Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Πολυτροπική Κατάτμηση Ταινιών σε Σκηνές

Διπλωματική Εργασία

της

Ολίβιας Σ. Καραθάνου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2015

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Πολυτροπική Κατάτμηση Ταινιών σε Σκηνές

Διπλωματική Εργασία

της

Ολίβιας Σ. Καραθάνου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 20η Ιουλίου 2015.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Πέτρος Μαραγκός	Ευίτα-Σταυρούλα Φωτεινέα	Γεράσιμος Ποταμιάνος
Καθηγητής Ε.Μ.Π.	Ερευνήτρια Α' Ι.Ε.Λ.	Αναπληρωτής Καθηγητής Παν/μίου Θεσσαλίας

Αθήνα, Ιούλιος 2015

(Υπογραφή)

.....
Ολίβια Σ. Καραθάνου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Θα ήθελα κατ' αρχάς να ευχαριστήσω τον καθηγητή κ. Πέτρο Μαραγκό για την εμπιστοσύνη που μου έδειξε με την ανάθεση της παρούσας διπλωματικής εργασίας. Τον ευχαριστώ θερμότατα για το ενδιαφέρον που μου έδειξε, τις δημιουργικές συναντήσεις που είχαμε και τις χρήσιμες συμβουλές του.

Θα ήθελα, ακόμη, να ευχαριστήσω ιδιαίτερα το Νάσο Κατσαμάνη, του οποίου η καθοδήγηση ήταν καθοριστική για την πρόοδο της εργασίας μου. Ο χρόνος που μου αφιέρωσε, οι ιδέες και οι γνώσεις του με οδήγησαν με επιτυχία στην εκπόνηση αυτής της διπλωματικής.

Επιπλέον, ένα μεγάλο ευχαριστώ στον Πέτρο και τη Νάνσυ, για την πολύτιμη βοήθειά τους και τις οδηγίες τους σε διάφορα πρακτικά ζητήματα που αντιμετώπισα στο πλαίσιο της διπλωματικής, αλλά και όλα τα μέλη του εργαστηρίου για τις ευχάριστες συναντήσεις και ενδιαφέρουσες παρουσιάσεις τους.

Πολλά ευχαριστώ σε όλους μου τους φίλους, που με στηρίζουν και είναι δίπλα μου. Ιώ, Ειρήνη, Ντορίνα, σας ευχαριστώ για την ηθική υποστήριξη και την αγάπη σας. Πάνο, σ' ευχαριστώ για τις συμβουλές σου, και Χρήστο, σ' ευχαριστώ για την αγάπη και την κατανόησή σου.

Θα ήθελα, τέλος, να ευχαριστήσω την οικογένειά μου, Έλενα, Στέλιο, Νάσο, για τη στήριξή, την υπομονή και την εμπιστοσύνη που μου δείχνουν όλα αυτά τα χρόνια. Η παρουσία τους ήταν και είναι καθοριστική για την εκπλήρωση των ονείρων μου.

Περίληψη

Σκοπός της παρούσας διπλωματικής εργασίας είναι η αυτόματη κατάτμηση μιας ταινίας σε σκηνές, αξιοποιώντας την εικόνα, τον ήχο και την πληροφορία από το σενάριο της ταινίας. Πρόκειται για ένα πρόβλημα που μελετάται ευρέως και έχει ιδιαίτερο ενδιαφέρον, καθώς η κατάτμηση μιας ταινίας σε στοιχειώδεις θεματικές ενότητες αποτελεί βασικό στάδιο προεπεξεργασίας σε εφαρμογές video indexing, μη γραμμικής πλοήγησης, ταξινόμησης βίντεο κ.α. Η κατάτμηση της ταινίας σε σκηνές προϋποθέτει την κατάτμησή της σε λήψεις. Η προσέγγιση που εφαρμόζεται στο πλαίσιο της παρούσας διπλωματικής για την κατάτμηση σε λήψεις επικεντρώνεται αποκλειστικά σε χαμηλού επιπέδου χαρακτηριστικά, όπως είναι τα ιστογράμματα χρώματος και οι ακμές της εικόνας (καρέ της ταινίας). Στη συνέχεια, δοκιμάζονται υπάρχοντες αλγόριθμοι της βιβλιογραφίας για την κατάτμηση σε σκηνές, που βασίζονται είτε στην κατασκευή ενός συνεκτικού γράφου μεταβάσεων είτε στην ομαδοποίηση λήψεων με βάση τη φασματική τους ομοιότητα (Spectral Clustering). Αφού γίνει η αρχική αυτή κατάτμηση, προτείνονται τρόποι βελτίωσης του αποτελέσματος, εμπνευσμένοι από τη θεωρία πληροφορίας (Bayesian Information Criterion) ή τη γλωσσική μοντελοποίηση (Bag of Words). Στο στάδιο αυτό εισάγεται η ακουστική πληροφορία (συντελεστές MFCC) καθώς και βελτιωμένοι περιγραφητές της οπτικής πληροφορίας (GIST ή SIFT). Για την αξιοποίηση του σεναρίου, παρουσιάζεται μια μεθοδολογία για τη χρονική ευθυγράμμιση του με τους υπότιτλους, ώστε να αποδοθούν χρονικές ετικέτες σε γεγονότα και ομιλητές από το σενάριο.

Λέξεις Κλειδιά

Χρονική ευθυγράμμιση σεναρίου-υποτίτλων, κατάτμηση σε λήψεις, αντιπροσωπευτικά καρέ, κατάτμηση σε σκηνές, γράφος μεταβάσεων λήψεων, ιστόγραμμα χρώματος, λόγος αλλαγής ακμών (ECR), φασματική ομαδοποίηση, κριτήριο πληροφορίας του Bayes, Bag of Visual Words.

Abstract

The aim of this diploma thesis is to deal with the problem of multi-modal movie scene segmentation. This task is widely studied and its interest lies in the fact that segmentation of a video into fundamental semantic units is an essential pre-processing stage in applications such as video indexing, non-linear browsing, classification etc. Shot segmentation is a prerequisite for scene segmentation. Our approach focuses on low-level features, such as color histograms and edges of the image (movie frames), in order to initially segment the movie into shots. Subsequently, existing algorithms, based on the construction of a connected graph or the grouping of shots using Spectral Clustering, are tested. Initial segmentation results are further refined through our proposed methods, based on the Bayesian Information Criterion and Bag of Words techniques. At this point acoustic information is also used (MFCCs) and improved descriptors of visual information (GIST or SIFT features). To exploit information from the movie script, a temporal alignment of the subtitles and the script is performed, in order to assign temporal labels to events and speakers.

Keywords

Temporal alignment of script and subtitles, shot segmentation, Key Frames, scene segmentation, shot transition graph, color histogram, Edge Change Ratio (ECR), Spectral Clustering, Bayesian Information Criterion (BIC), Bag of Visual Words.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Σημασία της Κατάτμησης σε Σκηνές

Στην εποχή των ψηφιακών μέσων που ζούμε υπάρχει μια ραγδαία αύξηση της διαθεσιμότητας οπτικοακουστικού υλικού. Η αποτελεσματική πρόσβαση σε αυτό τον τεράστιο όγκο πληροφοριών απαιτεί τη σωστή οργάνωσή του, αλλά και εργαλεία για την πλοήγηση και την ανάκτηση σημείων ενδιαφέροντος, δηλαδή περιεχομένων που ενδιαφέρουν το χρήστη.

Ένας αρχικός τρόπος οργάνωσης αυτού του υλικού είναι ο εντοπισμός των διαφορετικών θεματικών ενοτήτων που περιέχει και η εξαγωγή μιας περιγραφής για κάθε μια από αυτές. Στην περίπτωση των ταινιών, ως θεματικές ενότητες μπορούν να θεωρηθούν οι σκηνές, η περιγραφή των οποίων λαμβάνεται από τις πληροφορίες του σεναρίου. Κάθε σκηνή μπορεί να θεωρηθεί ως μια ανεξάρτητη θεματική ενότητα, που διαδραματίζεται σε ένα συγκεκριμένο χώρο και χρόνο. Εξάγοντας τα χρονικά όρια των σκηνών και συνδυάζοντάς τα με πληροφορία (διαλόγους, περιγραφή σκηνής/προσώπων) από το σενάριο δημιουργείται μια περιγραφή της ταινίας με χρόνους έναρξης και λήξης των επιμέρους σκηνών καθώς και μια σύντομη ανάλυση του περιεχομένου τους.

Η κατάτμηση μιας ταινίας (ή ενός βίντεο) σε σκηνές μπορεί να γίνει από επισημειωτές, αλλά αυτή είναι μια χρονοβόρα και υψηλού κόστους διαδικασία. Για το λόγο αυτό, ανακύπτει η ανάγκη της αυτόματης κατάτμησης των ταινιών και του εντοπισμού των χρονικών ορίων των σκηνών τους.

1.2 Εφαρμογές

Η κατάτμηση μιας ταινίας σε σκηνές έχει πρακτικές εφαρμογές που σχετίζονται με την οργάνωση του οπτικοακουστικού υλικού και τη διευκόλυνση του χρήστη που επιθυμεί να πλοηγηθεί και να ανακτήσει δεδομένα που τον ενδιαφέρουν μέσα από το υλικό αυτό. Κάποιες από τις εφαρμογές, που αναφέρονται στη βιβλιογραφία, είναι οι ακόλουθες:

Video Indexing: Πρόκειται για την ανάθεση μιας περιγραφής σε κάθε διαθέσιμο βίντεο. Η κατάτμηση σε σκηνές οδηγεί σε μικρότερης χρονικής διάρκειας αποσπάσματα, κοινού θεματικού περιεχομένου, τα οποία μπορούν να περιγραφούν σύντομα και περιεκτικά με χρήση κάποιων λέξεων-κλειδιών.

Video Retrieval: Εφόσον έχει γίνει η περιγραφή κάθε αποσπάσματος, μπορεί εύκολα να γίνει και η ανάκτησή του (retrieval), με μια απλή αναζήτηση των λέξεων-κλειδιών. Με αυτό τον τρόπο, μάλιστα, σκηνές διαφορετικών ταινιών μπορούν να συσχετιστούν μεταξύ τους. Έτσι, αναζητώντας, για παράδειγμα, σκηνές με χορό, θα εμφανιστούν αποσπάσματα από πολλές διαφορετικές ταινίες στα οποία έχει ανατεθεί η ετικέτα "χορός".

Non-linear Browsing: Η προσπέλαση μιας ταινίας μπορεί να γίνει γραμμικά, δηλαδή προσπερνώντας κάθε καρέ της ταινίας μέχρι να εντοπίσουμε αυτό που μας ενδιαφέρει, ή μη γραμμικά, εντοπίζοντας απευθείας τα σημεία ενδιαφέροντος. Η κατάτμηση μιας ταινίας σε σκηνές εξυπηρετεί στη γρηγορότερη και ευκολότερη πλοήγησή μας σε αυτή, καθώς βασίζεται στο περιεχόμενο και όχι τη χρονική ακολουθία των καρέ.

Video Classification: Η ταξινόμηση των αποσπασμάτων των ταινιών μπορεί εύκολα να γίνει, εφόσον έχει προηγηθεί η κατάτμηση σε σκηνές. Οι ετικέτες που έχουν ανατεθεί σε κάθε απόσπασμα (π.χ. σκηνή διαλόγου, δράσης, εξωτερικός/εσωτερικός χώρος κ.α.) συμβάλλουν στην ομαδοποίηση σκηνών με κοινά στοιχεία και στην ταξινόμησή τους.

Video Summarization: Η αυτόματη εξαγωγή της περίληψης μιας ταινίας συνίσταται στη δημιουργία ενός σύντομου βίντεο που περιέχει όλες τις σημαντικές πληροφορίες του αρχικού αποσπάσματος, διατηρώντας, όμως, τον αισθητικό σκοπό του. Για να είναι πλήρης η περίληψη καλό είναι να περιέχει αποσπάσματα από όλες (ή τουλάχιστον τις περισσότερες) σκηνές της ταινίας. Για το λόγο αυτό η κατάτμηση σε

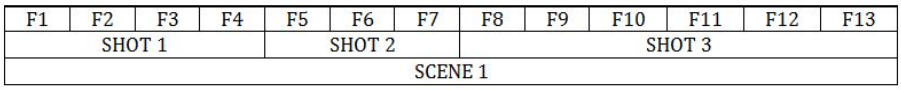
σκηνές μπορεί να θεωρηθεί ως ένα στάδιο προ επεξεργασίας πριν την εξαγωγή της περίληψης.

Σε όλες τις παραπάνω εφαρμογές πρέπει να τονιστεί η σημασία της εισαγωγής της πληροφορίας από το σενάριο, για την περιγραφή των επιμέρους σκηνών. Πέρα από οπτικά και ακουστικά χαρακτηριστικά, που μπορούν να αναγνωριστούν απευθείας από το βίντεο (π.χ. πρόσωπα, ομιλητές, κίνηση, αντικείμενα), είναι σημαντικό κάθε σκηνή να λαμβάνει μια περιγραφή από το σενάριο, η οποία να είναι λεπτομερής, όσον αφορά τα πρόσωπα, την τοποθεσία, το χρόνο διεξαγωγής, αλλά και να συμπληρώνεται από στοιχεία που μπορεί να αντιληφθεί ο θεατής και περιγράφονται στο σενάριο.

1.3 Περιγραφή του Προβλήματος

1.3.1 Δομή μιας Ταινίας

Μια ταινία αποτελείται από την αλληλουχία κάποιων καρέ (frames), η οποία εμφανιζόμενη στην οθόνη με την κατάλληλη συχνότητα δημιουργεί την ψευδαίσθηση της κινούμενης εικόνας. Μια χρονικά αδιάκοπη σειρά από καρέ που έχουν προέλθει από την ίδια κάμερα ορίζουν μια λήψη (shot). Ο συνδυασμός κάποιων συνεχόμενων λήψεων, με κοινό θεματικό περιεχόμενο, ορίζουν μια σκηνή (scene). Τα τρία αυτά στοιχεία - καρέ, λήψεις, σκηνές - συγκροτούν μια ταινία και απεικονίζονται στο Σχήμα 1.1.



Σχήμα 1.1: Δομή μιας ταινίας

1.3.2 Διατύπωση του Προβλήματος

Με βάση την προηγούμενη περιγραφή μπορούμε να διασπάσουμε το πρόβλημα της κατάτμησης σε σκηνές σε δύο υποπροβλήματα. Το πρώτο είναι η κατάτμηση σε λήψεις και το δεύτερο η σωστή ομαδοποίηση των λήψεων, ώστε να συγκροτούν μια σκηνή. Λήψεις που διεξάγονται στον ίδιο χώρο και χρόνο ενώνονται σε μια μεγαλύτερη ενότητα και αποτελούν τις σκηνές. Καθένα από τα δύο αυτά προβλήματα θα περιγραφεί αναλυτικά στα κεφάλαια που ακολουθούν.

	Title					
	BM	T-II	TG	G-60	GE	SF
Duration(min)	36	55	50	58	60	18
Shots	219	994	754	1332	879	245
RESULTS WITH THE HUMAN GENERATED GROUND TRUTH						
Scenes in ground truth	18	36	23	39	25	28
Scenes Detected	28	39	26	57	44	27
Correct Detection	15	27	18	29	22	23
False negative	3	9	5	10	3	5
False positive	13	12	8	28	22	4
Recal(%)	83	75	78	74	88	82
Precision(%)	54	69	69	51	50	85

Πίνακας 1.1: Αποτελέσματα μεθόδου Normalized Cuts, από το [20].

1.4 Σχετική Βιβλιογραφία

Στη βιβλιογραφία έχουν προταθεί διάφορες προσεγγίσεις για το πρόβλημα της κατάτμησης μιας ταινίας σε σκηνές. Μια κατηγορία μεθόδων χρησιμοποιεί αποκλειστικά οπτικά χαρακτηριστικά για να αντιμετωπίσει το πρόβλημα της κατάτμησης.

Στο [28] κατασκευάζεται ένας γράφος μεταβάσεων, που αναπαριστά την αλληλουχία των λήψεων. Ο γράφος αυτός χωρίζεται σε ανεξάρτητα υπογραφήματα, τα οποία αποτελούν τις τελικές σκηνές.

Μια παρόμοια μέθοδος ακολουθείται και στο [20], όπου το πρόβλημα της κατάτμησης ανάγεται σε ένα πρόβλημα διαμερισμού γράφων. Κάθε κόμβος ενός γράφου αναπαριστά μια λήψη και οι ακμές του γράφου δείχνουν την ομοιότητα μεταξύ διαφορετικών λήψεων, υπολογισμένη με βάση την πληροφορία του χρώματος και της κίνησης. Ενδεικτικά κάποια αποτελέσματα της μεθόδου φαίνονται στον Πίνακα 1.1 [1].

Στο [19] παρουσιάζεται ένας αλγόριθμος δύο περασμάτων, ο οποίος χρησιμοποιεί τις πληροφορίες χρώματος και κίνησης και τη διάρκεια των λήψεων για να εντοπίσει τα χρονικά όρια των σκηνών. Αρχικά, εντοπίζει πιθανές αλλαγές σκηνής, βασιζόμενος στη χρωματική ομοιότητα μιας λήψης με τις προηγούμενές της. Στη συνέχεια, απορρίπτονται κάποιες από τις εντοπισμένες σκηνές με βάση δυναμικά χαρακτηριστικά, όπως η κίνηση. Ενδεικτικά, κάποια αποτελέσματα, στην ίδια βάση ταινιών με την

[1] Η βάση ταινιών που αναφέρεται στον Πίνακα 1.1 είναι: A Beautiful Mind (BM), Terminator II (T-II), Top Gun (TG), Gone in 60 sec (G-60), Golden Eye (GE), Seinfeld (SF)

Video	Duration (min)	# Shots	G.Truth Scenes	Detected Scenes	False −ve	False +ve	Recall (%)	Precision (%)
Terminator 2	55	1632	36	38	5	7	86.1	81.6
Golden Eye	60	1519	25	35	3	13	88.0	62.9
Gone in 60 sec	58	1869	39	43	6	10	84.6	76.7
TopGun	50	1103	26	30	3	7	88.5	76.7
A Beautiful Mind	36	446	17	21	2	6	88.2	71.4
Seinfeld	21	318	22	27	3	8	86.4	70.0

Πίνακας 1.2: Αποτελέσματα μεθόδου αλγορίθμου δύο περασμάτων, από το [19].

Measures	Gone in 60 sec	Dr. No - 007	Mummy Returns
Length	01:46:09	01:30:55	01:45:33
Num. of Shots	2237	677	1600
Num. of Scenes	29	17	18
Detected Scenes	25	20	18
Match	24	14	15
Insertion	1	3	3
Deletion	5	6	3
Precision(%)	96.0	70.0	83.3
Recall(%)	82.8	82.4	83.3

Πίνακας 1.3: Αποτελέσματα μεθόδου με χρήση Μαρκοβιανής αλυσίδας Monte Carlo, από το [30].

προηγούμενη μέθοδο, φαίνονται στον Πίνακα 1.2.

Στο [30] οι συγγραφείς προτείνουν τη χρήση της Μαρκοβιανής αλυσίδας Monte Carlo για τον εντοπισμό των χρονικών ορίων των σκηνών. Η αρχική τυχαία κατάτμηση (σε αυθαίρετο αριθμό σκηνών) ανανεώνεται σε κάθε πέρασμα με χρήση δύο διαδικασιών, που ονομάζονται διαχύσεις (diffusions) και άλματα(jumps). Οι διαχύσεις απλά μετακινούν τα υπάρχοντα όρια, ενώ τα άλματα είτε προσθέτουν νέα όρια, χωρίζοντας μια υπάρχουσα σκηνή σε δύο, είτε ενώνουν δύο υπάρχουσες σκηνές. Τα αποτελέσματα της μεθόδου φαίνονται στον Πίνακα 1.3.

Οι περισσότερες προαναφερθείσες μέθοδοι λαμβάνουν υπόψιν τη χρονική απόσταση των λήψεων για να εξάγουν ένα μέγεθος ομοιότητάς τους. Κάτι τέτοιο δεν συμβαίνει στο [4], όπου οι λήψεις ομαδοποιούνται χρησιμοποιώντας τη μεθοδολογία της φασματικής ομαδοποίησης (Spectral Clustering). Σε κάθε λήψη ανατίθεται μια ετικέτα, ανάλογα με το cluster στο οποίο ανήκει και ακολουθίες ετικετών συγκρίνονται, ώστε να εντοπιστούν αλλαγές στα μοτίβα διαδοχικών ετικετών. Οι αλλαγές αυτές λαμβάνονται ως

Video	Recall(%)	Precision(%)	$F_1(\%)$
V_1	86.67	92.85	89.70
V_2	100.00	90.00	94.74
V_3	87.50	73.68	80.00
V_4	76.92	83.33	80.00
V_5	85.72	92.31	88.89
V_6	82.35	93.33	87050
V_7	86.67	81.25	83.87
V_8	76.00	95.00	84.44
V_9	72.00	75.00	73.43
V_{10}	70.00	93.33	84.26
Mean	82.38	87.01	84.26
Standard Deviation	8.46	7.64	5.85

Πίνακας 1.4: Αποτελέσματα μεθόδου εντοπισμού επαναλαμβανόμενων προτύπων, από το [4].

	Movie M_1			Movie M_2			Movie M_3		
Method	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1
sift10	83.33	83.33	83.33	86.67	83.33	78.79	77.03	71.25	74.03
sift20	83.33	83.33	83.33	77.78	83.33	72.92	81.08	68.97	74.53
sift50	88.89	84.21	86.49	82.22	84.21	75.51	81.08	70.59	75.47
sift100	88.89	88.89	88.89	82.22	88.89	78.72	82.43	69.32	75.31
sift200	83.33	88.24	85.71	80.00	88.24	78.26	87.84	73.03	79.75
sift500	88.89	88.89	88.89	91.11	88.89	87.23	82.43	76.62	77.22

Πίνακας 1.5: Αποτελέσματα μεθόδου Bag of Visual words, από το [3].

αλλαγές σκηνής. Στον Πίνακα 1.4 φαίνονται τα αποτελέσματα της μεθόδου σε μια βάση 10 βίντεο από τηλεοπτικές σειρές και ταινίες.

Στο [3] για κάθε λήψη εξάγεται ένα ιστόγραμμα οπτικών λέξεων και γίνεται σύγκριση των διαδοχικών ιστογραμμάτων, ώστε να βρεθούν οι αλλαγές σκηνής, ως τα σημεία στα οποία διαδοχικά ιστογράμματα παρουσιάζουν μεγάλη διαφορά. Οι οπτικές λέξεις προέρχονται από την ομαδοποίηση των περιγραφητών SIFT των καρέ όλων των λήψεων. Στον Πίνακα 1.5 φαίνονται τα αποτελέσματα της μεθόδου για τρεις διαφορετικές ταινίες [2] ως προς το μέγεθος του οπτικού λεξιλογίου.

Μια δεύτερη κατηγορία μεθόδων εισάγει στη διαδικασία της κατάτμησης και την ακουστική πληροφορία. Στο [23] προτείνονται δύο τρόποι για

[2]Οι ταινίες στις οποίες γίνεται ο πειραματισμός είναι: A Beautiful Mind (M_1), Sex and the City (M_2), Gone in 60 seconds (M_3)

Method	VSTG	SASTG	AVSTG
Coverage(%)	80.98	82.44	88.78
Overflow(%)	12.09	8.47	9.56

Πίνακας 1.6: Αποτελέσματα μεθόδου SASTG και AVSTG, από το [23].

την αξιοποίηση της ακουστικής πληροφορίας. Ο πρώτος (SASTG-Speaker Assisted Scene Transition Graph) βασίζεται στην αρχική οπτική κατάτμηση και απορρίπτει με βάση τα ακουστικά χαρακτηριστικά κάποιες από τις εντοπισμένες σκηνές. Ο δεύτερος (AVSTG-Audio Visual Scene Transition Graph) κατασκευάζει παράλληλα με τον οπτικό γράφο και έναν ακουστικό και συνδυάζει τα αποτελέσματα της κατάτμησης που προκύπτουν και από τους δύο. Τα αποτελέσματα της μεθόδου σε σχέση με τον απλό γράφο μεταβάσεων (VSTG-Visual Scene Transition Graph, [28]) φαίνονται στον Πίνακα 1.6. Τα κριτήρια αξιολόγησης που χρησιμοποιούνται (Overflow και Coverage) ορίζονται στο [26].

1.5 Βάση Δεδομένων

Η βάση των ταινιών που χρησιμοποιείται είναι η Movie Summarization (MovSum) Database [6], η οποία αποτελείται από τριαντάλεπτα αποσπάσματα επτά ταινιών. Τα αποτελέσματα της παρούσας εργασίας ελέγχθηκαν συγκρίνοντάς τα με την επισημείωση που έχει γίνει από μέλη του εργαστηρίου πάνω στη συγκεκριμένη βάση.

Οι ταινίες που περιλαμβάνονται στη βάση είναι οι ακόλουθες:

- **Beautiful Mind (BMI)** (2001) βιογραφία, δράμα.
- **Chicago (CHI)** (2002) μιούζικαλ.
- **Crash (CRA)** (2004) δράμα.
- **The Departed (DEP)** (2006) δράμα, θρίλερ.
- **Finding Nemo (FNE)** (2003) κινούμενα σχέδια, περιπέτεια.
- **Gladiator (GLA)** (2000) δράσης, δράμα.
- **Lord of The Rings (LOR)** (2003) φαντασίας, περιπέτειας.

Η μορφή των αρχείων είναι .avi (Xvid κωδικοποίηση) με frame rate 25fps και συχνότητα δειγματοληψίας του ακουστικού καναλιού 44100Hz. Η

	Duration(min)	Total Frames
BMI	31:17	46937
CHI	30:08	45202
CRA	26:37	39926
DEP	30:28	45707
FNE	30:17	45440
GLA	30:02	45062
LOR	37:33	56339

Πίνακας 1.7: Διάρκεια ταινιών Βάσης σε λεπτά και συνολικός αριθμός καρέ.

	Number of Shots	Mean Duration (frames)	Max Duration (frames)	Min Duration (frames)	Mean Number Shots per Scene	Max Number Shots per Scene	Min Number Shots per Scene
BMI	473	99	901	10	40	81	9
CHI	698	64	723	5	44	182	1
CRA	372	107	1710	1	34	139	2
DEP	575	79	1111	7	26	77	3
FNE	482	94	1150	1	69	120	3
GLA	585	77	647	3	59	384	1
LOR	663	84	801	6	29	93	2

Πίνακας 1.8: Στατιστικά Λήψεων για τη Βάση των Ταινιών

διάρκειά τους σε λεπτά και καρέ αναγράφεται στον Πίνακα 1.7, ενώ κάποια στατιστικά τους φαίνονται στους Πίνακες 1.8 και 1.9.

Επιπλέον, σε ορισμένα σημεία χρησιμοποιήθηκε για την αξιολόγηση των αποτελεσμάτων η ταινία **Gone With The Wind** (1939). Έτσι τα αποτελέσματα που αναφέρονται αποκλειστικά σε αυτή θα αναγράφονται με εμφανή τρόπο.

Η επισημείωση της βάσης έγινε με χρήση του λογισμικού ANVIL[3] [11] και για κάθε ταινία υπάρχουν δύο ξεχωριστά αρχεία, ένα για την κατάτμηση σε λήψεις και ένα για την κατάτμηση σε σκηνές. Κάθε τέτοιο αρχείο περιλαμβάνει ένα πίνακα με τα καρέ έναρξης και λήξης κάθε λήψης/σκηνής.

