ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Μέθοδοι Συμπαγούς Αναπαράστασης για Αναζήτηση Εικόνων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βασίλειος Π. Χατζηπάνος

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

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

Μέθοδοι Συμπαγούς Αναπαράστασης για Αναζήτηση Εικόνων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βασίλειος Π. Χατζηπάνος

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Οκτωβρίου 2015

.....	.....	.....
Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.	Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γιώργος Στάμου Επ. Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

.....

Βασίλειος Π. Χατζηπάνος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών

Περίληψη

Αντικείμενο της διπλωματικής εργασίας είναι η μελέτη του προβλήματος μεγάλης κλίμακας αναζήτησης εικόνων βάσει περιεχομένου (CBIR). Το πρόβλημα αυτό ανάγεται στον τρόπο σχηματισμού των αναπαραστάσεων κάθε εικόνας. Η βασική μέθοδος που εξετάζουμε στα πλαίσια της διπλωματικής είναι η VLAD. Παρουσιάζουμε και αναπαράγουμε τον βασικό αλγόριθμο που χρησιμοποιείται για τον σχηματισμό της VLAD αναπαράστασης, καθώς και διάφορες μεθόδους της υπάρχουσας βιβλιογραφίας για τη βελτίωσή της. Τέλος, προτείνουμε νέες συμπαγείς αναπαραστάσεις που βασίζονται στην VLAD και νέες μεθόδους για περαιτέρω βελτίωση της αναπαράστασης.

Λέξεις Κλειδιά

αναζήτηση εικόνων βάσει περιεχομένου, συμπαγείς διανυσματικές αναπαραστάσεις εικόνων, VLAD αναπαράσταση εικόνας, power-law κανονικοποίηση, intranorm, whitening, PCA

Abstract

This thesis addresses the problem of content based large scale image retrieval (CBIR). We study the algorithms and methods that are being used to produce compact image vector representations and primarily the VLAD image representation. We showcase the main algorithm used to produce the VLAD vector and known methods to improve it. Finally, we examine novel image representation vectors based on VLAD and a normalization scheme that can be used for further improvement.

Keywords

content based image retrieval (CBIR), compact image vector representation, VLAD image representation, power-law normalization, intranorm, whitening, PCA

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα Καθηγητή κ. Στέφανο Κόλλια για την εμπιστοσύνη που μου έδειξε και μου ανέθεσε την διπλωματική αυτή εργασία. Επίσης θέλω να ευχαριστήσω τον Ερευνητή Δρ Ιωάννη Αβρίθη, τον Ερευνητή Δρ Γεώργιο Τόλλια και τον Ερευνητή Δρ. Ιωάννη Καλαντίδη, για τη βοήθεια και το ενδιαφέρον τους από την διαμόρφωση του θέματος της εργασίας μέχρι και την τελική συγγραφή της. Τέλος, θέλω να ευχαριστήσω την οικογένεια μου για την στήριξή τους καθ' όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Εισαγωγή **6**

1 Αναζήτηση Εικόνων **9**

1.1 Εισαγωγή στην Αναζήτηση Εικόνων 9

1.2 Αναζήτηση Εικόνων Βάσει Περιεχομένου 11

1.2.1 Εξαγωγή Χαρακτηριστικών 12

1.2.2 Αλγόριθμοι παραγωγής αναπαράστασης εικόνας 14

1.2.3 Τρόπος αξιολόγησης των μεθόδων 16

2 Διάνυσμα αναπαράστασης εικόνων με βάση τοπικούς περιγραφείς **19**

2.1 Βασική μέθοδος 19

2.2 Βελτιώσεις της βασικής μεθόδου 29

3 Υλοποίηση και σύγκριση διανυσματικών αναπαραστάσεων εικόνων **37**

3.1 Πειραματική διάταξη 37

3.2 Υλοποίηση των μεθόδων του κεφαλαίου 2 39

3.3 Συγκριτική Αξιολόγηση 42

4 Νέες μέθοδοι σχηματισμού αναπαραστάσεων εικόνων **47**

4.1 Αναπαραστάσεις με βάση την αναπαράσταση VLAD 47
4.1.1 Τράπεζα Αντικειμένων 48
4.1.2 Επαναταξινόμηση 51
4.2 Μέθοδοι βελτίωσης της αναπαράστασης VLAD 53

Συμπεράσματα **59**

Εισαγωγή

Στην παρούσα διπλωματική εργασία θα μας απασχολήσει το πρόβλημα της αναζήτησης-ανάκτησης εικόνων. Αφορά την εύρεση όμοιων εικόνων σε ένα σύνολο από εικόνες. Ένα σύστημα ανάκτησης εικόνων δέχεται σαν είσοδο μια εικόνα, την συγκρίνει με ένα σύνολο από εικόνες, που αποτελούν τη βάση δεδομένων του συστήματος, και σαν έξοδο επιστρέφει μια λίστα που περιέχει όλες τις εικόνες της βάσης ταξινομημένες σε φθίνουσα σειρά σύμφωνα με την ομοιότητα τους με την εικόνα του ερωτήματος.

Αρχικά παρουσιάζουμε συνοπτικά την ιστορική εξέλιξη του προβλήματος και το πως δομείται ένα σύγχρονο σύστημα ανάκτησης εικόνων. Τα βασικά χαρακτηριστικά που επηρεάζουν τον σχεδιασμό των συστημάτων αυτών, είναι ο τρόπος με τον οποίο γίνεται το ερώτημα στο σύστημα, το μέγεθος του συνόλου πάνω στο οποίο γίνεται η αναζήτηση και το περιεχόμενο των εικόνων. Στα πλαίσια της εργασίας εξετάζουμε διαφορετικούς τρόπους σχεδιασμού CBIR (Συστήματα αναζήτησης εικόνων βάσει περιεχομένου) συστημάτων μεγάλης κλίμακας. Αφορούν αυτόματα συστήματα που ως είσοδο παίρνουν μια εικόνα και αναζητούν όμοιες σε σύνολα από εκατομμύρια εικόνες. Για κάθε εικόνα σχηματίζουμε μια αναπαράσταση που την περιγράφει, ώστε η ομοιότητα ανάμεσα σε δυο εικόνες να ανάγεται στη σύγκριση των αναπαραστάσεών τους. Επομένως, για να περιγράψει κανείς ένα

σύστημα αναζήτησης εικόνων αρκεί να περιγράψει τον τρόπο που σχηματίζεται η αναπαράσταση. Στόχος μας είναι η αναπαράσταση να έχει όσο το δυνατό μικρότερο μέγεθος, ώστε να μπορεί να χρησιμοποιηθεί σε προβλήματα μεγάλης κλίμακας. Για το λόγο αυτό εξετάζουμε, αξιολογούμε και προτείνουμε αλγορίθμους παραγωγής συμπαγών αναπαραστάσεων. Η βασική αναπαράσταση που εξετάζουμε στα πλαίσια της διπλωματικής ονομάζεται VLAD (Vector of Locally Aggregated Descriptors).

Στο πρώτο κεφάλαιο παρουσιάζουμε το πρόβλημα της αναζήτησης εικόνων και κάποιες από τις βασικές μεθόδους που έχουν χρησιμοποιηθεί και χρησιμοποιούνται ακόμη σήμερα για την αντιμετώπισή του, στις οποίες βασίζεται η VLAD αναπαράσταση. Στη συνέχεια του κεφαλαίου περιγράφουμε τη διαδικασία με την οποία εξάγονται τα χαρακτηριστικά από κάθε εικόνα, γιατί είναι κοινό για όλες τις μεθόδους που παρουσιάζουμε. Τέλος, αναφέρουμε τον τρόπο με τον οποίο αξιολογούμε τις μεθόδους και ποια σύνολα εικόνων χρησιμοποιούμε.

Στο δεύτερο κεφάλαιο της εργασίας γίνεται μια αναλυτική παρουσίαση της VLAD μεθόδου. Πρώτα παρουσιάζουμε τη βασική μέθοδο και στη συνέχεια όλες τις βελτιώσεις που έχουν προταθεί στην υπάρχουσα βιβλιογραφία. Οι βελτιώσεις αφορούν ως επί το πλείστον διαφορετικούς τρόπους κανονικοποίησης των αναπαραστάσεων, ελαχιστοποίησης των διαστάσεων και κωδικοποίησης του τελικού διανύσματος.

Στο τρίτο της διπλωματικής περιγράφουμε την πειραματική διάταξη που χρησιμοποιήσαμε για την σύγκριση των αλγορίθμων. Έπειτα, υλοποιούμε τις μεθόδους που έχουμε παρουσιάσει θεωρητικά, τις αξιολογούμε και συγκρίνουμε τα αποτελέσματα που είχαμε με τα αναμενόμενα.

Στο τέταρτο και τελευταίο κεφάλαιο προτείνουμε μεθόδους κατασκευής νέων

διανυσματικών συμπαγών αναπαραστάσεων που χρησιμοποιούν την VLAD αναπαράσταση ως βάση καθώς και νέες μεθόδους για τη περαιτέρω βελτίωσή της. Σημειώνουμε πως δεν είχαν όλες οι μέθοδοι θετικά αποτελέσματα αλλά θεωρήσαμε πως παρόλα αυτά αξίζει να τις συμπεριλάβουμε στην εργασία.

Κεφάλαιο 1

Αναζήτηση Εικόνων

Στην ενότητα αυτή παρουσιάζεται το πρόβλημα της αναζήτησης εικόνων, οι βασικοί αλγόριθμοι που χρησιμοποιούνται για την αντιμετώπισή του και ο τρόπος αξιολόγησής τους.

1.1 Εισαγωγή στην Αναζήτηση Εικόνων

Η σημερινή εποχή είναι γνωστή ως η εποχή της πληροφορίας. Καθημερινά παράγεται τεράστια ποσότητα πληροφοριών και ένα μεγάλο κομμάτι αυτής αφορά ψηφιακές εικόνες. Οι συλλογές των ψηφιακών εικόνων που έχουν δημιουργηθεί σαν αποτέλεσμα είναι πολύ μεγάλες σε μέγεθος και διαρκώς αυξάνονται. Το πρόβλημα που προκύπτει είναι ότι δεν μπορεί κανείς να διαχειριστεί εύκολα μια βάση από ψηφιακές εικόνες αν δεν υπάρχει αποδοτικός τρόπος για αναζήτηση ή ανάκτηση εικόνων.

Η ανάκτηση εικόνων αποτελεί ερευνητική περιοχή από την δεκαετία του 70 σε διάφορα επιστημονικά πεδία. Η πρώτη προσπάθεια για αντιμετώπιση του προβλή-

ματος έγινε σχεδιάζοντας ένα σύστημα ανάκτησης εικόνων να λειτουργεί όπως και ένα σύστημα ανάκτησης κειμένου(text-based). Συγκεκριμένα έπρεπε κανείς με το χέρι να εκτιμήσει το περιεχόμενο της εικόνας και να αναθέσει για κάθε εικόνα ένα σύνολο από λέξεις που περιγράφουν το περιεχόμενο. Στη συνέχεια η κάθε εικόνα αποθηκεύεται στο σύστημα με ένα σύνολο από λέξεις που την χαρακτηρίζουν και έτσι το πρόβλημα ανάγεται σε πρόβλημα ανάκτησης βάσει κειμένου. Η χρήση ανθρώπινου δυναμικού για την επίλυση του προβλήματος δεν είναι αποδοτική για πολλαπλούς λόγους, όπως:

- Η χρήση ανθρώπινου δυναμικού για την διαδικασία είναι πολυέξοδη.
- Η ταχύτητα εισαγωγής νέων εικόνων σε ένα τέτοιο σύστημα δεν μπορεί να ακολουθήσει την εισροή νέων εικόνων.
- Η αντίληψη του ανθρώπου είναι υποκειμενική και δεν είναι πάντα η βέλτιστη.
- Ο άνθρωπος δεν μπορεί αν περιγράψει επαρκώς μια εικόνα με πλούσιο περιεχόμενο.

Στη συνέχεια προτάθηκε να αντικατασταθεί η χρήση συστημάτων ανάκτησης βάσει κειμένου με συστήματα που βασίζονται στην ανάκτηση βάσει περιεχομένου (Content-Based Image Retrieval, CBIR). Το οπτικό περιεχόμενο αφορά τα χαρακτηριστικά, όπως σχήμα, υφή, χρώμα. Η προσέγγιση αυτή χρησιμοποιείται μέχρι και σήμερα και αναλύεται στην επόμενη ενότητα. Βασική διαφορά που δεν παρατηρήσαμε προηγούμενα είναι ότι τα συστήματα ανάκτησης βάσει περιεχομένου διαφέρουν και ως προς τον τρόπο που γίνεται το ερώτημα. Στην περίπτωση που

έχουμε μια βάση με μεταδεδομένα για κάθε εικόνα, τα οποία είναι σε μορφή κειμένου ο τρόπος αναζήτησης γίνεται δίνοντας έναν όρο σαν είσοδο στο σύστημα που αφορά το περιεχόμενο της εικόνας ή κάποιο χαρακτηριστικό. Στα CBIR συστήματα το ερώτημα γίνεται συνήθως με τη χρήση μιας εικόνας για ερώτημα.

1.2 Αναζήτηση Εικόνων Βάσει Περιεχομένου

Ο όρος CBIR φαίνεται να χρησιμοποιήθηκε πρώτη φορά από τον T. Kato[7] σε πειράματα που αφορούσαν την αυτόματη ανάκτηση εικόνων βάσει του χρώματος και του σχήματος τους. Για να μοντελοποιήσουμε ένα τέτοιο συστήματα σε ποιοτικό επίπεδο χρειαζόμαστε έναν αλγόριθμο που θα εξάγει αυτόματα μια αναπαράσταση για κάθε εικόνα και μια βάση από εικόνες και τις αντίστοιχες αναπαραστάσεις τους. Το βασικό πρόβλημα που ανακύπτει στη σχεδίαση ενός συστήματος CBIR είναι ότι μας ενδιαφέρει η αναπαράσταση κάθε εικόνας να είναι όσο πιο πλούσια γίνεται και επίσης το σύνολο της πληροφορίας που παράγεται για όλες τις εικόνες της βάσης να είναι διαχειρίσιμο. Το πρόβλημα αυτό είναι ιδιαίτερα εμφανές όταν η βάση των εικόνων είναι πάνω από 10 εκατομμύρια εικόνες, όπου πολλοί αλγόριθμοι αναζήτησης δεν είναι πλέον υλοποιήσιμοι.

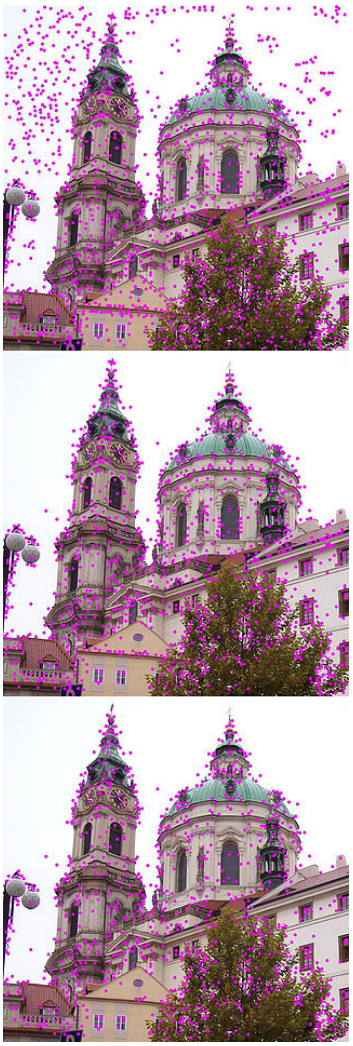
Συγκεκριμένα μας ενδιαφέρει η αναζήτηση σε ένα τέτοιο σύστημα να δίνει όσο το δυνατό καλύτερο αποτελέσματα σε σύγκριση με το μέγεθος της αναπαράστασης και την ταχύτητα της εύρεσης. Συνήθως όσο μεγαλώνει το μέγεθος της αναπαράστασης επηρεάζει αρνητικά την ταχύτητα της αναζήτησης αλλά θετικά την ποιότητα του αποτελέσματος. Υπάρχουν πολλοί αλγόριθμοι που έχουν πολύ καλά αποτελέσματα αλλά είναι αδύνατον να χρησιμοποιηθούν σε μεγάλη κλίμακα. Στην εργασία αυτή εξετάζουμε το πρόβλημα της ανάκτησης εικόνων σε με-

γάλη κλίμακα. Για το λόγο αυτό μας ενδιαφέρουν οι αλγόριθμοι που δημιουργούν συμπαγείς αναπαραστάσεις για κάθε εικόνα. Στη συνέχεια του κεφαλαίου παρουσιάζουμε την διαδικασία που ακολουθεί κανείς για την δημιουργία της αναπαράστασης κάθε εικόνας, γνωστές μεθόδους και αλγορίθμους καθώς και τον τρόπο με τον οποίο θα αξιολογούμε τις μεθόδους.

