ΤΙΤΛΟΣ ΔΙΑΤΡΙΒΗΣ

ΠΕΡΙΛΗΨΗ ΑΜΟΝΤΑΡΙΣΤΟΥ ΒΙΝΤΕΟ ΜΕ ΤΗ ΧΡΗΣΗ ΣΗΜΑΣΙΟΛΟΓΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Η

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης

του Τμήματος Μηχανικών Η/Υ και Πληροφορικής

Εξεταστική Επιτροπή

από την

Αθηνά Παππά

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ

ΣΤΟ ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Νοέμβριος 2013

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα Καθηγητή κ. Λύκα Αριστείδη και το Διδάκτορα κ. Χασάνη Βασίλειο για την πολύτιμη βοήθειά τους κατά τη διάρκεια της εργασίας αυτής. Επίσης, θα ήθελα να ευχαριστήσω τα αγαπημένα μου πρόσωπα για την υποστήριξή τους σε όλη τη διάρκεια των σπουδών μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

ΠΕΡΙΛΗΨΗ

Αθηνά Παππά του Δημητρίου και της Κωνσταντίας. MSc, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Νοέμβριο 2013. Περίληψη αμοντάριστου βίντεο με τη χρήση σημασιολογικών χαρακτηριστικών. Επιβλέπων: Λύκας Αριστείδης.

Στις μέρες μας ένας τεράστιος αριθμός βίντεο υπάρχει στη διάθεση του χρήστη. Το γεγονός αυτό, οδήγησε τους ερευνητές στην ανάπτυξη τεχνικών για όσο γίνεται πιο αξιόπιστη περίληψη, αναζήτηση, επεξεργασία και ανάκτηση βίντεο. Το πρόβλημα της περίληψης ενός βίντεο, δηλαδή μίας σύντομης και περιεκτικής αναπαράστασης του, είναι ένα από τα σημαντικότερα ζητήματα ανάλυσης και επεξεργασίας ψηφιακού βίντεο. Μια προσέγγιση για την περίληψη αμοντάριστου βίντεο, που αποτελεί και το αντικείμενο μελέτης της παρούσας εργασίας, είναι η ομαδοποίηση των όμοιων πλάνων που περιέχει, ώστε η αναπαράσταση κάθε ομάδας να γίνεται από ένα μόνο πλάνο. Έτσι, ένα ολόκληρο βίντεο μπορεί να αναπαρασταθεί από λίγα αλλά μοναδικά ως προς το οπτικό περιεχόμενο πλάνα, διατηρώντας ένα μεγάλο ποσοστό πληροφορίας. Το γεγονός αυτό βοηθάει στην γρήγορη κατανόηση του περιεχομένου, χωρίς να είναι απαραίτητη η παρακολούθηση ολόκληρου του βίντεο. Επιπλέον, η ομαδοποίηση των πλάνων κάνει ευκολότερη την οργάνωση και την ανάκτηση ακολουθιών βίντεο.

Στην εργασία αυτή, αρχικά αναφέρονται εργασίες που μελετούν την περίληψη, ομαδοποίηση και ανάκτηση περιεχομένου βίντεο, καθώς και τον τρόπο εντοπισμού σημασιολογικών εννοιών και επιπλέον ορίζεται η έννοια της περίληψης αμοντάριστου βίντεο. Στη συνέχεια, παρουσιάζεται ο τρόπος που εξάγονται τα σημασιολογικά χαρακτηριστικά με τα οποία γίνεται η ομαδοποίηση όμοιων πλάνων.

Επιπρόσθετα, αναλύεται η μέθοδος ομαδοποίησης όμοιων πλάνων, καθώς και η βελτίωση της ομαδοποίησης με χρήση μεθόδου εντοπισμού προσώπου και σώματος. Τέλος, παρουσιάζονται τα πειραματικά αποτελέσματα της μεθόδου και αποτιμάται η επίδοσή της σε ακολουθίες βίντεο από το διαγωνισμό TRECVID 2008.

EXTENDED ABSTRACT IN ENGLISH

Athina Pappa, MSc, Computer Science Department, University of Ioannina, Greece.
November, 2013. Rushes video summarization using semantic concepts. Thesis Supervisor: Likas Aristidis

Nowadays, a huge amount of videos is available to everyone. This motivated the researchers in developing techniques for reliable video summarization, search and retrieval. The subject of video summarization, which is a brief and comprehensive representation of the video content, is one of the most important topics in analysis and processing of digital videos. An approach used for video rushes summarization, which is the aim of this thesis, is the clustering of similar shots and representing each group with only one shot. As result, it becomes possible to represent the whole rushes video content using only a few, cautiously picked, shots maintaining at the same time a great percentage of information. This fact is of great help in obtaining a rapid assessment of the video content without needing to watch the whole video. In addition, the video summary is considerably useful in performing indexing and retrieval of similar videos.

In this thesis, previous research related to video summarization, clustering and retrieval and also ways to find and create semantic concepts is presented. The definition of video rushes summarization is given, too. Further on, the way that the semantic features are extracted is described and the way that similar shots are clustered. Moreover, the improvement of clustering using face and body detection are analyzed. Finally, experimental results are represented, along with evaluation of the performance of the proposed method on several video sequences.

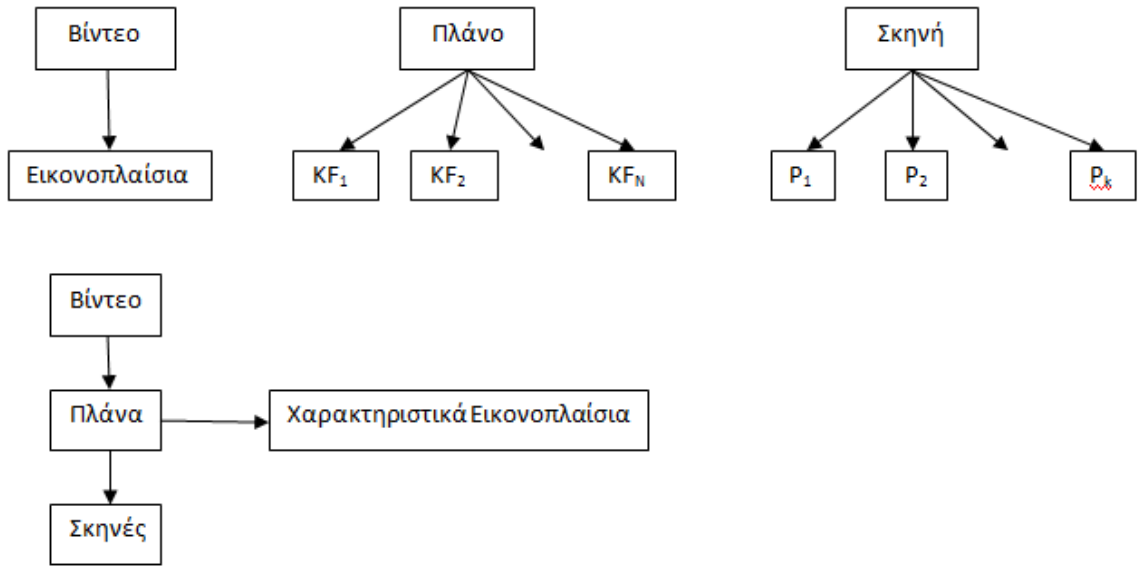
ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1. Περιγραφή του Προβλήματος

Λόγω της ραγδαίας ανάπτυξης τα τελευταία χρόνια σε διάφορους τομείς της τεχνολογίας όπως για παράδειγμα την υπολογιστική ισχύ, τα ταχύτερα δίκτυα, τη μεγαλύτερη και οικονομικότερη χωρητικότητα των αποθηκευτικών μέσων, έχει παρατηρηθεί τεράστια ανάπτυξη στον τομέα του ψηφιακού βίντεο. Τα παραπάνω, σε συνδυασμό με την δημιουργία πολυμεσικών συσκευών μικρού μεγέθους οδήγησε στην κατασκευή φορητών συσκευών τεχνολογίας όπως *notepads* και *smartphones,* δίνοντας έτσι τη δυνατότητα σε πολλούς χρήστες να έχουν ανά πάσα στιγμή έναν ηλεκτρονικό υπολογιστή μαζί τους. Ένα από τα χαρακτηριστικά των συσκευών αυτών είναι η εγγραφή και η αναπαραγωγή βίντεο. Το γεγονός αυτό σε συνδυασμό με τις ήδη υπάρχουσες ψηφιακές φωτογραφικές μηχανές και βιντεοκάμερες οδήγησε στη δημιουργία τεράστιου όγκου βίντεο και παρουσιάστηκε η ανάγκη για ανάπτυξη εφαρμογών για την καλύτερη διαχείριση αυτού του όγκου πληροφορίας από παραγόμενα βίντεο όπως είναι η αρχειοθέτηση, η δημιουργία ευρετηρίου και η ανάκτησή του. Για να γίνει πιο εύκολα η επεξεργασία του τεραστίου όγκου από αμοντάριστα βίντεο θα πρέπει να αναπτυχθεί ένας μηχανισμός ο οποίος θα επιτρέπει στον χρήστη να έχει άποψη για το περιεχόμενο ενός αμοντάριστου βίντεο χωρίς να

είναι απαραίτητο να παρακολουθήσει ολόκληρο το βίντεο. Αυτό μπορεί να επιτευχθεί με την περίληψη αμοντάριστου βίντεο.

Προκειμένου να είμαστε σε θέση να εξάγουμε την περίληψη ενός αμοντάριστου βίντεο, πρέπει πρώτα να αναλύσουμε τη δομή του. Τα δεδομένα ενός βίντεο, οργανώνονται σε μία ιεραρχική δομή. Η δομή αυτή προκύπτει χωρίζοντας το βίντεο σε τρία επίπεδα που περιέχουν σημαντικές πληροφορίες. Αυτά είναι τα εικονοπλαίσια (*frames*) τα οποία αποτελούν τις πρωταρχικές συνιστώσες ενός βίντεο και αναπαριστούν οπτικά περιεχόμενα. Στη συνέχεια, είναι το πλάνο (*shot*) το οποίο ορίζεται ως μία συνεχής ακολουθία από εικονοπλαίσια που έχουν καταγραφεί από μία μόνο κάμερα. Το τρίτο επίπεδο είναι η σκηνή (*scene*) που αποτελείται από διαδοχικά πλάνα τα οποία περιγράφουν μία ενέργεια ή ένα γεγονός. Κάθε πλάνο μπορεί να αναπαρασταθεί από αντιπροσωπευτικά εικονοπλαίσια, που ονομάζονται χαρακτηριστικά εικονοπλαίσια (*keyframes*) που περιγράφουν ικανοποιητικά το περιεχόμενο του πλάνου. Στην Εικόνα 1.1 φαίνεται η δομή του βίντεο όπου KF_i τα χαρακτηριστικά εικονοπλαίσια ενός πλάνου, P_i τα πλάνα μιας σκηνής, N το πλήθος των εικονοπλαισίων του πλάνου και k το πλήθος των πλάνων μιας σκηνής.



Εικόνα 3.1 Δομή του βίντεο.

Η περίληψη ενός βίντεο είναι μια συμπυκνωμένη εκδοχή του αρχικού βίντεο με την οποία μπορεί να γίνει ταχύτερα και με λιγότερη προσπάθεια η κατανόηση του περιεχομένου του. Επίσης γίνεται πιο εύκολη η αναζήτηση σε μια βάση δεδομένων από βίντεο. Επιπρόσθετα, η περίληψη αμοντάριστου βίντεο (*video rushes summarization*) είναι χρήσιμη σε διάφορες εφαρμογές επεξεργασίας βίντεο, όπως το μοντάζ, αλλά μπορεί να χρησιμοποιηθεί και ανεξάρτητα βελτιώνοντας τον απαιτούμενο χώρο αποθήκευσης, το εύρος ζώνης αλλά και το χρόνο παρακολούθησης.

Το αμοντάριστο βίντεο έχει κάποια ιδιαίτερα χαρακτηριστικά σε σχέση με το μονταρισμένο. Τα πλάνα ενός αμοντάριστου βίντεο περιέχουν επαναλαμβανόμενη πληροφορία, καθώς η ίδια σκηνή γυρίζεται πολλές φορές μέχρι να παραχθεί το επιθυμητό αποτέλεσμα. Επιπλέον, καθώς πρόκειται για αμοντάριστο βίντεο, περιέχει πολλά ανεπιθύμητα εικονοπλαίσια όπως δεσμίδες χρωμάτων (*colorbars*), μονόχρωμα εικονοπλαίσια και κλακέτες (*clapboards*). Οι δεσμίδες χρωμάτων (*colorbars*) είναι ένα τεχνητό ηλεκτρονικό σήμα που παράγεται από μια κάμερα ή από εξοπλισμό παραγωγής (*post production equipment*). Μια κλακέτα (*clapboard*) είναι μια συσκευή που χρησιμοποιείται για να βοηθήσει στο συγχρονισμό των εικόνων και του ήχου. Επιπλέον, η κλακέτα (*clapboard*) χρησιμοποιείται για να ορίσει και να επισημάνει την έναρξη μιας συγκεκριμένης σκηνής ώστε να γίνει η καταγραφή της κατά τη διάρκεια της παραγωγής. Σε ένα μονταρισμένο βίντεο, αντιθέτως, έχουν αφαιρεθεί όλα τα περιττά εικονοπλαίσια που αναφέραμε προηγουμένως όπως εικόνες με κλακέτες ή μονόχρωμες εικόνες και επιπλέον δεν υπάρχουν πολλαπλές λήψεις της ίδιας σκηνής παρά μόνο μοναδικά ως προς το περιεχόμενο πλάνα.

Επομένως, είναι φανερό πως η περίληψη αμοντάριστου βίντεο είναι πολύ σημαντική στον τομέα της επεξεργασίας βίντεο. Ιδιαίτερα τα τελευταία χρόνια έχει αποτελέσει σημαντικό αντικείμενο έρευνας η αυτόματη εξαγωγή αποτελεσματικής περίληψης. Αν και υπάρχουν τεχνικές με ικανοποιητικά αποτελέσματα είναι ακόμα αντικείμενο μελέτης. Βασικό πρόβλημα είναι ότι υπάρχει δυσκολία στην ομαδοποίηση πλάνων, δηλαδή στον εντοπισμό του σωστού σημείου εναλλαγής μιας ομάδας όμοιων πλάνων. Για παράδειγμα, η χρήση ενός οπτικού χαρακτηριστικού,

όπως το ιστόγραμμα χρώματος, δεν αναμένεται να έχει ικανοποιητικά αποτελέσματα όταν εφαρμόζεται σε μία συλλογή διαφορετικών ακολουθιών βίντεο. Για αυτό το λόγο στην εργασία αυτή γίνεται χρήση σημασιολογικών χαρακτηριστικών φιλοδοξώντας έτσι να αντιμετωπίσουμε το παραπάνω σημαντικό πρόβλημα.

1.2. Αντικείμενο της Εργασίας

Στην εργασία αυτή ασχοληθήκαμε με την ομαδοποίηση όμοιων πλάνων που προέρχονται από ένα μεγάλο εύρος ακολουθιών αμοντάριστου βίντεο και σκοπός μας είναι η ομαδοποίηση των πλάνων και εν συνεχεία η περίληψη του βίντεο. Για να γίνει πιο αποτελεσματική και ταχύτερη η περίληψη, έγινε εξαγωγή σημασιολογικών χαρακτηριστικών για κάθε πλάνο. Έπειτα, ακολούθησε η ομαδοποίηση όμοιων πλάνων συγκρίνοντας τα σημασιολογικά χαρακτηριστικά διαδοχικών πλάνων. Στη συνέχεια, η απόδοση της παραπάνω μεθοδολογίας βελτιώθηκε με τη χρήση ενός αλγορίθμου εντοπισμού προσώπου και σώματος, για να αποφευχθεί τυχόν εσφαλμένος εντοπισμός σημείων εναλλαγής ομάδας όμοιων πλάνων. Επιπρόσθετα, για να αξιολογηθεί η απόδοση της μεθόδου μας έγινε σύγκριση με την ομαδοποίηση όμοιων πλάνων, η οποία προκύπτει από τη χρήση τριών διαφορετικών χαρακτηριστικών. Τα χαρακτηριστικά αυτά είναι το ιστόγραμμα χρώματος, *SIFT* χαρακτηριστικά [14], καθώς επίσης και τα χαρακτηριστικά που προκύπτουν από την ανίχνευση προσώπου και σώματος.

1.3. Δομή της Εργασίας

Αρχικά, στο $2^{ο}$ κεφάλαιο περιγράφεται η εξαγωγή σημασιολογικών χαρακτηριστικών.

Στο $3^{ο}$ κεφάλαιο αναλύεται η μέθοδος που χρησιμοποιήθηκε για την ομαδοποίηση όμοιων πλάνων. Παράλληλα, περιγράφονται οι διαφορετικές αναπαραστάσεις των πλάνων που εξετάστηκαν, η σύγκριση αυτών και η αξιολόγησή τους.

Στο 4ο κεφάλαιο παρουσιάζεται ο αλγόριθμος των *Viola-Jones* στον οποίο βασίζεται η μέθοδος εντοπισμού προσώπου και σώματος. Επιπλέον, εξηγείται ο τρόπος βελτίωσης των αποτελεσμάτων ομαδοποίησης πλάνων με χρήση του εντοπισμού προσώπου και σώματος.

Στο 5ο κεφάλαιο περιγράφονται οι ακολουθίες βίντεο που χρησιμοποιήθηκαν στα πειράματα και παρουσιάζονται οι πίνακες των αποτελεσμάτων από τις σειρές πειραμάτων που εκτελέστηκαν.

Στο 6ο κεφάλαιο και τελευταίο κεφάλαιο αναφέρονται συνοπτικά τα συμπεράσματα που προέκυψαν από κάθε σειρά πειραμάτων και προτείνονται κάποιες ιδέες για μελλοντική έρευνα.

1.4. Συναφείς Εργασίες

Οι Cees etal. [1] μελετούν το γενικό πρόβλημα της ανάκτησης περιεχομένου βίντεο προκειμένου να αποκτηθεί άποψη για τα ενδιάμεσα βήματα που επηρεάζουν την απόδοση των μεθόδων ανάλυσης πολυμεσικού περιεχομένου. Το αποτέλεσμα είναι να παρέχουν στους ερευνητές ένα λεξικό που περιέχει 101 σημασιολογικές έννοιες, προϋπολογισμένα χαμηλού επιπέδου χαρακτηριστικά, εκπαιδευμένα μοντέλα ταξινομητή SVM και πειραματικά αποτελέσματα πάνω σε βίντεο 85 ωρών. Οι Zhu etal. [30] μελέτησαν την επίπτωση της εκμετάλλευσης των εικόνων με ετικέτα για την μάθηση της έννοιας και ερεύνησαν το ζήτημα του πώς η ποιότητα των εικόνων αυτών επηρεάζει την απόδοση του εντοπισμού εννοιών. Επίσης, πρότειναν και μια προσέγγιση για πρόβλεψη της σχέσης μεταξύ της έννοιας-στόχου και της λίστας με ετικέτες που σχετίζεται με την εικόνα. Η προτεινόμενη μέθοδος να παρουσιάζει καλύτερη απόδοση στην εκμάθηση εννοιών από άλλες προσεγγίσεις όπως αυτές που βασίζονται σε δειγματοληψία λέξεων κλειδιά και σε ψηφοφορία ετικετών. Οι Jiang etal. [9] υλοποίησαν μια μελέτη πάνω στις επιλογές αναπαράστασης *BoW* (*Bag of words*), που περιλαμβάνει μελέτη χαρακτηριστικών όπως το μέγεθος του λεξιλογίου, το σύστημα βαρύτητας, τη μη μετακίνηση λέξεων, την επιλογή χαρακτηριστικών, τη χωρική πληροφορία και το οπτικό δίγραμμα. Το αποτέλεσμα ήταν ότι το *soft-weighting* (δηλαδή για μια εικόνα να χρησιμοποιήσουμε τις *N* κοντινότερες οπτικές

λέξεις και όχι την μια κοντινότερη) υπερτερεί των άλλων συστημάτων βαρύτητας και τα χαρακτηριστικά *BoW* από μόνα τους με τις κατάλληλες επιλογές αναπαράστασης επιτυγχάνουν υψηλής ανταγωνιστικότητας απόδοση στην ανίχνευση εννοιών. Στο [4] πρότεινεται μια προσέγγιση που υπολογίζει τον αριθμό των *key-frames* χρησιμοποιώντας στοιχεία της φασματικής θεωρίας γραφημάτων. Οι Chasanis etal. [5] πρότειναν μια μέθοδο όπου τοπικοί αμετάβλητοι περιγραφείς χρησιμοποιούνται για να αναπαραστήσουν τα εικονοπλαίσια (key-frames) και επιπλέον δημιούργησαν ένα οπτικό λεξιλόγιο από τους περιγραφείς το οποίο είχε σαν αποτέλεσμα μια αναπαράσταση οπτικών λέξεων σε ιστόγραμμα (*bag of visual words*) για κάθε λήψη. Το κλειδί σε αυτή τη μέθοδο είναι ότι το ιστόγραμμα των οπτικών λέξεων που αντιστοιχεί σε κάθε εικόνα εξομαλύνεται προσωρινά λαμβάνοντας υπόψη τα ιστογράμματα από γειτονικές εικόνες. Η μέθοδος παρέχει υψηλά ποσοστά εντοπισμού και διατηρεί καλή εναλλαγή ανάμεσα σε ανάκληση και ακρίβεια. Οι Over etal. [16] στην εργασία τους περιγράφουν ένα σύστημα αυτόματης εξαγωγής της περίληψης βίντεο από πολλές δραματικές σειρές του BBC. Οι ερευνητές έκαναν περιλήψεις βίντεο από 40 αρχεία βίντεο με στόχο να απομακρύνουν περιττό και ασήμαντο υλικό. Το αποτέλεσμα ήταν η περίληψη του βίντεο να γίνεται ταχύτερα και να αποφεύγεται άχρηστο υλικό. Οι Smoliar και Zhang [21] στην εργασία τους προτείνουν έναν πιο προχωρημένο τρόπο ανάκτησης περιεχομένου του βίντεο. Η εργασία τους περιλαμβάνει μια αποτελεσματική ανάλυση περιεχομένου βίντεο, καθώς και σύστημα αναπαράστασης για χαρακτηρισμό εννοιών υψηλού επιπέδου και μια τεχνική ιεραρχικής ταξινόμησης βίντεο για την κάλυψη του σημασιολογικού κενού ανάμεσα σε χαμηλού επιπέδου οπτικά χαρακτηριστικά και υψηλού επιπέδου σημασιολογικές οπτικές έννοιες. Οι Tirilly etal. [23] πρότειναν τρόπους για τη βελτίωση του χαρακτηρισμού των εικόνων βασισμένη στην αναπαράσταση με *bag of words*. Συγκεκριμένα, πρότειναν την ύπαρξη οπτικών προτάσεων που αποτελούνται από οπτικές λέξεις σε συγκεκριμένη σειρά, όπως στην περίπτωση κειμένου. Επιπλέον, παρουσίασαν ένα σύστημα ταξινόμησης εικόνων που εκμεταλλεύεται τη σχέση μεταξύ των λέξεων. Το αποτέλεσμα της μελέτης είναι ότι βελτιώνεται η ταξινόμηση των εικόνων σε σύγκριση με την ταξινόμηση βασισμένη στο SVM [27].

Οι Yang etal. [25] εκμεταλλευόμενοι την αναλογία ανάμεσα στην αναπαράσταση *bag of words* των εγγράφων κειμένου, εφαρμόζουν διάφορες τεχνικές γνωστές στην κατηγοριοποίηση κειμένου (βαρύτητές, αποφυγή μετακίνησης λέξης, επιλογή χαρακτηριστικών) για να δημιουργήσουν αναπαραστάσεις εικόνας που διαφέρουν ως προς τη διάσταση, επιλογή και βαρύτητα των οπτικών λέξεων. Αυτή η εργασία παρέχει μια εμπειρική βάση για σχεδιασμό αναπαραστάσεων οπτικών λέξεων που μπορεί να οδηγήσουν σε υψηλές επιδόσεις ταξινόμησης.

ΚΕΦΑΛΑΙΟ 2. ΕΞΑΓΩΓΗ ΣΗΜΑΣΙΟΛΟΓΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

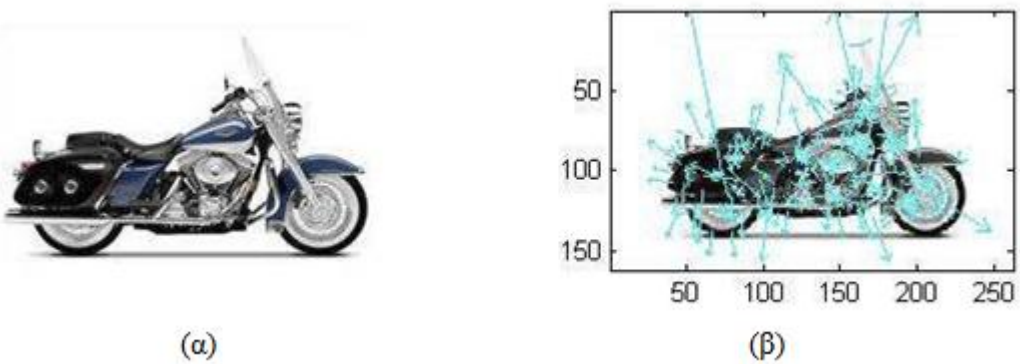
Σε αυτό το κεφάλαιο αναλύεται ο τρόπος εξαγωγής των σημασιολογικών χαρακτηριστικών. Για την εξαγωγή των σημασιολογικών χαρακτηριστικών χρησιμοποιήθηκαν οι περιγραφείς *SIFT* [14]. Ο περιγραφέας *SIFT* δημιουργείται μέσω της κλιμάκωσης, της μετατόπισης της εικόνας στην περιοχή γύρω από ένα σημαντικό σημείο και της δημιουργίας κατάλληλων ιστογραμμάτων κατεύθυνσης. Στη συνέχεια περιγράφεται ο τρόπος δημιουργίας ιστογράμματος οπτικών λέξεων (*BoW*) [20] για κάθε εικονοπλαίσιο. Το ιστόγραμμα οπτικών λέξεων είναι ένα διάνυσμα που περιέχει την συχνότητα εμφάνισης της κάθε οπτικής λέξης μέσα σε ένα εικονοπλαίσιο με τη χρήση και του *soft-weighting* [10]. Έπειτα, παρουσιάζονται τα μοντέλα σημασιολογικών εννοιών που υπάρχουν στη βιβλιογραφία [24],[9] [6]. Από κάθε μοντέλο αντλείται μια πρόβλεψη σχετικά με την ύπαρξη της έννοιας σε ένα εικονοπλαίσιο. Οι προβλέψεις από το σύνολο των μοντέλων συνθέτουν το

σημασιολογικό χαρακτηριστικό διάνυσμα για κάθε εικονοπλαίσιο. Τέλος, περιγράφεται ο ταξινομητής *SVM* [27] που χρησιμοποιήθηκε για τη δημιουργία των μοντέλων των σημασιολογικών εννοιών, καθώς και ο πυρήνας *Chi-square* [18] που ενσωματώθηκε στην παραπάνω διαδικασία.