1.6 Σκοπός της Διπλωματικής

Σκοπός της παρούσας διπλωματικής εργασίας είναι να παρουσιάσει μεθόδους κατάτμησης ταινιών σε σκηνές και να αξιολογήσει τα

[3] `http://www.anvil-software.org/`

	Number of Scenes	Mean Duration (frames)	Max Duration (frames)	Min Duration (frames)
BMI	12	3911	9258	1345
CHI	16	2825	9298	63
CRA	11	3629	10234	467
DEP	22	2077	8931	144
FNE	7	6491	11270	290
GLA	10	4506	23515	216
LOR	23	2449	7994	81

Πίνακας 1.9: Στατιστικά Σκηνών για τη Βάση των Ταινιών

αποτελέσματα που κάθε μέθοδος δίνει στη βάση των 7 ταινιών. Επιχειρείται να βρεθούν οι κατάλληλες παράμετροι που θα δίνουν καλά αποτελέσματα για όλες τις ταινίες της βάσης και, άρα, θα επιτρέψουν τη γενίκευση και την εφαρμογή αυτών των μεθόδων (με τις συγκεκριμένες παραμέτρους) σε οποιαδήποτε ταινία.

Κατά τεκμήριο, οι υπάρχουσες μέθοδοι της βιβλιογραφίας οδηγούν σε μια υπερκατάτμηση των ταινιών και για το λόγο αυτό κρίνεται απαραίτητο να γίνει μια βελτίωση αυτού του αποτελέσματος. Στο πλαίσιο της διπλωματικής προτείνονται και εφαρμόζονται κάποιες μέθοδοι με αυτό το σκοπό. Χρησιμοποιώντας το ακουστικό κανάλι καταφέρνουμε κάποια ικανοποιητικά αποτελέσματα κατάτμησης, ενώ δοκιμάζουμε και κάποιες μεθόδους που βασίζονται αποκλειστικά σε οπτικά χαρακτηριστικά ή συνδυασμό οπτικοακουστικής πληροφορίας.

1.7 Οργάνωση του Περιεχομένου της Διπλωματικής

Το περιεχόμενο της διπλωματικής είναι οργανωμένο σε κεφάλαια ως εξής:

- Στο **Κεφάλαιο 2** αναλύεται η τεχνική χρονικής ευθυγράμμισης του σεναρίου με τους υπότιτλους. Η διαδικασία αυτή είναι απαραίτητη για την εισαγωγή χρονικής πληροφορίας στα αποσπάσματα του σεναρίου. Παρουσιάζονται τα εργαλεία και οι αλγόριθμοι που χρησιμοποιούνται για το σκοπό αυτό, καθώς και κάποια παραδείγματα των αποτελεσμάτων, όπως αυτά προέκυψαν για το σενάριο και τους υπότιτλους της ταινίας Gone With The Wind.
- Στο **Κεφάλαιο 3** αναλύονται οι υπάρχουσες μέθοδοι κατάτμησης

μιας ταινίας σε λήψεις. Αναλύονται κατά κύριο λόγο μέθοδοι που βασίζονται σε χαρακτηριστικά χαμηλού επιπέδου, όπως ιστογράμματα χρώματος και ακμές, και παρουσιάζονται τα αποτελέσματα αυτών των μεθόδων στη βάση των 7 ταινιών. Επιπλέον, εισάγεται η έννοια των Key Frames, δηλαδή των αντιπροσωπευτικών καρέ κάθε λήψης. Περιγράφονται συνοπτικά οι μέθοδοι εξαγωγής των αντιπροσωπευτικών καρέ και δίνονται κάποιοι δείκτες αξιολόγησής τους.

- Στο **Κεφάλαιο 4** αναλύονται οι υπάρχουσες μέθοδοι κατάτμησης μιας ταινίας σε σκηνές. Χρησιμοποιούνται τα αποτελέσματα του προηγούμενου κεφαλαίου, με στόχο να γίνει μια σωστή ομαδοποίηση των λήψεων σε σκηνές. Αναλύονται μέθοδοι που βασίζονται σε τεχνικές κατάτμησης γράφων, φασματικής ομαδοποίησης (Spectral Clustering), αλλά και τεχνικές εμπνευσμένες από την επεξεργασία κειμένου (Bag of Visual Words). Τέλος, παρουσιάζονται τα αποτελέσματα της κατάτμησης για τη βάση των 7 ταινιών.

- Στο **Κεφάλαιο 5** επιχειρείται να γίνει μια βελτίωση των αποτελεσμάτων της κατάτμησης σε σκηνές, ώστε να εντοπίζεται μικρότερος αριθμός σκηνών, που να αντιστοιχεί σε ευρύτερες θεματικές ενότητες. Αυτό επιτυγχάνεται με χρήση είτε της ακουστικής πληροφορίας (συντελεστές MFCC) είτε της οπτικής. Προτείνονται διάφοροι τρόποι χρήσης του κριτηρίου πληροφορίας του Bayes (BIC) για την επίτευξη αυτού του σκοπού και παρουσιάζονται υπάρχουσες μέθοδοι. Επιπλέον, δοκιμάζονται οι περιγραφητές SIFT και GIST για τη βελτίωση του αποτελέσματος. Τέλος, εξάγονται τα αντίστοιχα πειραματικά αποτελέσματα.

- Στο **Κεφάλαιο 6** παρουσιάζονται τα συμπεράσματα που προκύπτουν από το σύνολο της διπλωματικής και συνοψίζονται οι επιστημονικές συνεισφορές της. Επίσης, αναφέρονται και κάποιες μελλοντικές κατευθύνσεις και προεκτάσεις της.

Κεφάλαιο 2

Χρονική Ευθυγράμμιση Σεναρίου

2.1 Γενικά

Οι περισσότερες εφαρμογές της κατάτμησης των ταινιών σε σκηνές συνδέονται άμεσα με την ανάθεση μιας περιγραφής σε κάθε σκηνή, όπως αυτή δίνεται από το σενάριο. Επομένως, είναι απαραίτητο να αντληθεί αυτή η πληροφορία από το σενάριο και με τρόπο εύκολο και άμεσο να μπορεί να αντιστοιχηθεί κάθε απόσπασμα του σεναρίου σε ένα τμήμα της ταινίας.

Το σενάριο, γενικά, περιέχει μια αναλυτική περιγραφή της πλοκής και της εξέλιξης της ταινίας, με επισήμανση της τοποθεσίας στην οποία εκτυλίσσεται μια σκηνή, των ηθοποιών που εμφανίζονται, των διαλόγων που πραγματοποιούνται και των γεγονότων που λαμβάνουν χώρα, μπροστά και πίσω από την κάμερα. Το πρόβλημα έγκειται στο γεγονός οτι στο σενάριο δεν αναφέρονται οι χρόνοι κατά τους οποίους εκτυλίσσονται τα γεγονότα. Η μόνη πηγή γραπτής πληροφορίας, η οποία περιέχει το στοιχείο του χρόνου είναι οι υπότιτλοι, οι οποίοι περιέχουν αποκλειστικά τους διαλόγους μεταξύ των ηθοποιών, καθώς και το χρονικό διάστημα στο οποίο οι υπότιτλοι εμφανίζονται στην οθόνη.

Στόχος μας είναι η εισαγωγή της χρονικής πληροφορίας των υποτίτλων στο πλούσιο σε πληροφορία σενάριο, ώστε κάθε τμήμα του σεναρίου να έχει ένα χρόνο έναρξης και λήξης και να μπορεί να γίνει η αντιστοίχιση τμημάτων του σεναρίου, με όλες τις λεπτομέρειες που αυτά περιέχουν, σε αποσπάσματα της ταινίας. Για να επιτευχθεί αυτό χρειάζεται να αντιστοιχηθούν οι χρόνοι των υποτίτλων στους διαλόγους του σεναρίου. Επομένως, πρέπει πρώτα να εντοπισθούν οι διάλογοι στο σενάριο και στη συνέχεια να ευθυγραμμιστούν με τους υπότιτλους.

Στο κεφάλαιο αυτό θα παρουσιαστεί, αρχικά, ο τρόπος εντοπισμού και εξαγωγής των διαλόγων από το σενάριο. Στη συνέχεια, θα αναλυθεί

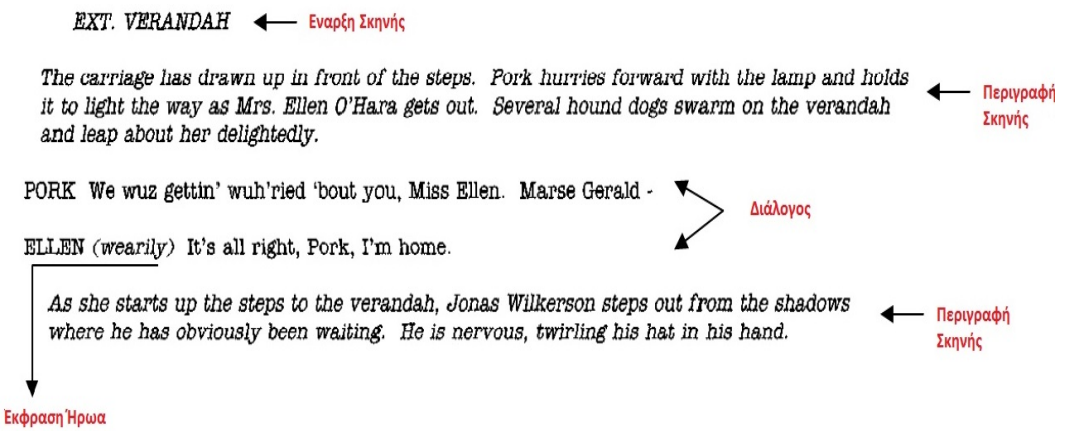
ο Dynamic Time Warping (DTW) αλγόριθμος, ο οποίος χρησιμοποιείται για να επιτευχθεί η αντιστοίχιση των λέξεων των διαλόγων με τις λέξεις των υποτίτλων, και, τέλος, θα παρουσιαστούν τα αποτελέσματα της ευθυγράμμισης.

2.2 Επεξεργασία Σεναρίου

2.2.1 Δομή Σεναρίου

Ο τρόπος γραφής ενός σεναρίου είναι αυστηρά ορισμένος και έτσι επιτρέπει την επεξεργασία του για την εξαγωγή διάφορων αποσπασμάτων, όπως περιγραφή μιας σκηνής, διάλογοι και θέσεις λήψεων της κάμερας.

Η έναρξη μιας νέας σκηνής υποδηλώνεται με τα αναγνωριστικά EXT ή INT, τα οποία υποδεικνύουν αν η σκηνή εκτυλίσσεται σε εξωτερικό ή εσωτερικό χώρο, αντίστοιχα. Στο ίδιο σημείο μπορεί να αναγράφεται η ονομασία της τοποθεσίας ή/και ο χρόνος στον οποίο εκτυλίσσεται η σκηνή [25]. Στη συνέχεια, γίνεται μια σύντομη περιγραφή του χώρου, της σκηνής και των νέων χαρακτήρων που εμφανίζονται. Στο σώμα του σεναρίου κυρίαρχο ρόλο παίζουν οι διάλογοι. Κάθε διάλογος ξεκινάει αναγράφοντας το όνομα του ομιλητή με κεφαλαία γράμματα και, έπειτα, τα λεγόμενά του, ανάμεσα στα οποία πολλές φορές παρεμβάλλονται παρενθέσεις, όπου επεξηγούνται κάποιες κινήσεις ή εκφράσεις του ήρωα. Ένα απόσπασμα σεναρίου φαίνεται στο Σχήμα 2.1. Ανάμεσα στους διαλόγους βρίσκονται επιπρόσθετες περιγραφές της σκηνής και διευκρινίσεις για τις ενέργειες των ηθοποιών ή για περιστατικά που συμβαίνουν εκτός κάμερας.



Σχήμα 2.1: Παράδειγμα δομής σεναρίου, Gone With The Wind

2.2.2 Εξαγωγή Διαλόγων

Ο εντοπισμός των διαλόγων στο σενάριο αποτελεί βασικό στοιχείο της επεξεργασίας. Για την επίτευξή του εντοπίζονται, αρχικά, οι γραμμές του σεναρίου που ξεκινούν με μια λέξη με κεφαλαία γράμματα (εξαιρούνται κάποιες δεσμευμένες λέξεις όπως CAMERA, EXT, INT και άλλα). Αφού εντοπιστούν με αυτό τον τρόπο τα ονόματα των ομιλητών, διατηρούνται οι γραμμές που τα ακολουθούν, μέχρι να εμφανιστεί κενή σειρά. Τα τμήματα που διατηρούνται θεωρούνται τα λεγόμενα του συγκεκριμένου ομιλητή. Σε αυτή τη διαδικασία παραβλέπονται οι τυχόν παρενθέσεις που παρεμβάλλονται. Έτσι, δημιουργείται ένας πίνακας $N \times 2$, όπου N είναι το πλήθος των διαλόγων που εντοπίστηκαν. Κάθε στοιχείο $(i, 1)$ του πίνακα περιέχει το όνομα του i-οστού ομιλητή, ενώ το στοιχείο $(i, 2)$ τα λεγόμενά του.

Πολλές φορές κατά τη διαδικασία αυτή εντοπίζονται λανθασμένα κάποιες λέξεις ως ονόματα ομιλητών. Για παράδειγμα, στο Σχήμα 2.2 η λέξη EXPLOSION εντοπίζεται λανθασμένα ως όνομα ομιλητή. Είναι εφικτό, για τη λήψη ακόμη καλύτερων αποτελεσμάτων, να γίνει ένας έλεγχος των ονομάτων των ομιλητών που έχουν εντοπισθεί και όσα θεωρούνται λανθασμένα να προστεθούν στη λίστα των δεσμευμένων λέξεων. Με την επανάληψη της διαδικασίας, χρησιμοποιώντας το νέο σύνολο δεσμευμένων λέξεων, θα έχουν διατηρηθεί μόνο τα σωστά ονόματα. Έτσι, σε ο,τι αφορά στο απόσπασμα του Σχήματος 2.2 η λέξη EXPLOSION θα έχει ενταχθεί στο σύνολο των δεσμευμένων λέξεων και δεν θα αναγνωρίζεται πλέον ως όνομα ομιλητή.

AUNT PITTYPAT *(coming down steps quickly)* I can't bear it! Those cannonballs right in my ears! I'll faint every time I hear one.

EXPLOSION
Aunt Pittypat closes her eyes and rocks - opens them and looks at Uncle Peter.

AUNT PITTYPAT Uncle Peter - look out for that trunk!

Σχήμα 2.2: Λανθασμένος εντοπισμός ομιλητή

2.3 Επεξεργασία Υποτίτλων

Η δομή των υποτίτλων είναι αρκετά πιο απλή από αυτή του σεναρίου. Όπως φαίνεται στο Σχήμα 2.3 κάθε τμήμα των υποτίτλων χαρακτηρίζεται από έναν αύξοντα αριθμό, τους χρόνους εμφάνισής και απομάκρυνσής από την οθόνη, καθώς και το κύριο σώμα των διαλόγων.

```
109
00:13:16,295 --> 00:13:18,505
We was worried about you, Miss Ellen.

110
00:13:18,756 --> 00:13:21,090
It's all right, Pork. I'm home.
```

Σχήμα 2.3: Δομή υποτίτλων

Η επεξεργασία τους περιορίζεται στη δημιουργία ενός πίνακα ο οποίος περιέχει στην πρώτη στήλη τους χρόνους έναρξης σε sec, στη δεύτερη στήλη τους χρόνους λήξης σε sec και στην τρίτη τους διαλόγους που πραγματοποιούνται σε εκείνο το χρονικό διάστημα, όπως φαίνεται και στον Πίνακα 2.1, ο οποίος αφορά στους διαλόγους του Σχήματος 2.3.

	Start Time sec	**End Time sec**	**Subtitle**
109	796.295	798.505	we was worried about you miss ellen
110	798.756	801.09	its all right pork im home

Πίνακας 2.1: Αποτέλεσμα επεξεργασίας των υποτίτλων

Κατά την επεξεργασία, τόσο των υποτίτλων όσο και του σεναρίου, αφαιρούνται οποιαδήποτε σημεία στίξης και όλα τα γράμματα μετατρέπονται από κεφαλαία σε μικρά.

2.4 Ευθυγράμμιση Σεναρίου-Υποτίτλων

Η συγγραφή ενός σεναρίου προηγείται συνήθως των γυρισμάτων της ταινίας, για το λόγο αυτό πολλές φορές οι διάλογοι του σεναρίου δεν ταυτίζονται απόλυτα με αυτούς των υποτίτλων. Επίσης, η προφορά

κάποιων λέξεων επηρεάζει το πώς αυτές καταγράφονται είτε στο σενάριο είτε στους υπότιτλους, με αποτέλεσμα η ίδια λέξη να εμφανίζεται γραμμένη με δύο διαφορετικούς τρόπους, για παράδειγμα "*We wuz gettin' wuh'ried*" στο σενάριο (Σχήμα 2.1) και "*We was worried*" στους υπότιτλους (Σχήμα 2.3). Ένας Dynamic Time Warping (DTW) αλγόριθμος χρησιμοποιείται για να εξαλειφθούν οι "ασυνέπειες" μεταξύ λέξεων των υποτίτλων και των διαλόγων του σεναρίου και να γίνει ευθυγράμμιση των δύο.

2.4.1 Ο αλγόριθμος DTW

Ο αλγόριθμος DTW χρησιμοποιείται για την εύρεση μιας βέλτιστης ευθυγράμμισης μεταξύ δύο χρονικών ακολουθιών [14], σε εφαρμογές όπως αναγνώριση φωνής και αναγνώριση ομιλητή. Πραγματοποιείται σύγκριση των δύο ακολουθιών $X := (x_1, x_2, \ldots, x_N)$ και $Y := (y_1, y_2, \ldots, y_M)$ μήκους $N \in \mathbb{N}$ και $M \in \mathbb{N}$, αντίστοιχα, με $x_n, y_m \in \mathcal{F}$, όπου $\mathcal{F}$ είναι ο χώρος χαρακτηριστικών, $n \in [1, \ldots, N]$ και $m \in [1, \ldots, M]$.

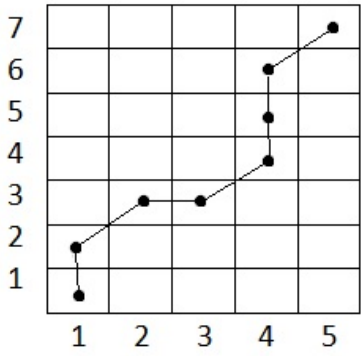
Για να συγκριθούν δύο διαφορετικά στοιχεία απαιτείται ένα μέτρο κόστους $c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0}$, το οποίο θα λαμβάνει μεγάλες τιμές όταν τα δύο στοιχεία διαφέρουν αρκετά μεταξύ τους και μικρές τιμές όταν μοιάζουν μεταξύ τους. Αφού υπολογισθεί το τοπικό κόστος για κάθε ζευγάρι στοιχείων των ακολουθιών X και Y, κατασκευάζεται ο πίνακας κόστους $C \in \mathbb{R}^{N \times M}$, ο οποίος ορίζεται ως $C(n, m) := c(x_n, y_m)$. Στόχος είναι να βρεθεί ένα βέλτιστο μονοπάτι $p = (p_1, \ldots, p_L)$ ευθυγράμμισης μεταξύ των ακολουθιών X και Y, το οποίο θα παρουσιάζει το ελάχιστο συνολικό κόστος. Το μονοπάτι αυτό, όπου $p_l = (n_l, m_l) \in [1, \ldots, N] \times [1, \ldots, M]$ για $l \in [1, \ldots, L]$ θα πρέπει να πληροί τους ακόλουθους περιορισμούς:

1. Οριακές συνθήκες: $p_1 = (1, 1)$ και $p_L = (N, M)$.

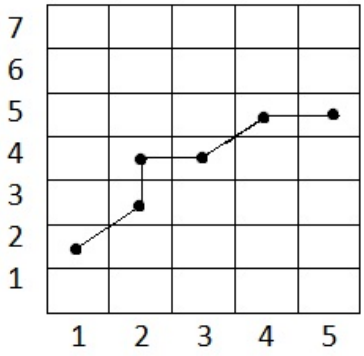
2. Μονοτονία: $n_1 \leq n_2 \leq \cdots \leq n_L$ και $m_1 \leq m_2 \leq \cdots \leq m_L$.

3. Μήκος βήματος: $p_{l+1} - p_l \in \{(1, 0), (0, 1), (1, 1)\}$ για $l \in [0, \ldots, L - 1]$.

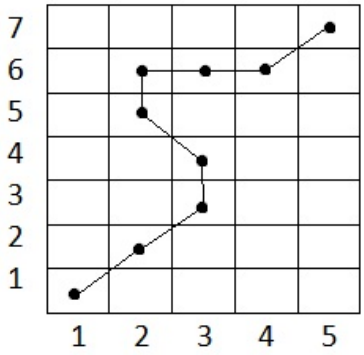
Με βάση αυτούς τους περιορισμούς κάθε στοιχείο x_{n_l} της ακολουθίας X αντιστοιχείται σε ένα στοιχείο y_{m_l} της ακολουθίας Y, ενώ τα πρώτα και τα τελευταία στοιχεία της κάθε ακολουθίας ευθυγραμμίζονται μεταξύ τους (Συνθήκη 1). Η τρίτη συνθήκη επιβάλλει μια "συνέχεια" στο αποτέλεσμα και εξασφαλίζει οτι κανένα στοιχείο των ακολουθιών δεν θα παραλειφθεί από την ευθυγράμμιση. Το Σχήμα 2.4 παρουσιάζει μια σωστή εκδοχή ευθυγράμμισης καθώς και το πώς μπορούν να παραβιαστούν καθεμία από τις παραπάνω συνθήκες.



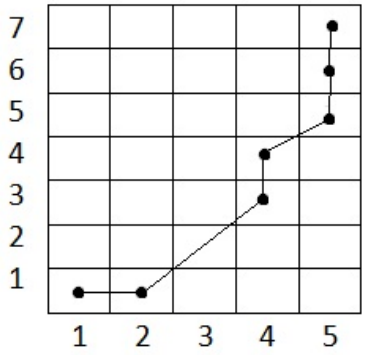
(i) Σωστή ευθυγράμμιση



(ii) Παραβίαση της πρώτης συνθήκης



(iii) Παραβίαση της δεύτερης συνθήκης



(iv) Παραβίαση της τρίτης συνθήκης

Σχήμα 2.4: Παραδείγματα ευθυγράμμισης δύο ακολουθιών

Εφόσον έχουν διατυπωθεί οι συνθήκες ευθυγράμμισης που πρέπει να ικανοποιούνται και έχει οριστεί το μέτρο κόστους που θα χρησιμοποιηθεί, εφαρμόζεται ένας αλγόριθμος δυναμικού προγραμματισμού για την ευθυγράμμιση. Αρχικά, κατασκευάζεται ένας πίνακας συσσωρευμένου κόστους D μεγέθους $N \times M$, ο οποίος ορίζεται ως εξής:

$$D(n,1) = \sum_{i=1}^{n} c(x_i, y_1) \quad \text{για} \quad n \in [1, \ldots, N] \tag{2.1}$$

$$D(1,m) = \sum_{i=1}^{m} c(x_1, y_i) \quad \text{για} \quad m \in [1, \ldots, M] \tag{2.2}$$

$$D(n,m) = \min\{D(n-1,m), D(n,m-1), D(n-1,m-1)\} + c(x_n, y_m) \quad \text{για} \quad n \in [2, \ldots, N], m \in [2, \ldots, M] \tag{2.3}$$

Κάθε στοιχείο $D(n,m)$ ισούται με το συνολικό κόστος ευθυγράμμισης των n πρώτων στοιχείων της ακολουθίας X με τα m πρώτα στοιχεία της ακολουθίας Y. Ο αλγόριθμος που εφαρμόζεται για τον υπολογισμό του βέλτιστου μονοπατιού $p* = (p_1, p_2, \ldots p_L)$ έχει ως εξής:

Βήμα 1 Ανάθεση τελικού σημείου, $p_L = (N, M)$.

Βήμα 2 Αν το τελευταίο στοιχείο που υπολογίστηκε ήταν το $p_l = (n, m)$ και $(n, m) = (1, 1)$ τότε $l = 1$ και τερματισμός της διαδικασίας, αλλιώς μετάβαση στο Βήμα 3 .

Βήμα 3

$$p_{l-1} := \begin{cases} (1, m-1), & \text{αν } n = 1 \\ (n-1, 1), & \text{αν } m = 1 \\ \text{argmin}\{D(n-1,m), & \\ D(n,m-1), D(n-1,m-1)\}', & \text{αλλιώς} \end{cases} \tag{2.4}$$

και επιστροφή στο Βήμα 2 .

Κατά τον υπολογισμό του βέλτιστου μονοπατιού υπάρχουν τρεις εναλλακτικές:

1. Δύο στοιχεία ευθυγραμμίζονται μεταξύ τους, είτε είναι ίδια είτε είναι διαφορετικά, οπότε πρόκειται για μια ταύτιση ή μια αντικατάσταση, αντίστοιχα. Αυτό συμβαίνει όταν στο τρίτο σκέλος της σχέσης 2.4 ελάχιστο κόστος έχει το στοιχείο $D(n-1, m-1)$.

2. Υπάρχει διαγραφή ενός στοιχείου από την ακολουθία X στην ακολουθία Y, όταν στη σχέση 2.4 ελάχιστο κόστος έχει το στοιχείο $D(n-1, m)$.

3. Υπάρχει προσθήκη ενός στοιχείου από την ακολουθία X στην ακολουθία Y, όταν στη σχέση 2.4 ελάχιστο κόστος έχει το στοιχείο $D(n, m-1)$.

2.4.2 Το εργαλείο SCLITE

[1] Το εργαλείο που χρησιμοποιήθηκε για την ευθυγράμμιση και αποτελεί υλοποίηση του DTW είναι το εργαλείο SCLITE, μέρος του NIST SCTK Scoring Toolkit, που έχει αναπτυχθεί για την αξιολόγηση συστημάτων αναγνώρισης φωνής. Συγκρίνει την έξοδο του συστήματος αναγνώρισης με το κείμενο αναφοράς, που αποτελεί τα πραγματικά λεγόμενα του ομιλητή. Το πρώτο κείμενο ονομάζεται hypothesized text (HYP), ενώ το κείμενο αναφοράς αποτελεί το reference text (REF). Αφού γίνει ευθυγράμμιση των δύο κειμένων, προκύπτει μια πληθώρα στατιστικών, που χρησιμοποιούνται για την αξιολόγηση του συστήματος, καθώς και το σκορ της ευθυγράμμισης. Στην προκειμένη περίπτωση ως κείμενο αναφοράς θεωρήθηκε το σενάριο και ως hypothesized text οι υπότιτλοι.

Η διαδικασία της ευθυγράμμισης βασίζεται στο δυναμικό προγραμματισμό και υλοποιεί μια συνολική ελαχιστοποίηση της Levenshtein απόστασης μεταξύ των λέξεων. Η Levenshtein απόσταση μεταξύ δύο ακολουθιών (λέξεων στη συγκεκριμένη περίπτωση) a και b, ισούται με $\text{lev}(|a|, |b|)$ και ορίζεται αναδρομικά ως εξής [27]:

$$\text{lev}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j) + \text{cost of insertion} \\ \text{lev}_{a,b}(i,j-1) + \text{cost of deletion} \\ \text{lev}_{a,b}(i-1,j-1) + \\ \text{cost of substitution} \times 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise} \end{cases} \tag{2.5}$$

όπου

$$1_{(a_i \neq b_j)} = \begin{cases} 1, & \text{if } a_i = b_j \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

είναι η συνάρτηση δείκτης (indicator function).