1.2.1 Εξαγωγή Χαρακτηριστικών

Για την εξαγωγή των χαρακτηριστικών χρησιμοποιήθηκε η μέθοδος SIFT [17] (Scale Invariant Feature Transform). Από κάθε εικόνα εξάγεται ένα σύνολο από διανύσματα 128 διαστάσεων. Τα διανύσματα αυτά ονομάζονται περιγραφείς και αντιστοιχούν σε περιοχές ενδιαφέροντος της εικόνας. Οι περιοχές αυτές αφορούν τις μέγιστες και ελάχιστες τιμές στο χώρο κλίμακας μιας εικόνας. Στο σύνολο τους περιγράφουν την χωρική κατανομή των ακμών της εικόνας και είναι ανεξάρτητοι από αφινικούς μετασχηματισμούς και αλλαγές της κλίμακας του αντικειμένου χωρίς να επηρεάζονται σε μεγάλο βαθμό από αλλαγές στην φωτεινότητα. Στην πρώτη εικόνα 1.1 φαίνονται όλα τα οριακά σημεία της εικόνας που έχει αναγνωρίσει ο αλγόριθμος στην εικόνα. Σαν δεύτερο βήμα αφαιρούνται τα σημεία που έχουν χαμηλή αντίθεση. Αυτά που έμειναν φαίνονται στην δεύτερη εικόνα και σαν τελευταίο βήμα ο αλγόριθμος αφαιρεί επίσης αυτά που βρίσκονται πάνω στις γωνίες. Τα σημεία ενδιαφέροντος που θα κρατήσει φαίνονται στην τελευταία εικόνα.

Για κάθε σημείο ενδιαφέροντος της εικόνας, όπως αναφέραμε, σχηματίζεται ένα διάνυσμα 128 διαστάσεων που περιέχει τις κυρίαρχες κατευθύνσεις της αλλαγής της φωτεινότητας της περιοχής του σημείου. Ο ακριβής αλγόριθμος για την παραγωγή του διανύσματος δε θα μας απασχολήσει στα πλαίσια της διπλωματικής αυτής.



Σχήμα 1.1: Στην εικόνα αυτή φαίνονται τα σημεία ενδιαφέροντος που αναγνωρίζει ο αλγόριθμος εξαγωγής χαρακτηριστικών SIFT στην εικόνα.

Αυτό έχει σαν αποτέλεσμα σε διαφορετικές εικόνες του ίδιου αντικειμένου να εξάγονται παρόμοιες περιοχές ενδιαφέροντος ανεξάρτητα από την γωνία του αντικειμένου και την κλίμακά του. Επομένως η σύγκριση δυο εικόνων ανάγεται στην απόσταση των διανυσματικών περιγραφέων τους. Το μέτρο που χρησιμοποιείται στον SIFT διανυσματικό χώρο είναι η ευκλείδεια απόσταση.

1.2.2 Αλγόριθμοι παραγωγής αναπαράστασης εικόνας

Γνωρίζουμε πλέον πως για να συγκρίνουμε δυο εικόνες ως προς την ομοιότητα τους αρκεί να συγκρίνουμε τους περιγραφείς τους. Για κάθε εικόνα, ειδικά αν είναι πλούσια σε περιεχόμενο, ένας ανιχνευτής περιγραφέων (feature detector), όπως ο SIFT, θα εξάγει χιλιάδες περιγραφείς. Επομένως, είναι υπολογιστικά αδύνατο να συγκρίνουμε ένα προς ένα τους περιγραφείς των δυο εικόνων μεταξύ τους. Το πρόβλημα της σύγκρισης δυο εικόνων ανάγεται στο πως θα επεξεργαστεί κανείς το σύνολο αυτό για να εξάγει πλέον μια αναπαράσταση για κάθε εικόνα που θα συγκεντρώνει την πληροφορία που περιέχει το σύνολο των SIFT περιγραφέων που αντιστοιχεί στην εικόνα. Στις επόμενες παραγράφους παρουσιάζουμε κάποιες βασικές προσεγγίσεις στο πρόβλημα αυτό, στις οποίες στηρίζεται και ο αλγόριθμος VLAD, που θέλουμε να εξετάσουμε.

Ένα βασικό μοντέλο που χρησιμοποιείται στην αναζήτηση εικόνων είναι το μοντέλο συνόλου λέξεων BOF (Bag of Features). Το μοντέλο αυτό χρησιμοποιείται κατά κόρων στην επεξεργασία φυσικής γλώσσας. Ονομάζεται, στο χώρο αυτό, μοντέλο συνόλου λέξεων γιατί μπορεί κανείς να αναπαραστήσει μια πρόταση σαν το σύνολο από τις λέξεις που συντελούν την πρόταση. Για την ακρίβεια στο μοντέλο αυτό κρατάμε τις διαφορετικές λέξεις που χρησιμοποιούνται και την πολλαπλότητα του. Στην ανάκτηση εικόνων δεν μπορεί κανείς να διαχειριστεί τους

περιγραφείς με τον ίδιο τρόπο γιατί ο χώρος των οπτικών περιγραφέων είναι πολύ αραιός με αποτέλεσμα σπανίως να εμφανίζονται οι ίδιες τιμές, όπως συμβαίνει στη φυσική γλώσσα.

Για την υλοποίηση του, στην ανάλυση εικόνων, το πρώτο βήμα είναι να ομαδοποιήσουμε τους περιγραφείς σε σύνολα με οπτικές λέξεις (visual words) που θα ονομάζουμε οπτικό λεξικό. Αυτό επιτυγχάνεται με χρήση ενός αλγορίθμου συσταδοποίησης, συνήθως του kmeans, στο χώρο που σχηματίζουν οι περιγραφείς. Στόχος είναι η τμηματοποίηση του συνόλου των SIFT σε k συστάδες που ονομάζουμε οπτικές λέξεις. Το σύνολο των οπτικών λέξεων που έχουμε βρει σχηματίζει ένα οπτικό λεξικό για το σύνολο των SIFT. Έχοντας σχηματίσει το λεξικό, κάθε περιγραφέας μιας εικόνας αντιστοιχείται με την κοντινότερη k οπτική λέξη. Κάθε εικόνα περιγράφεται από το ιστόγραμμα της αντιστοίχησης των οπτικών περιγραφέων της στις k οπτικές λέξεις, το οποίο σχηματίζει ένα διάνυσμα k διαστάσεων. Επόμενο βήμα του αλγορίθμου είναι η κανονικοποίηση του διανύσματος αυτού, συνήθως με L_2-κανονικοποίηση. Στη συνέχεια δίνονται βάρη σε κάθε στοιχείο του διανύσματος που στηρίζονται στη μετρική idf που διαιρεί κάθε στοιχείο ανάλογα με το πόσο συχνά εμφανίζεται στην εικόνα. Υπάρχουν πολλές προτάσεις για διαφορετικούς τρόπους κανονικοποίησης και υπολογισμού βαρών.

Ένας ακόμη δημοφιλής τρόπος για την δημιουργία της αναπαράστασης είναι με χρήση του Fisher Kernel [21]. Χρησιμοποιείται για την δημιουργία διανυσματικών αναπαραστάσεων σταθερού μήκους, όταν για δεδομένα έχουμε ένα σύνολο από ανεξάρτητα δείγματα μεταβλητού μεγέθους και γνωρίζουμε ότι τα δείγματα περιγράφονται από ένα παραμετρικό μοντέλο που έχει προκύψει από ένα σύνολο εκπαίδευσης. Η μέθοδος αυτή χρησιμοποιήθηκε για το πρόβλημα της κατηγοριοποίησης εικόνων από τον Perronnin [9]. Χρησιμοποίησε μείγματα Γκαουσιανών

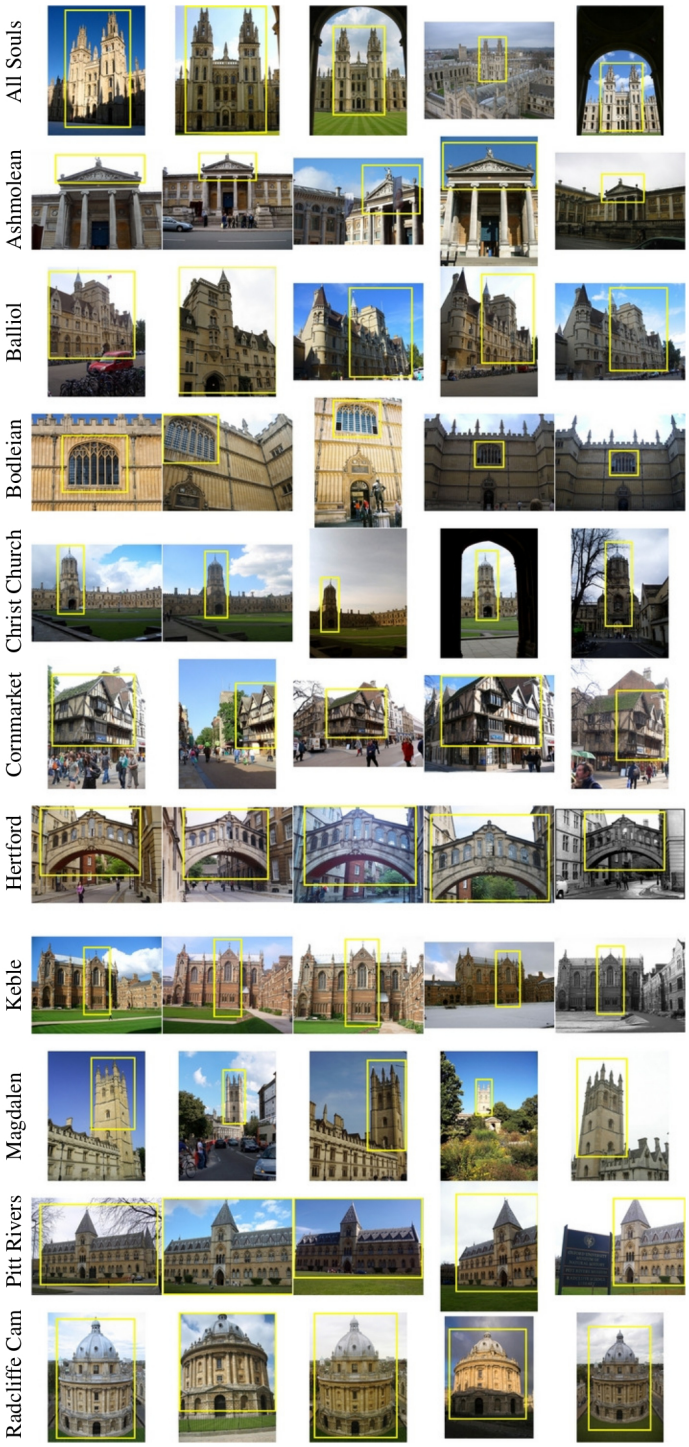
κατανομών για να μοντελοποιήσει τις οπτικές λέξεις, όπου ο πίνακας διασπορών κάθε Γκαουσιανής είχε συνιστώσες μόνο στη διαγώνιο. Έτσι η διανυσματική αναπαράσταση που παρήγαγε για κάθε εικόνα είχε μόνο $k \times d$ όπου k είναι ο αριθμός των Γκαουσιανών στο μοντέλο και d ο αριθμός των παραμέτρων για κάθε μια. Ενδιαφέρον χαρακτηριστικό της μεθόδου είναι ότι προσπαθεί να μοντελοποιήσει την τοπολογία των περιγραφέων σε κάθε οπτική λέξη. Κάτι που, όπως θα δούμε στη συνέχεια προσπαθεί να επιτύχει και η VLAD αναπαράσταση.

1.2.3 Τρόπος αξιολόγησης των μεθόδων

Για την σωστή αξιολόγηση των μεθόδων που θα μελετήσουμε, φροντίζουμε να χρησιμοποιήσουμε το ίδιο σύνολο εικόνων για την εκπαίδευση των οπτικών λεξικών. Συγκεκριμένα στα περισσότερα πειράματα η εκπαίδευση έγινε στο Oxford 100k [20], που αποτελεί ένα σύνολο από εικόνες κτηρίων της Οξφόρδης, τα οποία έχουν συλλεχθεί από το Flickr. Επίσης έχουμε ένα σύνολο από 5062 εικόνες που ονομάζεται Oxford5k [20] το οποίο είναι και το σύνολο στο οποίο γίνεται η αξιολόγηση. Συγκεκριμένα έχουμε 55 εικόνες που αντιστοιχούν σε 11 διαφορετικά κτήρια, σχήμα 1.2, στις οποίες είναι σημειωμένο με ένα τετράγωνο πλαίσιο το αντικείμενο ενδιαφέροντος που θα χρησιμοποιηθεί σαν ερώτημα στο σύστημα αυτό. Επίσης γνωρίζουμε για κάθε ερώτημα ποιες εικόνες του Oxford5k θεωρούνται σωστές απαντήσεις, ποιες λάθος και ποιες "άσχετες". Οι "άσχετες" δεν επηρεάζουν την βαθμολόγηση του αλγορίθμου. Το μέτρο που χρησιμοποιείται για την αξιολόγηση είναι το mAP (mean Average Precision). Αποτελεί το εμβαδό της καμπύλης ακρίβειας-ανάκτησης (precision-recall curve). Η ακρίβεια ορίζεται ως ο λόγος των σωστών εικόνων προς το σύνολο εικόνων που επέστρεψε το σύστημα και η ανάκτηση ως ο λόγος των σωστών εικόνων που επέστρεψε το σύστημα προς

το σύνολο των σωστών εικόνων. Για κάθε ερώτημα υπολογίζουμε την ακρίβεια της απάντησης για κάθε επίπεδο ανάκτησης, έπειτα βρίσκουμε τη μέση ακρίβεια για κάθε ερώτημα (AP) και τέλος το μέσο όρο της μέσης ακρίβειας στο σύνολο των ερωτημάτων.

Εκτός του συνόλου κτηρίων της Οξφόρδης (Oxford Buildings Dataset) χρησιμοποιούνται και άλλα σύνολα όπως το INRIA Holidays [11] για την αξιολόγηση μεθόδων. Αποτελείται από 1491 εικόνες υψηλής ανάλυσης με τοποθεσίες και αντικείμενα. Οι 500 από αυτές χρησιμοποιούνται σαν ερωτήματα στο σύστημα. Η ποιότητα της αναζήτησης όπως και στο Oxford υπολογίζεται με χρήση του mAP. Αξίζει στο σημείο αυτό να σημειώσουμε ότι στα δικά μας πειράματα χρησιμοποιήθηκε μόνο το Oxford για την αξιολόγηση των μεθόδων.



Σχήμα 1.2: Οι εικόνες που χρησιμοποιούνται σαν ερωτήματα στο Oxford5k.

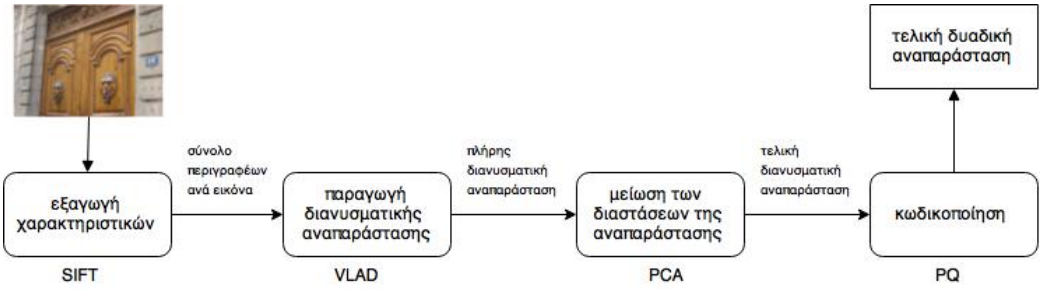
Κεφάλαιο 2

Διάνυσμα αναπαράστασης εικόνων με βάση τοπικούς περιγραφείς

Στο κεφάλαιο αυτό παρουσιάζουμε την μέθοδο VLAD (Διάνυσμα τοπικά συναθροισμένων περιγραφέων - Vector of Locally Aggregated Dscriptors). Αποτελεί μια διανυσματική αναπαράσταση εικόνων, με στοιχεία από τις μεθόδους BOF και Fisher Kernel που αναφέραμε στο προηγούμενο κεφάλαιο. Στη συνέχεια αναλύουμε την βασική μέθοδο και κάποιες βελτιώσεις της.