2.1. Περιγραφείς SIFT

Προκειμένου να εφαρμοστούν οι αλγόριθμοι ταξινόμησης είναι συνήθως αναγκαίο τα δεδομένα να αναπαρασταθούν ως διανύσματα. Ειδικά, στην περίπτωση των εικόνων οι αναπαραστάσεις ονομάζονται διανύσματα περιγραφής (*image descriptors*), που αφορούν είτε την εμφάνιση (*appearance*) είτε το σχήμα (*shape*). Για την εμφάνιση, τα τελευταία χρόνια χρησιμοποιείται ευρέως ο αλγόριθμος *SIFT* (*Scale Invariant Feature Transform*) [14], ο οποίος εξάγει από την εικόνα ορισμένα σημεία-κλειδιά, και ορίζει περιοχές ενδιαφέροντος (*patches*).

Στο παρακάτω σχήμα δίνεται ένα παράδειγμα με τα *patches*, που εντοπίζει ο αλγόριθμος *SIFT*. Συγκεκριμένα, στην Εικόνα 2.1(α) δίνεται η αρχική εικόνα, ενώ στην Εικόνα 2.1(β) απεικονίζονται (με μπλε βέλη) τα *patches* που εξάγει ο αλγόριθμος από αυτήν την αρχική εικόνα.



Εικόνα 4.1 (α)Αρχική Εικόινα (β) Εικόνα με τα 219 Patches που έχουν εξαχθεί από τον Αλγόριθμο *SIFT*.

Ο αλγόριθμος *SIFT* στηρίζεται σε μια προσέγγιση διαδοχικών φίλτρων, δηλαδή αρχικά χρησιμοποιεί αποτελεσματικούς αλγόριθμους για να εντοπίσει τοποθεσίες στην εικόνα, όπου είναι πιθανό να υπάρχουν *patches*, και στη συνέχεια τις εξετάζει με μεγαλύτερη λεπτομέρεια για να εντοπίσει ακριβώς τα *patches*. Ο αλγόριθμος αποτελείται από τα ακόλουθα τέσσερα βήματα:

- Ανίχνευση ακρότατων στο χώρο της κλιμάκωσης (*Scale-space extrema detection*)
- Εντοπισμός σημείων-κλειδιών (*Patch Localization*)
- Ανάθεση προσανατολισμού (*Orientation Assignment*)
- Περιγραφή των σημείων-κλειδιών (*Patch Description*)

Ανίχνευση ακρότατων στο χώρο της κλιμάκωσης

Δεδομένης μιας εικόνας *I (x, y)* , για τον εντοπισμό τοποθεσιών όπου είναι πιθανό να υπάρχουν *patches*, εφαρμόζεται ένα φίλτρο σε κάθε *pixel* της εικόνας, ώστε να μειωθεί ο θόρυβος. Το φίλτρο που χρησιμοποιεί ο αλγόριθμος *SIFT* είναι *Gaussian*

$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$, όπου σ είναι η τυπική απόκλιση. Παίρνοντας τη συνέλιξη (*convolution*) (εξίσωση 2.1) του φίλτρου με την αρχική εικόνα, για τιμές του *σ=1* έως *σ=2* παράγεται μια «οικογένεια» από εικόνες στις οποίες έχει μειωθεί ο θόρυβος (*Gaussian blurred images*):

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \qquad (2.1)$$

Στις εικόνες του πραγματικού κόσμου η εμφάνιση ενός αντικειμένου και συγκεκριμένα η κλίμακά του επηρεάζεται σημαντικά από την απόσταση του από το φωτογράφο. Δηλαδή, όταν η απόσταση αντικειμένου-φωτογράφου είναι μικρή, το αντικείμενο φαίνεται μεγάλο, ενώ όταν η απόσταση αντικειμένου-φωτογράφου είναι μεγάλη, το ίδιο αντικείμενο φαίνεται μικρό. Ωστόσο, τα *patches* που εντοπίζει ο αλγόριθμος θα πρέπει να είναι τα ίδια, ανεξάρτητα από την κλίμακα του αντικειμένου. Επειδή ένα σύστημα τεχνητής όρασης δεν είναι

δυνατόν να γνωρίζει την κλίμακα του αντικειμένου εκ των προτέρων, χρειάζονται αναπαραστάσεις της ίδιας εικόνας, αλλά σε διαφορετική κλίμακα. (*scale space representation*).

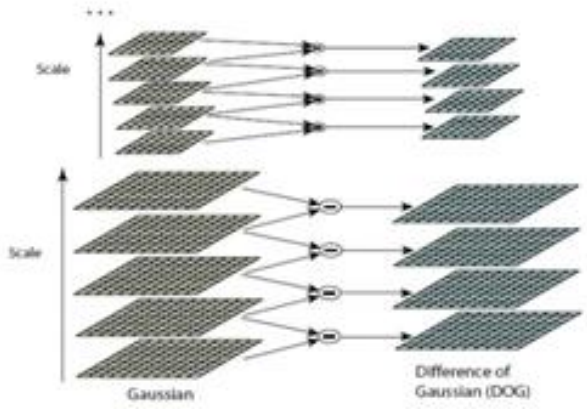
Για να είναι εφικτό το παραπάνω, δηλαδή για να παραχθεί μια σειρά φωτογραφιών με διαφορετική κλίμακα από την αρχική εικόνα, εφαρμόζουμε το φίλτρο *Gauss* με ένα πολλαπλασιαστικό παράγοντα $k \geq 0$, $G(x,y,k\sigma)$ και παίρνουμε τη συνέλιξη του με την αρχική εικόνα $I(x,y)$, οπότε προκύπτει:

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

Στη συνέχεια χρησιμοποιείται η συνάρτηση διαφοράς του *Gauss* (*difference-of-Gauss*) $D(x, y, \sigma)$ και προκύπτει η σχέση (2.3):

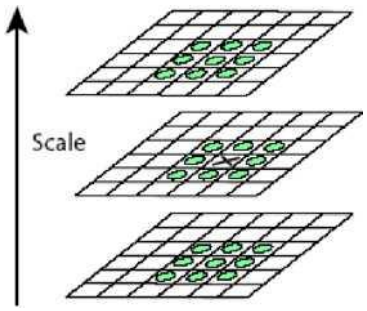
$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{2.3}$$

Από την παραπάνω σχέση γίνεται φανερό ότι αφαιρούμε «γειτονικές» εικόνες, δηλαδή εικόνες που διαφέρουν κατά τον πολλαπλασιαστικό παράγοντα k και παράγεται ένα νέο σύνολο από εικόνες, όπως φαίνεται και από την παρακάτω Εικόνα.



Εικόνα 2.2. (πάνω) Η αρχική εικόνα αλλάζει κλίμακα και γίνεται συνέλιξη των εικόνων που παράγονται με το φίλτρο Gauss. (κάτω) «Γειτονικές» εικόνες αφαιρούνται και παράγεται η συνάρτηση διαφοράς του Gauss.

Το επόμενο βήμα αφορά στον εντοπισμό του τοπικού ελαχίστου και μεγίστου της συνάρτησης διαφοράς του *Gauss*. Για να γίνει αυτό, συγκρίνουμε κάθε *pixel* της εικόνας με τα 8 γειτονικά της, καθώς και με τα 26 κοντινότερα *pixels* των «γειτονικών» της εικόνων όπως φαίνεται στην Εικόνα 2.3. Επιλέγεται το pixel του οποίου η συνάρτηση διαφορών του *Gauss* έχει τιμή μεγαλύτερη (εύρεση μεγίστου) ή μικρότερη (εύρεση ελαχίστου) από τη συνάρτηση διαφορών του *Gauss* των υπόλοιπων *pixels* με τα οποία συγκρίνεται.



Εικόνα 2.3 Το Pixel που σημειώνεται με X συγκρίνεται με τα 8 γειτονικά του στην ίδια κλίμακα και με τα 18 Pixels στις «γειτονικές» του εικόνες.

Εντοπισμός σημείων-κλειδιών

Για να γίνει ο εντοπισμός των σημείων-κλειδιών (key-points), αρχικά χρησιμοποιείται για κάθε υποψήφιο σημείο-κλειδί παρεμβολή των κοντινών σημείων. Η παρεμβολή υπολογίζεται με τετραγωνικό ανάπτυγμα *Taylor* της συνάρτησης Διαφοράς του *Gauss* $D(x,y,\sigma)$ ως εξής:

$$D(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x \qquad (2.4)$$

Όπου η συνάρτηση D και οι παράγωγοί της υπολογίζονται για το υποψήφιο σημείο-κλειδί και $x = (x, y, \sigma)$ είναι η απόσταση από το σημείο κλειδί. Η τοποθεσία του ακρότατου x' καθορίζεται υπολογίζοντας την παράγωγο ως προς x και θέτοντάς την ίση με 0.

Εάν η απόσταση του ακρότατου είναι μεγαλύτερη από 0.5 σε οποιαδήποτε διάσταση, αυτό είναι ένδειξη ότι το ακρότατο βρίσκεται πιο κοντά σε άλλο σημείο-κλειδί. Σε αυτήν την περίπτωση το υποψήφιο σημείο-κλειδί αλλάζει και η παρεμβολή υπολογίζεται στο σημείο αυτό. Αλλιώς, η απόσταση προστίθεται στο υποψήφιο σημείο-κλειδί για να πάρουμε την τοποθεσία του ακρότατου.

Για να απορρίψουμε τα σημεία-κλειδιά με μικρή αντίθεση υπολογίζεται το ανάπτυγμα *Taylo*r δεύτερης τάξης στο ακρότατο. Αν η τιμή είναι μικρότερη από 0.03 τότε το σημείο-κλειδί απορρίπτεται. Αλλιώς διατηρείται με τελική θέση που δίνεται από τον τύπο:

$$y + x' \tag{2.5}$$

και κλίμακα *σ,* όπου *y* είναι η αρχική τοποθεσία και του σημείου-κλειδιού σε κλίμακα *σ*.

Η συνάρτηση Διαφοράς του *Gauss* έχει υψηλή τιμή κατά μήκος των ακμών, ακόμη και αν το υποψήφιο σημείο-κλειδί δεν είναι ανθεκτικό σε μικρά επίπεδα θορύβου. Επομένως, για να αυξηθεί η ευρωστία πρέπει να περιοριστούν τα σημεία-κλειδιά τα οποία έχουν μεγάλη απόκριση, αλλά δεν έχει προσδιοριστεί επαρκώς η θέση τους.

Για αυτό το λόγο υπολογίζονται τα ιδιοδιανύσματα του Εσσιανού πίνακα δεύτερης τάξης:

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{2.6}$$

Ανάθεση προσανατολισμού

Σε αυτό το βήμα απαιτείται να ανατεθεί προσανατολισμός σε κάθε *patch* που εντοπίστηκε. Ο προσανατολισμός είναι απαραίτητος, γιατί είναι πιθανόν σε δύο εικόνες το ίδιο αντικείμενο να έχει φωτογραφηθεί με διαφορετικές γωνίες περιστροφής. Ωστόσο, ο αλγόριθμος θα πρέπει να είναι σε θέση να εντοπίζει *patches*

σε ένα αντικείμενο, ακόμη και αν αυτό είναι περιστραμμένο. Για την ανάθεση του προσανατολισμού ακολουθείται η διαδικασία που περιγράφεται στη συνέχεια.

Δεδομένης μιας *Gaussian blurred* εικόνας *L(x,y)* στην οποία έχουμε εντοπίσει το *patch*, για κάθε *pixel* αυτής υπολογίζουμε το μέγεθος κλίσης m(*x,y)* και τον προσανατολισμό *θ*(*x,y)* χρησιμοποιώντας τα τέσσερα γειτονικά *pixels*. Ο υπολογισμός γίνεται με τις ακόλουθες σχέσεις:

$$m(x,y) = \sqrt{(L(x+1,y) - L(\mathrm{x}-1,y))^2 + (L(\mathrm{x},\mathrm{y}+1) - \mathrm{L}(\mathrm{x},\mathrm{y}-1))^2} \qquad (2.7)$$

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1)) / (L(x+1,y) - L(x-1,y))) \qquad (2.8)$$

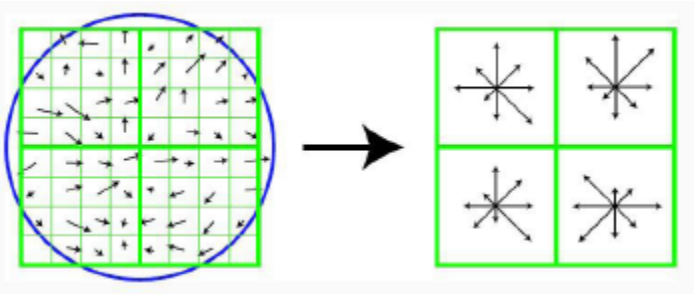
Στη συνέχεια επιλέγονται ορισμένα *pixels* γύρω από το *patch* που έχουν εντοπιστεί και για τα *pixels* αυτά φτιάχνεται ένα ιστόγραμμα με τον προσανατολισμό τους.

Τέλος, από το ιστόγραμμα επιλέγεται ο προσανατολισμός με την υψηλότερη συχνότητα, καθώς και οι προσανατολισμοί που αντιστοιχούν έως το 80% της υψηλότερης συχνότητας και ανατίθενται στο *patch*.

Περιγραφή των σημείων-κλειδιών

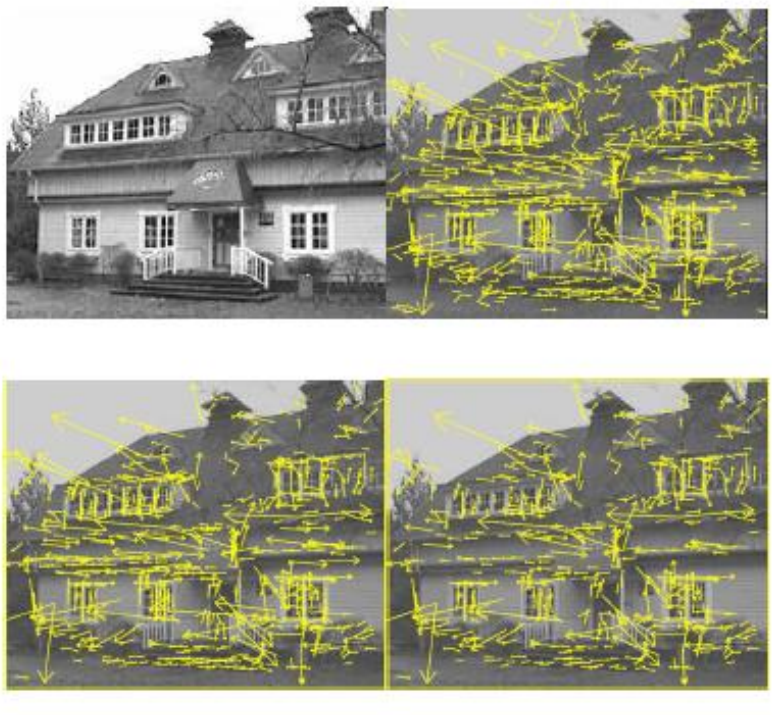
Το τελικό στάδιο είναι η περιγραφή κάθε *patch* με ένα διάνυσμα, το οποίο να περιλαμβάνει πληροφορία για την θέση, τον προσανατολισμό και την κλίμακά του.

Αρχικά υπολογίζεται το μέγεθος κλίσης και ο προσανατολισμός για κάθε *pixel*, που είναι γειτονικό του *patch* που θέλουμε να περιγράψουμε, όπως φαίνεται και στην παρακάτω Εικόνα (αριστερά) και στις υπολογιζόμενες τιμές εφαρμόζεται γκαουσιανό παράθυρο (μπλε κύκλος). Στη συνέχεια για κάθε μία από τις 4 υποπεριοχές κατασκευάζεται ένα ιστόγραμμα με τους προσανατολισμούς και αυτό έχει ως αποτέλεσμα να παράγεται ένα διάνυσμα περιγραφής διάστασης 2x2, όπως φαίνεται και δεξιά στην Εικόνα 2.4.



Εικόνα 2.4 Υπολογισμός ενός περιγραφέα σημαντικού σημείου (key-point).

Ο τυπικός περιγραφέας *SIFT* ενός σημαντικού σημείου, ο οποίο χρησιμοποιείται και στη παρούσα εργασία, δημιουργείται μέσω της κλιμάκωσης, της μετατόπισης της εικόνας στην περιοχή γύρω από το σημαντικό σημείο και της δημιουργίας κατάλληλων ιστογραμμάτων κατεύθυνσης (Εικόνα 2.5). Πιο συγκεκριμένα δημιουργείται ένα 4x4 διάνυσμα από ιστογράμματα με 8 κάδους κατεύθυνσης το καθένα. Τέλος, το τελικό διάνυσμα 128 στοιχείων κανονικοποιείται για να γίνει ανθεκτικότερο στις αλλαγές φωτεινότητας. Για επιπλέον ευρωστία εφαρμόζεται φίλτρο που δεν επιτρέπει να υπάρχουν στοιχεία με τιμή πάνω από 0.2 και στη συνέχεια γίνεται και πάλι κανονικοποίηση. Πρέπει να αναφερθεί ότι οι περιγραφείς *SIFT* αποτελούν μία πολύ αποτελεσματική και δημοφιλή προσέγγιση για πολλά προβλήματα υπολογιστικής όρασης.



Εικόνα 2.5 Στάδια επιλογής των σημαντικών σημείων. 1) Η αρχική εικόνα 2) Οι αρχικές 832 θέσεις των σημαντικών σημείων στα ελάχιστα και στα μέγιστα των διαφορών της Γκαουσιανής συνάρτησης. Τα σημεία ενδιαφέροντος απεικονίζονται ως διανύσματα που δηλώνουν την κλίμακα, την κατεύθυνση και τη θέση 3) Τα 729 σημαντικά σημεία που απομένουν εφαρμόζοντας ένα κατώφλι ελάχιστης φωτεινότητας 4) Τα τελικά 536 σημαντικά σημεία που απομένουν αφού εφαρμοστεί και ένα επιπλέον κατώφλι στον λόγο της κύριας καμπυλότητας.

2.2. Ιστόγραμμα Οπτικών Λέξεων

Ο εντοπισμός σημασιολογικών εννοιών είναι ένα θέμα μεγάλου ενδιαφέροντος καθώς παρέχει σημασιολογικά φίλτρα που βοηθούν στην ανάλυση πολυμεσικών δεδομένων. Είναι στην ουσία μια διαδικασία ταξινόμησης που καθορίζει αν μια εικόνα ή μια λήψη ενός βίντεο σχετίζονται με μια συγκεκριμένη σημασιολογική έννοια. Οι σημασιολογικές έννοιες καλύπτουν μια μεγάλη ποικιλία θεμάτων όπως αυτά που σχετίζονται με αντικείμενα (π.χ. αυτοκίνητο, αεροπλάνο), εσωτερικά/ εξωτερικά τοπία (π.χ. αίθουσα συνεδριάσεων, έρημος), γεγονότα (π.χ. διαδήλωση ανθρώπων) κ.α. Ο αυτόματος εντοπισμός των εννοιών είναι μια μεγάλη πρόκληση εξαιτίας της μεταβολής των κατηγοριών, της μη σταθερότητας του

υπόβαθρου, τις αλλαγές της στάσης και του φωτισμού στις εικόνες και στις λήψεις βίντεο. Ολικά χαρακτηριστικά, όπως τα ιστογράμματα χρώματος, δεν αντιμετωπίζουν αυτές τις δυσκολίες γεγονός που οδηγεί στη χρησιμοποίηση και εξέλιξη τοπικών αμετάβλητων χαρακτηριστικών (*keypoints*) τα τελευταία χρόνια. Τα χαρακτηριστικά σημεία (*keypoints*), είναι σημαντικά κομμάτια που περιέχουν πλούσια τοπική πληροφορία σχετικά με μια εικόνα. Η πιο γνωστή αναπαράσταση σημείων –κλειδιά είναι η *bag of visual words* (*BoW*) [20]. Για κάθε εικονοπλαίσιο υπολογίζεται ένας διαφορετικός αριθμός περιγραφέων που περιγράφουν συγκεκριμένα αντικείμενα ή σημεία ενδιαφέροντος. Συγκεκριμένα, υποθέτουμε ότι μας δίνεται ένα σύνολο από *N* εικονοπλαίσια $\{f_1, ..., f_N\}$. Για κάθε εικονοπλαίσιο f_i, $i=1,...,N$, εξάγεται ένα σύνολο από *SIFT* περιγραφείς D_i. Έπειτα, όλα τα σύνολα των περιγραφέων συνενώνοται για να περιγράψει ολόκληρο το σύνολο περιγραφέων D_s από τα *N* εικονοπλαίσια.

$$D_s = D_1 \bigcup ... \bigcup D_N \tag{2.9}$$

Για να εξαχθούν οπτικές λέξεις από τους περιγραφείς, το σύνολο των περιγραφέων ομαδοποιείται σε *k* ομάδες $\{C_1, C_2, ..., C_k\}$, όπου *k* είναι το σύνολο του μεγέθους του λεξικού των οπτικών λέξεων. Ο αλγόριθμος που χρησιμοποιήθηκε για την ομαδοποίηση είναι ο *k-means*. Ο αλγόριθμος ομαδοποίησης *k-means* είναι αρκετά απλός. Αρχικά, επιλέγουμε k αρχικά κέντρα, το *k* ορίζεται από τον χρήστη και αντιπροσωπεύει το επιθυμητό πλήθος των ομάδων (*clusters*). Στη συνέχεια, κάθε σημείο ανατίθεται στο κοντινότερό του κέντρο, και κάθε σύνολο σημείων που έχει ανατεθεί στο ίδιο κέντρο αποτελεί και μία ομάδα. Αφού ολοκληρωθεί αυτή η διαδικασία, τα κέντρα κάθε ομάδας επαναπροσδιορίζονται με βάση τα νέα σημεία της κάθε ομάδας. Η διαδικασία της ανάθεσης των σημείων και του επαναπροσδιορισμού των κέντρων επαναλαμβάνεται μέχρις ότου να μην αλλάζουν οι ομάδες, ή ισοδύναμα, τα κέντρα των ομάδων να παραμένουν σταθερά. Για την κατασκευή του ιστογράμματος των οπτικών λέξεων (*bag of visual words*) για κάθε εικονοπλαίσιο, το αντίστοιχο σύνολο περιγραφέων του D_i, αντιστοιχίζεται στις *k* οπτικές λέξεις καταλήγοντας σε ένα διάνυσμα που περιέχει τον σταθμισμένο αριθμό εμφάνισης κάθε οπτικής λέξης σε κάθε εικονοπλαίσιο.

Επομένως, γνωρίζοντας ότι το εικονοπλαίσιο f_i έχει D περιγραφείς $d_1,...,d_D$, το ιστόγραμμα των οπτικών λέξεων για κάθε εικονοπλαίσιο ορίζεται ως:

$$VH_i(l) = \frac{\left|\{d_j \in C_l,\ j = 1,...,D\}\right|}{D}, l = 1,...\mathrm{k} \qquad (2.10)$$

2.3. Vireo-374: Ανιχνευτές Σημασιολογικών Εννοιών βασισμένοι σε Χαρακτηριστικά Σημεία

Ο εντοπισμός εννοιών σε βίντεο έχει σαν στόχο να κατατάξει λήψεις βίντεο σύμφωνα με την παρουσία σημασιολογικών εννοιών (όπως «αθλήματα», «γραφήματα», «διαδήλωση ανθρώπων» κ.α.). Αυτές οι έννοιες μπορούν να λειτουργήσουν σαν σημασιολογικά φίλτρα για την αναζήτηση βίντεο στο διαδίκτυο. Για παράδειγμα, μια ερώτηση «βρες στρατιωτικό όχημα» μπορεί εύκολα να απαντηθεί αν επιστραφούν λήψεις του βίντεο οι οποίες πολύ πιθανό θα περιέχουν τις έννοιες «στρατός» και «όχημα». Ταξινομητές (όπως *SVM* [27]) εκπαιδεύονται με διάφορα χαρακτηριστικά που εξάγονται από παραδείγματα εκπαίδευσης, και οι ταξινομητές που εκπαιδεύτηκαν μπορούν στη συνέχεια να χρησιμοποιηθούν για εντοπισμό εννοιών. Με ένα μεγάλο σύνολο από ισχυρούς ανιχνευτές εννοιών, μπορεί να επιτευχθεί σημαντική βελτίωση σε πολλές απαιτητικές εφαρμογές, όπως αναζήτηση εικόνας και περίληψης [24].

Για να εκπαιδευτούν οι ανιχνευτές εννοιών, ένα κρίσιμο βήμα είναι να αποκτηθεί ένα αρκετά μεγάλο πλήθος δεδομένων εκπαίδευσης, το οποίο δεν είναι μια εύκολη διαδικασία. Ευτυχώς με τη πληθώρα των μέσων μαζικής ενημέρωσης, υπάρχουν όλο και περισσότερες ψηφιακές εικόνες διαθέσιμες στο δίκτυο. Πολλά σύνολα δεδομένων περιέχουν χιλιάδες εικόνες οι οποίες συλλέγονται από ιστοσελίδες όπως το *Flickr* και έχουν πρόσφατα διαθέτει για την έρευνα. Μια προσπάθεια για εκπαίδευση ανιχνευτών εννοιών έγινε από την *LSCOM* η οποία όρισε 1000+ σημασιολογικές έννοιες και χαρακτήρισε 400+ από αυτές πάνω σε ένα σύνολο από βίντεο με ειδήσεις [31].

Ένα από τα συστήματα ανίχνευσης εννοιών που χρησιμοποιήθηκε στην εργασία μας, είναι από το εργαστήριο *DVMM* στο πανεπιστήμιο της Κολούμπια το οποίο έδωσε για κοινή χρήση 374 *LSCOM* ανιχνευτές σημασιολογικών εννοιών (Columbia374) [24]. Οι *VIREO-374* ανιχνευτές σημασιολογικών εννοιών εκπαιδεύτηκαν στα πλαίσια του διαγωνισμού *TRECVID-2005* [16] χρησιμοποιώντας *LSCOM* χαρακτηρισμό [31].