[1] http://www.itl.nist.gov/iad/mig/tools/

Τα κόστη προσθήκης και διαγραφής (insertion και deletion αντίστοιχα) που χρησιμοποιούνται ισούνται με 3, ενώ το κόστος αντικατάστασης (substitution) ισούται με 4. Με αυτό τον τρόπο κατασκευάζεται ο πίνακας κόστους C που είναι απαραίτητος για την υλοποίηση του DTW αλγόριθμου και έχει οριστεί στην ενότητα 2.4.1.

Εφόσον έχει υπολογιστεί ο πίνακας C, υλοποιείται ο αλγόριθμος DTW, όπως αναλύθηκε στην προηγούμενη ενότητα. Στην έξοδο του συστήματος λαμβάνεται η αντιστοίχιση των δύο κειμένων, το σκορ ευθυγράμμισης και τα ποσοστά των σωστά ευθυγραμμισμένων λέξεων, των αντικαταστάσεων, των προσθηκών και των διαγραφών. Στο Σχήμα 2.5 φαίνεται πώς γίνεται η ευθυγράμμιση των αποσπασμάτων που εμφανίζονται στα Σχήματα 2.1 και 2.3. Τα γράμματα που φαίνονται στην πρώτη στήλη υποδηλώνουν αν δύο λέξεις έχουν ευθυγραμμιστεί σωστά (**C**orrect), αν έχει γίνει αντικατάσταση μιας λέξης του REF text από μια λέξη του HYP text (**S**ubstitution), αν μια λέξη του REF text έχει διαγραφεί (**D**eletion) ή, τέλος, αν έχει προστεθεί μια λέξη στο HYP text (**I**nsertion).

```
C   we  we
D   wuz *
S   gettin  was
S   wuhried worried
S   bout    about
C   you you
C   miss    miss
C   ellen   ellen
D   marse   *
D   gerald  *
C   its its
C   all all
C   right   right
C   pork    pork
C   im  im
C   home    home
```

Σχήμα 2.5: Απόσπσαμα ευθυγράμμισης Σεναρίου-Υποτίτλων

2.4.3 Επεξεργασία της Ευθυγραμμισμένης Εξόδου

Αφού έχει δημιουργηθεί ο πίνακας με τους ευθυγραμμισμένους διαλόγους σεναρίου και υποτίτλων, Σχήμα 2.5, οι χρόνοι που αντιστοιχούν σε κάθε υπότιτλο, και έχουν ήδη εξαχθεί βάσει της διαδικασίας στην ενότητα 2.3,

αντιστοιχίζονται στους διαλόγους του σεναρίου που έχουν ευθυγραμμιστεί με τον κάθε υπότιτλο. Με τον τρόπο αυτό τμηματικά το σενάριο έχει αποκτήσει χρονική πληροφορία, μιας και μόνο οι διάλογοι περιέχουν ακριβείς χρόνους. Για να ολοκληρωθεί η διαδικασία, τα τμήματα του σεναρίου που δεν περιέχουν χρόνους και θεωρείται ότι συμβαίνουν μεταξύ των διαλόγων λαμβάνουν τους χρόνους που μεσολαβούν των διαλόγων.

Το τελικό αποτέλεσμα που επιθυμούμε είναι ένας πίνακας που θα περιέχει κάθε γραμμή του σεναρίου και τους χρόνους έναρξης και λήξης που της έχουν αντιστοιχηθεί. Επιπλέον, αν πρόκειται για διάλογο πρέπει να εμφανίζεται και το όνομα του ομιλητή, όπως φαίνεται στον Πίνακα 2.2.

Start Time	End Time	Speaker	Script Line
795.919	796.295		S3 EXT VERANDAH
795.919	796.295		The carriage has drawn up in front of the steps Pork hurries forward with the lamp and holds
795.919	796.295		it to light the way as Mrs Ellen OHara gets out Several hound dogs swarm on the verandah
795.919	796.295		and leap about her delightedly
796.295	798.505	PORK	we wuz gettin wuhried bout you miss ellen
798.756	801.09	PORK	marse gerald
798.756	801.09	ELLEN	its all right pork im home
801.09	801.592		As she starts up the steps to the verandah Jonas Wilkerson steps out from the shadows
801.09	801.592		where he has obviously been waiting He is nervous twirling his hat in his hand

Πίνακας 2.2: Τελικό Αποτέλεσμα Ευθυγράμμισης

Κεφάλαιο 3

Κατάτμηση σε Λήψεις (Shots)

Η κατάτμηση μιας ταινίας σε λήψεις αποτελεί το πρώτο στάδιο της επεξεργασίας πριν την τελική κατάτμησή της σε σκηνές. Αφού γίνει ο διαχωρισμός σε λήψεις, εξάγονται από κάθε μια κάποια αντιπροσωπευτικά καρέ (γνωστά ως Key Frames), τα οποία συνοψίζουν το περιεχόμενο της λήψης και χρησιμοποιούνται στα επόμενα στάδια της επεξεργασίας, ώστε να επιταχύνεται η όλη διαδικασία.

Στο παρόν κεφάλαιο θα παρουσιαστούν δύο βασικές μέθοδοι κατάτμησης σε λήψεις, που χρησιμοποιούν χαμηλού επιπέδου οπτικά χαρακτηριστικά (ιστογράμματα χρώματος και ακμές). Για το πρόβλημα της κατάτμησης σε λήψεις τέτοιου είδους χαρακτηριστικά μπορούν να δώσουν ικανοποιητικά αποτελέσματα, καθώς η μετάβαση από μια λήψη στην επόμενη συνεπάγεται έντονες αλλαγές στη χρωματική κατανομή και στις ακμές της εικόνας. Αφού δοκιμαστεί κάθε μέθοδος ξεχωριστά, θα συνδυαστούν τα αποτελέσματα των δύο μεθόδων για να ληφθεί ένα καλύτερο αποτέλεσμα κατάτμησης. Με τον τρόπο αυτό οι λήψεις των ταινιών εντοπίζονται με μεγάλη ακρίβεια και τα αποτελέσματα είναι πολύ ικανοποιητικά για το σύνολο των ταινιών της βάσης.

Στη συνέχεια, περιγράφονται δύο μέθοδοι εξαγωγής των αντιπροσωπευτικών καρέ, που αντιμετωπίζουν το πρόβλημα από τελείως διαφορετική σκοπιά. Η μια προτείνει τη μη γραμμική χρονική δειγματοληψία και η άλλη αξιοποιεί τη φασματική ομαδοποίηση (Spectral Clustering) για την ομαδοποίηση παρόμοιων καρέ μιας λήψης. Για κάθε μια μέθοδο εξάγονται κάποιοι δείκτες αξιολόγησης και, τελικά, επιλέγεται η μέθοδος του Spectral Clustering, για τη μεγάλη συμπίεση της πληροφορίας που πετυχαίνει.

3.1 Είδη Μεταβάσεων

Η εναλλαγή λήψεων γίνεται με δύο διαφορετικές τεχνικές:

- Hard Cuts
- Ομαλές Μεταβάσεις
  - Fade Out
  - Fade In
  - Dissolve
  - Wipe

Στα *hard cuts* το τελευταίο καρέ μιας λήψης ακολουθείται από το πρώτο καρέ της επόμενης. Στα *fade out* οι εικόνες σταδιακά εκφυλίζονται σε μια μονοχρωματική εικόνα, συνήθως μαύρου χρώματος. Κατά τη διάρκεια ενός *fade in* συμβαίνει η αντίστροφη διαδικασία, δηλαδή από μια μονοχρωματική εικόνα εμφανίζονται σταδιακά τα καρέ που αποτελούν την αρχή μιας λήψης. Ως συνδυασμός των δύο προηγούμενων, προκύπτει το *dissolve*, όπου μέσα από τα τελευταία καρέ μιας λήψης εμφανίζονται τα πρώτα καρέ της επόμενης. Τέλος, στα *wipes* η μετάβαση γίνεται με τα καρέ της επόμενης λήψης να εμφανίζονται από τα καρέ της προηγούμενης λήψης με τη βοήθεια ενός κινούμενου ορίου. Τα διάφορα είδη μεταβάσεων εμφανίζονται στο Σχήμα 3.1.

Έστω μια ακολουθία καρέ $G(x, y, t)$ στα οποία θα εφαρμοστεί μια επεξεργασία, ώστε να δημιουργηθεί ένα οπτικό εφέ ομαλής μετάβασης. Τότε, η μετάβαση αυτή, διάρκειας l_{trans}, μπορεί να μοντελοποιηθεί σύμφωνα με το [8], ως εξής:

Fade Out

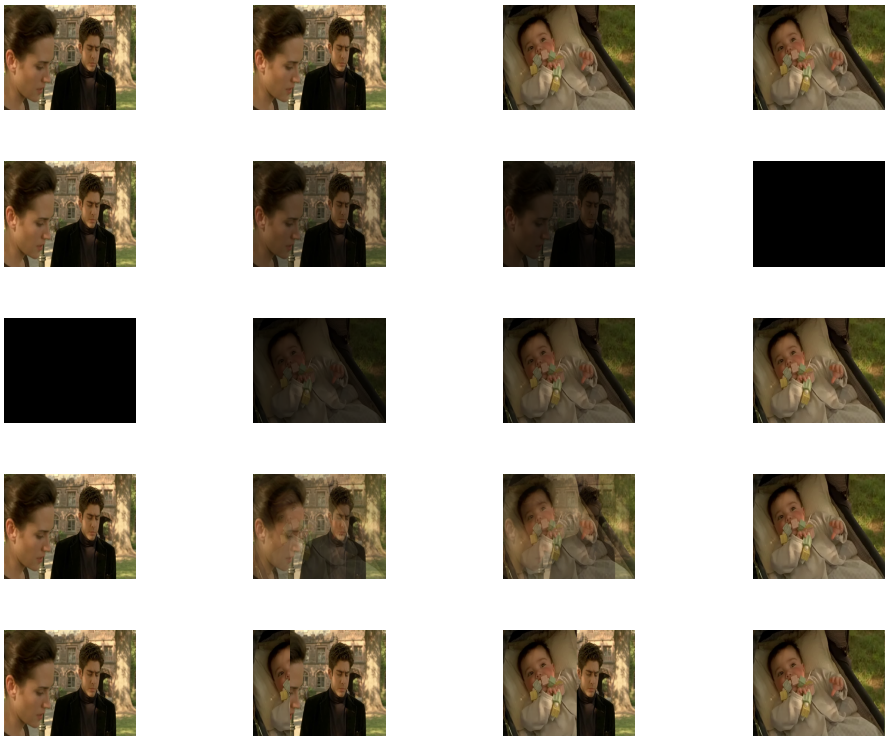
$$E_{fo} = G(x, y, t)\Big(1 - \frac{t}{l_1}\Big)\Big|_{t\in[t_1, t_1+l_1]} \tag{3.1}$$

Fade In

$$E_{fi} = G(x, y, t)\Big(\frac{t}{l_2}\Big)\Big|_{t\in[t_2, t_2+l_2]} \tag{3.2}$$

Dissolve

$$\begin{aligned} E_d = {} & G_1(x, y, t)\Big(1 - \frac{t}{l_1}\Big)\Big|_{t\in[t_1, t_1+l_1]} \\ & + G_2(x, y, t)\Big(\frac{t}{l_2}\Big)\Big|_{t\in[t_2, t_2+l_2]} \end{aligned} \tag{3.3}$$



Σχήμα 3.1: Είδη μεταβάσεων λήψεων - Hard Cut, Fade Out, Fade In, Dissolve, Wipe

3.2 Υπάρχουσες Μέθοδοι

Οι μέθοδοι που χρησιμοποιούνται για τον εντοπισμό των ορίων μεταξύ διαδοχικών λήψεων χρησιμοποιούν κυρίως οπτικά χαρακτηριστικά, όπως χρώμα και ακμές. Τα οπτικά χαρακτηριστικά αναμένεται να αλλάζουν δραματικά σε ένα hard cut, ενώ σε μια ομαλή μετάβαση, από μια λήψη στην επόμενη, αναμένεται μια σταδιακή μεταβολή τους. Τα διάφορα χαρακτηριστικά που έχουν χρησιμοποιηθεί κατά καιρούς στη βιβλιογραφία για τον εντοπισμό των εναλλαγών λήψεων αναλύονται παρακάτω.

3.2.1 Ιστόγραμμα Χρώματος

Στο [31] για κάθε καρέ, έστω i, εξάγεται το ιστόγραμμα φωτεινότητας H_i και συγκρίνεται με το αντίστοιχο του επόμενου καρέ, χρησιμοποιώντας την L^1 νόρμα, όπως φαίνεται στην εξίσωση (3.4).

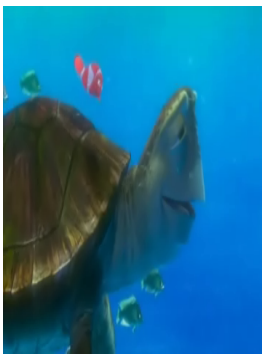
$$SD_i = \sum_{j=1}^{G} |H_i(j) - H_{i+1}(j)| \tag{3.4}$$

όπου G το σύνολο των bins του ιστογράμματος. Αφού γίνει κανονικοποίηση του SD, διαιρώντας με το γινόμενο του αριθμού των εικονοστοιχείων της εικόνας με το πλήθος των επιπέδων φωτεινότητας G, κάποιο καρέ i θεωρείται όριο μιας λήψης σε περίπτωση που η συνολική διαφορά SD_i υπερβαίνει ένα κατώφλι T. Αντίστοιχη διαδικασία μπορεί να ακολουθηθεί χρησιμοποιώντας και τα τρία κανάλια χρώματος (RGB). Εξάγονται για κάθε καρέ i το ιστόγραμμα φωτεινότητας για καθένα από τα κανάλια χρώματος H_i^{RED}, H_i^{GREEN}, H_i^{BLUE}, τα οποία συγκρίνονται μεταξύ τους, ώστε να εξαχθούν τα SD_i^{RED}, SD_i^{GREEN}, SD_i^{BLUE} και λαμβάνεται ο μέσος όρος τους, όπως φαίνεται στην εξίσωση (3.5). Ενδεικτικά, στο Σχήμα 3.2 παρουσιάζεται ένα καρέ της ταινίας FNE με τα αντίστοιχα ιστογράμματα χρώματος και στο Σχήμα 3.3 απεικονίζονται οι τιμές του SD για ένα τμήμα της ίδιας ταινίας. Οι κορυφές του διαγράμματος αντιστοιχούν στις αλλαγές λήψης.

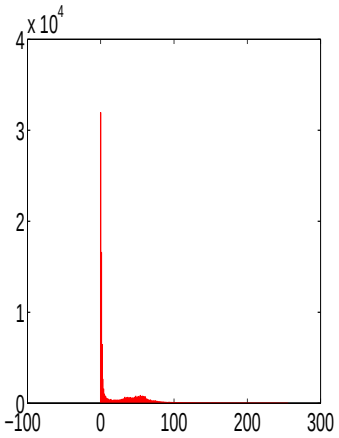
$$SD_i = \frac{SD_i^{RED} + SD_i^{GREEN} + SD_i^{BLUE}}{3} \tag{3.5}$$

3.2.2 Ακμές

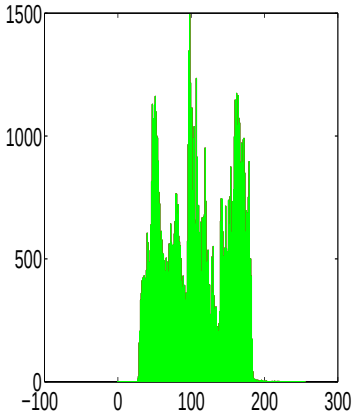
Κατά την εναλλαγή λήψεων, τόσο σε ένα hard cut όσο και σε μια ομαλή μετάβαση, αναμένεται να εμφανίζονται στην εικόνα νέες ακμές, δεδομένου οτι αλλάζουν τα αντικείμενα που εμφανίζονται. Στο [29] εισάγεται η έννοια του *Edge Change Ratio (ECR)*, το οποίο προσδιορίζει σε ποιο ποσοστό προστίθενται ή αφαιρούνται ακμές σε ένα καρέ. Για την εξαγωγή αυτού του λόγου εφαρμόζεται, αρχικά, σε δύο εικόνες, I και I', ένας τελεστής ανίχνευσης ακμών, με αποτέλεσμα να προκύπτουν δύο δυαδικές εικόνες E και E'. Ως *εισερχόμενες ακμές*, X^{in}, θεωρούνται οι ακμές της E' που απέχουν πάνω από μια απόσταση r από τις κοντινότερες ακμές τους στην E. Αντίστοιχα, ως *εξερχόμενες ακμές*, X^{out}, θεωρούνται οι ακμές της E που απέχουν πάνω από μια απόσταση r από τις κοντινότερες ακμές τους στην



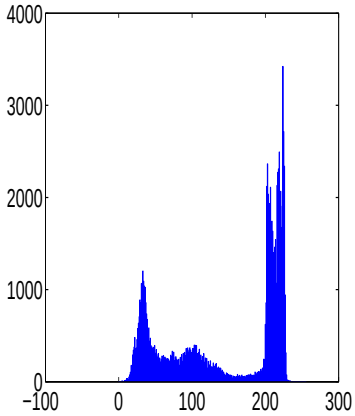
(i) Original Frame



(ii) Red Histogram



(iii) Green Histogram



(iv) Blue Histogram

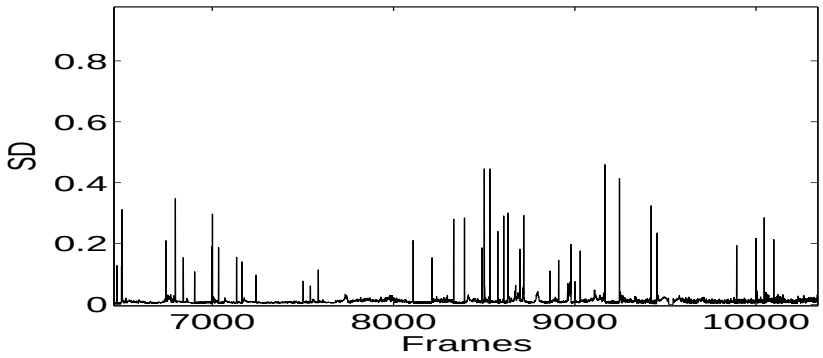
Σχήμα 3.2: Ιστογράμματα χρώματος ενός καρέ.

E'. Αν οριστεί $\sigma = \#edge\ pixels\ in\ E$ και $\sigma' = \#edge\ pixels\ in\ E'$, τότε ως ECR ορίζεται:

$$ECR = \max(\frac{X^{in}}{\sigma'}, \frac{X^{out}}{\sigma}) \tag{3.6}$$

Η εξίσωση (3.6) υποδεικνύει πως σε κάθε καρέ μπορεί να εξαχθεί το ποσοστό εισερχόμενων, ρ_{in}, και εξερχόμενων, ρ_{out}, ακμών και να οριστεί ως ECR το μεγαλύτερο εκ των δύο.

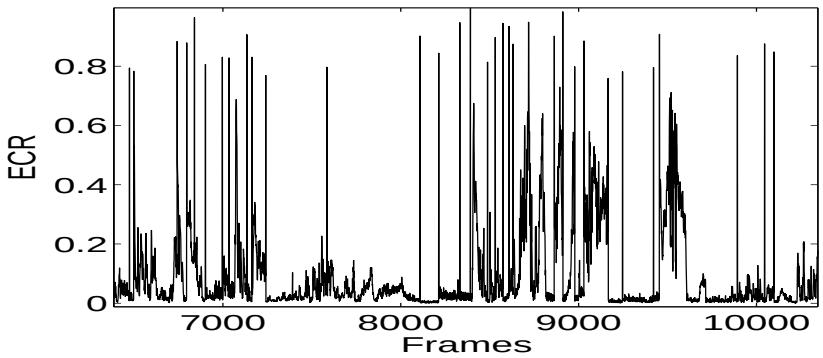
Στην πράξη, όπως φαίνεται και στο Σχήμα 3.5, αφού εξαχθούν οι ακμές δύο διαδοχικών καρέ (E και E'), γίνεται μια διαστολή τους (dilation) με ένα δομικό στοιχείο μεγέθους r και οι εικόνες που προκύπτουν αντιστρέφονται, οπότε προκύπτουν οι εικόνες ID και ID'. Στη συνέχεια, εφαρμόζεται ο



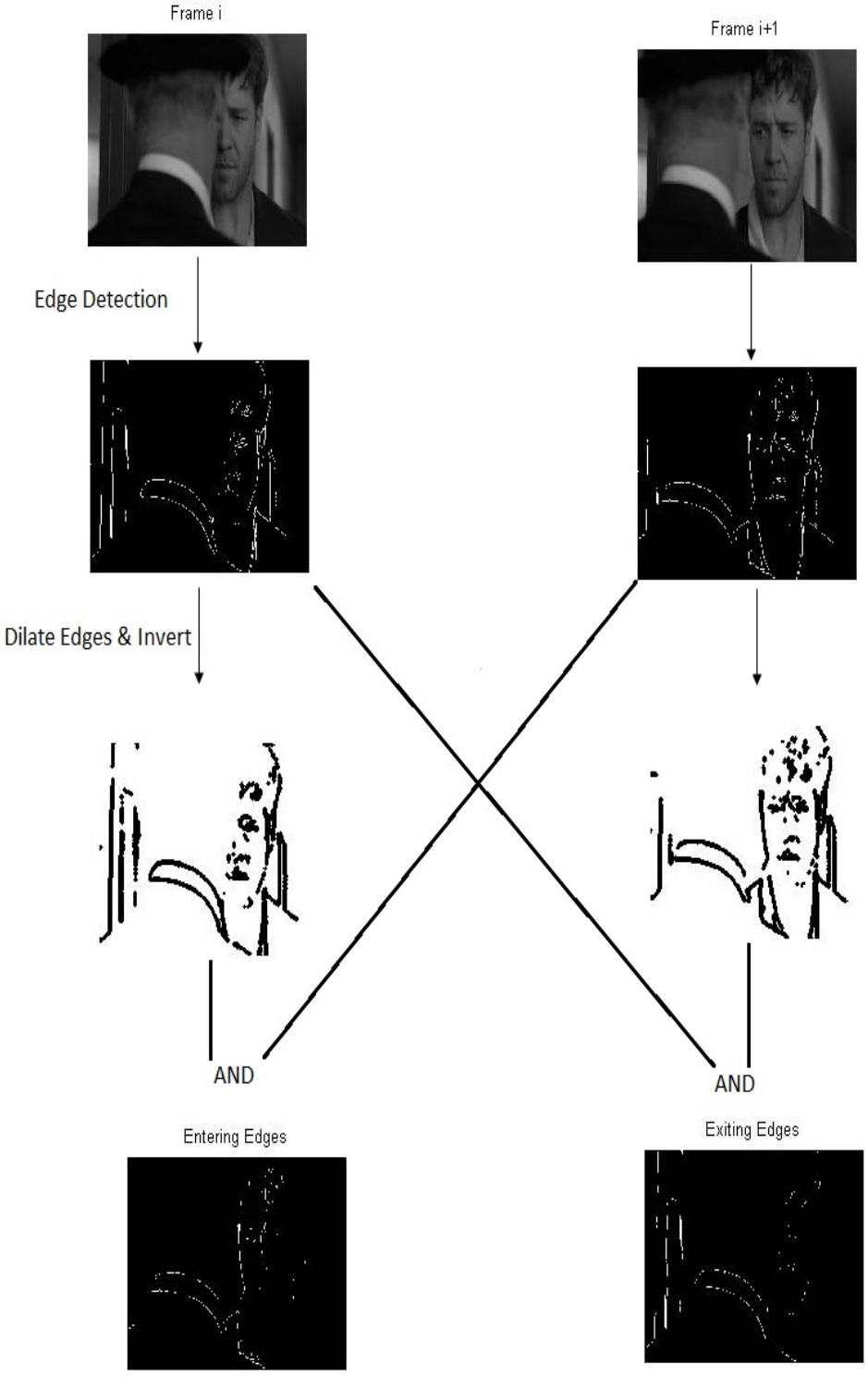
Σχήμα 3.3: Histogram Differences

λογικός τελεστής AND μεταξύ των εικόνων E' και ID, ώστε να προκύψουν οι εισερχόμενες ακμές X^{in} και μεταξύ των εικόνων E και ID', ώστε να προκύψουν οι εξερχόμενες ακμές X^{out}. Υπολογίζοντας το ECR για κάθε καρέ της ταινίας, όπως φαίνεται στο Σχήμα 3.4, ως αλλαγές λήψης μπορούν να θεωρηθούν τα καρέ στα οποία το ECR λαμβάνει τιμή μεγαλύτερη από ένα κατώφλι.

Εξετάζοντας τα διάφορα είδη μεταβάσεων είναι εφικτό να προβλεφθεί σε ποιες περιπτώσεις τα ρ_{in} και ρ_{out} θα λαμβάνουν μεγάλες τιμές. Σε ένα hard cut, για παράδειγμα, υπάρχει μεγάλο ποσοστό τόσο εισερχόμενων όσο και εξερχόμενων ακμών. Αντίστοιχα, σε ένα fade in, όπου εμφανίζονται νέα αντικείμενα, μόνο το ρ_{in} λαμβάνει μεγάλες τιμές, σε αντίθεση με ένα fade out, όπου σταδιακά τα αντικείμενα εξαφανίζονται από την εικόνα και το ρ_{out} υπερτερεί σε μέγεθος. Τέλος, λόγω της δομής του dissolve, στο αρχικό του στάδιο αναμένονται μεγάλες τιμές του ρ_{out} και στο τελικό του στάδιο μεγάλες τιμές του ρ_{in}.



Σχήμα 3.4: ECR



Σχήμα 3.5: Κύρια Βήματα για τον Υπολογισμό του ECR

3.2.3 Χρήση Πληροφορίας Σεναρίου

Το σενάριο, στο οποίο έχει εισαχθεί η χρονική πληροφορία από τους υπότιτλους, περιλαμβάνει μια εκτίμηση των χρονικών διαστημάτων στα οποία υπάρχουν ομαλές μεταβάσεις, καθώς περιέχει αναφορές των λέξεων FADE και DISSOLVE. Είναι εφικτό να βρεθούν αυτές οι λέξεις στο σενάριο, με τους αντίστοιχους χρόνους τους και, στη συνέχεια, να εντοπισθούν ως αλλαγές λήψης τα καρέ που εμφανίζουν μέγιστη τιμή ECR εντός αυτού του χρονικού διαστήματος. Το αποτέλεσμα αυτής της διαδικασίας συνδυάζεται με το αποτέλεσμα που προκύπτει κάνοντας αποκλειστικά χρήση του ECR για την κατάτμηση σε λήψεις.