2.1 Βασική μέθοδος

Στο σχήμα 2.1 φαίνεται ο βασικός αλγόριθμος [13] που χρησιμοποιείται για την παραγωγή της VLAD αναπαράστασης. Το πρώτο βήμα είναι η εξαγωγή των περιγραφέων της εικόνας με χρήση της SIFT μεθόδου που αναφέραμε στην ενότητα 1.2.1. Όπως και στην BOF το πρώτο βήμα είναι να δημιουργήσουμε ένα λεξικό $C = c_1, ...c_k$, όπου k ο αριθμός των οπτικών λέξεων, με χρήση ενός αλ-



Σχήμα 2.1: Τα βήματα του βασικής μεθόδου παραγωγής της VLAD διανυσματικής αναπαράστασης.

γορίθμου συσταδοποίησης και πιο συγκεκριμένα τον kmeans. Το λεξικό αυτό εκπαιδεύεται σε ένα σύνολο από εικόνες, διαφορετικό από το σύνολο εικόνων στο οποίο γίνεται η αξιολόγηση του αλγορίθμου.

$$q : \Re^d \rightarrow C \subset \Re^d \tag{2.1}$$

$$\boldsymbol{x} \mapsto q(\boldsymbol{x}) = argmin \|\boldsymbol{x} - \boldsymbol{c}\|^2 \tag{2.2}$$

Στη συνέχεια για κάθε περιγραφέα x βρίσκουμε τη κοντινότερη οπτική λέξη $c_i = q(x)$, όπου q ένας αλγόριθμος κβάντισης βάσει απόστασης, εξίσωση 2.2. Για τον εύρεση των κοντινότερων κέντρων στον χώρο των SIFT χρησιμοποιούνται συνήθως για υπολογιστικούς λόγους μη εξαντλητικοί αλγόριθμοι εύρεσης κοντινότερου γείτονα (ANN). Αναλυτικότερα για κάθε x βρίσκουμε το κοντινότερο c_i και ως μέτρο απόστασης στο χώρο των SIFT χρησιμοποιούμε την ευκλείδεια απόσταση. Σκοπός είναι να δημιουργήσουμε k διανύσματα, ένα για κάθε οπτική λέξη, ως το διανυσματικό άθροισμα των υπολοίπων $x - c_i$ των διανυσμάτων x ως προς

τα αντίστοιχα c_i.

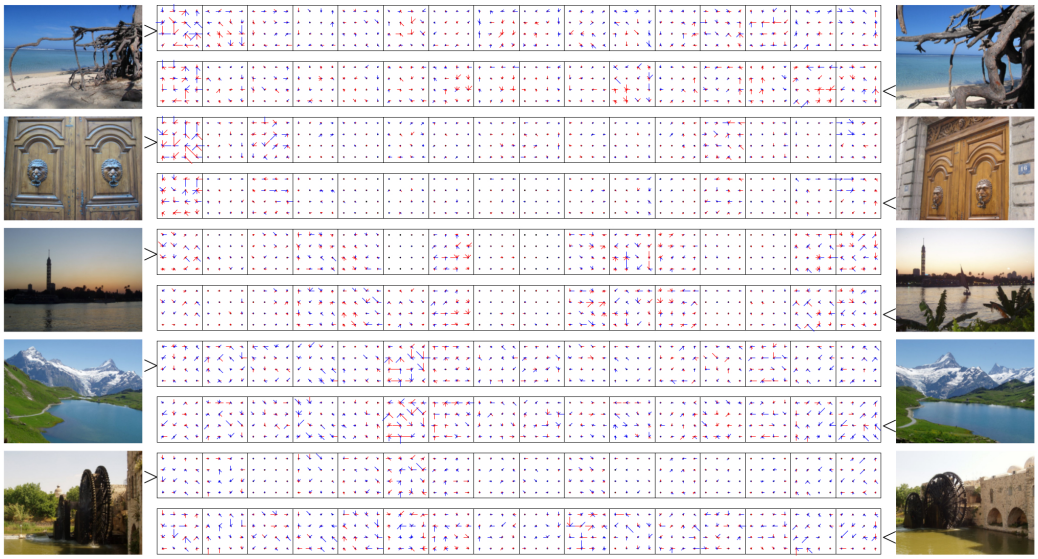
$$v_i = \sum_{q(x)=c_i} x - c_i \ , \qquad i = 1, .., k \tag{2.3}$$

Θεωρώντας ότι οι SIFT περιγραφείς έχουν d διαστάσεις τότε και τα διανύσματα v_i έχουν d διαστάσεις αφού προκύπτουν από το διανυσματικό άθροισμα των υπολοίπων τους. Στη συνέχεια δημιουργούμε ένα διάνυσμα V για κάθε εικόνα που προκύπτει από την ένωση των τοπικών συνισταμένων υπολοίπων ανά λέξη v_i, εξίσωση 2.4. Το τελικό διάνυσμα V έχει $D = k \times d$ διαστάσεις. Το τελευταίο βήμα είναι η κανονικοποίηση του διανύσματος V ως προς την L_2 νόρμα $V \leftarrow \frac{V}{\|V\|_2}$.

$$V = [v_1^T ... v_k^T] \tag{2.4}$$

Στο παραπάνω διάνυσμα V αναφερόμαστε ως πλήρες VLAD διάνυσμα και στα επιμέρους v_i ως τοπικά VLAD. Το πλήρες διάνυσμα VLAD έχει $k \times d$ διαστάσεις. Στο σημείο αυτό αξίζει να σημειώσουμε πως το d προκύπτει από τις διαστάσεις του χώρου SIFT και όπως αναφέραμε είναι 128 διαστάσεων. Επίσης το k είναι το μέγεθος του λεξικού που χρησιμοποιούμε και συνήθως παίρνει τιμές $k = 16, 32, 64, 128$. Αυτό έχει σαν αποτέλεσμα το πλήρες διάνυσμα VLAD να έχει πολλές χιλιάδες διαστάσεις. Για το λόγο αυτό δεν μπορεί να χρησιμοποιηθεί για προβλήματα μεγάλης κλίμακας ως έχει.

Ακόμη και στο σημείο αυτό, που έχουμε παράξει ένα διάνυσμα που περιγρά-



Σχήμα 2.2: Το πλήρες VLAD διάνυσμα 10 εικόνων, για λεξικό μεγέθους $k = 16$ ($D = 16 \times 128$). Με κόκκινο φαίνονται οι αρνητικές συνιστώσες.

φει μια ολόκληρη εικόνα, η αναπαράσταση που έχουμε σχηματίσει για κάθε εικόνα είναι εξαιρετικά μεγάλη για να χρησιμοποιηθεί σε προβλήματα αναζήτησης πολύ μεγάλης κλίμακας. Το επόμενο βήμα είναι να προσπαθήσουμε να μειώσουμε ακόμη περισσότερο τη διάσταση της.

Στο σχήμα 2.2 φαίνεται το πλήρες διάνυσμα VLAD στην περίπτωση που έχουμε λεξικό μεγέθους $k = 16$ (άρα $D = 16 \times 128$). Οι SIFT περιγραφείς έχουν θετικές τιμές σε κάθε συνιστώσα όπως αναφέραμε και στην ενότητα 1.2.1. Η VLAD αναπαράσταση επειδή προκύπτει από άθροισμα διαφορών θα έχει αρνητικές τιμές σε κάποιες συνιστώσες, οι οποίες φαίνονται με κόκκινο στο σχήμα. Παρατηρώντας κανείς την δομή των περιγραφέων γίνεται εμφανές ότι πολύ λίγες συνιστώσες συγκεντρώνουν σχεδόν όλη την ενέργεια του διανύσματος. Επίσης, φαίνεται να έχουν κάποια σταθερή δομή, καθώς αν συγκρίνει κανείς τις αναπαραστάσεις των

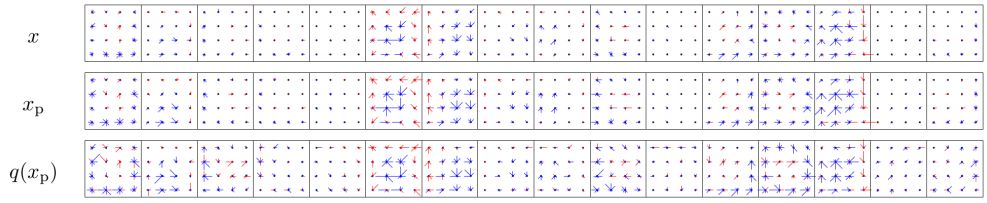
διαφορετικών εικόνων μεταξύ τους παρατηρεί πως μεγάλες τιμές εμφανίζονται στην ίδια οπτική λέξη συνήθως. Θεωρούμε πως με ανάλυση βασικών συνιστωσών PCA (Principal Component Analysis) μπορεί κανείς να μειώσει τις διαστάσεις του VLAD διανύσματος χωρίς να χάσει σημαντικό κομμάτι της πληροφορίας. Η μέθοδος PCA χρησιμοποιείται ευρέως όταν θέλουμε να μειώσουμε τις διαστάσεις ενός διανύσματος ειδικότερα στις περιπτώσεις που είναι εμφανές ότι λίγες συνιστώσες συγκεντρώνουν την περισσότερη πληροφορία, καθώς ο στόχος της ανάλυσης κυρίαρχων συνιστωσών είναι να βρει τις συνιστώσες του διανύσματος που έχουν την μεγαλύτερη διακριτική ικανότητα.

Συγκεκριμένα υπολογίζουμε, για ένα σύνολο από εικόνες, τον πίνακα συνδιακύμανσης και από αυτόν βρίσκουμε τα D' πιο ενεργά (μεγαλύτερες ιδιοτιμές) ιδιοδιανύσματα του πίνακα και σχηματίζουμε έναν πίνακα $M_{D' \times D}$. Ο πίνακας αυτός αποτελεί μια αντιστοιχία ενός διανύσματος $x \in \Re^D$ σε ένα διάνυσμα $x' = Mx \in \Re^{D'}$. Μπορεί κανείς να θεωρήσει ότι είναι μια προβολή σε ένα άλλο χώρο μικρότερης διάστασης. Μετά την εφαρμογή του PCA όπως είναι λογικό όσο λιγότερες διαστάσεις κρατήσουμε στο τελικό διάνυσμα τόσο μικρότερη διακριτική ικανότητα θα έχει η τελική αναπαράσταση.

Αξίζει να σημειωθεί πως το σύνολο εικόνων που χρησιμοποιείται για τον υπολογισμό του πίνακα συνδιακύμανσης είναι διαφορετικό από το σύνολο στο οποίο θα γίνει η αξιολόγηση του αλγορίθμου. Θα ήταν λάθος κανείς να κάνει ανάλυση βασικών συνιστωσών στο σύνολο που θα γίνει η αξιολόγηση γιατί οι συνιστώσες που θα κυριαρχούσαν θα ήταν αυτές που κυριαρχούν στο συγκεκριμένο σύνολο και όχι σε ένα γενικευμένο σύνολο. Η χρήση διαφορετικού συνόλου εικόνων για την εκπαίδευση βοηθάει και στην αντιμετώπιση του προβλήματος της υπερπροσαρμογής (overfitting) όπου το σύστημα προσαρμόζεται στις εικόνες στις οποίες

έγινε η εκπαίδευση του συστήματος και δεν λειτουργεί σωστά σε άλλα σύνολα. Το τελικό διάνυσμα V', θα το ονομάζουμε μειωμένο VLAD, έχει D' διαστάσεις και κανονικοποιείται ως προς την L_2 νόρμα όπως και το πλήρες διάνυσμα. Στο σημείο αυτό έχουμε καταφέρει να παραστήσουμε κάθε εικόνα με ένα διάνυσμα D' όπου στην πραγματικότητα αντιστοιχεί συνήθως σε 96 με 128 διαστάσεις και για να συγκρίνουμε δυο εικόνες απλώς υπολογίζουμε το εσωτερικό γινόμενο των αναπαραστάσεων τους.

Στην βασική μέθοδο ως τελευταίο βήμα κβαντίζουν το τελικό διάνυσμα μετά το PCA, για να σχηματίσουν μια ακόμη πιο συμπαγή απεικόνιση. Το βήμα αυτό ονομάζεται PQ (product quantizer) και αφορά την τελική κωδικοποίηση των αναπαραστάσεων στη βάση. Συγκεκριμένα, έστω x η τελική διανυσματική αναπαράσταση της εικόνας. Χωρίζουν το διάνυσμα x σε m υποδιανύσματα $x^1, \ldots x^m$ ίσου μεγέθους D/m. Στη συνέχεια ορίζουμε ένα PQ $q(x)$, όπως φαίνεται στην εξίσωση 2.5. Κάθε επιμέρους συνάρτηση κβάντισης $q_j(.), j = 1, .., .m$ κβαντίζει το επιμέρους x_j σε k_s διαφορετικές τιμές οι οποίες έχουν υπολογιστεί με τη βοήθεια του kmeans. Το k_s παίρνει τιμές ώστε κάθε επιμέρους x_j να μπορείς να αποθηκευτεί σε b_s bits, ώστε $b_s = log_2 k_s$, π.χ. $k_s = 256$ ώστε κάθε x_j να αποθηκεύεται σε 1 byte. Προκύπτει επομένως πως κάθε εικόνα στο σύνολο μπορεί να αντιστοιχηθεί μέσω της $q(x)$ σε $k = (k_s)^m$ διαφορετικά κέντρα, το οποίο είναι αποτελεί μια αρκετά ικανοποιητική προσέγγιση. Η μέθοδος αυτή ονομάζεται μέθοδος ασύμμετρου υπολογισμού απόστασης (ADC - Asymmetric Distance Computation) και ένα από τα σημαντικά προτερήματά της είναι ότι κωδικοποιεί μόνο τις εικόνες της βάσης, ενώ η εικόνα που δίνουμε σαν ερώτημα δεν χάνει πληροφορία λόγω κβάντισης.



Σχήμα 2.3: Επίδραση της κωδικοποίησης στην διανυσματική αναπαράσταση. Πάνω εικόνα: πλήρες VLAD διάνυσμα για $k = 16\,(D = 2048)$. Μεσαία εικόνα: VLAD διάνυσμα μετά το PCA σε $D' = 128$. Κάτω εικόνα: τελική διανυσματική αναπαράσταση μετά από ADC 16×8.

$$q(x) = (q_1(x^1), ..., q_m(x^m)), \tag{2.5}$$

Έστω ένα σύνολο $Y = \{y_1, ..., y_n\}$ που αποτελεί το σύνολο των διανυσματικών αναπαραστάσεων των εικόνων της βάσης και ένα διάνυσμα $x \in \Re^{D'}$ που αποτελεί τη διανυσματική αναπαράσταση της εικόνας του ερωτήματος για την οποία προσπαθούμε να βρούμε τον κοντινότερο γείτονα στο σύνολο NN(x) (Nearest Neighbor). Για να το πετύχουμε αυτό υπολογίζουμε το $argmin_i \|x - q(y_i)\|^2, i = 1, ..., n$, όπου ο υπολογισμός της τετραγωνικής απόστασης αναλύεται όπως φαίνεται στην εξίσωση 2.6, όπου το y_i^j είναι το j-οστό υποδιάνυσμα του y_i.

$$\|x - q(y_i)\|^2 = \sum_{j=1,...,m} \|x^j - q(y_i^j)\|^2, \tag{2.6}$$

Στην εικόνα 2.3 φαίνεται πως επηρεάζει την αναπαράσταση της εικόνας το

βήμα της μείωσης των διαστάσεων με το PCA και στη συνέχεια η κωδικοποίηση. Το διάνυσμα φαίνεται να μην έχει επηρεαστεί σε μεγάλο βαθμό ούτε από τη μείωση των διαστάσεων σε $D' = 128$, ούτε από την κωδικοποίηση ADC 16×8 που αντιστοιχεί σε μόλις 16 bytes τελικής αναπαράσταση. Η κωδικοποίηση ADC $m \times b_s$ δηλώνει ότι έχουμε χωρίσει το διάνυσμα σε 16 υποδιανύσματα όπου το κάθε ένα αποθηκεύεται σε 8 bits. Στο [13] παρατηρούν, πως αν προσέξει κανείς τον τρόπο που θα κβαντίσει το τελικό αποτέλεσμα μπορεί να βελτιώσει ακόμη περισσότερο την απόδοση της αναπαράστασης παρότι μειώνει το μέγεθος της αναπαράστασης. Ο λόγος είναι πως όταν επιλέξουμε τις κυρίαρχες συνιστώσες για να σχηματίσουμε το μειωμένο διάνυσμα VLAD(συνήθως $D' = 128$), η περισσότερη ενέργεια θα συγκεντρωθεί στις πρώτες λίγες τιμές, για την ακρίβεια όπως θα παρουσιάσουμε και στη συνέχεια το 50% της συνολικής ενέργειας του διανύσματος συγκεντρώνεται στις πρώτες 20-25 τιμές. Με αποτέλεσμα οι πρώτες συνιστώσες να επηρεάζουν πολύ περισσότερο την αναζήτηση από τις υπόλοιπες. Με τον τρόπο που γίνεται η κωδικοποίηση στο [13] κάθε συνιστώσα του D' θα έχει το ίδιο μέγεθος στον τελική αναπαράσταση με αποτέλεσμα η κβάντιση να είναι πολύ πιο αυστηρή στις πρώτες συνιστώσες, κάτι που λειτουργεί σαν idf στην αναπαράσταση. Στις περισσότερες περιπτώσεις βέβαια το κωδικοποιημένο διάνυσμα θα έχει χειρότερη απόδοση.