Επιπλέον, οι ανιχνευτές *DoG*[14] και *SIFT*[14] χρησιμοποιήθηκαν για την ανίχνευση και περιγραφή των σημείων κλειδιών που εξήχθησαν από τα εικονοπλαίσια. Ένα οπτικό λεξικό 500 οπτικών λέξεων δημιουργήθηκε από την ομαδοποίηση ενός συνόλου ~500χιλ. χαρακτηριστικών *SIFT*. Με αυτό οπτικό λεξιλόγιο, κάθε εικονοπλαίσιο του συνόλου εκπαίδευσης μπορεί να αναπαρασταθεί από ένα διάνυσμα χαρακτηριστικών διάστασης 500, το οποίο αναπαριστά τη συχνότητα εμφάνισης της κάθε οπτικής λέξης στο εικονοπλαίσιο. Χρησιμοποιήθηκε ακόμα το σύστημα soft-weighting [10] για να υπολογίσει τη σημαντικότητα γειτονικών οπτικών λέξεων στο εικονοπλαίσιο (key-frame). Κάθε χαρακτηριστικό σημείο σε ένα εικονοπλαίσιο, αντί να αντιστοιχίζεται μόνο στην κοντινότερη οπτική λέξη, επιλέγονται οι *N*-κοντινότερες οπτικές λέξεις: Υποθέτουμε ότι έχουμε ένα οπτικό λεξικό από *K* οπτικές λέξεις, χρησιμοποιούμε ένα διάνυσμα *K* διαστάσεων $T=[t_1,\ldots,t_k,\ldots,t_K]$, με κάθε t_k να αναπαριστά το βάρος της οπτικής λέξης *k* σε μια εικόνα:

$$t_k = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(\mathrm{j},\mathrm{k}), \tag{2.11}$$

όπου M_i είναι ο αριθμός των χαρακτηριστικών σημείων των οποίων ο *i*-οστός κοντινότερος γείτονας είναι η οπτική λέξη *k*. Η μετρική *sim(j,k)* αναπαριστά την ομοιότητα ανάμεσα στο χαρακτηριστικό σημείο *j* και στην οπτική λέξη *k*. Σημειώνεται επίσης ότι η συμβολή ενός χαρακτηριστικού σημείου εξαρτάται από την ομοιότητά του με μια λέξη *k* με βάρος $\frac{1}{2^{i-1}}$, που αναπαριστά τη λέξη με τον *i*-οστό κοντινότερο γείτονα.

Ο ταξινομητής που χρησιμοποιήθηκε για τον εντοπισμό των σημασιολογικών εννοιών είναι ο *SVM* [27]. Ο *SVM* ψάχνει για γραμμικώς διαχωριζόμενα υπερεπίπεδα χρησιμοποιώντας διανύσματα υποστήριξης («κρίσιμα» παραδείγματα εκπαίδευσης) και περιθώρια (margins) (που καθορίζονται από τα διανύσματα υποστήριξης). Ο *SVM* είναι ο πιο γνωστός ταξινομητής για ταξινόμηση εικόνων που βασίζονται σε σύνολο οπτικών λέξεων (*BoW*). Για SVM δυο κατηγοριών, η συνάρτηση απόφασης για ένα παράδειγμα ελέγχου *x* έχει την ακόλουθη μορφή:

$$g(x) = \sum_i a_i y_i K(x_i, x) - b, \tag{2.12}$$

όπου $K(x_i,x)$ είναι η έξοδος της συνάρτησης πυρήνα για το παράδειγμα εκπαίδευσης x_i και το παράδειγμα ελέγχου *x*, το οποίο μετράει την ομοιότητα ανάμεσα σε δυο παραδείγματα. Το y_i είναι η ετικέτα της κατηγορίας του x_i, a_i είναι το βάρος του παραδείγματος εκπαίδευσης x_i και *b* είναι η πόλωση. Ο πυρήνας που χρησιμοποιείται είναι ο *Chi-square* [18], ο οποίος είναι ένας γενικευμένος πυρήνας *RBF* και δίνει βέλτιστο αποτέλεσμα για την ταξινόμηση ιστογραμμάτων οπτικών λέξεων και αναλύεται στο υποκεφάλαιο 2.5. Το σύνολο των 374 σημασιολογικών εννοιών φαίνεται στον παρακάτω πίνακα:

Πίνακας 2.1 Οι 374 σημασιολογικές έννοιες βάσει των οποίων κατασκευάστηκαν οι ανιχνευτές vireo-374.

'Actor'	'Demonstration_Or_Protest'	'Kitchen'	'Shopping_Mall'
'Address_Or_Speech'	'Desert'	'Judge'	'Sidewalks'
'Administrative_Assistant'	'Dining_Room'	'Laboratory'	'Singing'
'Adobehouses'	'Dirt_Gravel_Road'	'Landlines'	'Single_Family_Homes'
'Adult'	'Ditch'	'Lakes'	'Single_Person_Female'
'Agent'	'Dogs'	'Landscape'	'Single_Person_Male'
'Agricultural_People'	'Donald_Rumsfeld'	'Laundry_Room'	'Single_Person'
'Aircraft_Cabin'	'Dredge_Powershovel_Dragline'	'Lawn'	'Sitting'
'Airplane_Flying'	'Dresses_Of_Women'	'Lawyer'	'Sketches'
'Airplane_Landing'	'Dresses'	'Logos_Full_Screen'	'Sky'
'Airplane_Takeoff'	'Driver'	'Machine_Guns'	'Smoke_Stack'
'Airplane'	'Earthquake'	'Male_Anchor'	'Smoke'
'Airport_Or_Airfield'	'Election_Campaign_Address'	'Male_News_Subject'	'Snow'

'Airport'	'Election_Campaign_Convention'	'Male_Person'	'Soccer'
'Alley'	'Election_Campaign_Debate'	'Male_Reporter'	'Soldiers'
'Animal_Pens_And_Cages'	'Election_Campaign_Greeting'	'Maps'	'Speaker_At_Podium'
'Animal'	'Election_Campaign'	'Medical_Personnel'	'Speaking_To_Camera'
'Antenna'	Emergency_Medical'	'Meeting'	'Sports'
'Apartment_Complex'	'Emergency_Room'	'Microphones'	'Stadium'
'Apartments'	'Emergency_Vehicles'	'Military_Base'	'Standing'
'Armed_Person'	'Entertainment'	'Military_Buildings'	'Steeple'
'Armored_Vehicles'	'Exiting_Car'	'Military_Personnel'	'Still_Image'
'Artillery'	'Exploding_Ordinance'	'Military'	'Stock_Market'
'Asian_People'	'Explosion_Fire'	'Moonlight'	'Store'
'Athlete'	'Eyewitness'	'Mosques'	'Street_Battle'
'Attached_Body_Parts'	'Face'	'Motorcycle'	'Streets'
'Baby'	'Factory_Worker'	'Mountain'	'Striking_People'
'Backpack'	'Factory'	'Muddy_Scenes'	'Studio_With_Anchorperson'
'Backpackers'	'Farms'	'Mug'	'Studio'
'Baker'	'Female_Anchor'	'Muslims'	'Suburban'
'Bar_Pub'	'Female_News_Subject'	'Natural-Disaster'	'Suits'
'Baseball'	'Female_Person'	'Natural_Disasters'	'Sunglasses'
'Basketball'	'Female_Reporter'	'Network_Logo'	'Sunny'
'Bathroom'	'Fields'	'News_Studio'	'Supermarket'
'Bazaar'	'Fighter_Combat'	'Newspapers'	'Swimmer'
'Beach'	'Finance_Busines'	'Nighttime'	'Swimming_Pools'
'Beards'	'Firefighter'	'Non-uniformed_Fighters'	'Swimming'
'Bicycle'	'First_Lady'	'Non-us_National_Flags'	'Talking'
'Bicycles'	'Flag-US'	'Observation_Tower'	'Tanks'
'Birds'	'Flags'	'Oceans'	'Telephones'
'Blank_Frame'	'Flood'	'Office_Building'	'Television_Tower'
'Boat_Ship'	'Flowers'	'Officers'	'Tennis'
'Body_Parts'	'Flying_Objects'	'Office'	'Tent'
'Bomber_Bombing'	'Food'	'Oil_Drilling_Site'	'Text_Labeling_People'
'Boy'	'Football'	'Oil_Field'	'Text_On_Artificial_Background'
'Bride'	'Forest'	'Old_People'	'Throwing'
'Bridges'	'Foxhole'	'Outdoor'	'Ties'
'Briefcases'	'Free_Standing_Structures'	'Outer_Space'	'Tony_Blair'
'Building'	'Freighter'	'Overlaid_Text'	'Tower'
'Bus'	'Funeral'	'Parade'	'Traffic'
'Business_People'	'Furniture'	'Parking_Lot'	'Trees'
'Cables'	'Gas_Station'	'Pavilions'	'Tropical_Settings'
'Camera'	'George_Bush'	'Peacekeepers'	'Truck'
'Canal'	'Girl'	'Pedestrian_Zone'	'Tunnel'

'Canoe'	'Glass'	'People-Marching'	'Underwater'
'Capital'	'Glasses'	'People_Crying'	'Urban_Park'
'Car_Crash'	'Golf_Course'	'People_Marching'	'Urban_Scenes'
'Car_Racing'	'Golf_Player'	'Photographers'	'Urban'
'Car'	'Golf'	'Person'	'Us_Flags'
'Cart_Path'	'Government-Leader'	'Pickup_Truck'	'Valleys'
'Castle'	'Government_Leader'	'Pipes'	'Vegetation'
'Caucasians'	'Grandstands_Bleachers'	Police_Private_Security'	'Vehicle'
'Celebration_Or_Party'	'Grassland'	'Police_Security'	'Walking_Running'
'Celebrity_Entertainment'	'Graveyard'	'Police'	'Walking'
'Cell_Phones'	'Greeting'	'Politics'	'Warehouse'
'Charts'	'Groom'	'Power_Plant'	'Water_Tower'
'Cheering'	'Ground_Combat'	Power_Transmission_Line'	'Waterscape_Waterfront'
'Child'	'Ground_Crew'	'Powerlines'	'Waterways'
'Cigar_Boats'	'Ground_Vehicles'	'Powerplants'	'Weapons'
'Cityscape'	'Group'	'Press_Conference'	'Weather'
'Civilian_Person'	'Guard'	'Prisoner'	'White_House'
'Classroom'	'Guest'	'Processing_Plant'	'Windows'
'Clearing'	'Gym'	'Protesters'	'Windy'
'Clock_Tower'	'Hand'	'Radar'	'Yasser_Arafat'
'Clouds'	'Handshaking'	'Raft'	
'Cloverleaf'	'Harbors'	'Railroad'	
'Coal_Powerplants'	'Head_And_Shoulder'	'Rainy'	
'Colin_Powell'	'Head_Of_State'	'Religious_Figures'	
Commentator_Or_StudioExpert'	'Helicopter_Hovering'	'Reporters'	
'Commercial_Advertisement'	'Helicopters'	'Residential_Buildings'	
ComputerOrTelevisionScreens'	'High_Security_Facility'	'Rifles'	
'Computer_TV-screen'	'Highway'	'Riot'	
'Computers'	'Hill'	'River_Bank'	
'Conference_Buildings'	'Horse'	'River'	
'Conference_Room'	'Hospital'	'Road_Block'	
'Congressman'	'Host'	'Road_Overpass'	
'Construction_Site'	'Hotel'	'Road'	
'Construction_Vehicles'	'House_Of_Worship'	'Rocky_Ground'	
'Construction_Worker'	'House'	'Rowboat'	
'Cordless'	'Hu_Jintao'	'Room'	
'Corporate-Leader'	'Individual'	'Rpg'	
'Corporate_Leader'	'Indoor_Sports_Venue'	'Ruins'	
'Court'	'Industrial_Setting'	'Running'	
'Courthouse'	'Insurgents'	'Runway'	
'Crowd'	'Infants'	'Scene_Text'	

'Cul-de-sac'	'Interview_On_Location'	'School'	
'Dancing'	'Interview_Sequences'	'Science_Technology'	
'Dark-skinned_People'	'Islands'	'Security_Checkpoint'	
'Daytime_Outdoor'	'John_Edwards'	'Ship'	
'Dead_Bodies'	'John_Kerry'	'Shooting'	

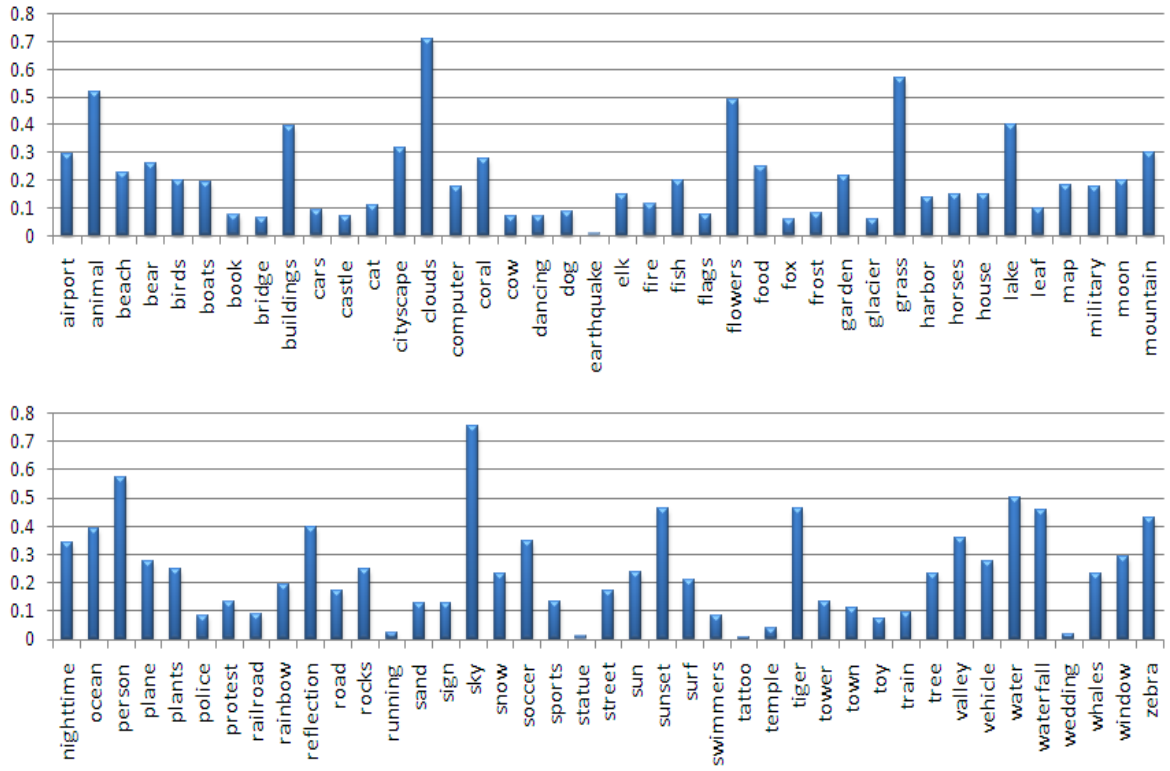
2.4. Vireo-Web81: Ανιχνευτές Σημασιολογικών Εννοιών από την Εκπαίδευση Εικόνων του Διαδικτύου

Γενικά όπως προαναφέραμε, ο εντοπισμός εννοιών είναι μια διαδικασία ταξινόμησης που καθορίζει αν ένα αντικείμενο πολυμέσων (π.χ. μια εικόνα) σχετίζεται με την έννοια-στόχο.

Για την εκπαίδευση ανιχνευτών 81 σημασιολογικών εννοιών αξιοποιήθηκε το *NUS-WIDE* [6], ένα δημοφιλές σύνολο δεδομένων με εικόνες διαδικτύου οι οποίες συγκεντρώθηκαν από ερευνητές από το Εθνικό Πανεπιστήμιο της Σιγκαπούρης, και το οποίο περιέχει περίπου 260 χιλιάδες εικόνες με χειρονακτικό χαρακτηρισμό από 81 εννοιολογικές κατηγορίες. Συγκεκριμένα, το σύνολο δεδομένων που αξιοποιήθηκε περιέχει ένα σύνολο από εικόνες που συγκεντρώθηκαν από το *Flickr*, μαζί με τις σχετικές ετικέτες, όπως επίσης και το *ground truth* για 81 έννοιες για αυτές τις εικόνες. Για κάθε έννοια, τρεις ταξινομητές έχουν εκπαιδευτεί αντίστοιχα βασιζόμενοι στα σύνολα οπτικών λέξεων, σε πλέγμα χρωματικών στιγμιότυπων (color moment) και την κυματική υφή (Gabor texture). Το ιστόγραμμα οπτικών λέξεων αξιοποιεί το ίδιο οπτικό βιβλίο με κώδικες με το Vireo-374 το οποίο περιέχει 500 οπτικές λέξεις και εφαρμόζεται το soft-weighting ως διάνυσμα ποσοτικοποίησης. Και στην περίπτωση αυτή για να εξαχθούν τα χαρακτηριστικά του συνόλου των οπτικών λέξεων χρησιμοποιήθηκε ο πυρήνας Chi- square [18]. Το σύνολο των ανιχνευτών ονομάζεται *vireo web 81*[30].

Στη συνέχεια δίνεται ένα παράδειγμα για 81 έννοιες που εκπαιδεύτηκαν χρησιμοποιώντας το σύνολο δεδομένων *NUS-WIDE* (~160k) και το οποίο φαίνεται στην Εικόνα 2.6. Οι εικόνες αυτές συγκεκριμένα, δείχνουν τη μέση ακρίβεια της

απόδοσης για κάθε έννοια από τους ανιχνευτές πάνω στο σύνολο δεδομένων *NUS-WIDE* (~100k).



Εικόνα 2.6 Απόδοση για κάθε έννοια από τους 81 ανιχνευτές πάνω στο σύνολο δεδομένων *NUS-WIDE*.

2.5. Ταξινομητής SVM

Η μέθοδος των Μηχανών Διανυσμάτων Υποστήριξης *SVM (Support Vector Machine)* [27], έχει εδραιωθεί ως μια από τις πιο διαδεδομένες μεθόδους ταξινόμησης, αποτελώντας συνήθως τη βέλτιστη επιλογή για προβλήματα, όπως η ταξινόμηση κειμένων (*text categorization*), η αναγνώριση γραφής (*handwriting recognition*) και η ταξινόμηση δεδομένων έκφρασης γονιδίων (*gene expression data*). Η μέθοδος χρησιμοποιείται τόσο για δυαδικά προβλήματα (*two-class*) όσο και για προβλήματα πολλών κατηγοριών (*multiclass*). Στο στάδιο της εκπαίδευσης το *SVM* προσπαθεί να βρει ένα υπερεπίπεδο απόφασης για να διαχωρίσει τα δεδομένα του συνόλου εκπαίδευσης. Μάλιστα, η ιδιαιτερότητα του SVM έγκειται στο ότι

προσπαθεί να βρει το καλύτερο υπερεπίπεδο, το οποίο περιγράφεται με τον όρο υπερεπίπεδο μέγιστου περιθωρίου (*maximum margin hyperplane*) και είναι εκείνο που διαχωρίζει τις κατηγορίες με τη μεγαλύτερη δυνατή απόσταση (περιθώριο).

Δεδομένου ενός συνόλου $\{(x_i, y_i)\}_{i=1}^{l}$ και δύο κατηγοριών $y_i \in \{-1,1\}$ η επιφάνεια απόφασης έχει τη μορφή:

$$a \cdot x + b = 0 \tag{2.13}$$

όπου *α* και *b* είναι οι παράμετροι του μοντέλου και το περιθώριο (margin) δίνεται από τον τύπο $\text{margin} = 2 / \|a\|$.

Δεδομένου ότι πρέπει τα παραδείγματα να ταξινομούνται σωστά και το υπερεπίπεδο να έχει το μεγαλύτερο εύρος, το πρόβλημα ελαχιστοποίησης που επιχειρεί να επιλύσει το SVM είναι:

$$\min \frac{\| a \|^2}{2} \tag{2.14}$$

με τους περιορισμούς ότι $y_i(a * x_i + b) \geq 1, \ \forall i = 1,2,\ldots,l$

Το παραπάνω είναι ένα πρόβλημα βελτιστοποίησης με περιορισμούς (*constraint optimization problem*), για το οποίο απαιτείται προσδιορισμός των παραμέτρων (α,b). Το παραπάνω πρόβλημα επιλύεται με πολλαπλασιαστές *Lagrange* λ_i και η διαχωριστική επιφάνεια απόφασης είναι η

$$(\sum_{i=1}^{l} \lambda_i y_i x_i \bullet x) + b = 0 \tag{2.15}$$

Επειδή στα περισσότερα προβλήματα τα δεδομένα δεν είναι γραμμικώς διαχωρίσιμα, ο αλγόριθμος που συζητήθηκε παραπάνω δεν μπορεί να χρησιμοποιηθεί για την εύρεση της επιφάνειας απόφασης. Αντίθετα, χρησιμοποιείται μια λίγο διαφορετική προσέγγιση στην εκπαίδευση του *SVM*, η οποία προσπαθεί να βρει το υπερεπίπεδο χαλαρού περιθωρίου (*soft margin hyperplane*), δηλαδή την επιφάνεια απόφασης που διαχωρίζει τα δεδομένα κάνοντας τα λιγότερα λάθη. Σε αυτή την περίπτωση χρησιμοποιούνται οι βοηθητικές μεταβλητές $\xi_i \geq 0$ (*stack variables*).

Σε κάθε παράδειγμα του συνόλου εκπαίδευσης αντιστοιχεί μια τιμή ξ_i , ανάλογα με τη θέση του ως προς τη διαχωριστική επιφάνεια. Συγκεκριμένα υπάρχουν τρεις περιπτώσεις: Όταν ένα παράδειγμα βρίσκεται στη σωστή πλευρά του υπερεπιπέδου και σε απόσταση μεγαλύτερη ή ίση της ελάχιστης απαιτούμενης απόστασης, τότε $\xi_i=0$. Αν το πρότυπο βρίσκεται στη σωστή πλευρά του υπερεπιπέδου, αλλά όχι σε ικανοποιητική απόσταση, τότε $0<\xi_i<1$. Τέλος, αν το πρότυπο βρίσκεται στη λάθος πλευρά του υπερεπιπέδου, τότε $\xi_i\geq 1$. Το χαλαρό λάθος (*soft error*) είναι το άθροισμα των ξ_i, των παραδειγμάτων του συνόλου εκπαίδευσης, δηλαδή:

$$\sum_i \xi_i \tag{2.16}$$

Για την εύρεση του υπερεπιπέδου χαλαρού εύρους, απαιτείται η επίλυση του εξής προβλήματος βελτιστοποίησης με περιορισμούς (*constant optimization problem*):

$$\min(\frac{\| a \|^2}{2} + C\sum_{i=1}^{l} \xi_i), \tag{2.17}$$

με τους περιορισμούς ότι $\xi_i \geq 0$ *και* $y_i(a \bullet x_i + b) \geq 1 - \xi_i, \forall i = 1,2,...,l$

Δηλαδή απαιτείται ο προσδιορισμός των παραμέτρων ξ_i και της μεταβλητής *C*, η οποία καθορίζει το πόσο αυστηροί είμαστε στο να επιτρέπονται τα λάθη.

Το παραπάνω πρόβλημα βελτιστοποίησης λύνεται και αυτό με τη βοήθεια των πολλαπλασιαστών *Langrange* $\lambda_i \geq 0,\ i=1,2,...,l$ και η επιφάνεια απόφασης που προκύπτει είναι:

$$(\sum_{i=1}^{l} \lambda_i y_i x_i \bullet x) + b = 0 \tag{2.18}$$

Στην περίπτωση που τα δεδομένα του συνόλου εκπαίδευσης είναι μη γραμμικά διαχωρίσιμα, τότε αυτά απεικονίζονται πρώτα σε ένα χώρο μεγαλύτερης διάστασης με μια συνάρτηση *Φ(x)* και στη συνέχεια σε αυτόν τον χώρο γίνεται προσπάθεια για γραμμικό διαχωρισμό.

Σε αυτήν την περίπτωση η γραμμική επιφάνεια απόφασης είναι της μορφής:

$$\alpha * \Phi(x) + b = 0 \tag{2.19}$$

Για την εύρεση της διαχωριστικής επιφάνειας απαιτείται η επίλυση του εξής προβλήματος ελαχιστοποίησης με περιορισμούς:

$$\min(\frac{\| a \|^2}{2} + C\sum_{i=1}^{l}\xi_i), \tag{2.20}$$

με τους περιορισμούς ότι $\xi_i \geq 0$ *και* $y_i(a \bullet \Phi(\mathrm{x}_i) + \mathrm{b}) \geq 1 - \xi_i, \forall i$

, το οποίο καταλήγει στο εξής δυϊκό πρόβλημα (*dual problem*):

$$L_l = \sum_{i=1}^{l}\lambda_i - \frac{1}{2}\sum_{i,j}^{l}\lambda_i\lambda_j y_i y_j \Phi(\mathrm{x}_i) \bullet \Phi(\mathrm{x}_j), \tag{2.21}$$

με τους περιορισμούς ότι $\xi_i \geq 0$ *και* $\sum_{i=1}^{l}\lambda_i y_i = 0$

Μετά τον προσδιορισμό των λ_i, χρησιμοποιούνται οι *KKT* συνθήκες για να μετατραπούν οι ανισότητες σε ισότητες και εν συνεχεία υπολογίζονται οι παράγωγοι ως προς *α, b* και τίθενται ίσα με μηδέν. Για το παραπάνω πρόβλημα, οι συνθήκες *KKT* μπορούν να αποδοθούν ως εξής:

$$\frac{\partial}{\partial a} L = \alpha - \sum_i \lambda_i y_i \Phi(\mathrm{x}_i), \quad v = 1, ..., d \tag{2.22}$$

$$\frac{\partial}{\partial b} L = -\sum_i \lambda_i y_i = 0 \tag{2.23}$$

$$y_i(\Phi(\mathrm{x}_i) \bullet \mathrm{a} + \mathrm{b}) - 1 \geq 0, \quad \mathrm{i} = 1, ..., \mathrm{l} \tag{2.24}$$

$$\lambda_i \geq 0, \ \forall i \tag{2.25}$$

$$\lambda_i(y(a \bullet \Phi(\mathrm{x}_i) + b) - 1) = 0, \ \forall i \tag{2.26}$$

Έτσι προκύπτουν οι ακόλουθες δύο σχέσεις:

$$a = \sum_{i=1}^{l} \lambda_i y_i \Phi(\mathrm{x}_i) \tag{2.27}$$

$$\lambda_i(\mathrm{y}_i(\alpha \bullet \Phi(\mathrm{x}_i) + \mathrm{b}) - 1 + \xi_i) = 0 \Rightarrow \lambda_i(\mathrm{y}_i(\sum_{j=1}^{l} \lambda_i \, \mathrm{y}_i \Phi(\mathrm{x}_j) \bullet \Phi(\mathrm{x}_i) + \mathrm{b} - 1 + \xi_i) = 0 \tag{2.28}$$

Η επιφάνεια απόφασης που προκύπτει έχει την παρακάτω μορφή:

$$(\sum_{i=1}^{l} \lambda_i y_i \Phi(\mathrm{x}_i) \bullet \Phi(\mathrm{x})) + \mathrm{b} = 0 \tag{2.29}$$

Το εσωτερικό γινόμενο στο χώρο *Φ(x)* ονομάζεται συνάρτηση πυρήνα (*kernel function*).