3.3 Πειραματικά Αποτελέσματα

Για την αξιολόγηση των αποτελεσμάτων της κατάτμησης χρησιμοποιείται, αρχικά, ως ground truth η επισήμανση της ταινίας "Gone With The Wind". Αξιολογούνται οι διάφορες μέθοδοι κατάτμησης σε λήψεις, καθώς και συνδυασμός αυτών, και στη συνέχεια, παρουσιάζονται τα αποτελέσματα για τη βάση των 7 ταινιών. Τα καρέ που έχουν επισημειωθεί ως αλλαγές λήψης αποτελούν το σύνολο των Relevant Frames, ενώ αυτά που εντοπίσθηκαν από τους αλγορίθμους αποτελούν το σύνολο των Retrieved Frames, αντίστοιχα. Επειδή πέρα από τα hard cuts είναι αδύνατον να επιτευχθεί πλήρης συμφωνία των επισημειωμένων καρέ με αυτά που εντοπίζουν οι αλγόριθμοι (για παράδειγμα σε ένα dissolve), επιτρέπουμε μια απόσταση $\sim 5-100$ καρέ μεταξύ των δύο. Τα μεγέθη αξιολόγησης που χρησιμοποιούνται είναι τα γνωστά:

$$\text{Precision} = \frac{|\{\text{relevant}\} \cap \{\text{retrieved}\}|}{|\{\text{retrieved}\}|} \tag{3.7}$$

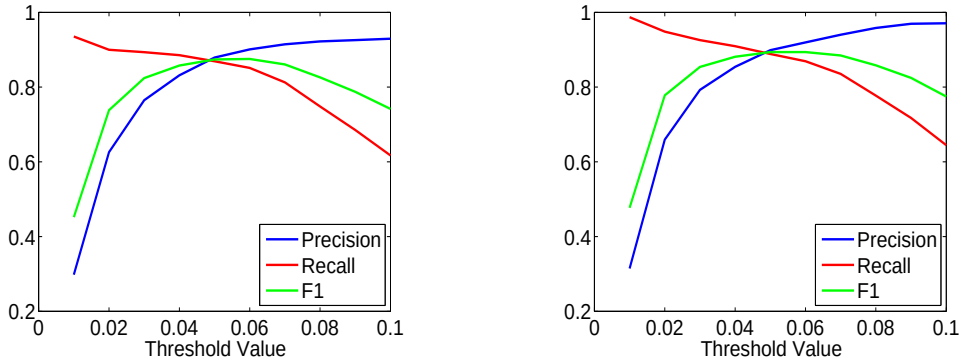
$$\text{Recall} = \frac{|\{\text{relevant}\} \cap \{\text{retrieved}\}|}{|\{\text{relevant}\}|} \tag{3.8}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.9}$$

3.3.1 Μέθοδος Ιστογράμματος

Αρχικά, εξετάζεται η μέθοδος του ιστογράμματος, όπως αναλύεται στην ενότητα 3.2.1, μόνο που στην εξίσωση (3.4) λαμβάνεται η ευκλείδεια απόσταση μεταξύ των ιστογραμμάτων δύο διαδοχικών καρέ και όχι η

L^1 νόρμα. Επιπλέον, εντοπίζονται τα τοπικά μέγιστα του SD και δεν επιλέγονται απλώς τα καρέ των οποίων το SD ξεπερνά το κατώφλι T. Ωστόσο, και για αυτόν τον τρόπο πρέπει να βρεθεί ένα βέλτιστο κατώφλι τ, που θα ορίζει την ελάχιστη απόσταση μιας κορυφής από τις διπλανές της, ώστε να θεωρηθεί τοπικό μέγιστο. Γραφικά τα αποτελέσματα εμφανίζονται στο Σχήμα 3.6. Στο σχήμα αυτό εμφανίζονται οι τιμές των μέτρων αξιολόγησης ως προς το κατώφλι, ενώ σωστά εντοπισμένα θεωρούνται τα καρέ που απέχουν 5 καρέ από τα επισημειωμένα (Σχήμα 3.6i) και 50 καρέ (Σχήμα 3.6ii).



(i) Παράθυρο 5 καρέ (ii) Παράθυρο 50 καρέ

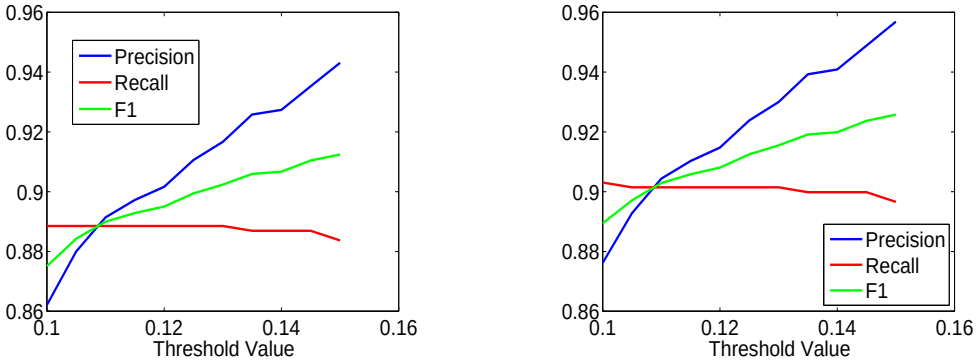
Σχήμα 3.6: Precision, Recall, F_1-measure ως προς το κατώφλι απόφασης για την αξιολόγηση της μεθόδου Ιστογράμματος στην ταινία Gone With The Wind.

Το βέλτιστο κατώφλι που ικανοποιεί το tradeoff για υψηλές τιμές τόσο Precision όσο και Recall είναι το $\tau = 0.05$. Παρατηρούμε, επίσης, πως παρά την πιο χαλαρή συνθήκη στο Σχήμα 3.6ii δεν υπάρχει δραματική βελτίωση στο αποτέλεσμα. Αυτό συμβαίνει διότι το ιστόγραμμα χρώματος εντοπίζει πολύ δύσκολα τις ομαλές μεταβάσεις, επομένως, και στις δύο περιπτώσεις εντοπίζει κατά κύριο λόγο τα hard cuts.

3.3.2 Μέθοδος ECR

Με παρόμοιο τρόπο εξετάζεται και η μέθοδος ECR για την υλοποίηση της οποίας επιλέχθηκε ο τελεστής sobel [24] για την ανίχνευση των ακμών. Τα αποτελέσματα φαίνονται στο Σχήμα 3.7. Παρατηρούμε πως η τιμή του Recall παραμένει σχεδόν σταθερή για όλο το εύρος των τιμών του κατωφλίου. Αυτό σημαίνει πως η μεταβολή του κατωφλίου δεν επηρεάζει τον αριθμό των σωστών λήψεων που εντοπίζονται, παρά μόνο το συνολικό

αριθμό των λήψεων που εντοπίζονται. Επιπλέον, παρατηρείται μια μικρή βελτίωση των ποσοστών σε σχέση με με τη μέθοδο του ιστογράμματος.



(i) Παράθυρο 5 καρέ (ii) Παράθυρο 50 καρέ

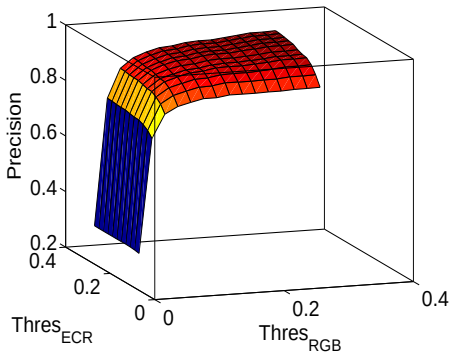
Σχήμα 3.7: Precision, Recall, F_1-measure ως προς το κατώφλι απόφασης για την αξιολόγηση της μεθόδου ECR στην ταινία Gone With The Wind.

3.3.3 Συνδυασμός Μεθόδου Ιστογράμματος και ECR

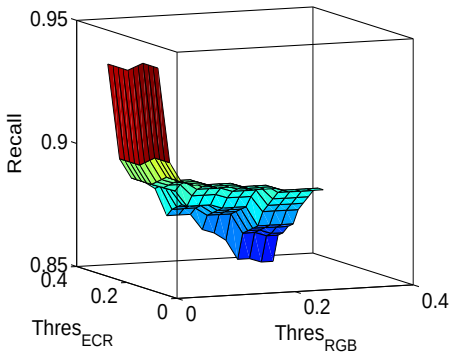
Οι παραπάνω αλγόριθμοι παρουσιάζουν αρκετά υψηλά ποσοστά επιτυχίας. Εύλογα, δημιουργείται η απορία αν καταφέρνουν να εντοπίσουν τις ίδιες λήψεις, ή καθένας εντοπίζει και κάποιες διαφορετικές. Για το λόγο αυτό επιδιώκεται ο συνδυασμός τους, όπου διατηρούνται τόσο οι κοινές λήψεις που εντοπίζουν όσο και οι λήψεις που εντοπίζονται μόνο από έναν εκ των δύο. Με βάση τα αποτελέσματα που εμφανίζονται στο Σχήμα 3.8 διαπιστώνεται πως ο συνδυασμός των δύο μεθόδων αυξάνει τα ποσοστά επιτυχίας κατά $4-5\%$.

3.3.4 Χρήση Πληροφορίας Σεναρίου

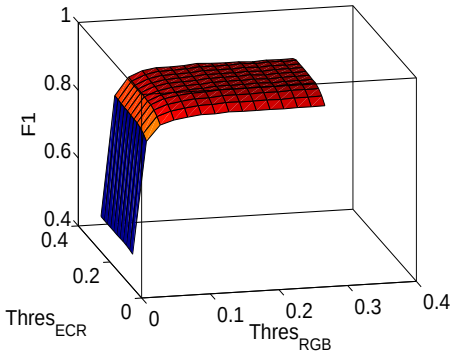
Στον Πίνακα 3.1 φαίνονται τα αποτελέσματα της κατάτμησης σε λήψεις για εντοπισμό των ομαλών μεταβάσεων μέσω του σεναρίου, όπως αυτό αναλύθηκε στην ενότητα 3.2.3. Παρατηρείται ότι με τον τρόπο αυτό μειώνεται το Precision, αλλά αυξάνεται το Recall σε σχέση με την τελευταία μέθοδο.



(i) Precision



(ii) Recall



(iii) F_1

Σχήμα 3.8: Δείκτες αξιολόγησης της κατάτμησης σε λήψεις με συνδυασμό μεθόδων ως προς τις παραμέτρους Thres_{RGB} και Thres_{ECR} στην ταινία Gone With The Wind.

3.3.5 Αποτελέσματα στη Βάση Ταινιών

Για την αξιολόγηση και εξαγωγή του καλύτερου αποτελέσματος της κατάτμησης σε λήψεις πάνω στη βάση δεδομένων, χρησιμοποιήθηκε ο συνδυασμός των μεθόδων ιστογράμματος και ECR. Στόχος είναι να εξαχθεί ένα σύνολο παραμέτρων (Thres_{RGB} και Thres_{ECR}) που θα δίνει ικανοποιητικά αποτελέσματα κατάτμησης για το σύνολο των ταινιών. Στο Σχήμα 3.9 φαίνονται οι μέσες τιμές των δεικτών αξιολόγησης. Ως βέλτιστες παράμετροι επιλέγονται οι $(\text{Thres}_{RGB}, \text{Thres}_{ECR}) = (0.58, 0.20)$, για τις οποίες επιτυγχάνεται η μέγιστη τιμή $F_1 = 94.63\%$ και $Recall = 94.8\%$.

	Precision	**Recall**	**F1**
5	0.84	0.87	0.86
50	0.90	0.93	0.91

Πίνακας 3.1: Δείκτες αξιολόγησης της κατάτμησης σε λήψεις με χρήση της πληροφορίας του σεναρίου για την ταινία Gone With The Wind.

3.4 Εξαγωγή Αντιπροσωπευτικών Καρέ (Key Frames)

Κάθε λήψη περιλαμβάνει ένα μεγάλο αριθμό καρέ, γεγονός που καθιστά δύσκολη και χρονοβόρα τη διαδικασία της επεξεργασίας της για την περαιτέρω ανάλυση και το διαχωρισμό των σκηνών. Για το λόγο αυτό είθισται από κάθε λήψη να εξάγεται ένας περιορισμένος αριθμός αντιπροσωπευτικών καρέ, που ονομάζονται Key Frames, τα οποία θα συμπυκνώνουν την πληροφορία ολόκληρης της λήψης και θα επιταχύνουν την επεξεργασία. Πρόκειται για ένα είδος δειγματοληψίας, που στόχο έχει να διατηρήσει το οπτικό περιεχόμενο της λήψης και, παράλληλα, να μειώσει τον όγκο των δεδομένων. Για την εξαγωγή των αντιπροσωπευτικών καρέ έχουν προταθεί διάφοροι τρόποι, οι οποίοι αναλύονται στις επόμενες ενότητες.

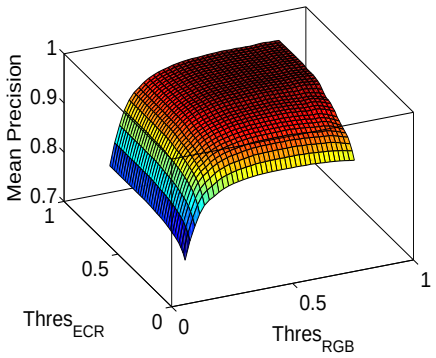
3.4.1 Μη γραμμική Χρονική Δειγματοληψία

Η μέθοδος αυτή [28] χρησιμοποιεί την τομή των ιστογραμμάτων χρώματος, εξίσωση (3.10), για να υπολογίσει τη χρωματική ομοιότητα ($ColSim$) δύο καρέ i και j μιας λήψης. Στη συνέχεια, εξάγει ένα δείκτη ανομοιότητας D μεταξύ των δύο αυτών καρέ, όπως φαίνεται στην εξίσωση (3.11).

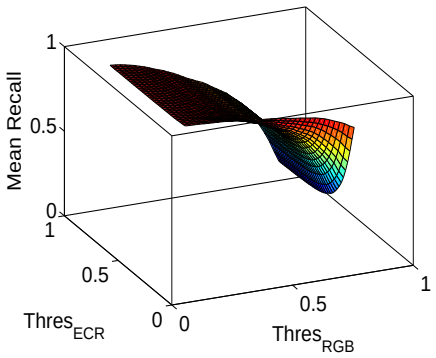
$$ColSim(i,j) = \frac{\sum_{k=1}^{G} \min(H_i(k), H_j(k))}{\sum_{k=1}^{G} H_j(k)} \tag{3.10}$$

$$D(i,j) = 1 - ColSim(i,j) \tag{3.11}$$

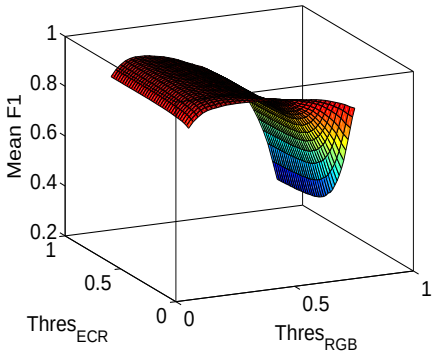
Εφόσον έχουν υπολογιστεί οι συντελεστές ανομοιότητας μεταξύ όλων των καρέ της λήψης, επιλέγεται ως πρώτο αντιπροσωπευτικό καρέ το πρώτο καρέ της λήψης και στη συνέχεια τα καρέ των οποίων η ανομοιότητα με το τελευταίο επιλεγμένο αντιπροσωπευτικό καρέ ξεπερνά ένα κατώφλι $\epsilon \in [0,1]$.



(i) Precision



(ii) Recall



(iii) F_1

Σχήμα 3.9: Δείκτες αξιολόγησης της κατάτμησης σε λήψεις με συνδυασμό μεθόδων ως προς τις παραμέτρους Thres_{RGB} και Thres_{ECR} στη βάση των 7 ταινιών.

Αυτή η μη γραμμική δειγματοληψία υπερτερεί σε σχέση με μια ομοιόμορφη, καθώς οι σχετικά σταθερές λήψεις τελικά συμπυκνώνονται σε ένα πολύ μικρό αριθμό καρέ (κάποιες φορές ένα ή δύο), ενώ στις λήψεις με έντονη κίνηση διατηρούνται περισσότερα καρέ, τα οποία, όμως, εμφανίζουν πλούσιο περιεχόμενο.

3.4.2 Φασματική Ομαδοποίηση (Spectral Clustering)

Σε αυτή τη μέθοδο [4] από κάθε καρέ μιας λήψης εξάγεται το κανονικοποιημένο HSV ιστόγραμμα. Στόχος είναι η ομαδοποίηση των καρέ της λήψης και, στη συνέχεια, η επιλογή ενός καρέ από κάθε ομάδα, που θα αποτελεί και το αντιπροσωπευτικό καρέ της συγκεκριμένης ομάδας. Ο αλγόριθμος που υλοποιείται για την εξαγωγή του συνόλου των

αντιπροσωπευτικών καρέ είναι ένας αλγόριθμος φασματικής ομαδοποίησης (Spectral Clustering).

Η φασματική ομαδοποίηση βασίζεται σε μεθόδους διαμερισμού γράφων. Αφού οριστεί ένα μέτρο ομοιότητας των δεδομένων, κατασκευάζεται ένας γράφος με βάρη των ακμών τις τιμές της ομοιότητας μεταξύ των κόμβων. Στην προκειμένη περίπτωση, τα καρέ μιας λήψης αποτελούν τους κόμβους του γράφου και τα βάρη των ακμών υπολογίζονται με βάση τη χρωματική ομοιότητα κάθε ζεύγους καρέ. Τα δεδομένα χωρίζονται με τέτοιο τρόπο ώστε εντός ενός υπογραφήματος οι ακμές να έχουν μεγάλα βάρη, ενώ οι ακμές μεταξύ διαφορετικών υπογραφημάτων να έχουν μικρά βάρη. Το πλεονέκτημα της μεθόδου είναι οτι χρησιμοποιεί τις ιδιοτιμές του γράφου ομοιότητας για να πετύχει μια μείωση της διάστασης των δεδομένων και στη συνέχεια ένα clustering σε λιγότερες διαστάσεις.

Τα βήματά του αλγορίθμου είναι τα ακόλουθα, [16]:

Βήμα 1 Υπολογισμός του πίνακα ομοιότητας A μεταξύ των καρέ της λήψης, με βάση το [4], ως εξής:

$$a(i,j) = 1 - \frac{1}{\sqrt{2}}\sqrt{\sum_{k=1}^{G}(HSV_i(k) - HSV_j(k))^2} \tag{3.12}$$

Βήμα 2 Ορισμός του διαγώνιου πίνακα D, όπου κάθε διαγώνιο στοιχείο (i,i) ισούται με το άθροισμα των στοιχείων της $i-$οστής γραμμής του πίνακα A. Κατασκευή του Λαπλασιανού πίνακα, σύμφωνα με την εξίσωση (3.13):

$$L = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \tag{3.13}$$

Βήμα 3 Υπολογισμός των K πρωτευόντων ιδιαδιανυσμάτων, $x_1, x_2, \ldots, x_K$ του πίνακα L και κατασκευή του πίνακα $X = [x_1\ x_2\ \ldots\ x_K]$. Ως πρωτεύοντα ιδιοδιανύσματα θεωρούνται αυτά των οποίων η αντίστοιχη ιδιοτιμή ξεπερνά ένα κατώφλι λ.

Βήμα 4 Κανονικοποίηση των γραμμών του πίνακα X, ώστε να έχουν ευκλείδειο μήκος ίσο με τη μονάδα και κατασκευή του πίνακα Y.

$$y_{ij} = \frac{x_{ij}}{\sqrt{\sum_j x_{ij}^2}} \tag{3.14}$$

Βήμα 5 Ομαδοποίηση των γραμμών του Y σε K ομάδες, χρησιμοποιώντας τον αλγόριθμο k-means [1].

Βήμα 6 Ανάθεση ενός καρέ i στο cluster j αν και μόνο αν η γραμμή i του πίνακα Y ανήκει στο cluster j.

Βήμα 7 Επιλογή ενός καρέ από κάθε cluster ως αντιπροσωπευτικό καρέ. Η επιλογή αυτού του καρέ γίνεται έτσι ώστε η ομοιότητα του σε σχέση με τα υπόλοιπα καρέ του cluster να είναι η μέγιστη δυνατή.

3.5 Αξιολόγηση Εξαγωγής Αντιπροσωπευτικών Καρέ

Για την αξιολόγηση των αντιπροσωπευτικών καρέ που έχουν εξαχθεί υπάρχουν κάποια μέτρα αξιοπιστίας, που ορίζονται στο [7] και αυτά είναι τα ακόλουθα:

Λόγος Συμπίεσης (Compression Ratio): Το ποσοστό συμπίεσης της ακολουθίας των καρέ είτε για μια λήψη είτε για ολόκληρη την ταινία προκύπτει από την ακόλουθη σχέση:

$$\text{CR} = 1 - \frac{\#\text{KeyFrames}}{\#\text{Frames}} \tag{3.15}$$

Είναι επιθυμητό για την καλύτερη δυνατή περίληψη μιας λήψης (και κατ' επέκταση μιας ταινίας) να διατηρούνται όσο το δυνατόν λιγότερα καρέ, πλούσια σε περιεχόμενο. Επομένως, ο δείκτης CR είναι επιθυμητό να λαμβάνει μεγάλες τιμές, κοντά στη μονάδα.

Μέση Πιστότητα Λήψης (Average Shot Fidelity): Πρόκειται για ένα δείκτη που δείχνει κατά πόσο τα αντιπροσωπευτικά καρέ μοίαζουν με τα υπόλοιπα καρέ της λήψης. Ο δείκτης αυτός ορίζεται ως εξής:

$$\text{ASF}(F, KF) = 1 - \frac{1}{N}\sum_{n=1}^{N} S(F_n, KF) \tag{3.16}$$

Στην εξίσωση (3.16) $F = \{F_1, F_2, \ldots, F_N\}$ είναι το σύνολο των καρέ της λήψης, $KF = \{KF_1, KF_2, \ldots, KF_{N_{kf}}\}$ είναι το σύνολο των αντιπροσωπευτικών καρέ που εξήχθησαν και S είναι η απόσταση μεταξύ ενός καρέ F_n και του συνόλου των αντιπροσωπευτικών καρέ, η οποία ορίζεται ως εξής:

$$S(F_n, KF) = \min_j \text{Diff}(F_n, KF_j) \tag{3.17}$$

όπου $\text{Diff}(F_i, F_j)$ είναι ένα μέτρο απόστασης μεταξύ δύο καρέ που θα μπορούσε να οριστεί όπως στην εξίσωση (3.11). Για να εξαχθεί ο δείκτης ASF για ολόκληρη την ταινία αρκεί να υπολογιστεί η μέση τιμή του από όλες τις λήψεις:

$$\text{ASF} = \frac{1}{\text{NumShots}} \sum_{i=1}^{\text{NumShots}} \text{ASF}(F_i, KF_i) \tag{3.18}$$

Βαθμός Ανακατασκευής της Λήψης (Shot Reconstruction Degree): Δοσμένης της ακολουθίας KF των αντιπροσωπευτικών καρέ κάθε λήψης επιχειρείται η ανακατασκευή όλων των καρέ της συγκεκριμένης λήψης χρησιμοποιώντας έναν αλγόριθμο παρεμβολής IA(). Κάθε καρέ ανακατασκευάζεται με βάση ένα ζευγάρι αντιπροσωπευτικών καρέ ως εξής:

$$\tilde{F}_n = \text{IA}(KF_{n_j}, KF_{n_{j+1}}) \tag{3.19}$$

Στη συνέχεια, τα ανακατασκευασμένα καρέ συγκρίνονται με τα αντίστοιχα πραγματικά και προκύπτει ο βαθμός επιτυχίας της ανακατασκευής:

$$\text{SRD}(F, KF) = \sum_{n=1}^{N} \text{Sim}(F_n, \tilde{F}_n) \tag{3.20}$$

όπου $\text{Sim}(F_n, \tilde{F}_n) = \log(\text{MaxDiff}/\text{Diff}(F_n, \tilde{F}_n))$. Η απόσταση Diff των δύο καρέ ορίζεται και πάλι όπως στην εξίσωση (3.11) και MaxDiff η μέγιστη τιμή που η απόσταση αυτή λαμβάνει. Αντίστοιχα με το ASF, πρέπει να ληφθεί η μέση τιμή του SRD για την αξιολόγηση ολόκληρης της ταινίας:

$$\text{SRD} = \frac{1}{\text{NumShots}} \sum_{i=1}^{\text{NumShots}} \text{SRD}(F_i, KF_i) \tag{3.21}$$

Με βάση αυτά τα τρία κριτήρια αξιολογήθηκαν οι δύο μέθοδοι εξαγωγής αντιπροσωπευτικών καρέ που αναλύθηκαν στην ενότητα 3.4. Για τις παραμέτρους των αλγορίθμων που χρησιμοποιήθηκαν, $\epsilon = 0.09$ για την πρώτη μέθοδο και $\lambda = 0.005$ για τη δεύτερη μέθοδο, τα αποτελέσματα φαίνονται στον Πίνακα 3.2. Η μέθοδος της μη γραμμικής χρονικής δειγματοληψίας διατηρεί μεγαλύτερο ποσοστό καρέ σε κάθε λήψη και έτσι επιτυγχάνει καλύτερα ποσοστά ανακατασκευής. Αντιθέτως, η μέθοδος Spectral Clustering διατηρεί μόνο το 2% περίπου των καρέ σε κάθε λήψη και παρόλα αυτά πετυχαίνει και αυτή πολύ υψηλά ποσοστά ASF και SRD. Για το λόγο αυτό προτιμάται στο πλαίσιο της παρούσας εργασίας και τα αντιπροσωπευτικά καρέ που εξήχθησαν με αυτή τη μέθοδο χρησιμοποιούνται για την περαιτέρω επεξεργασία. Ενδεικτικά, στον Πίνακα 3.3 φαίνεται ο μέσος αριθμός των αντιπροσωπευτικών καρέ που διατηρείται για κάθε λήψη με τη μέθοδο Spectral Clustering.

3.6 Συμπεράσματα

Και οι δύο μέθοδοι που παρουσιάστηκαν για την κατάτμηση σε λήψεις δίνουν πολύ υψηλά αποτελέσματα, ακόμη και αν χρησιμοποιηθούν

	CR	**ASF**	**SRD**
Non-Linear Sampling	94.14%	96.21%	377.7dB
Spectral Clustering	98.76%	91.10%	201.1dB

Πίνακας 3.2: Δείκτες αξιολόγησης των αντιπροσωπευτικών καρέ που εξήχθησαν από την ταινία Gone With The Wind.

	Mean Number of Key Frames
BMI	2
CHI	2
CRA	3
DEP	2
FNE	2
GLA	2
LOR	2

Πίνακας 3.3: Μέσος αριθμός αντιπροσωπευτικών καρέ για κάθε ταινία της βάσης, με τη μέθοδο Spectral Clustering.

μεμονωμένα. Ωστόσο, κρίνεται απαραίτητο να συνδυαστούν, ώστε να εντοπίζονται όσο το δυνατόν περισσότερες λήψεις και, μάλιστα, να εντοπίζονται και κάποιες από τις ομαλές μεταβάσεις. Για το λόγο αυτό, επιλέχτηκαν ως παράμετροι τα $(\text{Thres}_{RGB}, \text{Thres}_{ECR}) = (0.58, 0.20)$, που δίνουν συγχρόνως υψηλό ποσοστό Recall και F_1.

Όσον αφορά την εξαγωγή των αντιπροσωπευτικών καρέ, και οι δύο μέθοδοι επιτυγχάνουν αρκετά υψηλή συμπύκνωση της πληροφορίας, παράλληλα με υψηλή ικανότητα ανάκατασκευής της λήψης. Η μέθοδος που βασίζεται στον αλγόριθμο του Spectral Clustering προτιμάται για τον υψηλότερο βαθμό συμπίεσης που πετυχαίνει, το οποίο σημαίνει μικρότερος αριθμός αντιπροσωπευτικών καρέ για κάθε λήψη, και, άρα, ταχύτερη επεξεργασία στα επόμενα στάδια.