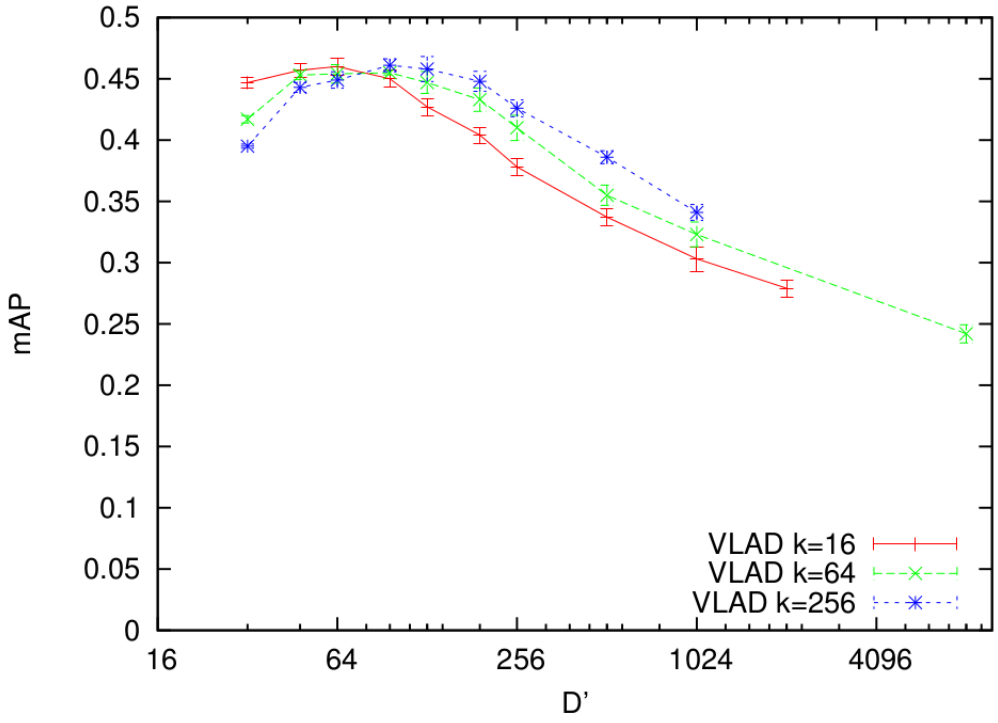
Το βήμα αυτό είναι πολύ σημαντικό γιατί επιτρέπει την χρήση του αλγορίθμου σε προβλήματα αναζήτησης εικόνων πολύ μεγάλης κλίμακας. Παρόλα αυτά στα πλαίσια της διπλωματικής εργασίας δεν θα μας απασχολήσει το βήμα αυτό της κωδικοποίησης. Θεωρούμε πως μπορεί να προσθέσει το βήμα αυτό κανείς σε κάθε μέθοδο αν χρειαστεί να πειραματιστεί σε τόσο μεγάλη κλίμακα. Στον πίνακα 2.1 συγκρίνουμε την απόδοση της βασικής μεθόδου VLAD με τις μεθόδους BOF

Περιγραφέας	k	D	Holidays (mAP)			
			D	$\rightarrow D' = 128$	$\rightarrow D' = 64$	$\rightarrow D' = 32$
BOF	1 000	1 000	0.401	0.444	0.434	0.408
	20 000	20 000	0.404	0.452	0.445	0.416
Fisher (μ)	16	2048	0.497	0.490	0.475	0.452
	64	8192	0.495	0.492	0.464	0.424
VLAD	16	2048	0.496	0.495	0.494	0.451
	64	8192	0.526	0.510	0.477	0.421

Πίνακας 2.1: Σύγκριση των BOF, Fisher και VLAD αναπαραστάσεων, πριν και μετά την μείωση των διαστάσεων του διανύσματος.

και Fisher στις οποίες βασίστηκε, όπως αναφέραμε και στο κεφάλαιο 1.2.2. Όπως φαίνεται οι μέθοδοι Fisher και VLAD που συγκρατούν την τοπική δομή κάθε των περιγραφέων έχουν καλύτερα αποτελέσματα και στο μειωμένο αλλά και στο πλήρες διάνυσμα σε σχέση με την BOF. Επίσης παρατηρεί κανείς πως όταν χρησιμοποιούμε το πλήρες διάνυσμα VLAD το μέγεθος του λεξικού επηρεάζει θετικά το αποτέλεσμα. Όμως όταν μειώνουμε το μέγεθος της τελικής αναπαράστασης όπως φαίνεται στη στήλη για $D' = 64$ το λεξικό $k = 16$ έχει καλύτερα αποτελέσματα. Όπως φαίνεται το σφάλμα, που προκύπτει από την προβολή του πλήρους διανύσματος στον υπόχωρο που σχηματίζει το PCA, επηρεάζει δραματικά την απόδοση της αναπαράστασης.

Στο σχήμα 2.4 φαίνεται πως αν σκοπεύουμε να κωδικοποιήσουμε το τελικό αποτέλεσμα, η επιλογή ενός μεγαλύτερου λεξικού δεν επηρεάζει θετικά την αναπαράσταση. Ο λόγος είναι πως το σφάλμα που προκύπτει από την προβολή του πλήρους VLAD στο D' που σχηματίζουμε με το PCA είναι πολύ μεγάλο όταν οι συνιστώσες που κρατάμε στο μειωμένο είναι πολύ λίγες συγκριτικά με τις αρ-



Σχήμα 2.4: Αποτελέσματα πειράματος του [13], όπου φαίνεται η απόδοση της VLAD για διαφορετικά D' κωδικοποιημένα με ADC 16×8. Το αποτέλεσμα προκύπτει ως ο μέσος όρος για 5 διαφορετικά λεξικά. Σαν σφάλμα στην μέτρηση φαίνεται η απόκλιση των τιμών στα διαφορετικά αυτά λεξικά.

χικές. Συγκεκριμένα παρατηρούμε ότι ένα λεξικό $k = 64$ είναι αρκετό για την δημιουργία ενός συμπαγούς διανύσματος VLAD, π.χ. $D' = 128$ διαστάσεων. Για το λόγο αυτό στο κεφάλαιο 3 θα χρησιμοποιήσουμε και εμείς λεξικό $k = 64$ για να συγκρίνουμε τις μεθόδους.

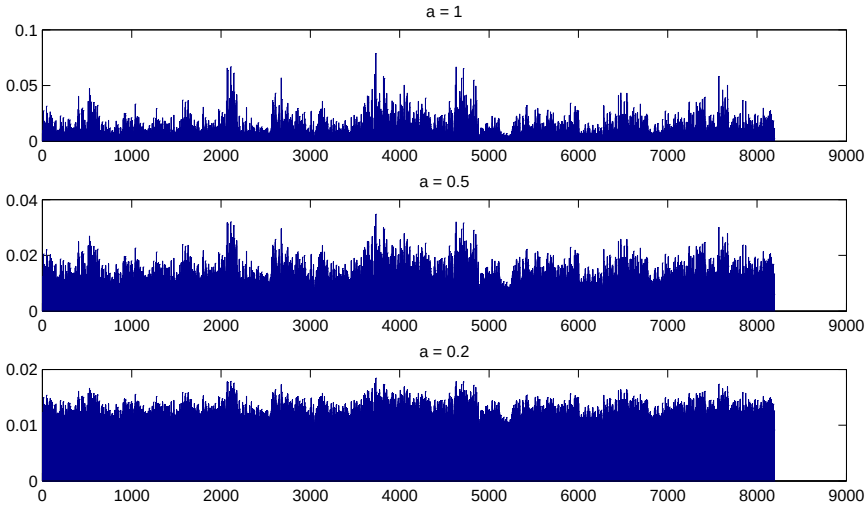
2.2 Βελτιώσεις της βασικής μεθόδου

Στην υπάρχουσα βιβλιογραφία έχουν προταθεί πολλοί τρόποι για περαιτέρω βελτίωση της VLAD διανυσματικής αναπαράστασης. Στην ενότητα αυτή θα παρουσιάσουμε θεωρητικά τις πιο υποσχόμενες από τις μεθόδους και στη συνέχεια στο κεφάλαιο 3 θα προχωρήσουμε στην υλοποίηση και την αξιολόγησή τους.

Η πρώτη βελτίωση που προτάθηκε στο [14] είναι να γίνει κανονικοποίηση με ύψωση σε δύναμη (power-law) ανά συνιστώσα του πλήρους διανύσματος V, εξίσωση 2.7. Η power-law κανονικοποίηση χρησιμοποιείται ευρέως σε αντίστοιχες αναπαραστάσεις, όπως την FV [8] για να αντιμετωπίσει την συσσώρευση της ενέργειας των SIFT σε συγκεκριμένες διαστάσεις. Το α παίρνει τιμές $0 \leq \alpha \leq 1$, και θεωρητικά εξισορροπεί τη διακύμανση του μέτρου των στοιχείων του διανύσματος. Το βήμα αυτό δίνει θετικά αποτελέσματα στις μεθόδους BOF και FV και φαίνεται να βελτιώνει τα αποτελέσματα και του VLAD αλγορίθμου.

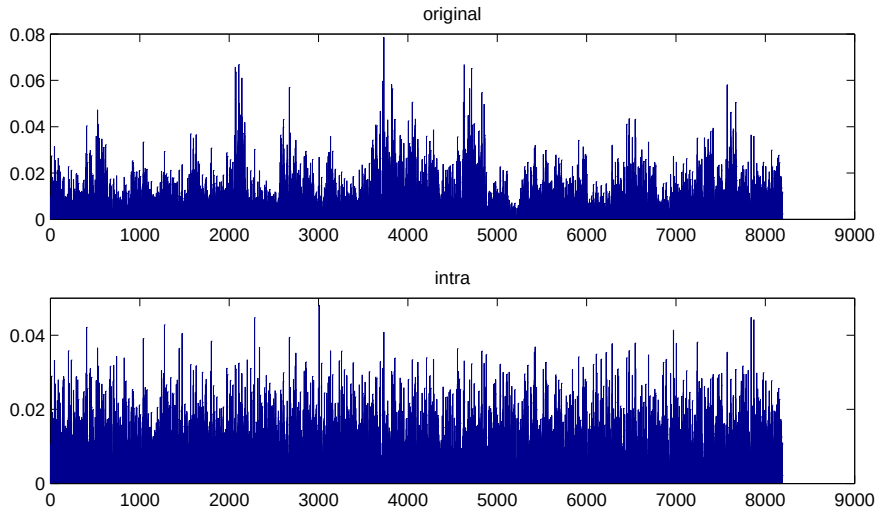
$$V_u := sign(V_u)|V_u|^\alpha \ , \qquad \forall u = 1, ..., kd \tag{2.7}$$

Η επίδραση του γίνεται εμφανής στο σχήμα 2.5, στο οποίο φαίνεται το πλάτος των διαστάσεων μιας VLAD αναπαράστασης πριν και μετά την εφαρμογή του power-law για διάφορες τιμές της παραμέτρου α. Η τιμή της παραμέτρου a που επιλέγουν να χρησιμοποιήσουν στο [14] είναι $a = 0.5$ και για το λόγο αυτό ονομάζεται και SSR-κανονικοποίηση (Signed Square Root). Η δεύτερη πρόταση [14] είναι να επεξεργαστούμε τους SIFT περιγραφείς πριν σχηματίσουμε τα τοπικά VLAD διανύσματα. Συγκεκριμένα προτείνεται να εφαρμόσουμε PCA στον χώρο



Σχήμα 2.5: Πλήρες VLAD για λεξικό k = 64. Επίδραση της power-κανονικοποίησης.

των SIFT πριν τον υπολογισμό των υπολοίπων. Η περιστροφή αυτή φαίνεται να ενισχύει την power-law κανονικοποίηση. Για να σιγουρέψουμε ότι δεν είναι τυχαίο φαινόμενο εφαρμόσαμε μια τυχαία περιστροφή στους περιγραφείς μετά το PCA και είδαμε πως χειροτερεύει την απόδοση της αναπαράστασης. Επίσης στο [14] προτείνεται μαζί με την περιστροφή να μειώσουμε τις διαστάσεις των διανυσμάτων SIFT από 128 σε 64 και συνεπώς την διάσταση των τοπικών VLAD που σχηματίζουμε ανά οπτική λέξη. Αυτό έχει ως στόχο να κόψει τα λιγότερο ενεργητικά στοιχεία από κάθε i, γιατί τα στοιχεία θεωρείται πως λειτουργούν ως θόρυβος στο τελικό PCA του πλήρες VLAD. Η μείωση αυτή των διαστάσεων φαίνεται να επηρεάζει θετικά το τελικό αποτέλεσμα για τις μεθόδους FV και BOF, όταν γίνεται μαζί με power-law κανονικοποίηση, αλλά όχι τόσο στην περίπτωση του VLAD διανύσματος, κάτι που θα εξετάσουμε με περισσότερη λεπτομέρεια στο κεφάλαιο

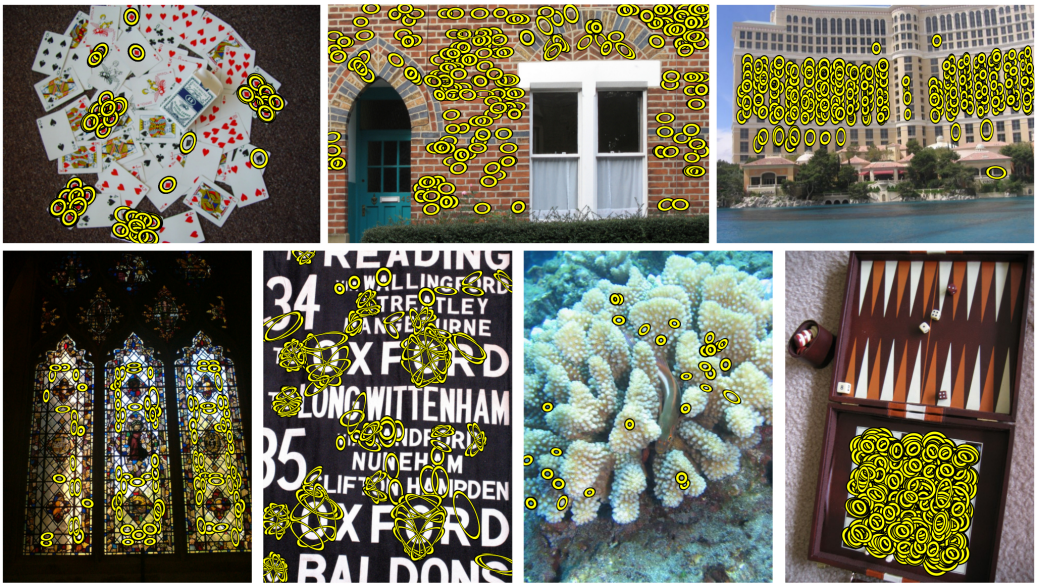


Σχήμα 2.6: Πλήρες VLAD για λεξικό k = 64. Επίδραση της intra-κανονικοποίησης.