Η συνάρτηση πυρήνα που χρησιμοποιήθηκε για τη δημιουργία των ανιχνευτών σημασιολογικών εννοιών *vireo374* και *web81* ονομάζεται *Chi-square*. Ο πυρήνας *Chi-square*, [18], είναι ένας πυρήνας απλός και αποτελεσματικός σε περιπτώσεις σύγκρισης ιστογραμμάτων και είναι στη ουσία ο γενικευμένος *RBF* πυρήνας. Επίσης, έχει αποδειχθεί ότι είναι ανθεκτικός σε περιπτώσεις διαχωρισμού εικόνων που στηρίζονται στο χρώμα και την υφή. Ο πυρήνας *Chi-square* περιγράφεται από την σχέση:

$$k(x, y) = \sum_{i} \frac{(x_i - y_i)^2}{x_i + y_i}, \tag{2.30}$$

x_i και y_i είναι οι συνιστώσες των x και y αντίστοιχα.

Σε αυτήν τη παράγραφο εξετάζεται η περίπτωση του προβλήματος ταξινόμησης με πολλές (περισσότερες από δυο) κατηγορίες. Σε αυτήν την περίπτωση, υπάρχουν δυο προσεγγίσεις, ένας-εναντίον-όλων και ένας-εναντίον-ένα.

Στην προσέγγιση «ένας-εναντίον-όλων», δεδομένου ενός συνόλου *M* προκαθορισμένων κατηγοριών $C=\{C_1,C_2,\ldots,C_M\}$, για το διαχωρισμό των παραδειγμάτων απαιτείται εκπαίδευση *M* δυαδικών *SVM*, με τρόπο ώστε κάθε *SVM* να βρίσκει μια επιφάνεια απόφασης, που διαχωρίζει τα παραδείγματα της κατηγορίας *i* από τα παραδείγματα των υπόλοιπων *M-1* κατηγοριών. Ουσιαστικά, το πρόβλημα

των πολλών κατηγοριών διαιρείται σε *M* δυαδικά προβλήματα όπου για την εύρεση του υπερεπιπέδου απόφασης κάθε δυαδικού *SVM* χρησιμοποιεί τους γνωστούς αλγορίθμους εκπαίδευσης.

Η απόφαση για την κατηγορία του προτύπου προκύπτει με ένα είδος «ψηφοφορίας»: υπολογίζεται η έξοδος για κάθε δυαδικό *SVM* και το πρότυπο ταξινομείται στην κατηγορία C_j, αν η έξοδος του ταξινομητή *j* είναι μεγαλύτερη από τις εξόδους των υπόλοιπων ταξινομητών.

Στην περίπτωση της προσέγγισης «ένας-εναντίον-ένα», κατασκευάζεται ένας ταξινομητής *SVM* για κάθε ζεύγος κατηγοριών. Δεδομένου ενός συνόλου *M* προκαθορισμένων κατηγοριών $C=\{C_1,C_2,...C_M\}$ κατασκευάζονται $M(M-1)/2$ *SVM* ταξινομητές και κάθε ταξινομητής διακρίνει τα παραδείγματα της μιας κατηγορίας από τα παραδείγματα κάθε άλλης κατηγορίας. Για τη λήψη της απόφασης ισχύει το σύστημα της ψηφοφορίας, όπου κάθε *SVM* ψηφίζει για μια κατηγορία.

2.6. Εξαγωγή Σημασιολογικών Χαρακτηριστικών

Για να εξάγουμε τα σημασιολογικά χαρακτηριστικά στην παρούσα εργασία αρχικά χρησιμοποιήσαμε 10 βίντεο, τα οποία προήλθαν από τα δεδομένα ελέγχου του TRECVID 2008 [16] στο διαγωνισμό Rushes Summarization του TRECVID 2008. Τα χαρακτηριστικά εικονοπλαίσια εξάγονται σύμφωνα τη μέθοδο που προτείνεται στο [3]. Αρχικά, για κάθε εικονοπλαίσιο Kf_i, $i=1,...,N$, εξάγεται ένα σύνολο από *SIFT* περιγραφείς D_i. Έπειτα, υπολογίζεται το ιστόγραμμα οπτικών λέξεων για κάθε εικονοπλαίσιο με τη μεθοδολογία που αναφέρθηκε στο Κεφάλαιο 2.2. Κάθε ένας από τους αρχικούς περιγραφείς κάθε εικονοπλαισίου *i*, ανατίθεται σε μία από τις 500 οπτικές λέξεις $\{C_1, C_2, ..., C_{500}\}$ οι οποίες υπολογίστηκαν στα πλαίσια των ανιχνευτών σημασιολογικών εννοιών vireo-374 και είναι διαθέσιμες στο [24]. Αυτό έχει ως αποτέλεσμα την δημιουργία ενός διανύσματος που περιέχει την συχνότητα εμφάνισης της κάθε οπτικής λέξης μέσα σε ένα εικονοπλαίσιο, δηλαδή δημιουργείται το ιστόγραμμα οπτικών λέξεων του κάθε εικονοπλαισίου. Επιπλέον για τον υπολογισμό του ιστογράμματος οπτικών λέξεων χρησιμοποιήθηκε και η τεχνική

soft-weighting όπως περιγράφεται στην εξίσωση 2.11. Άρα για κάθε εικονοπλαίσιο το ιστόγραμμα οπτικών λέξεων χρησιμοποιώντας την τεχνική *soft-weighting* είναι $HSBOW = T$, όπου $T=[t_1,...,t_k,...,t_K]$, με κάθε t_k να αναπαριστά το βάρος της οπτικής λέξης k στο εικονοπλαίσιο, σύμφωνα με την εξίσωση 2.11.

Επομένως για κάθε εικονοπλαίσιο έχουμε ένα διάνυσμα οπτικών λέξεων με διάσταση 500. Στη συνέχεια, θέλουμε να εξάγουμε τα σημασιολογικά χαρακτηριστικά για κάθε διάνυσμα. Τα σημασιολογικά χαρακτηριστικά προκύπτουν από την εφαρμογή των ανιχνευτών σημασιολογικών εννοιών στα ιστογράμματα οπτικών λέξεων των εικονοπλαισίων που μας ενδιαφέρουν. Τα σημασιολογικά χαρακτηριστικά στην ουσία είναι διανύσματα διάστασης όσα και το πλήθος των διαθέσιμων εννοιών και εκφράζουν την πιθανότητα ύπαρξης της κάθε σημασιολογικής έννοιας στο εικονοπλαίσιο. Έστω $HSBoW^j$ είναι το διάνυσμα οπτικών λέξεων του j εικονοπλαισίου και m η σημασιολογική έννοια που μελετάμε. Το διάνυσμα σημασιολογικών χαρακτηριστικών S^j για το j εικονοπλαίσιο προκύπτει από τη σχέση:

$$S^j(m) = P(m \mid HSBoW^j) \tag{2.31}$$

Όπου $P(m \mid HSBoW^j)$ είναι η πιθανότητα ύπαρξης της m έννοιας όταν το ιστόγραμμα οπτικών λέξεων της $HSBoW^j$ εφαρμόζεται ως είσοδο στον αντίστοιχο ανιχνευτή της σημασιολογικής έννοιας m. Το $m=1,...,81$ όταν χρησιμοποιούμε τους ανιχνευτές της βάσης web-81 σημασιολογικών εννοιών για την πρόβλεψη, ενώ όταν χρησιμοποιούμε τους ανιχνευτές της βάσης vireo-374 σημασιολογικών εννοιών το είναι $m=1,...,374$.

Το διάνυσμα των πιθανοτήτων ύπαρξης των εννοιών για κάθε εικονοπλαίσιο αποτελεί το διάνυσμα σημασιολογικών χαρακτηριστικών του εικονοπλαισίου, το οποίο έχει διάσταση 81 και 374, όταν χρησιμοποιούνται οι βάσεις web-81 και vireo-374, αντίστοιχα. Αυτά τα σημασιολογικά χαρακτηριστικά επεξεργαζόμαστε για να μπορέσουμε να πετύχουμε μια αξιόπιστη ομαδοποίηση των πλάνων και εν συνεχεία μια επιτυχημένη περίληψη του βίντεο. Στο εφεξής θα αναφερόμαστε σε αυτά τα διανύσματα σημασιολογικών χαρακτηριστικών ως S^{81} και S^{374}.

ΚΕΦΑΛΑΙΟ 3. ΟΜΑΔΟΠΟΙΗΣΗ ΟΜΟΙΩΝ ΠΛΑΝΩΝ

Σε αυτό το κεφάλαιο περιγράφουμε την προτεινόμενη μέθοδο βάσει της οποίας επιτυγχάνεται η ομαδοποίηση όμοιων πλάνων σε αμοντάριστο βίντεο. Αρχικά γίνεται μια γενική περιγραφή του αλγορίθμου και στη συνέχεια παρουσιάζεται ο τρόπος περιγραφής των πλάνων, ο οποίος έγινε θεωρώντας 12 διαφορετικές αναπαραστάσεις. Στη συνέχεια παρουσιάζεται η μεθοδολογία σύγκρισης και αξιολόγησης.

3.1. Γενικός Αλγόριθμος Ομαδοποίησης μέσω Κατάτμησης

Στην παρούσα εργασία, για να πετύχουμε την ομαδοποίηση όμοιων πλάνων, συγκρίνουμε δυο διαδοχικά πλάνα και θέτουμε ως μέτρο ομοιότητας μεταξύ τους την ελάχιστη απόσταση μεταξύ όλων των πιθανών συνδυασμών των χαρακτηριστικών εικονοπλαισίων που τα περιγράφουν. Συγκεκριμένα, αν ο αριθμός των πλάνων ενός βίντεο είναι *P*, συγκρίνουμε το $1^{ο}$ πλάνο με το $2^{ο}$ πλάνο, το $2^{ο}$ πλάνο με το $3^{ο}$ πλάνο , …, το *P-1* πλάνο με το *P* πλάνο. Με αυτό τον τρόπο προκύπτει ένα διάνυσμα

αποστάσεων μεταξύ διαδοχικών πλάνων. Στη συνέχεια θέτουμε ένα κατώφλι, και ως όρια αλλαγής ομάδας όμοιων πλάνων ορίζουμε εκείνες τις θέσεις στις οποίες αντιστοιχεί απόσταση μεγαλύτερη από το κατώφλι. Έπειτα για να γίνει η αξιολόγηση της ομαδοποίησης συγκρίνουμε τις θέσεις που εντοπίστηκαν από τη μέθοδό μας ότι γίνεται αλλαγή ομάδας πλάνων με τις θέσεις που πραγματικά αλλάζει.

3.2. Περιγραφή Πλάνου

Κάθε πλάνο περιγράφεται με ένα συγκεκριμένο αριθμό χαρακτηριστικών εικονοπλαισίων, τα οποία εξάγονται με τη μέθοδο που προτείνεται στο [3]. Εκτός από τα χαρακτηριστικά εικονοπλαίσια, επιλέγουμε σε ορισμένες περιπτώσεις και την αναπαράσταση του κάθε πλάνου με επιπλέον εικονοπλαίσια που είναι γειτονικά των χαρακτηριστικών εικονοπλαισίων. Η λογική πίσω από αυτήν την επιλογή είναι ότι μπορεί το εξαγόμενο χαρακτηριστικό εικονοπλαίσιο να μην αποδίδει στο μέγιστο τις έννοιες που πραγματικά υπάρχουν μέσα στο εικονοπλαίσιο λόγω εσφαλμένης εξαγωγής των χαρακτηριστικών SIFT. Με τη χρήση της γειτονιάς ενός χαρακτηριστικού εικονοπλαισίου στοχεύουμε στην πιο αξιόπιστη απόδοση ύπαρξης των εννοιών σε κάποιο χαρακτηριστικό εικονοπλαίσιο εφόσον αυτές υπάρχουν για το σύνολο της γειτονιάς του. Με άλλα λόγια εφόσον μία έννοια παρουσιάζεται σε όλα ή στα περισσότερα γειτονικά εικονοπλαίσια ενός χαρακτηριστικού εικονοπλαισίου, τότε με πολύ μεγάλη πιθανότητα η συγκεκριμένη έννοια όντως υπάρχει, και σωστά αποδίδεται στα σημασιολογικά χαρακτηριστικά του χαρακτηριστικού εικονοπλαισίου.

Πιο συγκεκριμένα, γύρω από ένα χαρακτηριστικό εικονοπλαίσιο Kf_i επιλέγεται ένα σύνολο εικονοπλαισίων $F_{Kfi} = \{f_{i-3d}, f_{i-2d}, f_{i-d}, Kf_i, f_{i+d}, f_{i+2d}, f_{i+3d}\}$ που αποτελεί τη γειτονιά του. Για το βήμα d, την απόσταση δηλαδή κάθε εικονοπλαισίου από το προηγούμενο και το επόμενο του στη γειτονιά δοκιμάσαμε τρεις τιμές, d=1,3,5. Επιλέγουμε N=3,5,7, γειτονικά εικονοπλαίσια για d=1 και N=7 για d=3,5 με $\left\lfloor \frac{N}{2} \right\rfloor$ εικονοπλαίσια σε κάθε πλευρά του χαρακτηριστικού εικονοπλαισίου.

Στη συνέχεια περιγράφουμε το πλάνο με 12 διαφορετικές αναπαραστάσεις. Επεξεργαζόμαστε δηλαδή με 12 διαφορετικούς τρόπους τα σημασιολογικά χαρακτηριστικά των εικονοπλαισίων όπως αυτά εξάγονται σύμφωνα με το κεφάλαιο 2. Για κάθε μια αναπαράσταση θεωρούμε ότι κάθε πλάνο έχει *K* χαρακτηριστικά εικονοπλαίσια και τα διανύσματα σημασιολογικών χαρακτηριστικών των χαρακτηριστικών εικονοπλαισίων συμβολίζονται ως $\{S_{KF_1},\ldots,S_{KF_K}\}$. Το πλήθος της γειτονιάς ενός χαρακτηριστικού εικονοπλαισίου KF_i (μαζί με το χαρακτηριστικό εικονοπλαίσιο) ορίζεται ως N_{KFi}. Τα διανύσματα σημασιολογικών χαρακτηριστικών της γειτονιάς ενός χαρακτηριστικού εικονοπλαισίου KF_i ορίζονται ως $\{S_1,\ldots,S_{N_{KFi}}\}$. Κάθε εικονοπλαίσιο της γειτονιάς αναπαρίσταται από το αντίστοιχο σημασιολογικό του χαρακτηριστικό S^{81} ή S^{374}, ανάλογα με τη βάση ανιχνευτών που έχει χρησιμοποιηθεί. Για λόγους ευκολίας, όταν αναφερόμαστε στο διάνυσμα σημασιολογικών χαρακτηριστικών *S* ενός εικονοπλαισίου, θα εννοούμε είτε το S^{81} είτε το S^{374}.

1η Αναπαράσταση: Για κάθε χαρακτηριστικό εικονοπλαίσιο KF_i υπολογίζουμε ένα διάνυσμα χαρακτηριστικών που περιέχει το μέσο όρο των διανυσμάτων των σημασιολογικών χαρακτηριστικών της γειτονιάς του F_{Kfi}.

$$R^1_{KFi}(m) = \frac{\sum_{j=1}^{N_{KFi}} S_j(m)}{N_{KFi}} \tag{3.1}$$

Όπου $j = 1, \ldots, N_{KFi}$ και $m=1,\ldots,81$ όταν χρησιμοποιούμε τους ανιχνευτές της βάσης *web-81* σημασιολογικών εννοιών για την πρόβλεψη, ενώ όταν χρησιμοποιούμε τους ανιχνευτές της βάσης *vireo-374* σημασιολογικών εννοιών ισχύει $m=1,\ldots,374$.

2η Αναπαράσταση: Κάθε χαρακτηριστικό εικονοπλαίσιο KF_i αναπαρίσταται από το αντίστοιχο διάνυσμα σημασιολογικών χαρακτηριστικών του.

$$R^2_{KFi}(m) = S_{KFi}(m) \tag{3.2}$$

3η Αναπαράσταση: Για κάθε χαρακτηριστικό εικονοπλαίσιο KF_i κατασκευάζουμε ένα διάνυσμα το οποίο παίρνει την τιμή 1 σε κάθε θέση/ έννοια όταν τουλάχιστον ένα από τα διανύσματα σημασιολογικών χαρακτηριστικών της γειτονιάς του έχει τιμή μεγαλύτερη του 0.5 στην αντίστοιχη θέση.

Άρα:

$R^3_{KFi}(m) = 1$, εάν υπάρχει έστω και ένα $S_j(m) > 0.5$, j=1,…,N_{KFi} (3.3)

$R^3_{KFi}(m) = 0$, διαφορετικά.

4η Αναπαράσταση: Για κάθε χαρακτηριστικό εικονοπλαίσιο δημιουργείται ένα διάνυσμα που περιέχει το κανονικοποιημένο πλήθος των σημασιολογικών χαρακτηριστικών της γειτονιάς του που σε κάθε θέση/ έννοια έχουν τιμή μεγαλύτερη του 0.5.

$C_i^j(m) = 1$, για τα j όπου $S_j(m) > 0.5$, j=1,…,N_{KFi}

$C_i^j(m) = 0$, διαφορετικά.

$$R^4_{KFi}(m) = \frac{\sum_{j=1}^{N_{KFi}} C_i{}^j(m)}{N_{KFi}} \quad (3.4)$$

5η Αναπαράσταση: Για κάθε χαρακτηριστικό εικονοπλαίσιο δημιουργείται ένα διάνυσμα που περιέχει το μέσο όρο των σημασιολογικών της γειτονιάς του όταν η τιμή τους σε κάθε θέση/έννοια είναι μεγαλύτερη του 0.5.

$$R^5_{KFi}(m) = \frac{\sum_{j=1}^{N_{KFi}} S_j(m)}{N_{KFi}}, \text{για τα } j \text{ όπου } S_j(m) > 0.5 \quad (3.5)$$

6η Αναπαράσταση: Σε αντίθεση με τις προηγούμενες αναπαραστάσεις, κάθε πλάνο αναπαρίσταται από ένα και μόνο διάνυσμα που προκύπτει από τον μέσο όρο των διανυσμάτων σημασιολογικών χαρακτηριστικών μόνο των χαρακτηριστικών εικονοπλαισίων του.

$$R^6(m) = \frac{\sum_{i=1}^{K} S_{KFi}(m)}{K} \tag{3.6}$$

7η Αναπαράσταση: Για κάθε πλάνο υπολογίζεται ο μέσος όρος των διανυσμάτων των χαρακτηριστικών εικονοπλαισίων του που προκύπτουν από την αναπαράσταση R^1.

$$R^7(m) = \frac{\sum_{i=1}^{K} R^1_{KFi}(m)}{K} \tag{3.7}$$

8η Αναπαράσταση: Για κάθε πλάνο υπολογίζεται ο μέσος όρος των διανυσμάτων των χαρακτηριστικών εικονοπλαισίων του που προκύπτουν από την αναπαράσταση R^5.

$$R^8(m) = \frac{\sum_{i=1}^{K} R^5_{KFi}(m)}{K} \tag{3.8}$$

9η Αναπαράσταση: Για κάθε χαρακτηριστικό εικονοπλαίσιο δημιουργείται ένα διάνυσμα χαρακτηριστικών το οποίο παίρνει την τιμή 1 όταν τα σημασιολογικά χαρακτηριστικά του χαρακτηριστικού εικονοπλαισίου έχουν τιμή μεγαλύτερη του 0.5.

$$R^9_{KFi}(m) = 1, \text{ εάν } S_{KF_i}(m) > 0.5 \tag{3.9}$$

$R^9_{KFi}(m) = 0$, διαφορετικά.

10η Αναπαράσταση: Για κάθε χαρακτηριστικό εικονοπλαίσιο δημιουργείται ένα διάνυσμα χαρακτηριστικών το οποίο παίρνει την τιμή 1 όταν ο μέσος όρος των διανυσμάτων σημασιολογικών χαρακτηριστικών της γειτονιάς του είναι μεγαλύτερος από 0.5.

$$R_{KFi}^{10}(m) = 1, \text{ εάν } R_{KFi}^{1}(m) > 0.5 \quad (3.10)$$

$R_{KFi}^{10}(m) = 0$, διαφορετικά.

11η Αναπαράσταση: Για κάθε πλάνο κατασκευάζεται ένα διάνυσμα που παίρνει την τιμή 1 όταν ο μέσος όρος των σημασιολογικών χαρακτηριστικών των χαρακτηριστικών εικονοπλαισίων του πλάνου είναι μεγαλύτερος από το 0.5.

$$R^{12}(m) = 1, \text{ εάν } R^{6}(m) > 0.5 \quad (3.11)$$

$R^{12}(m) = 0$, διαφορετικά.

12η Αναπαράσταση: Για κάθε πλάνο δημιουργείται ένα διάνυσμα το οποίο παίρνει την τιμή 1 στα σημεία όπου ο μέσος όρος των μέσων όρων των σημασιολογικών χαρακτηριστικών των χαρακτηριστικών εικονοπλαισίων και της γειτονιάς τους είναι μεγαλύτερος από 0.5.

$$R^{11}(m) = 1 \text{ , εάν } R^{7}(m) > 0.5 \quad (3.12)$$

$R^{11}(m) = 0$, διαφορετικά.

3.3. Σύγκριση Πλάνων

Σε αυτό το σημείο θα εξηγήσουμε τον τρόπο σύγκρισης των διαφορετικών αναπαραστάσεων των πλάνων ώστε να επιλέξουμε αυτό που οδηγεί σε καλύτερα αποτελέσματα. Κάθε πλάνο συγκρίνεται με το επόμενό του. Συγκεκριμένα, εξετάζουμε όλους τους πιθανούς συνδυασμούς των διανυσμάτων χαρακτηριστικών που τα περιγράφουν. Όπως είδαμε προηγουμένως, κάθε πλάνο μπορεί να αναπαρίσταται είτε από ένα μόνο διάνυσμα χαρακτηριστικών (αναπαραστάσεις R^6, R^7, R^8, R^{11}, R^{12}) είτε από τα διανύσματα χαρακτηριστικών των χαρακτηριστικών εικονοπλαισίων του. Έστω $S_1 = \{F_1^1, \ldots, F_{N_1}^1\}$ και $S_2 = \{F_1^2, \ldots, F_{N_2}^2\}$ δυο πλάνα όπου F_i^j το *i* διάνυσμα χαρακτηριστικών *j* πλάνου. Ως απόσταση δύο πλάνων ορίζεται η μικρότερη απόσταση όλων των δυνατών συνδυασμών (ζεύγη) των επιμέρους διανυσμάτων χαρακτηριστικών που τα αντιπροσωπεύουν.

$$D(S_1, S_2) = min\left(dist\left(F_i^1, F_j^2\right)\right) \mu\varepsilon \quad i = 1, \ldots, N_1 \; \kappa\alpha\iota \; j = 1, \ldots, N_2 \tag{3.13}$$

$$dist(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \tag{3.14}$$

Το dist είναι η ευκλείδεια απόσταση η οποία υπολογίζει το τετράγωνο της διαφοράς των διανυσμάτων χαρακτηριστικών.

Πρέπει να αναφέρουμε επίσης ότι συγκρίναμε δυο πλάνα μεταξύ τους με 4 διαφορετικούς τρόπους.

$1^η$ περίπτωση: Συγκρίνουμε τα πλάνα ξεχωριστά για τα S^{81} και S^{374} διανύσματα σημασιολογικών χαρακτηριστικών.

$2^η$ περίπτωση: Συγκρίνουμε τα πλάνα κάνοντας συνένωση των S^{81} και S^{374} διανυσμάτων σημασιολογικών χαρακτηριστικών, δηλαδή δημιουργήσαμε για κάθε εικονοπλαίσιο ένα διάνυσμα χαρακτηριστικών με 81+374=455 σημασιολογικά χαρακτηριστικά.

$3^{η}$ περίπτωση: Συνδυάζουμε τις αποστάσεις, που υπολογίσαμε από τη σύγκριση των πλάνων από τα S^{81} σημασιολογικά χαρακτηριστικά με τις αποστάσεις από τα S^{374} σημασιολογικά χαρακτηριστικά. Αν *dist81* είναι οι αποστάσεις που προέκυψαν από τα S^{81} σημασιολογικά χαρακτηριστικά και *dist374* οι αποστάσεις από τα S^{374} σημασιολογικά χαρακτηριστικά τότε η συνολική απόσταση δυο πλάνων προκύπτει ως εξής:

$$Distance = a * dist81 + (1-a) * dist374, \text{όπου } a = 0.9, 0.8, \ldots, 0.1 \qquad (3.15)$$

$4^{η}$ περίπτωση: Αρχικά, υπολογίζουμε το μέσο όρο όλων των διανυσματικών σημασιολογικών χαρακτηριστικών όλων των πλάνων σε όλα τα βίντεο που εξετάσαμε χρησιμοποιώντας την αναπαράσταση R^8. Έπειτα, η σύγκριση μεταξύ των πλάνων έγινε μόνο για τις έννοιες όπου ο μέσος όρος εμφάνισης τους σε όλο το σύνολο των βίντεο ήταν μεγαλύτερος από ένα προκαθορισμένο κατώφλι.

3.4. Αξιολόγηση Ομαδοποίησης

Για να αξιολογήσουμε την απόδοση της μεθόδου μας χρησιμοποιήθηκαν τα ακόλουθα κριτήρια:

$$\mathrm{Re}\,call = \frac{N_c}{N_c + N_m} \quad , \qquad (3.16)$$

$$\mathrm{Pr}\,ecision = \frac{N_c}{N_c + N_f} \quad , \qquad (3.17)$$

$$F_1 = \frac{2 \times \mathrm{Re}\,call \times \mathrm{Pr}\,ecision}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision} \quad , \qquad (3.18)$$

όπου N_c είναι ο αριθμός των σωστά εντοπιζόμενων ορίων αλλαγής ομάδας όμοιων πλάνων, N_m είναι ο αριθμός των ορίων που δεν εντοπίστηκαν και N_f είναι ο αριθμός των λανθασμένων εντοπισμών.

ΚΕΦΑΛΑΙΟ 4. ΒΕΛΤΙΩΣΗ ΜΕ ΕΝΤΟΠΙΣΜΟ ΠΡΟΣΩΠΟΥ ΚΑΙ ΣΩΜΑΤΟΣ

Σε αυτό το κεφάλαιο αναφέρουμε τον τρόπο αξιοποίησης του εντοπισμού προσώπου για τη βελτίωση της απόδοσης της μεθόδου μας, ενώ αναλύεται ο αλγόριθμος των *Viola & Jones* [28,29] που χρησιμοποιείται για την ανίχνευση προσώπου και σώματος.