Κεφάλαιο 4

Κατάτμηση σε Σκηνές (Scenes)

Εφόσον έχουν εντοπιστεί τα χρονικά όρια των λήψεων και για κάθε μια από αυτές έχουν εξαχθεί τα αντίστοιχα αντιπροσωπευτικά καρέ, το επόμενο στάδιο της επεξεργασίας είναι η σωστή εξαγωγή των χρονικών ορίων των σκηνών. Όπως προαναφέρθηκε, οι σκηνές αποτελούνται από ένα σύνολο διαδοχικών λήψεων οι οποίες συνδέονται μεταξύ τους νοηματικά. Για το σωστό εντοπισμό τους πρέπει να γίνει κατανοητή η δομή μιας σκηνής, η οποία περιλαμβάνει είτε λήψεις με παρόμοιο οπτικό περιεχόμενο είτε λήψεις που επαναλαμβάνονται με κάποιο συγκεκριμένο μοτίβο. Πέρα από τα αντιπροσωπευτικά καρέ, τα οποία συγκρίνονται μεταξύ τους για να ομαδοποιηθούν οι λήψεις και να σχηματιστούν οι σκηνές, χρησιμοποιείται κάποιες φορές και η πληροφορία που έχει εξαχθεί από την ευθυγράμμιση σεναρίου - υποτίτλων για τον προσδιορισμό των χρονικών ορίων των σκηνών.

Στο παρόν κεφάλαιο θα παρουσιαστούν μέθοδοι της βιβλιογραφίας για την κατάτμηση σε σκηνές και θα αξιολογηθούν στη βάση των 7 ταινιών. Οι μέθοδοι ακολουθούν διαφορετικές προσεγγίσεις για την επίτευξη της κατάτμησης. Κεντρικό ρόλο παίζουν οι μέθοδοι που βασίζονται στο διαμερισμό γράφων. Οι περισσότερες χρησιμοποιούν χαμηλού επιπέδου χαρακτηριστικά (ιστογράμματα χρώματος) για να υπολογίσουν την ομοιότητα μεταξύ διαφορετικών λήψεων. Ωστόσο, αναλύεται και μια μέθοδος που εξάγει τοπικούς περιγραφητές από κάθε καρέ και ομαδοποιεί λήψεις με κοινούς περιγραφητές.

Όπως θα φανεί, όλες οι μέθοδοι οδηγούν σε μια υπερκατάτμηση της ταινίας, δημιουργώντας μεγάλο αριθμό σκηνών και μικρής διάρκειας σκηνές. Στόχος μας είναι για κάθε μέθοδο να βρεθεί ένα σύνολο παραμέτρων που να οδηγεί σε ικανοποιητικά αποτελέσματα για όλες τις ταινίες της βάσης.

4.1 Υπάρχουσες Μέθοδοι

4.1.1 Όρια από Χρονικά Ευθυγραμμισμένο Σενάριο

Η μέθοδος αυτή [18] χρησιμοποιεί αποκλειστικά την πληροφορία που έχει προκύψει από την ευθυγράμμιση του σεναρίου με τους υπότιτλους. Συγκεκριμένα, εντοπίζει τις αλλαγές σκηνών, όπως αυτές περιγράφονται στο σενάριο με τη χρήση αναγνωριστικών (EXT., INT.), καθώς και τα χρονικά διαστήματα που τους έχουν ανατεθεί, $[T_{start}, T_{end}]$. Στη συνέχεια, αν μόνο μια αλλαγή λήψης έχει εντοπιστεί στο διάστημα $[T_{start}, T_{end}]$ τότε αυτή θεωρείται και το όριο μεταξύ των δύο σκηνών. Σε περίπτωση που περισσότερες από μια αλλαγές λήψεων εντοπίζονται στο συγκεκριμένο διάστημα (συμβαίνει με λήψεις οι οποίες δεν περιέχουν διαλόγους), τότε συγκρίνονται οι λήψεις μεταξύ τους (εξίσωση (4.1)) και αυτές που παρουσιάζουν μεγάλη ομοιότητα, ή αντίστοιχα μικρή ανομοιότητα, θεωρείται πως ανήκουν στην ίδια σκηνή, οπότε διαχωρίζονται με τις υπόλοιπες και προκύπτει το όριο της σκηνής.

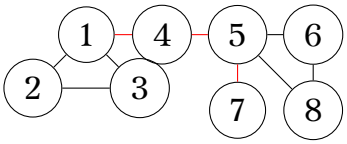
$$VisDiss(S_i, S_j) = d(S_i, S_j) \approx \min_{f_l \in KF_i, f_m \in KF_j} D(f_l, f_m) \tag{4.1}$$

Όπου το D έχει οριστεί στην εξίσωση (3.11) και KF_k είναι το σύνολο των αντιπροσωπευτικών καρέ της $k-$οστής λήψης.

4.1.2 Γράφοι Οπτικών Μεταβάσεων Σκηνών (Visual Scene Transition Graphs - VSTGs)

Η μέθοδος αυτή [28] εκμεταλλεύεται το γεγονός οτι σε μια σκηνή πολλές φορές υπάρχει εναλλαγή παρόμοιων λήψεων, για παράδειγμα σε μια σκηνή διαλόγου η κάμερα εναλλάσσεται ανάμεσα στους δύο ομιλητές. Έτσι, όλες οι λήψεις με παρόμοιο οπτικό περιεχόμενο μπορούν να ομαδοποιηθούν και τα cluster που θα προκύψουν να αποτελέσουν τους κόμβους ενός κατευθυνόμενου γράφου, του οποίου οι ακμές θα προσδιορίζουν τις μεταβάσεις μεταξύ των λήψεων. Αφού δημιουργηθεί ο γράφος αυτός κάθε σκηνή θα αποτελεί σε αυτόν ένα ανεξάρτητο υπογράφημα. Επομένως, η διαδικασία προσδιορισμού των χρονικών ορίων των σκηνών έγκειται στον εντοπισμό των ακμών τομής (cut edges) [2], οι οποίες χωρίζουν τον γράφο σε ανεξάρτητα υπογραφήματα. Ένα παράδειγμα cut edges φαίνεται στο Σχήμα 4.1.

Η ομαδοποίηση των λήψεων, ώστε να προκύψουν cluster, υλοποιείται με χρήση του αλγορίθμου που ακολουθεί. Σε κάθε cluster επιτρέπεται να ανήκουν λήψεις των οποίων η μεταξύ τους χρονική απόσταση d_t είναι



Σχήμα 4.1: Οι γέφυρες ενός μη κατευθυνόμενου γράφου

μικρότερη από ένα κατώφλι T. Η ανομοιότητα μεταξύ των cluster C_i και C_j υπολογίζεται από τη Σχέση (4.2).

$$\hat{d}_{max}(C_i, C_j) = \max_{S_l \in C_i, S_k \in C_j} \hat{d}(S_l, S_k) \tag{4.2}$$

Όπου

$$\hat{d}(S_l, S_k) = \left\{ \begin{array}{ll} d(S_l, S_k) & \text{αν } d_t(S_l, S_k) \leq T \\ \infty & \text{αλλιώς} \end{array} \right\}. \tag{4.3}$$

Με το d όπως έχει οριστεί στη Σχέση (4.1). Ο αλγόριθμος είναι ο ακόλουθος:

Βήμα 1 Αρχικά, κάθε λήψη αντιστοιχείται σε ένα cluster, οπότε $NumClusters \leftarrow NumShots$.

Βήμα 2 Τερματισμός όταν $\hat{d}_{max}(A, B) > \delta$ για όλα τα cluster A, B ή $NumClusters = 1$.

Βήμα 3 Εντοπισμός των cluster R και S που παρουσιάζουν τη μέγιστη ομοιότητα μεταξύ τους σε σχέση με τα υπόλοιπα ζευγάρια, δηλαδή $\hat{d}_{max}(R, S) \leqslant \hat{d}_{max}(A, B)$, για όλα τα cluster A, B.

Βήμα 4 Ενοποίηση των cluster R και S σε ένα νέο και $NumClusters \leftarrow NumClusters - 1$.

Βήμα 5 Επιστροφή στο Βήμα 2.

Η παράμετρος δ κανονικοποιείται στο διάστημα $[0, 1]$ και ορίζει την ελάχιστη ομοιότητα που απαιτείται για να ανήκουν δύο λήψεις στο ίδιο cluster. Στις ακραίες περιπτώσεις, δηλαδή $\delta = 0$ και $\delta = 1$, $NumClusters = 1$ και $NumClusters = NumShots$, αντίστοιχα.

4.1.3 Εντοπισμός Επαναλαμβανόμενων Μοτίβων

Ο τρόπος με τον οποίο μοντάρεται μια ταινία υποδεικνύει πως μέσα σε κάθε σκηνή θα υπάρχουν κάποια επαναλαμβανόμενα μοτίβα λήψεων. Για το λόγο αυτό στο [4], μετά την ομαδοποίηση των λήψεων σε cluster και την ανάθεση μιας ετικέτας στις λήψεις κάθε cluster,

εφαρμόζεται ένας αλγόριθμος ευθυγράμμισης [15], ώστε να εντοπιστούν αυτά τα επαναλαμβανόμενα μοτίβα. Σε κάθε επανάληψη του αλγορίθμου ευθυγραμμίζονται μεταξύ τους οι υποακολουθίες λήψεων $X_1 = L_1 L_2 \dots L_w$ και $X_2 = K_1 K_2 \dots K_w$, όπου $L_i, K_i \in ClusterLabels$ με $i \in [1, \dots, w]$ και εξάγεται ένα σκορ ευθυγράμμισης. Υψηλό σκορ ευθυγράμμισης υποδηλώνει την ύπαρξη κοινού μοτίβου και κατατάσσει τις λήψεις στην ίδια σκηνή.

Για να επιτευχθεί η ευθυγράμμιση και να υπολογιστεί το σκορ, κατασκευάζεται ένας $(w+1) \times (w+1)$ πίνακας N, όπου κάθε στοιχείο $N(i, j)$ εκφράζει το σκορ της βέλτιστης ευθυγράμμισης μεταξύ των ακολουθιών $X_1(1 \dots i)$ και $X_2(1 \dots j)$. Υπάρχουν τρεις εναλλακτικές για το αποτέλεσμα της ευθυγράμμισης μεταξύ δύο στοιχείων $X_1(i)$ και $X_2(j)$:

1. Το $X_1(i)$ αντιστοιχίζεται στο $X_2(j)$.
2. Το $X_1(i)$ αντιστοιχίζεται σε ένα κενό.
3. Το $X_2(j)$ αντιστοιχίζεται σε ένα κενό.

Επομένως, το σκορ ευθυγράμμισης θα λαμβάνει τη μέγιστη από τις ακόλουθες τιμές:

$$N(i, j) = \max \left\{ \begin{array}{l} N(i-1, j-1) + S(X_1(i), X_2(j)) \\ N(i-1, j) - d \\ N(i, j-1) - d \end{array} \right\} \tag{4.4}$$

Ο πίνακας S, μεγέθους $NumClusters \times NumClusters$ ονομάζεται *πίνακας αντικατάστασης* και υποδηλώνει τα κόστη αντικατάστασης - ευθυγράμμισης μιας λήψης που ανήκει στο cluster C_i με μια λήψη από το cluster C_j. Η παράμετρος d ορίζει το κόστος ύπαρξης ενός κενού.

Για τον ορισμό του πίνακα S λαμβάνονται υπόψιν τόσο η χρωματική ομοιότητα των δύο cluster όσο και η πιθανότητα αυτά τα δύο να ανήκουν στο ίδιο μοτίβο λήψεων. Για την αναπαράσταση αυτών των δύο μεγεθών ορίζονται οι πίνακες ομοιότητας των cluster (Cluster Similarity Matrix - CSM) και πιθανοτήτων ζευγών λήψεων (Pair Probability Matrix - PPM).

Cluster Similarity Matrix: Για την κατασκευή αυτού του πίνακα απαιτείται να εξαχθεί από κάθε cluster μια αντιπροσωπευτική λήψη, κατά τρόπο ανάλογο με την εξαγωγή των αντιπροσωπευτικών καρέ. Στην εξίσωση (4.1) έχει οριστεί η οπτική ανομοιότητα μεταξύ δύο λήψεων. Για τον υπολογισμό της οπτικής ομοιότητας μεταξύ δύο λήψεων λαμβάνεται η μέγιστη χρωματική ομοιότητα που εμφανίζουν τα αντιπροσωπευτικά τους καρέ, όπως αυτή έχει οριστεί στην εξίσωση (3.10):

$$VisSim(S_i, S_j) = \max_{f_l \in KF_i, f_m \in KF_j} ColSim(f_l, f_m) \tag{4.5}$$

Με βάση αυτή τη σχέση, ως αντιπροσωπευτική λήψη, m_i, κάθε cluster C_i ορίζεται αυτή που εμφανίζει τη μέγιστη μέση ομοιότητα με τις υπόλοιπες λήψεις του cluster. Έχοντας εντοπίσει τις αντιπροσωπευτικές λήψεις, ο πίνακας CSM ορίζεται όπως ακολουθεί:

$$CSM(i,j) = VisSim(m_i, m_j) \tag{4.6}$$

Pair Probability Matrix: Ο πίνακας αυτός εκφράζει την πιθανότητα μια λήψη του cluster C_i να ακολουθείται από μια λήψη του cluster C_j. Όσο μεγαλύτερη είναι αυτή η πιθανότητα, τόσο πιο πιθανό είναι αυτά τα δύο cluster να ανήκουν στην ίδια σκηνή, οπότε το κόστος ευθυγράμμισής τους θα είναι μικρότερο. Η πιθανότητα αυτή μπορεί να προσεγγιστεί από τη συχνότητα εμφάνισης του συγκεκριμένου ζεύγους.

$$PPM(i,j) = \frac{1}{NumShots - 1}\{\#pairs(L_1 = C_i, L_2 = C_j)\} \tag{4.7}$$

Στη σχέση (4.7) $NumShots$ είναι ο συνολικός αριθμός λήψεων και L_1, L_2 το πρώτο και δεύτερο στοιχείο, αντίστοιχα, ενός ζεύγους λήψεων.

Πίνακας Αντικατάστασης: Έχοντας κατασκευάσει τους δύο παραπάνω πίνακες, ορίζουμε τον πίνακα αντικατάστασης ως εξής:

$$S(i,j) = \begin{cases} CSM(i,j) + PPM(i,j) & \text{αν } i = j \\ -\alpha(1 - CSM(i,j)) - \beta(1 - PPM(i,j)) & \text{αν } i \neq j \end{cases} \tag{4.8}$$

Οι παράμετροι α, β είναι οι συντελεστές βαρύτητας που αποδίδονται στην χρωματική ομοιότητα και στην χρονική εγγύτητα, αντίστοιχα, και για τους οποίους ισχύει $\alpha + \beta = 1$.

Αλγόριθμος Ευθυγράμμισης Needlaman-Wunsch [15]

Έχοντας υπολογίσει τους παραπάνω πίνακες ο αλγόριθμος απαιτεί την κατασκευή του πίνακα N καθώς και ενός ακόμη πίνακα T, ο οποίος μετά την κατασκευή του πίνακα N καθοδηγεί στον εντοπισμό της βέλτιστης ευθυγράμμισης και για το λόγο αυτό ονομάζεται *traceback* πίνακας. Ο αλγόριθμος υλοποιείται ως εξής:

Βήμα 1 Αρχικοποίηση πινάκων N και T:

$N =$

0	$-d$	$\cdots$	$-wd$
$-d$			
$\vdots$			
$-wd$			

$T =$

$done$	$left$	$\cdots$	$left$
up			
$\vdots$			
up			

Βήμα 2 Γέμισμα του πίνακα N με βάση την εξίσωση (4.4). Σε κάθε στοιχείο (i, j), το κελί $T(i, j)$ λαμβάνει μια από τις τρεις πιθανές τιμές (diag, up, left) αναλόγως από πού έχει προέλθει η τιμή του $N(i, j)$.

Βήμα 3 Αφού έχουν γεμίσει και οι δύο πίνακες, τότε ξεκινάει η ευθυγράμμιση των δύο ακολουθιών αρχίζοντας από το τελευταίο κελί του πίνακα T που συμπληρώθηκε, δηλαδή από το κελί $(w + 1, w + 1)$, και ακολουθώντας τη διαδρομή που αυτό καταδεικνύει, καταλήγοντας στο κελί done, όπως φαίνεται στον Πίνακα 4.1.

Βήμα 4 Στο μονοπάτι που προκύπτει κάθε up αντιστοιχεί σε ένα κενό στην ακολουθία X_1, κάθε left σε ένα κενό στην ακολουθία X_2, ενώ κάθε diag σημαίνει πως τα στοιχεία $X_1(i)$ και $X_2(j)$ έχουν ευθυγραμμιστεί.

Βήμα 5 Το τελικό σκορ F της ευθυγράμμισης προκύπτει:

$$F = S(\text{matches}) - (\#\text{gaps}) \times d \tag{4.9}$$

Όπου, $S(\text{matches})$ οι τιμές του πίνακα αντικατάστασης στα στοιχεία που έχουν ευθυγραμμιστεί και #gaps = #up + #left. Τα τοπικά ελάχιστα αυτού του σκορ υποδεικνύουν πως δυο ακολουθίες λήψεων δεν παρουσιάζουν μεγάλη ομοιότητα και, επομένως, θα μπορούσαν να ανήκουν σε διαφορετικές σκηνές.

done	*left*	*left*	*left*
up	*diag*	*left*	*up*
up	*up*	*diag*	*left*
up	*left*	*up*	*up*

Πίνακας 4.1: Traceback πίνακας και εύρεση της βέλτιστης ευθυγράμμισης

4.1.4 Γράφοι Ομοιότητας Λήψεων (Shot Similarity Graphs - SSGs)

Η μέθοδος αυτή [20] βασίζεται σε μεθόδους διαμερισμού γράφων, λαμβάνοντας υπόψιν την ομοιότητα κάθε ζεύγους λήψεων και όχι μόνο ζευγών διαδοχικών λήψεων. Κατασκευάζει ένα μη κατευθυνόμενο γράφο

με βάρη, όπου κάθε λήψη αποτελεί έναν κόμβο και οι ακμές αναπαριστούν ένα σταθμισμένο δείκτη ομοιότητας μεταξύ των λήψεων, ο οποίος λαμβάνει υπόψιν την πληροφορία χρώματος και κίνησης. Με τον τρόπο αυτό εισέρχεται και η έννοια της κίνησης στη διαδικασία της κατάτμησης και μπορούν να αναγνωριστούν τόσο σκηνές που διαδραματίζονται σε ένα συγκεκριμένο περιβάλλον (setting) και έχουν κοινά χρώματα όσο και σκηνές δράσης που χαρακτηρίζονται από έντονη κίνηση, και όχι τόσο χρωματική ομοιότητα.

Αρχικά, για κάθε λήψη S_z ορίζεται ένας δείκτης που δείχνει το περιεχόμενο κίνησης της λήψης, ονομάζεται Shot Motion Content Feature και συμβολίζεται με Mot_z. Ο ορισμός του είναι ο ακόλουθος:

$$Mot_z = \frac{1}{b-a} \sum_{f=a}^{b-1} D(f, f+1) \tag{4.10}$$

όπου a, b το πρώτο και το τελευταίο καρέ μιας λήψης, αντίστοιχα, και D όπως έχει οριστεί στην εξίσωση (3.11). Κάνοντας χρήση αυτού του δείκτη μπορεί να οριστεί η ομοιότητα που εμφανίζουν δύο λήψεις S_i και S_j με βάση το κινητικό τους περιεχόμενο, ως εξής:

$$MotSim(S_i, S_j) = \frac{2 \cdot \min(Mot_i, Mot_j)}{Mot_i + Mot_j} \tag{4.11}$$

Έχοντας ήδη ορίσει την οπτική ομοιότητα δύο λήψεων στην εξίσωση (4.5) και την ομοιότητα κινητικού περιεχομένου είναι δυνατόν να οριστεί ένα νέο μέτρο της ομοιότητας μεταξύ δύο λήψεων, ως εξής:

$$ShotSim(S_i, S_j) = \alpha \cdot VisSim(S_i, S_j) + \beta \cdot MotSim(S_i, S_j) \tag{4.12}$$

όπου α και β τα βάρη που αποδίδονται σε καθένα από τα δύο μέτρα ομοιότητας και για τα οποία ισχύει $\alpha + \beta = 1$.

Εφόσον έχει υπολογιστεί η ομοιότητα μεταξύ όλων των λήψεων μπορεί να κατασκευαστεί ο γράφος SSG, έστω $G = (V, E)$, ώστε κάθε λήψη i να αποτελεί έναν κόμβο v_i και όλοι οι κόμβοι να συνδέονται μεταξύ τους. Αν $e(i, j) \in E$ είναι η ακμή μεταξύ των κόμβων i και j, τότε σε αυτή αντιστοιχείται ένα βάρος $W(i, j)$ το οποίο δείχνει την πιθανότητα οι δύο αυτές λήψεις να ανήκουν στην ίδια σκηνή. Σε αυτό το συντελεστή παίζει ρόλο και η χρονική εγγύτητα δύο λήψεων, μιας και δυο χρονικά απομακρυσμένες λήψεις δύσκολα θα ανήκουν στην ίδια σκηνή. Για το λόγο αυτό ο συντελεστής W ορίζεται ως εξής:

$$W(S_i, S_j) = w(i, j) \times ShotSim(S_i, S_j) \tag{4.13}$$

όπου $w(i,j)$ είναι μια φθίνουσα συνάρτηση της χρονικής απόστασης δύο λήψεων. Επιλέγεται μια εκθετική συνάρτηση της μορφής:

$$w(i,j) = \exp\big(-\frac{1}{d}\cdot\Big|\frac{m_i - m_j}{\sigma}\Big|^2\big) \tag{4.14}$$

όπου $|m_i - m_j|$ είναι η χρονική απόσταση των μεσαίων καρέ δύο λήψεων και σ η τυπική απόκλιση της διάρκειας των λήψεων σε ολόκληρη την ταινία.

Η κατάτμηση σε σκηνές βασίζεται στο διαχωρισμό του γράφου ομοιότητας των λήψεων (SSG) με βάση την τεχνική των Normalized Cuts που προτείνεται στο [22]. Η τεχνική αυτή επιτυγχάνει να μεγιστοποιεί την ανομοιότητα (disassociation) μεταξύ των υπογράφων που προκύπτουν, ενώ ταυτόχρονα μεγιστοποιεί την ομοιότητα (association) μέσα σε κάθε υπογράφο. Αν επιδιώκεται να χωριστεί ένας γράφος $G = (V, E)$ σε δύο υπογραφήματα $G' = (V', E')$ και $G'' = (V'', E'')$ για τα οποία να ισχύει $V' \cup V'' = V$ και $V' \cap V'' = \emptyset$, τότε αρκεί να υπολογιστεί για κάθε πιθανό διαχωρισμό ένα μέγεθος ανομοιότητας μεταξύ των δύο τμημάτων που προκύπτουν. Το μέγεθος αυτό ονομάζεται cut και ορίζεται ως εξής:

$$cut(V', V'') = \sum_{i \in V', j \in V''} W(i,j) \tag{4.15}$$

Ο διαχωρισμός που επιτυγχάνει την ελάχιστη τιμή cut θεωρείται πως αποτελεί και τη βέλτιστη λύση. Ωστόσο, όταν πρόκειται ο γράφος να διαχωριστεί σε περισσότερα του ενός υπογραφήματα η μέθοδος αυτή δεν δίνει ένα καθολικό βέλτιστο, καθώς τείνει να δημιουργεί υπογραφήματα από απομονωμένους κόμβους. Για το λόγο αυτό εισάγεται η έννοια της κανονικοποίησης, αφού πρώτα οριστεί ένα μέγεθος συσχέτισης μεταξύ των κόμβων ενός υπογραφήματος $G_n = (V_n, E_n)$ με όλους τους κόμβους του γράφου, ως εξής:

$$assoc(V_n, V) = \sum_{i \in V_n, j \in V} W(i,j) \tag{4.16}$$

Έχοντας αυτό τον ορισμό το Normalized Cut ορίζεται ως εξής:

$$Ncut(V', V'') = \frac{cut(V', V'')}{assoc(V', V)} + \frac{cut(V', V'')}{assoc(V'', V)} \tag{4.17}$$

Στο [22] προτείνεται μια μέθοδος επίλυσης του προβλήματος εύρεσης του ελάχιστου $Ncut$ για το βέλτιστο διαχωρισμό που περιλαμβάνει την εύρεση κάποιων ιδιοτιμών-ιδιοδιανυσμάτων, με τρόπο παρόμοιο με το Spectral Clustering. Ωστόσο, για το συγκεκριμένο πρόβλημα κατάτμησης σε σκηνές πρέπει να προστεθεί και ο ακόλουθος περιορισμός:

$$(i < j \quad \text{or} \quad i > j) \quad \text{for all } v_i \in V', v_j \in V'' \tag{4.18}$$

Ο περιορισμός αυτός εξασφαλίζει πως μόνο διαδοχικές λήψεις θα ομαδοποιούνται ώστε να συγκροτήσουν μια σκηνή και, έτσι, επιτυγχάνει ταχύτερη επίλυση του προβλήματος. Έτσι, μπορεί να εφαρμοστεί ένας αναδρομικός αλγόριθμος, ώστε σε κάθε βήμα να εντοπίζεται η ελάχιστη τιμή του $Ncut$ και ο γράφος να διαχωρίζεται σε δύο υπογραφήματα. Η διαδικασία επαναλαμβάνεται όσο το ελάχιστο $Ncut$ είναι μικρότερο από ένα κατώφλι λ. Όσο μεγαλύτερη είναι η τιμή αυτού του κατωφλίου τόσο περισσότερες σκηνές εντοπίζονται.

4.1.5 Bag of Visual Words

Η προηγούμενη μέθοδος εισάγει την έννοια της κίνησης για να μπορέσει να εντοπίσει λήψεις κοινού θεματικού περιεχομένου, που, όμως, δεν παρουσιάζουν χρωματική ομοιότητα λόγω της αλλαγής περιβάλλοντος. Σε σκηνές τέτοιου τύπου, για παράδειγμα καταδίωξη με αυτοκίνητα, υπάρχουν λήψεις που προέρχονται από διαφορετικές κάμερες και διαφορετικές οπτικές γωνίες και, κατά συνέπεια, οι χρωματικές κατανομές των λήψεων διαφέρουν δραματικά μεταξύ τους. Ωστόσο, υπάρχουν αντικείμενα ή σημεία που εμφανίζονται επανειλημμένα σε διαδοχικές λήψεις. Ο εντοπισμός αυτών των αντικειμένων σε διαδοχικές λήψεις δείχνει οτι οι λήψεις αυτές συνδέονται νοηματικά και, άρα, ανήκουν στην ίδια σκηνή.

Για τον εντοπισμό των αντικειμένων πρέπει να χρησιμοποιηθούν τοπικοί περιγραφητές που παραμένουν αμετάβλητοι σε αλλαγές κλίμακας και προσανατολισμού. Ένας τέτοιος περιγραφητής είναι ο περιγραφητής SIFT.

Αφού εξαχθούν οι περιγραφητές SIFT από κάθε αντιπροσωπευτικό καρέ μιας λήψης, ομαδοποιούνται σε cluster, καθένα από τα οποία αποτελεί μια οπτική λέξη. Στόχος είναι περιγραφητές του ίδιου αντικειμένου να ενταχθούν στο ίδιο cluster. Έτσι, κάθε καρέ θα χαρακτηρίζεται από ένα σύνολο περιγραφητων, γεγονός που οδηγεί στην υιοθέτηση της μεθόδου Bag of Words για την αναπαράσταση κάθε καρέ και κάθε λήψης. Αυτή η αναπαράσταση επιτρέπει το συσχετισμό διαδοχικών λήψεων με βάση το σημασιολογικό τους περιεχόμενο και τον εντοπισμό πιθανών αλλαγών σκηνής.