3.

Στο [2] προτείνεται ένας διαφορετικός τρόπος κανονικοποίησης από αυτούς που αναφέραμε προηγουμένως. Η μέθοδος ονομάζεται intra-κανονικοποίηση (Intra normalization) και προτείνεται να γίνει μια L_2 κανονικοποίηση των τοπικών VLAD v_i και στη συνέχεια η L_2 κανονικοποίηση του πλήρους VLAD V. Η επίδραση της κανονικοποίησης αυτής στο διάνυσμα φαίνεται στο σχήμα 2.6. Όπως παρατηρεί κανείς δεν μας ενδιαφέρει πλέον να μειώσουμε τη διασπορά του μέτρου των συνιστωσών αλλά τη διασπορά του πλήθους των περιγραφέων ανά οπτική λέξη.

Οι κανονικοποιήσεις αυτές που αναφέραμε στοχεύουν στην αντιμετώπιση του φαινομένου που παρατηρείται στο [10] και ονομάζεται ”burstiness” των οπτικών χαρακτηριστικών. Όταν σε μια εικόνα επαναλαμβάνεται το ίδιο αντικείμενο (π.χ. ένα πάτωμα με τετράγωνα πλακάκια), είτε το αντικείμενο της εικόνας έχει κά-



Σχήμα 2.7: Παραδείγματα εικόνων στις οποίες επαναλαμβάνονται τα ίδια αντικείμενα. Φαινόμενο burstiness.

ποια υφή, κατά την εξαγωγή των περιγραφέων θα προκύψουν πολλοί περιγραφείς στην ίδια περιοχή. Αυτό έχει σαν αποτέλεσμα να συγκεντρωθούν στην ίδια οπτική λέξη και στην διανυσματική αναπαράσταση που θα σχηματιστεί να συγκεντρώνεται η ενέργεια του διανύσματος στις διαστάσεις που αφορούν το επαναλαμβανόμενο αυτό χαρακτηριστικό. Παραδείγματα τέτοιων εικόνων φάινονται στην εικόνα 2.7. Οι αναπαραστάσεις των εικόνων αυτών θα έχουν συγκεντρωμένη ενέργεια σε πολύ λίγες συνιστώσες με αποτέλεσμα οι υπόλοιπες συνιστώσες που μεταφέρουν και αυτές χρήσιμη πληροφορία να εκμηδενίζονται αν δεν διαχειριστεί κανείς το φαινόμενο αυτό.

Ο στόχος της power-law (SSR) κανονικοποίησης είναι να μειώσει το μέτρο των συνιστωσών που έχουν συγκεντρώσει την περισσότερη ενέργεια. Αντίθετα η intra κανονικοποίηση στοχεύει στο να αντιμετωπίσει το πρόβλημα της συγκέ-

ντρωσης των περιγραφέων όχι σε συγκεκριμένες συνιστώσες αλλά σε μια οπτική λέξη. Στο [2] θεωρούν πως με power-law κανονικοποίηση, παρότι βελτιώνει το αποτέλεσμα, το μόνο που πετυχαίνει κανείς είναι να μειώσει το πρόβλημα, ενώ με την intra κανονικοποίηση εξαλείφεται πλήρως. Ο λόγος είναι πως κανονικοποιώντας τα τοπικά VLAD διανύσματα κάθε οπτικής λέξης αναγκάζεις το τελικό διάνυσμα να έχει ίση ενέργεια ανά οπτική λέξη. Η μέθοδος αυτή επηρεάζεται σε μεγάλο βαθμό από το μέγεθος και την ποιότητα του λεξικού, όπως θα παρουσιάσουμε εκτενέστερα στην ενότητα 3.3.

Στο [6] θεωρούν πως κάθε περιγραφέας πρέπει να έχει την ίδια συνεισφορά στη δημιουργία του τοπικού VLAD. Συγκεκριμένα, όπως είναι φυσικό, το υπόλοιπο που θα σχηματιστεί από ένα περιγραφέα που βρίσκεται μακριά από την οπτική λέξη θα έχει πολύ μεγαλύτερο μέτρο από έναν που βρίσκεται κοντά. Αυτό θα έχει σαν αποτέλεσμα να επηρεάσει σε πολύ μεγαλύτερο βαθμό το διάνυσμα που θα σχηματιστεί από την διανυσματικό άθροισμα των υπολοίπων. Ταυτόχρονα, το σύνολο των περιγραφέων που βρίσκονται πολύ κοντά στην οπτική λέξη, θα έχουν πολύ μικρό μέτρο και δε θα επηρεάσουν σχεδόν καθόλου το τοπικό VLAD διάνυσμα. Για την αποφυγή του φαινομένου αυτού προτείνεται, στο [6], να γίνει L_2-κανονικοποίηση στο υπόλοιπο κάθε περιγραφέα πριν το σχηματισμό του τοπικού VLAD για κάθε περιγραφέα x όπως φαίνεται στην εξίσωση 2.8. Η κανονικοποίηση αυτή ονομάζεται κανονικοποίηση υπολοίπου (Residual Normalization) στην οποία θα αναφερόμαστε ως RN.

$$v_i = \sum_{q(x)=c_i} \frac{x - c_i}{\|x - c_i\|} , \qquad i = 1, .., k \tag{2.8}$$

Για την σωστότερη λειτουργία της μεθόδου προτείνεται να γίνει μια περιστροφή του υπολοίπου πριν το διανυσματικό άθροισμα, εξίσωση 2.9. Η περιστροφή αυτή προκύπτει από την εφαρμογή ενός PCA ανά οπτική λέξη στο χώρο των SIFT των εικόνων του συνόλου εκπαίδευσης. Η περιστροφή αυτή διαφέρει από αυτή που εφαρμόσαμε σε ολόκληρο το χώρο των SIFT και ονομάζεται LCS (Local Coordinate System - Τοπικό σύστημα συντεταγμένων). Ταυτόχρονα με την περιστροφή όπως και στην περίπτωση του PCA σε όλο τον χώρο μας δίνεται η δυνατότητα να μειώσουμε τις διαστάσεις των διανυσμάτων μετά την περιστροφή. Έστω Q_i η περιστροφή ανά οπτική λέξη, η εξίσωση 2.8 μετατρέπεται στην 2.9.

$$v_i = \sum_{q(x)=c_i} Q_i \frac{x - c_i}{\|x - c_i\|} , \qquad i = 1, .., k \tag{2.9}$$

Σημειώνουμε ότι μπορούμε να εφαρμόσουμε την μέθοδο LCS ανεξάρτητα από την μέθοδο κανονικοποίησης που θα χρησιμοποιήσουμε. Όπως θα δούμε και στο κεφάλαιο 3 επιλέγουμε να επεξεργαστούμε τους SIFT περιγραφείς είτε σε όλο τον χώρο είτε ανά οπτική λέξη. Βέβαια, όπως φαίνεται και από τα πειραματικά αποτελέσματα του κεφαλαίου 3, η μέθοδος LCS φαίνεται να επιδρά θετικά μόνο στην περίπτωση της RN κανονικοποίησης.

Μέχρι το σημείο αυτό έχουμε παρουσιάσει μεθόδους της υπάρχουσας βιβλιογραφίας που αφορούν τον τρόπο με τον όποιο σχηματίζεται το πλήρες VLAD διάνυσμα. Αναφέρονται στο πως μπορεί κανείς να διαχειριστεί την συγκέντρωση της ενέργειας του διανύσματος σε λίγες συνιστώσες ή σε συγκεκριμένες οπτικές λέξεις. Εκτός από τα προβλήματα αυτά στο [12] θεωρούν πως και ο τρόπος που γίνεται η μείωση των διαστάσεων επηρεάζει αρνητικά το αποτέλεσμα. Το πρό-

βλημα που εξετάζουν, παρατηρήθηκε πρώτα στο [3], εμφανίζεται όταν δύο ή παραπάνω περιγραφείς εμφανίζονται μαζί σε εικόνες (visual word co-occurrences). Επίσης, ο ανιχνευτής περιγραφέων μπορεί να εισάγει τεχνητές ομάδες περιγραφέων σε μια εικόνα, όταν π.χ. έχουμε μια εικόνα που εμφανίζει την περιοχή με διαφορετικές κατευθύνσεις. Το πρόβλημα αυτό δεν αντιμετωπίζεται με το PCA, γιατί συγκεντρώνει στις κυρίαρχες συνιστώσες του τελικού μειωμένου διανύσματος αποκρίσεις που προκύπτουν από τό φαινόμενο αυτό, επηρεάζοντας αρνητικά την αναπαράσταση.

Για την αντιμετώπισή του προτείνουν στο [12] να γίνει μια εξισορρόπηση των ιδιοτιμών (whitening) των δεδομένων, όπως στην ανάλυση ανεξάρτητων συνιστωσών στο [4]. Συγκεκριμένα στην παραγωγή του τελικού διανύσματος, αφού έχουμε βρει τον πίνακα $M_{D' \times D}$ με ανάλυση κυρίαρχων συνιστωσών, σχηματίζουμε το μειωμένο διάνυσμα V' όπως φαίνεται στην εξίσωση 2.10, όπου λ_i οι ιδιοτιμές που αντιστοιχούν στα i μεγαλύτερα ιδιοδιανύσματα του πίνακα συνδιακύμανσης. Μετά τη διαδικασία αυτή L_2-κανονικοποιούμε το μειωμένο τελικό διάνυσμα για να μπορούμε να συγκρίνουμε τις τελικές αναπαραστάσεις με το εσωτερικό τους γινόμενο.

$$V' = \frac{diag(\lambda_1^{-\frac{1}{2}}, ..., \lambda_{D'}^{-\frac{1}{2}})MV}{\|diag(\lambda_1^{-\frac{1}{2}}, ..., \lambda_{D'}^{-\frac{1}{2}})MV\|} \tag{2.10}$$

Μια ακόμη πρόταση για τη βελτίωση της αναπαράστασης μετά τη μείωση των διαστάσεων παρουσιάζεται στο [15]. Προτείνεται να γίνει μια κανονικοποίηση με ύψωση σε δύναμη (power-law) μετά την περιστροφή της αναπαράστασης στο βήμα που γίνεται το PCA. Αποτελεί έναν διαφορετικό τρόπο αντιμετώπισης του

προβλήματος που παρουσιάσαμε στην προηγούμενη παράγραφο. Οι συγγραφείς του [15], θεωρούν πως ο υπολογισμός των ιδιοτιμών που απαιτεί το ”whitening” προσθέτει υπολογιστικό κόστος στον αλγόριθμο χωρίς να επιτυγχάνει καλύτερα αποτελέσματα από μια απλή (υπολογιστικά) power-law κανονικοποίηση και επίσης θεωρούν πως μια μέθοδος που εξαρτάται από τις ιδιοτιμές, που προκύπτουν από το σύνολο εκπαίδευσης, θα έχει λιγότερη σταθερή απόδοση. Σαν τελευταίο βήμα συγκρατούν τις D' συνιστώσες του παραχθέντος διανύσματος και όπως όλες οι μέθοδοι L_2- κανονικοποιούν το τελικό διάνυσμα.

Κεφάλαιο 3

Υλοποίηση και σύγκριση διανυσματικών αναπαραστάσεων εικόνων

Στο κεφάλαιο αυτό υλοποιούμε τις μεθόδους που παρουσιάσαμε στο προηγούμενο κεφάλαιο και αξιολογούμε την επίδραση τους στα αποτελέσματα της αναζήτησης. Επίσης παρουσιάζουμε και δικές μας προτάσεις για τη δημιουργία της αναπαράστασης. Σαν πρώτο βήμα παρουσιάζουμε την πειραματική διάταξη που χρησιμοποιήσαμε για την σύγκριση των διαφόρων μεθόδων.

3.1 Πειραματική διάταξη

Τα πειράματα γίνονται πάνω στο ίδιο σύνολο εκπαίδευσης, με το ίδιο λεξικό και το mAP μετριέται πάνω στο ίδιο σύνολο αξιολόγησης. Το οπτικό λεξικό δημιουργήθηκε με τον ίδιο αλγόριθμο ομαδοποίησης και συγκεκριμένα τον kmeans.

Όπως αναφέραμε και στην ενότητα 1.2.3 το σύνολο της εκπαίδευσης είναι το Oxford100k [20]. Το μέτρο της απόστασης είναι η ευκλείδεια απόσταση, που ανάγεται σε εσωτερικό γινόμενο μετά την L_2- κανονικοποίηση, και όλες οι αναζητήσεις στον χώρο των περιγραφέων γίνονται εξαντλητικά. Βρέθηκαν λεξικά για $k = 16, 32, 64$ και για κάθε k βρήκαμε τρία διαφορετικά λεξικά. Η αρχικοποίηση του αλγορίθμου δεν έγινε τυχαία όπως συνηθίζεται.

Στην δική μας υλοποίηση επιλέγεται τυχαία η πρώτη λέξη και έπειτα υπολογίζεται η απόσταση της από όλους τους περιγραφείς. Η επόμενη οπτική λέξη επιλέγεται με μια πιθανότητα που είναι ανάλογη της απόστασης από την προηγούμενη λέξη. Στη συνέχεια ο αλγόριθμος kmeans εκπαιδεύεται μέχρι να συγκλίνουν οι τιμές. Το κριτήριο της σύγκλισης που επιλέξαμε είναι το άθροισμα της μετατόπισης των k λέξεων μετά από κάθε επανάληψη. Στόχος ήταν το λεξικό να είναι όσο το δυνατό καλύτερο. Σε ένα πραγματικό σύστημα θα ήταν υπολογιστικά αδύνατο να υλοποιηθεί μια τέτοια εξαντλητική μέθοδος, όμως στα πλαίσια της διπλωματικής ένα ποιοτικό αρχικό οπτικό λεξικό μας είναι χρήσιμο για μια δίκαιη σύγκριση των διαφορετικών μεθόδων.

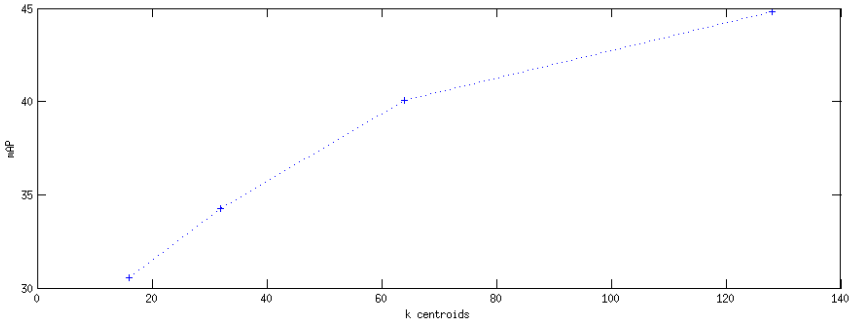
Σε όποιο σημείο χρησιμοποιήθηκε ανάλυση βασικών συνιστωσών PCA, η κατασκευή του πίνακα συνδιακύμανσης έγινε σε ένα υποσύνολο εικόνων του Oxford 100k. Δεν χρησιμοποιήσαμε όλο το σύνολο των εικόνων για μεγαλύτερη ταχύτητα στην κατασκευή του πίνακα. Θεωρούμε πως η επιλογή αυτή δεν επηρεάζει τα συγκριτικά αποτελέσματα των διαφορετικών μεθόδων, δεδομένου ότι τα σύνολα εκπαίδευσης ήταν κοινά για όλες τις μεθόδους και η διαδικασία επαναλήφθηκε πολλαπλές φορές με τυχαία επιλογή του υποσυνόλου. Η τελική αναπαράσταση αξιολογήθηκε στο Oxford5k, όπου το μέτρο της αξιολόγησης είναι το mAP, για κάθε διαφορετική μέθοδο.

