύονται σε έναν πίνακα ευρετήριο που πλησιάζει την δομή των *αντεστραμμένων αρχείων* (*inverted files*) που έχουν ευρεία χρήση στην ταχεία ανάκτηση κειμένου [Harman et al., 1992], [Zobel et al., 1998], [Squire et al., 2000], [Witten et al., 1999].

Κάθε εικόνα αναπαρίσταται λοιπόν από το ποιες οπτικές λέξεις και με ποια συχνότητα εμφανίζονται ως κοντινότερες στα σημεία της. Η μορφή του διανύσματος αναπαράστασης παρουσιάζεται στον πίνακα 4.1. Όταν αναφέρεται από εδώ και πέρα ότι μια οπτική λέξη εμφανίζεται σε μια εικόνα, εννοείται ότι είναι κοντινότερος γείτονας σε ένα ή περισσότερα από τα σημεία της.

image ID	term list	term frequencies

Πίνακας 4.1: Τα πεδία του διανύσματος αναπαράστασης των εικόνων. Στο πεδίο term list αποθηκεύεται μια λίστα από τους όρους (οπτικές λέξεις) που εμφανίζονται σε αυτήν και στο πεδίο term frequencies η λίστα με τις αντίστοιχες συχνότητες εμφάνισης.

Μιας και η όλη τεχνική δεικτοδότησης είναι εμπνευσμένη από την αναζήτηση κειμένου, πολλές φορές μπορεί να χρησιμοποιηθεί εκτός από τον όρο της συχνότητας εμφάνισης (tf - term frequency) και ένας όρος αντίστροφης συχνότητας εμφάνισης (idf - inverse document frequency), περίπτωση που μελετάται στην υποενότητα 4.3.4.

Η διαδικασία ερωτήματος σε μια συλλογή εικόνων με και χωρίς οπτικό λεξικό φαίνεται στο σχήμα 4.10. Στην πρώτη περίπτωση κατά την οποία δεν υπάρχει λεξικό, η σύγκριση των τοπικών περιγραφέων γίνεται άμεσα και βρίσκονται οι κοντινότερες περιοχές εξαντλητικά. Στην δεύτερη περίπτωση, στη δεξιά μεριά της εικόνας, κάθε μια από τις εικόνες της βάσης έχει εξ' αρχής αντιστοιχίσει κάθε σημείο τις με κάποιο από τα σημεία-λέξεις του οπτικού λεξικού. Έτσι, σε μια νέα εικόνα-ερώτημα αρκεί να αντιστοιχηθούν τα σημεία της στις κοντινότερες οπτικές λέξεις του λεξικού. Τότε, πα-

ρόμοιες θα είναι οι εικόνες της συλλογής, τα σημεία των οποίων αντιστοιχούνται επίσης στις ίδιες λέξεις.

4.3.3 Αναζήτηση κοντινότερου γείτονα με k-d Δέντρα

Τα *k-d δέντρα* (*k-d trees*) έχουν αρκετές δεκαετίες που εφαρμόζονται στην αναζήτηση πληροφοριών [Freidman et al., 1977], [Bentley, 1975]. Τα δέντρα αυτά είναι *δυαδικά δέντρα* (binary trees) και είναι μια δομή δεδομένων που αποθηκεύει έναν πεπερασμένο αριθμό σημείων διάστασης k (το k-d είναι συντομογραφία του k-dimentional) και έχουν πληθώρα εφαρμογών στον τομέα της υπολογιστικής μάθησης [Moore, a] και των νευρωνικών δικτύων [Omohundro, 1987]. Εμείς τα χρησιμοποιούμε για την εύρεση της κοντινότερης σε κάθε σημείο μιας εικόνας οπτικής λέξης, πρόβλημα ιδιαίτερα δύσκολο και χρονοβόρο λόγω της υψηλής διάστασης.

Το k-d δέντρο κατασκευάζεται διαιρώντας τον πολυδιάστατο χώρο σε ημιεπίπεδα διαδοχικά ως προς τις διάφορες διαστάσεις. Ως είσοδο για την κατασκευή του παίρνει n στοιχεία διάστασης k. Διαδοχικά σε κάθε μια από τις k διαστάσεις κυκλικά, ο χώρος διαιρείται στη μέση σε δύο ημιεπίπεδα και κάθε ένα από αυτά έπειτα διαιρούνται ξανά ως προς την επόμενη διάσταση και η διαδικασία αυτή τερματίζεται όταν κάθε ένα από τα σημεία της εισόδου έχει μόνο του την δικιά του περιοχή.

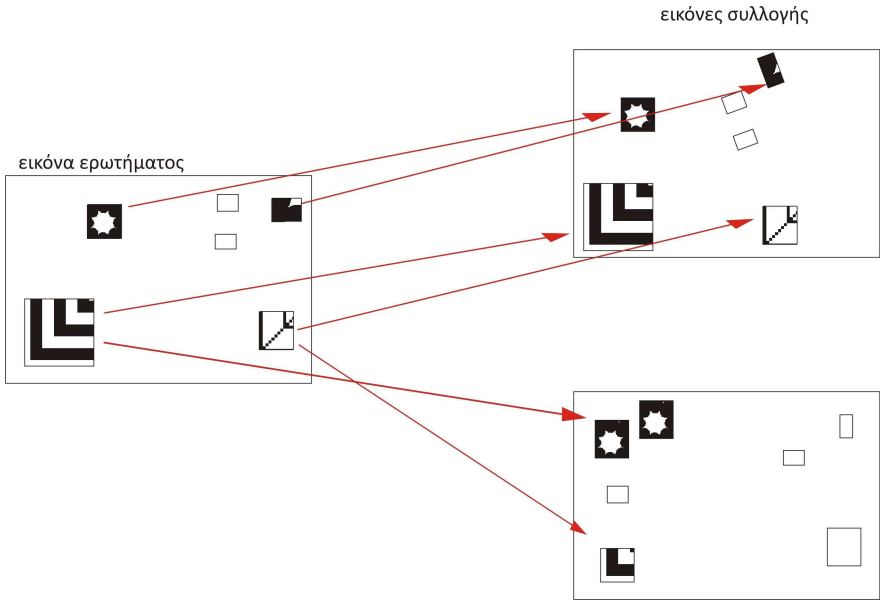
Στο σχήμα 4.12 απεικονίζεται ένα τρισδιάστατο k-d δέντρο. Ο πρώτος διαχωρισμός (κόκκινο χρώμα) κόβει τον αρχικό χώρο (λευκό χρώμα) σε δύο υπο-χώρους, κάθε ένας από τους οποίους διαχωρίζεται (με πράσινο χρώμα) επίσης σε δύο υπο-χώρους. Τέλος κάθε ένας από αυτούς τους τέσσερις υπο-χώρους ξανά χωρίζεται (με μπλέ χρώμα). Καθώς δεν μπορεί να υπάρξει επιπλέον διαχωρισμός, οι τελικοί οκτώ υπο-χώροι, καλούνται *φύλλα του δέντρου*.

Δημιουργείται έτσι ένα δέντρο, το οποίο μας παρέχει μια πολύ γρήγορη αναζήτηση σε όλα τα σημεία ως προς την θέση τους. Το δέντρο αυτό έχει ύψος $log(n)$.

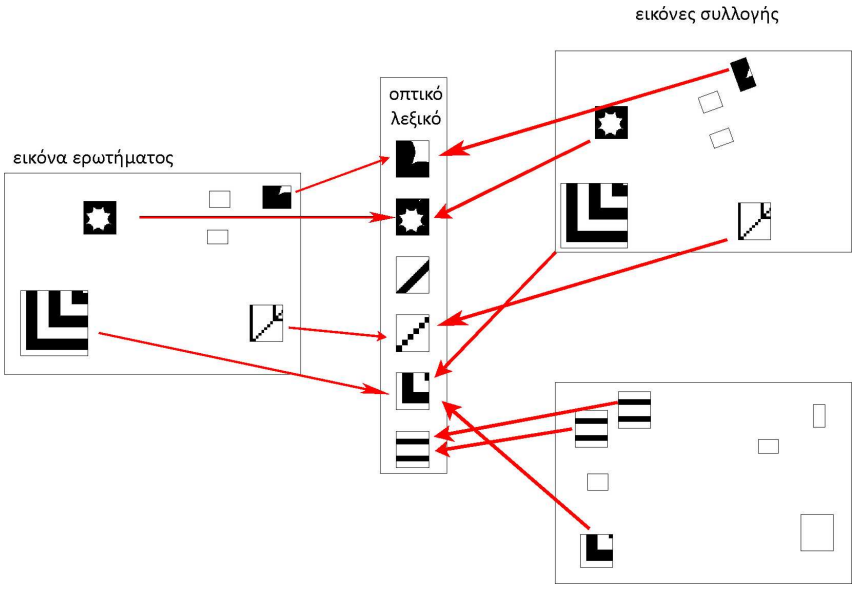
Ένα τέτοιο δέντρο, για σημεία διάστασης 2, φαίνεται στο σχήμα 4.11 στην αριστερή μεριά. Στην δεξιά μεριά παρουσιάζεται ο διδιάστατος χώρος των σημείων και οι περιοχές του καθενός. Σε μια περίπτωση ερωτήματος για τον πλησιέστερο γείτονα ενός σημείου q εκτός δέντρου, αρκεί να προσπελαστούν οι χρωματιστοί, στο αριστερό τμήμα κόμβοι του δέντρου.

Στην περίπτωση που εξετάζεται εδώ, το k-d δέντρο δημιουργείται από όλα τα 64-διάστατα κέντρα των συστάδων που εξάγονται από την συσταδοποίηση (ενότητα 4.3) και αποτελούν τις οπτικές λέξεις του λεξικού μας. Το δέντρο αυτό δημιουργείται μια φορά, όσες εικόνες και να θέλουμε να δεικτοδοτήσουμε.

Κάθε σημείο της εικόνας, για την οποία θέλουμε να βρούμε το διάνυσμα αναπαράστασης, έρχεται ως ερώτημα για πλησιέστερο γείτονα στο δέντρο και αυτό βρίσκει την κοντινότερη σε αυτό το σημείο περιοχή, η οποία αντιστοιχεί σε κάποιο κέντρο συστάδας, σε κάποια οπτική λέξη. Μια πολύ πρόσφατη ανασκόπηση των μεθόδων αναζήτησης πλησιέστερου γείτονα, θεωρεί ότι τα k-d δέντρα δεν είναι τα καταλληλότερα σε μεγάλο αριθμό διαστάσεων, ενώ τα καλύτερα αποτελέσματα εξάγονται με την χρήση *vantage point δέντρων* [Kumar et al., 2008].

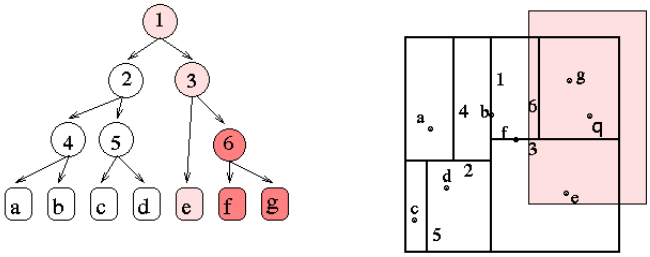


(a)

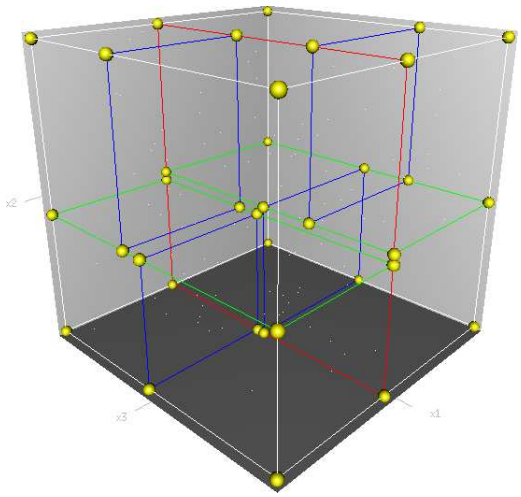


(b)

Σχήμα 4.10: Ταίριασμα εικόνων χωρίς οπτικό λεξικό (4.10(a)) και με λεξικό (4.10(b)).



Σχήμα 4.11: Ερώτημα στο δέντρο για τον πλησιέστερο γείτονα του q.



Σχήμα 4.12: Τρισδιάστατο k-d δέντρο (σχήμα από την Wikipedia).

4.3.4 Όρος αντίστροφης συχνότητας εμφάνισης και η λίστα τερματισμού

Ο *όρος αντίστροφης συχνότητας*, είναι μια ακόμη τεχνική δανεισμένη από την θεωρία αναζήτησης και αναγνώρισης κειμένου [McGill & Salton, 1983] [Rui et al., 1997a], η οποία έχει τα τελευταία χρόνια εφαρμοστεί και στην ανάκτηση εικόνων είτε έμμεσα σε συνδυασμό με γλωσσική ανάλυση [Kanade, 1998], είτε άμεσα στην αναζήτηση με οπτικά λεξικά [Sable & Hatzivassiloglou, 2000] [Sivic & Zisserman, 2003] [Chum et al., 2008].

Μέχρι τώρα, το διάνυσμα αναπαράστασης μιας εικόνας λάμβανε υπόψη μονάχα μια τιμή για κάθε οπτική λέξη που περιέιχε η εικόνα: την τιμή της συχνότητας εμφάνισης της λέξης αυτής ως πλησιέστερος γείτονας στα σημεία της εικόνας αυτής. Τώρα, προστίθεται και ο όρος αντίστροφης συχνότητας -*idf (inverse document frequency)* ο οποίος ορίζεται ως:

$$d_k = log\frac{N}{n_k} \tag{4.4}$$

	εικόνα	εικόνα ερωτήματος
οπτικές λέξεις	...	
	VW34	0.13
	VW35	0.1
	VW36	
	VW37	0.4
	VW38	0.2
	VW39	0.3
	...	

	εικόνα	1	2	3	4	...
οπτικές λέξεις	...	...	...	...	...	...
	VW34	0.12	0.12	0.05	0.22	...
	VW35			0.1		...
	VW36			0.4		...
	VW37	0.3			0.3	...
	VW38		0.6		0.12	...
	VW39	0.4		0.1	0.33	...
	...	...	...	...	...	....

Σχήμα 4.13: ένα μέρος του πίνακα οπτικών λέξεων - εικόνων της βάσης δεδομένων, και ένα μέρος του διανύσματος αναπαράστασης της εικόνας του ερωτήματος.

όπου N είναι ο αριθμός των εικόνων της συλλογής και n_k ο αριθμός των φορών που εμφανίζεται η λέξη VW_k ως κοντινότερος γείτονας σε όλα τα σημεία από όλες της εικόνες της συλλογής.

Είναι δηλαδή ένας συντελεστής βάρους ο οποίος μας δείχνει πόσο «δημοφιλής» είναι οι λέξεις σε όλη την συλλογή. Αυτή είναι μια σημαντική παράμετρος, καθώς λέξεις που εμφανίζονται πάρα πολύ συχνά, άρα και σε πολλές εικόνες, δεν μας βοηθάνε στη σωστή ανάκτηση. Παρόμοια, σπάνιες λέξεις οι οποίες εμφανίζονται πολύ αραιά, είναι πολύ πιθανό να είναι θόρυβος από της εικόνες και όχι κάποιο πραγματικά χαρακτηριστικό σημείο των εικόνων.

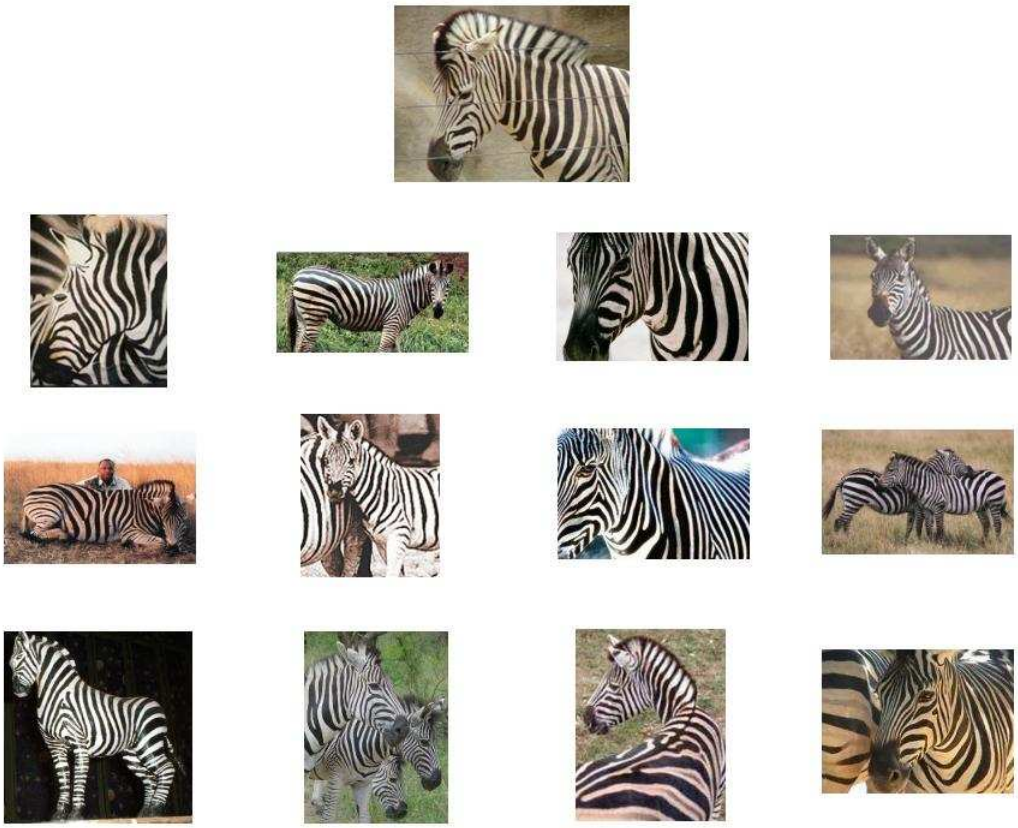
Γι' αυτό τον λόγο επιλέγεται σε κάποιες περιπτώσεις να δημιουργηθεί μια /emphλίστα τερματισμού (*stop list*) σύμφωνα με την οποία μηδενίζεται η επιρροή των οπτικών λέξεων που περιέχονται σε αυτήν. Τέτοιες αποτελούν οπτικές λέξεις οι οποίες εμφανίζονται είτε πολύ συχνά είτε πολύ σπάνια. Η stop list αυτή έρχεται σε άμεση αντιστοιχία με την αντίστοιχη λίστα στην ανάλυση κειμένου, η οποία περιέχει λέξεις που δεν μας βοηθάνε στην κατανόηση (όπως πχ άρθρα ή σύνδεσμους). Με την stop list θα αμελούταν, για παράδειγμα η οπτική λέξη 34 στο σχήμα 4.13 η οποία είναι κοινή σε πολλές εικόνες (στο απλουστευμένο παράδειγμα του σχήματος) και δεν παρέχει ουσιαστικά βοήθεια για να διαχωρίσουμε τις εικόνες.

4.4 Ταίριασμα των εικόνων

Και στην προκειμένη περίπτωση, όπως και στην περίπτωση με τον οπτικό θησαυρό από περιοχές (ενότητα 3.6), για να υπολογιστεί η απόσταση μεταξύ δύο εικόνων αρκεί να υπολογιστεί η Ευκλείδεια απόσταση μεταξύ των διανυσμάτων αναπαράστασης των δύο αυτών εικόνων.