4.1.5.1 Περιγραφητής SIFT

Η εξαγωγή του περιγραφητή SIFT (Scale Invariant Feature Transform) βασίζεται σε τέσσερα βασικά στάδια [12].

Εντοπισμός υποψήφιων σημείων ενδιαφέροντος: Αυτά ορίζονται ως τα ακρότατα στο χώρο κλίμακας. Για να επιτευχθεί αυτό, η εικόνα

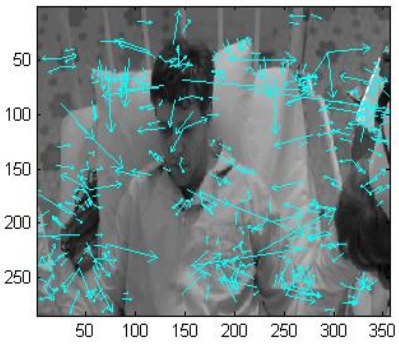
συνελίσσεται με Γκαουσιανές σε διαφορετική κλίμακα και εξάγονται οι διαφορές των Γκαουσιανών εικόνων (difference-of-Gaussians) από διαδοχικές κλίμακες. Ως σημεία ενδιαφέροντος θεωρούνται οι τοπικές κορυφές αυτών των διαφορών.

Απόρριψη ασταθών σημείων: Απορρίπτονται κάποια από τα εντοπισμένα σημεία ενδιαφέροντος και σε κάθε σημείο που απομένει προσαρμόζεται ένα μοντέλο που περιγράφει την κλίμακα και τη θέση του.

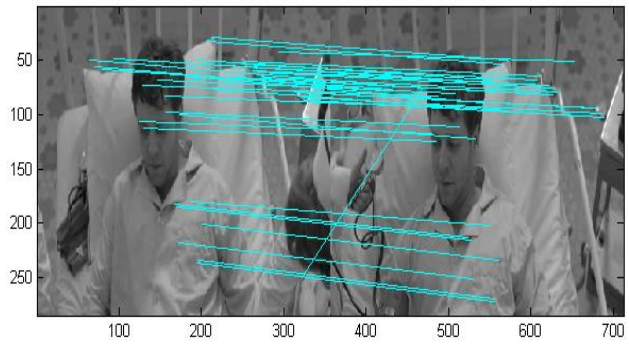
Ανάθεση προσανατολισμού: Σε κάθε σημείο ενδιαφέροντος ανατίθεται ένας ή περισσότεροι προσανατολισμοί, με βάση τις κατευθύνσεις των τοπικών gradient της εικόνας.

Εξαγωγή του περιγραφητή: Για κάθε σημείο ενδιαφέροντος εξάγεται ένας τοπικός περιγραφητής στην επιλεγμένη κλίμακα βασιζόμενος στα τοπικά διανύσματα κλίσης.

Στο Σχήμα 4.2 φαίνονται τα σημεία ενδιαφέροντος που έχουν εξαχθεί για μια εικόνα. Τα διανύσματα που εικονίζονται δείχνουν την κλίμακα, τον προσανατολισμό και την θέση των σημείων ενδιαφέροντος. Επιπλέον, στο Σχήμα 4.3 φαίνεται πώς γίνεται το ταίριασμα μεταξύ σημείων ενδιαφέροντος δύο διαφορετικών εικόνων. Να σημειωθεί πως οι εικόνες που απεικονίζονται ανήκουν σε δύο διαφορετικές λήψεις της ταινίας. Γίνεται ξεκάθαρο πως τα σημεία ενδιαφέροντος στην ανθρώπινη μορφή ταυτίζονται και ταιριάζονται μεταξύ τους, ενώ δεν υπάρχει αντιστοίχιση για τα σημεία των αντικειμένων δεξιά και αριστερά της ανθρώπινης μορφής.



Σχήμα 4.2: Περιγραφητής SIFT.



Σχήμα 4.3: Matching (ταίριασμα) των σημείων ενδιαφέροντος δύο εικόνων.

4.1.5.2 Μέθοδος

Στο [3] οι συγγραφείς χρησιμοποιούν μεθόδους από την επεξεργασία κειμένου για την κατάτμηση μιας ταινίας σε σκηνές. Από ολόκληρη την ταινία εξάγονται κάποιες οπτικές λέξεις (visual words) και για κάθε λήψη εξάγεται ένα ιστόγραμμα οπτικών λέξεων. Σύγκριση διαδοχικών ιστογραμμάτων δείχνει κατά πόσο διαδοχικές λήψεις μοιάζουν μεταξύ τους.

Αρχικά, η μέθοδος περιλαμβάνει την εξαγωγή ενός συνόλου περιγραφητών SIFT από κάθε αντιπροσωπευτικό καρέ μιας λήψης, έστω D_{kf}. Έτσι, κάθε λήψη S_i που αντιπροσωπεύεται από ένα σύνολο καρέ μεγέθους n, $KF_i = \{kf_{i_1}, \ldots, kf_{i_n}\}$, τελικά περιγράφεται από τους περιγραφητές όλων των αντιπροσωπευτικών καρέ της:

$$D_{S_i} = D_{kf_i} \cup \ldots\ldots D_{kf_n} \tag{4.19}$$

Εφόσον έχουν εξαχθεί οι περιγραφητές για κάθε λήψη της ταινίας, κατασκευάζεται ο περιγραφητής για ολόκληρη την ταινία: $D_S = D_{S_1} \cup D_{S_2} \cup \cdots \cup D_{S_N}$ και οι περιγραφητές ομαδοποιούνται σε k ομάδες χρησιμοποιώντας τον αλγόριθμο k-means[1]. Τα cluster που προκύπτουν από αυτή τη διαδικασία αποτελούν το οπτικό λεξιλόγιο και χρησιμοποιούνται για την κατασκευή του οπτικού ιστογράμματος, VH_i, για κάθε λήψη. Δεδομένου οτι μια λήψη έχει P περιγραφητές $D_{S_i} = \{d_1, \ldots, d_P\}$ (που έχουν προκύψει από τα αντίστοιχα αντιπροσωπευτικά καρέ) και έχουν ομαδοποιηθεί στα cluster $\{C_1, \ldots, C_k\}$, το οπτικό ιστόγραμμα για τη συγκεκριμένη λήψη υπολογίζεται ως εξής:

$$VH_i(l) = \frac{\{d_j \in C_l, j = 1, \ldots, P\}}{P} \tag{4.20}$$

όπου $l = 1, \ldots, k$ οι οπτικές λέξεις που έχουν προκύψει.

Μετά τη δημιουργία των οπτικών ιστογραμμάτων γίνεται μια χρονική εξομάλυνσή τους με ένα γκαουσιανό φίλτρο K_σ μηδενικής μέσης τιμής και τυπικής απόκλισης σ, ως εξής:

$$SH_t = \sum_{n=-\infty}^{\infty} (VH_{t-n}) K_\sigma (t-n) \tag{4.21}$$

Τέλος, υπολογίζονται οι αποστάσεις μεταξύ των SH διαδοχικών λήψεων και τα τοπικά μέγιστα λαμβάνονται ως αλλαγές σκηνής.

Για τον υπολογισμό της απόστασης μεταξύ διαδοχικών ιστογραμμάτων μπορούν να χρησιμοποιηθούν τρία διαφορετικά μετρικά:

Ευκλείδεια Απόσταση

$$V_i = \sqrt{\sum_{h=1}^{k} (SH_i(h) - SH_{i+1}(h))^2} \tag{4.22}$$

Τομή Ιστογραμμάτων

$$V_i = 1 - \frac{\sum_{h=1}^{k} \min(SH_i(h), SH_{i+1}(h))}{\sum_{h=1}^{k} SH_i(h)} \tag{4.23}$$

Chi-squared

$$V_i = 0.5 \frac{\sum_{h=1}^{k} (SH_i(h) - SH_{i+1}(h))^2}{\sum_{h=1}^{k} (SH_i(h) + SH_{i+1}(h))^2} \tag{4.24}$$

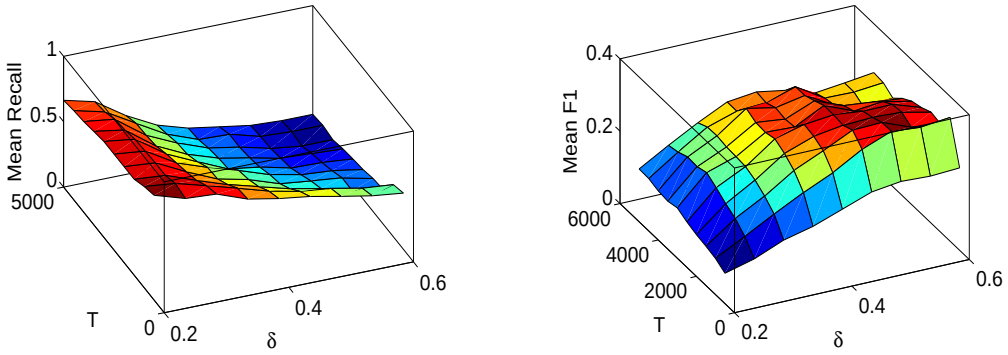
Οι συγγραφείς του [3] επιλέγουν να χρησιμοποιήσουν την ευκλείδεια απόσταση για τον υπολογισμό της απόστασης διαδοχικών ιστογραμμάτων.

4.2 Πειραματικά Αποτελέσματα

Η πρώτη μέθοδος κατάτμησης δεν ελέγχθηκε πάνω στη βάση των ταινιών, δεδομένου οτι δεν ήταν διαθέσιμα όλα τα σενάρια στη λεπτομερή τους μορφή. Οι περισσότερες από τις υπόλοιπες μεθόδους αξιολογήθηκαν πάνω στις ταινίες της βάσης και παρακάτω θα παρουσιαστούν οι μέσες τιμές των αποτελεσμάτων που επιτεύχθηκαν. Παράλληλα, θα εξαχθεί για κάθε μέθοδο ένα σύνολο παραμέτρων που οδηγεί στην επίτευξη των βέλτιστων αποτελεσμάτων. Στην ανάλυση που ακολουθεί, για την εξαγωγή των δεικτών αξιολόγησης ως Relevant Frames θεωρήθηκαν οι εντοπισμένες αλλαγές σκηνής που απείχαν έως 50 καρέ από μια επισημειωμένη αλλαγή σκηνής.

4.2.1 Γράφοι Οπτικών Μεταβάσεων Σκηνών

Τα αποτελέσματα αυτής της μεθόδου παρουσιάζονται στο Σχήμα 4.4. Οι παράμετροι που ελέγχονται είναι το χρονικό κατώφλι T και η ελάχιστη επιτρεπτή ανομοιότητα δύο cluster δ, που χρησιμοποιούνται κατά την ομαδοποίηση των λήψεων. Παρατηρείται πως οι παράμετροι, για τις οποίες παρατηρούνται υψηλές τιμές Recall ($\approx 80\%$), εμφανίζουν χαμηλά ποσοστά F_1 ($\approx 10\%$). Οι τιμές των παραμέτρων που οδηγούν στα προαναφερθέντα αποτελέσματα είναι $T = 500 - 1500$ και $\delta = 0.2 - 0.3$.



(i) Recall (ii) F1

Σχήμα 4.4: Δείκτες αξιολόγησης της κατάτμησης σε σκηνές με τη μέθοδο VSTGs ως προς τις παραμέτρους και δ στη βάση των 7 ταινιών.

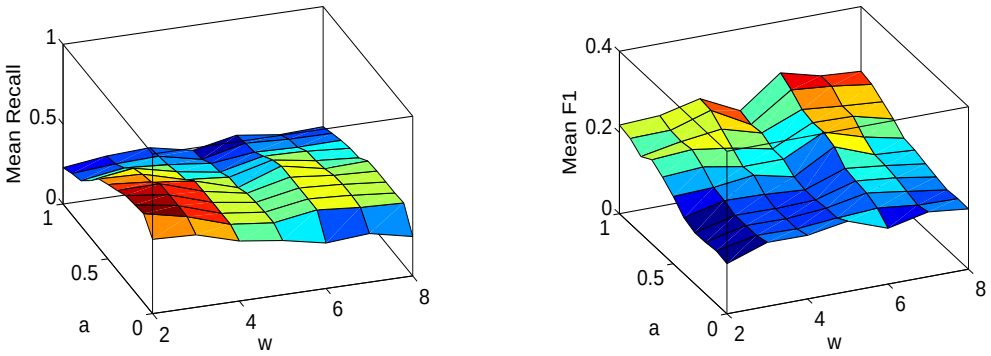
4.2.2 Εντοπισμός Επαναλαμβανόμενων Μοτίβων

Για τη μέθοδο αυτή απαιτείται να προσδιοριστούν κάποιες παράμετροι, όπως το d που αντιστοιχεί στο κόστος ύπαρξης ενός κενού στην ευθυγραμμισμένη ακολουθία, καθώς και οι παράμετροι α και β, που χρησιμοποιούνται για τον ορισμό του πίνακα αντικατάστασης S. Επιπλέον, όπως προτείνεται στο [4] για την ομαδοποίηση των λήψεων σε cluster υλοποιήθηκε και πάλι ένας αλγόριθμος φασματικής ομαδοποίησης, όπως αυτός που περιγράφηκε στην ενότητα 3.4.2, με μόνη διαφορά πως ο πίνακας ομοιότητας ορίστηκε ως εξής:

$$a(i, j) = VisSim(S_i, S_j) \tag{4.25}$$

όπου η οπτική ομοιότητα $VisSim$ μεταξύ δύο λήψεων έχει οριστεί στην εξίσωση (4.5). Επομένως, πρέπει να επιλεγεί και η παράμετρος λ, που καθορίζει πόσα θα είναι τα πρωτεύοντα ιδιοδιανύσματα. Τέλος, πρέπει να

προσδιοριστεί η τιμή του παραθύρου w που καθορίζει πόσες λήψεις πρέπει να ευθυγραμμίζονται σε κάθε πέρασμα του αλγορίθμου. Αρχικά, επιλέγεται $d = 1$ και $\lambda = 0.005$ και οι υπόλοιπες παράμετροι δοκιμάζονται σε ένα εύρος τιμών, ώστε να προκύψει ο καλύτερος δυνατός συνδυασμός.



(i) Recall (ii) F1

Σχήμα 4.5: Δείκτες αξιολόγησης της κατάτμησης σε σκηνές με τη μέθοδο επαναλαμβανόμενων προτύπων ως προς τις παραμέτρους w και a στη βάση των 7 ταινιών.

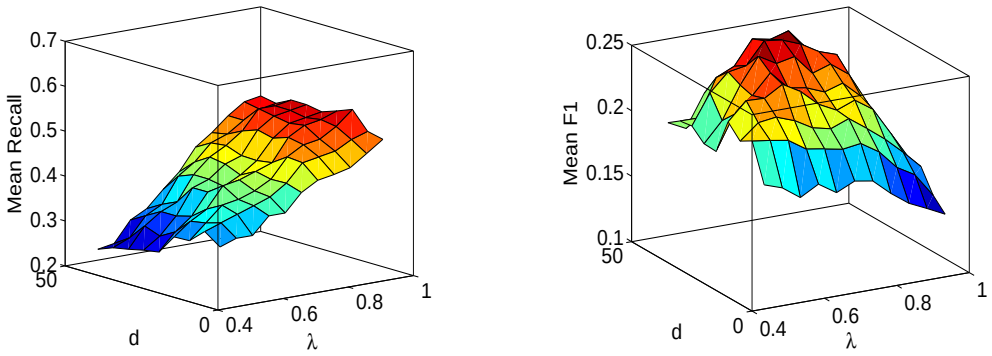
Παρατηρείται πως η μέθοδος αυτή δίνει χειρότερα αποτελέσματα από την προηγούμενη, με εμφανώς μειωμένο Recall. Τα καλύτερα αποτελέσματα λαμβάνονται για $a = 0.1$ και $w = 2$ και είναι Recall$= 58\%$ και $F_1 = 18\%$.

4.2.3 Γράφοι Ομοιότητας Λήψεων

Οι παράμετροι αυτής της μεθόδου είναι το d που εμφανίζεται στην εξίσωση (4.14) και το κατώφλι λ που καθορίζει πότε θα τερματιστεί ο διαχωρισμός του γράφου. Επίσης, να σημειωθεί πως στην εξίσωση (4.12) λήφθηκε $\alpha = \beta = 0.5$. Τα αποτελέσματα είναι αντίστοιχα με αυτά της προηγούμενης μεθόδου και εμφανίζονται στο Σχήμα 4.6. Η μέγιστη τιμή του Recall είναι 53%, η αντίστοιχη τιμή του F_1 είναι 19% και λαμβάνονται για $d = 20$ και $\lambda = 1$.

4.2.4 Bag of Visual Words

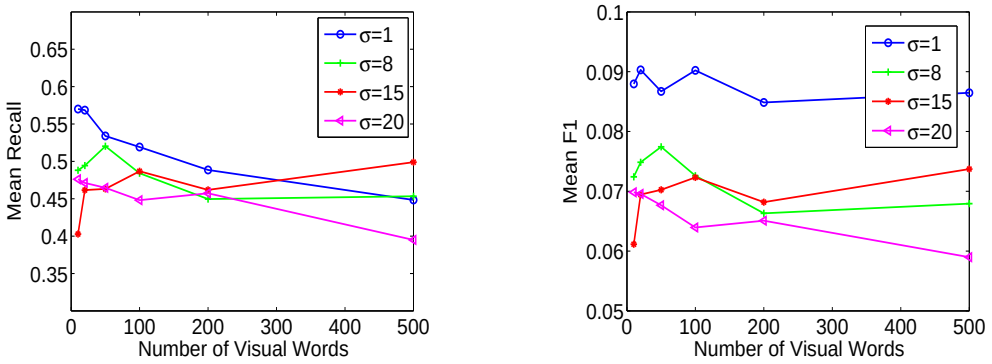
Τα αποτελέσματα αυτής της μεθόδου απεικονίζονται το Σχήμα 4.7 ως προς τον αριθμό των οπτικών λέξεων που χρησιμοποιούνται και την τυπική απόκλιση του γκαουσιανού φίλτρου. Τα αποτελέσματα φαίνεται να είναι χειρότερα από τα αντίστοιχα των προηγούμενων μεθόδων.



(i) Recall (ii) F1

Σχήμα 4.6: Δείκτες αξιολόγησης της κατάτμησης σε σκηνές με τη μέθοδο SSGs ως προς τις παραμέτρους d και λ στη βάση των 7 ταινιών.

Συγκεκριμένα, το Recall της μεθόδου λαμβάνει τιμές αντίστοιχες με αυτές των δύο προηγούμενων μεθόδων, αλλά οι εξαιρετικά χαμηλές τιμές του Precision οδηγούν σε χαμηλές τιμές F_1, πολύ χαμηλότερες από αυτές των προηγούμενων μεθόδων. Αξίζει να σημειωθεί πως τα καλύτερα αποτελέσματα επιτυγχάνονται για μικρή τιμή του σ και μέγεθος λεξιλογίου $50 - 150$ οπτικές λέξεις. Τέλος, οι τιμές των δεικτών αξιολόγησης φαίνεται να μην επηρεάζονται ιδιαίτερα από το μετρικό που θα χρησιμοποιηθεί για τον υπολογισμό της απόστασης διαδοχικών ιστογραμμάτων.



(i) Recall (ii) F1

Σχήμα 4.7: Δείκτες αξιολόγησης της κατάτμησης σε σκηνές με τη μέθοδο Bag Of Visual Words ως προς τον αριθμό των οπτικών λέξεων και την παράμετρο σ στη βάση των 7 ταινιών.

4.3 Συμπεράσματα

Από τις μεθόδους που αναλύθηκαν υψηλότερο Recall δίνει η μέθοδος των VSTGs και τα αποτελέσματα αυτής της μεθόδου επιλέγονται για την περαιτέρω ανάλυση που ακολουθεί στο επόμενο κεφάλαιο.

Όλες οι παραπάνω μέθοδοι εντοπίζουν με ακρίβεια τις σκηνές, αλλά χωρίζουν την ταινία σε πολύ μικρές υποενότητες. Κάποιες από τις υποενότητες αυτές παρουσιάζουν μια θεματική συνοχή και κατά κάποιο τρόπο διαχωρίζονται νοηματικά από τις προηγούμενες και τις επόμενες. Ωστόσο, υπάρχουν και εντοπισμένες αλλαγές οι οποίες δεν μπορούν να δικαιολογηθούν, καθώς χωρίζουν μια ενιαία θεματική ενότητα, όπως, για παράδειγμα, μια σκηνή διαλόγου.

Αξίζει, τέλος, να σημειωθεί πως σε όλες τις εφαρμογές, στις οποίες είναι απαραίτητη η κατάτμηση σε σκηνές, προτιμάται μια υπερκατάτμηση της ταινίας σε σχέση με μια κατάτμηση που εντοπίζει μόνο ένα μικρό αριθμό σωστών σκηνών. Παρόλ᾽αυτά θα επιχειρηθεί στο επόμενο κεφάλαιο να γίνει μια βελτίωση του αποτελέσματος της κατάτμησης, απορρίπτοντας κάποιες από τις λάθος εντοπισμένες σκηνές.

Κεφάλαιο 5

Βελτίωση του Αποτελέσματος της Κατάτμησης

Όπως παρατηρήθηκε στο προηγούμενο κεφάλαιο οι μέθοδοι κατάτμησης δίνουν ικανοποιητικά αποτελέσματα Recall αλλά χαμηλά ποσοστά Precision. Αυτό σημαίνει οτι γίνεται μια υπερκατάτμηση της ταινίας σε σκηνές. Επομένως, αναζητούνται τρόποι ώστε οι λάθος εντοπισμένες σκηνές να απορρίπτονται και να διατηρούνται μόνο οι σωστές. Έχουν προταθεί κάποιες μέθοδοι για την επίλυση αυτού του προβλήματος, οι οποίες θα παρουσιαστούν παρακάτω. Επιπλέον, σε αυτό το σημείο της επεξεργασίας επιχειρείται να εισαχθεί και η ακουστική πληροφορία στη διαδικασία της κατάτμησης, ώστε να ενισχύσει τα αποτελέσματα που έχουν προκύψει από την επεξεργασία αποκλειστικά της οπτικής πληροφορίας.

Για αυτό τον πειραματισμό θα θεωρήσουμε ως αποτέλεσμα του πρώτου σταδίου της κατάτμησησς σε σκηνές την κατάτμηση που προκύπτει από τη μέθοδο γράφων για το σετ παραμέτρων $(T, \delta) = (1000, 0.3)$. Επιπλέον, στα πειραματικά αποτελέσματα που εμφανίζονται σε αυτό το κεφάλαιο ως Retrieved Frames θεωρούνται οι εντοπισμένες αλλαγές σκηνής που απέχουν έως και 250 καρέ από κάποια επισημειωμένη αλλαγή σκηνής.

5.1 Απόρριψη Σκηνών Μικρής Διάρκειας

5.1.1 Υλοποίηση

Πολλές από τις σκηνές που έχουν προκύψει στο πρώτο στάδιο της επεξεργασίας είναι πολύ μικρής διάρκειας. Ουσιαστικά, πρόκειται για ενότητες στα πλαίσια της ευρύτερης σκηνής. Για να απορριφθούν αυτές οι τόσο μικρές σκηνές πρέπει να αποφασιστεί αν θα ενταχθούν στην

προηγούμενη ή στην επόμενη τους σκηνή. Πρέπει, επομένως, να βρεθεί η ομοιότητά τους με την προηγούμενη και την επόμενη σκηνή και να ενοποιηθούν με αυτή με την οποία εμφανίζουν τη μέγιστη ομοιότητα.

Έστω ότι στην ακολουθία σκηνών $\{Sc_{i-1}, Sc_i, Sc_{i+1}\}$ η σκηνή Sc_i είναι σύντομης διάρκειας και κρίνεται απαραίτητο να ενοποιηθεί είτε με την προηγούμενη είτε με την επόμενη. Υπολογίζεται ο συντελεστής ανομοιότητάς της με κάθε μια από τις γειτονικές της σκηνές, σύμφωνα με την εξίσωση (5.1), όπου d είναι η ανομοιότητα δύο λήψεων, όπως έχει οριστεί στην εξίσωση (4.1).

$$\text{SceneDiss}(Sc_i, Sc_j) = \max_{S_l \in Sc_i, S_k \in Sc_j} d(S_l, S_k) \tag{5.1}$$

Η σκηνή Sc_i θα ενσωματωθεί στη σκηνή με την οποία εμφανίζει την ελάχιστη ανομοιότητα.

5.1.2 Πειραματικά Αποτελέσματα

Η τεχνική αυτή μειώνει σημαντικά τον αριθμό των εντοπισμένων σκηνών, οδηγώντας, έτσι, στη δημιουργία μεγαλύτερων θεματικών ενοτήτων, που προσεγγίζουν καλύτερα την έννοια της σκηνής. Παράλληλα, διατηρεί το υψηλό Recall και οδηγεί σε αύξηση του F_1, όπως φαίνεται στον Πίνακα 5.1.

	Precision(%)	**Recall(%)**	**F1(%)**
Before	9.2292	87.7456	16.4182
After	19.8475	84.3972	31.5091

Πίνακας 5.1: Σύγκριση Δεικτών Αξιολόγησης Πριν και Μετά την Απόρριψη των Σκηνών Μικρής Διάρκειας

Τα αποτελέσματα αυτής της μεθόδου είναι δυνατόν να χρησιμοποιηθούν ως η αρχική κατάτμηση για την αξιολόγηση των μεθόδων που ακολουθούν. Έτσι, στις επόμενες ενότητες θα παρουσιαστούν τα αποτελέσματα των διαφόρων μεθόδων τόσο ως προς την κατάτμηση που προκύπτει από τη μέθοδο γράφων όσο και ως προς την κατάτμηση μετά την απόρριψη των σύντομων σκηνών.

5.2 Ομαδοποίηση Λήψεων μεταξύ Διαδοχικών Σκηνών

5.2.1 Υλοποίηση

Η μέθοδος αυτή περιγράφεται στο [28] και βασίζεται σε μια εκ νέου ομαδοποίηση των λήψεων μεταξύ διαδοχικών σκηνών. Σε περίπτωση που μετά την ομαδοποίηση υπάρχει έστω και ένα ζεύγος λήψεων που ανήκουν στο ίδιο cluster αλλά όχι στην ίδια σκηνή, τότε οι δύο αυτές σκηνές πρέπει να ενοποιηθούν.

Για την ομαδοποίηση λαμβάνεται υπόψιν η οπτική ομοιότητα των λήψεων, αλλά και η χρονική τους εγγύτητα. Σε αντίθεση με τη μέθοδο που περιγράφηκε στην Ενότητα 4.1.2 το χρονικό κατώφλι προσαρμόζεται στα χαρακτηριστικά κάθε σκηνής, ενώ το οπτικό κατώφλι δ είναι προκαθορισμένο στην τιμή δ^*.