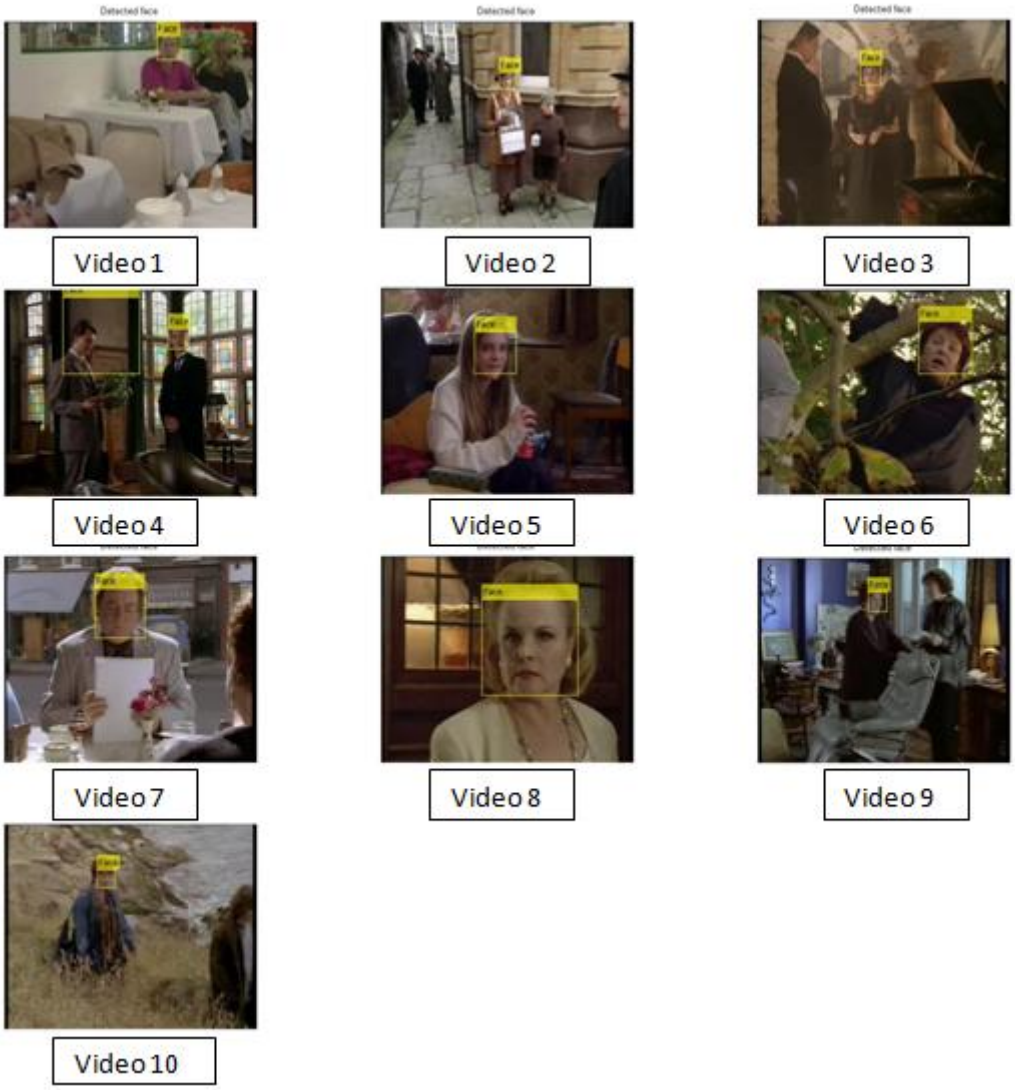
4.1. Βελτίωση με Εντοπισμό Προσώπου και Σώματος

Έχοντας εξάγει τα χρονικά σημεία στα οποία αλλάζει η ομάδα όμοιων πλάνων με τη μέθοδο που περιγράφουμε στο προηγούμενο κεφάλαιο, προσπαθήσαμε να βελτιώσουμε την απόδοση της μεθόδου χρησιμοποιώντας μια μέθοδο εντοπισμού προσώπου και σώματος. Συγκεκριμένα, η μέθοδος εντοπισμού προσώπου και σώματος μας επιστρέφει τις συντεταγμένες μιας περιοχής της εικόνας που αντιστοιχεί στο πρόσωπο ή το σώμα που εντοπίστηκε. Αφού γίνει η εξαγωγή του ιστογράμματος χρώματος των παραπάνω περιοχών, συγκρίνουμε διαδοχικά πλάνα με βάση τα παραπάνω ιστογράμματα. Έστω ότι τα ιστογράμματα συμβολίζονται με Η και N_1,N_2 είναι το πλήθος των χαρακτηριστικών εικονοπλαισίων κάθε πλάνου.

$$D(S_1, S_2) = min\left(dist\left(H_i^1 H_j^2\right)\right) \mu\varepsilon \quad i = 1,\ldots,N_1 \kappa\alpha\iota\ j = 1,\ldots,N_2 \qquad (4.1)$$

Αφού πάρουμε τη μικρότερη απόσταση όλων των δυνατών συνδυασμών (ζεύγη) των επιμέρους διανυσμάτων χαρακτηριστικών που αντιπροσωπεύουν τα πλάνα, θέτουμε ένα κατώφλι και διατηρούμε μόνο τις θέσεις όπου οι αποστάσεις είναι μικρότερες από αυτό το κατώφλι. Δοκιμάζουμε 51 διαφορετικές τιμές για το κατώφλι από 0 έως 0.05 με βήμα 0.01. Για κάθε ένα κατώφλι λαμβάνουμε θέσεις οι οποίες δείχνουν τα σημεία όπου ο αλγόριθμος εντοπισμού προσώπου και σώματος εντοπίζει μεγάλη αλλαγή ομάδων με όμοια πλάνα. Αφαιρούμε αυτές τις θέσεις από τις θέσεις που εμείς εντοπίσαμε ότι γίνεται εναλλαγή ομάδας όμοιων πλάνων. Εφόσον, δύο πλάνα δε διαφέρουν ως προς αυτά τα ιστογράμματα, θεωρούνται όμοια και δεν θα έπρεπε μεταξύ τους να υπάρχει αλλαγή της ομάδας όμοιων πλάνων. Έτσι, απομακρύνουμε τυχόν λανθασμένους εντοπισμούς θέσεων αλλαγής ομάδων όμοιων πλάνων που εξήγαγε η μέθοδος μας. Έπειτα, συγκρίνουμε τις καινούριες θέσεις με τις πραγματικές θέσεις (*ground truth*) και παίρνουμε το ποσοστό επιτυχίας με βάση τις εξισώσεις 3.15, 3.16 και 3.17.

Μερικά από τα πλαίσια που μας επιστρέφει ο ανιχνευτής προσώπου και σώματος δίνονται στις εικόνες 4.1 και 4.2 αντίστοιχα.



Εικόνα 4.1 Παραδείγματα εντοπισμού προσώπου από τα 10 βίντεο.

Εικόνα 4.2 Παραδείγματα εντοπισμού σώματος από τα 10 βίντεο.

Για να βελτιωθεί η μέθοδός μας αφαιρέσαμε από τις αρχικές θέσεις εναλλαγής μιας ομάδας όμοιων πλάνων, τις θέσεις που η μέθοδος εντοπισμού προσώπου και σώματος βρίσκει ότι τα πλάνα είναι όμοια, δηλαδή τα ιστογράμματα χρώματος του πλαισίου που έχουν μικρή Ευκλείδεια απόσταση. Με αυτό τον τρόπο απομακρύνουμε τις θέσεις όπου γίνεται εσφαλμένη ανίχνευση εναλλαγής μιας ομάδας όμοιων πλάνων. Ένα παράδειγμα της βελτίωσης φαίνεται στην εικόνα 4.3.

Εικόνα 4.3 Εσφαλμένος εντοπισμός αλλαγής πλάνου.

Σύμφωνα με τη δική μας μέθοδο χωρίς τη βελτίωση, η εναλλαγή μιας ομάδας όμοιων πλάνων γίνεται σε αυτά τα πλάνα ενώ με τη βελτίωση ο εντοπισμός προσώπου εντοπίζει το ίδιο πρόσωπο επομένως θεωρεί τη συγκεκριμένη θέση όπου εντοπίστηκε αρχικά εναλλαγή ομάδας όμοιων πλάνων ως εσφαλμένη και άρα την απομακρύνει.

Για να επιτευχθεί όμως αυτή η βελτίωση πρέπει να χρησιμοποιηθεί ο κατάλληλος ανιχνευτής προσώπου ο οποίος θα είναι ικανός να ανακαλύπτει και να εντοπίζει όλα τα πρόσωπα που υπάρχουν στην εικόνα ανεξάρτητα από τις συνθήκες φωτισμού, τη θέση, την κλίμακα, τον προσανατολισμό και τους μορφασμούς των προσώπων. Την καλύτερη απόδοση ως προς την αποτελεσματικότητα και την ταχύτητα μέχρι σήμερα εμφανίζεται να έχει η μέθοδος των *Viola & Jones* [28,29] που στηρίζεται σε εκπαίδευση με την τεχνική *AdaBoost* και την χρήση χαρακτηριστικών γνωρισμάτων της μορφής *Haar*. Για τον εντοπισμό σώματος χρησιμοποιείται ο ίδιος

αλγόριθμος με τον εντοπισμό προσώπου αλλάζοντας τα οπτικά χαρακτηριστικά που ανιχνεύονται τα οποία πρέπει να αντιπροσωπεύουν μέρη του σώματος ώστε να δημιουργηθούν αποτελεσματικοί ταξινομητές.

4.2. Ανίχνευση Προσώπου κατά Viola & Jones

Οι ιδέες-κλειδιά της μεθόδου αυτής είναι οι εξής [28,29]:

- Χρήση απλών χαρακτηριστικών τύπου *Haar* για απόκτηση γνώσης από δεδομένα μάθησης, που αποτελούνται από εικόνες προσώπων και μη-προσώπων.
- Χρήση μιας νέας εικόνας αναπαράστασης των προσώπων που ονομάζεται Εικόνα Ολοκλήρωμα (*Integral Image*), και που επιτρέπει τα χαρακτηριστικά που χρησιμοποιούνται από τον ανιχνευτή να υπολογίζονται πολύ γρήγορα.
- Χρήση ενός αλγόριθμου μάθησης, που στηρίζεται στον *AdaBoost*, ο οποίος επιλέγει ένα μικρό αριθμό από κρίσιμα οπτικά χαρακτηριστικά και αποδίδει άκρως αποτελεσματικούς ταξινομητές.
- Συνδυασμός των ταξινομητών σε ακολουθιακή διάταξη (*cascade*), που επιτρέπει περιοχές υποβάθρου της εικόνας να απορρίπτονται γρήγορα, αναλώνοντας περισσότερο υπολογιστικό χρόνο σε περιοχές που μοιάζουν περισσότερο σε πρόσωπα.

4.2.1. Υπολογισμός Χαρακτηριστικών

Για τον υπολογισμό χαρακτηριστικών ισχύει ότι:

• Η ταξινόμηση των εικόνων στηρίζεται στην τιμή απλών βαθμωτών χαρακτηριστικών (*features*), κατά τη διαδικασία ανίχνευσης.

• Η χρήση των χαρακτηριστικών αντί της εικόνας έντασης (*intensity image*) έχει ως σκοπό την μείωση των διαφοροποιήσεων μέσα στην κλάση (*intra-class variability*) και την αύξηση των διαφοροποιήσεων μεταξύ των κλάσεων (*inter-class variability*), ώστε η ταξινόμηση να καταστεί ευκολότερη.

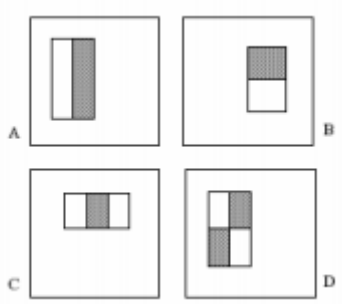
• Τα χαρακτηριστικά αυτά περιέχουν γνώση από συγκεκριμένες περιοχές της εικόνας.

Τα χαρακτηριστικά υπολογίζονται χρησιμοποιώντας τις συναρτήσεις βάσης τύπου *Haar* (*Haar basis*) [17].

• Τα χαρακτηριστικά εφαρμόζονται σε ασπρόμαυρες εικόνες. Η τιμή τους εξαρτάται από την τιμή της υπολογιζόμενης σταθμισμένης, ανάλογα με το εμβαδόν, διαφοράς των αθροισμάτων των εντάσεων των εικονοστοιχείων πάνω σε ορθογώνιες περιοχές. Επιπλέον οι γκρίζες περιοχές θεωρούνται θετικές και οι λευκές αρνητικές, όπως φαίνεται στην Εικόνα 4.4.

• Τα χαρακτηριστικά καθορίζονται από την θέση, τις διαστάσεις και την τιμή τους.

• Το πλήθος των χαρακτηριστικών που δημιουργούνται για παραδείγματα προσώπων 24x24 pixels είναι ~ 45.000, που είναι ένα πολύ μεγαλύτερο σύνολο σε σχέση με τις 576 τιμές έντασης του παραδείγματος. Για αυτόν τον λόγο απαιτείται μια διαδικασία επιλογής των κυριότερων χαρακτηριστικών από αυτά. Σύμφωνα με τους *Viola & Jones* ακόμα και ανιχνευτές με 2 χαρακτηριστικά είναι αρκετά αποτελεσματικοί.

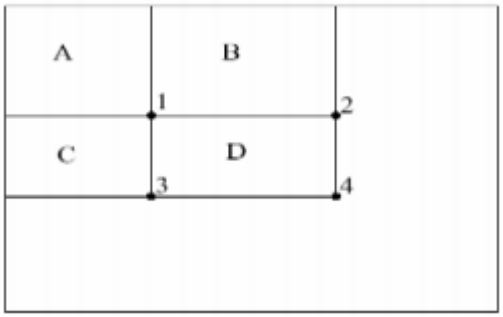


Εικόνα 4.4 Μορφή των χαρακτηριστικών Viola & Jones.

Τα ορθογώνια χαρακτηριστικά των *Viola & Jones* μπορούν να υπολογιστούν πολύ αποτελεσματικά, με τη χρήση μιας βοηθητικής εικόνας, ως "εικόνα ολοκλήρωμα" (*integral image*). Η εικόνα ολοκλήρωμα, *I*, έχει τιμή στη θέση (*x, y*) που καθορίζεται ως άθροισμα των εντάσεων των pixels του ορθογωνίου που ορίζεται από την πάνω αριστερή κορυφή (0, 0) και την κάτω δεξιά κορυφή (x, y):

$$I(x,y) = \sum i(x',y'), x' \le x, y' \le y \tag{4.2}$$

όπου *i* είναι η αρχική εικόνα εισόδου.



Εικόνα 4.5 Αναπαράσταση της Εικόνας Ολοκλήρωμα.

Κάθε ορθογώνιο άθροισμα μπορεί να υπολογιστεί σε σταθερό χρόνο με τέσσερις αναφορές στις τιμές ενός πίνακα, χρησιμοποιώντας την εικόνα ολοκλήρωμα. Έτσι το άθροισμα εντός του *D* (Εικόνα 4.5) μπορεί να υπολογιστεί σαν I4+I1-(I2+I3)[28].

4.2.2. Επιλογή Χαρακτηριστικών με *AdaBoost*

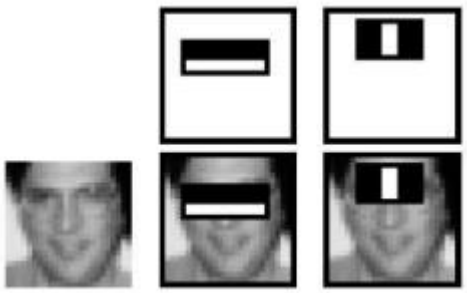
Στη μεθοδολογία *boosting* συνδυάζεται ένα μεγάλο σύνολο μοντέλων ταξινόμησης, αποδίδοντας μεγαλύτερο βάρος σε κάθε καλό μοντέλο ταξινόμησης και μικρότερο βάρος σε υποδέστερα μοντέλα. Αδύναμοι ταξινομητές (*weak classifiers*), ονομάζονται οι απλοί ταξινομητές χαμηλής επίδοσης που συνδυάζονται, ώστε να αποτελέσουν τον τελικό ισχυρό ταξινομητή (*strong classifier*). Ο αλγόριθμος προσαρμοστικής ενίσχυσης (*adaptive boosting*) *AdaBoost*, είναι ο πιο δημοφιλής αλγόριθμος *boosting* και ονομάζεται και διακριτός (*discrete*) *AdaBoost*, μιας και αποδίδει διακριτές τιμές εξόδου [22].

Το πλήθος των χαρακτηριστικών που δημιουργούνται, όπως ήδη έχει αναφερθεί, είναι υπερβολικά μεγάλο, ώστε να απαιτείται η επιλογή των πλέον αποτελεσματικών από αυτά ώστε να χρησιμοποιηθούν στον τελικό ταξινομητή. Οι *Viola & Jones* χρησιμοποίησαν τον αλγόριθμο *AdaBoost* [7] σαν μια αποτελεσματική

διαδικασία για την ανεύρεση ενός μικρού αριθμού "καλών" χαρακτηριστικών που επιπλέον είναι σημαντικά διαφοροποιημένα. Ο περιορισμός του αδύναμου ταξινομητή σε λειτουργίες ταξινόμησης που εξαρτώνται από ένα μόνο χαρακτηριστικό, είναι μια απλή και πρακτική μέθοδος για την επίτευξη αυτού του στόχου [28,29].

Έτσι, η μέθοδος *AdaBoost* στοχεύει στην επίλυση των παρακάτω 3 θεμελιωδών προβλημάτων [13]:

1. Εντοπισμός των πιο αποτελεσματικών χαρακτηριστικών από ένα μεγάλο σύνολο χαρακτηριστικών
2. Κατασκευή αδύναμων ταξινομητών, καθένας από τους οποίους στηρίζεται σε ένα μόνο από τα δημιουργηθέντα χαρακτηριστικά, και
3. Συνδυασμός των αδύναμων ταξινομητών για την κατασκευή ενός ισχυρού ταξινομητή.



Εικόνα 4.6 Τα 2 κυριότερα χαρακτηριστικά εφαρμοσμένα σε ένα τυπικό πρόσωπο.

4.2.3. Κατασκευή του Αδύναμου Ταξινομητή

Η "ρίζα" ("*stump*") ενός δένδρου απόφασης (*decision tree*), είναι ο απλούστερος τύπος ενός αδύναμου ταξινομητή. Μία ρίζα απόφασης, μπορεί να κατασκευαστεί όταν το χαρακτηριστικό παίρνει πραγματικές τιμές, και συγκρίνοντας απλά την τιμή του επιλεγμένου χαρακτηριστικού με μια συγκεκριμένη τιμή

κατωφλίου. Έτσι ο αδύναμος ταξινομητής σχεδιάζεται ώστε να μπορεί να επιλέγει εκείνο το μοναδικό χαρακτηριστικό που διαχωρίζει καλύτερα τα θετικά από τα αρνητικά παραδείγματα. Για κάθε χαρακτηριστικό, ο αδύναμος ταξινομητής καθορίζει το ιδανικό κατώφλι λειτουργίας της ταξινόμησης, έτσι ώστε να ελαχιστοποιείται ο αριθμός παραδειγμάτων που ταξινομείται εσφαλμένα [28,29,13].

Εικόνα 4.7 Υπολογισμός της τιμής των χαρακτηριστικών πάνω σε πρόσωπα και σε μη-πρόσωπα.

Έτσι ο αδύναμος ταξινομητής $h_j(x)$ αποτελείται από ένα χαρακτηριστικό j και ένα κατώφλι θ_j:

• Για κάθε χαρακτηριστικό j, υπολογίζεται η $f_j(x)$, μια βαθμωτή τιμή του χαρακτηριστικού (εδώ οι διαφορές αθροισμάτων), όπου x είναι ένα θετικό ή αρνητικό παραδείγμα

• Κάθε χαρακτηριστικό χρησιμοποιείται σαν ένας αδύναμος ταξινομητής

• Καθορισμός τιμής κατωφλίου θ_j για κάθε χαρακτηριστικό έτσι ώστε τα περισσότερα παραδείγματα να ταξινομούνται σωστά:

$$h(x) = \begin{cases} 1, if \;\; f_j(x) < \theta_j \; or \;\; f_j(x) > \theta_j, x : positive \\ 0, otherwise, \; x : negative \end{cases} \qquad (4.3)$$

• Επιλογή χαρακτηριστικού και κατωφλίου με το χαμηλότερο σταθμισμένο σφάλμα ταξινόμησης.

• Διαδοχική αποτίμηση όλων των χαρακτηριστικών.

4.2.4. Ταξινόμηση με Ακολουθία Ταξινομητών

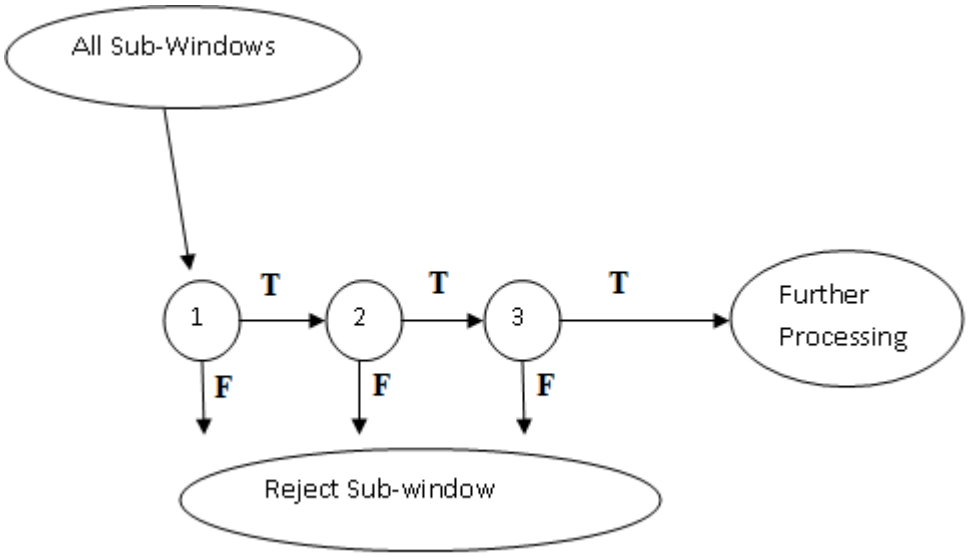
Οι *Viola & Jones* εισήγαγαν την έννοια της ακολουθίας ταξινομητών (*cascade of classifiers*), ώστε να βελτιώσουν την επίδοση του *AdaBoost* που απαιτεί μεγάλο αριθμό δειγμάτων και καταλήγει σε μεγάλο πλήθος αδύναμων ταξινομητών που απαιτούν μεγάλο υπολογιστικό κόστος, [28,29].

Με τη μέθοδο αυτή κατασκευάζεται μία διαδοχή ταξινομητών (cascade) που επιτυγχάνει αυξημένη απόδοση στην ανίχνευση και μειώνει ριζικά τον χρόνο υπολογισμού. Η ιδέα είναι ότι μπορούν να κατασκευαστούν μικροί και ωστόσο αποτελεσματικοί, συνδυασμένοι ταξινομητές που απορρίπτουν πολλά από τα αρνητικά, ενώ ανιχνεύουν σχεδόν όλα τα θετικά παραδείγματα.

Η μέθοδος της ακολουθίας ταξινομητών στηρίζεται στο γεγονός ότι σε μία οποιαδήποτε εικόνα η πλειονότητα των παραθύρων ανίχνευσης δεν περιλαμβάνει πρόσωπα. Έτσι πιο απλοποιημένοι και λιγότερο χρονοβόροι ταξινομητές χρησιμοποιούνται για να απορρίψουν την πλειονότητα των παραθύρων ανίχνευσης ως αρνητικά, προτού χρησιμοποιηθούν οι πιο σύνθετοι και περισσότερο χρονοβόροι ταξινομητές που θα επεξεργαστούν τις πιο πολύπλοκες περιπτώσεις και θα επιτύχουν χαμηλά επίπεδα εσφαλμένων θετικών ανιχνεύσεων.

Παράδειγμα: Ακολουθία Ταξινομητών 32 επιπέδων
• Ταξινομητής 2-χαρακτηριστικών στο πρώτο επίπεδο Απορρίπτει το 60% των μη-προσώπων ενώ ανιχνεύει 100% τα πρόσωπα • Ταξινομητής 5-χαρακτηριστικών στο δεύτερο επίπεδο Απορρίπτει το 80% των μη-προσώπων ενώ ανιχνεύει 100% τα πρόσωπα • Ταξινομητής 20-χαρακτηριστικών στα επίπεδα 3,4 και 5 • Ταξινομητής 50-χαρακτηριστικών στα επίπεδα 6 και 7 • Ταξινομητής 100-χαρακτηριστικών στα επίπεδα 8 έως και 12 • Ταξινομητής 200-χαρακτηριστικών στα επίπεδα 13 έως και 32

Η συνολική διαδικασία ανίχνευσης είναι παρόμοια με ένα δένδρο απόφασης (*decision tree*). Ένα θετικό αποτέλεσμα από τον ταξινομητή πρώτου επιπέδου οδηγείται στον ταξινομητή δευτέρου επιπέδου, του οποίου το θετικό αποτέλεσμα οδηγείται στον ταξινομητή τρίτου επίπεδου κ.ο.κ. όπως στην Εικόνα 4.8. Τα αρνητικά αποτελέσματα σε κάθε επίπεδο απορρίπτονται χωρίς να επανελέγχονται. Έτσι οι ταξινομητές των αρχικών επιπέδων ασχολούνται με τα εύκολα περιστατικά, ενώ οι επόμενοι αντιμετωπίζουν πιο δύσκολες περιπτώσεις.



Εικόνα 4.8 Σχηματική παράσταση μιας ανίχνευσης με ακολουθία ταξινομητών.

Η εκπαίδευση του ακολουθιακού ταξινομητή γίνεται χρησιμοποιώντας τον *AdaBoost*, και καθορίζει:

- τον αριθμό των επιπέδων του καταρράκτη ταξινομητή
- τον αριθμό των χαρακτηριστικών σε κάθε επίπεδο
- το κατώφλι σε κάθε επίπεδο

ώστε να ελαχιστοποιείται ο αριθμός των χρησιμοποιούμενων χαρακτηριστικών.