Όταν εισάγεται στο σύστημα ένα νέο ερώτημα αναζήτησης, τότε, αν η εικόνα είναι εικόνα της συλλογής, φορτώνεται από την βάση το διάνυσμα αναπαράστασης και υπολογίζεται η απόσταση του από όλα τα αντίστοιχα διανύσματα όλων των εικόνων της συλλογής. Οι αποστάσεις ταξινομούνται και στον χρήστη του συστήματος επιστρέφονται ως παρόμοιες οι εικόνες που έχουν την μικρότερη απόσταση.

Ένα απλοποιημένο παράδειγμα αυτής της διαδικασίας αποτυπώνεται σχηματικά στο σχήμα 4.13, όπου φαίνεται ένα μέρος του πίνακα οπτικών λέξεων - εικόνων της βάσης δεδομένων, και ένα μέρος του διανύσματος αναπαράστασης της εικόνας του ερωτήματος. Σύμφωνα με αυτό το κομμάτι των διανυσμάτων αναπαράστασης, η εικόνα του ερωτήματος τείνει να μοιάζει με τις εικόνες 1 και 4 της



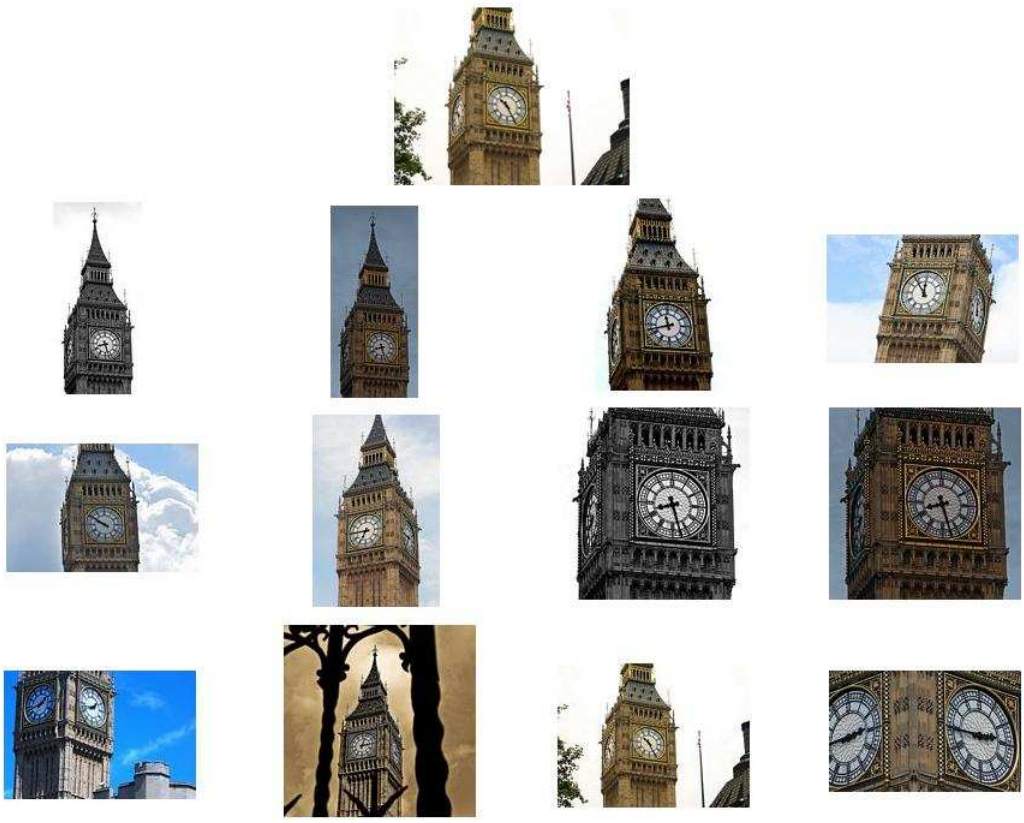
Σχήμα 4.14: Αποτελέσματα αναζήτησης στην συλλογή από εικόνες του Caltech. Ζέβρα.

βάσης.

Μερικά παραδείγματα αναζήτησης φαίνονται στα σχήματα 4.14 και 4.15. Η εικόνα του ερωτήματος φαίνεται στο πάνω μέρος των σχημάτων.

4.5 Έλεγχος γεωμετρίας

Δυστυχώς, μετά το στάδιο αναζήτησης με σύγκριση των διανυσμάτων αναπαράστασης των εικόνων είναι σύνηθες το φαινόμενο να επιστρέφονται ως κοντινότερες αρκετές εικόνες οι οποίες μπορεί να μοιάζουν στην εικόνα του ερωτήματος ως προς τις οπτικές λέξεις που περιέχουν, τα σημεία όμως που αντιστοιχούν στις λέξεις αυτές είναι κατανεμημένα με διαφορετική χωρική δομή στην ανακτημένη εικόνα σε σχέση με την εικόνα του ερωτήματος. Θα ήταν λοιπόν ιδιαίτερα χρήσιμο να μπορούσε να συμμετέχει στην εξαγωγή του τελικού αποτελέσματος και κάποιο γεωμετρικό κριτήριο για τα σημεία ανάμεσα στις δύο εικόνες που μοιάζουν με βάση τα χαρακτηριστικά των γύρω τους περιοχών. Δοκιμάστηκε με θετικά αποτελέσματα η μέθοδος *RANSAC* που προτάθηκε από τους Fischler και Bolles το 1981 [Fischler & Bolles, 1981].



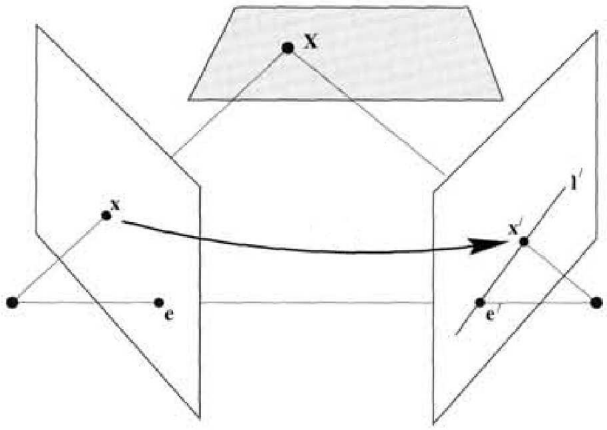
Σχήμα 4.15: Αποτελέσματα αναζήτησης σε συλλογή με εικόνες από το Flickr. Big Ben.

Μιας και η μέθοδος αυτή είναι απαιτητική υπολογιστικά, δεν εκτελείται σε όλες της εικόνες της βάσης, αλλά στις πρώτες εικόνες που επιστρέφονται από την σύγκριση των διανυσμάτων αναπαράστασης. Είναι σημαντικό λοιπόν να τονιστεί ότι δεν συμμετέχει αμιγώς στην αναζήτηση των εικόνων, αλλά στην *επαναταξινόμηση* (*re-ranking*) των καλύτερων αποτελεσμάτων, διορθώνοντας σε μεγάλο βαθμό τις λάθος ανακτήσεις εικόνων (*false positives*). Με κατωφλίωση των αποτελεσμάτων στο τέλος, μπορούν να επιτευχθούν υψηλές αποδόσεις στην ανάκτηση, με κάποιο επιπρόσθετο κόστος, βέβαια, στον χρόνο αναζήτησης.

4.5.1 Εκτίμηση της ομογραφίας με RANSAC

Η μέθοδος *RANSAC* (RANdom SAmple Consensus) έχει σαν στόχο την εύρεση της *ομογραφίας* (*homography*) ανάμεσα σε δύο εικόνες, δηλαδή τον προοπτικό μετασχηματισμό που μετασχηματίζει κάθε σημείο x_i της μίας εικόνας σε κάθε σημείο x'_i της άλλης, στην περίπτωση που υπάρχουν πολλές λάθος αντιστοιχίες.

Με δεδομένες τις αντιστοιχίες των σημείων ανάμεσα σε δύο εικόνες, δηλαδή τα ζεύγη $x_i \leftrightarrow x'_i$,



Σχήμα 4.16: Τις δύο εικόνες μιας στερεοσκοπικής κάμερας μπορεί εύκολα κανείς να τις δει σαν δύο όψεις του ίδιου αντικείμενου από δυο διαφορετικές οπτικές γωνίες.

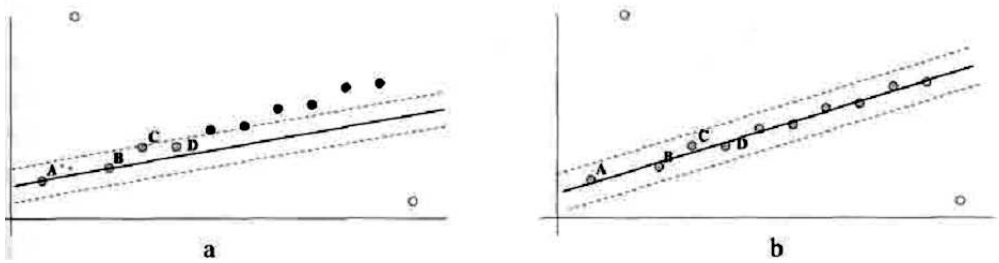
ο *πίνακας ομογραφίας* H ορίζεται ως η μήτρα:

$$x_i' = H x_i \tag{4.5}$$

Η εκτίμηση της ομογραφίας έχει άμεση χρησιμότητα στην ρύθμιση στερεοσκοπικών καμερών όπου εκεί η μία εικόνα διαφέρει από την άλλη μόνο ως προς έναν προοπτικό μετασχηματισμό. Άμεσα, όμως, μπορεί κάποιος να ανάγει την αντιστοιχία μεταξύ δύο εικόνων από στερεοσκοπική κάμερα σε δύο εικόνες του ίδιου αντικείμενου από δυο διαφορετικές οπτικές γωνίες (σχήμα 4.16). Επειδή στην δεύτερη περίπτωση οι μεταβολές μπορεί να είναι αρκετά μεγαλύτερες, κάτι που θα οδηγήσει σε λίγες σωστές και αρκετές λάθος αντιστοιχίες μεταξύ των σημείων των δύο εικόνων, χρησιμοποιείται η μέθοδος RANSAC που είναι ιδιαίτερα εύρωστη και μπορεί να εκτιμήσει την ομογραφία ακόμα και σε περιβάλλοντα με μεγάλο ποσοστό από λάθος αντιστοιχίες.

Στόχος της μεθόδου είναι όχι μόνο να προσδιορίσει μια μήτρα ομογραφίας μεταξύ των δυο εικόνων, αλλά και να ταξινομήσει τα σημεία σε δύο κατηγορίες: σε αυτά με σωστές αντιστοιχίες (*inliers*) και σε αυτά με λάθος αντιστοιχίες (*outliers*).

Ένα απλό παράδειγμα, που φαίνεται στο σχήμα 4.17, είναι το πρόβλημα εκτίμησης ενός μοντέλου το οποίο προσαρμόζει μια γραμμή σε μία ομάδα από σημεία. Αυτό το πρόβλημα μπορεί να θεωρηθεί αντίστοιχο με την εκτίμηση ενός μονοδιάστατου αφινικού μετασχηματισμού $x' = ax + b$ ανάμεσα σε αντιστοίχοντα σημεία που ανήκουν σε δύο γραμμές. Τα μαύρα σημεία είναι οι inliers και τα άδεια οι outliers. Η μέθοδος ελαχίστων τετραγώνων στο παράδειγμα αυτό θα έδινε την εκτίμηση του σχήματος a. Η μέθοδος RANSAC δίνει την, κατά πολύ σωστότερη, εκτίμηση του σχήματος b. Η ιδέα του RANSAC είναι πολύ απλή: Διαλέγουμε δύο σημεία τυχαία, τα οποία προφανώς ορίζουν μια ευθεία. Η *υποστήριξη* (*support*) της ευθείας αυτής, δίνεται τον αριθμό των σημείων που βρίσκονται σε απόσταση μικρότερη από t από την ευθεία. Αφότου εκτελέσουμε την τυχαία επιλογή σημείων για έναν αριθμό φορών, θεωρούμε ως λύση την ευθεία με την μεγαλύτερη υποστήριξη. Ως inliers θεωρούνται όσα σημεία βρίσκονται σε απόσταση μικρότερη από t από την ευθεία (αυτά ορίζουν και το σύνολο consensus που αναφέρει και το όνομα του RANSAC) και τα υπόλοιπα θεωρούνται outliers.



Σχήμα 4.17: Πρόβλημα εκτίμησης ενός μοντέλου το οποίο προσαρμόζει μια ευθεία στα παραπάνω δεδομένα (από το βιβλίο των Hartley και Zisserman [Hartley & Zisserman, 2004]).

Ορίζοντας ως καταλληλότερο μοντέλο την ευθεία με την μεγαλύτερη υποστήριξη, ευνοούμε την καλύτερα προσαρμοσμένη στα δεδομένα ευθεία. Για παράδειγμα, η ευθεία $\langle a, b\rangle$ του σχήματος έχει υποστήριξη 10 ενώ η ευθεία $\langle a, d\rangle$ μονάχα 4.

Γενικά, όταν θέλουμε να προσαρμόσουμε ένα μοντέλο σε δεδομένα το αρχικό τυχαίο μας δείγμα θα πρέπει να περιέχει τόσα σημεία όσα είναι αναγκαία για να οριστεί το μοντέλο. Στο παραπάνω παράδειγμα το μοντέλο ήταν η ευθεία και χρειάστηκαν δύο σημεία. Όταν το μοντέλο θα είναι η ομογραφία θα απαιτούνται τουλάχιστον τέσσερα σημεία για να οριστεί.

Με δεδομένες τις αντιστοιχίες των σημείων στις δύο εικόνες, ο βασικός *αλγόριθμος του RANSAC* για τον εντοπισμό της ομογραφίας μεταξύ δύο εικόνων έχει ως εξής:

- Διαδοχικά επιλέγονται με τυχαίο τρόπο 4 σημεία, δηλαδή ο ελάχιστος αριθμός σημείων που χρειάζεται για να οριστεί μια ομογραφία, και με αυτά υπολογίζεται το μοντέλο ομογραφίας σύμφωνα με τη σχέση (4.5). Υποθέτουμε δηλαδή αρχικά ότι τα 4 αυτά τυχαία σημεία ανήκουν στο μοντέλο ομογραφίας που θέλουμε να βρούμε (υποθέτουμε δηλαδή ότι είναι inliers).

- Με βάση το μοντέλο που υπολογίστηκε προηγουμένως και μια παράμετρο απόστασης t, υπολογίζεται ο αριθμός των σημείων που επαληθεύουν το μοντέλο αυτό. Για να επαληθεύει ένα σημείο το μοντέλο, θα πρέπει η αντιστοιχία του στην δεύτερη εικόνα να βρίσκεται σε απόσταση μικρότερη από t από την θέση που προβλέπει το μοντέλο ομογραφίας. Δηλαδή:

$$d^2_{κθετη} < t^2 \tag{4.6}$$

  όπου $t^2 = F_m^{-1}(\alpha)\sigma^2$ και $F_m(k^2) = \int_0^{k^2} \chi^2_m(\xi)d\xi$ είναι η συνάρτηση πυκνότητας πιθανότητας για την κατανομή του σφάλματος που θεωρούμε ότι ακολουθεί κατανομή χ^2_m με m βαθμούς ελευθερίας. Τα σημεία για τα οποία ισχύει η παραπάνω συνθήκη, θεωρούνται *inliers* σε σχέση με το συγκεκριμένο μοντέλο. Το α ορίζεται συνήθως ίσο με 0.95 ώστε να υπάρχει 95% πιθανότητα κάθε σημείου να είναι inlier.

- Τετράδες σημείων επιλέγονται διαδοχικά και ανάμεσα στις επαναλήψεις κρατιέται ο μέχρι τότε μέγιστος αριθμός inliers που έχει βρεθεί καθώς και ο πίνακας ομογραφίας για το αντίστοιχο μοντέλο.

- Η διαδικασία επαναλαμβάνεται είτε μέχρι να ολοκληρωθεί ένας προκαθορισμένος αριθμός επαναλήψεων είτε μέχρι να πέσει η πιθανότητα να βρεθούν περισσότεροι inliers σε κάποιο μοντέλο κάτω από η τοις εκατό. Η πιθανότητα αυτή ορίζεται ως εξής:

$$\eta = (1 - P_I)^k \tag{4.7}$$

όπου k είναι ο αριθμός των σημείων - αντιστοιχιών μεταξύ των δύο εικόνων και το P_I είναι η πιθανότητα να επιλεκτεί ένα «αμόλυντο» από outliers δείγμα, και ισούται με

$$P_I = \frac{\binom{I}{m}}{\binom{N}{m}} = \prod_{j=0}^{m-1} \frac{I-j}{N-j} \approx \varepsilon^m \tag{4.8}$$

όπου $\varepsilon = I/N$ είναι ο λόγος inliers / σημεία. Το η συνήθως ορίζεται ίσο με 0.99.

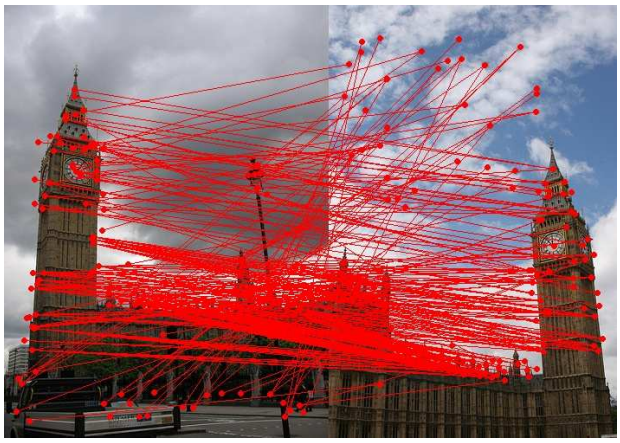
- Όταν ικανοποιηθεί η παραπάνω συνθήκη επαναπροσδιορίζεται το μοντέλο ομογραφίας με όλα τα σημεία που θεωρήθηκαν inliers από τα παραπάνω βήματα.

4.5.2 Εφαρμογή του RANSAC με δεδομένα τα διανύσματα αναπαράστασης των εικόνων

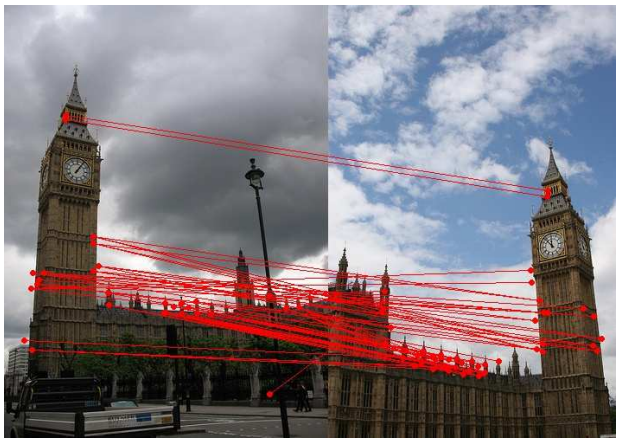
Είναι εμφανές ότι ο αλγόριθμος του RANSAC, όπως παρουσιάστηκε πριν, εξαρτάται σε μεγάλο βαθμό από τις αντιστοιχίες των σημείων που θα του δοθούν για να βγάλει σωστή εκτίμηση της ομογραφίας. Οι αντιστοιχίες αυτές δεν είναι διαθέσιμες άμεσα, και το να εκτελεστεί μια διαδικασία εύρεσης των κοντινότερων γειτόνων ανάμεσα σε όλα τα σημεία των δύο εικόνων είναι μια ιδιαίτερα χρονοβόρα διαδικασία. Θα εκμεταλλευτούμε λοιπόν τις αντιστοιχίες σημείων - οπτικών λέξεων για να δημιουργήσουμε αντιστοιχίες σημείων μεταξύ των δύο εικόνων. Με μια επιπλέον μορφή δεικτοδότησης, σύμφωνα με την οποία για κάθε εικόνα αποθηκεύεται το ποια σημεία έχουν ως κοντινότερο γείτονα την κάθε οπτική λέξη, η διαδικασία αυτή είναι ιδιαίτερα γρήγορη.