Ας θεωρηθεί ότι η πρώτη κατάτμηση σε σκηνές έχει χωρίσει την ταινία σε K θεματικές ενότητες (story units) $\{U_i\}_{i=1}^{K}$. Η διάρκεια κάθε θεματικής ενότητας συμβολίζεται με $\tau(U_i)$. Σε κάθε επανάληψη του αλγορίθμου η χρονική παράμετρος προσαρμόζεται ώστε να ισούται με το άθροισμα της διάρκειας των δύο θεματικών ενοτήτων που ελέγχονται. Στην περιγραφή που ακολουθεί κάθε θεματική ενότητα U_i αναφέρεται στην αρχική κατάτμηση, ενώ τα U'_m στην τελική κατάτμηση. Ο αλγόριθμος που χρησιμοποιείται είναι ο ακόλουθος:

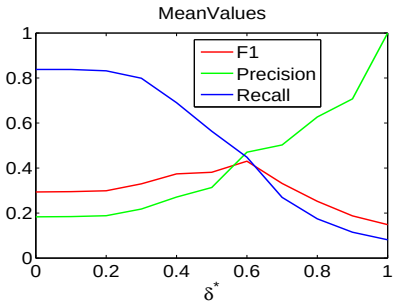
Βήμα 1 Αρχικοποίηση, $i \leftarrow 1$, $m \leftarrow 1$, $U'_1 \leftarrow U_1$.

Βήμα 2 **while** $i \leq K$ **do**
$i \leftarrow i + 1$
$T \leftarrow \tau(U_i) + \tau(U'_m)$
Ομαδοποίηση όλων των λήψεων $S_j \in U_i \cup U'_m$ με παραμέτρους (T, δ^*).
if $\exists$ cluster που να περιέχει λήψεις από το U_i και το U'_m **then**
$U'_m \leftarrow U_i \cup U'_m$
else
$m \leftarrow m + 1$
$U'_m \leftarrow U_i$
end if
end while

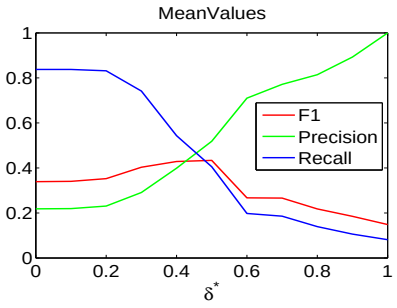
Βήμα 3 Επιστροφή της τελικής κατάτμησης $\{U'_1, U'_2, \ldots, U'_m\}$.

5.2.2 Πειραματικά Αποτελέσματα

Τα αποτελέσματα της μεθόδου φαίνονται στο Σχήμα 5.1. Παρατηρείται πως για οπτικό κατώφλι $\delta^* = 0.3$ λαμβάνονται τα καλύτερα αποτελέσματα, που είναι Recall $= 79.91\%$ και $F_1 = 33\%$. Επιπλέον, στο Σχήμα 5.2 απεικονίζονται τα αποτελέσματα εφαρμογής της μεθόδου στα αποτελέσματα της ενότητας 5.1.2. Σε αυτή την περίπτωση, τα καλύτερα αποτελέσματα λαμβάνονται για $\delta^* = 0.2$ και είναι Recall $= 83.15\%$ και $F_1 = 35.25\%$, αρκετά καλύτερα από αυτά που προέκυψαν βασιζόμενα στην αρχική κατάτμηση.



Σχήμα 5.1: Δείκτες Αξιολόγησης του Scene Refinement με τη μέθοδο του Shot Clustering μεταξύ διαδοχικών σκηνών ως προς την παράμετρο δ.



Σχήμα 5.2: Δείκτες Αξιολόγησης του Scene Refinement στο αποτέλεσμα την ενότητας 5.1.2 με τη μέθοδο του Shot Clustering μεταξύ διαδοχικών σκηνών ως προς την παράμετρο δ.

5.3 Κριτήριο Πληροφορίας του Bayes (Bayesian Information Criterion - BIC)

5.3.1 Θεωρητικό Υπόβαθρο

Το κριτήριο πληροφορίας του Bayes [21] είναι ένα κριτήριο επιλογής μοντέλου για την περιγραφή ενός δοσμένου συνόλου δεδομένων. Η επιλογή αυτή γίνεται ανάμεσα σε ένα πλήθος μοντέλων, τα οποία διαφέρουν ως προς τον αριθμό των παραμέτρων τους, μεγιστοποιώντας την πιθανοφάνειά τους. Όσο μεγαλύτερη είναι η πιθανοφάνεια ενός μοντέλου, τόσο καλύτερα περιγράφει τα δεδομένα. Επίσης, είναι προφανές ότι αυξάνοντας τις παραμέτρους του μοντέλου, θα αυξάνεται και η πιθανοφάνεια του. Αυτό οδηγεί σε πιο πολύπλοκα μοντέλα και προβλήματα υπεριεκπαίδευσης (overtraining). Για το λόγο αυτό στο κριτήριο εισάγεται και ένας παράγοντας κόστους που εξαρτάται από τον αριθμό των παραμέτρων του κάθε μοντέλου. Τελικά, επιλέγεται το μοντέλο για το οποίο το BIC λαμβάνει τη μέγιστη τιμή.

Πιο συγκεκριμένα, έστω $\mathcal{X} = \{x_i : i = 1, \ldots, N\}$ το σύνολο των δεδομένων που μοντελοποιείται και $\mathcal{M} = \{M_i : i = 1, \ldots, K\}$ τα υποψήφια μοντέλα. Για κάθε μοντέλο M, με d παραμέτρους, υπολογίζεται η πιθανοφάνεια $\mathcal{L}(\mathcal{X}, M)$ και το BIC ορίζεται ως εξής:

$$BIC(M) = \log(\mathcal{L}(\mathcal{X}, M)) - 0.5\lambda d \log(N) \tag{5.2}$$

όπου λ ένας συντελεστής βάρους που συνήθως επιλέγεται ίσος με τη μονάδα. Η διαδικασία ολοκληρώνεται επιλέγοντας το μοντέλο που μεγιστοποιεί το BIC.

Σε περίπτωση που πρέπει να διαπιστωθεί ποιο από δύο μοντέλα M_i και M_j περιγράφουν καλύτερα τα δεδομένα, τότε λαμβάνεται η διαφορά $\Delta BIC = BIC(M_i) - BIC(M_j)$ και αν είναι θετική τότε προτιμάται το M_i, ενώ σε αντίθετη περίπτωση το M_j.

Βασιζόμενοι σε αυτό, και ακολουθώντας τη μεθοδολογία που αναπτύσσεται στο [5], επιχειρούμε να εντοπίσουμε σε μια ακολουθία δεδομένων ένα σημείο αλλαγής, που σηματοδοτεί πως τα δεδομένα προέρχονται από διαφορετική κατανομή. Συγκεκριμένα, για μια ακολουθία δεδομένων $\mathcal{X} = \{x_i \in \mathbb{R}^k : i = 1, \ldots, N\}$, ελέγχονται δύο υποθέσεις: $H_0 : x_1 \ldots x_N \sim f(\theta)$ και $H_1 : x_1 \ldots x_i \sim f(\theta_1); x_{i+1} \ldots x_N \sim f(\theta_2)$. Ουσιαστικά, θεωρούμε ότι τα δεδομένα έχουν ληφθεί από μια στοχαστική διαδικασία, $x_i \sim f(\theta_i)$, στην οποία υπάρχει το πολύ ένα σημείο αλλαγής, το οποίο και πρέπει να εντοπιστεί. Για κάθε πιθανό σημείο αλλαγής i εξάγεται το

$\Delta BIC(i)$, ως εξής:

$$\begin{aligned}\Delta BIC(i) = BIC(M_1) - BIC(M_0) = \log \mathcal{L}(\mathcal{X}, \theta_1) + \log \mathcal{L}(\mathcal{X}, \theta_2) - \\ 0.5\lambda(2d)\log(N) - (\log \mathcal{L}(\mathcal{X}, \theta) - 0.5\lambda d \log(N)) \\ = \log \mathcal{L}(\mathcal{X}, \theta_1) + \log \mathcal{L}(\mathcal{X}, \theta_2) - \log \mathcal{L}(\mathcal{X}, \theta) - 0.5\lambda d \log(N)\end{aligned} \tag{5.3}$$

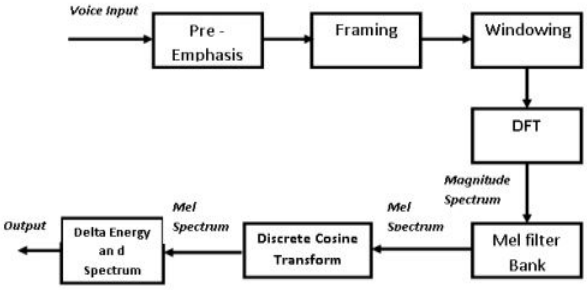
Όταν η εξίσωση (5.3) λαμβάνει θετικές τιμές τότε προτιμάται το μοντέλο με τις δύο κατανομές. Ως σημείο αλλαγής λαμβάνεται το $\hat{t} = \text{argmax}\, \Delta BIC(i)$.

5.3.2 Προσαρμογή του BIC στην Κατάτμηση σε Σκηνές

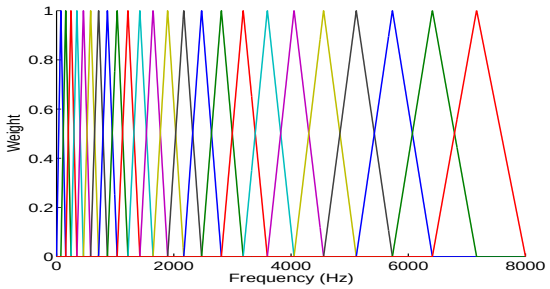
Το BIC έχει χρησιμοποιηθεί αρκετά για την κατάτμηση ενός ηχητικού αποσπάσματος ([5], [32]), ενώ στο [10] εισάγεται η χρήση του για κατάτμηση των τηλεοπτικών εκπομπών σε προγράμματα. Οι μέθοδοι αυτές εξάγουν κάποια χαρακτηριστικά είτε από το οπτικό είτε από το ακουστικό κανάλι και πάνω σε αυτά εφαρμόζουν το BIC. Ως ακουστικά χαρακτηριστικά χρησιμοποιούνται συνήθως οι συντελεστές MFCC (Mel Frequency Cepstral Coefficients), ενώ ως οπτικά, τα ιστογράμματα χρώματος.

Για τη βελτίωση της κατάτμησης σε σκηνές το BIC μπορεί να χρησιμοποιηθεί με τον ακόλουθο τρόπο πάνω σε ένα σύνολο χαρακτηριστικών (οπτικών/ακουστικών) λίγο πριν και λίγο από κάθε αλλαγή σκηνής. Ελέγχονται δύο διαφορετικά μοντέλα M_1 και M_0. Το M_1 χρησιμοποιεί δύο ξεχωριστές κατανομές για να περιγράψει τα δεδομένα (μια κατανομή για τα δεδομένα πριν την αλλαγή σκηνής και μια για μετά). Το M_0 περιγράφει με μια κοινή κατανομή όλα τα δεδομένα. Υπολογίζουμε το BIC κάθε μοντέλου και εξάγουμε τη διαφορά $\Delta BIC = BIC(M_1) - BIC(M_0)$. Αν $\Delta BIC > 0$ τότε τα δεδομένα περιγράφονται καλύτερα με δύο κατανομές και η αλλαγή σκηνής διατηρείται. Σε αντίθετη περίπτωση, τα δεδομένα μπορούν να περιγραφούν με μια κατανομή, άρα η αλλαγή σκηνής απορρίπτεται.

Αυτή η διαδικασία περιλαμβάνει αρκετές παραμέτρους που πρέπει να ληφθούν υπόψιν. Τα χαρακτηριστικά που θα εξαχθούν από κάθε κανάλι, το είδος των κατανομών που θα χρησιμοποιηθούν, η παράμετρος βάρους λ στην εξίσωση (5.2) είναι όλα στοιχεία που πρέπει να αποφασιστούν και να ορισθούν κατάλληλα. Για το λόγο αυτό στις επόμενες ενότητες θα παρουσιαστούν οι διάφοροι τρόποι υλοποίησης που δοκιμάστηκαν, καθώς και τα αποτελέσματά τους.



Σχήμα 5.3: Βασικά Βήματα για τον Υπολογισμό των MFCC, από το [13].



Σχήμα 5.4: Τριγωνική Συστοιχία 24 φίλτρων σε κλίμακα Mel.

5.4 Κριτήριο Πληροφορίας του Bayes (BIC) σε Ακουστικά Χαρακτηριστικά

5.4.1 Συντελεστές MFCC

Οι Mel Frequency Cepstral Coefficients (MFCC) είναι ένα σύνολο συντελεστών cepstrum που εξάγονται μετά από ανάλυση του ακουστικού σήματος με μια ειδικά σχεδιασμένη συστοιχία φίλτρων (filterbank) σε κλίμακα Mel, η οποία έχει προκύψει από ψυχοακουστικές μελέτες. Η εξαγωγή τους βασίζεται στον τρόπο με τον οποίο ο άνθρωπος αντιλαμβάνεται τα ακουστικά σήματα. Ένα block διάγραμμα που δείχνει τον τρόπο υπολογισμού των συντελεστών αυτών φαίνεται στο Σχήμα 5.3, ενώ η συστοιχία φίλτρων σε κλίμακα Mel απεικονίζεται στο Σχήμα 5.4.

5.4.2 Υλοποίηση

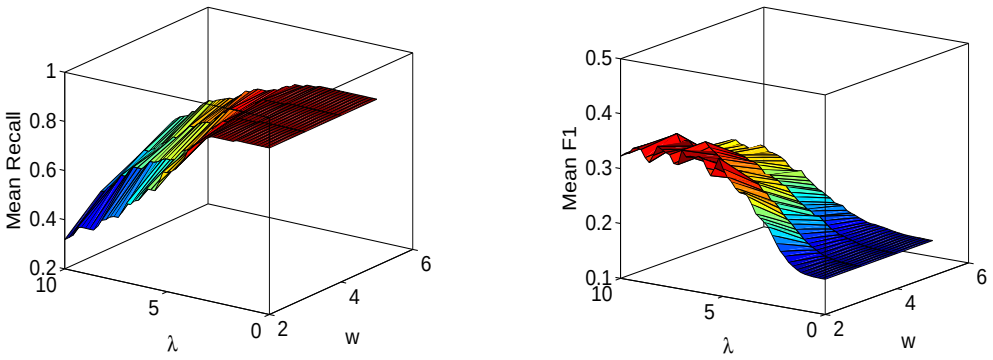
Το σύνολο των ακουστικών χαρακτηριστικών πάνω στο οποίο εφαρμόζεται το BIC είναι οι συντελεστές MFCC του ακουστικού καναλιού και τα μοντέλα που συγκρίνονται είναι είτε απλές γκαουσιανές είτε μίγματα

γκαουσιανών.

5.4.3 Πειραματικά αποτελέσματα

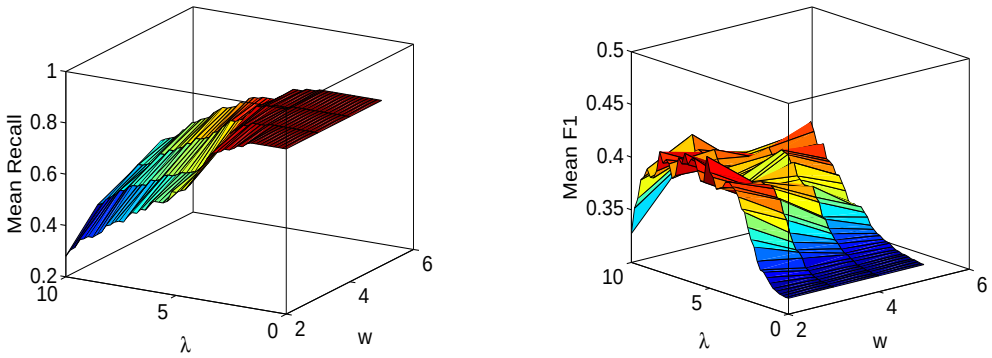
Οι παράμετροι που πρέπει να ελεγχθούν είναι: ο αριθμός των γκαουσιανών για τη μοντελοποίηση, το μέγεθος w του παραθύρου και ο συντελεστής λ στην εξίσωση (5.2). Στο Σχήμα 5.5 φαίνονται τα αποτελέσματα της μεθόδου ως προς το λ και το w, που καθορίζει το μέγεθος του παραθύρου ελέγχου, δηλαδή πόσες λήψεις πριν και μετά την εντοπισμένη αλλαγή θα χρησιμοποιηθούν για την εξαγωγή του BIC. Για τη μοντελοποίηση των συντελεστών χρησιμοποιήθηκαν απλές γκαουσιανές.

Παρατηρείται πως οι παράμετροι που δίνουν υψηλές τιμές Recall (της τάξης του 87%) έχουν παράλληλα σχετικά χαμηλές τιμές F_1 (της τάξης του 16%) και δεν επιφέρουν κάποια ουσιαστική αλλαγή στην κατάτμηση. Εφαρμόζοντας τη μέθοδο BIC Audio στα αποτελέσματα που έχουν ληφθεί από την ενότητα 5.1.2 λαμβάνουμε τα αποτελέσματα που φαίνονται στο Σχήμα 5.6. Τα αποτελέσματα αυτά είναι αρκετά βελτιωμένα σε σχέση με τα προηγούμενα (ενδεικτικά, βέλτιστο Recall $= 82.23\%$ και $F_1 = 37.92\%$). Αξίζει να σημειωθεί, ωστόσο, πως για καθεμία ταινία ξεχωριστά υπάρχει ένας συνδυασμός παραμέτρων (λ^*, w^*) που δίνει βέλτιστα αποτελέσματα. Οι τιμές αυτές των παραμέτρων φαίνονται στον Πίνακα 5.2. Όπως φαίνεται, για τις περισσότερες ταινίες το βέλτιστο παράθυρο w^* έχει μήκος δύο (δηλαδή δύο λήψεις πριν την αλλαγή και δύο λήψεις μετά την αλλαγή), ενώ το λ^* είναι διαφορετικό για κάθε ταινία.



(i) Recall (ii) F1

Σχήμα 5.5: Δείκτες αξιολόγησης του Refinement με χρήση του BIC Audio ως προς το λ και το w.



(i) Recall (ii) F1

Σχήμα 5.6: Δείκτες αξιολόγησης του Refinement με χρήση του BIC Audio ως προς το λ και το w στα αποτελέσματα της ενότητας 5.1.2.

	Recall(%)	**F1(%)**	λ^*	w^*
BMI	91.67	53.66	6.2	2
CHI	100	44.44	3.4	2
CRA	90.91	50	5.2	2
DEP	77.27	57.63	3.4	3
FNE	85.71	32.43	4	2
GLA	80	34.04	4.4	4
LOR	65.22	36.59	2.8	5

Πίνακας 5.2: Βέλτιστες τιμές των παραμέτρων για κάθε ταινία της βάσης, για τη μέθοδο BIC Audio.

Σε αυτό το σημείο αξίζει να σημειωθεί πώς αύξηση του αριθμού των γκαουσιανών χειροτερεύει τα αποτελέσματα. Μια πιθανή αιτιολογία είναι πως δύο ή περισσότερες γκαουσιανές προσαρμόζονται στα δεδομένα των διαφορετικών σκηνών και καταφέρνουν να τα περιγράψουν επαρκώς, χωρίς να απαιτείται το πιο πολύπλοκο μοντέλο, με τα δύο μείγματα γκαουσιανών.

5.5 Κριτήριο Πληροφορίας του Bayes (BIC) σε Οπτικά Χαρακτηριστικά

5.5.1 Υλοποίηση

Τα χαρακτηριστικά που χρησιμοποιούνται για την εφαρμογή του BIC σε δεδομένα από την εικόνα είναι τα ιστογράμματα χρώματος. Με βάση το [10] εξάγεται για κάθε καρέ το ιστόγραμμα χρώματος (256 bins) κάθε καναλιού, τα οποία στη συνέχεια ενώνονται σε ένα κοινό διάνυσμα χαρακτηριστικών (μεγέθους $256 \times 3 = 768$).

Αρχικά, στο διάνυσμα χαρακτηριστικών εφαρμόζεται ανάλυση σε πρωτεύουσες συνιστώσες (Principal Components Analysis - PCA) για να μειωθεί η διάστασή του. Ο αριθμός των συνιστωσών που διατηρούνται αποτελεί μια παράμετρο της μεθόδου.

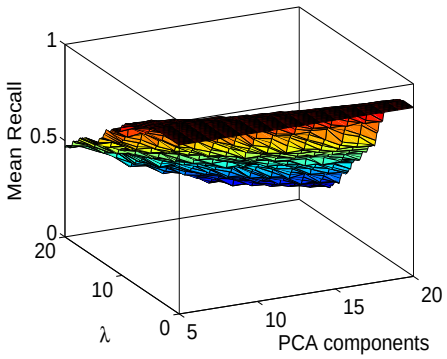
Ο αλγόριθμος PCA [9] εφαρμόζει έναν ορθογώνιο μετασχηματισμό στα δεδομένα-παρατηρήσεις, ώστε να εντοπίσει πιθανές συσχετίσεις μεταξύ των διάφορων μεταβλητών και να εξάγει ένα σύνολο ασυσχέτιστων μεταβλητών, που ονομάζονται πρωτεύουσες συνιστώσες (principal components). Ο αριθμός των πρωτευουσών συνιστωσών είναι το πολύ ίσος με τον αρχικό αριθμό των μεταβλητών, ενώ η μέθοδος ορίζεται με τέτοιο τρόπο, ώστε οι πρώτες συνιστώσες να περικλείουν τη μεγαλύτερη διακύμανση των δεδομένων. Έτσι, διατηρώντας ένα (μικρότερο) αριθμό πρωτευουσών συνιστωσών, που οδηγεί σε μικρότερη διάσταση των δεδομένων, τα δεδομένα μπορούν να περιγραφούν με μεγάλη ακρίβεια.

Στη συνέχεια, εφαρμόζεται το BIC, όπως και στην περίπτωση των ακουστικών χαρακτηριστικών, λαμβάνοντας ένα παράθυρο δύο λήψεων πριν και δύο λήψεων μετά την εντοπισμένη αλλαγή.

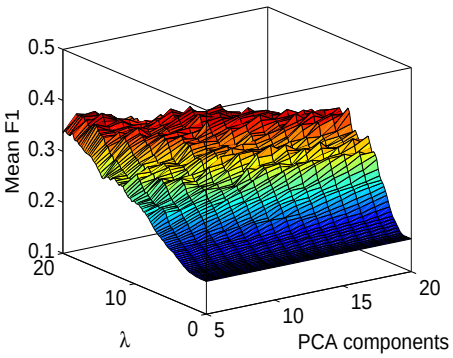
5.5.2 Πειραματικά αποτελέσματα

Τα αποτελέσματα αυτής της μεθόδου φαίνονται στο Σχήμα 5.7 και στο Σχήμα 5.8 ως προς το συντελεστή λ και τον αριθμό των συνιστωσών του PCA που διατηρούνται. Οι υψηλές τιμές του Recall και οι αντίστοιχες τιμές του F_1 είναι ίσες με αυτές της αρχικής κατάτμησης, και απ' ο,τι φαίνεται η εφαρμογή του BIC σε οπτικά χαρακτηριστικά δεν επιφέρει κάποια ουσιαστική βελτίωση.

Εν μέρει, αυτό μπορεί να αιτιολογείται από το γεγονός οτι όλη η προηγούμενη διαδικασία της κατάτμησης έχει βασιστεί σε οπτικά δεδομένα, επομένως, μια επιπλέον επεξεργασία των δεδομένων της εικόνας δεν μπορεί να προσφέρει ουσιαστική βελτίωση.



(i) Recall



(ii) F1

Σχήμα 5.7: Δείκτες αξιολόγησης του Refinement με χρήση του BIC Visual ως προς το λ και τον αριθμό των πρωτευουσών συνιστωσών.

5.6 Σύμμειξη Ροών Πληροφορίας

5.6.1 Υλοποίηση

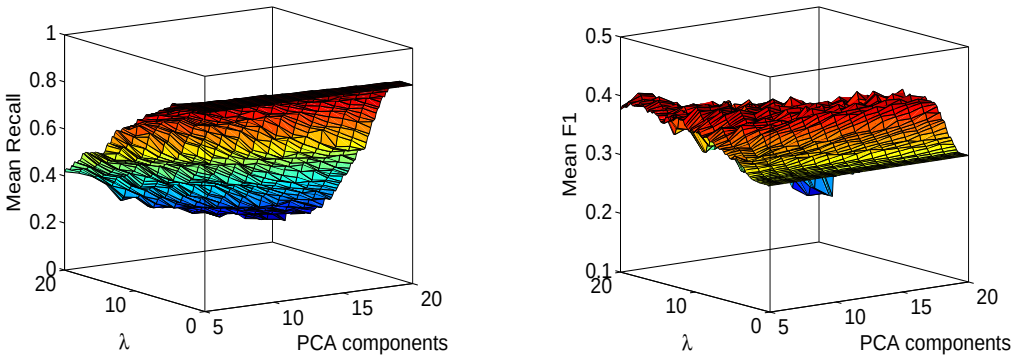
Σε αυτό το σημείο θα εφαρμοστεί το BIC σε δεδομένα που προέρχονται τόσο από το ακουστικό κανάλι όσο και από το οπτικό. Υπάρχουν δύο τρόποι να συνδυαστούν αυτά τα δεδομένα, όπως αναλύεται στο [10]:

Early Fusion: Πρόκειται για μια σύμμειξη στο επίπεδο των χαρακτηριστικών. Πιο αναλυτικά, τα ακουστικά χαρακτηριστικά (MFCC) και τα οπτικά (ιστογράμματα χρώματος) ενώνονται σε ένα κοινό διάνυσμα χαρακτηριστικών και το BIC εφαρμόζεται σε αυτά τα δεδομένα.

Ένα πρόβλημα που υπάρχει σε αυτή την υλοποίηση είναι η διαφορετική συχνότητα που έχουν μεταξύ τους το οπτικό και το ακουστικό κανάλι. Για το λόγο αυτό οι συντελεστές MFCC κανονικοποιούνται, κβαντίζονται και εξάγονται τα ιστογράμματά τους με συχνότητα ίση με τη συχνότητα του βίντεο.

Τα ιστογράμματα των συντελεστών MFCC και χρώματος συνδυάζονται σε ένα κοινό διάνυσμα χαρακτηριστικών, πάνω στο οποίο εφαρμόζεται αρχικά μείωση της διάστασης μέσω του PCA και, στη συνέχεια, το BIC για να απορριφθούν κάποιες από τις εντοπισμένες σκηνές.

Late Fusion: Σε αυτό το σημείο συνδυάζεται το BIC που έχει εξαχθεί μεμονωμένα από τα οπτικά (ΔBIC_V) και ακουστικά (ΔBIC_A)



(i) Recall (ii) F1

Σχήμα 5.8: Δείκτες αξιολόγησης του Refinement στα αποτελέσματα της ενότητας 5.1.2 με χρήση του BIC Visual ως προς το λ και τον αριθμό των πρωτευουσών συνιστωσών.

χαρακτηριστικά, με τον εξής τρόπο:

$$\Delta BIC_{AV} = \Delta BIC_A + \Delta BIC_V \tag{5.4}$$

Εξετάζοντας το πρόσημο του ΔBIC_{AV} η αλλαγή σκηνής διατηρείται ή απορρίπτεται.

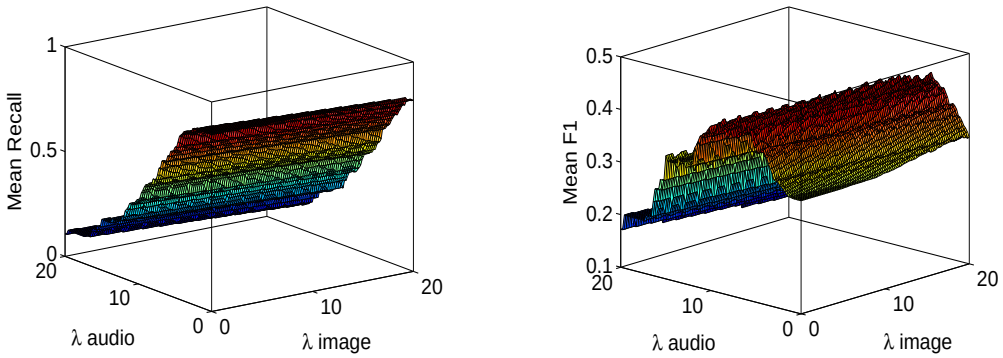
5.6.2 Πειραματικά αποτελέσματα

Η σύμμειξη των ροών πληροφορίας στο πρώτο στάδιο (early fusion) δεν δίνει κάποια αξιόλογα αποτελέσματα. Για το λόγο αυτό στο σημείο αυτό θα παρουσιαστούν μόνο τα αποτελέσματα του late fusion, που απεικονίζονται στο Σχήμα 5.9. Για την εξαγωγή αυτών των αποτελεσμάτων χρησιμοποιήθηκε παράθυρο $w = 2$ στον υπολογισμό του BIC Audio και 10 πρωτεύουσες συνιστώσες κατά την εφαρμογή του PCA για τον υπολογισμό του BIC Image. Τα βέλτιστα αποτελέσματα κυμαίνονται στις ίδιες τιμές με αυτά του BIC Audio, γεγονός που επιβεβαιώνει την ένδειξη οτι τα ιστογράμματα χρώματος δεν μπορούν να επιφέρουν περαιτέρω βελτίωση στην κατάτμηση.

5.7 GIST Χαρακτηριστικά

5.7.1 Θεωρητικό Υπόβαθρο

Ο περιγραφητής GIST, που προτείνεται στο [17], έχει ως βασική ιδέα την κατάτμηση της εικόνας σε υπο-εικόνες και την εξαγωγή πληροφοριών (από



(i) Recall (ii) F1

Σχήμα 5.9: Δείκτες αξιολόγησης του Refinement στα αποτελέσματα της ενότητας 5.1.2 με χρήση του Late Fusion ως προς το λ .

κάθε υπό-εικόνα) σχετικών με την κλίση (gradient) σε διάφορες κλίμακες και προσανατολισμούς. Ο περιγραφητής αυτός δίνει μια γενική περιγραφή της εικόνας και χρησιμοποιείται για την αναγνώριση παρόμοιων εικόνων, όπως βουνά, ψηλά κτήρια, δρόμοι κ.α. Ένα παράδειγμα του περιγραφητή φαίνεται στο Σχήμα 5.10.

Για την εξαγωγή του περιγραφητή GIST ακολουθούνται τα παρακάτω βήματα:

Βήμα 1 Συνέλιξη της εικόνας με 32 Gabor φίλτρα σε 4 κλίμακες και 8 προσανατολισμούς. Παράγονται 32 χάρτες χαρακτηριστικών (feature maps).

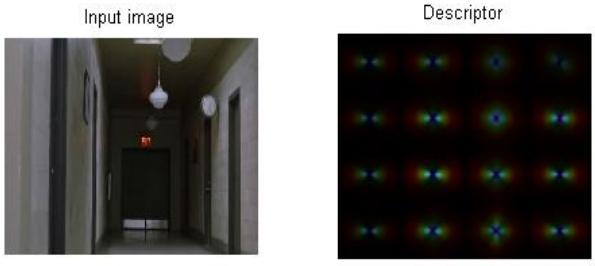
Βήμα 2 Κατάτμηση κάθε χάρτη σε 16 περιοχές (4×4 grid). Υπολογισμός της μέσης τιμής από κάθε περιοχή.

Βήμα 3 Ένωση όλων των τιμών σε ένα διάνυσμα χαρακτηριστικών μεγέθους $32 \times 16 = 512$, το οποίο αποτελεί τον περιγραφητή GIST.

5.7.2 Υλοποίηση

Η ιδέα της βελτίωσης της κατάτμησης σε σκηνές με χρήση του περιγραφητή GIST βασίζεται στη μεθοδολογία που αναπτύχθηκε στην Ενότητα 4.1.5, ως προς την εξαγωγή οπτικών λέξεων και οπτικών ιστογραμμάτων.

Για την υλοποίηση αυτή από κάθε καρέ εξάγεται ο περιγραφητής GIST, όπως αναλύθηκε παραπάνω, και γίνεται μια ομαδοποίησή τους σε k ομάδες



Σχήμα 5.10: Περιγραφητής GIST.

- οπτικές λέξεις. Στη συνέχεια, εξάγεται το οπτικό ιστόγραμμα λέξεων από κάθε υπολογισμένη σκηνή και γίνεται σύγκριση των ιστογραμμάτων διαδοχικών σκηνών. Και σε αυτή την υλοποίηση τα οπτικά ιστογράμματα ομαλοποιούνται φιλτράροντάς τα με μια γκαουσιανή μηδενικής μέσης τιμής και τυπικής απόκλισης σ.

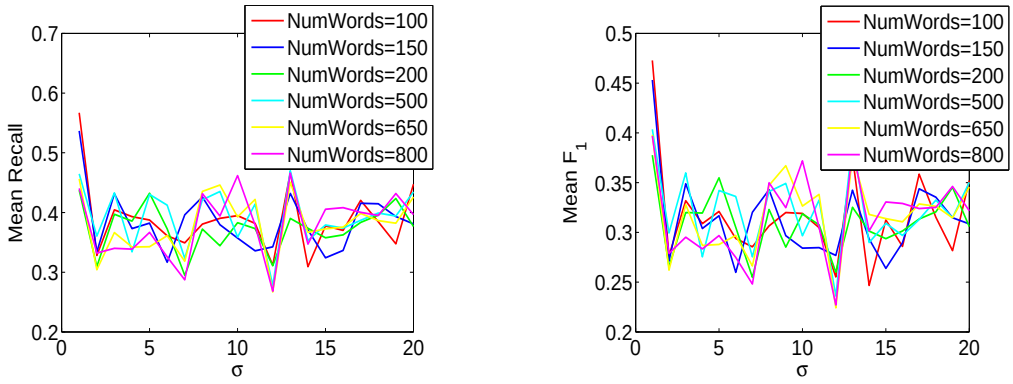
5.7.3 Πειραματικά Αποτελέσματα

Στο Σχήμα 5.11 φαίνονται τα αποτελέσματα της μεθόδου με χρήση της chi-squared απόστασης μεταξύ διαδοχικών ιστογραμμάτων. Από τον πειραματισμό προέκυψε πως η τομή ιστογραμμάτων, εξίσωση (4.23), και η chi-squared απόσταση, εξίσωση (4.24), δίνουν παραπλήσια αποτελέσματα. Αντιθέτως, η ευκλείδεια απόσταση, εξίσωση (4.22), δίνει ελαφρώς χειρότερα αποτελέσματα για όλες τις ταινίες εκτός από μια (LOR).

Στον Πίνακα 5.3 φαίνονται τα καλύτερα αποτελέσματα που λαμβάνονται επιμέρους για κάθε ταινία ξεχωριστά. Το Recall της μεθόδου είναι εμφανώς μειωμένο για τις περισσότερες ταινίες. Ωστόσο, για την ταινία των κινουμένων σχεδίων (FNE) το Recall διατηρεί υψηλή τιμή, έχοντας ταυτόχρονα και υψηλό F_1. Η παρατήρηση αυτή είναι ιδιαίτερα σημαντική, καθώς δείχνει πως ο περιγραφητής GIST ίσως είναι καταλληλότερος να χρησιμοποιείται σε τέτοιου είδους ταινίες.

Τέλος, στον Πίνακα 5.4 αναγράφονται οι μέσες τιμές των δεικτών

αξιολόγησης κάνοντας χρήση των τριών διαφορετικών μετρικών για τον υπολογισμό της απόστασης, για μέγεθος οπτικού λεξιλογίου $k = 100$ και τυπική απόκλιση του γκαουσιανού φίλτρου $\sigma = 1$.



(i) Recall (ii) F1

Σχήμα 5.11: Δείκτες αξιολόγησης του Refinement στα αποτελέσματα της ενότητας 5.1.2 με χρήση του Bag of Visual Words του περιγραφητή GIST ως προς το μέγεθος του λεξιλογίου και την τυπική απόκλιση σ.

	Recall(%)	**F1(%)**	σ^*	k^*
BMI	66	55.17	3	100
CHI	68.75	56.41	1	150
CRA	54.56	46.15	1	150
DEP	63.64	65.12	1	150
FNE	85.71	57.14	1	100
GLA	50	40	1	100
LOR	39.13	46.15	4	500

Πίνακας 5.3: Βέλτιστες τιμές των παραμέτρων για κάθε ταινία της βάσης, για τη μέθοδο Bag of Visual Words του περιγραφητή GIST.

5.8 Περιγραφητής SIFT

Ο περιγραφητής SIFT έχει ήδη δοκιμαστεί για την κατάτμηση της ταινίας σε σκηνές (ενότητα 4.1.5), χωρίς να έχει δώσει ικανοποιητικά αποτελέσματα. Στο σημείο αυτό, επιχειρείται να χρησιμοποιηθεί για τη βελτίωση της

	Recall(%)	**Precision(%)**	**F1(%)**
Histogram Intersection	55.77	40.22	45.48
Chi-Squared Distance	56.71	42.49	47.28
Euklidean Distance	47.11	41.53	42.72

Πίνακας 5.4: Μέσες τιμές των δεικτών αξιολόγησης για τη μέθοδο Bag of Visual Words του περιγραφητή GIST, με χρήση των τριών διαφορετικών μετρικών υπολογισμού της απόστασης.

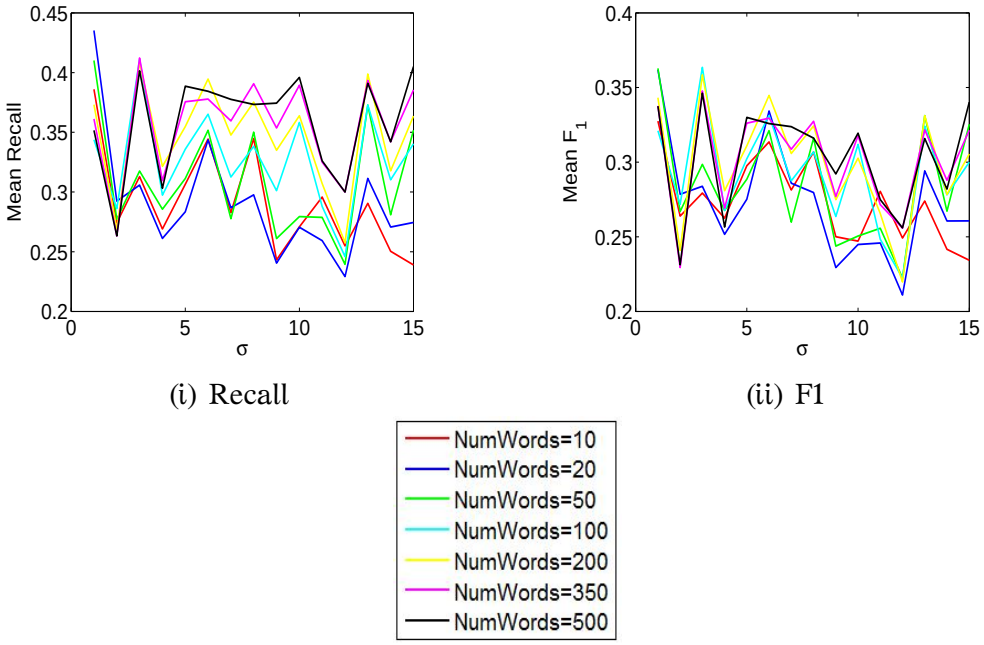
ήδη υπάρχουσας κατάτμησης, με τρόπο αντίστοιχο αυτού με τον οποίο χρησιμοποιήθηκε και ο περιγραφητής GIST.

Εξάγεται ο περιγραφητής SIFT από κάθε αντιπροσωπευτικό καρέ όλων των λήψεων, και, έτσι, δημιουργείται ο περιγραφητής D_S για ολόκληρη την ταινία. Μετά την εφαρμογή του k-means [1] εξάγεται το οπτικό ιστόγραμμα όχι για κάθε λήψη, αλλά για κάθε σκηνή, και αυτά τα ιστογράμματα συγκρίνονται μεταξύ τους, χρησιμοποιώντας ένα από τα τρία διαθέσιμα μετρικά απόστασης.

Στο Σχήμα 5.12 φαίνονται οι μέσες τιμές των δεικτών αξιολόγησης της μεθόδου για τη βάση των 7 ταινιών. Τα αποτελέσματα, κατά μέσο όρο, είναι λίγο χαμηλότερα από αυτά της προηγούμενης ενότητας. Ωστόσο, όπως φαίνεται και στον Πίνακα 5.5, για κάποιες από τις ταινίες (CRA, GLA, LOR) αυτή η μέθοδος οδηγεί σε καλύτερα αποτελέσματα (με το βέλτιστο συνδυασμό παραμέτρων).

	Recall(%)	**F1(%)**	σ^*	k^*
BMI	58.33	48.28	1	100
CHI	62.5	47.62	3	500
CRA	63.64	46.67	1	20
DEP	40.91	50	7	500
FNE	71.43	55.56	1	500
GLA	50	41.67	10	500
LOR	47.62	43.48	8	500

Πίνακας 5.5: Βέλτιστες τιμές των παραμέτρων για κάθε ταινία της βάσης, για τη μέθοδο Bag of Visual Words του περιγραφητή SIFT.



(i) Recall (ii) F1

Σχήμα 5.12: Δείκτες αξιολόγησης του Refinement στα αποτελέσματα της ενότητας 5.1.2 με χρήση του Bag of Visual Words του περιγραφητή SIFT ως προς το μέγεθος του λεξιλογίου και την τυπική απόκλιση σ.

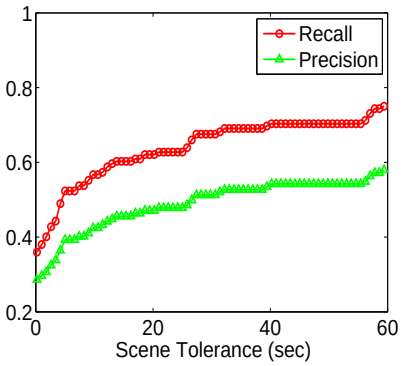
5.9 Επιτρεπτή απόσταση

Στο κεφάλαιο 4 τα αποτελέσματα που παρουσιάζονται έχουν προκύψει λαμβάνοντας ως σωστά εντοπισμένες αλλαγές σκηνής αυτές που απέχουν έως 50 καρέ από αυτές που έχουν επισημειωθεί, ενώ στον παρόν κεφάλαιο η αντίστοιχη ανοχή είναι 250 καρέ. Στο Σχήμα 5.13 φαίνεται η επίδραση της αύξησης αυτού του ορίου στους δείκτες αξιολόγησης. Τα αποτελέσματα αυτά είναι ενδεικτικά και έχουν προέλθει από τη μεθοδολογία της ενότητας 5.7 για $\sigma = 1$ και μέγεθος λεξιλογίου 100 οπτικών λέξεων.

Παρατηρείται μια δραματική αύξηση των δεικτών αξιολόγησης μέσα στο διάστημα $0 - 15sec$, που αντιστοιχούν σε $0 - 375$ καρέ, και μια πιο ήπια αύξηση μέχρι τα $60sec$ (δηλαδή $375 - 1500$ καρέ).

5.10 Συμπεράσματα

Από τις μεθόδους που παρουσιάστηκαν σε αυτό το κεφάλαιο διαπιστώνουμε πως η εισαγωγή νέων χαρακτηριστικών, και κυρίως η χρήση της ακουστικής πληροφορίας, είναι ιδιαίτερα χρήσιμη για τη βελτίωση του



Σχήμα 5.13: Μεταβολή των δεικτών αξιολόγησης σε σχέση με το επιτρεπτό όριο απόστασης εντοπισμένων και επισημειωμένων σκηνών.

αποτελέσματος της κατάτμησης, όπως φαίνεται στον Πίνακα 5.2. Αντιθέτως, μέθοδοι που χρησιμοποιούν τα ιστογράμματα χρώματος, με διαφορετική προσέγγιση από αυτή που χρησιμοποιήθηκε στην αρχική κατάτμηση, δεν είναι ικανές να βελτιώσουν το αποτέλεσμα.

Ο πειραματισμός με αυτά τα χαρακτηριστικά και οι διαφορετικοί τρόποι αντιμετώπισης που παρουσιάστηκαν οδηγούν σε μια σημαντική βελτίωση του αποτελέσματος. Για κάθε ταινία υπάρχει ένα σύνολο παραμέτρων που δίνει πολύ ικανοποιητικά αποτελέσματα, χωρίς να έχει βρεθεί ένα σύνολο παραμέτρων για όλες τις ταινίες της βάσης.

Κεφάλαιο 6

Σύνοψη

Το κεφάλαιο αυτό αποτελεί την κατακλείδα της παρούσας διπλωματικής και αξιοποιείται ώστε αφενός να παρουσιαστεί μια ανακεφαλαίωση των βασικών σημείων που αναπτύχθηκαν και αφετέρου να προτείνει κάποιες μελλοντικές κατευθύνσεις για σχετική έρευνα, πέρα από τα όρια αυτής της εργασίας.

6.1 Ανακεφαλαίωση-Συνεισφορά

Στην παρούσα διπλωματική εργασία επικεντρωθήκαμε στο πρόβλημα της πολυτροπικής κατάτμησης μιας ταινίας σε σκηνές.

- Αρχικά, μελετήσαμε υπάρχουσες μεθόδους της βιβλιογραφίας που βασίζονται σε χαμηλού επιπέδου χαρακτηριστικά (ιστογράμματα χρώματος και ακμές) για την κατάτμηση μιας ταινίας σε λήψεις. Τα αποτελέσματα αυτής της κατάτμησης ήταν άκρως ικανοποιητικά, αλλά δοκιμάσαμε και ένα συνδυασμό των μεθόδων της βιβλιογραφίας και καταφέραμε να λάβουμε βελτιωμένα αποτελέσματα.

- Στη συνέχεια, εξετάσαμε μεθόδους ομαδοποίησης των λήψεων, ώστε να συγκροτούν μια σκηνή. Η ομαδοποίηση των λήψεων γίνεται κατά κύριο λόγο με βάση την οπτική ομοιότητα τους. Εξετάστηκαν μέθοδοι που βασίζονται στη δημιουργία ενός γράφου και την κατάλληλη κατάτμησή του με Normalized Cuts ή Cut Edges, ώστε να προκύψουν τα όρια των σκηνών. Δοκιμάστηκε, επίσης, η ομαδοποίηση των λήψεων με βάση τη φασματική τους ομοιότητα, μέσω του Spectral Clustering αλγορίθμου και ο εντοπισμός επαναλαμβανόμενων μοτίβων λήψεων κατά τη διάρκεια της ταινίας. Τέλος, εξετάστηκε μια μέθοδος Bag of Visual Words, όπου κάθε λήψη αναπαραστάθηκε με ένα ιστόγραμμα

οπτικών λέξεων. Όλες οι παραπάνω μέθοδοι, που περιγράφονται στη βιβλιογραφία, αξιολογήθηκαν πάνω σε μια διαθέσιμη βάση 7 ταινιών και εξήχθησαν τα μέτρα αξιολόγησης για κάθε μια μέθοδο, αλλά και ένα σύνολο παραμέτρων που δίνει ικανοποιητικά αποτελέσματα για το σύνολο των ταινιών της βάσης.

- Το καλύτερο αποτέλεσμα της κατάτμησης, όπως αυτό προέκυψε από τις μεθόδους της βιβλιογραφίας, περιείχε μεγάλο αριθμό εντοπισμένων σκηνών, ενώ εντόπιζε σωστά και με μεγάλη ακρίβεια τις περισσότερες επισημειωμένες σκηνές. Για το λόγο αυτό επιχειρήσαμε να κάνουμε μια περαιτέρω βελτίωση του αποτελέσματος, πειραματιζόμενοι με διάφορες μεθόδους. Στο σημείο αυτό χρησιμοποιήθηκαν οι περιγραφητές SIFT και GIST της εικόνας, για την εξαγωγή ενός οπτικού ιστογράμματος για κάθε εντοπισμένη σκηνή. Διαδοχικά ιστογράμματα σκηνών συγκρίθηκαν μεταξύ τους για να διαπιστωθεί εάν παρουσιάζουν μικρή ή μεγάλη ομοιότητα, και, αντίστοιχα, η εντοπισμένη σκηνή να διατηρηθεί ή να απορριφθεί. Το καλύτερο αποτέλεσμα, ωστόσο, προέκυψε με χρήση του κριτηρίου πληροφορίας του Bayes πάνω στα ακουστικά χαρακτηριστικά (MFCC). Το BIC προσαρμόστηκε ώστε να ελέγχει ένα παράθυρο γύρω από κάθε εντοπισμένη σκηνή και αναλόγως να διατηρεί ή να απορρίπτει τη σκηνή αυτή. Επιμέρους, για κάθε ταινία της βάσης η προσέγγιση αυτή έδωσε ικανοποιητικά αποτελέσματα, χωρίς να βρεθεί ένα σύνολο παραμέτρων που να δίνει καλά αποτελέσματα συγχρόνως για όλες τις ταινίες.

- Τέλος, δημιουργήθηκε ένα εργαλείο για την εισαγωγή χρονικής πληροφορίας στο σενάριο μιας ταινίας, μέσω της ευθυγράμμισης των διαλόγων του σεναρίου και των υποτίτλων. Η διαδικασία αυτή θα είναι ιδιαίτερα χρήσιμη κατά την ανάθεση περιγραφής σε κάθε σκηνή για εφαρμογές ταξινόμησης και κατηγοριοποίησης βίντεο.

6.2 Μελλοντικές Κατευθύνσεις

Από τις ιδέες που αναπτύχθηκαν και από τα ερευνητικά αποτελέσματα που προέκυψαν στο πλαίσιο της παρούσας διπλωματικής εργασίας αναδύονται πολλαπλές πιθανές προεκτάσεις για μελλοντική έρευνα. Οι ιδέες αυτές συνοψίζονται σε τρεις μεγάλες κατηγορίες:

Πειραματισμός με υψηλού επιπέδου χαρακτηριστικά. Η εισαγωγή πληροφορίας όπως η αναγνώριση προσώπων και αντικειμένων μπορεί

να συνεισφέρει θετικά στη διαδικασία της κατάτμησης. Όπως είδαμε τα χαμηλού επιπέδου χαρακτηριστικά είναι ικανά να εντοπίσουν τα όρια των λήψεων, αλλά δεν αποδίδουν εξίσου καλά στον εντοπισμό των ορίων των σκηνών.

Πειραματισμός σε μεγαλύτερη βάση ταινιών. Οι 7 ταινίες της διαθέσιμης βάσης ποικίλουν ως προς το είδος. Θεωρούμε πως πειραματισμός σε ταινίες ενός είδους θα μπορέσει να δώσει καλύτερα αποτελέσματα και παραμέτρους κατάλληλες για κάθε είδος ταινίας ξεχωριστά. Για την επίτευξη του σκοπού αυτού θα μπορούσαν να χρησιμοποιηθούν και machine learning τεχνικές, εφόσον το μέγεθος της βάσης είναι αρκετά μεγάλο.

Αξιοποίηση του Σεναρίου. Η επεξεργασία της γραπτής πληροφορίας του σεναρίου μπορεί να φανεί χρήσιμη στο πρόβλημα της κατάτμησης. Αξιοποιώντας μεθόδους επεξεργασίας κειμένου, μπορούν να δημιουργηθούν αποσπάσματα του σεναρίου τα οποία να αντιστοιχούν σε ανεξάρτητες θεματικές ενότητες και να δίνουν μια αρχική εκτίμηση της κατάτμησης.

Βιβλιογραφία

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[2] J. A. Bondy and U. S. R. Murty, *Graph theory with applications*. Macmillan London, 1976, vol. 290.

[3] V. Chasanis, A. Kalogeratos, and A. Likas, "Movie segmentation into scenes and chapters using locally weighted bag of visual words," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, ACM, 2009.

[4] V. Chasanis, C. Likas, and N. Galatsanos, "Scene detection in videos using shot clustering and sequence alignment," *IEEE Transactions on Multimedia*, vol. 11, no. 1, Jan. 2009.

[5] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[6] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, Nov. 2013.

[7] C. Gianluigi and S. Raimondo, "An innovative algorithm for key frame extraction in video summarization," *Journal of Real-Time Image Processing*, vol. 1, no. 1, 2006.

[8] A. Hampapur, T. Weymouth, and R. Jain, "Digital video segmentation," in *Proceedings of the Second ACM International Conference on Multimedia*, San Francisco, California, USA, 1994.

[9] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[10] E. El-Khoury, C. Senac, and P. Joly, "Unsupervised tv program boundaries detection based on audiovisual features," in *5th International Conference on Visual Information Engineering*, 2008.

[11] M. Kipp, "Multimedia annotation, querying and analysis in anvil," *Multimedia information extraction*, vol. 19, 2010.

[12] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.

[13] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *Journal of Computing*, vol. 2, Mar. 2010.

[14] M. Müller, *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc., 2007.

[15] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, 1970.

[16] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Advances in Neural Information Processing Systems*, MIT Press, 2002.

[17] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, 2001.

[18] S.-B. Park, H.-N. Kim, H. Kim, and G.-S. Jo, "Exploiting script-subtitles alignment to scene boundary dectection in movie," in *IEEE International Symposium on Multimedia*, Dec. 2010.

[19] Z. Rasheed and M. Shah, "Scene detection in hollywood movies and tv shows," in *Int. Conf. Computer Vision and Pattern Recognition*, 2003.

[20] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Transactions on Multimedia*, vol. 7, no. 6, 2005.

[21] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, Mar. 1978.

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, 2000.

[23] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, and I. Trancoso, "Multi-modal scene segmentation using scene transition graphs," in *Proceedings of the 17th ACM international conference on Multimedia*, ACM, 2009.

[24] I. Sobel, "An isotropic 3× 3 image gradient operator," *Machine Vision for three-demensional Sciences*, 1990.

[25] R. Turetsky and N. Dimitrova, "Screenplay alignment for closed-system speaker identification and analysis of feature films," in *Procs. IEEE Int. Conf. Multimedia and Expo, ICME*, 2004.

[26] J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," *IEEE Transactions on Multimedia*, vol. 4, no. 4, 2002.

[27] Wikipedia, *Levenshtein distance — wikipedia, the free encyclopedia*, [Online; accessed 3-September-2014], 2014. [Online]. Available: `http://en.wikipedia.org/w/index.php?title=Levenshtein_distance&oldid=623739638`.

[28] M. Yeung, B.-L. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Comput. Vis. Image Underst.*, vol. 71, no. 1, 1998.

[29] R. Zabih, J. Miller, and K. Mai, "A feature-based algorithm for detecting and classifying scene breaks," in *Proceedings of the Third ACM International Conference on Multimedia*, San Francisco, California, USA, 1995.

[30] Y. Zhai and M. Shah, "Video scene segmentation using markov chain monte carlo," *IEEE Transactions on Multimedia*,, vol. 8, no. 4, Aug. 2006.

[31] H. J. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, vol. 1, no. 1, 1993.

[32] B. Zhou and J. Hansen, "Unsupervised audio stream segmentation and clustering via the bayesian information criterion," in *Proc. International Conference on Spoken Language Processing*, 2000.