ΚΕΦΑΛΑΙΟ 5. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

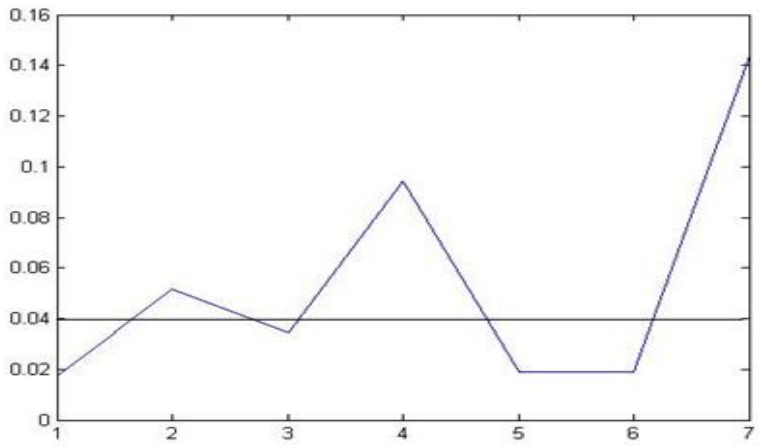
Σε αυτό το κεφάλαιο θα παρουσιάσουμε τα πειραματικά αποτελέσματα για τις 12 αναπαραστάσεις που προτείναμε στο Κεφάλαιο 3. Παρουσιάζονται επίσης και οι τέσσερις διαφορετικοί τρόποι με τους οποίους συγκρίνονται τα πλάνα μεταξύ τους. Στη συνέχεια, παρουσιάζεται η βελτίωση των αποτελεσμάτων με τη χρήση του εντοπισμού προσώπου και σώματος. Τέλος, γίνεται σύγκριση με τη χρήση ιστογραμμάτων χρώματος και απλών περιγραφέων SIFT.

5.1. Σύνολο Δεδομένων

Για την υλοποίηση των πειραμάτων γίνεται επεξεργασία δέκα βίντεο τα οποία προήλθαν από τα δεδομένα ελέγχου της *TRECVID* 2008 [16], στα πλαίσια του διαγωνισμού *Rushes Summarization* της *TRECVID* 2008. Πρόκειται για αμοντάριστα βίντεο, κυρίως σκηνές από πέντε σειρές προγραμμάτων δράματος του *BBC* που διατέθηκαν από το αρχείο του *BBC* στα πλαίσια του διαγωνισμού *TRECVID*. Οι δραματικές σειρές περιλαμβάνουν ένα ιστορικό δράμα στο Λονδίνο στις αρχές του 1900, μια σειρά για την αρχαία Ελλάδα, ένα σύγχρονο αστυνομικό πρόγραμμα, ένα πρόγραμμα με υπηρεσίες έκτακτης ανάγκης, ένα αστυνομικό δράμα και διάφορες σκηνές από άλλα προγράμματα. Ο χωρισμός των βίντεο σε πλάνα και η απομάκρυνση πλάνων με ανεπιθύμητα εικονοπλαίσια έγινε χειρονακτικά.

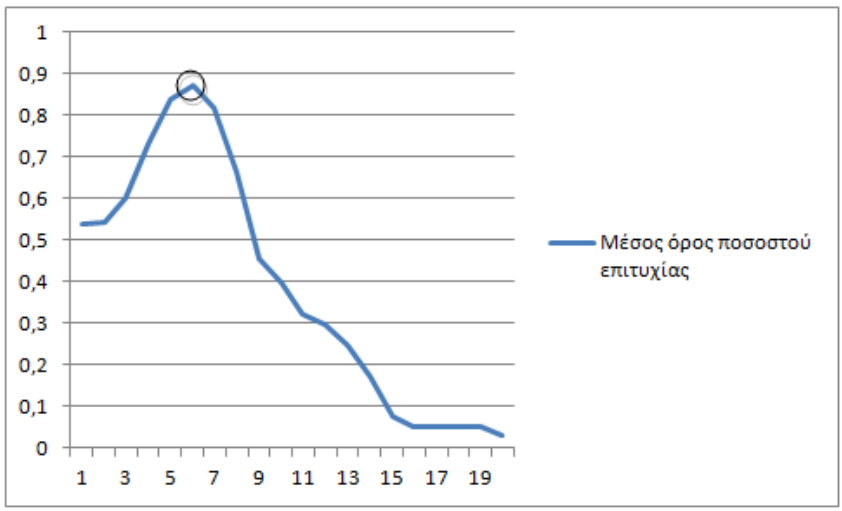
5.2. Πειραματικά Αποτελέσματα για τις 12 Αναπαραστάσεις

Αρχικά, εξήχθησαν τα χαρακτηριστικά εικονοπλαίσια (*key-frames*) από τα βίντεο με τον αλγόριθμο που προτείνεται στο [3]. Στη συνέχεια όπως προαναφέραμε υπολογίστηκαν και οι γειτονιές αυτών ενώ εξήχθησαν οι περιγραφείς *SIFT* [14]. Έπειτα, υπολογίσαμε τα ιστογράμματα των οπτικών λέξεων και στη συνέχεια εφαρμόσαμε τους ανιχνευτές σημασιολογικών εννοιών *web-81* και *vireo-374* οι οποίοι μας επέστρεψαν τα αντίστοιχα σημασιολογικά χαρακτηριστικά όπως περιγράφηκε στο κεφάλαιο 2. Δώδεκα διαφορετικές αναπαραστάσεις χρησιμοποιήθηκαν για την αναπαράσταση κάθε πλάνου. Για κάθε μια αναπαράσταση συγκρίναμε διαδοχικά πλάνα μεταξύ τους, θωρώντας ως μεταξύ τους απόσταση την μικρότερη απόσταση όλων των δυνατών συνδυασμών (ζεύγη) των επιμέρους διανυσμάτων χαρακτηριστικών που τα αντιπροσωπεύουν. Με αυτό τον τρόπο προέκυψε ένα διάνυσμα αποστάσεων μεταξύ διαδοχικών πλάνων. Ως όρια αλλαγής ομάδας όμοιων πλάνων ορίζονται εκείνες οι θέσεις που έχουν τιμή μεγαλύτερη από ένα προκαθορισμένο κατώφλι. Το κατώφλι παίρνει τιμές ανάμεσα στο 0 και στη μεγαλύτερη απόσταση που παρατηρείται σε όλα τα βίντεο t_{max}, με βήμα που προκύπτει από την σχέση $t_{max}/20$. Για παράδειγμα, θέτοντας ως κατώφλι την τιμή 0.04 στο διάνυσμα αποστάσεων της Εικόνας 5.1, οι θέσεις [2,4,7], οι οποίες έχουν τιμή μεγαλύτερη από το κατώφλι, υποδεικνύουν θέσεις αλλαγής της ομάδας όμοιων πλάνων.



Εικόνα 5.1 Αναπαράσταση των αποστάσεων μεταξύ διαδοχικών πλάνων.

Στη συνέχεια συγκρίνουμε αυτές τις θέσεις με τις πραγματικές θέσεις (*ground truth*) και παίρνουμε το ποσοστό επιτυχίας χρησιμοποιώντας την Εξίσωση 3.17. Το τελικό αποτέλεσμα είναι ο μέσος όρος των ποσοστών επιτυχίας στα δέκα βίντεο. Διατηρούμε μόνο τον υψηλότερο μέσο όρο του ποσοστού επιτυχίας από τα ποσοστά επιτυχίας που υπολογίσαμε για τα 20 διαφορετικά κατώφλια. Για παράδειγμα για τις 20 τιμές του μέσου όρου ποσοστού επιτυχίας που απεικονίζονται στην Εικόνα 5.2, παίρνουμε μόνο την υψηλότερη τιμή η οποία και μας δίνει το καλύτερο αποτέλεσμα. Σε αυτή την περίπτωση η υψηλότερη τιμή είναι 87%.



Εικόνα 5.2 Μέσος όρος ποσοστού επιτυχίας.

Στη συνέχεια θα παρουσιάσουμε σε πίνακες τα αποτελέσματα των βέλτιστων μέσων όρων του ποσοστού επιτυχίας για τις 12 αναπαραστάσεις για τα σημασιολογικά χαρακτηριστικά που ανιχνεύτηκαν από τους ανιχνευτές σημασιολογικών εννοιών *web-81* και *vireo-374*. Διακρίνουμε τέσσερις διαφορετικές περιπτώσεις. Στην πρώτη περίπτωση (Πίνακας 5.1) αναπαριστούμε τα αποτελέσματα των 12 αναπαραστάσεων των σημασιολογικών χαρακτηριστικών στα οποία εφαρμόσαμε τους ανιχνευτές σημασιολογικών εννοιών *web-81*. Αντίστοιχα, στον Πίνακα 5.2 παρουσιάζονται τα αποτελέσματα των σημασιολογικών χαρακτηριστικών εφαρμόζοντας τους ανιχνευτές σημασιολογικών εννοιών *vireo-374*. Στην τρίτη περίπτωση (Πίνακας 5.3) εμφανίζονται τα αποτελέσματα από τη συνένωση των

σημασιολογικών χαρακτηριστικών που προέκυψαν από τους ανιχνευτές σημασιολογικών εννοιών *web-81* και *vireo-374*. Τέλος, στον Πίνακα 5.4 παρουσιάζεται ο καλύτερος μέσος όρος των ποσοστών επιτυχίας των σημασιολογικών χαρακτηριστικών, στα οποία εφαρμόστηκαν οι ανιχνευτές των 81 και 374 σημασιολογικών εννοιών, από το συνδυασμό των αποστάσεων αυτών.

Πίνακας 5.1 Αποτελέσματα με τη χρήση των ανιχνευτών σημασιολογικών εννοιών *web-81*.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	**84%**	**84%**	**87%**	**86%**	**87%**
Αναπαράσταση 2	80%	80%	80%	80%	80%
Αναπαράσταση 3	63%	68%	70%	67%	63%
Αναπαράσταση 4	67%	68%	73%	72%	72%
Αναπαράσταση 5	66%	69%	72%	71%	69%
Αναπαράσταση 6	78%	78%	78%	78%	78%
Αναπαράσταση 7	78%	79%	80%	80%	79%
Αναπαράσταση 8	72%	71%	74%	74%	74%
Αναπαράσταση 9	66%	66%	66%	66%	66%
Αναπαράσταση 10	66%	65%	67%	66%	65%
Αναπαράσταση 11	64%	64%	64%	64%	64%
Αναπαράσταση 12	66%	67%	65%	66%	62%

Πίνακας 5.2 Αποτελέσματα με τη χρήση των ανιχνευτών σημασιολογικών εννοιών *vireo-374*.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	**83%**	**84%**	**83%**	**86%**	**87%**
Αναπαράσταση 2	80%	80%	80%	80%	80%
Αναπαράσταση 3	73%	78%	80%	74%	73%
Αναπαράσταση 4	79%	77%	75%	79%	81%
Αναπαράσταση 5	77%	79%	79%	78%	77%
Αναπαράσταση 6	80%	80%	80%	80%	80%
Αναπαράσταση 7	79%	80%	80%	81%	80%
Αναπαράσταση 8	77%	75%	80%	79%	80%
Αναπαράσταση 9	73%	73%	73%	73%	73%
Αναπαράσταση 10	79%	74%	72%	70%	76%
Αναπαράσταση 11	68%	68%	68%	68%	68%
Αναπαράσταση 12	70%	70%	70%	72%	72%

Πίνακας 5.3 Αποτελέσματα από τη συνένωση των σημασιολογικών χαρακτηριστικών που προέκυψαν από τη χρήση των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374*.

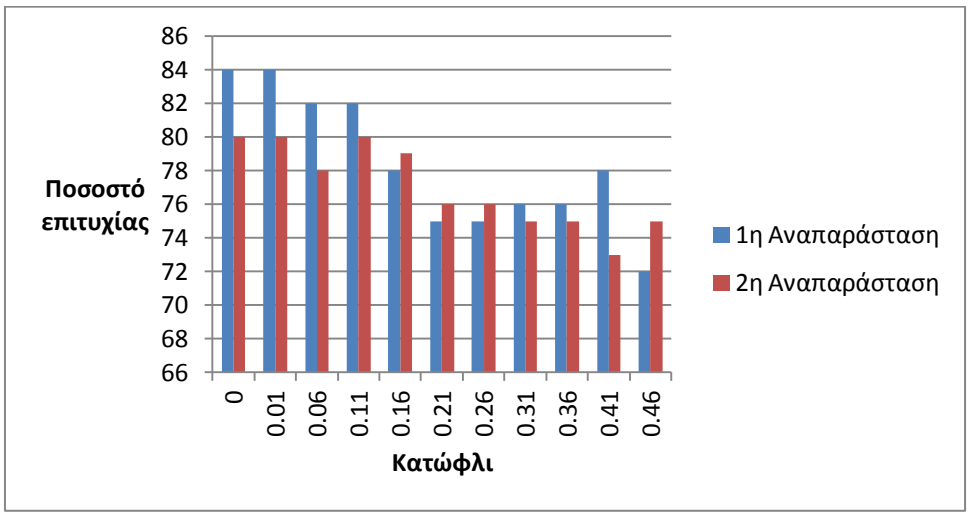
Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	**88%**	**88%**	**87%**	**88%**	**89%**
Αναπαράσταση 2	84%	84%	84%	84%	84%
Αναπαράσταση 3	75%	79%	80%	80%	79%
Αναπαράσταση 4	78%	78%	84%	86%	85%
Αναπαράσταση 5	78%	81%	85%	82%	80%
Αναπαράσταση 6	81%	81%	81%	81%	81%
Αναπαράσταση 7	80%	81%	82%	82%	81%
Αναπαράσταση 8	80%	81%	80%	81%	82%
Αναπαράσταση 9	73%	73%	73%	73%	73%
Αναπαράσταση 10	78%	76%	73%	79%	76%
Αναπαράσταση 11	71%	71%	71%	71%	71%
Αναπαράσταση 12	73%	72%	73%	73%	74%

Πίνακας 5.4 Αποτελέσματα με συνδυασμό των αποστάσεων που προέκυψαν από τη χρήση των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374*.

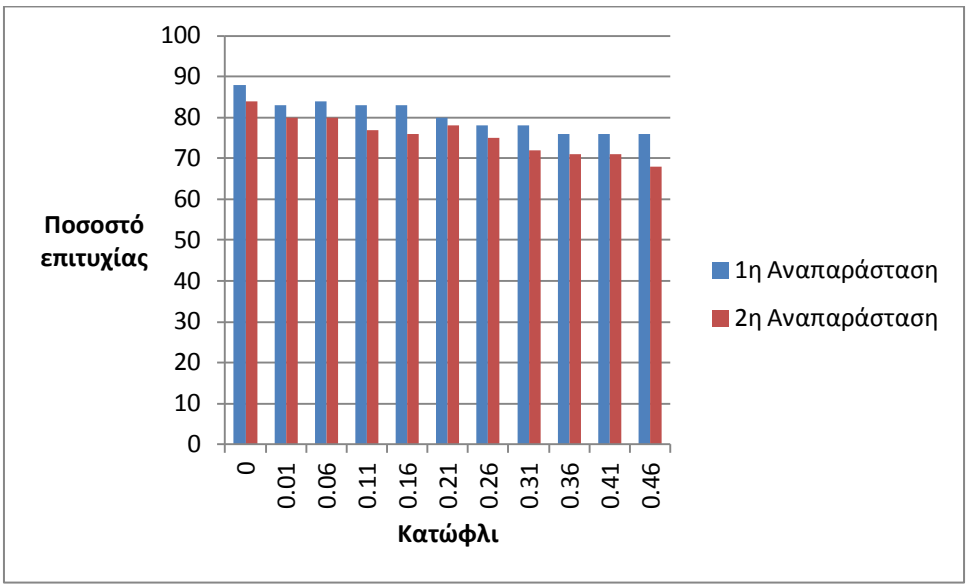
Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	**90%**	**89%**	**87%**	**87%**	**93%**
Αναπαράσταση 2	85%	85%	85%	85%	85%
Αναπαράσταση 3	76%	79%	80%	79%	77%
Αναπαράσταση 4	83%	81%	81%	85%	83%
Αναπαράσταση 5	79%	80%	80%	83%	81%
Αναπαράσταση 6	81%	80%	81%	81%	81%
Αναπαράσταση 7	82%	82%	83%	82%	82%
Αναπαράσταση 8	79%	81%	82%	81%	81%
Αναπαράσταση 9	77%	77%	77%	77%	77%
Αναπαράσταση 10	80%	76%	75%	73%	79%
Αναπαράσταση 11	70%	70%	70%	70%	70%
Αναπαράσταση 12	73%	73%	71%	72%	74%

Παρατηρούμε ότι οι δυο πρώτες αναπαραστάσεις έχουν την καλύτερη απόδοση. Η πρώτη αναπαριστά το μέσο όρο των σημασιολογικών χαρακτηριστικών εικονοπλαισίων μίας γειτονιάς ενός χαρακτηριστικού εικονοπλαισίου (*key-frame*) και η δεύτερη μόνο τα σημασιολογικά χαρακτηριστικά των χαρακτηριστικών εικονοπλαισίων. Για αυτό το λόγο στη συνέχεια προτείνουμε βελτιώσεις τις απόδοσης μόνο για τις δύο αυτές αναπαραστάσεις.

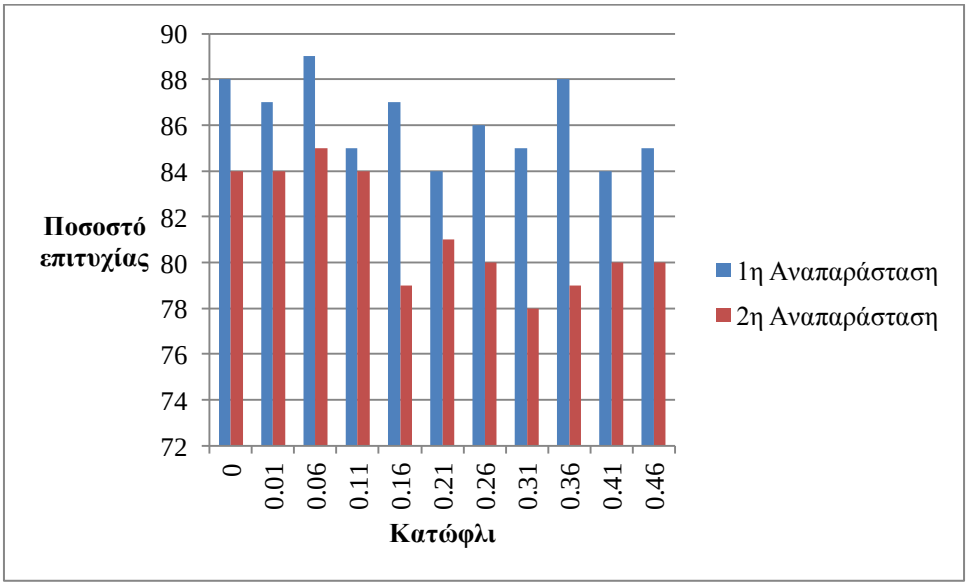
Επιπλέον, αναφέρουμε και τα αποτελέσματα που αντλήσαμε συγκρίνοντας μόνο τις τιμές των σημασιολογικών χαρακτηριστικών των πλάνων όπου ο μέσος όρος των σημασιολογικών χαρακτηριστικών για όλα τα πλάνα είναι μεγαλύτερος από ένα κατώφλι, όπως περιγράφεται στην 4^{η} περίπτωση (Κεφάλαιο 3). Με αυτό τον τρόπο, ένα υποσύνολο των εννοιών χρησιμοποιείται ουσιαστικά για τον υπολογισμό των σημασιολογικών χαρακτηριστικών. Το όριο που θέσαμε κυμαίνεται από 0.01 έως 0.5 με βήμα 0.05. Σε αυτά τα σημασιολογικά χαρακτηριστικά εφαρμόσαμε τη μέθοδό μας για τις δυο πρώτες αναπαραστάσεις και παρουσιάζουμε το βέλτιστο ποσοστό επιτυχίας για κάθε γειτονιά ξεχωριστά αλλά και τα αποτελέσματα όταν εφαρμόζουμε τους ανιχνευτές σημασιολογικών εννοιών web-81 και vireo-374, τη συνένωση αυτών αλλά και το συνδυασμό των αποστάσεών τους. Τα αποτελέσματα είναι τα ακόλουθα:



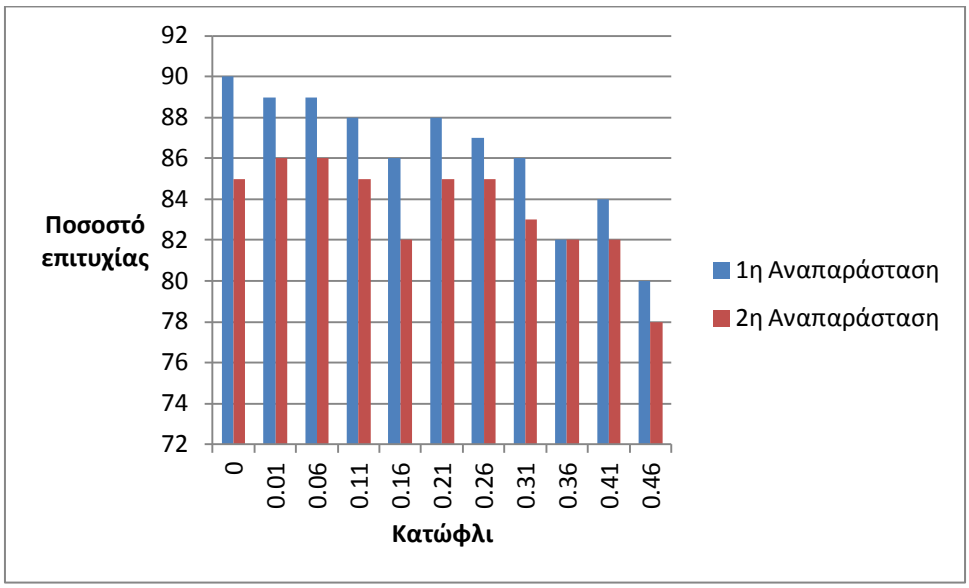
Εικόνα 5.3 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *web-81* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=3).



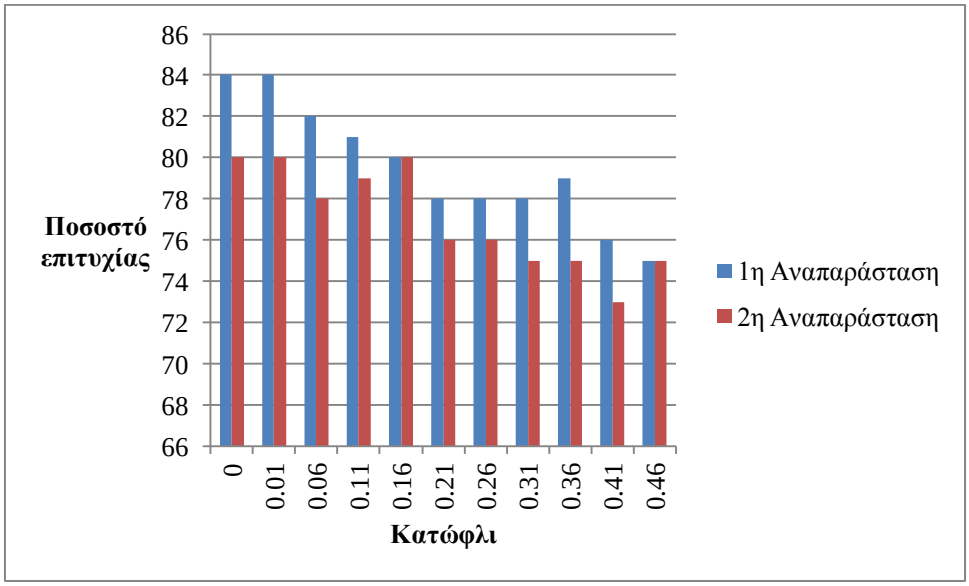
Εικόνα 5.4 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=3).



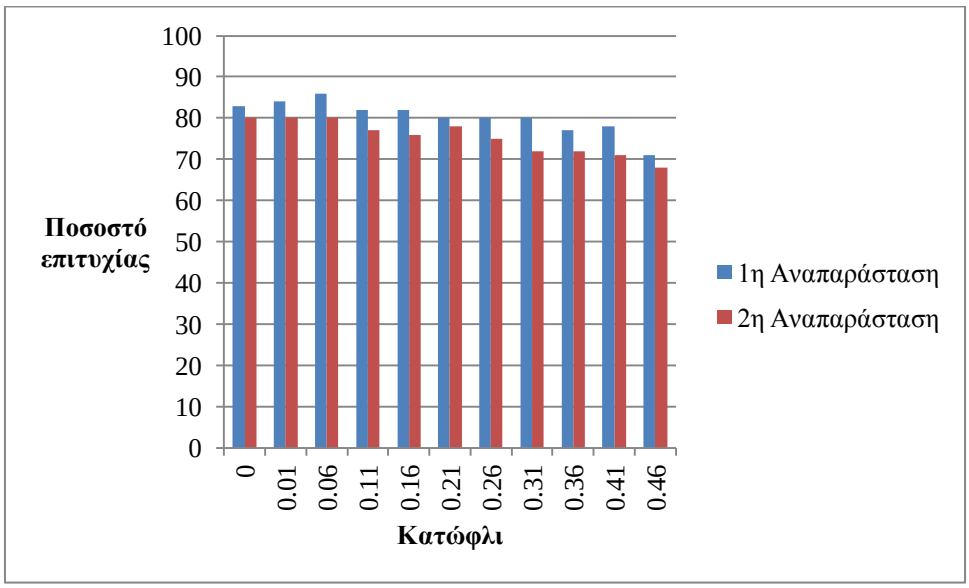
Εικόνα 5.5 Αποτελέσματα από συνένωση των σημασιολογικών χαρακτηριστικών των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=3).



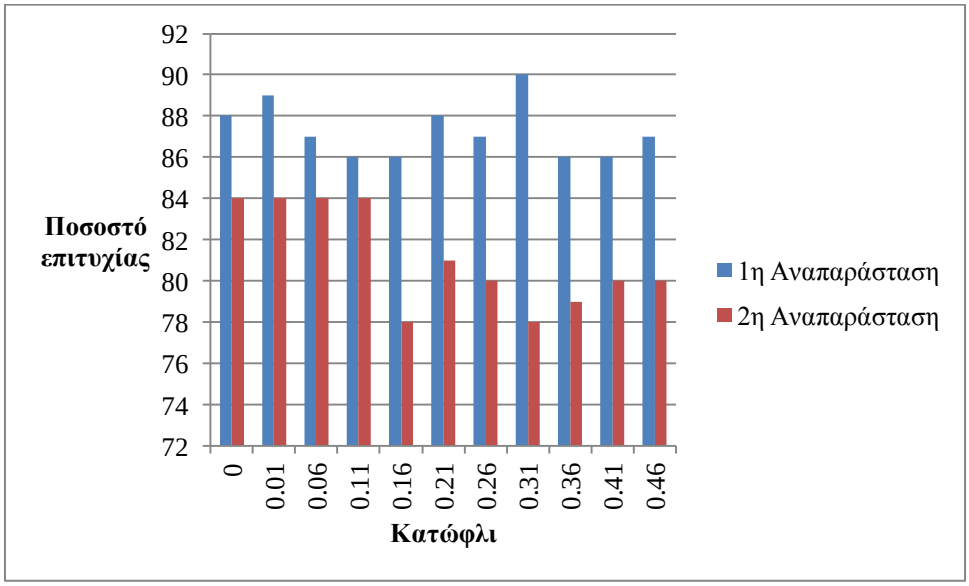
Εικόνα 5.6 Αποτελέσματα με συνδυασμό των αποστάσεων που προέκυψαν από τη χρήση των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=3).