Παρόλα αυτά, η ελαφριά αυτή διαδικασία αντιστοίχησης σημείων επιφέρει λόγω του χβαντισμού των οπτικών λέξεων, πολλές λάθος αντιστοιχίες. Αν σε μια οπτική λέξη αντιστοιχούν για παράδειγμα 10 σημεία της πρώτης εικόνας και 8 της δεύτερης, αντί για 8 σωστές αντιστοιχίες θα πάρουμε 80. Έτσι, τη διαδικασία αντιστοίχησης ακολουθεί και μια διαδικασία απόρριψης πολλών από αυτές ελέγχοντας την αντιστοιχία των γειτόνων του κάθε σημείου με τους γείτονες του αντίστοιχού του σημείου [Sivic & Zisserman, 2003].

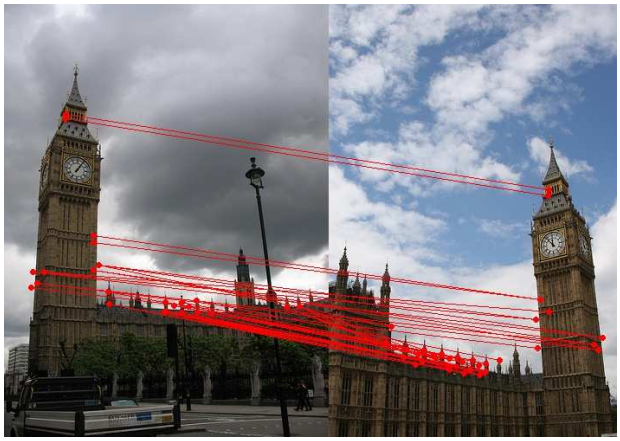
Παραδείγματα της όλης διαδικασίας φαίνονται στα σχήματα 4.18 και 4.19. Στις πάνω εικόνες φαίνονται οι αρχικές αντιστοιχίες που προκύπτουν από την αντιστοίχηση όλων των σημείων των δύο εικόνων που ανήκουν στην ίδια οπτική λέξη, στη μεσαία φαίνονται οι αντιστοιχίες που μένουν μετά τον έλεγχο αντιστοιχίας της γειτονιάς και στην κάτω εικόνα οι τελικές αντιστοιχίες που εντόπισε ο RANSAC. Στην πρώτη περίπτωση (σχήμα 4.18) υπάρχει αντιστοιχία μεταξύ των δύο εικόνων και βρέθηκαν πολλοί inliers, σε αντίθεση με την δεύτερη (σχήμα 4.19), όπου δεν υπάρχει και βρέθηκαν λίγοι.



(a)

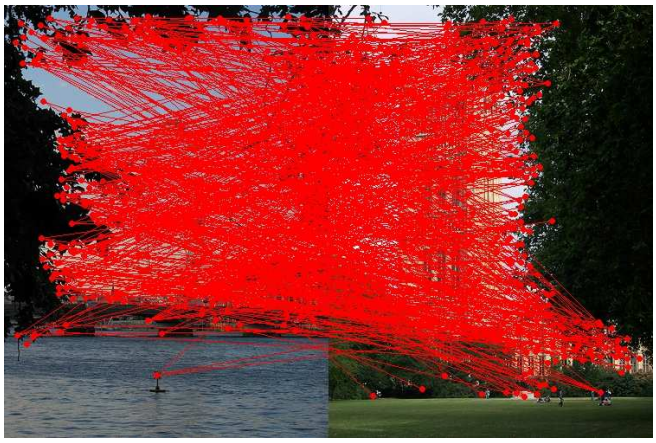


(b)

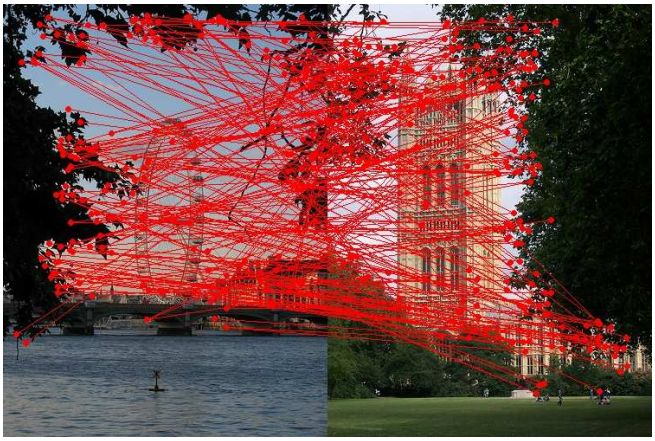


(c)

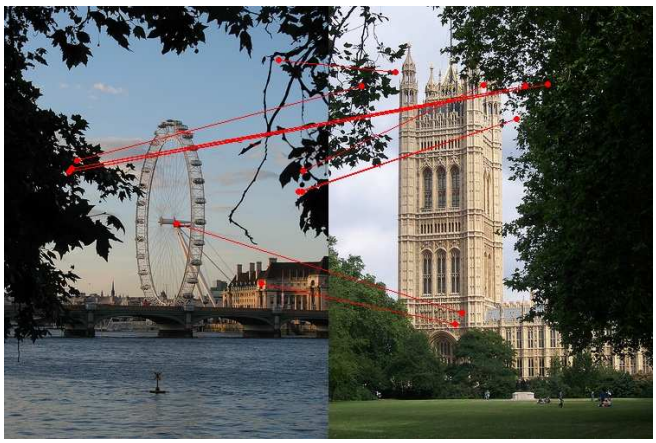
Σχήμα 4.18: Εφαρμογή του RANSAC σε περίπτωση ύπαρξης σχετισμού ανάμεσα στις εικόνες.



(a)



(b)



(c)

Σχήμα 4.19: Εφαρμογή του RANSAC σε περίπτωση μη ύπαρξης σχετισμού ανάμεσα στις εικόνες.

4.6 Εφαρμογή σε συλλογή φωτογραφιών με μεταδεδομένα γεωγραφικής θέσης

Η γιγαντοποίηση των συλλογών εικόνων στο διαδίκτυο, συνοδεύεται και από την προσθήκη-ενσωμάτωση στις εικόνες διαφόρων μεταδεδομένων για την πληρέστερη περιγραφή τους. Στην κατηγορία αυτή περιλαμβάνονται κάποιο κείμενο που περιγράφει το περιεχόμενο της εικόνας (description), χαρακτηριστικές λέξεις για την εικόνα (tags) (λέξεις στις οποίες συνήθως βασίζεται και η αναζήτηση εικόνων με βάση κείμενο που παρέχεται στις ιστοσελίδες με συλλογές) αλλά επίσης και μεταδεδομένα γεωγραφικής θέσης (*geo-tag*). Τα μεταδεδομένα θέσης είναι το γεωγραφικό μήκος και πλάτος του μέρους που τραβήχτηκε η φωτογραφία, τιμές που είτε εξάγονται αυτόματα μέσω GPS από ορισμένες ψηφιακές φωτογραφικές μηχανές είναι καθορίζονται χειροκίνητα από τους χρήστες μέσω της εκάστοτε ιστοσελίδας που φιλοξενεί την συλλογή φωτογραφιών. Στην τεράστια συλλογή της ιστοσελίδας Flickr[1] για παράδειγμα, εκτιμάται ότι καταχωρούνται κατά μέσο όρο κάθε μήνα περίπου τρία εκατομμύρια φωτογραφίες με μεταδεδομένα θέσης.

Εφόσον η αναζήτηση με οπτικά χαρακτηριστικά μπορεί να χρησιμοποιηθεί αποδοτικά για την ανάκτηση φωτογραφιών από κτίρια (βλέπε ενότητα 6.6.2), με κάποια επεξεργασία στην πληροφορία θέσης των εικόνων που ανακτώνται θα μπορούσε εύκολα να εξαχθεί πληροφορία θέσης για την εικόνα του ερωτήματος.

Οι σωστές εικόνες που ανακτήθηκαν, αν το ερώτημα απεικονίζει για παράδειγμα κάποιο συγκεκριμένο μνημείο μιας πόλης, αναμένεται να έχουν την ίδια ή σχεδόν την ίδια πληροφορία γεωγραφικής θέσης με το μνημείο αυτό. Έτσι με απλή ανάλυση των geo-tags των εικόνων που επιστρέφονται με την αναζήτηση, αναμένεται το μεγαλύτερο συνεκτικό υποσύνολο τους να έχει την ίδια γεωγραφική θέση με την εικόνα.

Πειράματα έγιναν σε συλλογή από 2000 περίπου εικόνες του Flickr, στις οποίες απεικονίζονται μνημεία της πόλης του Λονδίνου. Όλες οι εικόνες έχουν πληροφορία γεωγραφικής θέσης.

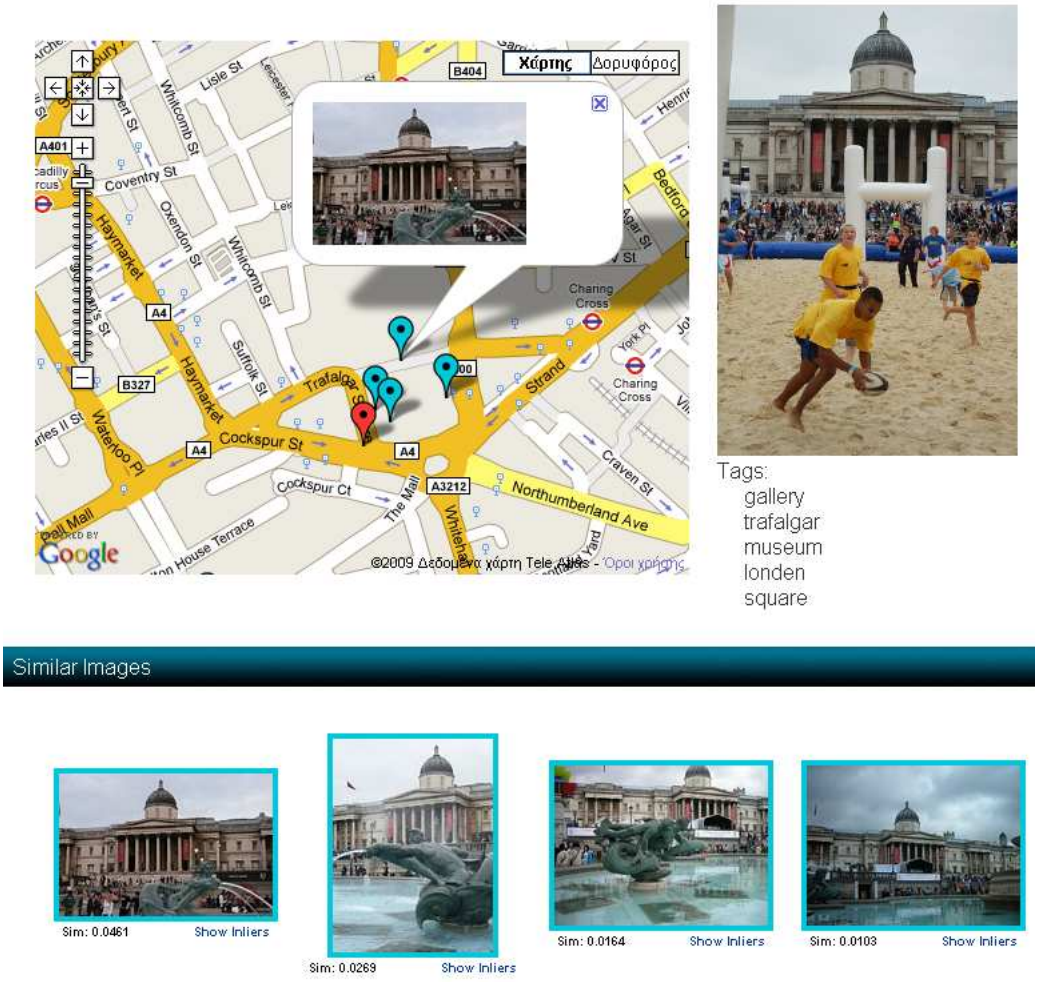
Στην διαδικτυακή σελίδα που αναπτύχθηκε, μπορεί ο χρήστης να εκτελέσει ερώτημα με μια φωτογραφία του που απεικονίζει μεταξύ άλλων κάποιο μνημείο του Λονδίνου, και μετά την αναζήτηση να πάρει μια εκτίμηση της πιθανού μέρους όπου μπορεί να τραβήχτηκε η φωτογραφία. Αυτό επιτυγχάνεται με μια διαδικασία συσταδοποίησης με σύμπτυξη πάνω στα διδιάστατα geo-tags των φωτογραφιών που επιστράφηκαν ως κοντινότερες.

Ο αλγόριθμος που χρησιμοποιείται για την συσταδοποίηση αυτή είναι ο αλγόριθμος *RNN - reciprocal nearest neighbor*[Olson, 1995],[Leibe et al., 2008] , από την έξοδο του οποίου επιλέγεται η πιο συνεκτική συστάδα ως πιθανή για πρόβλεψη θέσης.

Για την οπτικοποίηση των αποτελεσμάτων χρησιμοποιήθηκε το γραφικό περιβάλλον του Google maps όπως φαίνεται στο σχήμα 4.20. Στον χάρτη με κόκκινο marker φαίνεται η εκτιμώμενη γεωγραφική θέση και με γαλάζιο η θέση των εικόνων που εκτιμάται ότι απεικονίζουν το ίδιο μνημείο με το ερώτημα. Αποτελέσματα ανάκτησης φαίνονται στο σχήμα 4.21 και στο σχήμα 5.3.

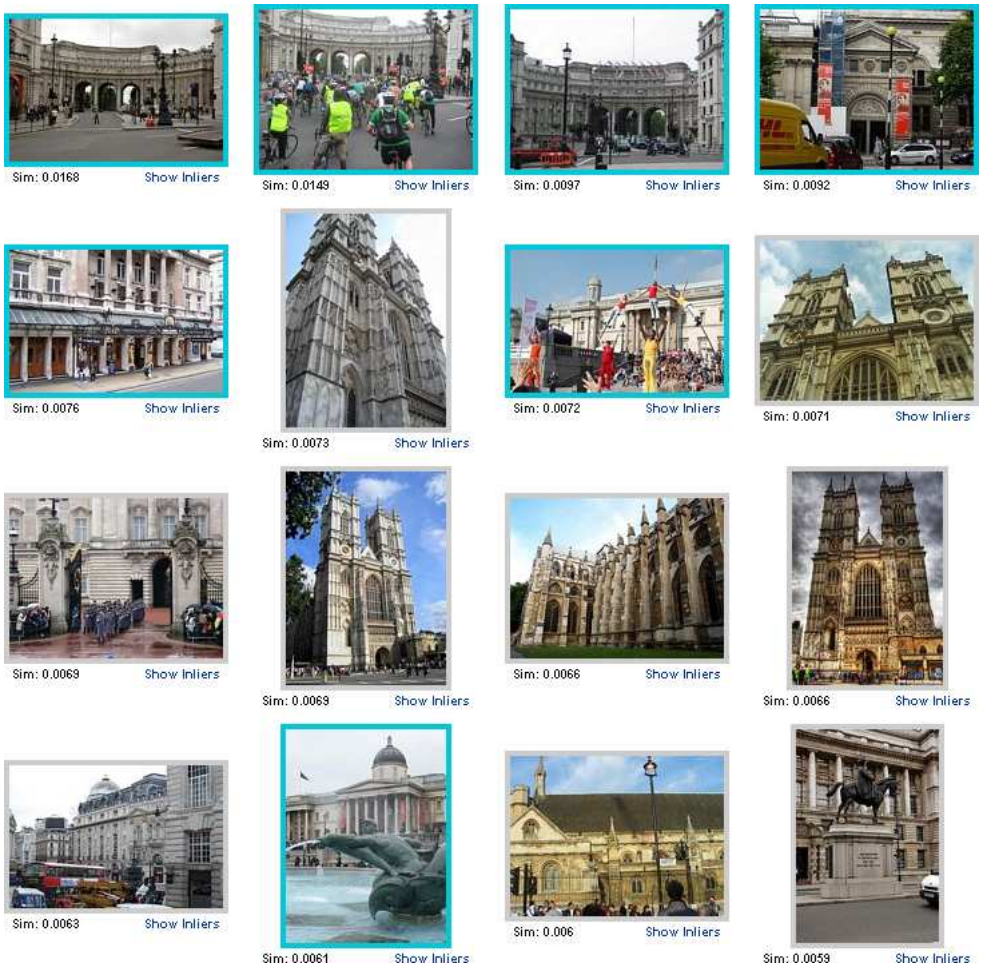
Αναλύθηκαν επίσης οι χαρακτηριστικές λέξεις – τα tags – των φωτογραφιών, λέξεις με τις οποίες οι χρήστες που ανεβάζουν τις φωτογραφίες στη συλλογή τις χαρακτηρίζουν. Έτσι, μαζί με τις εικόνες που επιστρέφονται στο χρήστη και την εκτίμηση της γεωγραφικής θέσης της εικόνας που έθεσε ως ερώτημα, επιστρέφονται επίσης και οι συχνότερες χαρακτηριστικές λέξεις των φωτογραφιών που

[1] `http://www.flickr.com`



Σχήμα 4.20: Ο χάρτης με την εκτίμηση της θέσης όπου τραβήχτηκε η φωτογραφία του ερωτήματος (δεξιά). Επίσης κάτω από την εικόνα φαίνονται και τα συχνότερα tags τωνεικόνων που επιστράφηκαν.

επιστράφηκαν ως κοντινότερες (σχήμα 4.20, κάτω από την εικόνα).



Σχήμα 4.21: Αποτελέσματα αναζήτησης. Οι εικόνες που εκτιμάται ότι τραβήχτηκαν στο ίδιο μέρος με την εικόνα του ερωτήματος έχουν ένα γαλάζιο περίβλημα.

Κεφάλαιο 5

Γραφικό περιβάλλον αναζήτησης

5.1 Εισαγωγή

Το γραφικό περιβάλλον είναι ιδιαίτερα σημαντικό σε ένα σύστημα ανάκτησης εικόνων, καθώς η αλληλεπίδραση του χρήστη με αυτό είναι συνεχής. Πρέπει λοιπόν να είναι καλαίσθητο και λειτουργικό, όπως επίσης και εύκολο στη χρήση, έτσι ώστε ο χρήστης να οδηγείται σε σωστά αποτελέσματα ανάκτησης γρήγορα, ακόμα και αν έχει μονάχα τις πλέον βασικές γνώσης χειρισμού.

Το γραφικό περιβάλλον που δημιουργήθηκε για να στεγάσει τις διάφορες τεχνικές ανάκτησης που αναπτύχθηκαν ήταν σε μορφή σελίδας ίντερνετ, γραμμένο σε γλώσσα php.

Η γενική του άποψη φαίνεται στο σχήμα 5.1, στο οποίο φαίνεται η αρχική σελίδα για την συλλογή εικόνων από το Caltech[1]. Ο χρήστης με ένα απλό κλικ σε μια φωτογραφία εκτελεί ερώτημα στη βάση με αυτή. Τα αποτελέσματα ανάκτησης για το ερώτημα με την μηχανή του παραπάνω σχήματος, φαίνεται στο σχήμα 5.2. Ανάλογα με την εκάστοτε αναζήτηση το περιβάλλον μεταβάλλεται, όπως για παράδειγμα όταν ο χρήστης θέλει να ανεβάσει δικιά του φωτογραφία για να ξεκινήσει ερώτημα ανάκτησης, όπως θα δούμε στη συνέχεια.

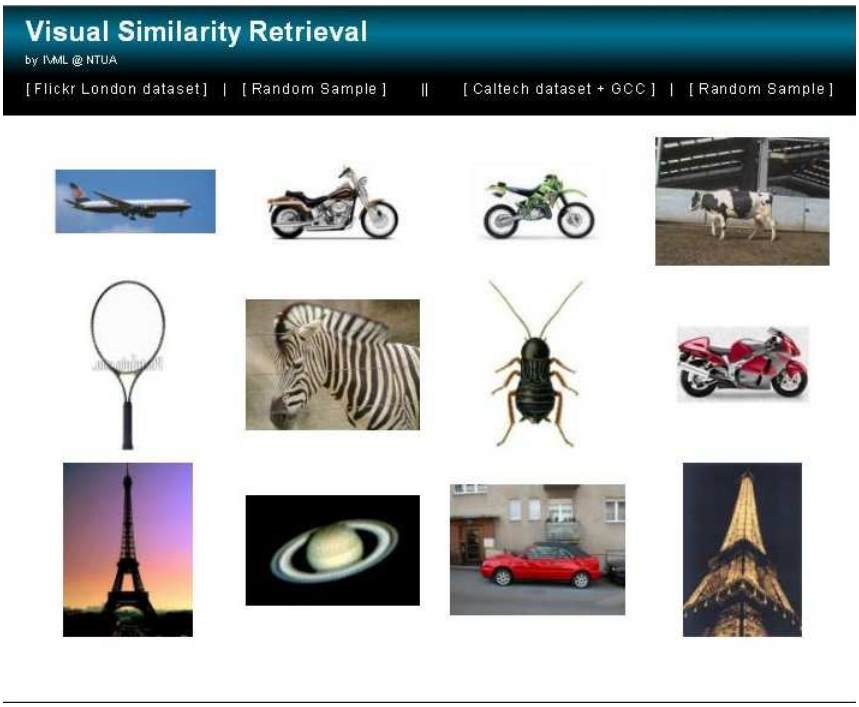
Ένα άλλο παράδειγμα φαίνεται στο σχήμα 5.3, όπου στην συλλογή αυτή με εικόνες από το flickr[2] φαίνεται δίπλα από την εικόνα του ερωτήματος στο πάνω μέρος και ο χάρτης με την γεωγραφική θέση των εικόνων που επιστράφηκαν. Η γεωγραφική θέση των εικόνων (Geo-tag) είναι ένα μεταδεδομένο που παρέχεται στο σύστημα από τη ίδια τη συλλογή.

5.2 Επιλογές βαρών στην ανάκτηση εικόνων με χαρακτηριστικά του MPEG7

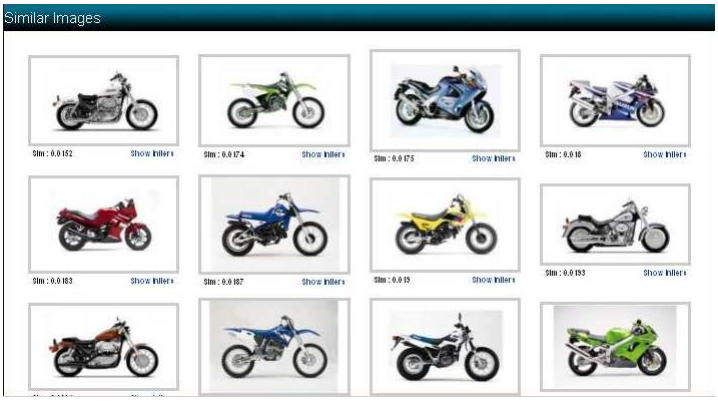
Στο σχήμα 5.4 φαίνεται το γραφικό περιβάλλον που εμφανίζεται όταν ο χρήστης ανεβάσει μια δικιά του εικόνα για ερώτημα. Σε παρόμοιο γραφικό περιβάλλον καταλήγει και ο χρήστης που διαλέγει μια από της εικόνες της συλλογής για ερώτημα, αλλά διαλέγει από το μενού την επιλογή «προηγμένη αναζήτηση» (βλέπε ενότητα 2.3.1). Εδώ μπορούν να καθοριστούν βάρη για κάθε έναν από

[1] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

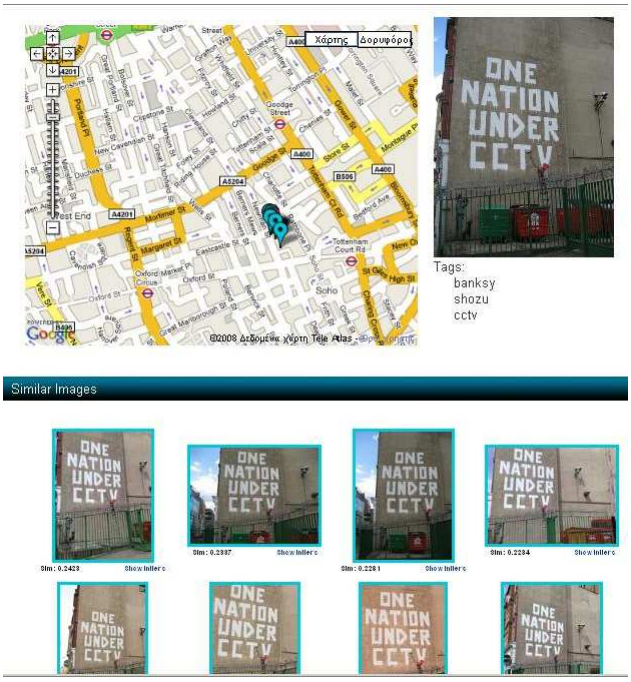
[2] http://www.flickr.com/



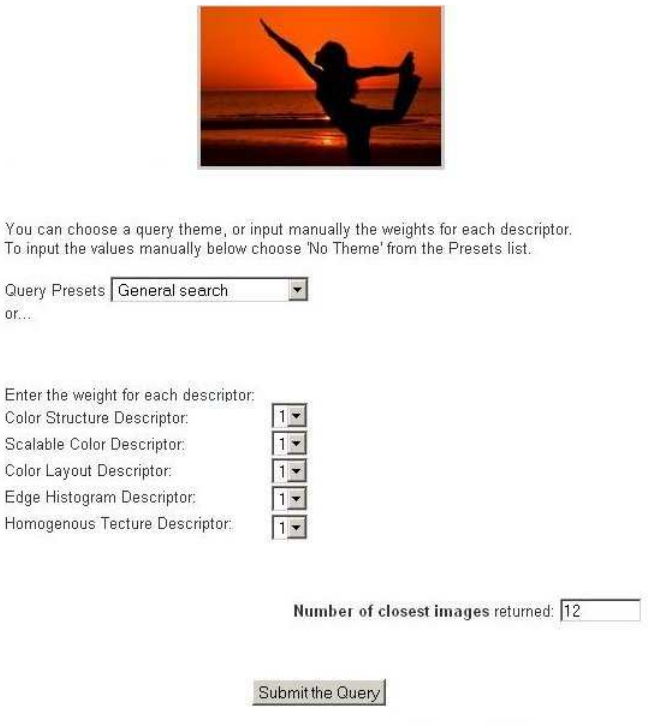
Σχήμα 5.1: Η αρχική σελίδα για την συλλογή του Caltech.



Σχήμα 5.2: Αποτελέσματα για ερώτημα με εικόνα μηχανής.



Σχήμα 5.3: Αποτελέσματα για ερώτημα με γνωστό graffiti του Λονδίνου. Στην εικόνα του ερωτήματος δίπλα, φαίνονται σημειωμένες στον χάρτη οι θέσεις των εικόνων που επιστράφηκαν από το σύστημα καθώς και μια εκτίμηση μέσω αυτών της θέσης το (βλέπε ενότητα 4.6).



Σχήμα 5.4: Μενού επιλογής των βαρών για τους περιγραφείς στην προηγμένη αναζήτηση με εικόνα ανεβασμένη από τον χρήστη.

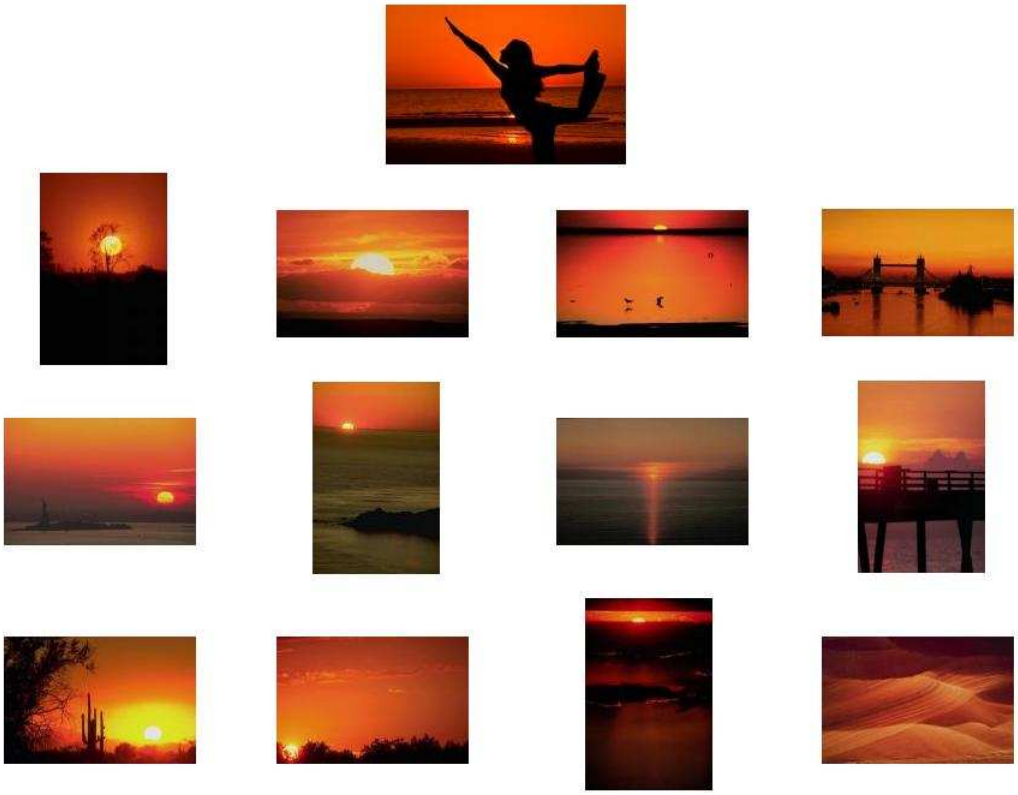
τους πέντε MPEG-7 περιγραφείς που διατίθενται, που παρουσιάζονται αναλυτικότερα στην ενότητα 2.2. Όπως φαίνεται και από το σχήμα, ο χρήστης μπορεί είτε να δώσει χειροκίνητα ένα βάρος για κάθε περιγραφέα είτε να επιλέξει από μια σειρά από *προτάσεις ερωτήματος* (*query presets*). Σε αυτές υπάρχουν έτοιμες επιλογές βαρών, όπου σε κάποια, για παράδειγμα, δίνεται βάρος μόνο στους χρωματικούς περιγραφείς ή σε κάποια άλλη μηδενίζονται όλοι οι χρωματικοί περιγραφείς.

Στο σχήμα 5.5 παρουσιάζονται τα αποτελέσματα από ερώτημα ανάκτησης με εικόνα του χρήστη, στην συλλογή εικόνων με χαρακτηριστικά που εξάγονται από ολόκληρη την εικόνα και ισορροπημένα βάρη περιγραφέων.

5.3 Γραφικό περιβάλλον λεπτομεριών στην ανάκτηση με χαρακτηριστικά από περιοχές των εικόνων

Παράλληλα με την υλοποίηση του συστήματος ανάκτησης, αναπτύχθηκε και ένα εποπτικό γραφικό περιβάλλον, στο οποίο μπορεί να δει ο εξελιγμένος χρήστης – γνώστης των τεχνολογιών περισσότερα στοιχεία για το σύστημα. Το περιβάλλον αυτό είναι επίσης σε μορφή σελίδων του ίντερνετ.

Στην ανάκτηση με χαρακτηριστικά που εξάγονται από περιοχές της εικόνας, θα ήταν ιδιαίτερα

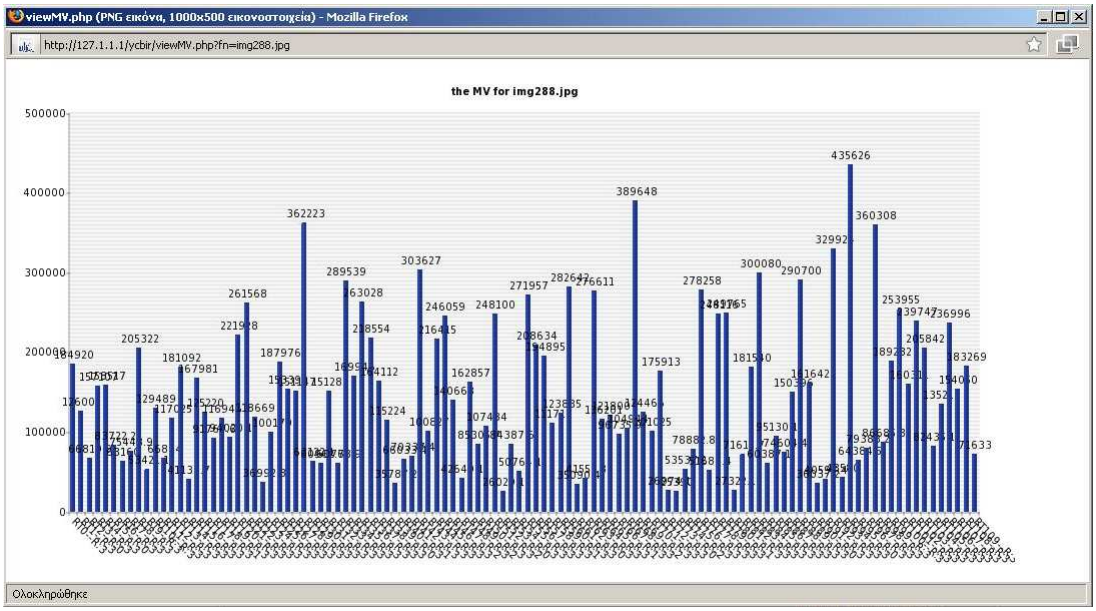


Σχήμα 5.5: Αποτελέσματα ανάκτησης με ισσορόπημένα βάρη στην εικόνα ηλιοβασιλέματος του χρήστη.

χρήσιμο να μπορεί να δει κανείς μερικά επιπλέον στοιχεία για την κάθε εικόνα. Στην αναζήτηση με χαρακτηριστικά από περιοχές για παράδειγμα μπορεί κανείς να δει μια εικόνα και τις εξαγόμενες από την κατάτμηση περιοχές της. Δίνεται επίσης η δυνατότητα να δει κάποιος το διάνυσμα αναπαράστασης για μια εικόνα (ενότητα 3.5), σε μορφή γραφήματος όπως φαίνεται στο σχήμα 5.6. Στον άξονα x βρίσκονται οι τύποι περιοχής (110 στην προκειμένη περίπτωση). Πάνω στον άξονα των τύπων περιοχής, αναγράφεται σε κάθε έναν από αυτούς και ο αριθμός της κοντινότερης σε αυτόν περιοχής.

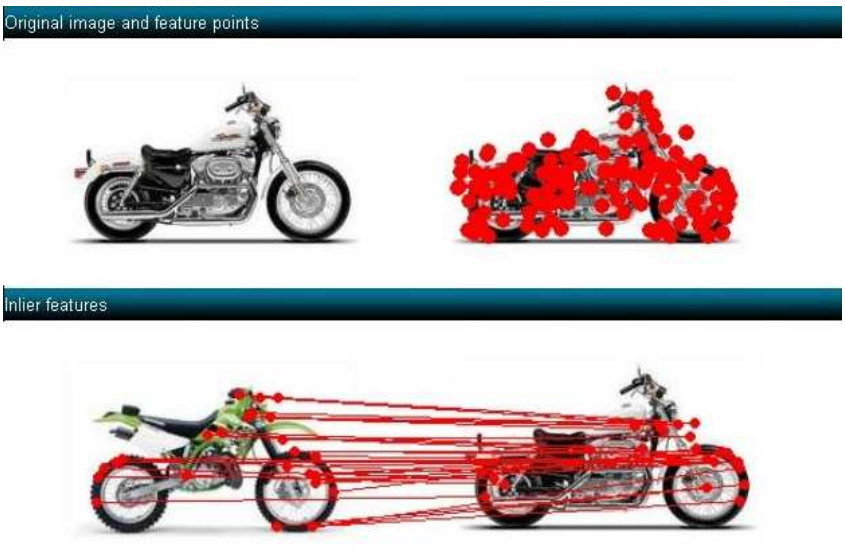
5.4 Γραφικό περιβάλλον λεπτομεριών στην ανάκτηση με χαρακτηριστικά από σημεία ενδιαφέροντος των εικόνων

Στο γραφικό περιβάλλον αποσφαλμάτωσης του συστήματος ανάκτησης εικόνων με χαρακτηριστικά εξαγόμενα από σημεία ενδιαφέροντος των εικόνων, μπορεί ο χρήστης που ενδιαφέρεται να δει, σε αναλογία με την ανάκτηση από περιοχές, μια από τις εικόνες με τα σημεία ενδιαφέροντος της τυπω-



Σχήμα 5.6: Το διάνυσμα αναπαράστασης της εικόνας img288.jpg από την συλλογή.

μένα καθώς και το διάνυσμα αναπαράστασης της (ενότητα 4.3.2). Επίσης σημαντική βοήθεια για την αποσφαλμάτωση του ελέγχου γεωμετρίας με RANSAC (ενότητα 4.5.1) είναι να μπορεί κα δεί κανείς το ποια σημεία έχουν αντιστοιχία σύμφωνα με τον εξαγόμενο από τον RANSAC, μετασχηματισμό. Μια άποψη του γραφικού αυτού περιβάλλοντος φαίνεται στο σχήμα 5.7.



Σχήμα 5.7: Το γραφικό περιβάλλον αποσφαλμάτωσης για ανάκτηση από σημεία ενδιαφέροντος. Φαίνεται η εικόνα με τα σημεία της καθώς και οι αντιστοιχίες από τον έλεγχο γεωμετρικής συνεκτικότητας με RANSAC.

Κεφάλαιο 6

Πειράματα και αξιολόγηση

6.1 Εισαγωγή

Η αξιολόγηση των συστημάτων ανάκτησης εικόνων είναι μια ιδιαίτερα δύσκολη αλλά παράλληλα και ουσιώδης διαδικασία. Μέσω των αποτελεσμάτων και της ανάδρασης που εξάγεται από αυτή μπορεί να μετρηθεί η απόδοση και ως εκ τούτου να κριθεί η βιωσιμότητα του συστήματος εμπορικά και κυρίως ερευνητικά.

Σαν εμπορικό προϊόν, ένα σύστημα ανάκτησης πρέπει, για να βρει εφαρμογή, να παράγει τα επιθυμητά αποτελέσματα στον πελάτη. Έτσι, η εξαγωγή θετικών αποτελεσμάτων είναι επιβεβλημένη και απαιτούμενη.

Σαν προϊόν έρευνας, μια ολοκληρωμένη τεχνική ανάκτησης εικόνων συνδυάζει ένα ιδιαίτερα ευρύ φάσμα επιμέρους τεχνικών. Από τεχνικές εξαγωγής σημείων και χαρακτηριστικών μέχρι τεχνικές δεικτοδότησης και ταιριάσματος εικόνων. Συνδυάζει βάσεις δεδομένων και συλλογές εικόνων σε μια ολοκληρωμένη οντότητα, τα αποτελέσματα, είτε συνολικά είτε επιμέρους, της οποίας είναι όχι μόνο μετρήσιμα, αλλά και οπτικώς εξαγόμενα. Σύμφωνα με αυτή την λογική, εκτός από την αξιολόγηση συνολικά του συστήματος αυτού καθ' αυτού σαν μια συνδυαστική και πολύπλοκη ερευνητική κατασκευή, το σύστημα μπορεί να χρησιμοποιηθεί και ως βάση για αξιολόγηση (*testbed*) άλλων επιμέρους τεχνικών, όπως για παράδειγμα τεχνικές συσταδοποίησης, εξαγωγής σημείων ενδιαφέροντος ή δεικτοδότησης.

6.2 Μέτρα αξιολόγησης

Τα μέτρα αξιολόγησης τα δανείζεται η ανάκτηση εικόνων από το γενικότερο και γειτονικό τομέα της ανάκτησης πληροφοριών. Υπάρχει η ανάγκη να χρησιμοποιηθούν κάποια αντικειμενικά μέτρα για την αξιολόγηση της απόδοσης των ανιχνευτών τα οποία θα εφαρμοστούν στις εικόνες που επιστρέφονται μετά την ανάκτηση. Τις προϋποθέσεις που πρέπει να πληρεί ένα καθολικό μέτρο σύγκρισης συγκεντρώνει και περιγράφει ο Datta [Datta et al., 2008].

Δύο γνωστά μέτρα που δανείζονται από την ανάκτηση πληροφοριών, εφαρμόσιμα σε προβλήματα σαν και το συγκεκριμένο, είναι το *μέτρο ακρίβειας* (*precision*) και το *μέτρο ανάκτησης* (*recall*). Ορίζοντας σαν $|\cdot|$ το πλήθος των στοιχείων ενός συνόλου, τα μέτρα precision και recall δίνονται

από τους τύπους (6.1) και (6.2) αντίστοιχα.

$$P_i = \frac{|D_c \bigcap G_c|}{|D_c|}, \quad c = 1 \dots N_C \tag{6.1}$$

$$R_i = \frac{|D_c \bigcap G_c|}{|G_c|}, \quad c = 1 \dots N_C \tag{6.2}$$

όπου N_C είναι ο αριθμός των περιεχόμενων στη συλλογή εννοιών.

Με απλά λόγια το πρώτο μέτρο είναι το ποσοστό των εικόνων που επιστράφηκαν από το σύστημα οι οποίες περιέχουν μια ή περισσότερες από τις έννοιες που περιείχε η εικόνα του ερωτήματος, προς όλες τις εικόνες που επιστράφηκαν στον χρήστη. Το δεύτερο μέτρο, το μέτρο ανάκτησης, μας δίνει το ποσοστό των εικόνων που επιστράφηκαν και περιέχουν μια ή περισσότερες από τις έννοιες που περιείχε η εικόνα του ερωτήματος, προς όλες τις εικόνες που την απεικονίζουν.

Από τα παραπάνω φαίνεται καθαρά ότι για την αξιολόγηση απαιτείται a priori γνώση των επιθυμητών αποτελεσμάτων. Ως *σύνολο δεδομένης αλήθειας* (*ground-truth*) ονομάζεται η πληροφορία για το σημασιολογικό περιεχόμενο των εικόνων, δηλαδή η επιθυμητή έξοδος της ανάκτησης. Περιλαμβάνει δηλαδή πληροφορία για το ποια έννοια απεικονίζεται σε κάθε μία από τις εικόνες.

Για κάθε έννοια (*concept*) προς εντοπισμό c το σύνολο των εικόνων της συλλογής μπορεί να χωριστεί σε δύο υποσύνολα. Το ένα θα αποτελείται από τις εικόνες που απεικονίζουν την έννοια c και το άλλο από εκείνες που δεν την απεικονίζουν. Τα δύο αυτά υποσύνολα συμβολίζονται με $G_c i$ και $\bar{G}_c$ αντίστοιχα:

$$G_c : \{i \in I : \quad c \in C(i)\} \tag{6.3}$$

$$\bar{G}_c : \{i \in I : \quad c \notin C(i)\} \tag{6.4}$$

όπου $C(i)$ είναι το σύνολο των εννοιών εκείνων οι οποίες απεικονίζονται στην εικόνα i και I είναι το σύνολο των εικόνων της συλλογής. Σύμφωνα με τον Smeulders η όλη διαδικασία αξιολόγησης είναι προβληματική καθώς το σύνολο δεδομένης αλήθεια δεν μπορεί να οριστεί πάντα ολοκληρωτικά και με ακρίβεια [Smeulders et al., 2000]. Στις περισσότερες περιπτώσεις, αυτό το ζήτημα ανάγεται στο πρόβλημα του ορισμού της τέλειας ανάκτησης, κάτι που αποδεικνύεται ότι είναι εξαρτώμενο από την περίσταση.

Εύκολα καταλαβαίνει κανείς ότι αν το σύστημα ανάκτησης επιστρέφει στον χρήστη ολόκληρη την συλλογή εικόνων, ταξινομημένη βέβαια ως προς την απόσταση από την εικόνα του ερωτήματος, τότε αυτά τα μέτρα δεν έχουν ιδιαίτερη σημασία, μιας και δεν εμπεριέχουν καθόλου πληροφορία για την θέση των σωστών και λάθος ανακτώμενων εικόνων. Το μέτρο ανάκτησης σε αυτή τη περίπτωση θα είναι πάντα 1 γιατί όλες οι εικόνες που περιέχουν τις έννοιες της εικόνας του ερωτήματος θα έχουν επιστραφεί, μιας και αποτελούν υποσύνολο των εικόνων της συλλογής. Το μέτρο ακρίβειας θα είναι επίσης σταθερό και ίσο με τον λόγο του αριθμού των εικόνων που περιέχουν τις έννοιες της εικόνας του ερωτήματος προς το σύνολο των εικόνων της συλλογής. Απαιτείται λοιπόν ένα μέτρο το οποίο θα λαμβάνει υπόψη και την θέση ανάκτησης των «σωστών» και «λάθος» σύμφωνα με το ερώτημα ανακτήσεων.

Γι' αυτό τον λόγο εφαρμόζεται το *μέτρο της μέσης ακρίβειας* (*Average Precision*), το οποίο ορίζεται από την σχέση (6.6. Για τις ανάγκες υπολογισμού της ορίζεται με την σχέση (6.5) η

ακρίβεια εάν πάρουμε τις m κοντινότερες στην εικόνα του ερωτήματος εικόνες.

$$p_m = \frac{1}{m} \sum_{k=1}^{m} x_k \tag{6.5}$$

Μπορεί να ερμηνευθεί ότι το p_m είναι η μέση τιμή των $x_1, x_2, \ldots, x_m$, και συμβολίζεται με $\bar{x}_m$, όπου $x_i \in \{0, 1\}$ και είναι 0 αν δεν υπάρχει στην εικόνα στη θέση i έννοια της εικόνας του ερωτήματος και 1 αν υπάρχει. Η μέση ακρίβεια ορίζεται σαν η μέση τιμή των ακριβειών μετά από κάθε σχετικό χαρακτηριστικό καρέ που συναντιέται στην λίστα. Μαθηματικά εκφράζεται από τον τύπο (6.6), ο οποίος αντιστοιχεί στην μέση ακρίβεια για την έννοια c με παράθυρο N.

$$AP_c^N = \frac{1}{|G_c|} \sum_{j=1}^{N} x_j p_j = \frac{1}{|G_c|} \sum_{j=1}^{N} \frac{x_j}{j} \sum_{k=1}^{j} x_k \tag{6.6}$$

Όπως γίνεται εύκολα αντιληπτό, το μέτρο της μέσης ακρίβειας επηρεάζεται πολύ περισσότερο από τις σωστές ανακτήσεις που επιστρέφονται σε υψηλή θέση, από αυτές που επιστρέφονται σε χαμηλότερες. Αν πάρουμε τον αριθμητικό μέσο όρο από μέσες ακρίβειες για μια σειρά από ερωτήματα τότε παίρνουμε το μέτρο *mean Average Precision* (mAP) το οποίο είναι δύσκολο να μεταφραστεί εύηχα καθώς δεν υπάρχουν δύο λέξεις για τον μέσο όρο στην ελληνική γλώσσα σε αντίθεση με την αγγλική (mean - average).

Για την δημιουργία των γραφημάτων και την εξαγωγή καμπυλών ακρίβειας, πρέπει να υπάρχει και μια παράμετρος που μεταβάλλεται. Αυτή είναι στις περισσότερες περιπτώσεις το *παράθυρο των επιστρεφόμενων αποτελεσμάτων* γνωστό και ως *scope*, αλλά όχι πάντα. Άλλες φορές σημαντικότατος παράγοντας για την απόδοση είναι το μέγεθος του οπτικού λεξικού, άρα ο αριθμός των συστάδων κατά τη διαδικασία της συσταδοποίησης [Jing et al., 2004]. Αλλού προτείνεται να μεταβάλλεται ο αριθμός των «σωστών» εικόνων ως προς το ερώτημα (size of the *embedding*) που υπάρχουν στη συλλογή εικόνων [Huijsmans & Sebe, 2005].

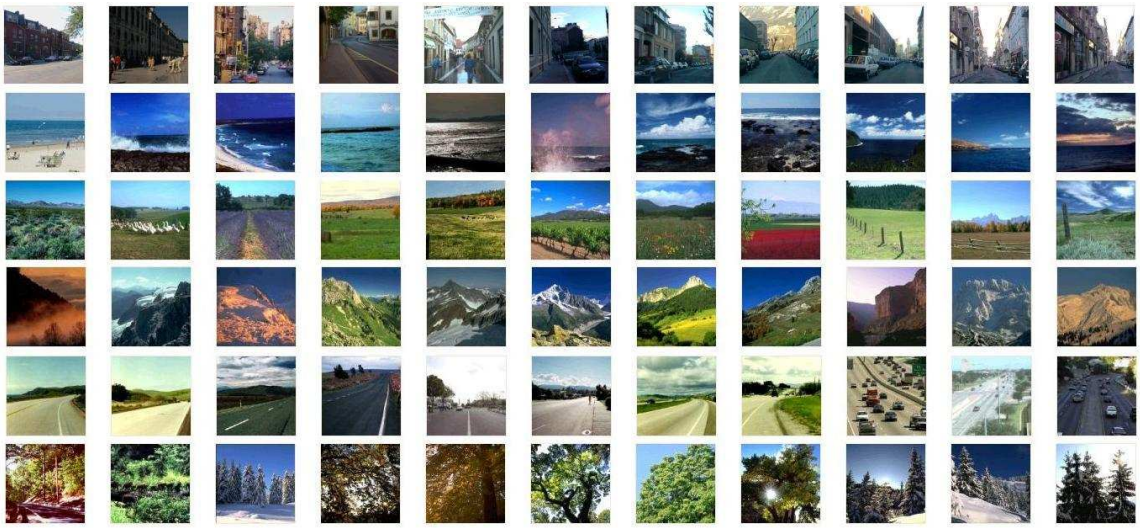
6.3 Συλλογές εικόνων

Για την πληρότητα της παρουσίασης και αξιολόγησης των τεχνικών που αναπτύχθηκαν, χρησιμοποιήθηκαν πολλές διαφορετικές συλλογές εικόνων. Οι συλλογές είναι διαφορετικές για του οπτικούς περιγραφείς του MPEG-7 και διαφορετικές για τους τοπικούς περιγραφείς γύρω από σημεία ενδιαφέροντος, καθώς κάθε τεχνική αναπαριστά σωστά διαφορετικούς τύπους εικόνων.

Στην πρώτη περίπτωση, όπου τα οπτικά χαρακτηριστικά ήταν οι περιγραφείς του MPEG-7, τον πρώτο λόγο έχουν εικόνες με φυσικά τοπία και αραιό σημασιολογικό περιεχόμενο, που περιέχουν μάλιστα έννοιες αναγνωρίσιμες ως προς την υφή και το χρώμα. Διαλέχτηκαν γι' αυτό το λόγο εικόνες από την τεράστια συλλογή της Corel, δείγμα των οποίων φαίνεται στο σχήμα 6.1. Για να δοκιμαστεί η τεχνική με οπτικά χαρακτηριστικά εξαγόμενα από περιοχές της εικόνας μετά από κατάτμηση, χρησιμοποιήθηκε επίσης η συλλογή του Torralba[1] με εικόνες από τις κατηγορίες: ακτή, δάσος, αυτοκινητόδρομος, και δρόμος πόλης. Δείγματα φαίνονται στο σχήμα 6.2.

[1] `http://people.csail.mit.edu/torralba/code/spatialenvelope/`

Σχήμα 6.1: Δείγμα από το υποσύνολο της συλλογής εικόνων Corel που χρησιμοποιήθηκε.



Σχήμα 6.2: Δείγμα από τη συλλογή εικόνων του Torralba που χρησιμοποιήθηκε.

Στη δεύτερη περίπτωση όπου χρησιμοποιήθηκαν τοπικοί περιγραφείς δοκιμάστηκαν διάφορες συλλογές, για να ελεγχθεί κατά πόσο μπορούν οι τεχνικές αυτές να χρησιμοποιηθούν αποδοτικά για γενικές συλλογές εικόνων, όπως για παράδειγμα για την ανάκτηση εικόνων με συγκεκριμένα αντικείμενα ή κτίρια, σε περιβάλλοντα με μεγάλη αλλαγή συνθηκών (αλλαγές κλίμακας, περιγραφή, αλλαγές στον φωτισμό).

Οι συλλογές εικόνων στις οποίες έγιναν πειράματα ανάκτησης με χαρακτηριστικά εξαγόμενα από σημεία ενδιαφέροντος είναι οι εξής:

- Ένα υποσύνολο από εικόνες του Caltech 101[2], δείγμα των οποίων φαίνεται στο σχήμα 6.6. Αποτελείται από 1025 εικόνες διαλεγμένες από 8 κατηγορίες: ζέβρες, αυτοκίνητα, ρακέτες

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Σχήμα 6.3: Δείγμα από τη συλλογή εικόνων UKBench που χρησιμοποιήθηκε.

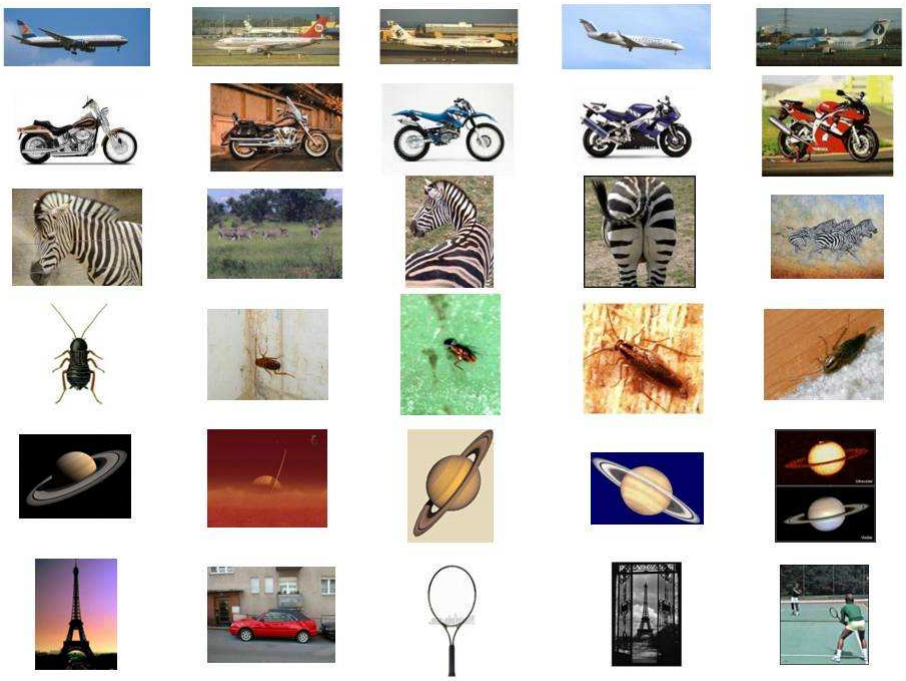
Σχήμα 6.4: Δείγμα από τη συλλογή εικόνων Zurich Buildings που χρησιμοποιήθηκε.

τένις, έντομα, αγελάδες, αεροπλάνα, μηχανές και πύργος του Άιφελ.

- Τη συλλογή εικόνων Zurich Buildings[3] (σχήμα 6.4) που αποτελείται από 1005 εικόνες, οι οποίες απεικονίζουν 201 κτίρια της Ζυρίχης.
- Τη πρόσφατη συλλογή εικόνων Oxford Buildings[4] (σχήμα 6.5) η οποία περιέχει 5063 εικόνες, που απεικονίζουν κτίρια της Οξφόρδης και αρκετές εικόνες-«σκουπίδια», δηλαδή εικόνες που δεν απεικονίζουν κτίρια και προκαλούν θόρυβο στην ανάκτηση.

[3] `http://www.vision.ee.ethz.ch/datasets/index.en.html`
[4] `http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/index.html`

Σχήμα 6.5: Δείγμα από τη συλλογή εικόνων Oxford Buildings που χρησιμοποιήθηκε.



Σχήμα 6.6: Δείγμα από το υποσύνολο της συλλογής εικόνων Caltech που χρησιμοποιήθηκε.

- Τη συλλογή εικόνων UKbench[5] (σχήμα 6.3) η οποία αποτελείται από 10000 εικόνες, που

[5] http://www.vis.uky.edu/~stewe/ukbench/

απεικονίζουν 2500 αντικείμενα, το καθένα από τέσσερις οπτικές γωνίες.

Σε όλες τις παραπάνω συλλογές εικόνων που χρησιμοποιήθηκαν, είτε υπήρχε είτε δημιουργήθηκε *σύνολο δεδομένης αλήθειας* (*ground-truth*). Προσδιορίστηκαν δηλαδή σε κάθε εικόνα οι περιεχόμενες έννοιες για να μπορεί να αξιολογηθεί η ανάκτηση με τα μέτρα που παρουσιάστηκαν στην ενότητα 6.2.

6.4 Αποτελέσματα αναζήτησης με περιγραφείς από ολόκληρη την εικόνα

Μιας και οι τεχνικές που αναπτύχθηκαν στο κεφάλαιο 2 είναι τεχνικές παλαιότερων ετών, πλήρως αναλυμένες και αξιολογημένες στη διεθνή βιβλιογραφία, δεν κρίθηκε χρήσιμο να αξιολογηθούν αυτές εκτενώς. Επιλέχτηκε γι' αυτό την σκοπό ένα υποσύνολο της συλλογής του Corel.

Οι έννοιες που υπήρχαν στη συλλογή με φυσικές εικόνες είναι οι εξής: χιόνι, βλάστηση, ηλιοβασίλεμα, δρόμος, έρημος, καταρράκτης, παραλία, καθώς και εικόνες που προσπαθούν να «μπερδέψουν» την ανάκτηση φυσικών εικόνων, όπως: νυχτερινές φωτογραφίες πόλης, εικόνες εσωτερικών χώρων, υποθαλάσσιες φωτογραφίες και εικόνες με γάτες.

Το υποσύνολο του Corel που επιλέχτηκε, περιέχει μεγάλη ποικιλία στην μορφή εμφάνισης των παραπάνω κύριων εννοιών. Για παράδειγμα, μια εικόνα στης οποίας μια άκρη εμφανίζεται μια συστάδα βλάστησης, θεωρείται ως εικόνα που περιέχει τη συγκεκριμένη έννοια. Αυτό έχει ως αποτέλεσμα μειωμένα συνολικά ποσοστά του μέτρου ανάκτησης για κάθε έννοια. Τα ποσοστά αυτά προκύπτουν ως μέσος όρος των μέτρων ανάκτησης για ερωτήματα με όλες τις εικόνες της συλλογής που περιέχουν την έννοια. Έτσι παρότι σε ένα μεγάλο υποσύνολο των εικόνων μπορεί να εμφανίζεται αξιοθαύμαστη ανάκτηση, αρκούν μερικές εικόνες οι οποίες δεν περιέχουν την συγκεκριμένη έννοια καθαρά ή σε μεγάλο ποσοστό για να μειωθούν τα συνολικά ποσοστά ανάκτησης.

Στο σχήμα 6.7 φαίνονται τα μέτρα ανάκτησης για τις έννοιες χιόνι, ηλιοβασίλεμα και βλάστηση.

Υπολογίστηκε επίσης το *μέτρο της μέσης ακρίβειας* για κάθε έννοια, που στην προκειμένη περίπτωση αποτελείται από τον μέσο όρο των μέτρων όλων των εικόνων της κάθε έννοιας (*mean Average Precision* - mAP) και τα αποτελέσματα παρουσιάζονται στον πίνακα 6.1.

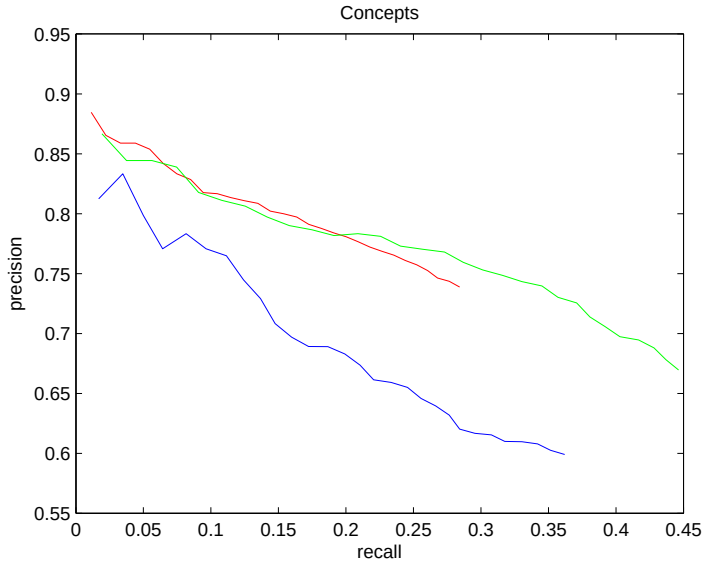
Η συνολική μέση τιμή των μέτρων για όλες τις έννοιες είναι:

Παρατηρείται εύκολα ότι οι έννοιες που εντοπίζονται πιο εύκολα από τους περιγραφείς του MPEG-7 εξαγόμενους από ολόκληρη την εικόνα είναι η βλάστηση, το ηλιοβασίλεμα ή η έρημος, έννοιες με κυρίαρχη θέση στο περιβάλλον τους. Με επιτυχία φαίνεται ότι ξεχωρίζει η συγκεκριμένη τεχνική και τις εικόνες εσωτερικού χώρου από τις υπόλοιπες.

Από την άλλη, έννοιες όπως ο δρόμος και ο καταρράκτης που μπορεί να έχουν χαρακτηριστική υφή και χρώμα, δεν καταλαμβάνουν όμως συνήθως μεγάλο ποσοστό των εικόνων, δεν είναι και ιδιαίτερα αναγνωρίσιμες από τα συνολικά οπτικά χαρακτηριστικά. Μικρό ποσοστό ανάκτησης έχει κ η έννοια «παραλία» καθώς διέπεται από μεγάλη οπτική ποικιλομορφία.

Καθώς η περιγραφή εξάγεται συνολικά από ολόκληρη την εικόνα είναι λογικό να εντοπίζονται καλύτερα οι έννοιες οι οποίες είναι οι μοναδικές στις εικόνες τους.

Με αλλαγή των βαρών μπορεί να παρατηρήσει κανείς την σημασία ύπαρξης ταυτόχρονα περιγραφέων υφής και χρώματος. Αν επαναλάβουμε τις μετρήσεις για την έννοια ηλιοβασίλεμα, μια έννοια που κανείς θα περίμενε ότι ο εντοπισμός της θα στηριζόταν πιο πολύ στους χρωματικούς περιγραφείς,



Σχήμα 6.7: Τα μέτρα ανάκτησης για τις έννοιες χιόνι (μπλέ), ηλιοβασίλεμα (πράσινο) και βλάστηση (κόκκινο).

εξαλείφοντας τους περιγραφείς υφής τότε τα αποτελέσματα είναι υποδεέστερα, όπως φαίνεται και στο σχήμα 6.8.

6.5 Αποτελέσματα αναζήτησης με περιγραφείς από περιοχές

6.5.1 Ανάκτηση φυσικών εικόνων στη συλλογή Corel

Για να εξετάσουμε την τεχνική σύμφωνα με την οποία οι περιγραφείς MPEG-7 εξάγονται από περιοχές της εικόνας μετά από κατάτμηση, χρησιμοποιήσαμε και πάλι τη συλλογή φυσικών εικόνων του Corel. Παρότι πιο σύνθετη, η τεχνική αυτή που περιγράφεται στο κεφάλαιο 3 έδωσε σε μερικές έννοιες της συλλογής χειρότερα αποτελέσματα, από ότι οι περιγραφείς που εξήχθησαν από ολόκληρη την εικόνα. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι οι εικόνες της συγκεκριμένης συλλογής περιγράφουν έννοιες υψηλού επιπέδου οι οποίες έχουν ομοιόμορφη υφή και χρώμα και κατανέμονται συνολικά σε ολόκληρη την εικόνα.

Το μέτρο μέσης ακρίβειας (mean Average Precision) για κάθε μια από τις έννοιες της συλλογής φαίνεται στον πίνακα 6.3. Στο συγκεκριμένο πείραμα χρησιμοποιήθηκε θησαυρός 70 τύπων περιοχής, καθώς αυτό βρέθηκε ότι βγάζει στις περισσότερες περιπτώσεις τα καλύτερα αποτελέσματα. Επίσης κρατήθηκαν στο διάνυσμα αναπαράστασης οι 3 κοντινότεροι τύποι περιοχών για κάθε περιοχή κατάτμησης.

Από τα αποτελέσματα παρατηρεί κανείς ότι ενώ το μέτρο της μέσης ανάκτησης έπεσε σε έννοιες όπως χιόνι, έρημος και ηλιοβασίλεμα, αυξήθηκε όμως στις πιο σύνθετες έννοιες δρόμος και βλάστηση που συνυπάρχουν στις εικόνες με άλλες έννοιες.

Έννοια	mAP
Χιόνι	0.525
Ηλιοβασίλεμα	0.633
Βλάστηση	0.609
Δρόμος	0.149
Έρημος	0.595
Καταρράκτης	0.262
Παραλία	0.244
Εικόνες από εσωτερικό χώρο	0.781

Πίνακας 6.1: Το μέτρο της μέσης ακρίβειας (mean Average Precision) για τις έννοιες της συλλογής φυσικών εικόνων.

mAP
0,4750

Πίνακας 6.2: Το συνολικό μέτρο της μέσης ακρίβειας (mean Average Precision) για τις όλες έννοιες της συλλογής φυσικών εικόνων.

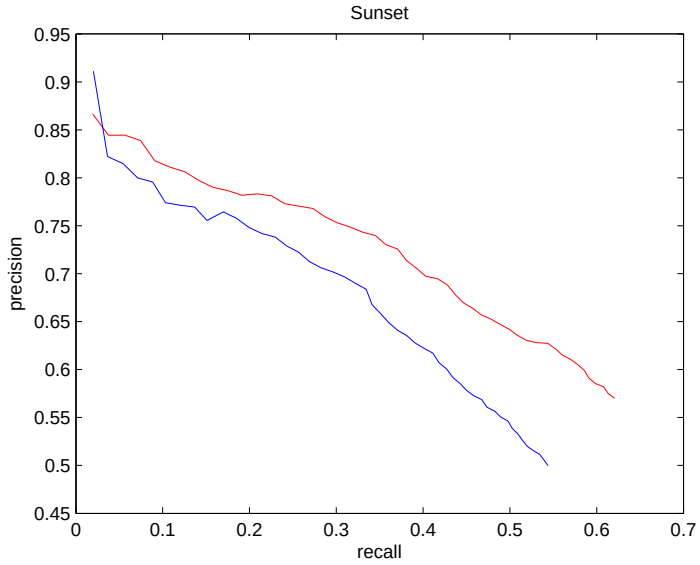
Έγιναν πειράματα με διαφορετικά μεγέθη θησαυρού, δηλαδή διαφορετικό αριθμό τύπων περιοχών που εξάγονται από την διαδικασία της συσταδοποίησης. Η εξέλιξη του μέτρου μέσης ακρίβειας για την έννοια βλάστηση όσο αλλάζει ο αριθμός των τύπων περιοχής στον θησαυρό φαίνεται στο σχήμα 6.9(a). Η καμπύλη ακρίβειας - ανάκτησης για την έννοια έρημος παρουσιάζεται στο σχήμα 6.9(b). Στη συγκεκριμένη συλλογή είναι λογικό να έχουμε μικρά ποσοστά ανάκτησης λόγω της ύπαρξης μεγάλου αριθμού εικόνων της έννοιας στην συλλογή.

6.5.2 Ανάκτηση φυσικών εικόνων στη συλλογή Torralba

Για να φανούν οι δυνατότητες της τεχνικής έγιναν πειράματα και στην απαιτητική συλλογή για κατηγοριοποίηση σκηνής του Torralba. Όπως εύκολα φαίνεται και στα δείγματα της συλλογής (σχήμα 6.2) οι εικόνες είναι ιδιαίτερα ανομοιογενείς και πιο σύνθετες από του Corel. Εδώ, οι περιγραφείς MPEG-7 εξαγόμενοι από ολόκληρη την εικόνα έδωσαν απογοητευτικά αποτελέσματα. Ας τονιστεί επίσης, ότι οι έννοιες που αξιολογήθηκαν ως προς την ανάκτηση είχαν πλέον μεγαλύτερη συσχέτιση. Μετρήσαμε τιμές του μέτρου μέσης ακρίβειας για τις έννοιες ακτή, δάσος, αυτοκινητόδρομος και δρόμος πόλης και ταυτόχρονα στη συλλογή υπήρχαν επίσης οι κατηγορίες επαρχία (open country), βουνό και εικόνες πόλης. Τα μέτρα μέσης ακρίβειας για αυτές τις έννοιες δίνονται στον πίνακα 6.4

Παρατηρούμε ότι η τεχνική αποδίδει πάρα πολύ καλά στις εικόνες με την έννοια ακτή καθώς και στις εικόνες δάσους, ενώ η απόδοση πέφτει αισθητά στις έννοιες αυτοκινητόδρομος και δρόμος πόλης. Δύο είναι οι κύριοι λόγου που μπορούν να εξηγήσουν αυτό το φαινόμενο.

Πρώτον, στην περίπτωση της ακτής και του δάσους η κατάτμηση είναι ευκολότερη και δίνει



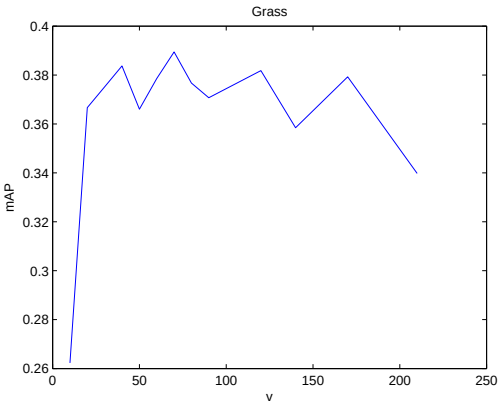
Σχήμα 6.8: Το μέτρο ανάκτησης για την έννοια ηλιοβασίλεμα με όλους τους περιγραφείς (κόκκινο) και μόνο με τους περιγραφείς χρώματος (μπλέ). Οι περιγραφείς εξάγονται από ολόκληρη την εικόνα.

πιο σωστά τις περιοχές για την εξαγωγή οπτικών χαρακτηριστικών. Καθώς η κατάτμηση είναι η βάση για την περαιτέρω ανάλυση και εξαγωγή χαρακτηριστικών, η σωστή κατάτμηση δίνει περιοχές με «σημασία» δηλαδή περιοχές χαρακτηριστικές και ομοιογενείς. Στις εικόνες δρόμου πόλης και αυτοκινητόδρομου από την άλλη η κατάτμηση είναι συνήθως «κακή» από την άποψη ότι δεν βγάζει περιοχές με «νόημα» δηλαδή περιοχές με κυρίαρχες εμπεριεχόμενες έννοιες. Ξεκινώντας από τέτοιου είδους περιοχές, οι περιγραφείς MPEG-7 δεν δίνουν αντιπροσωπευτικές και διακριτές απεικονίσεις των εικόνων με αποτέλεσμα τελικά την μείωση της απόδοσης.

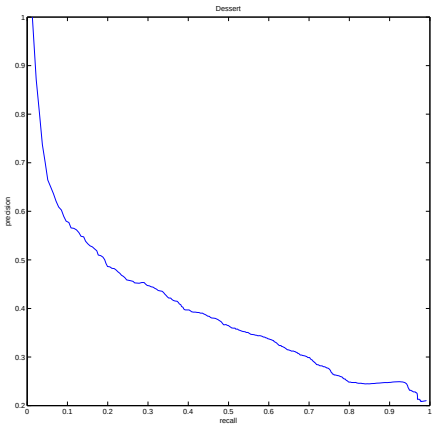
Ο δεύτερος λόγος που μπορεί να εξηγήσει την πτώσης της απόδοσης της τεχνικής στις έννοιες αυτοκινητόδρομος και δρόμος πόλης είναι η ιδιαίτερα υψηλή συσχέτιση των δύο κατηγοριών. Την ανάκτηση έμπλεκαν επίσης και οι εικόνες πόλης που περιείχε η συλλογή, με αποτέλεσμα την σύγχυση αυτών των τριών εννοιών κατά την ανάκτηση.

Την εξέλιξη του μέσου μέτρου ανάκτησης σε σχέση με τον αριθμό των κοντινότερων τύπων περιοχής που κρατούνται στο διάνυσμα αναπαράστασης (ν) φαίνεται στο σχήμα 6.10 για τις έννοιες ακτή και δρόμος πόλης. Όπως ανεβαίνει το ν ανεβαίνει και το μέτρο μέσης ανάκτησης για τον δρόμο πόλης καθώς με περισσότερους τύπους περιοχών τέτοιες έννοιες περιγράφονται καλύτερα. Το αντίθετο συμβαίνει για την έννοια ακτή.

Με την συγκεκριμένη τεχνική έγιναν τέλος πειράματα σε υποσύνολο της συλλογής UKBench (σχήμα 6.3) με απογοητευτικά αποτελέσματα. Η ανομοιομορφία των αντικειμένων οδήγησε σε τύπους περιοχών μη αντιπροσωπευτικούς και πολύ χαμηλά ποσοστά ανάκτησης. Το μέσο μέτρο ανάκτησης αναγράφεται στον πίνακα 6.6.



(a)



(b)

Σχήμα 6.9: Το μέτρο μέσης ανάκτησης της έννοιας βλάστηση για διάφορα μεγέθη θησαυρού και οι καμπύλη ακρίβειας - ανάκτησης για την έννοια έρημος.

Έννοια	**mAP**
Χιόνι	0.238364
Ηλιοβασίλεμα	0.296623
Βλάστηση	0.6198
Δρόμος	0.173303
Έρημος	0.430011
Καταρράκτης	0.198477
Παραλία	0.182617

Πίνακας 6.3: Το μέτρο της μέσης ακρίβειας (mean Average Precision) για τις έννοιες της συλλογής φυσικών εικόνων του Corel.

Έννοια	α	β	γ	δ
Ακτή	0.360817	0.459348	**0.659348**	0.451611
Δάσος	0.165435	**0.265507**	0.233548	0.177838
Δρόμος Πόλης	0.127017	0.0553128	0.0876473	**0.139885**
Αυτοκινητόδρομος	0.0877063	0.0619179	0.102364	**0.144771**

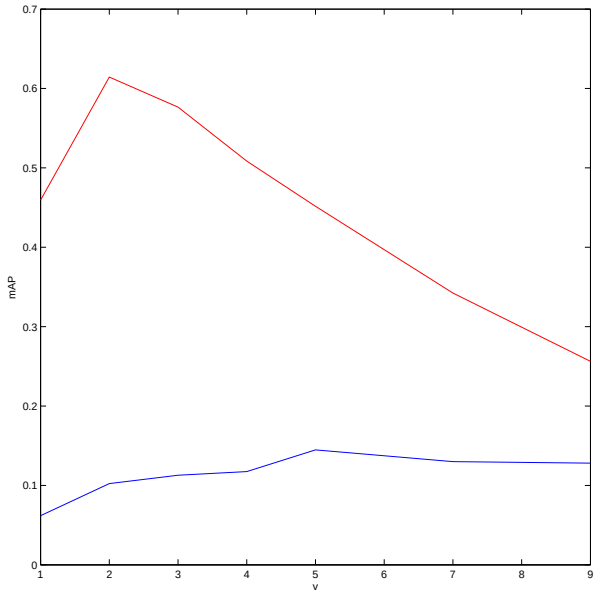
Πίνακας 6.4: Το μέτρο της μέσης ακρίβειας (mean Average Precision) για τις αντίστοιχες έννοιες σε διάφορες περιπτώσεις μεγέθους θησαυρού και αριθμού κρατούμενων κοντινότερων τύπων περιοχών. Στήλη α: 150 Τύποι περιοχής και ν=4, στήλη β: 270 Τύποι περιοχής και ν=1, στήλη γ: 270 Τύποι περιοχής και ν=2, στήλη δ: 270 Τύποι περιοχής και ν=5.

Από τα παραπάνω αποτελέσματα είναι φανερή η ανάγκη μιας πιο εύρωστης περιγραφής των εικόνων, που θα μπορεί να αναπαραστήσει αποτελεσματικά μια εικόνα με υψηλή πυκνότητα οπτικού περιεχομένου. Προς αυτή την κατεύθυνση μελετήθηκαν οι τεχνικές με περιγραφείς εξαγόμενους γύρω από σημεία ενδιαφέροντος της εικόνας, τα αποτελέσματα των οποίων ακολουθούν στην επόμενη υποενότητα.

6.6 Αποτελέσματα αναζήτησης με τοπικούς περιγραφείς

Παρότι η ανάκτηση με τοπικά χαρακτηριστικά των εικόνων είναι το State-of-the-art στην κατηγορία της ανάκτησης σήμερα, δεν υπάρχει μια αντικειμενική συλλογή αξιολόγησης για ανάλογες εφαρμογές. Η πλειονότητα των συλλογών που δημοσιεύονται, τείνουν περισσότερο προς την ανίχνευση αντικειμένων - τόπων παρά προς την απλή θεματική - σημασιολογική ανάκτηση.

Παρόλα αυτά επιλέχτηκαν συλλογές που μπορούν να εκτιμήσουν με αξιοπιστία την απόδοση των τεχνικών ανάκτησης.



Σχήμα 6.10: Το μέτρο ανάκτησης για τις έννοιες δρόμος πόλης (μπλέ)και ακτή (κόκκινο) όσο εξελίσσεται το ν.

6.6.1 Αναζήτηση αντικειμένων της συλλογής UKBench

Η συλλογή *UKBench* αποτελείται από 10000 εικόνες, που περιέχουν 2500 αντικείμενα από 4 εικόνες στο καθένα. Σκοπός της ανάκτησης εδώ είναι να επιστραφούν σε ένα ερώτημα ως πρώτες, οι 3 εναπομείναντες εικόνες του ίδιου αντικειμένου που υπάρχουν στη συλλογή. Τα αντικείμενα ανήκουν σε μια ευρεία γκάμα, όπως φαίνεται και στο σχήμα 6.3 και πολλά από αυτά έχουν φωτογραφηθεί στο ίδιο περιβάλλον και έχουν υποστεί μεγάλη αλλαγή στην οπτική γωνία λήψης.

Παρακάτω φαίνεται ο μέσος αριθμός των ανακτημένων εικόνων του ίδιου αντικειμένου σε όλη τη συλλογή UKBench, ένα μέτρο που έχει ως μέγιστη τιμή την τιμή 4 για τέλεια ανάκτηση. Η τιμή αυτή υπολογίστηκε ως μέσος όρος των ανακτήσεων με όλες τις εικόνες της συλλογής.

Μέτρο Ανάκτησης
2.5303

Πίνακας 6.5: Ο μέσος αριθμός σωστά ανακτημένων εικόνων στις 4 πρώτες που επιστρέφονται, για όλη τη συλλογή UKBench.

Τέλεια ανάκτηση, δηλαδή επιστροφή των 4 εικόνων του αντικειμένου πάντα ως πρώτες για ερώτημα με οποιαδήποτε από αυτές, επιτεύχθηκε σε αρκετά μεγάλο αριθμό από τις εικόνες.

Στην ανάκτηση εικόνων σημαντικό είναι επίσης το μέτρο μέσης ακρίβειας που για ολόκληρη τη συλλογή παίρνει μια ιδιαίτερα αποδοτική τιμή, όπως φαίνεται στον πίνακα 6.6.

λεξικό τύπων περιοχών	οπτικό λεξικό
mAP	mAP
0.0812	0.4463

Πίνακας 6.6: Το μέσο μέτρο της μέσης ακρίβειας για όλη τη συλλογή των 2500 αντικειμένων, αριστερά για την τεχνική που περιγράφεται στο κεφάλαιο 3 και δεξιά για την τεχνική που περιγράφεται στο κεφάλαιο 4.

Το μέτρο αυτό είναι μεγαλύτερο από το αντίστοιχο μέτρο ανάκτησης που παρουσιάζεται παραπάνω ($1/2,5303 = 0,395$), κάτι που σημαίνει ότι πολλές φορές η ανάκτηση μπορεί να μην είναι τέλεια, ο χρήστης όμως σχεδόν πάντα θα βρει αυτό που ζητάει μέσα στις πρώτες N εικόνες.

6.6.2 Αναζήτηση κτιρίων στις συλλογές Zurich Buildings και Oxford Buildings

Μια πρόσφατη τάση είναι ο συνδυασμός της ανάκτησης εικόνων με τεχνικές εντοπισμού - αναγνώρισης θέσης. Για να επιτευχθεί κάτι τέτοιο απαιτούνται πρόσθετα μεταδεδομένα [Hays & Efros, 2008], όμως ως πρώτο βήμα προτάσσεται η σωστή ανάκτηση. Προς αυτή την κατεύθυνση στρέφονται δύο γνωστές συλλογές που στοχεύουν στην αναγνώριση κτιρίων. Η μια αποτελείται από κτίρια της Οξφόρδης[6] (σχήμα 6.5) και η άλλη από κτίρια της Ζυρίχης (σχήμα 6.4). Πειράματα, έγιναν και στις δύο συλλογές, με ιδιαίτερα θετικά αποτελέσματα στη δεύτερη περίπτωση.

Παρακάτω φαίνεται ο μέσος αριθμός των ανακτημένων εικόνων του ίδιου κτιρίου στις πέντε πρώτες θέσεις, για όλη τη συλλογή Zurich Buildings. Το μέτρο αυτό έχει ως μέγιστη τιμή την τιμή 5 για τέλεια ανάκτηση, καθώς υπάρχουν 5 εικόνες του κάθε κτιρίου στη συλλογή. Συνολικά υπάρχουν 201 κτίρια. Η τιμή αυτή υπολογίστηκε ως μέσος όρος των ανακτήσεων με όλες τις εικόνες της συλλογής.

Μέτρο Ανάκτησης
3.7114

Πίνακας 6.7: Ο μέσος αριθμός σωστά ανακτημένων εικόνων στις 5 πρώτες που επιστρέφονται, για όλη τη συλλογή Zurich Buildings.

Πρακτικά αυτό σημαίνει ότι κατά μέσο όρο σε κάθε ανάκτηση οι 3.72 πρώτες εικόνες που επιστρέφονται είναι εικόνες του ίδιου κτιρίου. Τέλεια ανάκτηση, δηλαδή επιστροφή των 5 εικόνων του κτιρίου πάντα ως πρώτες για ερώτημα με οποιαδήποτε από αυτές, επιτεύχθηκε σε μεγάλο ποσοστό των κτιρίων και αυτό φαίνεται και από την υψηλή μέση τιμή του μέτρου μέσης ανάκτησης στον πίνακα 6.6.2.

[6] http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/index.html

Συλλογή	mAP
Zurich Buildings	0.796
Zurich Buildings συλλογή 2	0.7299

Έγιναν επίσης πειράματα για να βρεθεί το βέλτιστο μέγεθος οπτικού λεξικού. Το μέτρο μέσης ανάκτησης για τη συλλογή κτιρίων της Ζυρίχης φαίνεται στον πίνακα 6.6.2. Παρατηρείται ότι το βέλτιστο μέγεθος λεξικού είναι στις 350 οπτικές λέξεις.

Μέγεθος Λεξικού	mAP
30	0.2610
105	0.6457
350	0.7954
454	0.6443
640	0.3381

Πίνακας 6.8: Το μέσο μέτρο ανάκτησης όσο μεταβάλλεται το μέγεθος του οπτικού λεξικού.

Σε μια προσπάθεια να προσομοιώσει η αναζήτηση στη συλλογή την αναζήτηση εικόνων στο διαδίκτυο, προστέθηκαν στη συλλογή Zurich Buildings και εικόνες από άλλες συλλογές, άσχετες μεταξύ τους. Σε κάθε έννοια - κτίριο ο λόγος των σωστών αποτελεσμάτων προς τα σωστά είναι πλέον αρκετά μικρός: 10^{-4}. Η συλλογή αυτή ονομάστηκε *Zurich Buildings συλλογή 2* και το μέσο μέτρο ανάκτησης και σε αυτή την περίπτωση φαίνεται στον πίνακα 6.6.2. Παρατηρείται ότι παρά την εισαγωγή πολλών παραπλανητικών εικόνων η ανάκτηση στις περισσότερες φορές παρέμεινε σε υψηλά επίπεδα.

Τα πειράματα εκτελέστηκαν επίσης και για τα ερωτήματα που θέτει η συλλογή Oxford Buildings. Η συλλογή αποτελείται από 5063 εικόνες από τις οποίες 624 είναι ωφέλιμες (52 εικόνες σε κάθε ένα από 12 κτίρια της πόλης της οξφόρδης) και οι υπόλοιπες εικόνες παραπλανητικές για την ανάκτηση. Η αξιολόγηση στη συγκεκριμένη βάση προυποθέτει 5 ερωτήματα για κάθε ένα από τα 12 κτίρια και τέλεια ανάκτηση θεωρείται εκείνη κατά την οποία έχουμε τις 52 εικόνες του εκάστοτε κτιρίου να επιστρέφονται πρώτες και για τα 5 ερωτήματα.

Στο σύνολο δεδομένης αλήθειας που παρέχεται, εκτός από το ποιες εικόνες θα χρησιμοποιηθούν ως ερώτημα, ορίζονται επίσης και οι συντεταγμένες ενός ορθογωνίου στην εικόνα - ερώτημα μέσα στο οποίο περιέχεται το κτίριο. Στο σχήμα 6.5 φαίνονται σαν κίτρινα ορθογώνια. Στην τεχνική μας δεν περιλαμβάνεται η επιλογή υποσυνόλου της εικόνας ως ερώτημα για την ανάκτηση, και έτσι αυτή η ιδιαίτερα χρήσιμη πληροφορία αγνοήθηκε στα πειράματα που έγιναν. Αυτό έχει ως αποτέλεσμα αρκετά χαμηλότερα ποσοστά ανάκτησης από άλλες πρόσφατες αντίστοιχες επιστημονικές δουλείες και συμβαίνει διότι στην προκειμένη περίπτωση δεν έρχονται σαν ερώτημα μόνο σημεία του κτιρίου αλλά και σημεία από ολόκληρη την εικόνα.

Τα μέτρα μέσης ανάκτησης για κάθε ένα από τα κτίρια της συλλογής φαίνονται στον πίνακα 6.9. Είναι ο μέσος όρος των μέτρων μέσης ανάκτησης (mean Average Precision - mAP) για τα πέντε ερωτήματα για κάθε κτίριο.

Κτίριο	**mAP**
All souls	0.1207
Ashmolean	0.1156
Balliol	0.0967
Bodleian	0.0885
Christ church	0.1465
Cornmarket	0.1902
Hertford	0.3950
Keble	0.1944
Magdalen	0.0310
Pitt Rivers	0.0667
Radcliffe Camera	0.1505

Πίνακας 6.9: Τα μέσα μέτρα ανάκτησης για τα 11 κτίρια της συλλογής του πανεπιστημίου της Οξφόρδης.

6.6.3 Ανάκτηση στη συλλογή Caltech

Από την τεράστια συλλογή του Caltech επιλέχτηκαν εικόνες από τις κατηγορίες: Ζέβρα, ήλιοι, αεροπλάνο, αγελάδα, μηχανή, κατσαρίδα, ζέβρα και πλανήτης. Επίσης προστέθηκαν εικόνες από τις έννοιες αυτοκίνητα, ρακέτες τένις, και πύργος του Άιφελ. Το υποσύνολο στο οποί έγιναν πειράματα, περιείχε τελικά 1025 εικόνες από τις παραπάνω κατηγορίες. Η δυσκολία στην ανάκτηση εικόνων που περιέχουν τόσο γενικές έννοιες είναι μεγάλη για τεχνικές χωρίς επιβλεπόμενη μάθηση. Δεν αναζητούμε πλέον ένα συγκεκριμένο αντικείμενο, αλλά μια γενικότερη σημασιολογική έννοια που μπορεί να περιέχεται σε όλη η μέρος της οθόνης, μια ή παραπάνω από μια φορές. Το σχήμα 6.6 φανερώνει την ποικιλομορφία στις επιλεγμένες εικόνες.

Στον πίνακα 6.10 είναι συγκεντρωμένα τα μέσα μέτρα ανάκτησης για τις έξι έννοιες. Παρατηρούνται υψηλές τιμές ανάκτησης στις ζέβρες, στις μηχανές, τα αεροπλάνα και τους πλανήτες. Για τις ζέβρες βοηθητικός παράγοντας ήταν η χαρακτηριστική όψη που μοιράζεται ο πληθυσμός. Η έννοια πλανήτης έχει, όπως και η έννοια αεροπλάνο, σχετικά μικρή ανομοιομορφία στη συλλογή ενώ στις μηχανές θετικό για την ανάκτηση ήταν το γεγονός ότι οι περισσότερες φωτογραφίες ήταν τραβηγμένες σε λευκό φόντο. Σε αυτή την περίπτωση, όλα τα σημεία της εικόνας είναι ωφέλιμα σημεία της έννοιας.

Έννοια	**mAP**
Ήλιος	0.121425
Αεροπλάνο	0.299128
Αγελάδα	0.216099
Μηχανή	0.374383
Κατσαρίδα	0.223029
Ζέβρα	0.413085
Πλανήτης	0.281523

Πίνακας 6.10: Τα μέσα μέτρα ανάκτησης για έξι από τις έννοιες της συλλογής του πανεπιστημίου Caltech.

Κεφάλαιο 7

Συμπεράσματα και μελλοντικές τάσεις

Στα πλαίσια αυτής της διπλωματικής εργασίας μελετήθηκε η βιβλιογραφία γύρω από το θέμα της αναζήτησης εικόνων με βάση το οπτικό τους περιεχόμενο. Σε αυτή περιλαμβάνονται επιστημονικές δημοσιεύσεις για συστήματα ανάκτησης συνολικά αλλά και για ένα μεγάλο εύρος παράπλευρων τεχνικών τις οποίες χρησιμοποιούνται κατά την αναζήτηση, όπως εξαγωγή και περιγραφή οπτικών χαρακτηριστικών, συσταδοποίηση, έλεγχο γεωμετρίας, βάσεις δεδομένων, δεικτοδότηση και εύρεση πλησιέστερου γείτονα.

Με βάση το θεωρητικό αυτό πλαίσιο, υλοποιήθηκαν διάφορες τεχνικές για κάθε στάδιο της ανάκτησης εικόνων, με τελικό στάδιο την δημιουργία μιας εφαρμογής ιστού για αναζήτηση εικόνων με βάση οπτικά χαρακτηριστικά. Το ολοκληρωμένο σύστημα, μπορεί να χρησιμοποιηθεί στο μέλλον ως βάση (testbed) για τον έλεγχο απόδοσης νέων τεχνικών, σε όλα τα επιμέρους στάδια της αναζήτησης.

Τέλος, υλοποιήθηκε μια προηγμένη εφαρμογή ανάκτησης με τις αποδοτικότερες από τις τεχνικές που υλοποιήθηκαν σύμφωνα με την οποία είναι δυνατός ο προσδιορισμός της τοποθεσίας μιας εικόνας μέσω αναζήτησης σε μια συλλογή εικόνων με μεταδεδομένα θέσης (ενότητα 4.6).

Η κάθε τεχνική από αυτές που μελετήθηκαν, περιέγραφε και μια διαφορετική προσέγγιση για την εξαγωγή και την περιγραφή του οπτικού περιεχομένου των εικόνων. Η εικόνα αντιμετωπίστηκε στην αρχή συνολικά ως οντότητα, έπειτα περιγράφηκε από τις περιοχές κατάτμησής της και έφτασε να περιγράφεται από περιοχές γύρω από σημεία ενδιαφέροντος της. Κάθε μια από τις τεχνικές υπερτερούσε σε κάποιο επίπεδο επί των άλλων, οπότε δεν μπορούμε να μιλήσουμε γενικά για μια καλύτερη τεχνική. Μπορούμε όμως πάντα να επιλέξουμε την κατάλληλη για την εκάστοτε περίσταση.

Στο κεφάλαιο 2 υλοποιήθηκε μια τεχνική κατά την οποία εξάγονται οπτικοί περιγραφείς χρώματος και υφής του προτύπου MPEG-7 από ολόκληρη την εικόνα. Η τεχνική αυτή έχει πλέον εγκαταληφθεί από την προηγούμενη δεκαετία και θεωρείται ξεπερασμένη. Παρ' όλα αυτά, η ανάκτηση εννοιών όπως ηλιοβασίλεμα και χιόνι με αυτή την τεχνική σε συλλογή από φωσικές εικόνες δίνει ανταγωνιστικά αποτελέσματα. Είναι γεγονός πάντως ότι περιγραφείς εξαγόμενοι από ολόκληρη την εικόνα δεν μπορούν να περιγράψουν αποτελεσματικά εικόνες με πολύπλοκο σημασιολογικό περιεχόμενο, την πλειονότητα δηλαδή των εικόνων που συναντά κανείς.

Σύμφωνα με την τεχνική που παρουσιάζεται στο κεφάλαιο 3 οι εικόνες υπόκεινται πρώτα σε κατάτμηση και έπειτα χρωματικοί περιγραφείς και περιγραφείς υφής εξάγονται από τις περιοχές. Γίνεται με αυτό τον τρόπο προσπάθεια να διαιρεθεί η εικόνα στις περιεχόμενες έννοιες, κάτι που προσεγγίζεται ικανοποιητικά στις εικόνες με αραιό σημασιολογικό περιεχόμενο, όπως φαίνεται από

τα αποτελέσματα σε πειράματα με φυσικές εικόνες. Έπειτα δημιουργείται λεξικό τύπων περιοχών με αποτέλεσμα να μειώνεται δραστικά η υπολογιστική πολυπλοκότητα στην σύγκριση εικόνων. Δίνεται η δυνατότητα να μεγαλώσει το μέγεθος των συλλογών. Εξέλιξη σε παρόμοιες τεχνικές μπορεί να υπάρξει, με την χρησιμοποίηση διαφόρων τεχνικών κατάτμησης και διαφόρων οπτικών περιγραφέων. Επίσης χρήσιμο θα ήταν να μπορέσει να κρατηθεί πληροφορία για την θέση των περιοχών μέσα στην εικόνα.

Σε συλλογές με μεγαλύτερη πολυπλοκότητα, βέβαια οι περιοχές κατάτμησης δεν μπορούν να περιγράψουν αποτελεσματικά το σημασιολογικό διάκοσμο των εικόνων με αποτέλεσμα να αποτυγχάνει η ανάκτηση. Γι' αυτό μελετήθηκαν τεχνικές που βασίζονται σε περιγραφείς εξαγόμενους από σημεία ενδιαφέροντος της εικόνας, όπως περιγράφεται και στο κεφάλαιο 4. Τα πειραματικά αποτελέσματα έδειξαν ότι οι τοπικοί αυτοί περιγραφείς είναι πράγματι εύρωστοι σε αλλαγές κλίμακας και σε περιστροφή, με αποτέλεσμα υψηλά ποσοστά ανάκτησης σε εικόνες κτιρίων από διαφορετικές οπτικές γωνίες. Υψηλά επίπεδα ανάκτησης αντικειμένων επιτεύχθηκαν και σε γενικές συλλογές, κάτι που αποδεικνύει την γενικότητα της μεθόδου.

Η δημιουργία μεγάλων οπτικών λεξικών φέρνει την όλη τεχνική σε άμεση αντιστοιχία με την αναζήτηση λέξεων σε μηχανές αναζήτησης του διαδικτύου. Η δημιουργία ενός ανάλογου συστήματος αναζήτησης εικόνων σε μεγέθη διαδικτύου είναι ένα ανοικτό ερευνητικό θέμα που απασχολεί μερίδα της επιστημονικής κοινότητας σήμερα. Η ανάκτηση εικόνων από συλλογές εκατομμυρίων εικόνων, συλλογές οι οποίες ήδη είναι διαθέσιμες στο διαδίκτυο, σε ελάχιστο χρόνο, απαιτεί περαιτέρω ανάπτυξη σε μια πληθώρα επιμέρους τομέων, όπως της δεικτοδότησης, της σύγκρισης και της ανάλυσης των εικόνων. Σαν παράδειγμα αναφέρεται η πρόσφατη τάση για δενδρικής δομής οπτικά λεξικά ή λεξικά που εμπεριέχουν και πληροφορία θέσης των σημείων στην εικόνα.

Τέλος, με δεδομένη μια αποδοτική ανάκτηση με βάση το οπτικό περιεχόμενο, μπορεί να προχωρήσει κανείς στο επόμενο βήμα, όπου μέσω μιας τεράστιας παγκόσμιας βάσης, δισεκατομμυρίων εικόνων κλίμακας διαδικτύου, μια εικόνα θα αρκεί για την ακριβή αναγνώριση της θέσης στην οποία αυτή τραβήχτηκε. Επίσης, θα μπορεί να είναι δυνατή η πλήρως αυτοματοποιημένη διαχείριση και ταξινόμηση των προσωπικών συλλογών φωτογραφιών των χρηστών.

Το μέλλον της αυτοματοποιημένης αναζήτησης εικόνων, ή πολυμέσων γενικότερα, με την γιγαντοποίηση των αντίστοιχων συλλογών στο διαδίκτυο αναμένεται τουλάχιστον ενδιαφέρον.

Βιβλιογραφία

[Abbasi et al., 1999] Abbasi, S., Mokhtarian, F., & Kittler, J. (1999). Curvature scale space image in shape similarity retrieval. *Multimedia Syst.*, 7(6), 467–476.

[Adamek et al., 2005] Adamek, T., OǒConnor, N., & Murphy, N. (2005). Region-based segmentation of images using syntactic visual features. In *Workshop on Image Analysis for Multimedia Interactive Services* Montreux, Switzerland.

[Aksoy & Haralick, 1998] Aksoy, S. & Haralick, R. M. (1998). Textural features for image database retrieval. In *In Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries, in conjunction with CVPR'98* (pp. 45–49).

[Armitage & Enser, 1997] Armitage, L. H. & Enser, P. G. B. (1997). Analysis of user need in image archives. *Journal of Information Science*, 23(4), 287–299.

[Assfalg et al., 2002] Assfalg, J., Bimbo, A. D., & Pala, P. (2002). Three-dimensional interfaces for querying by example in content-based image retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 8(4), 305–318.

[Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., & Gool, L. V. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3), 346–359.

[Bennstrom & Casas, 2004] Bennstrom, C. & Casas, J. (2004). Binary-partition-tree creation using a quasi-inclusion criterion. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on* (pp. 259–264).

[Bentley, 1975] Bentley, J. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9), 509–517.

[Bertini et al., 2005] Bertini, E., Cali', A., Catarci, T., Gabrielli, S., & Kimani, S. (2005). Interaction-based adaptation for small screen devices. In L. Ardissono, P. Brna, & A. Mitrovic (Eds.), *User Modeling*, volume 3538 of *Lecture Notes in Computer Science* (pp. 277–281).: Springer.

[Bimbo & Pala, 1997] Bimbo, A. D. & Pala, P. (1997). Visual image retrieval by elastic matching of user sketches. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(2), 121–132.

[Bishop, 2006] Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.

[Bober, 2001] Bober, M. (2001). Mpeg-7 visual shape descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6), 716–719.

[Buijs & Lew, 1999] Buijs, J. M. & Lew, M. S. (1999). Visual learning of simple semantics in ImageScape. *Lecture notes in computer science*, 1614, 131–??

[Carson et al., 1997] Carson, C., Belongie, S., Greenspan, H., & Malik, J. (1997). Region-based image querying. In *CAIVL '97: Proceedings of the 1997 Workshop on Content-Based Access of Image and Video Libraries (CBAIVL '97)* (pp.42). Washington, DC, USA: IEEE Computer Society.

[Carson et al., 2002] Carson, C., Belongie, S., Greenspan, H., & Malik, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(8), 1026–1038.

[Chang et al.,] Chang, S., Sikora, T., & Puri, A. Overview of the MPEG-7 standard.

[Chen et al., 2003] Chen, L.-Q., Xie, X., Fan, X., Ma, W.-Y., Zhang, H., & Zhou, H.-Q. (2003). A visual attention model for adapting images on small displays. *Multimedia Syst.*, 9(4), 353–364.

[Chen et al., 2005] Chen, Y., Wang, J., & Krovetz, R. (2005). Clue: cluster-based retrieval of images by unsupervised learning. *Image Processing, IEEE Transactions on*, 14(8), 1187–1201.

[Chen et al., 2001] Chen, Y., Zhou, X. S., & Huang, T. (2001). One-class svm for learning in image retrieval. *Image Processing, 2001. Proceedings. 2001 International Conference on*, 1, 34–37 vol.1.

[Chum et al., 2007] Chum, O., Philbin, J., Sivic, J., Isard, M., & Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil*.

[Chum et al., 2008] Chum, O., Philbin, J., & Zisserman, A. (2008). Near duplicate image detection: min-hash and tf-idf weighting. In *Proceedings of the British Machine Vision Conference*.

[Conescu & Christel, 2005] Conescu, R. M. & Christel, M. G. (2005). Addressing the challenge of visual information access from digital image and video libraries. *jcdl*, 00, 69–78.

[Cullen et al., 1997] Cullen, J. F., Hull, J. J., & Hart, P. E. (1997). Document image database retrieval and browsing using texture analysis. In *ICDAR '97: Proceedings of the 4th International Conference on Document Analysis and Recognition* (pp. 718–721). Washington, DC, USA: IEEE Computer Society.

[Cunningham et al., 2004] Cunningham, S. J., Bainbridge, D., & Masoodian, M. (2004). How people describe their image information needs: a grounded theory analysis of visual arts queries. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries* (pp. 47–48). New York, NY, USA: ACM.

[Cunningham & Masoodian, 2006] Cunningham, S. J. & Masoodian, M. (2006). Looking for a picture: an analysis of everyday image information searching. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 198–199). New York, NY, USA: ACM.

[Datta et al., 2008] Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2).

[Egas et al., 1999] Egas, R., Huijsmans, D. P., Lew, M. S., & Sebe, N. (1999). Adapting k-d trees to visual retrieval. In *VISUAL '99: Proceedings of the Third International Conference on Visual Information and Information Systems* (pp. 533–540). London, UK: Springer-Verlag.

[El-Kwae & Kabuka, 2000] El-Kwae, E. A. & Kabuka, M. R. (2000). Efficient content-based indexing of large image databases. *ACM Trans. Inf. Syst.*, 18(2), 171–210.

[Fang, 1996] Fang, R. (1996). Periodicity, Directionality, and Randomness: Wold Features for Image Modeling and Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 722–733).

[Fischler & Bolles, 1981] Fischler, M. & Bolles, R. (1981). Random sample consensus - a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

[Flickner et al., 1995] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., & Yanker, P. (1995). Query by image and video content: The qbic system. *Computer*, 28(9), 23–32.

[Freidman et al., 1977] Freidman, J., Bentley, J., & Finkel, R. (1977). An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Transactions on Mathematical Software (TOMS)*, 3(3), 209–226.

[Frossyniotis et al., 2004] Frossyniotis, D., Likas, A., & Stafylopatis, A. (2004). A clustering method based on boosting. *Pattern Recognition Letters*, 25(6), 641–654.

[Gemert et al., 2008] Gemert, J. C. V., Geusebroek, J.-M., Veenman, C. J., & Smeulders, A. W. (2008). Kernel codebooks for scene categorization. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer Vision – ECCV 2008*, volume 5304 of *lncs* (pp. 696–709).: Springer.

[Greene et al., 1994] Greene, D., Parnas, M., & Yao, F. (1994). Multi-index hashing for information retrieval. In *Foundations of Computer Science, 1994 Proceedings., 35th Annual Symposium on* (pp. 722–731).

[H. Tamura & Yamawaki, 1978] H. Tamura, S. M. & Yamawaki, T. (1978). Texture features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-8, 460–473.

[Haralick et al., 1973] Haralick, R. M., Dinstein, & Shanmugam, K. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, 610–621.

[Harman et al., 1992] Harman, D., Baeza-Yates, R., Fox, E., & Lee, W. (1992). Inverted files. (pp. 28–43).

[Hartley & Zisserman, 2004] Hartley, R. I. & Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning*. Springer.

[Hays & Efros, 2008] Hays, J. & Efros, A. (2008). IM2GPS: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8).

[Hsu & Shih, 2002] Hsu, C. & Shih, M. (2002). Content-Based Image Retrieval by Interest Points Matching and Geometric Hashing. In *SPIE Photonics Asia Conference*, volume 4925 (pp. 80–90).

[Huijsmans & Sebe, 2005] Huijsmans, D. & Sebe, N. (2005). How to Complete Performance Graphs in Content-Based Image Retrieval: Add Generality and Normalize Scope. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 245–251).

[Jegou et al., 2008a] Jegou, H., Douze, M., & Schmid, C. (2008a). Hamming embedding and weak geometric consistency for large scale image search. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer Vision – ECCV 2008*, volume 5302 of *lncs* (pp. 304–317).: Springer.

[Jegou et al., 2008b] Jegou, H., Douze, M., & Schmid, C. (2008b). Hamming embedding and weak geometric consistency for large scale image search. In A. Z. David Forsyth, Philip Torr (Ed.), *European Conference on Computer Vision*, LNCS: Springer. to appear.

[Jing et al., 2004] Jing, F., Li, M., Zhang, H., & Zhang, B. (2004). An efficient and effective region-based image retrieval framework. *Image Processing, IEEE Transactions on*, 13(5), 699–709.

[Jolion, 2001] Jolion, J.-M. (2001). Feature similarity. (pp. 121–143).

[Kanade, 1998] Kanade, M. (1998). Video Skimming and Characterization through the Combination of Image and Language Understanding. In *Proceedings of the 1998 International Workshop on Content-Based Access of Image and Video Databases (CAIVD'98)* (pp.6̃1).: IEEE Computer Society Washington, DC, USA.

[Kanungo et al., 2002] Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., & Wu, A. (2002). An efficient k-means clustering algorithm - analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 881–892).

[Käster et al., 2003] Käster, T., Pfeiffer, M., & Bauckhage, C. (2003). Combining speech and haptics for intuitive and efficient navigation through image databases. In *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces* (pp. 180–187). New York, NY, USA: ACM.

[Kumar et al., 2008] Kumar, N., Zhang, L., & Nayar, S. (2008). What is a good nearest neighbors algorithm for finding similar patches in images? In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer Vision – ECCV 2008*, volume 5303 of *lncs* (pp. 364–378).: Springer.

[Lehmann et al., 2004] Lehmann, T., Gold, M., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., & Wein, B. (2004). Content-based Image Retrieval in Medical Applications. *Methods of Information in Medicine*, 43(4), 354–361.

[Lehmann et al., 2000] Lehmann, T., Wein, B., Dahmen, J., Bredno, J., Vogelsang, F., & Kohnen, M. (2000). Content-based image retrieval in medical applications: A novel multi-step approach. In *Proceedings SPIE*, volume 3972 (pp. 312–320).

[Leibe et al., 2008] Leibe, B., Leonardis, A., & Schiele, B. (2008). Robust Object Detection with Interleaved Categorization and Segmentation. *International Journal of Computer Vision*, 77(1), 259–289.

[Lew et al., 2006] Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1), 1–19.

[Li & Castelli, 1997] Li, C.-S. & Castelli, V. (1997). Deriving texture feature set for content-based retrieval of satellite image database. In *ICIP '97: Proceedings of the 1997 International Conference on Image Processing (ICIP '97) 3-Volume Set-Volume 1* (pp. 576). Washington, DC, USA: IEEE Computer Society.

[Li & Wang, 2003] Li, J. & Wang, J. (2003). Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 1075–1088).

[Lindeberg, 1994] Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*. Kluwer Academic Print on Demand.

[Lindeberg, 1998] Lindeberg, T. (1998). Feature Detection with Automatic Scale Selection. *International Journal of Computer Vision*, 30(2), 79–116.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2), 91–110.

[Ma & Manjunath, 1995] Ma, W. & Manjunath, B. (1995). A comparison of wavelet transform features for texture image annotation. *icip*, 2, 2256.

[MacQueen,] MacQueen, J. B. 1967.ôSome methods for classification and analysis of multivariate observationsŏ. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 1–297).

[Manjunath et al., 2001] Manjunath, B., Ohm, J., Vasudevan, V., & Yamada, A. (2001). Color and Texture Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), 703–715.

[McGill & Salton, 1983] McGill, M. & Salton, G. (1983). *Introduction to modern information retrieval.* McGraw-Hill.

[Mikolajczyk & Schmid, 2005] Mikolajczyk, K. & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10), 1615–1630.

[Moore, a] Moore, A. *An intoductory tutorial on kd-trees.* Technical report, Technical Report.

[Moore, b] Moore, A. K-means and hierarchical clustering - tutorial slides.

[Müller et al., 2004] Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications?clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1), 1–23.

[Nakazato & Huang, 2001] Nakazato, M. & Huang, T. (2001). 3d mars: immersive virtual reality for content-based image retrieval. *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, (pp. 44–47).

[Neubeck & Van Gool, 2006] Neubeck, A. & Van Gool, L. (2006). Efficient Non-Maximum Suppression. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3.

[Ng & Sedighian, 1996] Ng, R. T. & Sedighian, A. (1996). Evaluating multidimensional indexing structures for images transformed by principal component analysis. volume 2670 (pp. 50–61).: SPIE.

[Olson, 1995] Olson, C. (1995). Parallel algorithms for hierarchical clustering. *Parallel Computing*, 21(8), 1313–1325.

[Omohundro, 1987] Omohundro, S. (1987). Efficient algorithms with neural network behavior. *Complex Systems*, 1(2), 273–347.

[Pelleg & Moore, 1999] Pelleg, D. & Moore, A. (1999). Accelerating exact k-means algorithms with geometric reasoning. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 277–281). New York, NY, USA: ACM.

[Philbin et al., 2007] Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). : (pp. 1–8).

[Philbin et al., 2008] Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[Rodden & Wood, 2003] Rodden, K. & Wood, K. R. (2003). How do people manage their digital photographs? In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 409–416). New York, NY, USA: ACM.

[Rowley et al., 1998] Rowley, H., Baluja, S., & Kanade, T. (1998). Human face detection in visual scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[Rubner et al., 1998] Rubner, Y., Tomasi, C., & Guibas, L. J. (1998). A metric for distributions with applications to image databases. *iccv*, 00, 59.

[Rubner et al., 2000] Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2), 99–121.

[Rui et al., 1997a] Rui, Y., Huang, T., & Mehrotra, S. (1997a). Content-based image retrieval with relevance feedback in MARS. In *Image Processing, 1997. Proceedings., International Conference on*, volume 2.

[Rui et al., 1999] Rui, Y., Huang, T. S., & fu Chang, S. (1999). Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10, 39–62.

[Rui et al., 1997b] Rui, Y., Huang, T. S., & Mehrotra, S. (1997b). Relevance feedback techniques in interactive content-based image retrieval. volume 3312 (pp. 25–36).: SPIE.

[Sable & Hatzivassiloglou, 2000] Sable, C. & Hatzivassiloglou, V. (2000). Text-based approaches for non-topical image categorization. *International Journal on Digital Libraries*, 3(3), 261–275.

[Salton, 1971] Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

[Sapiro & Tannenbaum, 1993] Sapiro, G. & Tannenbaum, A. (1993). Affine invariant scale-space. *International Journal of Computer Vision*, 11(1), 25–44.

[Schmid & Mohr, 1997] Schmid, C. & Mohr, R. (1997). Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5), 530–535.

[Sclaroff et al., 2001] Sclaroff, S., Cascia, M. L., Sethi, S., & Taycher, L. (2001). Mix and match features in the imagerover search engine. (pp. 259–277).

[Scott & Shyu, 2003] Scott, G. J. & Shyu, C.-R. (2003). Ebs k-d tree: An entropy balanced statistical k-d tree for image databases with ground-truth labels. In *CIVR* (pp. 467–476).

[Sebe & Lew, 2001] Sebe, N. & Lew, M. S. (2001). Color-based retrieval. *Pattern Recogn. Lett.*, 22(2), 223–230.

[Sebe et al., 2000] Sebe, N., Lew, M. S., & Huijsmans, D. P. (2000). Toward improved ranking metrics. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10), 1132–1143.

[Sikora, 2001] Sikora, T. (2001). The MPEG-7 Visual Standard for Content Description - An Overview. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), 696–702.

[Sivic & Zisserman, 2003] Sivic, J. & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision* (pp. 1470). Washington, DC, USA: IEEE Computer Society.

[Smeulders et al., 2000] Smeulders, A. W., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), 1349–1380.

[Smith & Chang, 1994] Smith, J. & Chang, S.-F. (1994). Transform features for texture classification and discrimination in large image databases. *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, 3, 407–411 vol.3.

[Spyrou et al., 2008] Spyrou, E., Tolias, G., Mylonas, P., & Avrithis, Y. (2008). A semantic multimedia analysis approach utilizing a region thesaurus and lsa. *wiamis*, 0, 8–11.

[Squire et al., 2000] Squire, D., Müller, W., Müller, H., & Pun, T. (2000). Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters*, 21(13-14), 1193–1198.

[Stricker & Orengo, 1995a] Stricker, M. & Orengo, M. (1995a). : (pp. 381–392).

[Stricker & Orengo, 1995b] Stricker, M. & Orengo, M. (1995b). : (pp. 381–392).

[Swain & Ballard, 1991] Swain, M. J. & Ballard, D. H. (1991). Color indexing. *Int. J. Comput. Vision*, 7(1), 11–32.

[Tagare et al., 1997] Tagare, H., Jaffe, C., & Duncan, J. (1997). Medical Image Databases A Content-based Retrieval Approach.

[Tieu & Viola, 2000] Tieu, K. & Viola, P. (2000). Boosting image retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*, 1, 228–235 vol.1.

[Tolias, 2007] Tolias, G. (2007). Object detection and image classification using mpeg-7 descriptors and visual thesaurus teqniques.

[Torralba et al., 2007] Torralba, A., Fergus, R., & Freeman, W. T. (2007). *Tiny Images.* Technical Report MIT-CSAIL-TR-2007-024, Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology.

[Veltkamp & Tanase, 2002] Veltkamp, R. & Tanase, M. (2002). Content-based image retrieval systems: A survey. *Department of Computing Science, Utrecht University.*

[Veltkamp & Hagedoorn, 2001] Veltkamp, R. C. & Hagedoorn, M. (2001). State of the art in shape matching. (pp. 87–119).

[Vinay et al., 2005] Vinay, V., Cox, I. J., Milic-Frayling, N., & Wood, K. R. (2005). Evaluating relevance feedback algorithms for searching on small displays. In *ECIR* (pp. 185–199).

[Viola & Jones, 2001] Viola, P. & Jones, M. (2001). Rapid Object Detection Using a Boosted Cascade of Simple Features. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1: IEEE Computer Society; 1999.

[Wang et al., 1997] Wang, J., jann Yang, W., & Acharya, R. (1997). Color clustering techniques for color-content-based image retrieval from image databases. *icmcs*, 00, 442.

[Wang et al., 2001] Wang, J., Li, J., & Wiederhold, G. (2001). SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 947–963).

[Witten et al., 1999] Witten, I., Moffat, A., & Bell, T. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann.

[Wolf et al., 2000] Wolf, C., Kropatsch, W., Bischof, H., & Jolion, J.-M. (2000). Content based image retrieval using interest points and texture features. *icpr*, 04, 4234.

[Ye & Xu, 2003] Ye, H. & Xu, G. (2003). Fast search in large-scale image database using vector quantization. In *CIVR* (pp. 477–487).

[Zobel et al., 1998] Zobel, J., Moffat, A., & Ramamohanarao, K. (1998). Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems (TODS)*, 23(4), 453–490.