3.2 Υλοποίηση των μεθόδων του κεφαλαίου 2

Στην ενότητα αυτή υλοποιούμε τις μεθόδους που παρουσιάσαμε θεωρητικά στο κεφάλαιο 2. Η υλοποίηση αυτή γίνεται στην πειραματική διάταξη που παρουσιάσαμε στην προηγούμενη ενότητα. Στόχος του κεφαλαίου είναι να επιλέξουμε ποια μέθοδος έχει τα καλύτερα αποτελέσματα στο σύνολο εικόνων Oxford5k για να την χρησιμοποιήσουμε στη δημιουργία καινούργιων συμπαγών αναπαραστάσεων. Επίσης, μέσα από τη σύγκριση των μεθόδων συμπεραίνουμε ποιες από τις τεχνικές βελτίωσης που προτείναμε στο 2.2 επιδρούν καλύτερα στην δική μας πειραματική διάταξη. Πρώτο βήμα στη δημιουργία των αναπαραστάσεων είναι να επιλέξουμε το μέγεθος του οπτικού λεξικού που θα χρησιμοποιήσουμε.

Στο σχήμα 3.1 φαίνεται η απόδοση στο Oxford5k της βασική μέθοδου για λεξικά διαφορετικού μεγέθους στην οποία γίνεται μόνο L2-κανονικοποίηση του πλήρες VLAD και εφαρμόγη PCA για τη μείωση των διαστάσεων. Είναι εμφανές ότι στη βασική μέθοδο όσο μεγαλύτερο είναι το λεξικό τόσο καλύτερη είναι η απόδοση της τελικής αναπαράστασης. Το αποτέλεσμα αυτό είναι απόλυτα λογικό δεδομένου ότι οι διαστάσεις D του τελικού διανύσματος θα αυξάνουν όσο μεγαλύτερο είναι το λεξικό $D = k \times d$.

Στην περίπτωση που θέλουμε να σχηματίσουμε μια συμπαγή αναπαράσταση, όπως φαίνεται και στο σχήμα 2.4, πρέπει η τελική μειωμένη αναπαράσταση να είναι συνήθως από 96 έως 256 διαστάσεις. Αυτό έχει σαν αποτέλεσμα το σφάλμα που εισάγει η μείωση αυτή να επηρεάζεται όπως ήταν αναμενόμενο από το μέγεθος της αρχικής πλήρους αναπαράστασης. Για το λόγο αυτό και για υπολογιστικούς λόγους επιλέξαμε να χρησιμοποιήσουμε λεξικά μεγέθους $k = 64$ σε όλα τα υπόλοιπα πειράματα. Στο σημείο αυτό συγκρίνουμε τις διαφορετικές μεθόδους που έχουν προταθεί στην υπάρχουσα βιβλιογραφία για τη βελτίωση του αποτελέ-



Σχήμα 3.1: Η απόδοση της VLAD ως προς το μέγεθος του λεξικού k.

σματος. Στον πίνακα 3.1 φαίνονται τα αποτελέσματα των διαφορετικών μεθόδων. Σημειώνουμε πως οι μέθοδοι που εξετάζουμε αφορούν στον τρόπο που γίνεται η κανονικοποίηση του πλήρους διανύσματος καθώς και την επεξεργασία των SIFT περιγραφέων. Στο 2 είχαμε αναφέρει ότι προτείνεται να γίνει κάποια επεξεργασία των SIFT διαφορών με χρήση PCA πριν αθροιστούν για το σχηματισμό του VLAD διανύσματος. Το PCA μπορεί να αναλυθεί σε 3 βήματα C(Centering), R(Rotation) και D(Dimensionality Reduction). Η υποσημείωση CR σημαίνει πως έχει εφαρμοστεί PCA στους SIFT περιγραφείς για απλή περιστροφή, ενώ το CRD σημαίνει ότι έχει γίνει μείωση των διαστάσεων μετά την περιστροφή. Για την παραγωγή του μειωμένου διανύσματος $D' = 128$ εφαρμόσαμε PCA όπως περιγράψαμε στο 3.1, χωρίς καμία περαιτέρω βελτίωση.

Οι μέθοδοι που εξετάζονται στον πίνακα 3.1 είναι οι παρακάτω:

1. VLAD: Power-law και L2- κανονικοποίηση του πλήρους VLAD.

2. VLAD_{CRD}: Περιστροφή των SIFT περιγραφέων με χρήση PCA και μείωση των διαστάσεων από 128 σε 64. Power-law και L2- κανονικοποίηση του πλήρους VLAD.

Περιγραφέας	Πλήρες VLAD (mAP %) $a = 0.2$	$a = 0.5$	$D' = 128$ (mAP %) $a = 0.2$	$a = 0.5$
VLAD	43.1	43.7	34.2	34
VLAD_{CRD}	42.9	42.2	32.7	31.4
VLAD_{CR}	48.7	50.8	34.8	35.4
$\text{VLAD}_{\text{LCS+CRD}}$	43.2	46.4	31.8	32.9
$\text{VLAD}_{\text{LCS+CR}}$	46.9	49.2	32.2	33.7
$\text{VLAD}_{\text{LCS+RN}}$	52.4	49.2	33.1	31.9
$\text{VLAD}_{\text{Intra}}$	40.8		30.8	

Πίνακας 3.1: Υλοποίηση γνωστών μεθόδων που παρουσιάστηκαν στο κεφάλαιο 2, για διαφορετικά α και διαστάσεις του τελικού διανύσματος. Οι αναπαραστάσεις αξιολογούνται με τη μετρική mAP (%).

3. $VLAD_{CR}$: Περιστροφή των περιγραφέων με χρήση PCA. Power-law και L2-κανονικοποίηση του πλήρους VLAD.

4. $VLAD_{LCS+CRD}$: Περιστροφή των περιγραφέων ανά τοπικό κελί με χρήση PCA και μείωση των διαστάσεων από 128 σε 64. Power-law και L2- κανονικοποίηση του πλήρους VLAD.

5. $VLAD_{LCS+CR}$: Περιστροφή των περιγραφέων ανά τοπικό κελί με χρήση PCA. Power-law και L2- κανονικοποίηση του πλήρες VLAD.

6. $VLAD_{LCS+RN}$: Περιστροφή των περιγραφέων ανά τοπικό κελί και L2- κανονικοποίηση υπολοίπου. Power-law και L2- κανονικοποίηση του πλήρους VLAD.

7. $VLAD_{Intra}$: L2-κανονικοποίηση του πλήρους VLAD ανά κελί και L2- κανονικοποίηση ολόκληρου του διανύσματος.

3.3 Συγκριτική Αξιολόγηση

Για την δίκαιη σύγκριση των διαφορετικών αναπαραστάσεων υλοποιήσαμε ένα σύστημα ανάκτησης εικόνων πάνω στο ίδιο σύνολο εικόνων, συγκεκριμένα στο Oxford5k. Η κάθε αναπαράσταση αξιολογείται από την μετρική mAP. Εκτός από το σύνολο στο οποίο γίνεται η ανάκτηση πρέπει και τα σύνολα που χρησιμοποιούνται για την εκπαίδευση των αλγορίθμων να είναι ίδια. Για το λόγο αυτό δεν είναι δυνατό να συγκρίνουμε άμεσα τα αποτελέσματα των δικών μας πειραμάτων με αυτά που έχουν οι αλγόριθμοι στις αντίστοιχες δημοσιεύσεις που παρουσιάζονται.

Η βασική διαφορά είναι η ποιότητα του λεξικού. Στη δική μας υλοποίηση χρησιμοποιούμε εξαντλητική αναζήτηση κοντινότερου γείτονα καθώς και καλύτερη αρχικοποίηση στον σχηματισμό του λεξικού. Αποτέλεσμα της διαφοράς αυτής είναι ότι έχουμε καλύτερη απόδοση ακόμη και στη βασική μέθοδο του VLAD αλγορίθμου.

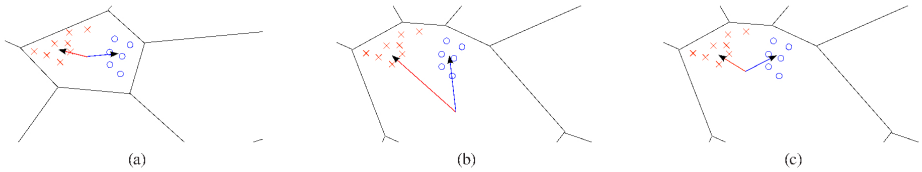
Η βασική VLAD αναπαράσταση έχει απόδοση 39% για το πλήρες διάνυσμα και 29,5% για το μειωμένο διάνυσμα 128 διαστάσεων. Επομένως, όπως φαίνεται από τον πίνακα 3.1, όλες οι αναπαραστάσεις που έχουμε υλοποιήσει έχουν καλύτερη απόδοση από την βασική μέθοδο. Ακόμη και μια απλή power-law κανονικοποίηση βελτιώνει σε μεγάλο βαθμό τα αποτελέσματα, ειδικότερα στην περίπτωση του μειωμένου διανύσματος. Συγκεκριμένα παρατηρούμε ότι η επιλογή του α δεν επηρεάζει σε μεγάλο βαθμό το αποτέλεσμα.

Από τα υπόλοιπα αποτελέσματα παρατηρούμε ότι όταν επεξεργαζόμαστε τους SIFT περιγραφείς πριν σχηματίσουμε τα τοπικά VLAD βελτιώνουμε σε μεγάλο βαθμό την ποιότητα της αναπαράστασης. Η επεξεργασία όπως έχουμε αναφέρει στοχεύει στο να περιστρέψουμε των χώρο τον SIFT των εικόνων του συνόλου εκπαίδευσης ώστε να λειτουργήσει καλύτερα η power-law κανονικοποίηση. Βέβαια, στην πρώτη υλοποίηση της μεθόδου προτάθηκε εκτός από την περιστροφή να κρατήσουμε μόνο τις μισές συνιστώσες δηλαδή να μειώσουμε τις διαστάσεις των SIFT από 128 σε 64 κρατώντας μόνο τις πρώτες μισές κυρίαρχες συνιστώσες. Αυτό λειτουργεί σωστά στις μεθόδους BOF και Fisher αλλά όπως φαίνεται από τα αποτελέσματα δεν λειτουργεί θετικά στην περίπτωση της VLAD αναπαράστασης.

Σημαντικό ρόλο στην απόδοση της αναπαράστασης παίζει η επιλογή του συστήματος συντεταγμένων στο οποίο επεξεργαζόμαστε τους SIFT. Παρατηρούμε πως όταν περιστρέφουμε τους SIFT περιγραφείς ανά οπτική λέξη έχουμε καλύ-

τερα αποτελέσματα στην περίπτωση του πλήρους διανύσματος αλλά χειρότερα στην περίπτωση του μειωμένου. Συγκεκριμένα παρατηρώντας τα αποτελέσματα του AP στις διαφορετικές εικόνες ερωτημάτων, η χρήση τοπικού LCS επηρεάζει θετικά τις εικόνες με λίγους περιγραφείς. Θεωρούμε πως ο λόγος για τον οποίο συμβαίνει αυτό είναι ότι χρησιμοποιώντας τοπικό σύστημα συντεταγμένων δεν επηρεάζεται η αναπαράσταση από την κατανομή των SIFT περιγραφέων στο οπτικό λεξικό. Το πλήθος των SIFT που αντιστοιχεί σε κάθε οπτική λέξη δεν είναι σταθερό σε όλες τις λέξεις. Επομένως, όταν χρησιμοποιεί κανείς συνολικό σύστημα συντεταγμένων για όλο των χώρο των SIFT οι περιγραφείς που ανήκουν στις ”αραιές” οπτικές λέξεις επηρεάζονται αρνητικά από την επεξεργασία.

Στην περίπτωση του $\text{VLAD}_{\text{Intra}}$ τα αποτελέσματα απέχουν σε μεγάλο βαθμό από εκείνα που παρουσιάζονται στο [2]. Ο λόγος είναι πως το λεξικό που χρησιμοποιούν στην παρουσίαση της μεθόδου είναι πολύ πιο πυκνό από αυτό που χρησιμοποιούμε στην δική μας πειραματική διάταξη. Συγκεκριμένα στη δημοσίευση χρησιμοποιούν λεξικό $k = 256$ και εκτός αυτού χρησιμοποιούν μεθόδους για περαιτέρω βελτίωση. Η μέθοδος που χρησιμοποιούν ονομάζεται μέθοδος προσαρμογής κέντρων (cluster center adaptation). Ως πρώτο βήμα, βρίσκουμε από το σύνολο εικόνων στο οποίο θα γίνει η αξιολόγηση του αλγορίθμου όλους του περιγραφείς που αντιστοιχούν σε μια οπτική λέξη και να προσαρμόσουμε το κέντρο της λέξης αυτής ως το μέσο όρο των περιγραφέων αυτών. Τα VLAD διανύσματα προκύπτουν ως άθροισμα των διαφορών και συγκρίνονται με χρήση εσωτερικού γινομένου. Στην περίπτωση του $\text{VLAD}_{\text{Intra}}$ τα τοπικά VLAD (βελάκια στο σχήμα) L2- κανονικοποιούνται, το οποίο σημαίνει ότι όταν συγκριθούν μεταξύ τους με χρήση εσωτερικού γινομένου, μόνο γωνία μεταξύ των διανυσμάτων θα συνεισφέρει. Όπως φαίνεται από το σχήμα 3.2 αν το κέντρο μιας κυψέλης είναι όπως στην



Σχήμα 3.2: Παράδειγμα της μεθόδου προσαρμογής κέντρων [2]. (α) Στην περίπτωση αυτή δεν είναι αναγκαία η προσαρμογή του κέντρου. Στο (β) φαίνεται μια περίπτωση που οι ίδιοι περιγραφείς σε μια χειρότερη ομαδοποίηση θα παράξουν VLAD διανύσματα με πολύ μικρότερη διακριτική ικανότητα. Στο (γ) φαίνεται το αποτέλεσμα της μεθόδου προσαρμογής κέντρων αν χρησιμοποιηθεί στο (β).

περίπτωση του (β) η διακριτική ικανότητα των διανυσμάτων που θα προκύψουν θα είναι ελλιπής. Αν όμως προσαρμόσουμε το κέντρο της κυψέλης όπως φαίνεται στο (γ) τα $\text{VLAD}_{\text{Intra}}$ διανύσματα θα είναι σχεδόν αντιδιαμετρικά.

Συνολικά, παρατηρούμε ότι για τα πλήρη διανύσματα το $\text{VLAD}_{\text{CR+SSR}}$ έχει την καλύτερη απόδοση μετά το $\text{VLAD}_{\text{LCS+RN}}$ για $a = 0.2$. Ο λόγος που θεωρούμε ότι προκύπτει αυτό είναι ότι το λεξικό που έχουμε παράξει είναι καλής ποιότητας με αποτέλεσμα η κανονικοποίηση υπολοίπου να έχει θεαματικά αποτέλεσμα όταν χρησιμοποιείται το πλήρες διάνυσμα. Βέβαια στην περίπτωση του μειωμένου διανύσματος η μέθοδος $\text{VLAD}_{\text{LCS+RN}}$ έχει πολύ καλή απόδοση. Επίσης, παρατηρούμε πως η μείωση των διαστάσεων των SIFT περιγραφέων από 128 σε 64 δεν επηρεάζει πολύ αρνητικά την τελική αναπαράσταση και στην περίπτωση που χρησιμοποιούμε τοπικό σύστημα συντεταγμένων. Επομένως αν θέλει να επεξεργαστεί κανείς τους SIFT πριν την παραγωγή του VLAD καλό είναι να χρησιμοποιήσει απλώς το PCA για περιστροφή χωρίς να μειώσει τις διαστάσεις τους. Τέλος, για την power-law κανονικοποίηση η επιλογή $a = 0.5$, γνωστή και ως SSR, είναι

προτιμότερη από το $a = 0.2$.

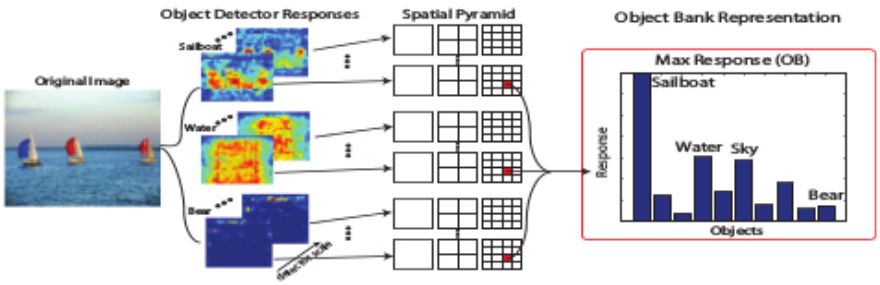
Κεφάλαιο 4

Νέες μέθοδοι σχηματισμού αναπαραστάσεων εικόνων

Στο κεφάλαιο αυτό της εργασίας προτείνουμε διαφορετικούς τρόπους για τον σχηματισμό διανυσματικών αναπαραστάσεων. Η πρώτη ενότητα αφορά μεθόδους που βασίζονται στην VLAD αναπαράσταση που έχουμε σχηματίσει, ενώ η δεύτερη αφορά μεθόδους για περαιτέρω βελτίωσή της.

4.1 Αναπαραστάσεις με βάση την αναπαράσταση VLAD

Στόχος της ενότητας αυτής είναι να σχηματίσουμε μια αναπαράσταση για κάθε εικόνα που θα την περιγράφει μέσω άλλων εικόνων του συστήματος. Οι αναπαραστάσεις αυτές παίρνουν ως δεδομένο μια διανυσματική αναπαράσταση για κάθε εικόνα και όχι ένα σύνολο από περιγραφείς όπως η VLAD. Στην ενότητα αυτή



Σχήμα 4.1: Μέθοδος κατασκευής μιας OB αναπαράστασης.

χρησιμοποιούμε τις VLAD διανυσματικές αναπαραστάσεις που έχουμε παράξει για όλες τις εικόνες της βάσης για να κατασκευάσουμε μια καινούργια διανυσματική αναπαράσταση.

4.1.1 Τράπεζα Αντικειμένων

Την πρώτη αναπαράσταση που σχηματίσαμε την ονομάζουμε Τράπεζα Αντικειμένων (OB - Object Bank) [16], στην όποια θα αναφερόμαστε πλέον ως OB αναπαράσταση. Η μέθοδος αυτή χρησιμοποιείται στο [16] στο πρόβλημα της κατηγοριοποίησης εικόνων-σκηνών. Αφορούν σκηνές που περιέχουν πολλά διαφορετικά αντικείμενα και ο στόχος είναι η τελική αναπαράσταση να περιγράφει όσο το δυνατό καλύτερα τα αντικείμενα που ανιχνεύονται στη σκηνή.

Η OB αναπαράσταση κατασκευάζεται από την απόκριση πολλών διαφορετικών ανιχνευτών αντικειμένων που εφαρμόζονται στην εικόνα. Η έννοια αντικείμενο στην περίπτωση αυτή δεν αφορά μόνο αντικείμενα του φυσικού κόσμου αλλά και ευρύτερες έννοιες όπως θάλασσα, ουρανός κτλ, όπως φαίνεται και από το σχήμα 4.1. Συγκεκριμένα χρησιμοποιούν τον ανιχνευτή SVM [19] για αντικείμενα όπως τραπέζι, αυτοκίνητο, άνθρωπος κτλ, και ένα κατηγοριοποιητή υφής [5]

για αντικείμενα που βασίζονται στην υφή και στο υλικό, όπως ουρανός, δρόμος, άμμος κτλ. Η μέθοδος για τη δημιουργία της αναπαράστασης είναι ανεξάρτητη από τους ανιχνευτές αντικειμένων που χρησιμοποιούμε.

Όπως φαίνεται και στο σχήμα 4.1 η τελική αναπαράσταση θυμίζει το μοντέλο BOF, διότι αντιστοιχεί σε ένα ιστόγραμμα όπου κάθε παράμετρος αντιστοιχεί σε ένα συγκεκριμένο αντικείμενο με τιμή την απόκριση του αντικειμένου αυτού στην εικόνα. Προφανώς, η διάσταση της τελικής αναπαράστασης εξαρτάται από το πλήθος των κατηγοριών που μπορούν να ανιχνεύσουν οι αλγόριθμοι που χρησιμοποιούνται.

Η μέθοδος που προτείνουμε βασίζεται στην ιδέα αυτή που παρουσιάσαμε. Το πρόβλημα που επεξεργαζόμαστε στα πλαίσια της διπλωματικής αυτής είναι αυτό της ανάκτησης εικόνων από ένα σύνολο από εικόνες. Οι εικόνες αυτές αποτελούν κτήρια της Οξφόρδης με αποτέλεσμα να μην μπορούμε να τις αντιμετωπίσουμε σαν σκηνές από διαφορετικά αντικείμενα. Το πρώτο βήμα είναι να μεταφράσουμε την έννοια των αντικειμένων ή κλάσεων. Χρησιμοποιήσαμε πολλές διαφορετικές προσεγγίσεις για τον σχηματισμό της βάσης της OB, όπου με τον όρο ”βάση” εννοούμε το μέγεθος του τελικού OB διανύσματος, που αντιστοιχεί όπως είπαμε στο πλήθος των διαφορετικών αντικειμένων που αναγνωρίζονται στην εικόνα.

Έχοντας σχηματίσει την VLAD αναπαράσταση για όλες τις εικόνες του συνόλου, επιλέγουμε ένα υποσύνολο τους που θα αποτελεί τη ”βάση” της OB αναπαράστασης που προσπαθούμε να σχηματίσουμε. Συγκεκριμένα στα πειράματα που κάναμε είχαμε σχηματίσει VLAD από λεξικό $k = 64$, το οποίο σχηματίζει VLAD διανύσματα 8192 διαστάσεων. Επιλέξαμε N διαφορετικές εικόνες από το Oxford5k τις οποίες και χρησιμοποιήσαμε ως βάση. Έστω B_i με $i = 1, .., N$ οι VLAD αναπαραστάσεις των εικόνων που σχηματίζουν τη βάση, OB η τελική δια-

νυσματική αναπαράσταση μιας εικόνας και OB_i η i-οστή διάσταση του διανύσματος, με $i = 1, .., N$. Αν V_{img} το VLAD διάνυσμα της εικόνας, η OB υπολογίζεται όπως στην εξίσωση 4.1. Δηλαδή ως το εσωτερικό γινόμενο της εικόνας αναφοράς και των βάσεων.

$$OB_i = V_{img} \cdot B_i,\ i = 1, .., N \tag{4.1}$$

Τέλος, το OB διάνυσμα L_2 - κανονικοποιείται και η σύγκριση δύο εικόνων γίνεται με μέτρο το εσωτερικό γινόμενο. Σημαντικό σημείο που παραλείψαμε είναι ο τρόπος με τον οποίο επιλέγουμε ποιες εικόνες θα χρησιμοποιηθούν σα βάση. Χρησιμοποιήσαμε αρκετούς διαφορετικούς τρόπους για την επιλογή των εικόνων βάσης όπως:

- Τυχαία επιλογή N διαφορετικών εικόνων από το Oxford5k.
- Ομαδοποίηση των VLAD αναπαραστάσεων του Oxford100k ή του Oxford5k σε N διαφορετικές ομάδες, με χρήση του k-means.
- Ομαδοποίηση των VLAD αναπαραστάσεων του Oxford100k ή του Oxford5k σε N διαφορετικές ομάδες, με χρήση του k-means και επιλογή της κοντινότερης πραγματικής εικόνας για κάθε κέντρο k.

Για την αξιολόγηση του αλγορίθμου η τυχαία επιλογή έγινε πολλαπλές φορές και για διαφορετικές τιμές N από 100 έως 1000. Η OB αναπαράσταση όπως φάνηκε από τα αποτελέσματα των πειραμάτων λειτουργεί πολύ χειρότερα από την VLAD αναπαράσταση. Η OB αναπαράσταση για την ακρίβεια ακόμη και στην

περίπτωση των 1000 διαστάσεων είχε 10 πόντους mAP χειρότερα αποτελέσματα από την VLAD. Τα χειρότερα αποτελέσματα είχε η δεύτερη μέθοδος στην οποία οι N εικόνες της βάσης δεν αποτελούσαν πραγματικές εικόνες.

Θεωρούμε πως ένας από τους λόγους που δε λειτουργεί μια τέτοια μέθοδος στο δικό μας πρόβλημα είναι ότι η VLAD αναπαράσταση δεν έχει από μόνη της αρκετά καλή απόδοση στο Oxford5k. Αυτό παίζει τόσο σημαντικό ρόλο είναι πως το εσωτερικό γινόμενο στην εξίσωση 4.1 παίζει το ρόλο του ανιχνευτή, με αποτέλεσμα να μην έχουμε καλής ποιότητας ανίχνευση των ψευδοκλάσεων που έχουμε δημιουργήσει για να περιγράψουμε κάθε εικόνα. Επίσης ο χώρος των VLAD διανυσμάτων του συνόλου εικόνων είναι πολύ αραιός για να μπορεί να περιγραφεί σωστά με τόσο λίγες εικόνες και μέτρο το εσωτερικό γινόμενο.

4.1.2 Επαναταξινόμηση

Μετά την αποτυχία σχηματισμού μιας OB αναπαράστασης προσπαθήσαμε να μελετήσουμε αν είναι γενικότερα δυνατό να περιγραφεί σωστά μια εικόνα από τη γειτονιά της στο χώρο που σχηματίζουν οι VLAD αναπαραστάσεις. Η ιδέα είναι να χρησιμοποιήσουμε τη λίστα που επιστρέφει το σύστημα ανάκτησης που έχουμε σχηματίσει, η οποία ταξινομεί όλες τις εικόνες του συνόλου βάσει της ομοιότητας τους, ως τη $k - NN$ γειτονιά της εικόνας του ερωτήματος q. Με το τρόπο αυτό μπορούμε να υπολογίσουμε την ομοιότητα δυο εικόνων συγκρίνοντας τις γειτονιές τους. Η μέθοδος αυτή ονομάζεται Reranking (Επαναταξινόμηση) στην οποία για συντομία θα αναφερόμαστε ως RR [1].

Ο στόχος είναι να συγκρίνουμε τις δυο εικόνες συγκρίνοντας τις γειτονιές τους. Το επιτυγχάνουμε αυτό θεωρώντας πως όταν έχουμε σχηματίσει μια λίστα $N_k(q)$ που αποτελεί τη γειτονιά μεγέθους k μίας εικόνας q μπορούμε να εξάγουμε τη

σχέση μεταξύ των γειτονιών οποιοδήποτε δύο εικόνων που εμφανίζονται στη λίστα. Συγκεκριμένα έστω δύο εικόνες t και u, το σύνολο που σχηματίζει την κοινή γειτονιάς τους ορίζεται ως ο αριθμός των εικόνων της τομής των γειτονιών $k-NN$ τους. Όπως φαίνεται και στην εξίσωση 4.2. Ο πληθάριθμος $|SNN_k(t,u)|$ αντιπροσωπεύει το πλήθος των κοινών εικόνων.

$$SNN_k(t,u) = N_k(t) \cap N_k(u) \tag{4.2}$$

Στο σημείο αυτό πρέπει να σχηματίσουμε τη μετρική με την οποία θα συγκρίνουμε τις εικόνες. Όπως είναι λογικό αν δυο εικόνες έχουν πολλές κοινές εικόνες στη γειτονιά τους θα είναι και πιο πιθανό να είναι όμοιες. Η μετρική που χρησιμοποιήσαμε για να συγκρίνουμε τα σύνολα είναι του Jaccard και φαίνεται στην εξίσωση 4.3. Οι τιμές της μετρικής αυτής κυμαίνονται από 0 έως 1, με $j_k(x,y) = 1$ να σημαίνει ότι οι εικόνες έχουν ακριβώς τις ίδιες γειτονιές.

$$j_k(x,y) = \frac{|SNN_k(x,y)|}{|N_k(x) \cup N_k(y)|} \tag{4.3}$$

Τα πειράματα έγιναν χρησιμοποιώντας τα πλήρη διανύσματα VLAD που έχουν καλύτερα αποτελέσματα. Όπως και στην περίπτωση της τράπεζας αντικειμένων τα αποτελέσματα της μεθόδου RR βασίζονται σε μεγάλο βαθμό από την ποιότητα των αρχικών αναπαραστάσεων. Τα πειράματα έγιναν για διάφορες τιμές του k και το μέγεθος της λίστας ήταν 5061. Το Oxford5k δηλαδή χωρίς την εικόνα του ερωτήματος. Δυστυχώς και αυτή η μέθοδος δεν είχε τα θετικά αποτελέσματα

που περιμέναμε στο Oxford5k. Θεωρούμε επομένως πως δεν είναι δυνατό να χρησιμοποιήσουμε τον χώρο που σχηματίζεται από τις VLAD αναπαραστάσεις του συνόλου Oxford5k ως βάση. Το RR λειτούργησε βέλτιστα με χρήση του πλήρους VLAD διανύσματος και k από 23-26, όπου είχε 3 πόντους mAP χειρότερη απόδοση από την αναπαράσταση στην οποία βασίστηκε.

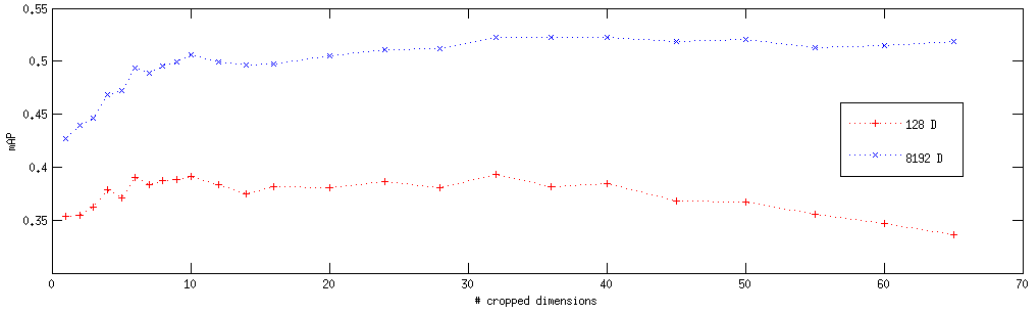
Στο σημείο αυτό αποφύγαμε να σχηματίσουμε κάποια άλλη διανυσματική αναπαράσταση που θα βασίζεται στη σύγκριση του συνόλου των εικόνων. Συνεχίσαμε με τη μελέτη της συμπεριφοράς των VLAD αναπαραστάσεων στο σύνολο του Oxford5k με στόχο να εντοπίσουμε ένα τρόπο για περαιτέρω βελτίωση της.

4.2 Μέθοδοι βελτίωσης της αναπαράστασης VLAD

Στην ενότητα αυτή μελετάμε τη συμπεριφορά των μειωμένων τελικών VLAD διανυσμάτων που έχουμε υλοποίησει στην ενότητα 2. Εξετάσαμε το πως επηρεάζεται η απόδοση της τελικής συμπαγούς αναπαράστασης από τις μεθόδους κανονικοποίησης που παρουσιάσαμε στο κεφάλαιο 2.2.

Οι μέθοδοι που θα μας απασχολήσουν είναι οι ”whitening” και ”power-law” καθώς και αυτή που προτείνουμε εμείς. Για την δημιουργία του πλήρους VLAD, πάνω στο οποίο θα γίνει η σύγκριση, χρησιμοποιούμε το $\text{VLAD}_{\text{CR+SSR}}$ που περιγράψαμε στο προηγούμενο κεφάλαιο. Στη συνέχεια θα αναφερόμαστε στην τελική VLAD διανυσματική αναπαράσταση ως VLAD_x όπου x η διάσταση στην οποία έχει μειωθεί η αναπαράσταση, π.χ. η VLAD_{8192} αφορά το πλήρες διάνυσμα χωρίς μείωση.

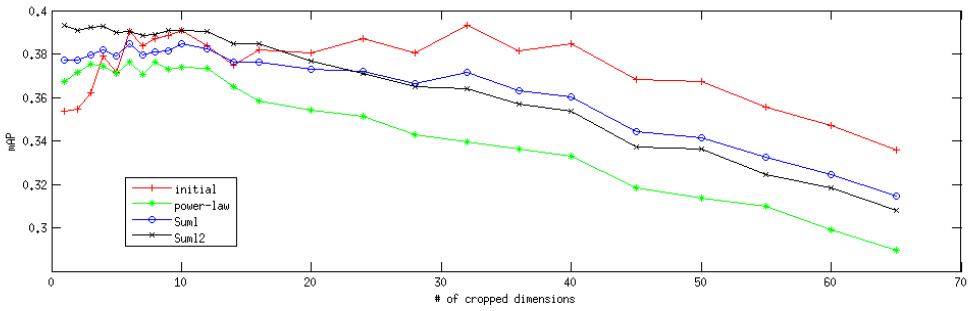
Όπως έχουμε αναφέρει για να παράγουμε το VLAD_{128} επιλέγουμε τα 128 διανύσματα που αντιστοιχούν στις 128 μεγαλύτερες ιδιοτιμές του πίνακα συνδιακύ-



Σχήμα 4.2: Σχέση μεταξύ των κυρίαρχων συνιστωσών που συγκρατούμε και της απόδοσης της τελικής αναπαράστασης για το $\text{VLAD}_{\text{CR+SSR}}$.

μανσης. Μελετώντας τη συμπεριφορά του VLAD_{128} παρατηρήσαμε πως αφαιρώντας τα μεγαλύτερα ιδιοδιανύσματα βελτιώνεται η απόδοση της αναπαράστασης. Στο σχήμα 4.2 φαίνεται ότι το αν το VLAD_{128} ξεκινάει από την 6η μεγαλύτερη ιδιοτιμή βελτιώνει την απόδοση της αναπαράστασης κατά . Το ίδιο χαρακτηριστικό παρατηρείται και για το VLAD_{8192} το οποίο σημαίνει ότι η μέθοδος μας βελτιώνει το αποτέλεσμα ακόμη και στην περίπτωση που θέλουμε να σχηματίσουμε τη βέλτιστη αναπαράσταση χωρίς να μας ενδιαφέρει το μέγεθος της αναπαράστασης, σε αντίθεση με τη μέθοδο ”whitening”.

Η συμπεριφορά αυτή είναι αξιοσημείωτη καθώς αφαιρώντας τις κυρίαρχες συνιστώσες του χώρου των VLAD_{8192}, αφαιρούμε τις συνιστώσες με τη μεγαλύτερη διασπορά τιμών, αρά και τη μεγαλύτερη, θεωρητικά, διακριτική ικανότητα. Στην προσπάθεια να εξηγήσουμε το φαινόμενο αυτό, υποθέσαμε πως ο βασικός λόγος για τον οποίο συμβαίνει αυτό είναι ότι σε ένα τόσο αραιό και μεγάλο χώρο οι πρώτες συνιστώσες έχουν τόσο μεγάλη ιδιοτιμή που ”θολώνουν” την αναπαράσταση. Με τον όρο αυτό εννοούμε πως επηρεάζουν σε τέτοιο βαθμό το εσωτερικό γινόμενο μεταξύ των VLAD_{128} διαφορετικών εικόνων που δεν επιτρέπει στις συ-

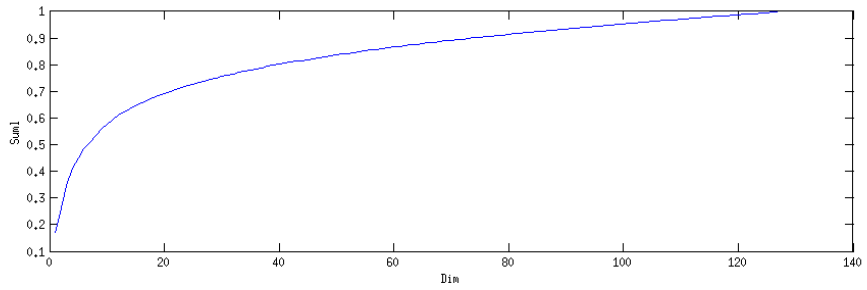


Σχήμα 4.3: Σχέση μεταξύ των κυρίαρχων συνιστωσών που συγκρατούμε και της απόδοσης της τελικής αναπαράστασης για διαφορετικές μεθόδους.

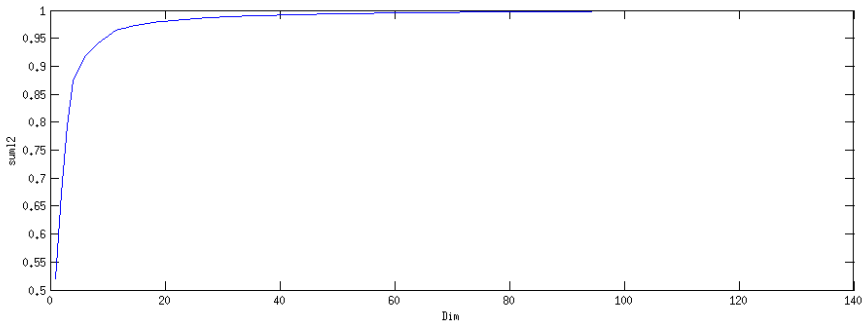
νιστώσες που όντως συγκεντρώνουν την διακριτική ικανότητα των εικόνων να επηρεάσουν το αποτέλεσμα. Αυτό γίνεται εμφανές στο σχήμα 4.3.

Στον πίνακα 4.3 φαίνεται το πόσο βελτιώνει το αποτέλεσμα η εφαρμογή ”power-law” κανονικοποίησης στο VLAD_{8192} καθώς επίσης και το ”whitening”. Παρά τη συνολική βελτίωση του αποτελέσματος παρατηρούμε ότι και πάλι αφαιρώντας τις πρώτες συνιστώσες η αναπαράσταση έχει καλύτερο αποτελέσματα. Ο λόγος που θεωρούμε ότι η εφαρμογή του power-law δεν εξαλείφει το φαινόμενο αυτό είναι πως ενώ μειώνει τη συνεισφορά των πρώτων συνιστωσών μειώνει ταυτόχρονα και τη συνεισφορά όλων των συνιστωσών που έχουν μεγάλη ενέργεια στο διάνυσμα. Επομένως, για να εξαλείψουμε το φαινόμενο αυτό θεωρήσαμε πως μια μέθοδος που δρα ανεξάρτητα σε κάθε αναπαράσταση όπως η power-law κανονικοποίηση δεν μπορεί να βοηθήσει.

Προτείνουμε μια μέθοδο κανονικοποίησης που χρησιμοποιεί τις ιδιοτιμές που έχουμε υπολογίσει από τον πίνακα συνδιακύμανσης που θεωρούμε ότι μειώνει τη συνεισφορά των πρώτων κυρίαρχων συνιστωσών χωρίς να επηρεάζει τις υπόλοιπες συνιστώσες. Έστω x_i οι συνιστώσες του διανύσματος VLAD_{128}, y_i οι συνι-



Σχήμα 4.4: Γραφική αναπαράσταση της μεθόδου suml που χρησιμοποιούμε για την κανονικοποίηση ενός διανύσματος VLAD_{128}



Σχήμα 4.5: Γραφική αναπαράσταση της μεθόδου suml2 που χρησιμοποιούμε για την κανονικοποίηση ενός διανύσματος VLAD_{128}

στώσες του $\text{VLAD}_{128+\text{norm}}$ που προκύπτει μετά την κανονικοποίηση που προτείνουμε και λ_i οι ιδιοτιμές που αντιστοιχούν στις συνιστώσες αυτές, όπου προφανώς για VLAD_{128} $i = 1, .., 128$. Προτείνουμε να πολλαπλασιάζεται κάθε συνιστώσα x_i με το άθροισμα των ιδιοτιμών 1 έως i, όπως φαίνεται στην εξίσωση 4.4 για $D = 128$. Τέλος, όπως γίνεται σε κάθε κανονικοποίηση, το τελικό διάνυσμα κανονικοποιείται ως προς την L_2 νόρμα.

$$y_i = \sum_{j=1}^{i} \lambda_j \cdot x_i,\ i = 1, .., D \tag{4.4}$$

Παρατηρούμε από το σχήμα 4.3 ότι η μέθοδος που προτείνουμε βελτιώνει σε μεγάλο βαθμό τα αποτελέσματα. Την μέθοδο αυτή την ονομάζουμε Suml, διότι κάθε συνιστώσα πολλαπλασιάζεται με το άθροισμα των ιδιοτιμών που αντιστοιχών στις συνιστώσες από την αρχή του διανύσματος. Όμως παρότι βελτιώνει το αποτέλεσμα δεν απαλείφει εντελώς το φαινόμενο που παρατηρήσαμε και παρουσιάσαμε. Έχοντας πλέον αποδείξει ότι αυτή η κανονικοποίηση βοήθησε στο πρόβλημα που αντιμετωπίζαμε προσπαθήσαμε να την διαμορφώσουμε ώστε να μειώνει ακόμη περισσότερο την συνεισφορά των κυρίαρχων συνιστωσών. Το πετύχαμε αυτό υψώνοντας τις ιδιοτιμές στο τετράγωνο πριν τις αθροίσουμε όπως φαίνεται στην εξίσωση 4.5.

$$y_i = \sum_{j=1}^{i} {\lambda_j}^2 \cdot x_i,\ i = 1, .., D \tag{4.5}$$

Τα αποτελέσματα της δεύτερης μεθόδου που ονομάσαμε Suml2 φαίνονται στο σχήμα 4.3 και πράγματι φαίνεται ότι με την κανονικοποίηση αυτή όσο αφαιρούμε κυρίαρχες συνιστώσες από το διάνυσμα μειώνεται η απόδοση της αναπαράστασης όπως ήταν επιθυμητό. Σημαντική παρατήρηση είναι πως η μέθοδός μας αυτή μειώνει την επίδραση των κυρίαρχων συνιστωσών σε αντίθεση με τη μέθοδο ”whitening” η οποία όπως αναφέραμε, ενισχύει την επίδραση των πρώτων συνιστωσών. Δεύτερη σημαντική διαφορά της μεθόδου Suml2 που αναπτύξαμε

είναι πως λειτουργεί θετικά ακόμη και όταν χρησιμοποιούμε το πλήρες διάνυσμα VLAD χωρίς να μειώσουμε τις διαστάσεις του. Προφανώς βέβαια μετά την περιστροφή του διανύσματος με χρήση PCA.

Για να γίνει πιο αντιληπτός ο τρόπος με τον οποίο επηρεάζουμε το τελικό διάνυσμα παρουσιάζουμε τους συντελεστές $\sum_1^i {\lambda_i}^2$ και $\sum_1^i \lambda_i$ ως συναρτήσεις στα σχήματα 4.4 και 4.5 αντίστοιχα. Μελετώντας κανείς την καμπύλη που σχηματίζουν οι συντελεστές κανονικοποίησης και ειδικά τη δεύτερη παρατηρούμε πως οι πρώτες 8-10 συνιστώσες που αν αφαιρεθούν εντελώς, σχήμα 4.3, βελτιώνουν την απόδοση επηρεάζονται σε μεγάλο βαθμό από την κανονικοποίηση το οποίο και επιζητούσαμε.

Συμπεράσματα

Στη διπλωματική αυτή εξετάσαμε πολλές διαφορετικές αναπαραστάσεις που έχουν αναπτυχθεί με βάση την VLAD και τις αξιολογήσαμε πάνω στο Oxford5k. Μελετήσαμε τη συμπεριφορά γνωστών αλλά και νέων αναπαραστάσεων πάνω στο συγκεκριμένο σύνολο εικόνων. Για το λόγο αυτό είναι αδύνατο να εξάγουμε γενικευμένα συμπεράσματα για τη λειτουργία των μεθόδων.

Παρατηρήσαμε πως για να έχει καλά αποτελέσματα η VLAD, όπως και οποιαδήποτε άλλη αναπαράσταση, στο σύνολο αυτό πρέπει να αντιμετωπίζει το πρόβλημα του ”burstiness” που παρουσιάσαμε καθώς σε εικόνες κτηρίων επηρεάζει σε μεγάλο βαθμό την αναπαράσταση. Είναι αναγκαίο, για το λόγο αυτό, να επεξεργαστεί κανείς τους SIFT περιγραφείς πριν τη δημιουργία του πλήρους VLAD. Συγκεκριμένα παρατηρήσαμε πως αν το οπτικό λεξικό που έχουμε σχηματίσει είναι καλής ποιότητας, όπως στη δική μας περίπτωση, δεν χρειάζεται η επεξεργασία των SIFT να γίνει ανεξάρτητα σε κάθε οπτική λέξη και πως είναι αρκετό να γίνει μια συνολική περιστροφή στον χώρο των SIFT και στη συνέχεια να κανονικοποιηθεί κάθε συνιστώσα με ύψωση σε δύναμη. Η διαδικασία αυτή ήταν αρκετή για να περιορίσουμε την επίδραση του ”burstiness”.

Σημαντικό αποτέλεσμα επίσης ήταν ότι οι μέθοδοι OB και RR που βασίστηκαν πάνω στη VLAD αναπαράσταση δεν είχαν τα αποτελέσματα που περιμέναμε.

Συγκεκριμένα για το OB οι λόγοι που θεωρούμε ότι επηρεάζουν την επίδοση των τεχνικών αυτών προκύπτουν από τη μορφή του χώρου που σχηματίζουν τα VLAD διανύσματα του Oxford5k και από την μέτρια ποιότητα της VLAD αναπαράστασης στο Oxford5k. Όσο για το RR θεωρούμε πως η κακή απόδοση του οφείλεται κατά κύριο λόγο στην κακή διακριτική ικανότητα της VLAD.

Το πιο ενδιαφέρον και σημαντικό συμπέρασμα είναι ότι στο σχηματισμό της μειωμένης VLAD αναπαράστασης οι κυρίαρχες συνιστώσες που προκύπτουν επιβαρύνουν την απόδοση της αναπαράστασης. Παρατηρήσαμε πως αν κανονικοποιήσουμε το τελικό μειωμένο διάνυσμα ώστε να μετριάσουμε την ένταση του πρώτων συνιστωσών η αναπαράσταση που προκύπτει έχει καλύτερα αποτελέσματα τόσο στο πλήρες όσο και στο μειωμένο διάνυσμα. Προτείναμε δύο μεθόδους κανονικοποίησης που βασίζονται έμμεσα στην τοπολογία του συνόλου εκπαίδευσης και λειτουργούν καλύτερα από αυτές που έχουν χρησιμοποιηθεί ως τώρα στην αντίστοιχη υπάρχουσα βιβλιογραφία.

Τέλος, οφείλουμε να σημειώσουμε πως τα τελευταία δυο χρόνια έχει αντικατασταθεί πλήρως ο τρόπος που αντιμετωπίζουμε τα προβλήματα ανάλυσης εικόνων, όπως αυτό της ανάκτησης που μελετήσαμε εμείς αλλά και αυτό της κατηγοριοποίησης, από τη χρήση CNN (Convolutional Neural Networks). Οι τεχνικές μηχανικής μάθησης που χρησιμοποιήσαμε παραμένουν χρήσιμες στο σχεδιασμό τέτοιων συστημάτων, όπως φάνηκε και από το [18].

Βιβλιογραφία

[1] L. Amsaleg A. Delvinioti, H. Jégou and M. E. Houle. Image retrieval with reciprocal and shared nearest neighbors. 2012.

[2] R. Arandjelović and A. Zisserman. All about VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[3] O. Chum and J. Matas. Unsupervised discovery of co-occurence in sparse high dimensional data. 2010.

[4] P. Comon. Independent component analysis, a new concept? 1994.

[5] A. A. Efors D. Hoiem and M. Hebert. Automatic photo pop-up. 2005.

[6] Jonathan Delhumeau, Philippe-Henri Gosselin, Hervé Jégou, and Patrick Pérez. Revisiting the vlad image representation. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 653–656, New York, NY, USA, 2013. ACM.

[7] John Eakins and Margaret Graham. Content-based image retrieval. 1999.

[8] J. Sanchez F. Peronnin, Y. Liu and H. Poirer. Large-scale image retrieval with compressed fisher vectors. 2010.

[9] F.Perrnonin and C.R.Dance. Fisher kernels on visual vocabularies for image categorization. 2007.

[10] M. Douze H. Jégou and Schmid C. On the burstiness of visual elements. 2009.

[11] M. Douze H. Jégou and C. Schmid. Improving bag-of-features for large scale image search. 2010.

[12] Hervé Jégou and Ondrej Chum. Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *ECCV - European Conference on Computer Vision*, Firenze, Italy, October 2012.

[13] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3304–3311, jun 2010.

[14] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, and Patrick Pérez. Aggregating local descriptors into compact codes. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1704–1716, 2012.

[15] Hervé Jégou and Andrew Zisserman. Triangulation embedding and democratic aggregation for image search. In *CVPR - International Conference on Computer Vision and Pattern Recognition*, Columbus, United States, June 2014.

[16] Eric P. Xing Li-Jia Li, Hao Su and Li Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. 2010.

[17] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, 1999.

[18] Yao Lu. Unsupervised learning on neural network outputs. 2015.

[19] D. McAllester P. Felzenszwalb, R. Girshick and D. Ramanan. Object detection with discriminatively trained part based models. 2007.

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

[21] T.Jaakkola and D. Haussle. Exploiting generative models in discriminative classifiers. 1998.