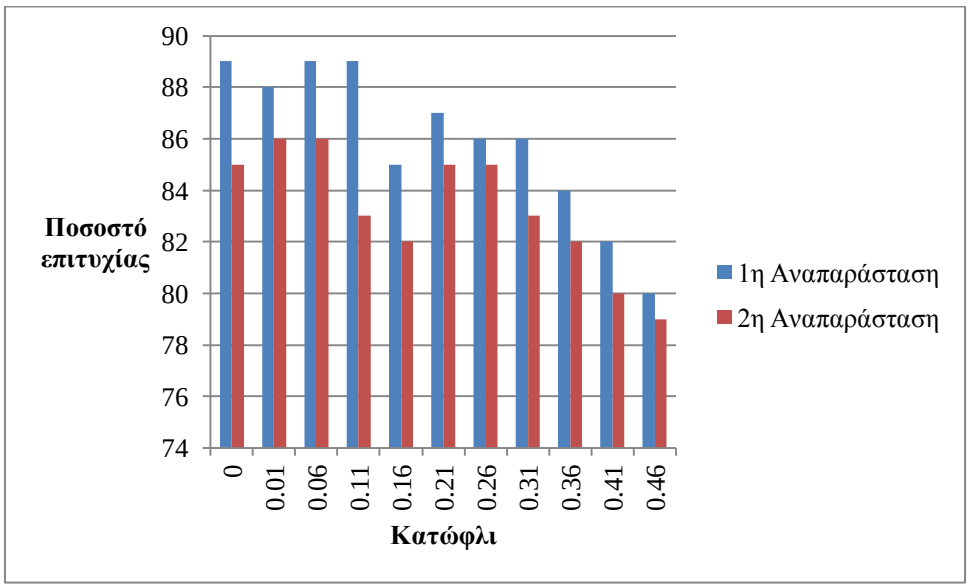
Εικόνα 5.7 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *web-81* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=5).



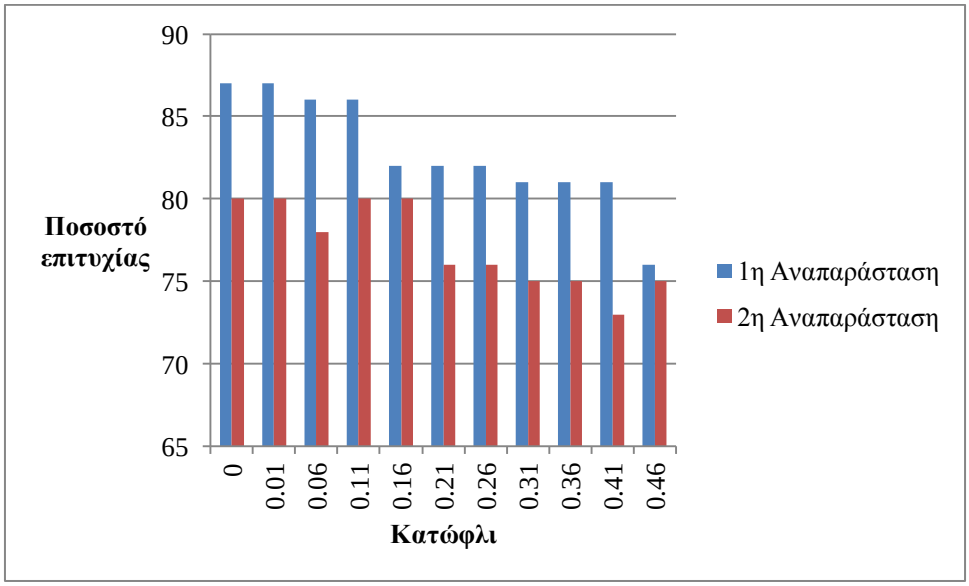
Εικόνα 5.8 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=5).



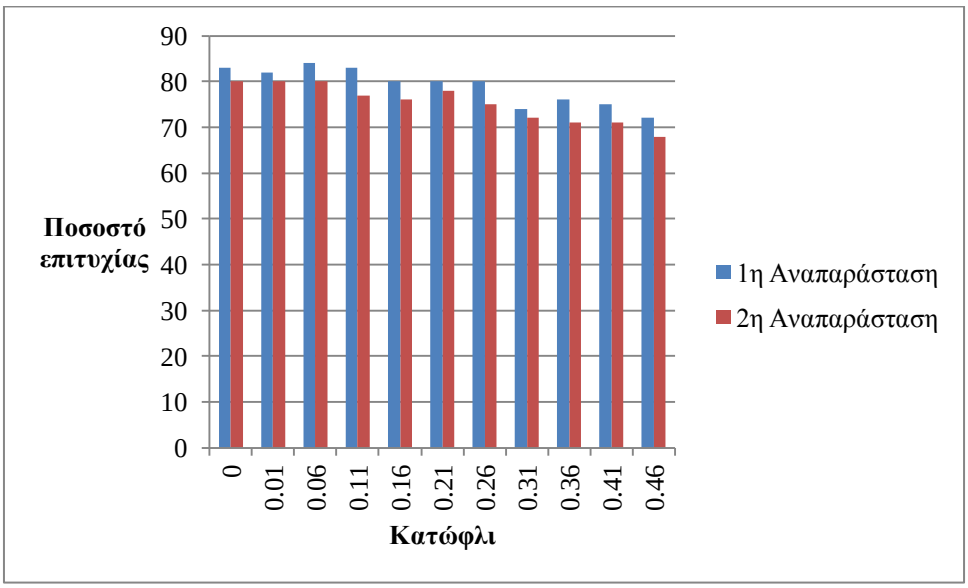
Εικόνα 5.9 Αποτελέσματα από συνένωση των σημασιολογικών χαρακτηριστικών των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=5).



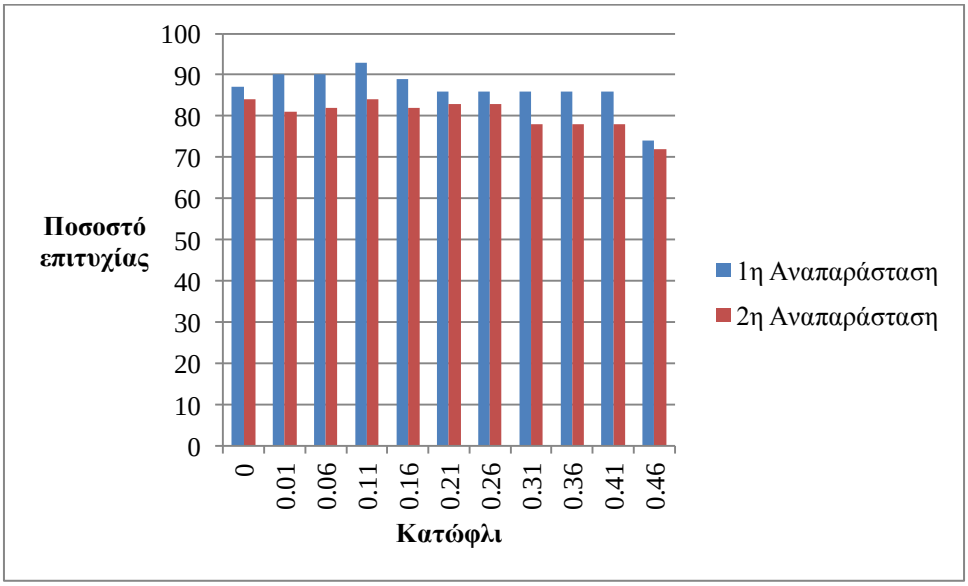
Εικόνα 5.10 Αποτελέσματα με συνδυασμό των αποστάσεων που προέκυψαν από τη χρήση των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=5).



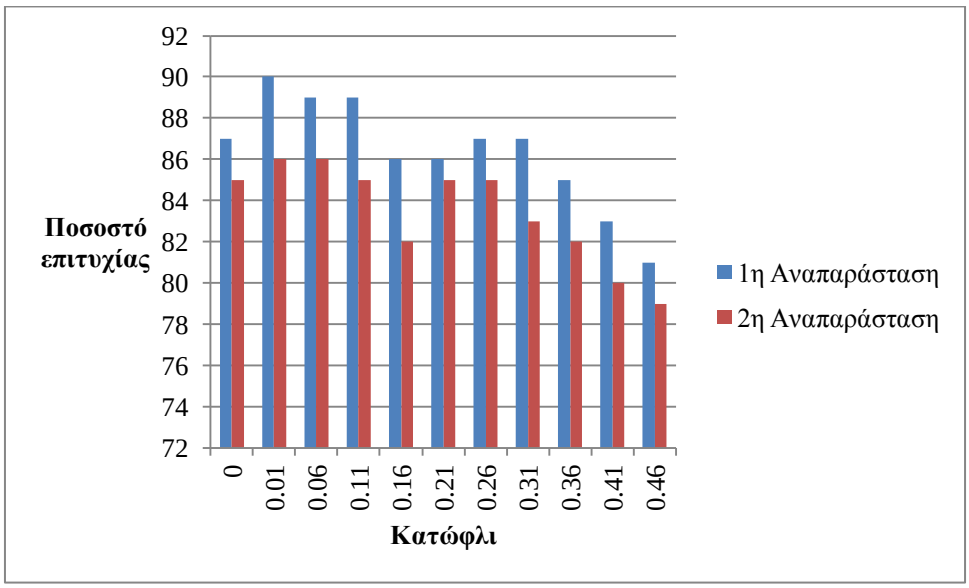
Εικόνα 5.11 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *web-81* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=7).



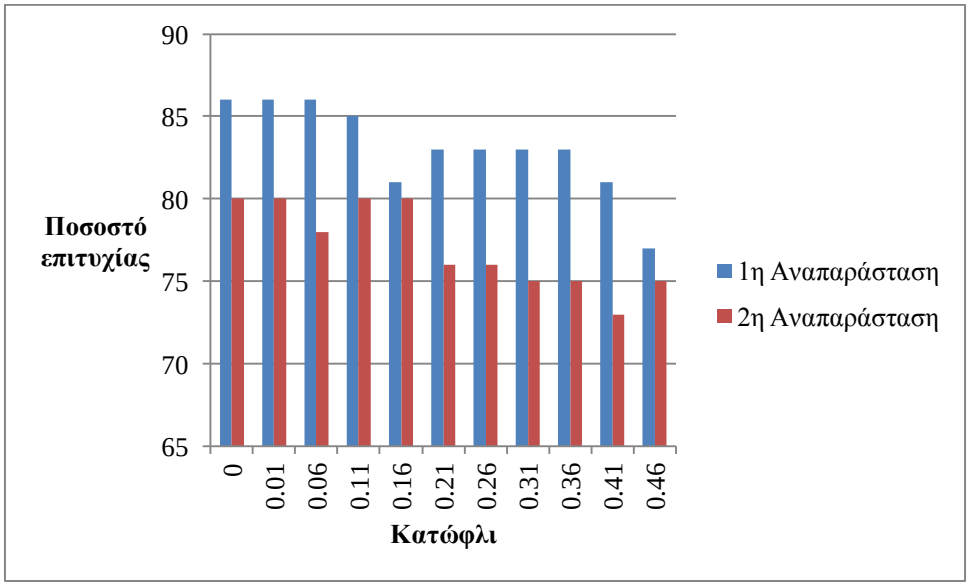
Εικόνα 5.12 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=7).



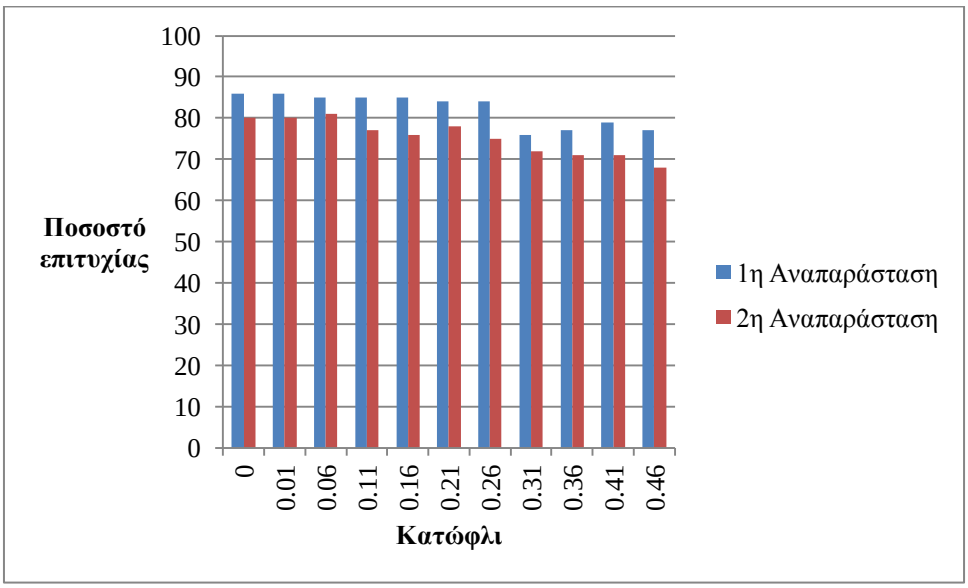
Εικόνα 5.13 Αποτελέσματα από συνένωση των σημασιολογικών χαρακτηριστικών των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=7).



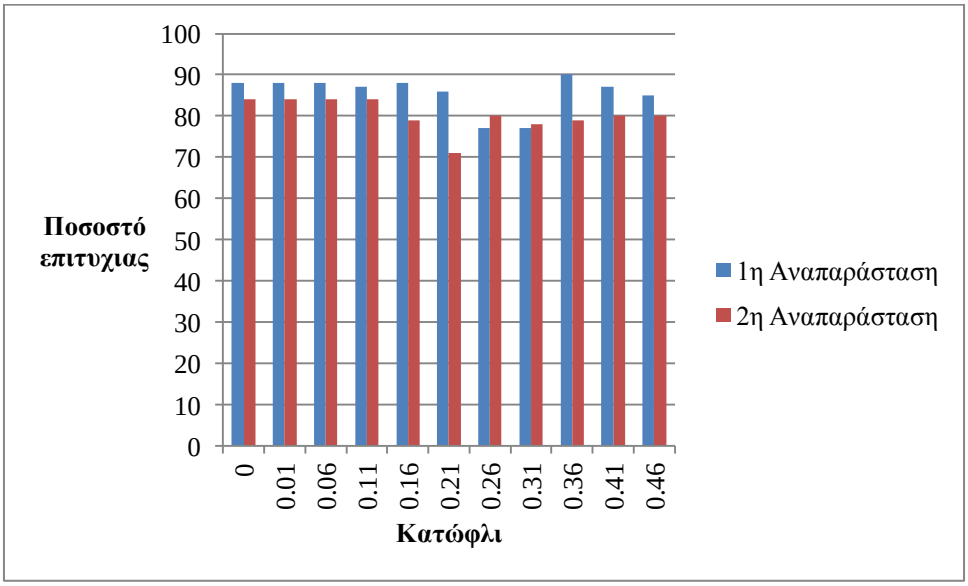
Εικόνα 5.14 Αποτελέσματα με συνδυασμό των αποστάσεων που προέκυψαν από τη χρήση των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=1) και γειτονιά (N=7).



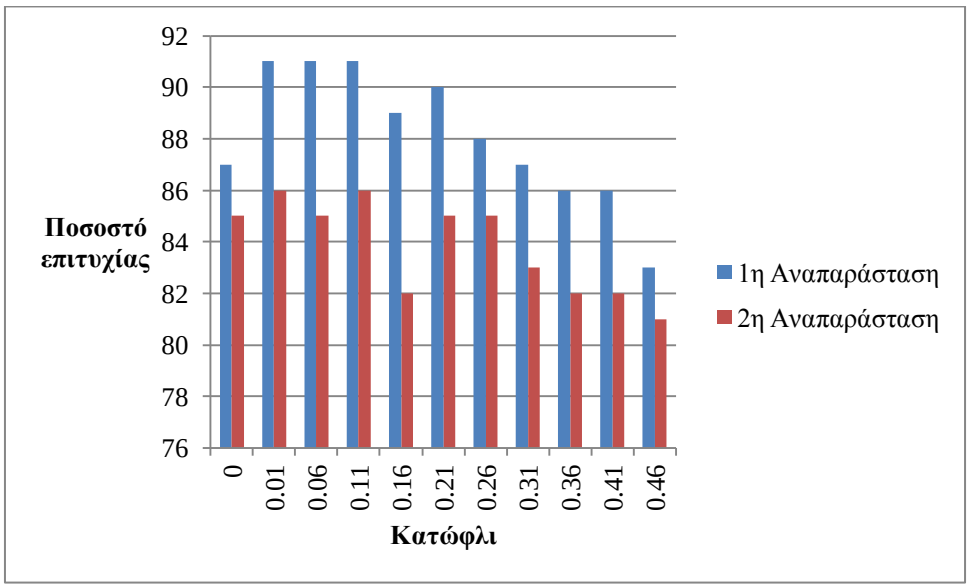
Εικόνα 5.15 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *web-81* για ένα υποσύνολο εννοιών όταν βήμα (d=3) και γειτονιά (N=7).



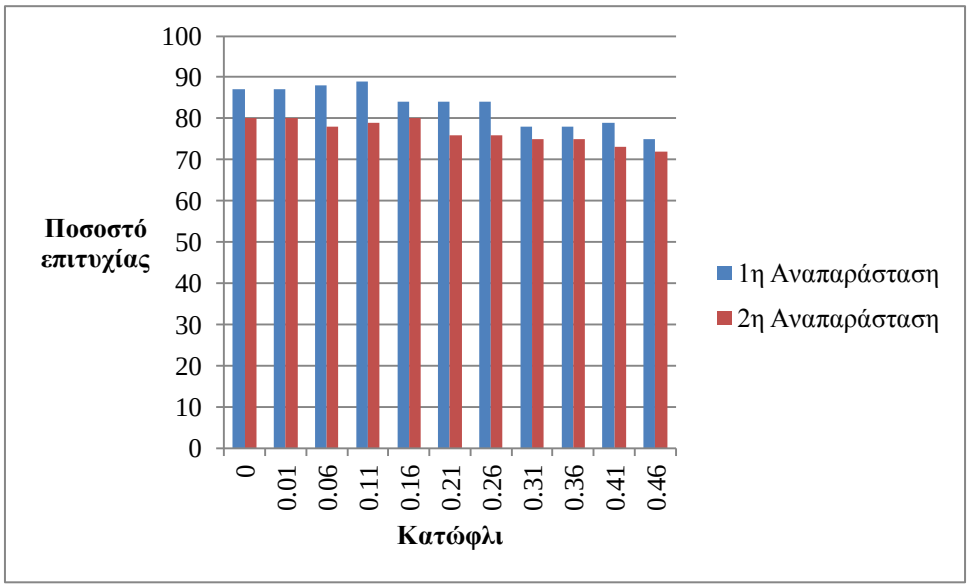
Εικόνα 5.16 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=3) και γειτονιά (N=7).



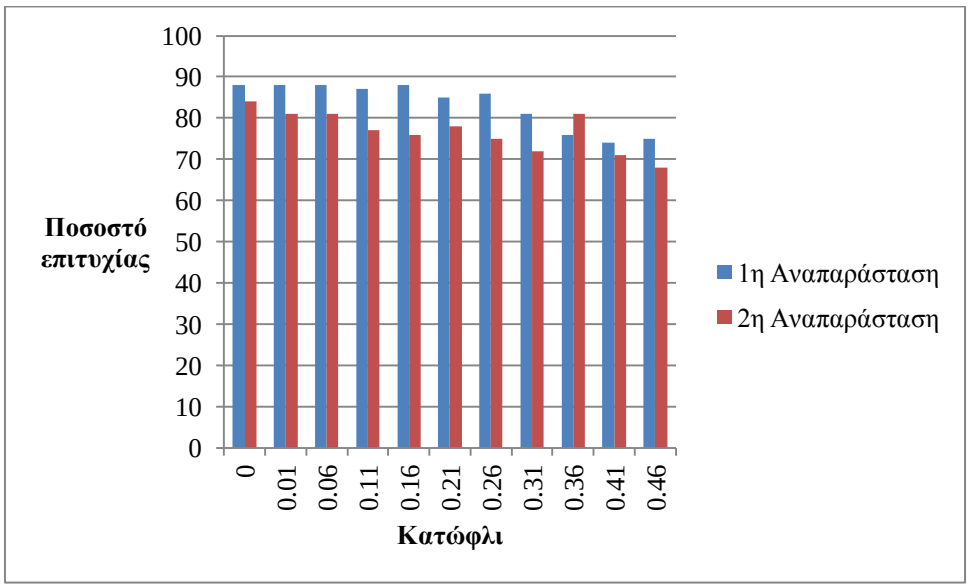
Εικόνα 5.17 Αποτελέσματα από συνένωση των σημασιολογικών χαρακτηριστικών των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=3) και γειτονιά (N=7).



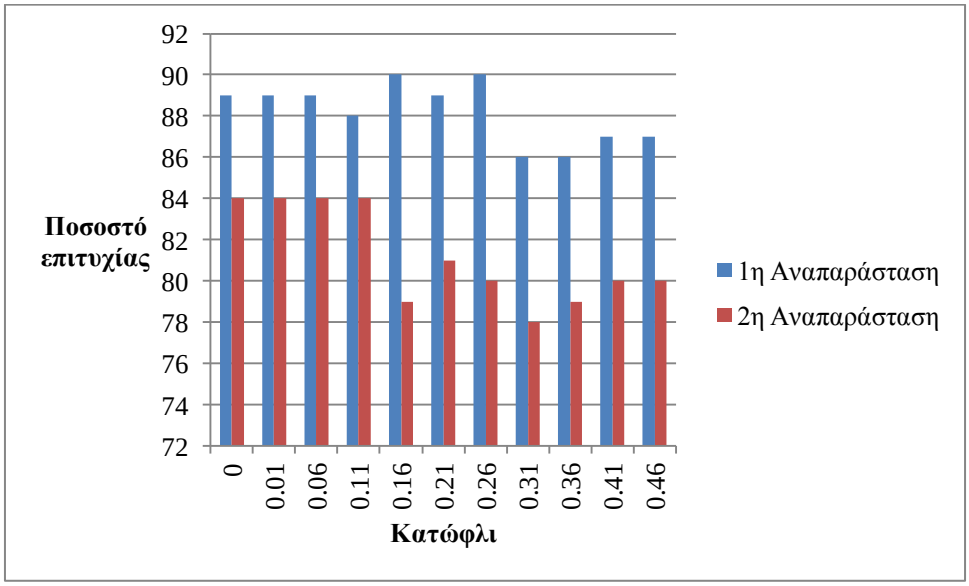
Εικόνα 5.18 Αποτελέσματα με συνδυασμό των αποστάσεων που προέκυψαν από τη χρήση των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=3) και γειτονιά (N=7).



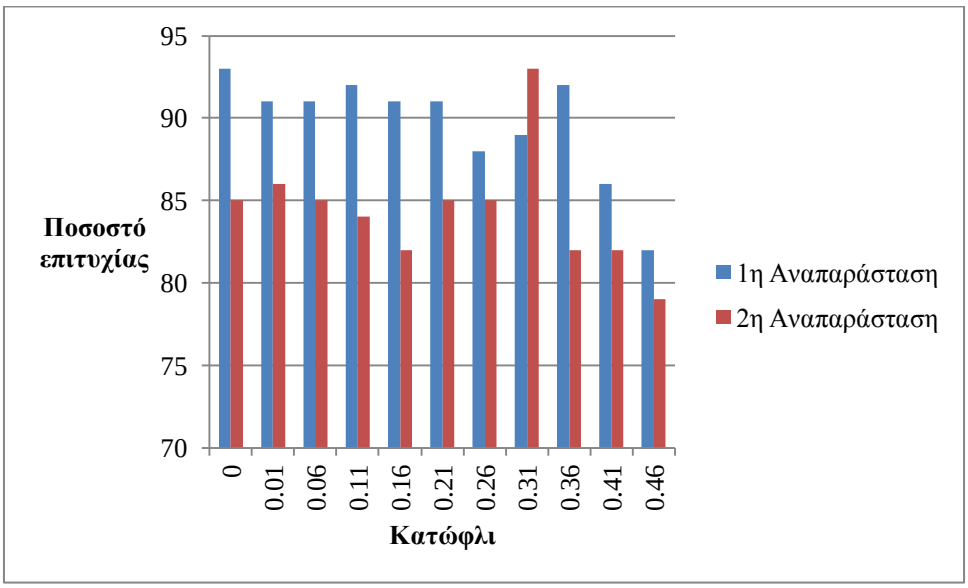
Εικόνα 5.19 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *web-81* για ένα υποσύνολο εννοιών όταν βήμα (d=5) και γειτονιά (N=7).



Εικόνα 5.20 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=5) και γειτονιά (N=7).



Εικόνα 5.21 Αποτελέσματα από συνένωση των σημασιολογικών χαρακτηριστικών των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=5) και γειτονιά (N=7).



Εικόνα 5.22 Αποτελέσματα με συνδυασμό των αποστάσεων που προέκυψαν από τη χρήση των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* για ένα υποσύνολο εννοιών όταν βήμα (d=5) και γειτονιά (N=7).

Παρατηρούμε ότι χρησιμοποιώντας λιγότερες σημασιολογικές έννοιες η μέθοδος έχει καλή απόδοση και σε ορισμένες περιπτώσεις καλύτερη και από τη αρχική, η οποία παρουσιάζεται στους πίνακες ως η περίπτωση όπου το κατώφλι έχει τιμή 0.

5.3. Βελτίωση Αποτελεσμάτων με Εντοπισμό Προσώπου και Σώματος

Για τη βελτίωση των αποτελεσμάτων εκμεταλλευτήκαμε μεθόδους εντοπισμού προσώπου και σώματος. Σε προηγούμενο κεφάλαιο αναλύσαμε πώς γίνεται ο εντοπισμός προσώπου. Ανάλογα με τη μέθοδο εντοπισμού προσώπου γίνεται και ο εντοπισμός σώματος. Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, αφαιρέσαμε από τις θέσεις που εντοπίσαμε ότι γίνεται αλλαγή πλάνου, αυτές, που σύμφωνα με τη μέθοδο εντοπισμού προσώπου ή σώματος, δε γίνεται αλλαγή.

Ο εντοπισμός προσώπου και σώματος μας επιστρέφει τις συντεταγμένες των πλαισίων που περιέχουν το πρόσωπο ή το σώμα αντίστοιχα. Στη συνέχεια υπολογίζεται το ιστόγραμμα χρώματος αυτού του πλαισίου. Έχοντας σα δεδομένα αυτά τα ιστογράμματα συγκρίνουμε τα πλάνα σειριακά και θεωρώντας μια τιμή κατωφλίου διατηρούμε μόνο τις θέσεις αλλαγής όπου οι αποστάσεις που προκύπτουν από τη σύγκριση των πλάνων είναι μικρότερες από αυτό το κατώφλι. Δοκιμάζουμε 51 διαφορετικές τιμές για το κατώφλι από 0 έως 0.05 με βήμα 0.01. Για κάθε τιμή του κατωφλίου λαμβάνουμε θέσεις όπου ο αλγόριθμος εντοπισμού προσώπου και σώματος εντοπίζει μεγάλη ομοιότητα. Αφαιρούμε αυτές τις θέσεις από τις θέσεις που αρχικά εντοπίσαμε ότι γίνεται εναλλαγή ομάδας όμοιων πλάνων. Με αυτό τον τρόπο αφαιρούνται οι θέσεις όπου γίνεται εσφαλμένος εντοπισμός εναλλαγής ομάδας όμοιων πλάνων. Στη συνέχεια συγκρίνουμε τις καινούριες θέσεις με τις πραγματικές θέσεις (*ground truth*) και παίρνουμε το ποσοστό επιτυχίας με βάση τις εξισώσεις 3.15, 3.16 και 3.17. Επιπλέον πρέπει να αναφέρουμε ότι αφαιρέσαμε από τα αποτελέσματά μας τις θέσεις που γινόταν εσφαλμένη αλλαγή πλάνου με τη χρήση του εντοπισμού προσώπου και ξεχωριστά τις θέσεις που γινόταν εσφαλμένη αλλαγή πλάνου με τη χρήση του εντοπισμού σώματος. Τα αποτελέσματά μας έμειναν ίδια ή βελτιώθηκαν. Καλύτερα αποτελέσματα στις περισσότερες περιπτώσεις μας δίνει η χρήση της μεθόδου εντοπισμού σώματος.

Πίνακας 5.5 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών web-81 με εντοπισμό προσώπου.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	84%	84%	87%	86%	88%
Αναπαράσταση 2	81%	81%	81%	81%	81%

Πίνακας 5.6 Αποτελέσματα χρησιμοποιώντας τους ανιχνευτές σημασιολογικών εννοιών vireo-374 με εντοπισμό προσώπου.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	83%	84%	83%	86%	87%
Αναπαράσταση 2	81%	81%	81%	81%	81%

Πίνακας 5.7 Αποτελέσματα από συνένωση των σημασιολογικών χαρακτηριστικών των ανιχνευτών σημασιολογικών εννοιών web-81 και vireo-374 με εντοπισμό σώματος.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	88%	88%	87%	88%	89%
Αναπαράσταση 2	84%	84%	84%	84%	84%

Πίνακας 5.8 Αποτελέσματα με συνδυασμό των αποστάσεων που πρόεκυψαν από τη χρήση των ανιχνευτών σημασιολογικών εννοιών web-81 και vireo-374 με εντοπισμό σώματος.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	90%	89%	90%	87%	93%
Αναπαράσταση 2	86%	86%	86%	86%	86%

Συγκεντρωτικά, τα καλύτερα αποτελέσματα ανεξάρτητα από την αναπαράσταση και τον τρόπο χρήσης των ανιχνευτών σημασιολογικών εννοιών *web-81* και *vireo-374* παρουσιάζονται στον ακόλουθο πίνακα. Ως thr ορίζουμε το κατώφλι που θέτουμε για να «κόψουμε» το διάνυσμα αποστάσεων μεταξύ διαδοχικών πλάνων, thrf (πρόσωπο) και thrb (σώμα) είναι το όριο κάτω από το οποίο αφαιρέσαμε τις αποστάσεις μεταξύ διαδοχικών πλάνων από την εφαρμογή του αλγορίθμου εντοπισμού προσώπου και σώματος αντίστοιχα, και a είναι η μετρική στην εξίσωση 3.15 όπου γίνεται συνδυασμός των αποστάσεων.

Πίνακας 5.9 Το καλύτερο αποτέλεσμα για ομαδοποίηση των πλάνων.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Web-81 Αναπαράσταση 1	84% (thr=0.06) (thrf=0.01)	84% (thr=0.06) (thrf=0.01)	87% (thr=0.06) (thrf=0.01)	86% (thr=0.06) (thrf=0.01)	88% (thr=0.04) (thrf=0.01)
Vireo-374 Αναπαράσταση 1	83% (thr=0.19) (thrb=0.02)	84% (thr=0.16) (thrb=0.02)	83% (thr=0.17) (thrb=0.02)	86% (thr=0.16) (thrb=0.02)	87% (thr=0.14) (thrb=0.02)
Συνένωση web-81 & vireo374 Αναπαράσταση 1	88% (thr=0.17) (thrb=0.02)	88% (thr=0.14) (thrb=0.02)	87% (thr=0.14) (thrb=0.02)	88% (thr=0.14) (thrb=0.02)	89% (thr=0.13) (thrb=0.02)
Συνδυασμό αποστάσεων web-81& vireo374 Αναπαράσταση 1	90% (thr=0.13) (thrb=0.02) (a=0.5)	89% (thr=0.09) (thrb=0.02) (a=0.8)	90% (thr=0.13) (thrb=0.02) (a=0.5)	87% (thr=0.13) (thrb=0.02) (a=0.5)	93% (thr=0.06) (thrb=0.02) (a=0.9)

Για να μπορέσουμε όμως να διαπιστώσουμε αν τα αποτελέσματά μας είναι τα καλύτερα τα συγκρίναμε με άλλες μεθόδους όπως με το μέσο όρο του ποσοστού επιτυχίας που προκύπτει από τα ιστογράμματα χρώματος του πλαισίου που επιστρέφει ο εντοπισμός προσώπου και σώματος.

Επιπλέον, εξάγαμε τα ιστογράμματα χρώματος των εικόνων και τα συνδυάσαμε και αυτά με τον εντοπισμό προσώπου και σώματος ώστε να τα συγκρίνουμε με τα δικά μας αποτελέσματα. Τέλος, δοκιμάσαμε να εξετάσουμε αν ο τρόπος που εντοπίζει ο αλγόριθμος *SIFT* τα όμοια πλάνα δίνει καλύτερα αποτελέσματα από τη δικά μας μέθοδο και βελτιώσαμε αυτή τη μέθοδο βάζοντας τον περιορισμό οι περιγραφείς που είναι όμοιοι να έχουν και κοντινές χωρικές συντεταγμένες.

5.4. Σύγκριση με Εντοπισμό Προσώπου και Σώματος

Για τον εντοπισμό προσώπου και σώματος επεξεργαστήκαμε τα χαρακτηριστικά εικονοπλαίσια αλλά και τις γειτονιές αυτών. Βρήκαμε με τη μέθοδό μας το μέσο όρο του ποσοστού επιτυχίας για τα 10 βίντεο ξεχωριστά για τον εντοπισμό προσώπου και σώματος αλλά και τη συνένωση αυτών.

Πίνακας 5.10 Αποτελέσματα από εντοπισμό προσώπου και σώματος.

Βήμα (*d*)	1	1	1	3	5	
Γειτονικά (*N*)	3	5	7	7	7	keyframes
Εντοπισμός προσώπου	52%	54%	55%	55%	55%	52%
Εντοπισμός σώματος	54%	54%	54%	54%	50%	52%
Συνένωση προσώπου & σώματος	54%	53%	54%	54%	50%	53%

Παρατηρούμε ότι μόνο η ανίχνευση προσώπου ή σώματος δεν είναι ικανή να ανιχνεύσει σωστά τα όρια αλλαγής ομάδων όμοιων πλάνων.

5.5. Σύγκριση με Ιστόγραμμα Χρώματος

Τα ιστογράμματα χρώματος είναι η αναπαράσταση της κατανομής των χρωμάτων σε μία εικόνα. Το σύνολο των χρωμάτων χωρίζεται σε κάδους (*bins*) που περιέχουν ένα προκαθορισμένο εύρος χρωμάτων. Ένα ιστόγραμμα χρώματος αναπαριστά την κατανομή των *pixels* της εικόνας στους κάδους αυτούς. Ένα ιστόγραμμα χρώματος μπορεί να δημιουργηθεί για οποιοδήποτε χώρο χρωμάτων, αλλά συνήθως χρησιμοποιείται στον τρισδιάστατο χώρο όπως είναι το *RGB* ή *HSV*. Στην εργασία χρησιμοποιήθηκαν κανονικοποιημένα ιστογράμματα στον χώρο χρώματος *HSV*. Δηλαδή, για κάθε εικονοπλαίσιο υπολογίζεται ένα κανονικοποιημένο ιστόγραμμα με 8 κάδους για την απόχρωση *H* (*Hue*) και από 4 κάδους για κάθε ένα από τα κορεσμός *S* (*Saturation*) και αξία *V* (*Value*). Τα τρία αυτά ιστογράμματα ενώνονται και σχηματίζουν ένα διάνυσμα διάστασης 4x4x8=128.

Το βασικότερο μειονέκτημα της μεθόδου αυτής είναι ότι περιγράφει μόνο την κατανομή χρώματος του αντικειμένου που μελετάται και αγνοεί σχήμα και επιφάνεια. Κάτι τέτοιο μπορεί να οδηγήσει σε ακριβώς ίδια ιστογράμματα χρώματος δύο αντικειμένων απλά και μόνο επειδή έχουν ίδια χρώματα. Ένα ακόμα πρόβλημα των ιστογραμμάτων χρώματος είναι η ευαισθησία τους στο θόρυβο, όπως αλλαγές στον φωτισμό.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα για τα 10 βίντεο για τις 2 πρώτες αναπαραστάσεις που μας δίνει το ιστόγραμμα χρώματος των εικονοπλαισίων (Πίνακας 5.11) και επιπλέον παρουσιάζονται τα αποτελέσματα που εξάγουμε αν συνδυάσουμε το ιστόγραμμα χρώματος με τη μέθοδο εντοπισμού προσώπου και σώματος όπως κάναμε και στα δικά μας αρχικά αποτελέσματα (Πίνακας 5.12).

Πίνακας 5.11 Αποτελέσματα από ιστόγραμμα χρώματος.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	79%	79%	78%	78%	78%
Αναπαράσταση 2	79%	79%	79%	79%	79%

Πίνακας 5.12 Αποτελέσματα από ιστόγραμμα χρώματος με εντοπισμό προσώπου και σώματος.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Αναπαράσταση 1	79%	79%	80%	80%	80%
Αναπαράσταση 2	79%	79%	79%	79%	79%

Είναι εμφανές από τον πίνακα ότι τα αποτελέσματά μας είναι πολύ καλύτερα σε σύγκριση με τα αποτελέσματα που δίνει ο αλγόριθμος που βασίζεται μόνο στα ιστογράμματα χρώματος για την ομαδοποίηση όμοιων πλάνων. Ακόμα και αν αφαιρέσουμε τις εσφαλμένες θέσεις εντοπισμού αλλαγής πλάνου σύμφωνα με τον εντοπισμό προσώπου και σώματος τα αποτελέσματα είναι χειρότερα σε σχέση με τη μέθοδό μας.

5.6. Σύγκριση με απλούς Περιγραφείς SIFT

Επειδή έχουμε έννοιες που προκύπτουν από SIFT [14] περιγραφείς, δοκιμάσαμε τη χρήση μόνο των SIFT περιγραφέων για τα χαρακτηριστικά εικονοπλαίσια (key-frames) για να εξετάσουμε αν βρίσκουμε καλύτερα αποτελέσματα. Συγκεκριμένα, υπολογίσαμε τα SIFT χαρακτηριστικά για όλα τα εικονοπλαίσια. Στη συνέχεια, τα χαρακτηριστικά κάθε εικονοπλαισίου συγκρίνονται με το επόμενό του διαδοχικά και υπολογίζεται το πλήθος των υποψήφιων χαρακτηριστικών που έχει το εικονοπλαίσιο με τον κοντινότερο γείτονά του εφαρμόζοντας την Ευκλείδεια απόσταση. Έπειτα, κανονικοποιείται αυτό το πλήθος και διατηρούμε μόνο τις μεγαλύτερες τιμές οι οποίες θεωρούνται αποστάσεις των πλάνων. Με αυτό τον τρόπο προκύπτει ένα διάνυσμα αποστάσεων μεταξύ διαδοχικών πλάνων. Ως όρια αλλαγής ομάδας όμοιων πλάνων ορίζονται εκείνες οι θέσεις που έχουν τιμή μεγαλύτερη από ένα προκαθορισμένο κατώφλι. Το κατώφλι παίρνει τιμές ανάμεσα στο 0 και στη μεγαλύτερη απόσταση που παρατηρείται σε όλα τα βίντεο St_{max}, με βημα που προκύπτει από την σχέση $St_{max}/20$. Για να βρούμε το ποσοστό επιτυχίας συγκρίνουμε τις θέσεις που εντοπίσαμε ότι γίνεται αλλαγή της ομάδας όμοιων πλάνων με τις πραγματικές θέσεις που γίνεται η αλλαγή με βάση τις εξισώσεις 3.15, 3.16 και 3.17 (Πίνακας 5.13). Αυτό τον τρόπο τον εφαρμόσαμε τόσο στα χαρακτηριστικά εικονοπλαίσια όσο και στις γειτονιές αυτών. Στη συνέχεια, βελτιώσαμε αυτό τον αλγόριθμο θεωρώντας περιορισμούς γειτνίασης στις συντεταγμένες των περιγραφέων SIFT. Δεν αρκεί να εντοπίζεται ομοιότητα με τη χρήση της Ευκλείδειας απόστασης στα διανύσματα χαρακτηριστικών των περιγραφέων SIFT αλλά πρέπει και οι συντεταγμένες αυτών να διαφέρουν κατά ένα όριο στο οποίο δώσαμε την τιμή 20 (Πίνακας 5.14). Στη τιμή του ορίου εξετάσαμε και τις τιμές 5, 10 για τις οποίες όμως τα αποτελέσματα δεν βελτιωνόταν. Λαμβάνοντας υπόψη και τις συντεταγμένες των περιγραφέων SIFT πήραμε καλύτερα αποτελέσματα και μειώσαμε τη πιθανότητα λάθους στον εντοπισμό αλλαγής πλάνου.

Πίνακας 5.13 Αποτελέσματα με περιγραφείς SIFT.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Χαρακτηριστικά εικονοπλαίσια με γειτονιές	53%	54%	54%	54%	54%
Χαρακτηριστικά εικονοπλαίσια	67%	67%	67%	67%	67%

Πίνακας 5.14 Αποτελέσματα με SIFT και περιορισμούς γειτνίασης των SIFT.

Βήμα (*d*)	1	1	1	3	5
Γειτονικά (*N*)	3	5	7	7	7
Χαρακτηριστικά εικονοπλαίσια με γειτονιές	64%	66%	64%	63%	66%
Χαρακτηριστικά εικονοπλαίσια	77%	77%	77%	77%	77%

Παρακάτω δίνουμε ένα παράδειγμα όπου φαίνεται η βελτίωση των αποτελεσμάτων που μας δίνουν οι περιγραφείς SIFT με χρήση της γειτονιάς του περιγραφέα. Στην Εικόνα 5.23 παρουσιάζονται κάποιοι περιγραφείς SIFT που έχουν ανιχνευτεί ως όμοιοι και αντίστοιχα στην Εικόνα 5.24 παρουσιάζονται κάποιοι περιγραφείς SIFT που έχουν ανιχνευτεί ως όμοιοι κάνοντας παράλληλα και έλεγχο ώστε να είναι στην ίδια γειτονιά οι συντεταγμένες των όμοιων περιγραφέων.

Εικόνα 5.23 Αποτέλεσμα από ενδεικτικά σημεία που επιστρέφει ο αλγόριθμος SIFT.

Εικόνα 5.24 Αποτέλεσμα από ενδεικτικά σημεία που επιστρέφει ο αλγόριθμος SIFT με έλεγχο στις γειτονικές συντεταγμένες.

Ακόμα και με τη βελτίωση, η χρήση μόνο των SIFT γα τη σύγκριση πλάνων δε δίνει καλύτερα αποτελέσματα από τη μέθοδο μας που χρησιμοποιεί τους σημασιολογικούς ανιχνευτές.

ΚΕΦΑΛΑΙΟ 6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΈΡΕΥΝΑ

6.1. Συμπεράσματα

Στην παρούσα εργασία μελετήθηκε το πρόβλημα της περίληψης αμοντάριστου βίντεο με τη χρήση σημασιολογικών χαρακτηριστικών. Η περίληψη αμοντάριστου βίντεο είναι πολύ σημαντική στον τομέα της επεξεργασίας βίντεο. Ιδιαίτερα τα τελευταία χρόνια, η αυτόματη και αποτελεσματική περίληψη έχει αποτελέσει σημαντικό αντικείμενο έρευνας. Αν και υπάρχουν τεχνικές με ικανοποιητικά αποτελέσματα υπάρχει δυσκολία στην ομαδοποίηση όμοιων πλάνων, δηλαδή στον εντοπισμό του σωστού σημείου εναλλαγής μεταξύ ομάδων όμοιων πλάνων. Στην εργασία αυτή προτείναμε μεθόδους για την αποτελεσματική ομαδοποίηση όμοιων πλάνων και κατά συνέπεια αποτελεσματική περίληψη βίντεο με τη χρήση σημασιολογικών χαρακτηριστικών.

Αρχικά, από κάθε βίντεο εξαγάγαμε χαρακτηριστικά εικονοπλαίσια και τις γειτονιές αυτών. Για κάθε ένα από αυτά τα εικονοπλαίσια υπολογίσαμε τους *SIFT* περιγραφείς τους και δημιουργήσαμε ιστογράμματα οπτικών λέξεων. Βασιζόμενοι σε ανιχνευτές σημασιολογικών εννοιών που είναι διαθέσιμοι στη βιβλιογραφία υπολογίσαμε σημασιολογικά χαρακτηριστικά για κάθε πλάνο βάση των οποίων έγινε η ομαδοποίηση τους σε όμοια πλάνα. Προτείναμε 12 διαφορετικές αναπαραστάσεις

περιγραφής κάθε πλάνου με βάση αυτά τα σημασιολογικά χαρακτηριστικά. Για κάθε μια αναπαράσταση συγκρίναμε διαδοχικά πλάνα, ανιχνεύσαμε αλλαγές στις ομάδες όμοιων πλάνων και έπειτα τις αξιολογήσαμε ώστε να διαπιστώσουμε ποια αναπαράσταση δίνει την καλύτερη ομαδοποίηση όμοιων πλάνων άρα και περίληψη βίντεο. Επιπρόσθετα, παρατηρήσαμε ότι η χρήση μεθόδων εντοπισμού προσώπου και σώματος βελτιώνει τα αποτελέσματά μας. Διαπιστώσαμε ότι η καλύτερη περίληψη βίντεο επιτυγχάνεται όταν κάθε χαρακτηριστικό εικονοπλαίσιο αναπαρίσταται από ένα διάνυσμα σημασιολογικών χαρακτηριστικών που προκύπτει από το μέσο όρο των σημασιολογικών χαρακτηριστικών της γειτονιάς του. Επιπλέον, τα καλύτερα αποτελέσματα της μεθόδου μας τα συγκρίναμε με αποτελέσματα που προκύπτουν από τη χρήση ιστογραμμάτων χρώματος και τη βελτίωση αυτών με τη χρήση του εντοπισμού προσώπου και σώματος, με αποτελέσματα από τον εντοπισμό προσώπου και σώματος καθώς και με αποτελέσματα από την επεξεργασία μόνο περιγραφέων SIFT και μια βελτίωση αυτών με τον έλεγχο των συντεταγμένων των περιγραφέων. Η απόδοση της μεθόδου μας υπερτερεί όλων των άλλων αποτελεσμάτων. Επομένως, η χρήση των σημασιολογικών χαρακτηριστικών σε μία γειτονιά των χαρακτηριστικών εικονοπλαισίων κάθε πλάνου προσφέρει τα καλύτερα αποτελέσματα για την ομαδοποίηση των όμοιων πλάνων και την περίληψη βίντεο. Με την προτεινόμενη μεθοδολογία επιτυγχάνεται αποτελεσματική περίληψη αμοντάριστου βίντεο.

6.2. Προτάσεις για Μελλοντική Έρευνα

Ένα ζήτημα που παρουσιάζει ενδιαφέρον για περαιτέρω έρευνα είναι αρχικά η αξιολόγηση της παραπάνω μεθόδου σε ένα αρκετά μεγαλύτερο πλήθος ακολουθιών βίντεο, προκειμένου τα αποτελέσματα να είναι περισσότερα και τα συμπεράσματα να είναι πιο αξιόπιστα. Επιπλέον, θα μπορούν να χρησιμοποιηθούν περισσότερες σημασιολογικές έννοιες ή να δοκιμαστούν γειτονιές των εικονοπλαισίων με διαφορετικό βήμα. Επίσης ένα ζήτημα έρευνας σε αμοντάριστο βίντεο σχετίζεται με την αφαίρεση εικονοπλαισίων με περιττή πληροφορία όπως για παράδειγμα εικόνες με κλακέτες.

ΑΝΑΦΟΡΕΣ

[1] Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*(MULTIMEDIA '06). ACM, New York, NY, USA, 421-430.

[2] C.-C.Chang, C.-J. Lin, "LIBSVM: a Library for Support Vector Machines", 2001

[3] V.T. Chasanis,A.C. Likas, N.P. Galatsanos, "Scene detection in videos using shot clustering and sequence alignment", IEEE Trans Multimedia,vol. 11, pp. 89-100, 2009

[4] V. Chasanis, A. Likas, and N. Galastanos. Efficient video shot summarizationusing an enhanced spectral clustering approach. In Proceedings of the 18th International Conference on Artificial Neural Networks, Part I, pp. 847-856, Prague, Czech Republic, September 2008.

[5] V. Chasanis, A. Kalogeratos, and A. Likas. Movie segmentation into scenes and chapters using locally wightened bag of visual words. In Proceedings of ACM International Conference on Image and Video Retrieval, Santorini, Greece, July 2009

[6] T.-S. Chua et al. NUS-WIDE: A real-world web image database from national university of Singapore, ACM CIVR, 2009

[7] Y.Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting, AT& Labs,1995

[8] A. Hanjalic, R. L. Lagendijk, and J. Biemond. Automated high-level movie segmentation for advanced video-retrieval systems. IEEE Transactions on Circuits and Systems for Video Technology, Vol.9, No.4. pp. 1280-1289, June 1999.

[9] Y.–G. Jiang, C.-W. Ngo, and J. Yang, VIREO-374: Keypoint-Based LSCOM Semantic Concept Detectors, 2010

[10] Y.–G. Jiang, C.-W. Ngo, and J. Yang, Towards optimal bag-of-features for object ctegorizationand semantic video retrieval, ACM International Conference on Image and Video Retrieval (CIVR'07), Amsterdam, The Netherlands, 2007.

[11] I. Koprinska and S. Carrato. Temporal video segmentation: A survey. Signal Processing: Image Communication, 16(5):477-500, January 2001.

[12] Lewis J.P.: Fast Normalized Cross Correlation. Available, 2005

[13] S.Z. Li and A.K. Jain(ed.). Handbook of Face Recognition, Springer 2005

[14] Lowe, David G. "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, Volume 60, Number 2, Pages 91–110, 2004.

[15] Mikolajczyk, K. and C. Shmid, "A Performance Evaluation of Local Descriptors," Journal IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 27, Issue 10, October 2005, pp1615–1630.

[16] P. Over, A. Smeaton, and G. Awad. The trecvid 2008 bbc rushes summarization evaluation. In TVS '08: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, pages 1-20, New York, NY, USA, 2008.

[17] C. P. Papageorgiou, M. Oren, and T. Poggio. A General Framework for Object Detection. Proceedings of IEEE International Conference on Computer Vision, 1998

[18] J. Puzicha, T. Hofmann, and J. Buhmann, "Non-Parametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval," in Proc. CVPR, 1997.

[19] Shiai Zhu, Gang Wang, Chong-Wah Ngo, and Yu-Gang Jiang. 2010. On the sampling of web images for learning visual concept classifiers. In *Proceedings of the ACM International Conference on Image and Video Retrieval* (CIVR '10). ACM, New York, NY, USA, 50-57.

[20] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in ICCV 2003

[21] S.Smoliar and H. Zhang. Content-based video indexing and retrieval. IEEE MultiMedia, 1(2)-62-72, 1994

[22] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 3d Ed. Academic Press, 2006

[23] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorization. In CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval, page 249-258, Niagara Falls, Canada, 2008

[24] A. Yanagawa, S-F. Chang, L. Kennedy, and W.Hsu, "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts", Columbia University ADVENT Technical Report #222-2006-8, Match 2007

[25] J. Yang, Y. G. Jiang, A. Hauptmann, and C.W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In MIR '07: Proceedings of the International Workshop on Multimedia Information Retrieval, pages 197-206, Augsburg, Bavaria, Germany, 2007.

[26] Yu-Gang Jiang; Yang, J.; Chong-Wah Ngo; Hauptmann, A.G., "Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study," *Multimedia, IEEE Transactions on* , vol.12, no.1, pp.42,53, Jan. 2010

[27] Vapnik, V, Cortes, C. Support Vector Networks, Machine Learning, vol. 20, no 3,pp 273-297, 1995

[28] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001

[29] P. Viola and M. Jones. Robust Real-time Object Detection. In IEEEICCV Workshop on Statistical and Computation Theories of Vision, 2001

[30] S. A. Zhu, G. Wang, C. W. Ngo, Y. G. Jiang. On the sampling of web images for learning visual concept classifiers, International Conference on Image and Video Retrieval (CIVR), Xi'an, China, July 2010.

[31] “LSCOM Lexicon Definitions and Annotations”, in DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVERT Technical Report #217-2006-3, 2006

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Η Αθηνά Παππά με καταγωγή από το Μεγάλο Περιστέρι Ιωαννίνων γεννήθηκε τον Ιούνιο του 1989. Το 2007 αποφοίτησε από το 4^{ο} Λύκειο Ιωαννίνων με βαθμό 19.4 και εισήχθη στο τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων στο οποίο αποφοίτησε με βαθμό 8.15 το 2011. Την ίδια χρονιά έγινε δεκτή στο Πρόγραμμα Μεταπτυχιακών Σπουδών του Τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων.