
Σχολη
Ηλεκτρολογων Μηχανικων
και
Μηχανικων Υπολογιστων

Τεχνολογια και Αναλυση Εικονων και Βιντεο
Εξαμηνιαια Εργασια

---

# Video Summarization
## με Χρήση Μοντέλων Οπτικής Προσοχής και Αντίληψης

---

Φλεμοτόμος Νικόλαος
A.M.: 03110082

9 Μαρτίου 2014

# Περιεχόμενα

# Κατάλογος Σχημάτων

# Κατάλογος Πινάκων

**Περίληψη**

Τα πρωτεύοντα θηλαστικά, με κυρίαρχο τον άνθρωπο, έχουν μια εκπληκτική και μοναδική ικανότητα ταχύτατης εξαγωγής των σημείων ενδιαφέροντος από ένα πολυμεσικό σήμα, είτε πρόκειται για ηχητικό, είτε οπτικό, είτε συνδυασμό αυτών. Εκμεταλλευόμενοι κάποια από τα νευροβιολογικά τεκμήρια που διαθέτουμε για το πώς ο ανθρώπινος εγκέφαλος εκτελεί τη διαδικασία αυτή, μπορούμε να παράγουμε αποδοτικούς και αποτελεσματικούς αλγορίθμους προς τη συγκεκριμένη κατεύθυνση. Στη συγκεκριμένη εργασία, θα εστιάσουμε σε μια μέθοδο αυτόματης παραγωγής περιλήψεων βίντεο (video summaries), βασισμένη σε μοντέλα οπτικής προσοχής. Η εργασία στηρίζεται κυρίως στη δημοσίευση [1], από όπου απομονώθηκε το μέρος εκείνο που αναφέρεται αποκλειστικά στο κομμάτι της κινούμενης εικόνας. Στο 1ο μέρος της εργασίας θα γίνει μια αναφορά στην έννοια του video summarization και στις πιθανές εφαρμογές. Στο 2ο μέρος θα γίνει συνοπτική παρουσίαση των νευροβιολογικών προτύπων που ώθησαν προς την υιοθέτηση του υπό μελέτη μοντέλου. Στο 3ο μέρος θα παρουσιαστεί αναλυτικά η μέθοδος που χρησιμοποιήθηκε, με σύντομες αναφορές σε υπολογιστικές λεπτομέρειες, όπου αυτό κρίνεται απαραίτητο. Στο 4ο και τελευταίο μέρος θα παρουσιαστεί μία σειρά πειραμάτων που έγινε προς αξιολόγηση του μοντέλου.

# 1 Γενική Επισκόπηση του Video Summarization

## 1.1 Γενικά Στοιχεία

Στη γενική τους μορφή, οι περιλήψεις των βίντεο παρέχουν συμπυκνωμένες αναπαραστάσεις του περιεχομένου ενός βίντεο μέσω ενός συνδυασμού ακίνητων εικόνων, τμημάτων του βίντεο, γραφικών αναπαραστάσεων και περιγραφών με χρήση κειμένου [2]. Η μέθοδος που ακολουθείται προς τη δημιουργία μιας περίληψης, έχοντας το αρχικό ολοκληρωμένο βίντεο, καλείται video summarization.

Η πολυτροπική φύση των βίντεο, υπό την έννοια ότι συνδυάζουν πολλά modalities, όπως ήχο, μουσική, κείμενο, στατική και κινούμενη εικόνα, κάνει την όλη διαδικασία αρκετά δυσκολότερη και υπολογιστικά δαπανηρή από την ανάλυση κειμένου ή ήχου για παράδειγμα. Ακόμη, δεν υπάρχει σαφώς καθορισμένο πλαίσιο που να ορίζει ποια στοιχεία θα πρέπει να προστεθούν ή να αφαιρεθούν από την περίληψη που πρόκειται να αφαιρεθεί, οπότε πρόκειται για ένα στοιχείο που σχετίζεται άμεσα με την τελική εφαρμογή για την οποία προορίζεται.

## 1.2 Κατηγοριοποίηση των Διαφορετικών Προσεγγίσεων

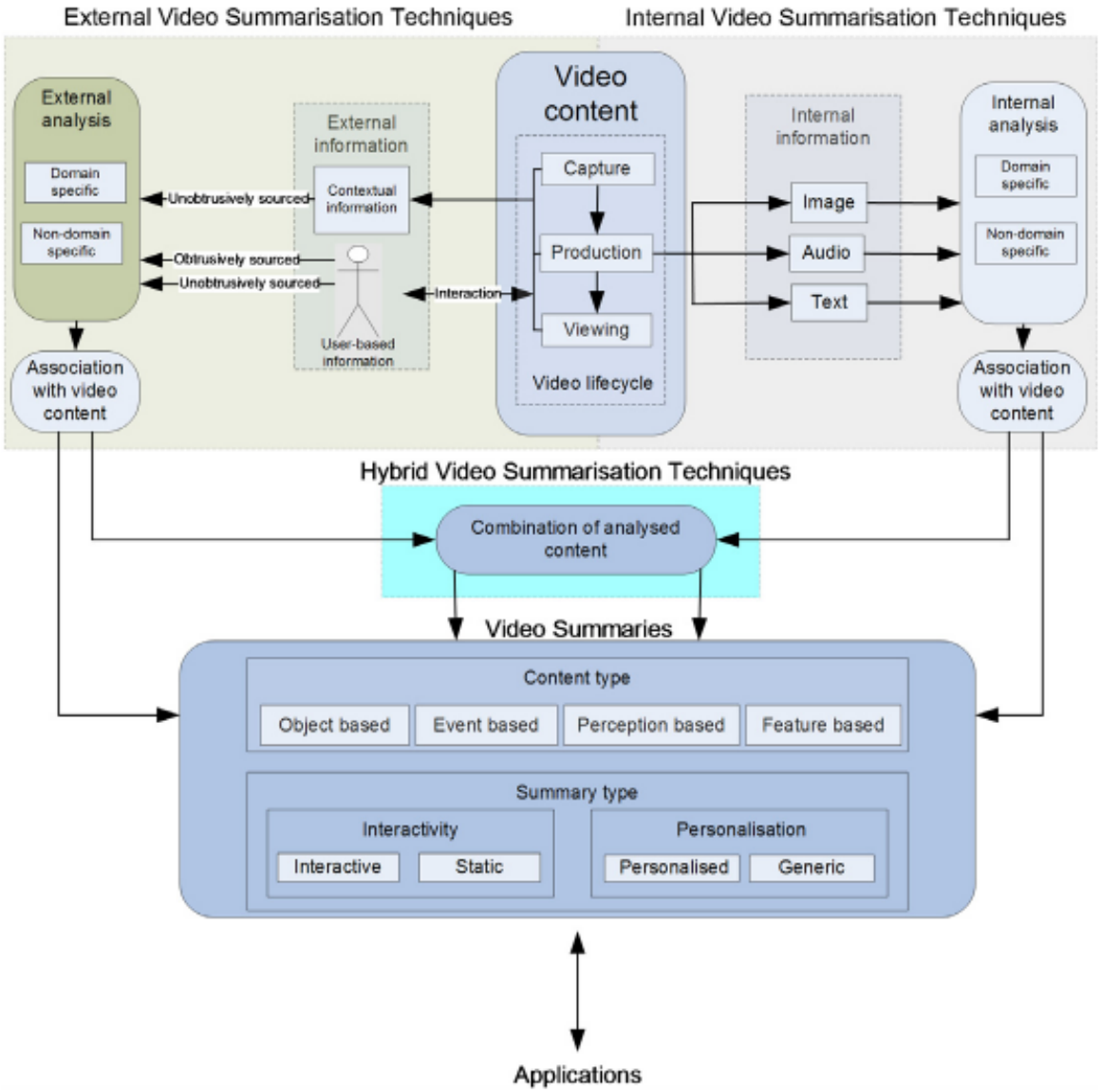
Στη βιβλιογραφία συναντώνται πολλές διαφορετικές προσεγγίσεις του video summarization, οι οποίες μπορούν να κατηγοριοποιηθούν σε συγκεκριμένες μεγάλες ομάδες. Μία αρκετά λεπτομερής κατηγοριοποίηση, τόσο των τεχνικών, όσο και των παραγόμενων περιλήψεων, που έχει προταθεί, παρουσιάζεται εποπτικά στο Σχήμα 1.

Κατ΄ αρχήν, οι χρησιμοποιούμενες τεχνικές μπορούν να διαχωριστούν σε εσωτερικές (internal, εξωτερικές (external) και υβριδικές (hybrid) που εκμεταλλεύονται τα πλεονεκτήματα και των δύο προηγούμενων κατηγοριών. Προτού γίνει περαιτέρω ανάλυση των διαφορετικών κατηγοριών, αναφέρεται πως ο κύκλος ζωής κάθε βίντεο θεωρείται πως αποτελείται από τρία στάδια, την καταγραφή, την παραγωγή, όπου λαμβάνει χώρα η όποια επεξεργασία πριν την τελική χρήση και την προβολή, όπου το βίντεο προβάλλεται σε ένα τελικό κοινό - στόχο.

Έτσι, λοιπόν, διακρίνονται οι εσωτερικές τεχνικές, όπου αναλύεται πληροφορία που πηγάζει άμεσα από το βίντεο κατά την παραγωγή του και οι εξωτερικές όπου η πληροφορία που αναλύεται πηγάζει από οποιοδήποτε στάδιο του κύκλου ζωής του βίντεο, αλλά όχι άμεσα από αυτό, αλλά από εξωτερικούς παράγοντες, όπως οι θεατές, παραγωγοί, κ.λπ. (και οι υβριδικές).

Κάθε μία από τρεις τεχνικές μπορεί να εστιάζει σε συγκεκριμένο πλαίσιο (π.χ. ταινίες δράσης) (domain specific) ή να είναι ανεξάρτητη περιεχομένου (non-domain specific). Οι domain-specific τεχνικές εκμεταλλεύονται a priori γνώση για την επίτευξη καλύτερων αποτελεσμάτων.

Οι πιο διαδεδομένες και ευρέως χρησιμοποιούμενες τεχνικές είναι οι internal. Αυτές εκμεταλλεύονται χαρακτηριστικά της εικόνας, του ήχου ή άλλων συνιστωσών του βίντεο, χωρίς να εμπλέκουν οποιαδήποτε πληροφορία σχετίζεται με το χρήστη. Χαρακτηριστικά εικόνας μπορεί να περιλαμβάνουν απότομες αλλαγές στο χρώμα ή στο σχήμα ή ακόμα και ανίχνευση αντικειμένων και κίνησης. Χαρακτηριστικά ήχου περιλαμβάνουν ομιλία, μουσική, κ.λπ. και μπορεί να αποδειχθούν εξαιρετικά βοηθητικά. Για παράδειγμα, σε έναν ποδοσφαιρικό αγώνα, οι επευφημίες και τα χειροκροτήματα μπορεί να σηματοδοτούν ένα σημείο ενδιαφέροντος όπως κάποιο γκολ. Πολλές φορές χρησιμοποιούνται και χαρακτηριστικά κειμένου, το οποίο κείμενο συνήθως δίνεται σε μορφή υποτίτλων, συγχρονισμένων με το υπόλοιπο βίντεο. Κατάλληλη επεξεργασία κειμένου μπορεί να οδηγήσει σε πλούσια σημασιολογική πληροφορία. Όμοιας φύσης πληροφορία, σε περίπτωση απουσίας κειμένου, θα μπορούσε να εξαχθεί με χρήση τενικών αναγνώρισης φωνής.




Σχήμα 1: Κατηγοριοποίηση των τεχνικών video summarization και των παραγόμενων αποτελεσμάτων.

Οι external τεχνικές χωρίζονται σε δύο επιμέρους κατηγορίες, αυτές που χρησιμοποιούν το χρήστη (user-based) και αυτές που δεν τον χρησιμοποιούν (contextual). Η user-based πληροφορία μπορεί να δοθεί είτε συνειδητά (για παράδειγμα με χρήση κάποιου ερωτηματολογίου σχετικά με τις προτιμήσεις του χρήστη), ή όχι (για παράγειγμα με καταγραφή EEG των εγκεφαλικών σημάτων ή βιντεοσκόπηση του χρήστη κατά την προβολή για αναγνώριση των σημείων ενδιαφέροντος αναλόγως των εκφράσεων του προσώπου). Από την άλλη, η contextual πληροφορία μπορεί να περιλαμβάνει γνώση της γεωγραφικής περιοχής όπου το βίντεο καταγράφηκε ή προβλήθηκε. Ακόμη, μπορεί να χρησιμοποιηθεί το διαδίκτυο για ανίχνευση ενδιαφέροντων συμβάντων. Για παράδειγμα, κατά το summarization αθλητικών μεταδόσεων, μπορεί να γίνει μία ''σάρωση'' αθλητικών sites ώστε γρήγορα να παρθεί μία εκτίμηση για τις χρονικές στιγμές όπου κάτι αξιοσημείωτο που

οφείλει να περιληφθεί στην τελική περίληψη συνέβη.

Όσον αφορά στις περιλήψεις που παράγονται, μία πρώτη κατηγοριοποίηση μπορεί να γίνει ως προς τον τύπο του περιεχομένου που περιλαμβάνουν. Έτσι, προκύπτουν αυτές που βασίζονται σε αντικείμενα (object-based), σε γεγονότα (event-based), σε αντίληψη του χρήστη (perception-based) και σε χαμηλού επιπέδου χαρακτηριστικά (feature-based). Οι πρώτες δύο γίνονται εύκολα κατανοητές από το όνομά τους, καθώς επίσης και η τελευταία. Οι perception-based περιλήψεις εξάγονται βάσει του τρόπου με τον οποίο οι χρήστες αντιλαμβάνονται, ή αναμένεται να αντιληφθούν, την πληροφορία του βίντεο. Σχετίζονται, λοιπόν, με τη σημαντικότητα που αναμένεται να αποδώσουν οι χρήστες στα διάφορα τμήματα του βίντεο και τα συναισθήματα που τους εγείρονται. Για να επιτευχθεί ένα τέτοιο αποτέλεσμα, καθίσταται προφανές ότι είναι συχνά απαραίτητη η συνεισφορά άλλων γνωστικών πεδίων, όπως η σημειωτική, η θεωρία ανθρώπινης αντίληψης από τη νευροεπιστήμη ή οι θεωρίες και τα μοντέλα προσοχής.

Επίσης, μπορούμε να διακρίνουμε στατικές ή διαδραστικές περιλήψεις και γενικές ή εξατομικευμένες. Η διαδραστικότητα μπορεί να εισαχθεί με πολλούς τρόπους, όπως για παράδειγμα με την παροχή της δυνατότητας στο χρήστη να ζητήσει από την εφαρμογή την εξαγωγή μιας περίληψης βάσει ορισμένων σημασιολογικών κριτηρίων. Από την άλλη, μια περίληψη θεωρείται εξατομικευμένη, όταν για την παραγωγή της, το βίντεο έχει φιλτραριστεί με βάση το προφίλ του χρήστη για τον οποίο προορίζεται.

Τέλος, μία σημαντική κατηγοριοποίηση αφορά στα στοιχεία από τα οποία αποτελείται μία περίληψη. Στη βιβλιογραφία εμφανίζονται εν γένει τέσσερις διαφορετικοί τύποι τέτοιων στοιχείων, υπό την έννοια ότι μια περίληψη μπορεί να αποτελείται από έναν εξ αυτών ή συνδυασμό τους. Αυτά είναι frames-κλειδιά, τμήματα βίντεο, γραφικές αναπαραστάσεις ή κείμενο. Χρησιμοποιώντας frames-κλειδιά, απομονώνεται μία σειρά από frames του αρχικού βίντεο που συγκεντρώνουν τη σημαντικότερη πληροφορία και παρουσιάζονται στο χρήστη σαν μια συλλογή από φωτογραφίες. Τα τμήματα βίντεο είναι μία δυναμική επέκταση της πρώτης προσέγγισης, όπου το τελικό αποτέλεσμα είναι ένα καινούριο βίντεο μικρότερης διάρκειας που αποτελείται από τμήματα του αρχικού. Συμπληρωματικό ρόλο συνήθως παίζουν οι γραφικές αναπαραστάσεις, όπου συμβολικά αποτυπώνονται διάφορες συσχετίσεις μέσα στο βίντεο και το κείμενο, με το οποίο δίνεται μια περίληψη της πληροφορίας του βίντεο σε μορφή προς ανάγνωση. Σαφώς σε αυτό το τελευταίο ιδιαίτερα χρήσιμη είναι η πληροφορία που δίνεται από τους υπότιτλους του βίντεο ή από τυχόν κείμενο που εμφανίζεται μέσα σε αυτό.

Σύμφωνα με τις κατηγοριοποιήσεις που προηγήθηκαν, η μέθοδος που θα αναλυθεί στην παρούσα εργασία χαρακτηρίζεται ως internal, non-domain specific. Οι παραγόμενες περιλήψεις είναι νέα βίντεο μικρότερης διάρκειας, είναι μη-εξατομικευμένες και στατικές (προσοχή στη σύγχυση του όρου με τη στατική εικόνα), ενώ, παρόλο που βασίζονται σε low-level χαρακτηριστικά του αρχικού βίντεο, μπορούν να θεωρηθούν ως perception-based, καθώς η τεχνική στηρίζεται εν πολλοίς σε μοντέλα της ανθρώπινης προσοχής.

## 1.3 Εφαρμογές

Οι πιθανές εφαρμογές όπου μπορεί να χρησιμοποιηθεί το video summarization χαρακτηρίζονται από τα δύο μεγάλα ευεργετήματα που η τεχνική αυτή μπορεί να προσφέρει: μείωση του χρόνου και μείωση των υπολογιστικών πόρων που απαιτούνται από το χρήστη για να έχει πρόσβαση στο χρήσιμο περιεχόμενο του βίντεο.

Μία από τις πιθανές εφαρμογές είναι η προσαρμογή ενός βίντεο στις προτιμήσεις του χρήστη (video adaptation). Δημιουργείται, έτσι ένα βίντεο που περιέχει όλη την πληροφορία εκείνη που ο χρήστης θεωρεί σημαντική και απαραίτητη. Ακόμη, μπορεί να βοηθήσει σε ευκολότερη

πλοήγηση μέσα στο σύνολο του βίντεο. Για παράδειγμα, διατηρώντας έναν συγκεκριμένο αριθμό από τα σημαντικότερα frames σε χρονολογική σειρά, ο χρήστης μπορεί εύκολα και γρόγορα σε μεταφερθεί στο επιθυμητό σημείο του βίντεο, χωρίς να είναι απαραίτητη η γνώση του επιθυμητού χρονικού σημείου στη διάρκεια του βίντεο. Πολλές μελέτες, όπως και η [1], στοχεύουν, επίσης, σε χρήση των περιλήψεων για διάφορους προωθητικούς σκοπούς, όπως, για παράδειγμα, για αυτόματη παραγωγή trailers κινηματογραφικών ταινιών.

Τέλος, δε θα πρέπει να θεωρείται αμελητέα και η μείωση των υπολογιστικών πόρων που απαιτούνται μέσω video summarization, ιδίως σε μία εποχή που η ανάγκη της φορητότητας σε συσκευές σχετικά μειωμένων δυνατοτήτων γίνεται ολοένα και μεγαλύτερη. Παρόλο που οι εν λόγω συσκευές γίνονται συνεχώς ισχυρότερες, συνεχής είναι παράλληλα και η αύξηση της ποιότητας, άρα και της απαιτούμενης χωρητικότητας για αποθήκευση, των παρεχόμενων βίντεο.

# 2 Νευροβιολογικά και Ψυχοφυσικά Κίνητρα

## 2.1 Μοντέλα Οπτικής Προσοχής

Τα πρωτεύοντα θηλαστικά έχουν μια εκπληκτική ικανότητα να αναγνωρίζουν και να ερμηνεύουν πολύπλοκες σκηνές σε πραγματικό χρόνο, παρά την περιορισμένη ταχύτητα του νευρολογικού υλισμικού που διαθέτουν για την εργασία αυτή [3]. Νευροεπιστημονικές έρευνες που έχουν γίνει στο θέμα αυτό έχουν αναδείξει ως υψηλής σημασίας το ρόλο της προσοχής για τη συγκεκριμένη εργασία. Η ανάπτυξη υπολογιστικών μοντέλων της ανθρώπινης προσοχής, λοιπόν, είναι μία σημαντική πρόκληση, τόσο για τους γνωστικούς νευροεπιστήμονες, όσο και για ερευνητές κλάδων οι οποίοι μπορούν να επωφεληθούν από τέτοια μοντέλα. Ένας τέτοιος κλάδος είναι και η επεξεργασία εικόνας και βίντεο, με μια εφαρμογή το video summariation.

Σημαντική έρευνα έχει γίνει στο δισδιάστατο χώρο, ενώ στην περίπτωση που εισάγεται και η διάσταση του χρόνου, η έρευνα είναι αρκετά πιο περιορισμένη. Θα εξετασθεί, λοιπόν, το κατά κόρον χρησιμοποιούμενο μοντέλο στις περιπτώσεις δισδιάστατων εικόνων, από όπου η επέκταση στις 3 διαστάσεις είναι σχεδόν τετριμμένη.

Ο εγκέφαλος χρησιμοποιεί τόσο bottom-up, οδηγούμενα από τα χαμηλού επιπέδου χαρακτηριστικά της σκηνής του οπτικού πεδίου, όσο και top-down, οδηγούμενα από την εκάστοτε εργασία που έχει να επιτελέσει, κριτήρια για την απόδοση ενός μέτρου προσοχής σε κάθε σημείο [4]. Η απόδοση υψηλού μέτρου προσοχής σε ένα αντικείμενο πρακτικά σημαίνει την αυτόματη (ή σχεδόν αυτόματη) στροφή του βλέμματος προς το αντικείμενο αυτό. Η ερευνητική προσοχή στρέφεται κυρίως προς την bottom-up προσέγγιση, καθώς από τη μία η μοντελοποίηση μπορεί πιο εύκολα να βασιστεί σε αντικειμενικά κριτήρια και από την άλλη, η top-down προσοχή μπορεί να χρειαστεί έως $200msec$ ανά αντικείμενο σε αντιδιαστολή με την εξαγωγή των χαρακτηριστικών που χρειάζονται για την bottom-up, που είναι 25 με $50msec$, οπότε μπορεί να θεωρηθεί ελάσσονος σημασίας.

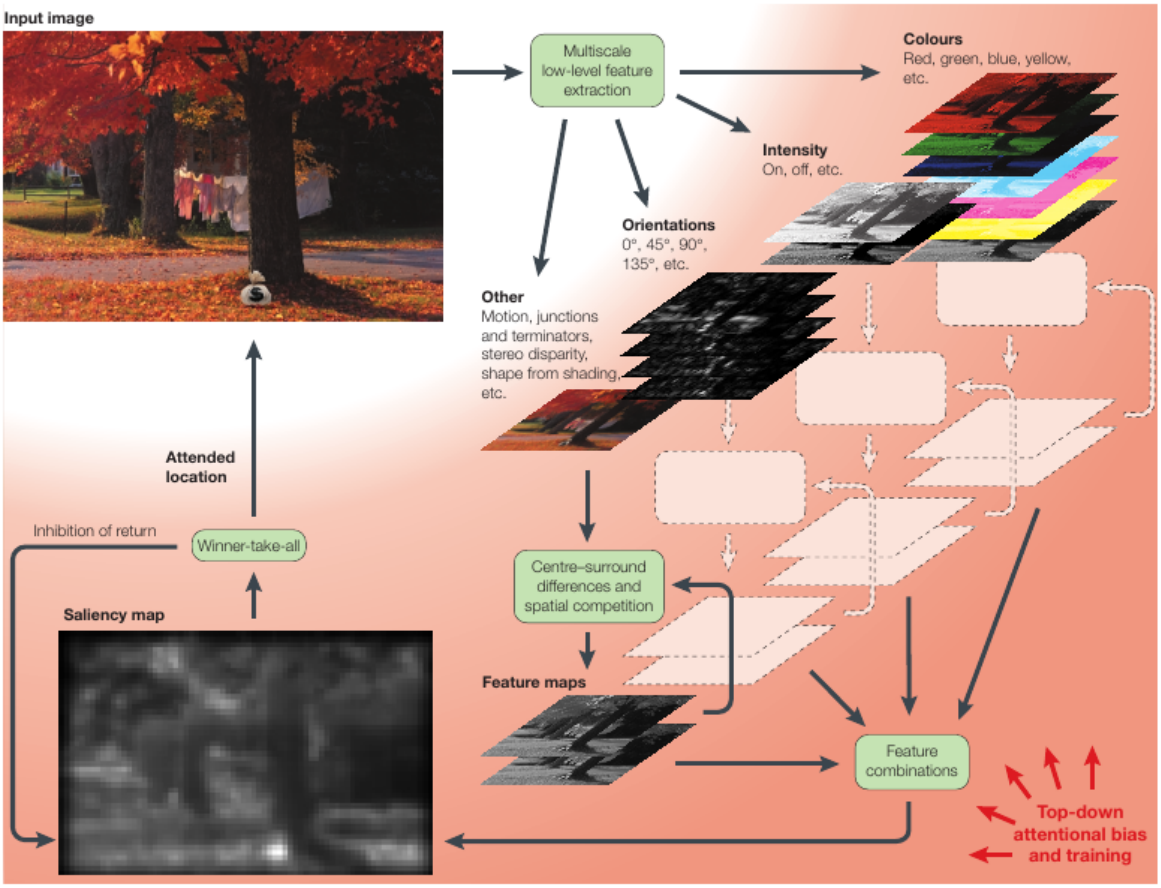
Η εισαγωγή ενός μοντέλου προσοχής σε προβλήματα επεξεργασίας εικόνας ή βίντεο μπορεί να αποδειχθεί σημαντική, καθώς στους βιολογικούς οργανισμούς λειτουργεί με τρόπο τέτοιο ώστε να επιτρέπει τη διαμέριση του προβλήματος της αναγνώρισης ενός οπτικού πεδίου σε επιμέρους προβλήματα τοπικής αναγνώρισης αντικειμένων ή ομάδων αντικειμένων που μπορούν να υπολογιστούν ταχύτερα και με υψηλή παραλληλία. Το πρώτο ολοκληρωμένο και ευρέως αποδεκτό μοντέλο προτάθηκε από τους Koch και Ullman [5] και απεικονίζεται γραφικά στο Σχήμα 2.

Στα πρώτα στάδια της οπτικής επεξεργασίας, οι νευρώνες είναι ευαίσθητοι σε απλά οπτικά χαρακτηριστικά, όπως η φωτεινότητα, η αντίθεση, οι χρωματικές διαφορές, η κατεύθυνση, ο προσανατολισμός, η ταχύτητα. Τα χαμηλού επιπέδου αυτά χαρακτηριστικά υπολογίζονται ταχύτατα και παράλληλα προτού λάβει μέρος η όποια απόφαση που αφορά την προσοχή. Με αυτόν τον τρόπο, η εικόνα εισόδου κατατμίζεται σε κανάλια χαρακτηριστικών. Οι νευρώνες που εξειδικεύονται στο κάθε χαρακτηριστικό δημιουργούν εν τέλει έναν χάρτη του συγκεκριμένου χαρακτηριστικού (feature map), όπου κωδικοποιούνται οι χωρικές αντιθέσεις στο εν λόγω κανάλι. Είναι αποδεκτό πως αυτό που έχει σημασία για την bottom-up προσοχή δεν είναι οι απόλυτες εντάσεις των χαρακτηριστικών, αλλά οι διαφορές μεταξύ γειτονικών σημείων. Οι χωρικές αντιθέσεις εξετάζονται σε πολλαπλές κλίμακες και δημιουργούνται έτσι διαδοχικοί υποχάρτες για κάθε ένα χαρακτηριστικό. Η όλη διαδικασία γίνεται παράλληλα για κάθε κανάλι, καθώς δεν υπάρχουν ενδείξεις για ισχυρές αλληλεπιδράσεις μεταξύ νευρώνων που είναι επιφορτισμένοι με διαφορετικά χαρακτηριστικά.

Ως προς το βιολογικό πρότυπο, ιδιαίτερη αναφορά αξίζει να γίνει στην αντίληψη των χρωμάτων, καθώς θα επηρεάσει το μοντέλο που θα επιλέξουμε εν τέλει. Τα νευρικά κύτταρα του οπτικού φλοιού παρουσιάζουν αντιθέσεις τόσο σε χρωματικό, όσο και σε χωρικό επίπεδο, γι΄ αυτό

και συναντιούνται στη βιβλιογραφία ως κύτταρα "double opponent". Στο κέντρο του δεκτικού τους πεδίου, οι νευρώνες αυτοί διεγείρονται από ένα χρώμα (π.χ. κόκκινο) και αναστέλλονται από ένα άλλο (π.χ. πράσινο), ενώ το αντίθετο ισχύει όσο απομακρυνόμαστε από αυτό. Στον οπτικό φλοιό του ανθρώπινου εγκεφάλου έχουν παρατηρηθεί τέτοιες αντιθέσεις για τα ζεύγη κόκκινου/πράσινου, πράσινου/κόκκινου, μπλε/κίτρινου και κίτρινου/μπλε χρώματος.

Οι δημιουργούμενοι feature maps συνδυάζονται κατάλληλα εν συνεχεία σε έναν μοναδικό χάρτη, γνωστό ως saliency map. Ο χάρτης αυτός αποτελεί μια αναπαράσταση της οπτικής προσοχής και δεν εξαρτάται από το χαρακτηριστικό που είναι υπεύθυνο για τη διέγερση της προσοχής. Ενεργοποιείται στο σημείο αυτό ένας WTA (Winner-Take-All) μηχανισμός που καθορίζει την περιοχή εκείνη που είναι πλέον σημαντική. Μία περιοχή, λοιπόν, χαρακτηρίζεται ως σημαντική, εάν νικήσει τον ανταγωνισμό των νευρώνων σε ένα ή περισσότερα κανάλια χαρακτηριστικών, σε μία ή περισσότερες χωρικές κλίμακες.

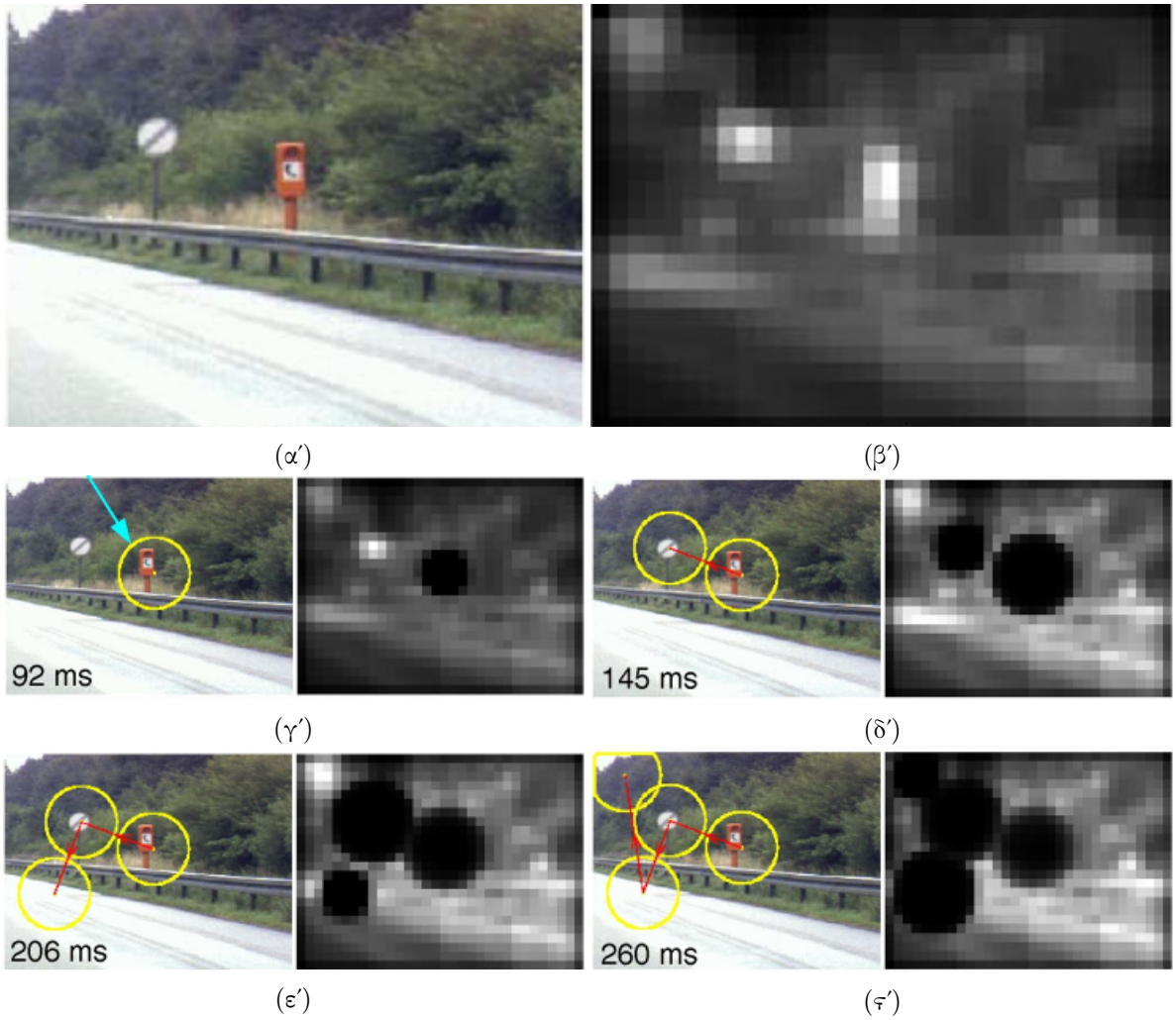



Σχήμα 2: Προτεινόμενο από τους Koch και Ullman μοντέλο οπτικής προσοχής για δισδιάστατες εικόνες.

Η διαδικασία που περιγράφεται, φυσικά, δεν είναι στιγμιαία, αλλά εξελίσσεται δυναμικά στο χρόνο. Υπό αυτή την έννοια, με το πέρας του χρόνου, ένα καινούριο σημείο προσοχής πρέπει να δημιουργείται ανά κάποια χρονικά διαστήματα (περίπου $50 - 100msec$ για τη bottom-up προσοχή που εξετάζουμε). Αυτό επιτρέπεται μέσω της αυτόματης ενεργοποίησης μιας λειτουργίας γνωστής ως αναστολή της επιστροφής (Inhibition of Return - IoR). Με τον τρόπο αυτό, εμποδίζεται το

WTA δίκτυο να αναδεικνύει συνεχώς την ίδια περιοχή του πεδίου ως σημαντική, αλλά κάθε φορά ωθείται στο να επιλέγει την αμέσως λιγότερο σημαντική σε σχέση με την προηγούμενη. Δημιουργείται τελικά ένα μονοπάτι στο χώρο της εικόνας που ορίζεται από τα διαδοχικά σημεία επιλογής του συστήματος προσοχής.

Η παραπάνω διαδικασία έχει μοντελοποιηθεί και υλοποιηθεί σε υπολογιστή στο [3]. Στο Σχήμα 3 παρουσιάζεται ένα αποτέλεσμα από την εν λόγω εργασία, ώστε να γίνει πιο κατανοητή η διαδικασία και η δημιουργία του τελικού μονοπατιού ανθρώπινης προσοχής.




Σχήμα 3: Ένα παράδειγμα του μοντέλου οπτικής προσοχής των Koch-Ullman. (α′): εικόνα εισόδου, (β′): παραγόμενος saliency map, (γ′)-(στ′): αριστερά το τρέχον μονοπάτι προσοχής μέχρι τη στιγμή που αναγράφεται, δεξιά οι περιοχές του saliency map που επιλέγονται από το WTA δίκτυο.

Το μοντέλο που περιγράφηκε είναι και αυτό που θα χρησιμοποιηθεί στην παρούσα εργασία. Η επέκταση στον τρισδιάστατο χώρο είναι απλή εάν σκεφτούμε πως πλέον ο αντίστοιχος saliency map δε θα είναι ένας δισδιάστατος χάρτης, αλλά ένας τρισδιάστατος κύβος, ενώ πλέον δεν ενδιαφερόμαστε για μονοπάτια της οπτικής προσοχής στις δύο διαστάσεις, αλλά για ένα μονοδιάστατο μονοπάτι στον άξονα του χρόνου, το οποίο εκτείνεται προς τη μία κατεύθυνση (από frames που

αντιστοιχούν σε μικρούς χρόνους, προς frames που αντιστοιχούν σε μεγαλύτερους χρόνους). Λεπτομέρειες αναφορικά με την υλοποίηση του μοντέλου θα δοθούν στην πορεία της εργασίας.

## 2.2 Βασικές Αρχές της Ψυχολογίας Gestalt

Το μοντέλο που μόλις περιγράφηκε εξηγεί πως κατά τα πρώτα στάδια της οπτικής επεξεργασίας, διαφορετικά χαρακτηριστικά αναλύνται ανεξάρτητα. Πρέπει, όμως, να καταστεί σαφές πως οι χάρτες χαρακτηριστικών που δημιουργούνται είναι αποτέλεσμα έντονων εσωτερικών αλληλεπιδράσεων, που οδηγούν σε ομαδοποίηση των περιοχών του οπτικού πεδίου ή αντίστοιχα των γενικευμένων περιοχών του κύβου που ορίζει τη χωροχρονική εξέλιξη ενός βίντεο. Υπό αυτή την έννοια, το μοντέλο που θα χρησιμοποιηθεί στην εργασία είναι συνεπές με θεωρίες που υποστηρίζουν πως το ανθρώπινο μάτι βλέπει τα αντικείμενα ως ολότητες, προτού τα διαχωρίσει σε επιμέρους μέρη.

Κύριος εκπρόσωπος της προσέγγισης αυτής είναι η ψυχολογία Gestalt, η οποία αν και βρίσκει εφαρμογή σε όλες τις αισθήσεις, παρουσιάζει ιδιαίτερο ενδιαφέρον όσον αφορά στην όραση. Υποστηρίζει, λοιπόν, ότι ο ανθρώπινος εγκέφαλος δεν αντιλαμβάνεται αυθαίρετες γραμμές και καμπύλες, αλλά ολοκληρωμένα σχήματα και φιγούρες. Η ψυχολογία Gestalt συνοψίζεται σε ένα σύνολο κανόνων, γνωστών ως Αρχές Gestalt, εκ των οποίων, αυτές που βρίσκουν εφαρμογή στην προς χρήση μέθοδο αξίζει να αναφέρουμε.

- *Αρχή της Εγγύτητας*. Όταν σε ένα άτομο παρουσιάζεται ένα σύνολο αντικειμένων, αντιλαμβάνεται αντικείμενα που βρίσκονται κοντά μεταξύ τους ως μία αυτοτελή ομάδα. Ένα παράδειγμα παρουσιάζεται στο Σχήμα 4α′. Συμφωνο με την αρχή αυτή, το μοντέλο μας προσπαθεί να ομαδοποιήσει κοινά χαρακτηριστικά με κριτήριο τη γειτνίαση τόσο στο χώρο όσο και το χρόνο.

- *Αρχή της Ομοιότητας*. Στοιχεία μέσα σε ένα σύνολο αντικειμένων γίνονται αντιληπτά ως ομάδα εάν μοιάζουν μεταξύ τους. Ένα παράδειγμα παρουσιάζεται στο Σχήμα 4γ′. Σύμφωνο με την αρχή αυτή, το μοντέλο μας στηρίζεται σε διαφορές μεταξύ σημείων στον ίδιο χάρτη χαρακτηριστικών για να εξαγάγει τις πλέον σημαντικές ως προς την έγερση της οπτικής προσοχής περιοχές, και όχι απόλυτες τιμές.

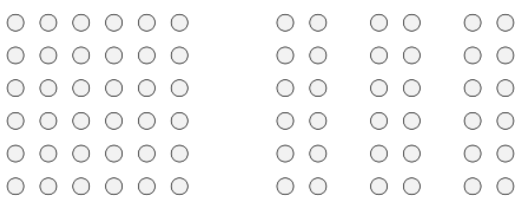
- *Αρχή της Κλειστότητας*. Τα άτομα αντιλαμβάνονται ως ολότητες τα αντικείμενα, ακόμα και αν στην πράξη αυτά εμφανίζουν ατέλειες και δεν είναι πλήρη. Για την ακρίβεια, όταν μέρη μιας εικόνας απουσιάζουν, η αντίληψή μας ''γεμίζει'' το εν λόγω κενό. Ένα παράδειγμα παρουσιάζεται στο Σχήμα 4β′. Σύμφωνες με την αρχή αυτή, οι παραγόμενες περιλήψεις δεν παρουσιάζουν μη-αναμενόμενα κενά. Πιο συγκεκριμένα, εάν αποφασιστεί πως δύο τμήματα του αρχικού βίντεο ανήκουν στην περίληψη, αλλά απέχουν μεταξύ τους λίγα μόνο frames, τότε θα περιληφθούν και τα frames αυτά στην περίληψη.

- *Αρχή του Prägnanz*. Στοιχεία ενός συνόλου αντικειμένων ομαδοποιούνται μαζί εάν σχηματίζουν ένα πρότυπο που είναι συνηθισμένο και απλό. Έτσι, τα άτομα αντιλαμβάνονται την πραγματικότητα στην πιο απλή της μορφή, εξαλείφοντας κατά το δυνατόν την όποια πολυπλοκότητα. Σύμφωνη με την αρχή αυτή, η μέθοδος δε βασίζεται μόνο σε ένα μοντέλο οπτικής προσοχής, αλλά εισάγονται κάποιοι όροι εξομάλυνσης σύμφωνα με κάποια κριτήρια, όπως θα δούμε και στη συνέχεια.

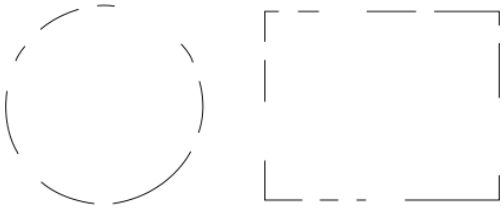
- *Αρχή της Κοινής Μοίρας*. Τα αντικείμενα γίνονται αντιληπτά ως γραμμές που κινούνται κατά μήκος της πιο ομαλής δυνατής γραμμής. Έτσι, όταν ένα σύνολο αντικειμένων κινούνται

στην ίδια κατεύθυνση και με τον ίδιο ρυθμό, η αντίληψη συνδέει την κίνηση ως μέρος του ίδιου του οπτικού ερεθίσματος. Για παράδειγμα, τα πουλιά διακρίνονται από το υπόλοιπο οπτικό πεδίο ως ένα μοναδικό σμήνος επειδή κινούνται μαζί, ακόμα και αν το κάθε ένα πουλί φαίνεται σαν μια μικρή κουκκίδα. Το σύνολο αυτό των κουκκίδων γίνεται αντιληπτό ως μια ενιαία οντότητα. Σύμφωνο με την αρχή αυτή, η μέθοδος προσεγγίζει το πρόβλημα όπως θα δούμε μέσω της ελαχιστοποίησης μιας ενέργειας, γεγονός που παράγει χωροχρονικές περιοχές που γίνονται αντιληπτές ως συνεκτικές και ομογενείς.

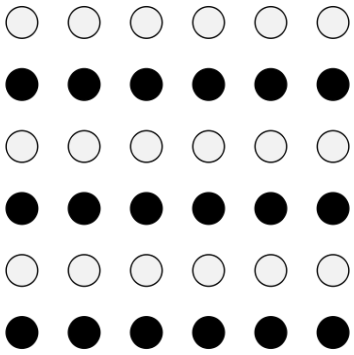
Κυρίαρχης σημασίας στην ψυχολογία Gestald είναι η έννοια του διαχωρισμού της εικόνας από το υπόβαθρο (figure/ground separation). Εικάζεται πως ο εγκέφαλος χρησιμοποιεί ένα πιθανοτικό μοντέλο για το διαχωρισμό αυτό που σχετίζεται με τις εμπειρίες του ατόμου, άρα με την έως τώρα μάθηση του εγκεφάλου. Για παράδειγμα, διαβάζοντας ένα κείμενο, μας φαίνεται αυτονόητο πως τα γράμματα αποτελούν την υπό αναγνώριση εικόνα και οι λευκές περιοχές του χαρτιού αποτελούν το υπόβαθρο, αλλά στο Σχήμα 5 η απάντηση δεν είναι προφανής ούτε ίδια για τον καθένα. Η μέθοδος που χρησιμοποιείται σχετίζεται στενά με την έννοια αυτή, καθώς στηρίζεται στην ελαχιστοποίηση κάποιας ενέργειας (όπως το πιθανοτικό μοντέλο του εγκεφάλου στηρίζεται στην ελαχιστοποίηση κάποιου σφάλματος), όπου σε κάθε επανάληψη η σημασία που προσδίδεται στο υπόβαθρο συνεχώς καταστέλλεται.



(α′) Αρχή της Εγγύτητας.

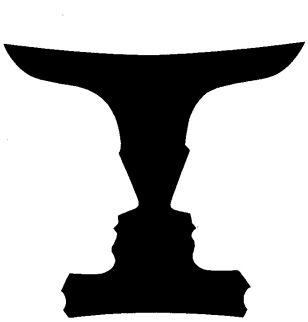


(β′) Αρχή της Κλειστότητας.



(γ′) Αρχή της Ομοιότητας.

Σχήμα 4: Κάποιες από τις Αρχές Gestalt. (α′): Αντιλαμβανόμαστε 1 ομάδα από 36 στοιχεία και 3 ομάδες από 12 στοιχεία. (β′): Αντιλαμβανόμαστε έναν κύκλο και ένα ορθογώνιο, παρόλο που δεν εμφανίζονται κλειστές καμπύλες. (γ′): Αντιλαμβανόμαστε 2 ομάδες, μία από μαύρα στοιχεία, μία από λευκά.



(α΄) Δύο πρόσωπα ή ένα βάζο;

(β΄) Πρόσωπο ή σαξοφωνίστας;

Σχήμα 5: Figure/Ground Separation. Η απάντηση στις ερωτήσεις (α΄) και (β΄) εξαρτάται από το τι θεωρούμε ως κυρίως σχήμα και τι ως υπόβαθρο.

Τέλος, ας σημειωθεί ότι, όπως ίσως φάνηκε από την έως τώρα ανάλυση, η ψυχολογία Gestalt στηρίζεται σε μεγάλο βαθμό στην περιεκτικότητα, στην οποία η ανθρώπινη αντίληψη οδηγείται μέσω της ομαδοποίησης των παρατηρούμενων αντικειμένων. Η μέθοδος που υλοποιούμε, οπότε, είναι φυσιολογικό να σχετίζεται άμεσα με την ψυχολογία Gestalt, αφού το video summarization δεν είναι παρά μία προσπάθεια περιεκτικής διατύπωσης της πληροφορίας του βίντεο εισόδου.

# 3 Υλοποίηση της Μεθόδου

## 3.1 Εξαγωγή Οπτικών Χαρακτηριστικών

Όπως είδαμε, πρώτο βήμα ενός υπολογιστικού μοντέλου της οπτικής προσοχής είναι η εξαγωγή των οπτικών χαρακτηριστικών και η δημιουργία των ανάλογων feature maps. Όταν εισάγεται και ο χρόνος ως τρίτη διάσταση, οι προκύπτοντες χάρτες δεν είναι δισδιάστατοι, αλλά βρίσκονται στον τρισδιάστατο χώρο. Οι δομές που προκύπτουν καλούνται conspicuity volumes. Δημιουργούνται, λοιπόν, conspicuity volumes που βασίζονται σε 3 χαμηλού επιπέδου χαρακτηριστικά, και συγκεκριμένα την ένταση, το χρώμα και τον προσανατολισμό.

Σημειώνουμε εδώ πως και το αρχικό βίντεο θεωρείται ως μια δομή, έστω $Q$, στον τρισδιάστατο χώρο. Το κάθε στοιχείο της δομής αυτής, έστω $q$, δεν είναι παρά το pixel ενός frame, στοιχείο το οποίο ονομάζεται voxel (< volume + pixel). Κάθε voxel χαρακτηρίζεται κατά τα γνωστά από τρεις αριθμούς που αντιστοιχούν στην ένταση του voxel στα τρία βασικά χρώματα κατά το πρωτόκολλο RGB.

### 3.1.1 Ένταση

Η ένταση σε κάθε ένα voxel $q$ δίνεται ως

$$I(q) = \frac{r(q) + g(q) + b(q)}{3} \tag{1}$$

, όπου $r$, $g$, $b$ οι εντάσεις στα τρία χρωματικά κανάλια. Βάσει αυτού, το conspicuity volume της έντασης $C_1$ δίνεται ως

$$C_1(q) = \left| I(q) - \frac{1}{|N_q|} \sum_{u \in N_q} I(u) \right| \tag{2}$$

Με $N_q$ συμβολίζεται η γειτονιά ενός voxel. Άμεση επέκταση της γειτονιάς 8 γειτόνων στο χώρο της εικόνας είναι η γειτονιά μεγέθους 26 γειτόνων στον τρισδιάστατο χώρο. Πρακτικά, ο αφαιρετέος της εξίσωσης (2) υπολογίζεται μέσω μιας παραλλαγής ενός (τρισδιάστατου) φιλτραρίσματος μέσου όρου, όπου όμως δεν περιλαμβάνεται στον υπολογισμό του μέσου όρου κάθε φορά το υπό εξέταση στοιχείο του $Q$. Με άλλα λόγια κάνουμε ένα τρισδιάστατο φιλτράρισμα με μία μάσκα $N \in \mathbf{\Pi}_{3\times3\times3}$, όπου

$$N(1,:,:) = \frac{1}{26}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, N(2,:,:) = \frac{1}{26}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}, N(3,:,:) = \frac{1}{26}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{3}$$

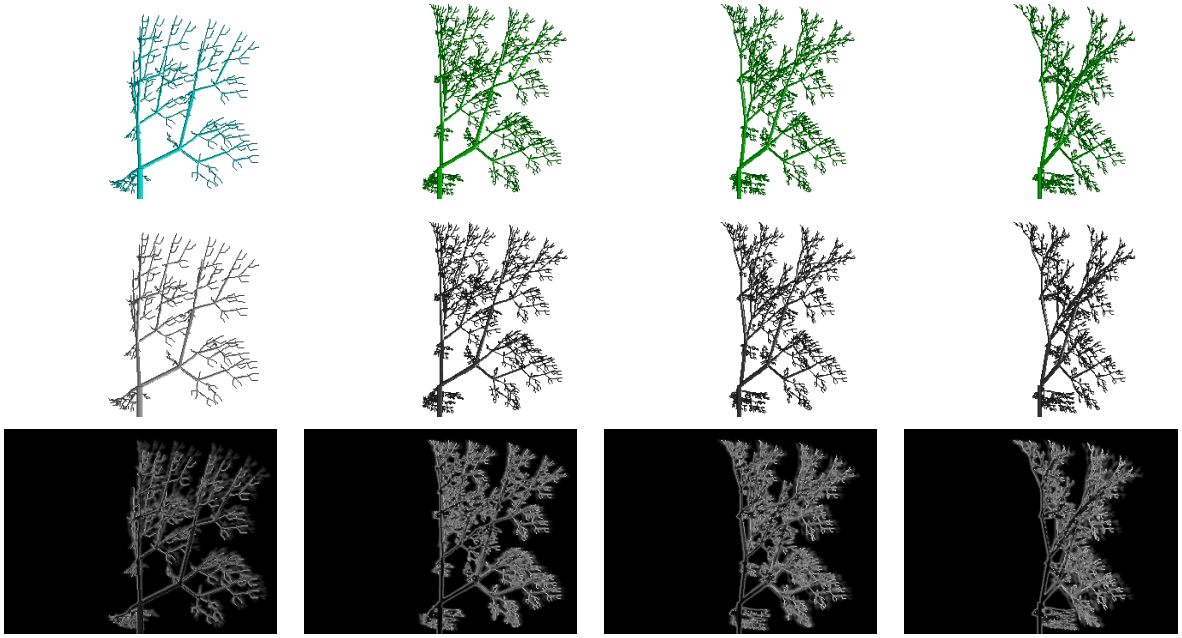
Η εξίσωση (2), λοιπόν, μετασχηματίζεται στην εξίσωση (4), χρησιμοποιώντας φορμαλισμό πινάκων.

$$C_1 = \left| I - I *^3 N \right| \text{[1]} \tag{4}$$

Ένα παράδειγμα εξαγωγής του conspicuity volume της έντασης παρουσιάζεται στο Σχήμα 6. Εκεί φαίνεται καθαρά η επίδραση του χωροχρονικού φιλτραρίσματος ''διαφορών και μέσου όρου''. Φαίνεται, για παράδειγμα πως παρόλο που η ένταση μπορεί να είναι μέγιστη (άσπρη απεικόνιση), από τη στιγμή που είναι σταθερή, η αντίστοιχη περιοχή του conspicuity volume της έντασης μηδενίζεται (μαύρη απεικόνιση). Ακόμη, όπως ήταν αναμενόμενο, οι ακμές αμβλύνονται

[1] Με $*^3$ συμβολίζουμε τη συνέλιξη στον τρισδιάστατο χώρο πινάκων.

και εξομαλύνονται. Η πληροφορία των ακμών, φυσικά, δε χάνεται, αλλά θα περιληφθεί στη συνέχεια της εξαγωγής χαρακτηριστικών (Ενότητα 3.1.3).



Σχήμα 6: Αποτελέσματα της εξαγωγής των χαρακτηριστικών μεταβολής έντασης. Ως είσοδος δίνεται ένα βίντεο διάρκειας 117 frames μιας περιστεφόμενης δομής που αλλάζει χρώμα και μέγεθος σε λευκό υπόβαθρο. Κάθε στήλη αντιστοιχεί σε ένα συγκεκριμένο frame. 1η γραμμή: αρχικό βίντεο. 2η γραμμή: ένταση $I$. 3η γραμμή: frames από το τελικό conspicuity volume της έντασης.

### 3.1.2 Χρώμα

Έχοντας τους τρισδιάστατους πίνακες $r$, $g$, $b$, και ακολουθώντας τη λογική του βιολογικού προτύπου που περιγράφηκε στην Ενότητα 2.1, άμεσα καταλήγουμε στο conspicuity volume των χρωματικών χαρακτηριστικών $C_2$ (δουλεύουμε κατευθείαν με φορμαλισμό πινάκων):

$$C_2 = (RG + BY) \tag{5}$$

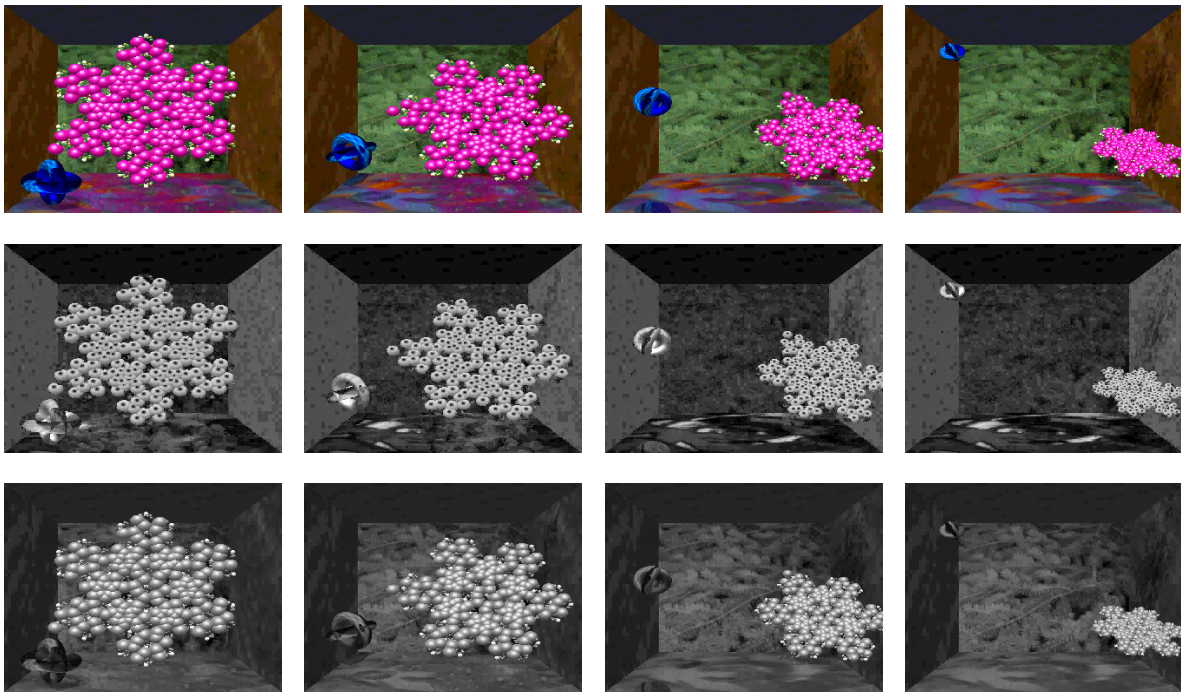
όπου

$$RG = |R - G|, \quad BY = |B - Y|$$

$$R = r - \frac{g+b}{2}, \quad G = g - \frac{r+b}{2}, \quad B = b - \frac{r+g}{2}, \quad Y = \frac{r+g}{2} - \frac{|r-g|}{2} - b$$

Εάν κάποιο στοιχείο εκ των $R$, $G$, $B$, $Y$ προκύψει σύμφωνα με τις παραπάνω εξισώσεις αρνητικό, τότε τίθεται ίσο με 0 [3].

Ένα χαρακτηριστικό παράδειγμα του conspicuity volume των χρωματικών χαρακτηριστικών παρουσιάζεται στο Σχήμα 7.



Σχήμα 7: Αποτελέσματα της εξαγωγής των χρωματικών χαρακτηριστικών. Ως είσοδος δίνεται ένα βίντεο διάρκειας 17 frames με έντονες χρωματικές αντιθέσεις. Κάθε στήλη αντιστοιχεί σε ένα συγκεκριμένο frame. 1η γραμμή: αρχικό βίντεο. 2η γραμμή: frames από το τελικό conspicuity volume για τα χρωματικά χαρακτηριστικά. 3η γραμμή: ένταση $I$, η οποία παρατίθεται για να καταστεί σαφής η διαφορά μεταξύ έντασης και χρωματικού saliency (ας παρατηρήσουμε π.χ. το δάπεδο).

### 3.1.3 Προσανατολισμός

Ο προσανατολισμός, και για την ακρίβεια οι διαφορές γειτονικών περιοχών ως προς τον προσανατολισμό τους, υπολογίζεται με χρήση χωροχρονικών κατευθυντικών φίλτρων (τρισδιάστατων, δηλαδή, φίλτρων) που αποκρίνονται σε ερεθίσματα κίνησης. Η επιλογή αυτή γίνεται γιατί έχει αποδειχθεί πως πληροφορία σχετική με την κίνηση μπορεί να εξαχθεί από ένα σύστημα που αποκρίνεται στην κατευθυνόμενη χωροχρονική ενέργεια [10]. Στη συνέχεια, κρίνεται απαραίτητο να γίνει μία σύντομη αναφορά στις βασικές ιδέες που αφορούν τα κατευθυντικά φίλτρα (oriented filters).

Πρόκειται για ειδική κατηγορία φίλτρων που έχουν την ιδιότητα να επιτρέπουν τη διέλευση συχνοτήτων μιας εικόνας μόνο κατά κάποια κατεύθυνση του πεδίου συχνοτήτων [12]. Βρίσκουν εφαρμογή σε μία ευρεία γκάμα εφαρμογών όπως ανάλυση υφής, ανίχνευση ακμών, ανάλυση κίνησης. Στις αναφερθείσες (αλλά και άλλες) περιοχές, είναι χρήσιμη η εφαρμογή φίλτρων αυθαίρετου προσανατολισμού και ο έλεγχος των αποκρίσεών τους. Ίδια προσέγγιση θα ακολουθηθεί και εδώ, αφού, όπως είδαμε, ο πρασανατολισμός είναι ένα από τα χαμηλού επιπέδου χαρακτηριστικά που παίζουν σημαντικό ρόλο στην οπτική προσοχή. Εγείρεται, λοιπόν, το πρόβλημα της υλοποίησης τέτοιων φίλτρων.

Μία πρώτη προσέγγιση είναι η δημιουργία πολλών παραλλαγών το ίδιου φίλτρου, η καθεμία διαφορετική από τις υπόλοιπες κατά μία μικρή γωνία. Προφανώς, μία τέτοια υλοποίηση είναι

εξαιρετικά υπολογιστικά ακριβή και μη αποδοτική. Μία άλλη στρατηγική θα ήταν η υλοποίηση ενός περιορισμένου αριθμού φίλτρων σε συγκεκριμένους προσανατολισμούς και η χρήση κάποιου αλγορίθμου παρεμβολής για τον έλεγχο της απόκρισης σε αυθαίρετη γωνία προσανατολισμού. Στο [11] εισάγεται η έννοια των στρεφόμενων φίλτρων (steerable filters), όπου φίλτρα αυθαίρετου προσανατολισμού μπορούν να δημιουργηθούν, εάν ικανοποιούν ορισμένες αυστηρές προϋποθέσεις, ως γραμμικός συνδυασμός ενός συνόλου ''φίλτρων βάσης''. Μία υποκατηγορία των στρεφόμενων κατευθυντικών φίλτρων είναι αυτά τα οποία είναι και διαχωρίσιμα, χαρακτηριστικό που οδηγεί σε πολύ μικρότερες ταχύτητες εκτέλεσης. Η διαδικασία υπολογισμού αναλύεται διεξοδικά για την περίπτωση δύο διαστάσεων, ενώ η επέκταση στις τρεις διαστάσεις γίνεται στο [13].

Το πρόβλημα με τα χωροχρονικά κατευθυνόμενα φίλτρα για την ανάλυση της κίνησης είναι πως πρόκειται για φίλτρα εξαιρετικά ευαίσθητα στη φάση [10]. Αυτό σημαίνει πως η απόκρισή τους κάποια στιγμή σε ένα κινούμενο πρότυπο εξαρτάται από το κατά πόσο το πρότυπο είναι ευθυγραμμισμένο με το δεκτικό τους πεδίο τη συγκεκριμένη στιγμή. Η έξοδος μπορεί να είναι θετική, αρνητική ή μηδενική, και άρα, η στιγμιαία απόκριση του φίλτρου δε σηματοδοτεί άμεσα την ύπαρξη (ή όχι) κίνησης. Πιθανώς προτιμούμε μία απόκριση σταθερής τιμής για σταθερή κίνηση.

Προς αυτή την κατεύθυνση, χρησιμοποιούμε δύο ξεχωριστά φίλτρα (για κάθε κατεύθυνση), με διαφορά φάσης $\pi/2$. Το κάθε ένα από αυτά τα φίλτρα συνεχίζει προφανώς να ''υποφέρει'' από το αναφερθέν πρόβλημα, αλλά το πρόβλημα παύει να υφίσταται εάν ως μέτρο της κίνησης θεωρήσουμε το άθροισμα των τετραγωνικών αποκρίσεων των δύο φίλτρων. Η ιδέα είναι απλή και βασίζεται στο γεγονός ότι $\sin^2 + \cos^2 = 1$. Συνήθης πρακτική είναι η χρήση δύο όμοιων Gabor φίλτρων, εκ των οποίων το ένα είναι διαμορφωμένο με ημίτονο και το άλλο με συνημίτονο.

Ωστόσο, θέλουμε να έχουμε μία εύκολη σχετικά υλοποίηση στρεφόμεων διαχωρίσιμων φίλτρων. Μία κατηγορία τέτοιων φίλτρων, είναι αυτά που γράφονται ως γινόμενο ενός πολυωνύμου άρτιας ή περιττής συμμετρίας επί έναν σφαιρικά συμμετρικό πυρήνα. Συμβολίζοντας ως $f^{\mathfrak{R}}$ το φίλτρο που έχει περισταφεί ώστε ο άξονας συμμετρίας του να είναι στη διεύθυνση που ορίζουν τα διευθύνοντα συνημίτονα $(\alpha, \beta, \gamma)$, $W(r)$ το συμμετρικό πυρήνα, όπου $r = \sqrt{x^2 + y^2 + z^2}$ και $P_N$ το πολυώνυμο τάξης $N$, έχουμε

$$f^{\mathfrak{R}}(x, y, z) = W(r)P_N(x') \tag{6}$$

$$x' = \alpha x + \beta y + \gamma z \tag{7}$$

Τότε, προκύπτει το εξής σημαντικό αποτέλεσμα:

$$f^{\mathfrak{R}}(x, y, z) = \sum_{j=1}^{M} k_j(\alpha, \beta, \gamma) f^{\mathfrak{R}_j}(x, y, z) \tag{8}$$

όπου

$$M \geq \frac{(N+1)(N+2)}{2} \tag{9}$$

$$\begin{bmatrix} \alpha^N \\ \alpha^{N-1}\beta \\ \alpha^{N-1}\gamma \\ \alpha^{N_2}\beta^2 \\ \vdots \\ \gamma^N \end{bmatrix} = \begin{bmatrix} \alpha_1^N & \alpha_2^N & \cdots & \alpha_M^N \\ \alpha_1^{N-1}\beta_1 & \alpha_2^{N-1}\beta_2^2 & \cdots & \alpha_M^{N-1}\beta_M \\ \alpha_1^{N-1}\gamma_1 & \alpha_2^{N-1}\gamma_2^2 & \cdots & \alpha_M^{N-1}\gamma_M \\ \alpha_1^{N-2}\beta_1^2 & \alpha_2^{N-2}\beta_2^2 & \cdots & \alpha_M^{N-2}\beta_M^2 \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_1^N & \gamma_2^N & \cdots & \gamma_M^N \end{bmatrix} \begin{bmatrix} k_1(\alpha, \beta, \gamma) \\ k_2(\alpha, \beta, \gamma) \\ \vdots \\ k_M(\alpha, \beta, \gamma) \end{bmatrix} \tag{10}$$

Μία επιλογή είναι η χρήση της δεύτερης Γκαουσιανής παραγώγου για το ένα εκ των δύο φίλτρων. Φορμαλιστικά, ένας Γκαουσιανός πυρήνας και η 2η παράγωγός του ως προς $x$ δίνονται από τις σχέσεις 11 και 12.

$$G(x, y, z) = e^{-(x^2+y^2+z^2)} \tag{11}$$

$$\frac{\partial^2 G}{\partial x^2} = (4x^2 - 2)e^{-(x^2+y^2+z^2)} \tag{12}$$

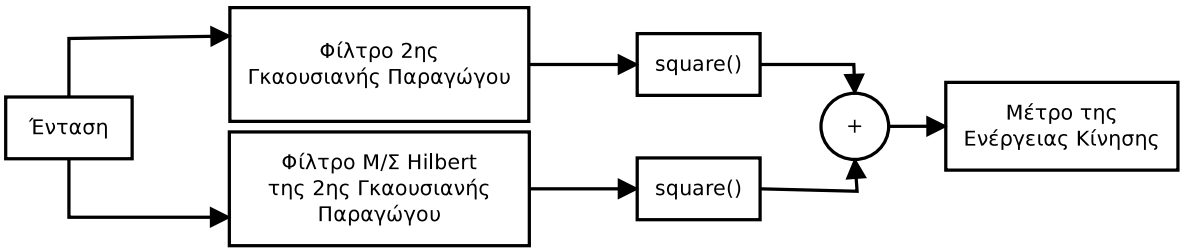
Οπότε, για το δεύτερο φίλτρο, όπου θέλουμε διαφορά φάσης από το πρώτο κατά $\pi/2$ χρησιμοποιούμε το Μετασχηματισμό Hilbert του (12).

$$H_2(x, y, z) = (-2.254x + x^3)e^{-(x^2+y^2+z^2)} \tag{13}$$

Η εξίσωση (13) προκύπτει ως μια προσέγγιση ελαχίστων τετραγώνων ώστε να προκύψει πολυώνυμο τρίτου βαθμού πολλαπλασιασμένο με Γκαουσιανή.

Προφανώς, για την ανίχνευση κίνησης βάσει προσανατολισμού δε μας ενδιαφέρει το χρώμα, παρά μόνο η ένταση $I$, όπως δίνεται από τη σχέση (1). Όλα τα παραπάνω συνοψίζονται στο διάγραμμα του Σχήματος 8, το οποίο αναφέρεται σε μία κατεύθυνση με χρήση ενός ζεύγους φίλτρων βάσης. Τονίζουμε στο σημείο αυτό ότι, εφόσον αναφερόμαστε σε χωροχρονικά φίλτρα, η έννοια ''κίνηση'' έχει μια γενικευμένη σημασία, καθώς, αναλόγως του προσανατολισμού του φίλτρου, μπορεί να δηλώνει μεταβολή στο χρόνο (οπότε κίνηση με τη συνήθη έννοια) ή μεταβολή στο χώρο (οπότε παρουσία ακμών).

Σύμφωνα με τη σχέση (9), για τη 2η Γκαουσιανή παράγωγο $G_2$ απαιτούνται 6 συναρτήσεις βάσης , ενώ για το Μετασχηματισμό Hilbert αυτής $H_2$ απαιτούνται 10. Χωρίς να υπεισέλθουμε σε επιπλέον λεπτομέρειες υπολογισμού, τα αποτελέσματα παρατίθενται στους Πίνακες 1 - 6.




Σχήμα 8: Διαγραμματική απεικόνιση της μεθόδου για την εξαγωγή ενός μέτρου της ενέργειας κίνησης αναφορικά με *μία* μόνο κατεύθυνση.

Το αποτέλεσμα από την όλη διαδικασία για κάποια αυθαίρετη κατεύθυνση, έστω $\Omega$, θα είναι μία δομή ίσων διαστάσεων με τον όγκο $Q$, την οποία συμβολίζουμε $E_\Omega$. Τελικά, εάν εξετάσουμε $|\Omega|$ διαφορετικές κατευθύνσεις στο χώρο, το conpicuity volume για τον προσανατολισμό θα είναι το

$$C_3 = \frac{\sum_\Omega E_\Omega}{|\Omega|} \tag{14}$$

Σημειώνουμε πως, σε αντίθεση με τα $C_1$ και $C_2$, οι τιμές των στοιχείων του $C_3$ δεν κυμαίνονται σε κάποιο προκαθορισμένο εύρος. Μία λύση για να προκύψουν συγκρίσιμες τιμές στα τρία conspicuity volumes είναι διαιρέσουμε με τη μέγιστη τιμή, ώστε να κανονικοποιηθούν και οι τιμές του $C_3$ στο [0,1]. Ωστόσο, η πιθανή ύπαρξη ενός πολύ μεγάλου μεγίστου εγκυμονεί κινδύνους ως

προς την προσέγγιση αυτή. Επιλέγουμε, λοιπόν, να πολλαπλασιάσουμε το $C_3$ με μια πολλαπλασιαστική σταθερά που θα καταστέλλει τις τιμές του (εφόσον σε όλα τα πειράματα βρέθηκαν τιμές μεγαλύτερες της μονάδας). Από τις δοκιμές που έγιναν, παρατηρήθηκε ότι υποδεκαπλασιασμός των τιμών δίνει ικανοποιητικά αποτελέσματα και συγκρίσιμα μεγέθη. Άρα, τελικά:

$$C_3 = \frac{\sum_\Omega E_\Omega}{10|\Omega|} \tag{15}$$

| Βάση | Παρεμβολή |
|---|---|
| $G_{2a} = N(2x^2-1)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = \alpha^2$ |
| $G_{2b} = N(2xy)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 2\alpha\beta$ |
| $G_{2c} = N(2y^2-1)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = \beta^2$ |
| $G_{2d} = N(2xz)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 2\alpha\gamma$ |
| $G_{2e} = N(2yz)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 2\beta\gamma$ |
| $G_{2f} = N(2z^2-1)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = \gamma^2$ |

Πίνακας 1: Συναρτήσεις βάσης και συναρτήσεις παρεμβολής για τη δεύτερη παράγωγο της 3-διάστατης Γκαουσιανής. Η $N$ είναι σταθερά κανονικοποίησης που ισούται με $\frac{2}{\sqrt{3}}\left(\frac{2}{\pi}\right)^{3/4}$ ώστε το ολοκλήρωμα του τετραγώνου της συνάρτησης να ισούται με τη μονάδα. Ένα φίλτρο με άξονα συμμετρίας στη διεύθυνση $\Omega = (\alpha,\beta,\gamma)$ (τα αντίστοιχα διευθύνοντα συνημίτονα) δημιουργείται ως $G_2^\Omega = \sum_{i\in\{a,\cdots,f\}} k_i(\Omega)G_{2i}$.

| | Μονοδιάστατη Συνάρτηση | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| $f_1$ | $N(2t^2-1)e^{-t^2}$ | -0.8230 | -0.0537 | 0.3540 | 0.1025 | 0.0084 |
| $f_2$ | $e^{-t^2}$ | 1.0000 | 0.6383 | 0.1660 | 0.0176 | 0.0008 |
| $f_3$ | $2Nte^{-t^2}$ | 0.0000 | 0.7039 | 0.3662 | 0.0582 | 0.0034 |
| $f_4$ | $te^{-t^2}$ | 0.0000 | 0.4277 | 0.2225 | 0.0354 | 0.0020 |

Πίνακας 2: Διακριτά φίλτρα 5 σημείων για το σύνολο συναρτήσεων βάσης για το $G_2$. Τα $f_1$, $f_2$ έχουν άρτια συμμετρία, τα $f_3$, $f_4$ περιττή. Τα φίλτρα αυτά είναι από τον Πίνακα 1 με δειγματοληψία ανά 0.67.

| Φίλτρο Βάσης | Φίλτρο στον $x$ | Φίλτρο στον $y$ | Φίλτρο στον $z$ |
|---|---|---|---|
| $G_{2a}$ | $f_1$ | $f_2$ | $f_2$ |
| $G_{2b}$ | $f_3$ | $f_4$ | $f_2$ |
| $G_{2c}$ | $f_2$ | $f_1$ | $f_2$ |
| $G_{2d}$ | $f_3$ | $f_2$ | $f_4$ |
| $G_{2e}$ | $f_2$ | $f_3$ | $f_4$ |
| $G_{2f}$ | $f_2$ | $f_2$ | $f_1$ |

Πίνακας 3: Κατασκευή των τρισδιάστατων φίλτρων βάσης για το $G_2$, εκμεταλλευόμενοι τη διαχωρισιμότητα.

| Βάση | Παρεμβολή |
|---|---|
| $H_{2a} = N(x^3 - 2.254x)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = \alpha^3$ |
| $H_{2b} = Ny(x^2 - 0.751333)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 3\alpha^2\beta$ |
| $H_{2c} = Nx(y^2 - 0.751333)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 3\alpha\beta^2$ |
| $H_{2d} = N(y^3 - 2.254y)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = \beta^3$ |
| $H_{2e} = Nz(x^2 - 0.751333)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 3\alpha^2\gamma$ |
| $H_{2f} = Nxyze^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 6\alpha\beta\gamma$ |
| $H_{2g} = Nz(y^2 - 0.751333)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 3\beta^2\gamma$ |
| $H_{2h} = Nx(z^2 - 0.751333)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 3\alpha\gamma^2$ |
| $H_{2i} = Ny(z^2 - 0.751333)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = 3\beta\gamma^2$ |
| $H_{2j} = N(z^3 - 2.254)e^{-(x^2+y^2+z^2)}$ | $k(\alpha,\beta,\gamma) = \gamma^3$ |

Πίνακας 4: Συναρτήσεις βάσης και συναρτήσεις παρεμβολής για το M/Σ Hilbert της δεύτερης παραγώγου της 3-διάστατης Γκαουσιανής. Η $N$ είναι σταθερά κανονικοποίησης που ισούται με 0.877776 ώστε το ολοκλήρωμα του τετραγώνου της συνάρτησης να ισούται με τη μονάδα. Ένα φίλτρο με άξονα συμμετρίας στη διεύθυνση $\Omega = (\alpha, \beta, \gamma)$ (τα αντίστοιχα διευθύνοντα συνημίτονα) δημιουργείται ως $H_2^\Omega = \sum_{i\in\{a,\cdots,j\}} k_i(\Omega)H_{2i}$.

| | Μονοδιάστατη Συνάρτηση | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| $f_1$ | $N(t^3 - 2.254t)e^{-t^2}$ | 0.0000 | -0.6776 | -0.0895 | 0.0554 | 0.0088 |
| $f_2$ | $N(t^2 - 0.751333t)e^{-t^2}$ | -0.6595 | -0.1695 | 0.1522 | 0.0508 | 0.0043 |
| $f_3$ | $e^{-t^2}$ | 1.0000 | 0.6383 | 0.1660 | 0.0176 | 0.0008 |
| $f_4$ | $Nte^{-t^2}$ | 0.0000 | 0.3754 | 0.1953 | 0.0310 | 0.0018 |
| $f_5$ | $te^{-t^2}$ | 0.0000 | 0.4277 | 0.2225 | 0.0354 | 0.0020 |

Πίνακας 5: Διακριτά φίλτρα 5 σημείων για το σύνολο συναρτήσεων βάσης για το $H_2$. Τα $f_2$, $f_3$ έχουν άρτια συμμετρία, τα $f_1$, $f_4$, $f_5$ περιττή. Τα φίλτρα αυτά είναι από τον Πίνακα 4 με δειγματοληψία ανά 0.67.

| Φίλτρο Βάσης | Φίλτρο στον $x$ | Φίλτρο στον $y$ | Φίλτρο στον $z$ |
|---|---|---|---|
| $H_{2a}$ | $f_1$ | $f_3$ | $f_3$ |
| $H_{2b}$ | $f_2$ | $f_5$ | $f_3$ |
| $H_{2c}$ | $f_5$ | $f_2$ | $f_3$ |
| $H_{2d}$ | $f_3$ | $f_1$ | $f_3$ |
| $H_{2e}$ | $f_2$ | $f_3$ | $f_5$ |
| $H_{2f}$ | $f_4$ | $f_5$ | $f_5$ |
| $H_{2g}$ | $f_3$ | $f_2$ | $f_5$ |
| $H_{2h}$ | $f_5$ | $f_3$ | $f_2$ |
| $H_{2i}$ | $f_3$ | $f_5$ | $f_2$ |
| $H_{2j}$ | $f_3$ | $f_3$ | $f_1$ |

Πίνακας 6: Κατασκευή των τρισδιάστατων φίλτρων βάσης για το $H_2$, εκμεταλλευόμενοι τη διαχωρισιμότητα.

Στο [3] προτείνεται η χρήση μιας γωνίας $\theta \in \{0°, 45°, 90°, 135°\}$ για τον άξονα συμμετρίας

του κατευθυντικού φίλτρου στον δισδιάστατο χώρο της εικόνας. Επεκτείνοντας την ιδέα αυτή στον τρισδιάστατο χώρο, εξετάζουμε κατευθύνσεις με πολικές και αζιμουθιακές γωνίες που μεταβάλλονται ανά 45°. Έτσι, έχουμε για την πολική και την αζιμουθιακή γωνία ($\theta$ και $\phi$ αντίστοιχα):

$$\theta \in \{0°, 45°, 90°, 135°\} \tag{16}$$

$$\phi \in \{0°, 45°, 90°, 135°\} \tag{17}$$

Δημιουργούνται, οπότε, 16 δυνατές κατευθύνσεις του άξονα συμμετρίας των φίλτρων ($|\Omega| = 16$). Τονίζουμε πως δε μας ενδιαφέρει η φορά, παρά μόνο η διεύθυνση του άξονα. Οπότε, δεν περιλαμβάνουμε, π.χ., την περίπτωση $\phi = 270°$, καθώς είναι πανομοιότυπη με την περίπτωση όπου $\phi = 90°$, η οποία έχει περιληφθεί.

Από τη σχέση (8), βέβαια, καθώς και όλη την ανάλυση που έχει ακολουθήσει, φαίνεται ότι χρειαζόμαστε τα συνημίτονα διεύθυνσης. Η μετάβαση είναι πολύ απλή, καθώς πρόκειται απλά για μετασχηματισμό σφαιρικών σε καρτεσιανές συντεταγμένες, εάν θεωρήσουμε τα συνημίτονα διεύθυνσης ως τις συντεταγμένες στο χώρο ενός διανύσματος μέτρου $r = 1$. Οπότε, έχουμε

$$\alpha = \sin(\theta)\cos(\phi) \tag{18}$$

$$\beta = \sin(\theta)\sin(\phi) \tag{19}$$

$$\gamma = \cos(\theta) \tag{20}$$

Για να γίνει καλύτερα αντιληπτή η λειτουργία των υπό χρήση στρεφόμενων κατευθυντικών φίλτρων, παρατίθενται χαρακτηριστικά αποτελέσματα στο Σχήμα 9.

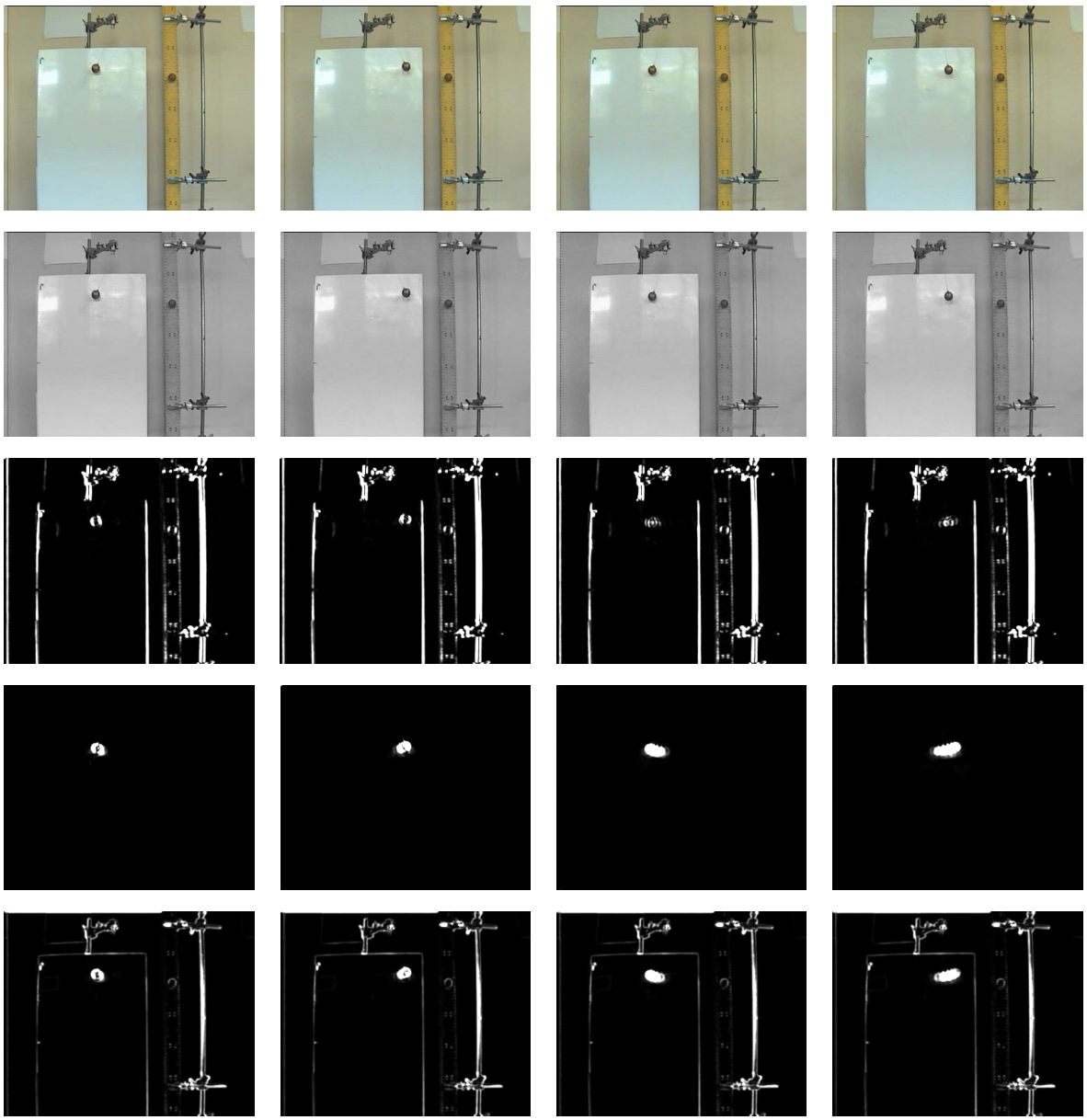
## 3.2 Αποσύνθεση Χαρακτηριστικών σε Κλίμακες

Όπως έχει αναφερθεί στην Ενότητα 2.1, αφού εξαχθούν τα διάφορα χαρακτηριστικά, αυτά εξετάζονται σε ένα σύνολο διαφορετικών κλιμάκων, γεγονός που οδηγεί στη δημιουργία διαδοχικών υποχαρτών (τρισδιάστατων εάν υπεισέρχεται και ο χρόνος) για κάθε χαρακτηριστικό. Η ιδέα είναι ο εντοπισμός ενός προτύπου ενδιαφέροντος που μπορεί να εμφανιστεί σε οποιαδήποτε κλίμακα. Σύμφωνα με την ιδέα αυτή, καθένα από τα τρία δημιουργούμενα conspicuity volumes αποσυντίθεται σε ένα συγκεκριμένο αριθμό χωροχρονικών κλιμάκων. Ένας τρόπος υλοποίησης της διαδοχικής αυτής κλιμάκωσης είναι η χρήση τρισδιάστατων Γκαουσιανών πυραμίδων.

Ας εξετάσουμε πρώτα τη δισδιάστατη περίπτωση. Μία πυραμίδα εικόνων είναι μια δομή αποτελούμενη από μια ακολουθία αντιγράφων της αρχικής εικόνας, όπου τόσο η συχνότητα δειγματοληψίας, όσο και η ανάλυση ελαττώνονται διαδοχικά [16]. Τα διαδοχικά αυτά επίπεδα εξάγονται μέσω μιας επαναληπτικής διαδικασίας. Το μηδενικό επίπεδο, έστω $G_0$, ταυτίζεται με την αρχική εικόνα. Αυτό φιλτράρεται από ένα βαθυπερατό φίλτρο και υποδειγματοληπτείται κατά έναν συντελεστή 2, ώστε να ληφθεί το επόμενο επίπεδο της πυραμίδας $G_1$. Το $G_1$ φιλτράρεται και δειγματοληπτείται κατά όμοιο τρόπο για να ληφθεί το $G_2$. Η διαδικασία επαναλαμβάνεται για τα $N$ επιθυμητά επίπεδα. Οπότε, συμβολίζοντας με $w$ τη μάσκα φιλτραρίσματος έχουμε:

$$G_l(i,j) = \sum_m \sum_n w(m,n) G_{l-1}(2i+m, 2j+n), \quad 0 < l < N \tag{21}$$

Το βαθυπερατό φιλτράρισμα είναι απαραίτητο για να αποφευχθεί το aliasing. Για παράδειγμα, στο Σχήμα 10 φαίνεται καθαρά το φαινόμενο του aliasing κατά την απλή υποδειγματοληψία και η αντιμετώπισή του εάν έχει προηγηθεί γκαουσιανό φιλτράρισμα.

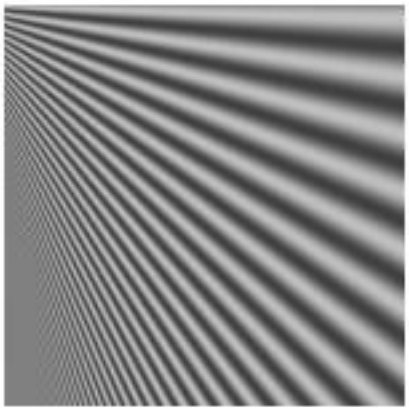


Σχήμα 9: Αποτελέσματα των υπό χρήση στρεφόμενων κατευθυντικών φίλτρων. Ως είσοδος δίνεται ένα βίντεο διάρκειας 180 frames ενός εκκρεμούς σε χώρο εργαστηρίου. Κάθε στήλη αντιστοιχεί σε ένα συγκεκριμένο frame. 1η γραμμή: αρχικό βίντεο. 2η γραμμή: ένταση $I$. 3η γραμμή: άθροισμα των τετραγωνικών αποκρίσεων των $G_2$ και $H_2$ για $\phi = 0°$, $\theta = 90°$. 4η γραμμή: άθροισμα των τετραγωνικών αποκρίσεων των $G_2$ και $H_2$ για $\phi = 0°$, $\theta = 0°$. 5η γραμμή: frames από το τελικό conspicuity volume για τον προσανατολισμό.

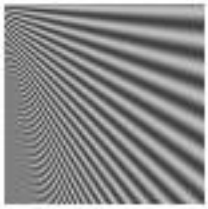
Η επέκταση στις 3 διαστάσεις είναι άμεση, εισάγοντας έναν τρισδιάστατο πυρήνα φιλτραρίσμα-

τος:

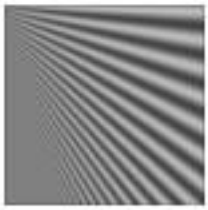
$$G_l(i,j,k) = \sum_m \sum_n \sum_p w(m,n,p) G_{l-1}(2i+m, 2j+n, 2k+p), \quad 0 < l < N \tag{22}$$



(α′) Αρχική εικόνα.



(β′) Υποδειγματοληψία χωρίς φιλτράρισμα.



(γ′) Βαθυπερατό φιλτράρισμα και υποδειγματοληψία.

Σχήμα 10: Υποδειγματοληψία εικόνας όταν έχει προηγηθεί βαθυπερατό φιλτράρισμα και όταν όχι. Η υποδειγματοληψία έγινε κατά έναν παράγοντα 2. Για το φιλτράρισμα χρησιμοποιήθηκε Γκαουσιανός πυρήνας διάστασης 10 και τυπικής απόκλισης 1. Το aliasing στο Σχήμα (β′) είναι εμφανές.

Το ερώτημα, τώρα, είναι πώς θα γίνει η επιλογή του πυρήνα $w$. Καταρχήν, επιλέγεται πυρήνας μεγέθους $5 \times 5 \times 5$, εφόσον παρέχει επαρκές φιλτράρισμα με χαμηλό υπολογιστικό κόστος [17]. Για περαιτέρω ταχύτητα υπολογισμών, είναι επιθυμητό ο πυρήνας να είναι διαχωρίσιμος, δηλαδή

$$w(m,n,p) = \hat{w}(m)\hat{w}(n)\hat{w}(p) \tag{23}$$

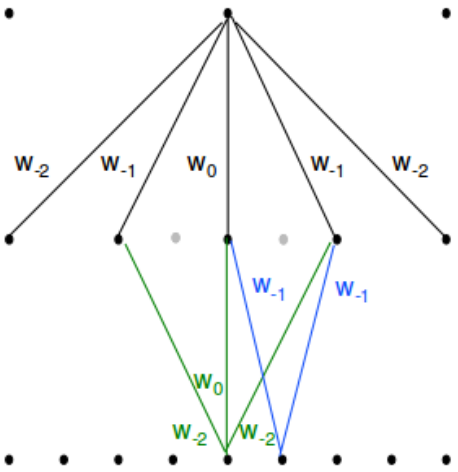
Ο μονοδιάστατος, μήκους 5, πυρήνας $\hat{w}$ είναι κανονικοποιημένος και συμμετρικός, δηλαδή

$$\sum_{m=-2}^{2} \hat{w}(m) = 1 \tag{24}$$

$$\hat{w}(i) = \hat{w}(-i), \quad i = 0, 1, 2 \tag{25}$$

Ένας συμπληρωματικός περιορισμός που τίθεται καλείται ''κοινή συνεισφορά''. Σύμφωνα με αυτό, όλοι οι κόμβοι (δηλαδή τα voxels) ενός επιπέδου πρέπει να συνεισφέρουν κατά το ίδιο ολικό βάρος $(=\frac{1}{8})$ στο αμέσως επόμενο επίπεδο. Βλέπουμε στο Σχήμα 11 πως ο κάποιοι κόμβοι συνεισφέρουν κατά $2b$ και κάποιοι κατά $a + 2c$ στο επόμενο επίπεδο, εάν θέσουμε $\hat{w} = [c, b, a, b, c]$. Οπότε, καταλήγουμε στο σύστημα:

$$\begin{cases} a + 2b + 2c = 1 \\ a + 2c = 2b \end{cases} \tag{26}$$




Σχήμα 11: Μονοδιάστατη γραφική αναπαράσταση της δημιουργίας μιας Γκαουσιανής πυραμίδας. Κάθε τελεία παριστά έναν κόμβο σε ένα επίπεδο της πυραμίδας (ένα voxel). Η απόσταση μεταξύ των κόμβων διπλασιάζεται σε κάθε νέο επίπεδο.

Καταλήγουμε εν τέλει στην εξής απειρία λύσεων:

$$\hat{w} = \left[\frac{1}{4} - \frac{a}{2}, \frac{1}{4}, a, \frac{1}{4}, \frac{1}{4} - \frac{a}{2}\right] \tag{27}$$

Θέτοντας τον επιπλέον περιορισμό για μια μονοκόρυφη συνάρτηση φιλτραρίσματος, έχουμε $a > 1/4$. Για λόγους αριθμητικής απλότητας, επιλέγουμε την τιμή $a = 0.375$ (που είναι η default τιμή της MATLAB).

Για κάθε ένα εκ των τριών conpicuity volumes που έχουμε δημιουργήσει, επιλέγουμε να κατασκευάσουμε 5 κλίμακες. Έχουμε, λοιπόν, 15 τρισδιάστατες δομές $C_{i,l}$, $i = 1, 2, 3$, $j = 0, \cdots, 4$. Ωστόσο, θέλουμε να κάνουμε πράξεις και συγκρίσεις μεταξύ διαφορετικών κλιμάκων, γεγονός που σημαίνει πως όλες οι δομές πρέπει να έχουν το ίδιο μέγεθος. Αυτό μπορεί να επιτευχθεί με interpolation κάθε δομής μικρότερων διαστάσεων, ώστε να έχει διαστάσεις ίσες με αυτές της προς σύγκριση δομής.

Πρακτικά, κάτι τέτοιο αποτελεί ανούσια υπολογιστική σπατάλη. Αυτό που κάνουμε, λοιπόν, είναι απλά να φιλτράρουμε διαδοχικά με τον πυρήνα $w$ την τρισδιάστατη δομή, χωρίς να ακολουθεί το στάδιο της υποδειγματοληψίας (το οποίο έτσι και αλλιώς θα ακυρωνόταν από την υπερδειγματοληψία που θα ακολουθούσε). Τονίζουμε για ακόμα μία φορά ότι εκμεταλλευόμαστε τη διαχωρισιμότητα του $w$ και φιλτράρουμε σε κάθε διάσταση διαδοχικά με το μονοδιάστατο πυρήνα $\hat{w}$, ο οποίος κάθε φορά είναι ευθυγραμμισμένος με την επιθυμητή διάσταση.

Τέλος, συνδυάζουμε όλες τις τρισδιάστατες δομές που έχουν δημιουργηθεί σε μία μοναδική 5-διάστατη δομή $\mathbf{C}$. Το τελικό αποτέλεσμα για ένα τυχαίο frame ενός βίντεο παρουσιάζεται στο Σχήμα 12.

## 3.3 Διατύπωση της Ενέργειας

Μέχρι στιγμής, έχουν δημιουργηθεί conspicuity volumes τα οποία κωδικοποιούν το μέτρο προσοχής που περιέχεται σε κάθε voxel σύμφωνα μόνο με το αντίστοιχο χαρακτηριστικό του volume. Αυτές οι δομές πρέπει να αλληλεπιδρούν ώστε να παραχθεί ένα μοναδικό μέτρο προσοχής για κάθε voxel. Στη μέθοδο που ακολουθείται, τα διάφορα voxels ανταγωνίζονται σε τρία διαφορετικά επίπεδα. Υπάρχει, λοιπόν, ανταγωνισμός μεταξύ διαφορετικών χαρακτηριστικών, μεταξύ διαφορετικών κλιμάκων, αλλά και μεταξύ γειτόνων στην ίδια κλίμακα του ίδιου χαρακτηριστικού.

Οι διάφορες αυτές αλληλεπιδράσεις που εκφράζονται ως ανταγωνισμοί προσεγγίζονται μέσω ελαχιστοποίησης μιας ολικής ενέργειας $E$ που συντίθεται από έναν όρο δεδομένων παρατηρήσεων $E_d$ και έναν όρο εξομάλυνσης $E_s$.

$$E(\mathbf{C}) = \lambda_d E_d(\mathbf{C}) + \lambda_s E_s(\mathbf{C}) \tag{28}$$

(α′) Υπό μελέτη frame μαζί με το προηγούμενο και το επόμενό του.

(β′) Οι 5 κλίμακες του conspicuity volume της έντασης.

(γ′) Οι 5 κλίμακες του conspicuity volume του χρώματος.

(δ′) Οι 5 κλίμακες του conspicuity volume του προσανατολισμού.

Σχήμα 12: "Frames" που αντιστοιχούν στις διαφορετικές κλίμακες των διαφορετικών conspicuity volumes για ένα τυχαίο frame ενός βίντεο. Σε κάθε περίπτωση, στα (β′), (γ′), (δ′), οι εικόνες είναι κανονικοποιημένες στο εύρος της ελάχιστης και της μέγιστης τιμής των pixels.

Ο όρος δεδομένων διατηρεί μία σχέση μεταξύ της αρχικής και της τρέχουσας εκτίμησης, ώστε να αποφευχθεί μία υπερβολική εξομάλυνση. Ο όρος αυτός εκφράζεται ως

$$E_d(\mathbf{C}) = \sum_{i=1}^{3} \sum_{l=1}^{5} \sum_{q} (C_{i,l}(q) - C_{i,l}^0(q))^2 \tag{29}$$

όπου $q \in Q$ και $C_{i,l}^0(q)$ η αρχική εκτίμηση για το voxel $q$. Λέγοντας αρχική εκτίμηση, αναφερόμαστε ακριβώς στην τιμή που έχει προκύψει στη δομή $\mathbf{C}$ που έχουμε κατασκευάσει έως τώρα.

Ο όρος εξομάλυσνης αποτυπώνει τους περιορισμούς ως προς τις αλληλεπιδράσεις που δημιουργούνται στα τρία επίπεδα που αναφέραμε. Έχουμε, οπότε

$$E_s(\mathbf{C}) = E_{s,1}(\mathbf{C}) + E_{s,2}(\mathbf{C}) + E_{s,3}(\mathbf{C}) \tag{30}$$

όπου ο όρος $E_{s,1}$ μοντελοποιεί τις αλληλεπιδράσεις σε ένα χαρακτηριστικό και κλίμακα, ο $E_{s,2}$ μεταξύ διαφορετικών χαρακτηριστικών και ο $E_{s,3}$ μεταξύ διαφορετικών κλιμάκων.

Ο όρος $E_{s,1}$ αποτυπώνει τη συνεκτικότητα εντός ενός χαρακτηριστικού, δηλαδή ορίζει την αλληλεπίδραση μεταξύ γειτονικών voxels του ίδιου χαρακτηριστικού και στην ίδια κλίμακα. Σκοπός του είναι η δημιουργία μικρών χωροχρονικών συστάδων αποτελούμενων από voxels με παραπλήσιες τιμές και παράλληλα η ενίσχυση των voxels εκείνων που ξεχωρίζουν και δεν είναι συνεπείς με τη γειτονιά τους. Ορίζεται ως εξής:

$$E_{s,1}(\mathbf{C}) = \sum_{i=1}^{3}\sum_{l=1}^{5}\sum_{q}\left(C_{i,l}(q) - \frac{1}{|N_q|}\sum_{r\in N_q} C_{i,l}(r)\right)^2 \tag{31}$$

Ο όρος $E_{s,2}$ επιτρέπει την αλληλεπίδραση μεταξύ διαφορετικών χαρακτηριστικών, ώστε voxels που σημαίνονται ως ''σημαντικά'' ως προς τα διάφορα χαρακτηριστικά να ομαδοποιούνται μαζί και να σχηματίζουν συμπαγείς περιοχές. Σχετίζεται με τον ανταγωνισμό ανάμεσα σε ένα voxel ενός feature volume και τα αντίστοιχα voxels σε όλα τα υπόλοιπα feature volumes. Ορίζεται ως εξής:

$$E_{s,2}(\mathbf{C}) = \sum_{i=1}^{3}\sum_{l=1}^{5}\sum_{q}\left(C_{i,l}(q) - \frac{1}{2}\sum_{j\neq i} C_{j,l}(q)\right)^2 \tag{32}$$

Ο όρος $E_{s,3}$ προσπαθεί να ενισχύσει τα voxels εκείνα που έχουν μεγάλο μέτρο προσοχής σε διαφορετικές κλίμακες των πυραμίδων που έχουν δημιουργηθεί. Έτσι, τα voxels που έχουν μεγάλες τιμές σε όλες τις κλίμακες είναι εκείνα που τελικά αναδεικνύονται ως σημαντικά. Ο όρος αυτός ορίζεται ως εξής:

$$E_{s,3}(\mathbf{C}) = \sum_{i=1}^{3}\sum_{l=1}^{5}\sum_{q}\left(C_{i,l}(q) - \frac{1}{4}\sum_{n\neq l} C_{i,n}(q)\right)^2 \tag{33}$$

Προφανώς, όσο μεγαλύτερη είναι η τιμή του συντελεστή $\lambda_s$, τόσο μεγαλύτερη είναι η εξομάλυνση, με κόστος τη χαμηλότερη επίδραση της αρχικής εκτίμησης, εφόσον υπάρχει ο περιορισμός

$$\lambda_d + \lambda_s = 1 \tag{34}$$

Πρέπει, επομένως, να βρεθεί ένας συμβιβασμός. Στην υλοποίησή μας, επιλέξαμε εν τέλει $\lambda_d = 0.8$ και $\lambda_s = 0.2$, δίνοντας ''προβάδισμα'' στον όρο δεδομένων.

## 3.4 Ελαχιστοποίηση της Ενέργειας

Σειρά, τώρα, έχει η ελαχιστοποίηση της ολικής ενέργειας $E$, που δίνεται από τη σχέση (28). Προς την κατεύθυνση αυτή, υιοθετούμε την προσέγγιση ενός αλγορίθμου gradient descent, όπου επαναληπτικά η τιμή του κάθε στοιχείου της δομής $\mathbf{C}$ ανανεώνεται ώστε να οδηγηθεί τελικά όλο το σύστημα προς το ελάχιστο στο χώρο ενέργειας. Συμβολίζοντας με $\gamma$ το ρυθμό μάθησης και $\mu$ έναν όρο ορμής ώστε να εξασφαλιστεί η ευστάθεια του αλγορίθμου, έχουμε σε κάθε επανάληψη $\tau$:

$$C_{i,l}^{\tau}(q) = C_{i,l}^{\tau-1}(q) + \Delta C_{i,l}^{\tau-1}(q) \tag{35}$$

$$\Delta C_{i,l}^{\tau-1}(q) = -\gamma\frac{\partial E(\mathbf{C}^{\tau-1})}{\partial C_{i,l}^{\tau-1}(q)} + \mu\Delta C_{i,l}^{\tau-2}(q) \tag{36}$$

Η μερική παράγωγος της σχέσης (36) υπολογίζεται ως εξής:

$$\frac{\partial E(\mathbf{C})}{\partial C_{k,m}(s)} = \lambda_d \frac{\partial E_d(\mathbf{C})}{\partial C_{k,m}(s)} + \lambda_s \sum_{c=1}^{3} \frac{\partial E_{s,c}(\mathbf{C})}{\partial C_{k,m}(s)} \quad (37)$$

Σύμφωνα με το [18], οι τέσσερις μερικές παράγωγοι που εμφανίζονται στην εξίσωση (37) υπολογίζονται ως εξής:

$$\frac{\partial E_d(\mathbf{C})}{\partial C_{k,m}(s)} = 2(C_{k,m}(s) - C^0_{k,m}(s)) \quad (38)$$

$$\frac{\partial E_{s,1}(\mathbf{C})}{\partial C_{k,m}(s)} = 2\left\{ C_{k,m}(s) - \frac{1}{|N_q|^2} \sum_{q \in N_s} \left( 2|N_s| C_{k,m}(q) - \sum_{r \in N_q} C_{i,l}(r) \right) \right\} \quad (39)$$

$$\frac{\partial E_{s,2}(\mathbf{C})}{\partial C_{k,m}(s)} = 2\left( C_{k,m}(s) - \frac{1}{2} \sum_{j \neq k} C_{j,m}(s) \right) \quad (40)$$

$$\frac{\partial E_{s,3}(\mathbf{C})}{\partial C_{k,m}(s)} = \frac{5}{2}\left( C_{k,m}(s) - \frac{1}{4} \sum_{n \neq l} C_{k,n}(s) \right) \quad (41)$$

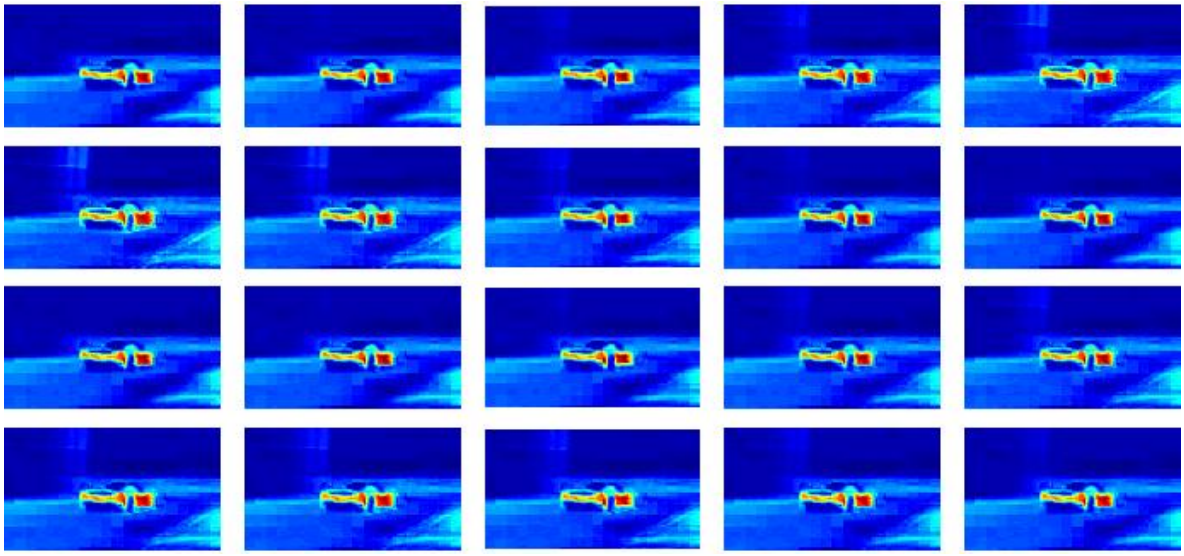
Προφανώς, οι σχέσεις (35) - (41) υπολογίζονται μαζικά για όλα τα voxels με πράξεις μεταξύ των αντίστοιχων πινάκων.

Η επίδραση και η σημασία της ενεργειακής προσέγγισης παρουσιάζεται στο Σχήμα 13. Παρατηρούμε, εκεί, πως όσο αυξάνονται οι επαναλήψεις, επιχειρείται μία εξομάλυνση του ''χάρτη'' προσοχής, υπό την έννοια ότι περιοχές με παραπλήσιες τιμές προσοχής τείνουν να ομαδοποιούνται μαζί. Αξίζει να παρατηρήσουμε, ακόμα, πως από την αρχική εκτίμηση των χαρακτηριστικών, η ύπαρξη του γκρι κτιρίου δε γίνεται εμφανής στο conspicuity volume του χρώματος (το οποίο και παρουσιάζεται στο συγκεκριμένο Σχήμα), γεγονός αναμενόμενο, εφόσον από καθαρά χρωματική άποψη, το κτίριο και ο ουρανός, όπως φαίνονται στο frame του αρχικού βίντεο, δε διαφέρουν ιδιαίτερα. Ωστόσο, λόγω της αλληλεπίδρασης με το conspicuity volume προσανατολισμού, το κτίριο εμφαίζεται από τις πρώτες κιόλας επαναλήψεις. Ωστόσο, όσο προχωράει η διαδικασία της εξομάλυνσης, αυτό και πάλι ''εξαφανίζεται'', στα πλαίσια της καταστολής της επίδρασης του υποβάθρου και της ανάδειξης των σημείων ενδιαφέροντος. Σε κάθε περίπτωση, είναι εμφανές πως το κεντρικό σημείο ενδιαφέροντος είναι το αυτοκίνητο.

Επανερχόμενοι στην επιρροή που έχει το conspicuity volume προσανατολισμού σε αυτό του χρώματος, είναι αναμενόμενο το κτίριο να πέρνει αρκετά μεγάλες τιμές για δύο λόγους. Πρώτον, παρουσιάζει μία αρκετά έντονη κάθετη ακμή, γεγονός που θα οδηγούσε σε μεγάλη απόκριση ενός χωρικού κατευθυντικού φίλτρου και επίσης κινείται, όπως φαίνεται στα διαδοχικά frames, γεγονός που θα οδηγούσε σε μεγάλη απόκριση ενός χρονικού κατευθυντικού φίλτρου. Συνεπώς, τα χωροχρονικά φίλτρα που έχουν χρησιμοποιηθεί προφανώς θα έχουν μια σχετικά μεγάλη απόκριση. Πράγματι, και εκ του αποτελέσματος (Σχήμα 14), επιβεβαιώνονται τα παραπάνω. Ως κύριο σημείο ενδιαφέροντος, βέβαια, αναδεικνύεται και πάλι το αυτοκίνητο.

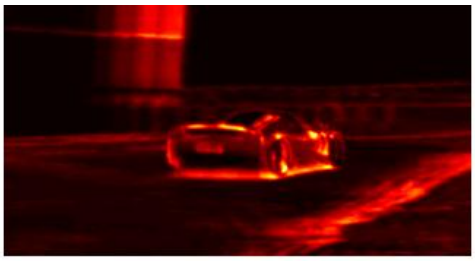
[2]Προσοχή, το γράμμα $s$ στο $E_s$ δηλώνει απλά τον όρο smoothing και δε σχετίζεται με τη μεταβλητή $s$ που δηλώνει το voxel.

(α′) Υπό μελέτη frame μαζί με το προηγούμενο και το επόμενό του.



(β′) Η τομή του $C_{2,2}$, δηλαδή της δεύτερης κλίμακας του χρώματος που αντιστοιχεί στο υπό μελέτη frame για διαδοχικές επαναλήψεις του αλγορίθμου ελαχιστοποίησης ενέργειας. Η 1η εικόνα αποτελεί την αρχική εκτίμηση, η 2η αντιστοιχεί σε μία επανάληψη, ..., η τελευταία σε 19 επαναλήψεις. Κάθε σειρά εικόνων αποτελεί συνέχεια της προηγούμενης. Οι εικόνες είναι grayscale, αλλά αποτυπώνονται με ψευδοχρώματα, ώστε το κόκκινο να αντιστοιχεί σε μεγάλη τιμή, ενώ το μπλε σε μικρή.

Σχήμα 13: Επίδραση της ελαχιστοποίησης ενέργειας σε διαδοχικές επαναλήψεις για ένα τυχαίο frame ενός βίντεο.



Σχήμα 14: Η τομή του $C_{3,2}$, δηλαδή της δεύτερης κλίμακας του προσανατολισμού που αντιστοιχεί στο υπό μελέτη frame (Σχήμα 13) για τη 19η επανάληψη του αλγορίθμου ελαχιστοποίησης ενέργειας. Η εικόνα είναι grayscale, αλλά αποτυπώνεται με ψευδοχρώματα στην κλίμακα του κόκκινου.

Ο αλγόριθμος επαναλαμβάνεται μέχρις ότου φτάσουμε σε μία πολύ μικρή απόσταση από το επιθυμητό ελάχιστο. Φορμαλιστικά, το κριτήριο σύγκλισης ορίζεται ως

$$\max_{q} |\Delta C_{i,l}^{\tau-1}(q)| < \epsilon \tag{42}$$

όπου $\epsilon$ μια μικρή σταθερά που έχει οριστεί εκ των προτέρων. Πρακτικά, παρατηρήθηκε ότι στα παραδείγματα που δοκιμάστηκαν, λίγες επαναλήψεις ήταν αρκετές για την εξαγωγή ικανοποιητι-

κών αποτελεσμάτων. Για το λόγο αυτό, ακολουθήσαμε την μία προσέγγιση σταθερού αριθμού επαναλήψεων για κάθε περίπτωση, ώστε να αποφύγουμε και πολύ μεγάλους χρόνους εκτέλεσης. Συγκεκριμένα, επιλέχθηκαν 10 επαναλήψεις του αλγορίθμου για κάθε block από frames που εξετάζονται.

## 3.5 Εξαγωγή Τελικού Saliency

Αφότου έχουμε εξαγάγει την τελική δομή **C**, μετά και την ελαχιστοποίηση της ενέργειας, ακολουθεί η συγχώνευση των διαφορετικών χαρακτηριστικών και κλιμάκων σε ένα ενιαίο saliency για το κάθε voxel. Προκύπτει, λοιπόν, μία νέα τρισδιάστατη δομή $S$ που ορίζεται ως ο μέσος όρος των τριών διαφορετικών χαρακτηριστικών και των πέντε διαφορετικών κλιμάκων:

$$S = \frac{1}{15} \sum_{i=1}^{3} \sum_{l=1}^{5} C_{i,l} \tag{43}$$

Στο Σχήμα 15 παρουσιάζονται οι 5 κλίμακες για τα 3 χαρακτηριστικά, καθώς και το τελικό saliency για ένα frame της ταινίας 300: Rise of an Empire.

Ωστόσο, τελικός μας στόχος, είναι η εξαγωγή της περίληψης για κάποιο βίντεο. Συνεπώς, θα πρέπει να αποδωθεί μία μοναδική τιμή ποσοτικοποίησης της προσοχής και του ενδιαφέροντος ανά frame και όχι ανά voxel, όπως έχει γίνει μέχρι τώρα. Για το σκοπό αυτό, χρησιμοποιούμε την πρώτη κλίμακα από το κάθε χαρακτηριστικό $C_{i,1}$, την οποία και κανονικοποιούμε στο διάστημα [0,1]. Εν συνεχεία, πολλαπλασιάζεται η δομή αυτή κατά σημείο με τη δομή του saliency $S$ και αθροίζουμε για όλα τα pixels που ανήκουν στο συγκεκριμένο frame. Έτσι, εάν συμβολίσουμε ένα voxel ως $q = (x, y, t)$, έχουμε:

$$S_{intensity}(t) = S_i(t) = \sum_x \sum_y S(x, y, t) C_{1,1}(x, y, t) \tag{44}$$

$$S_{color}(t) = S_c(t) = \sum_x \sum_y S(x, y, t) C_{2,1}(x, y, t) \tag{45}$$

$$S_{orientation}(t) = S_o(t) = \sum_x \sum_y S(x, y, t) C_{3,1}(x, y, t) \tag{46}$$

Έχουμε, λοιπόν, τρεις τιμές ανά frame, που αντιστοιχούν στα τρία διαφορετικά χαρακτηριστικά, δηλαδή την ένταση, το χρώμα και τον προσανατολισμό. Για τη συγχώνευση των τιμών αυτών, ακολουθούμε κάποιες διαφορετικές εναλλακτικές προσεγγίσεις.

- *Γραμμική Συγχώνευση.* Η πιο απλή προσέγγιση του προβλήματος είναι μία γραμμική συσχέτιση των τριών τιμών. Θεωρώντας, λοιπόν, τα αντίστοιχα βάρη $w_{intensity}$, $w_{color}$ και $w_{orientation}$, έχουμε:

  $$S_{lin} = w_i S_i + w_c S_c + w_o S_o \tag{47}$$

  Στη γενική περίπτωση, τα βάρη μπορεί να είναι άνισα, χρονικά μεταβαλλόμενα, προσαρμοζόμενα, εξαρτώμενα από a priori γνώση ή άλλους περιορισμούς, κ.λπ. Στα πλαίσια της εργασίας, θεωρούμε ίσοδύναμη εξάρτηση από τις τρεις συνιστώσες και θέτουμε

  $$w_i = w_c = w_o = \frac{1}{3} \tag{48}$$

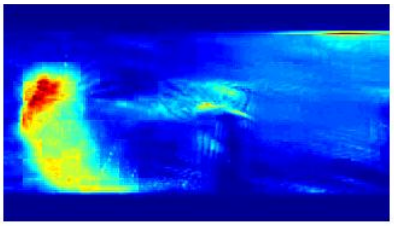
αν και η άθροιση των βαρών στη μονάδα δεν είναι απαραίτητη προϋπόθεση. Ισοδύναμα, δηλαδή, θα μπορούσαμε να τα θέσουμε ίσα με τη μονάδα.

(α′) Υπό μελέτη frame μαζί με το προηγούμενο και το επόμενό του.

(β′) Οι 5 κλίμακες του conspicuity volume της έντασης.

(γ′) Οι 5 κλίμακες του conspicuity volume του χρώματος.

(δ′) Οι 5 κλίμακες του conspicuity volume του προσανατολισμού.



(ε′) Τελικό saliency για το frame αφού έχουν συνδυαστεί όλα τα παραπάνω.

Σχήμα 15: "Frames" που αντιστοιχούν στις διαφορετικές κλίμακες των διαφορετικών conspicuity volumes για ένα τυχαίο frame ταινίας και τελικό saliency ψευδοχρωματισμένο.

- *Συγχώνευση Βασισμένη στη Διασπορά*. Με την προσέγγιση αυτή, σε κάθε χαρακτηριστικό προσδίδεται βάρος αντιστρόφως ανάλογο προς την αβεβαιότητά του, όπου η αβεβαιότητα προσεγγίζεται με τη διασπορά. Οπότε:

$$S_{var} = \frac{1}{var(S_i)} S_i + \frac{1}{var(S_c)} S_c + \frac{1}{var(S_o)} S_o \tag{49}$$

- *Μη-Γραμμική Συγχώνευση με Λογική Μεγίστου*. Με τη λογική αυτή, σε κάθε frame αποδίδεται τελικά η μέγιστη εκ των τριών τιμών που του έχουν αποδωθεί έως τώρα. Δηλαδή

$$S_{max} = \max\{S_i, S_c, S_o\} \tag{50}$$

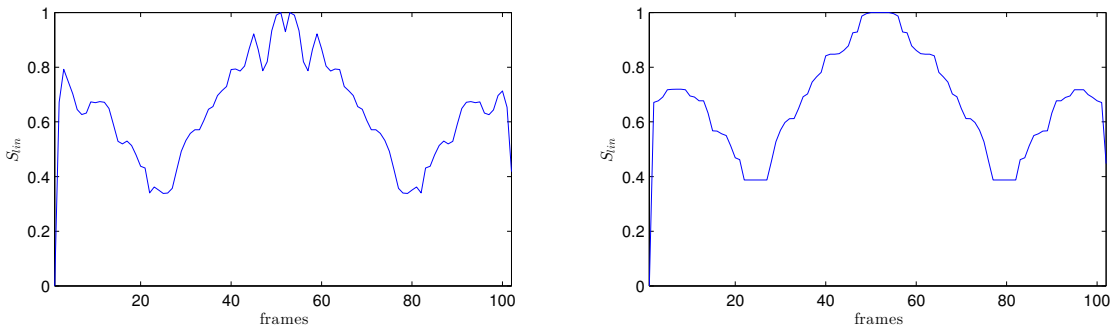
όπου, προφανώς, το max δηλώνει κατά σημείο μέγιστο.

- *Μη-Γραμμική Συγχώνευση με Λογική Ελαχίστου.* Όμοια με τη λογική μεγίστου, μπορούμε να αποδώσουμε σε κάθε frame την ελάχιστη εκ των τριών τιμών που του έχουν αποδωθεί έως τώρα. Δηλαδή

$$S_{min} = \min\{S_i, S_c, S_o\} \tag{51}$$

όπου το min δηλώνει κατά σημείο ελάχιστο.

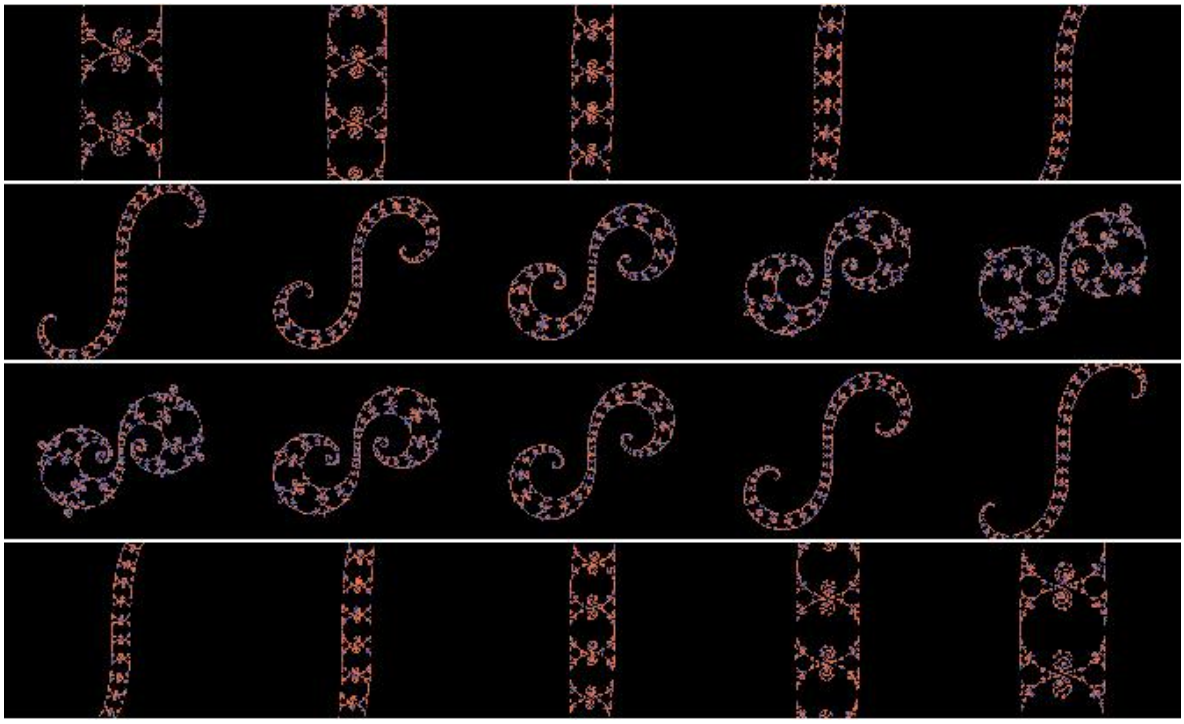
Ακολουθώντας έναν εκ των παραπάνω τρόπων, δημιουργείται ουσιαστικά μία καμπύλη ''σημαντικότητας'' συναρτήσει του χρόνου, την οποία θα ονομάζουμε saliency curve και θα συμβολίζουμε στη γενική περίπτωση ως $S_{total}$. Στο χρονικό σήμα που ορίζει την εκάστοτε saliency curve εφαρμόζεται ένα μονοδιάστατο φίλτρο μέσης τιμής, ώστε να εξαχθεί μία ομαλή καμπύλη, εφόσον δεν θέλουμε εν τέλει να κρατήσουμε μεμονωμένα frames-κλειδιά, αλλά ολόκληρα τμήματα διαδοχικών frames. Ένα παράδειγμα δίνεται στο Σχήμα 17, ενώ η χρησιμότητα του φιλτραρίσματος μέσης τιμής φαίνεται καθαρά στο Σχήμα 16.



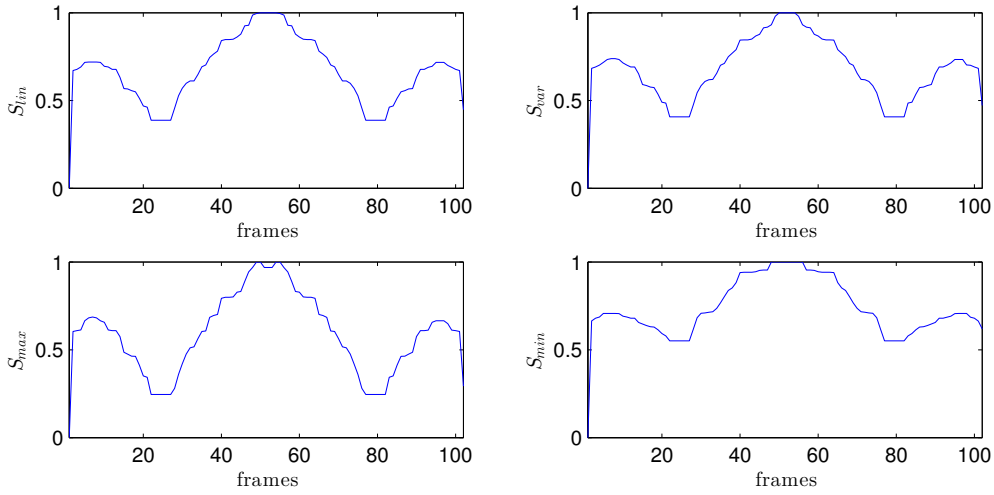

(α′) Χωρίς χρήση φίλτρου μέσης τιμής. (β′) Με χρήση φίλτρου μέσης τιμής.

Σχήμα 16: Saliency curves που προκύπτουν όταν χρησιμοποιείται φίλτρο μέσης τιμής και όταν όχι. Οι καμπύλες προέκυψαν από επεξεργασία του βίντεο του Σχήματος 17 με γραμμική συγχώνευση των χαρακτηριστικών. Χρησιμοποιήθηκε φίλτρο μήκους 11.

Παρατηρώντας το Σχήμα 17, βλέπουμε πως το γενικό pattern που δημιουργείται είναι παρόμοιο στις τέσσερις περιπτώσεις, υπάρχουν ωστόσο διαφορές. Οι διαφορές αυτές μπορεί να φαίνονται ασήμαντες στο συγκεκριμένο απλό παράδειγμα, όπου ως είσοδος δόθηκε ένα μικρής διάρκειας και συμμετρικό βίντεο, αλλά στην πράξη οδηγούν σε διαφορετικές περιλήψεις, όπως θα δούμε και στη συνέχεια. Αξίζει να σημειωθεί πως δεν πρέπει να δημιουργηθεί σύγχυση όσον αφορά στη σύγκριση μεταξύ $S_max$ και $S_min$, επειδή φαινομενικά η καμπύλη $S_min$ φαίνεται να περνά από πιο μεγάλες τιμές. Το γεγονός αυτό οφείλεται στην κανονικοποίηση που έχει προηγηθεί πριν την αναπαράσταση των καμπυλών. Στην πραγματικότητα, η απόλυτη τιμή της κορυφής της καμπύλης $S_min$ είναι αρκετά μικρότερη από αυτή της καμπύλης $S_max$ (ενώ στο Σχήμα παρουσιάζονται και οι δύο στη μονάδα).



(α΄) Διαδοχικά frames με βήμα 5 ενός βίντεο συνολικής διάρκειας 102 frames.




(β΄) Παραγόμενες καμπύλες σύμφωνα με τους 4 διαφορετικούς τρόπους που συζητήθηκαν.

Σχήμα 17: Saliency curves που παράγονται με χρήση τεσσάρων διαφορετικών τρόπων συγχώνευσης των τριών χαρακτηριστικών. Πριν την απεικόνση των αποτελεσμάτων, έχει προηγηθεί κανονικοποίηση στο διάστημα [0,1].

## 3.6 Δημιουργία της Περίληψης

Οι περιλήψεις δημιουργούνται, έχοντας καθορίσει εκ των προτέρων ένα ποσοστό συρρίκνωσης του βίντεο, υπό την έννοια της ελάττωσης της χρονικής διάρκειας. Έτσι, συμβολίζοντας το ποσοστό αυτό $c$, επιλέγεται ένα κατώφλι στο saliency, έστω $T_c$, ώστε να επιτευχθεί το επιθυμητό

αποτέλεσμα. Όλα τα frames $t$ με μέτρο ενδιαφέροντος $S_{total}(t) > T_c$ επιλέγονται (σαν μια πρώτη επιλογή) να περιληφθούν στην περίληψη. Για παράδειγμα, εάν επιθυμείται μια περίληψη 20%, δηλαδή $c = 0.2$, το $T_c$ επιλέγεται έτσι ώστε η πληθικότητα του συνόλου $D = \{t : S_{total}(t) > T_c\}$ να είναι 20% του συνόλου των frames του βίντεο.

Πρακτικά, έστω ότι έχουμε $M$ συνολικά frames και θέλουμε ένα ποσοστό συρρίκνωσης $c$. Θέλουμε, δηλαδή, μία περίληψη με $Mc$ frames. Το $S_{total}$ δεν είναι παρά ένα διάνυσμα $M$ στοιχείων. Βρίκουμε, οπότε, τους δείκτες των $Mc$ μέγιστων στοιχείων του $S_{total}$ και δημιουργούμε ένα νέο διάνυσμα, έστω $I$, που αντιστοιχεί στη (διακριτή) συνάρτηση:

$$I(t) = \begin{cases} 1 & \text{, αν } S_{total}(t) \text{ είναι μία εκ των } Mc \text{ μεγαλύτερων τιμών του } S_{total} \\ 0 & \text{, αλλιώς} \end{cases} \tag{52}$$

Η δημιουργούμενη συνάρτηση $I$ επεξεργάζεται περαιτέρω ώστε να σχηματιστούν γειτονικά blocks του βίντεο τα οποία θα περιληφθούν στην τελική περίληψη. [18]. Συγκεκριμένα, τα frames που μέχρι στιγμής έχουν επιλεγεί συνενώνονται σε τμήματα. Τμήματα που είναι μικρότερα από $N$ frames τελικά δε θα περιληφθούν στην περίληψη. Η διαδικασία αυτή μπορεί να υλοποιηθεί αποδοτικά ως το μορφολογικό άνοιγμα του $I$ με ένα μοναδιαίο διάνυσμα μήκους $N + 1$. Από τα εναπομείναντα τμήματα, εάν δύο απέχουν λιγότερο από $K$ frames, ενώνονται ώστε να σχηματίσουν ένα ενιαίο τμήμα. Η διαδικασία αυτή μπορεί να υλοποιηθεί αποδοτικά ως το μορφολογικό κλείσιμο του $I$ με ένα μοναδιαίο διάνυσμα μήκους $K + 1$. Τελικά, τα τμήματα που αποφασίζεται να περιληφθούν στην περίληψη ενώνονται για τη δημιουργία ενός νέου βίντεο με χρήση της τεχνικής fade-in και fade-out στα αρχικά και τελικά frames των τμημάτων, αντίστοιχα (αναλυτικά στοιχεία για την τεχνική αυτή θα δοθούν αργότερα).

Μέχρι στιγμής, έχουμε κάνει όλη την απαιτούμενη ανάλυση, θεωρώντας ότι το βίντεο επεξεργάζεται με ενιαίο τρόπο, δηλαδή η δομή $Q$ περιλαμβάνει όλα τα frames του προς ανάλυση βίντεο και αυτή δίνεται ως είσοδος στον αλγόριθμο. Προφανώς, κάτι τέτοιο είναι υπολογιστικά αδύνατο να υλοποιηθεί, καθώς θα είχε τεράστιες απαιτήσεις σε μνήμη. Για να αντιμετωπίσουμε το πρόβλημα αυτό, κάνουμε την επεξεργασία του βίντεο σε blocks από frames, δηλαδή βρίσκουμε για κάθε $T$ frames την αντίστοιχη saliency curve. Το ερώτημα τώρα που εγείρεται είναι πώς θα συνδυαστούν οι διαφορετικές αυτές καμπύλες σε μία.

Εάν απλά συνενώσουμε διαδοχικά τις καμπύλες, θα εξαχθεί μία τελική καμπύλη προφανώς ασυνεχής και που άρα δε θα ανταποκρίνεται στο πραγματικό ολικό saliency. Χρησιμοποιούμε, οπότε, την ιδέα της ανασύνθεσης σήματος με Overlap-Add, κάνοντας χρήση επικαλυπτόμενων blocks. Παρακάτω, δίνεται μία σύντομη ανάλυση της ιδέας σε ένα γενικότερο πλαίσιο για να γίνει αντιληπτή η χρήση της στη συγκεκριμένη εφαρμογή.

Ας θεωρήσουμε την short-time ανάλυση ενός μονοδιάστατου σήματος $x$ σε πλαίσια χρησιμοποιώντας παράθυρο $w$ πεπερασμένου μήκους $L$. Τότε μπορούμε να εκφράσουμε το $m$-οστό παραθυρωμένο πλαίσιο δεδομένων ως

$$x_m(n) = x(n)w(n - mR), \quad n \in (-\infty, \infty) \tag{53}$$

όπου $R$ το χρονικό βήμα ανάλυσης. Το χρονικό βήμα ανάλυσης είναι ο αριθμός των δειγμάτων μεταξύ των χρόνων έναρξης διαδοχικών πλαισίων. Συγκεκριμένα, είναι ο αριθμός των δειγμάτων κατά τον οποίο μετακινούμε κάθε επόμενο παράθυρο. Στο Σχήμα 18 φαίνεται το σήμα εισόδου και τρία διαδοχικά παραθυρωμένα πλαίσια ανάλυσης χρησιμοποιώντας ένα αιτιατό παράθυρο Hamming μήκους $L = 128$ με 50% επικάλυψη ($R = L/2 = 64$).

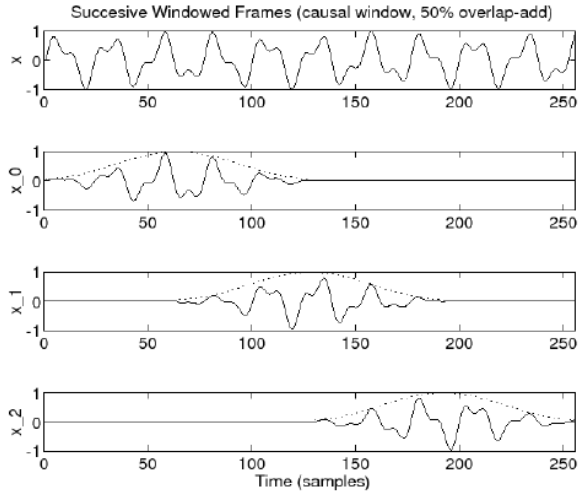
Για να δουλέψει, όμως, η ανάλυση σε πλαίσια θα πρέπει να μπορούμε να ανακατασκευάσουμε το σήμα $x$ από τα επιμέρους επικαλυπτόμενα παράθυρα, ιδανικά με απλή πρόσθεσή τους στις

αρχικές χρονικές τους θέσεις. Αυτό μπορεί να γραφτεί ως

$$x(n) = \sum_{m=-\infty}^{\infty} x_m(n) = x(n) \sum_{m=-\infty}^{\infty} w(m - nR) \tag{54}$$

Οπότε, η ανακατασκευή είναι δυνατή αν και μόνο αν

$$\sum_{m\in\mathbb{Z}} w[n - mR] = 1, \quad \forall n \in \mathbb{Z} \tag{55}$$




Σχήμα 18: Ανάλυση σε επικαλυπτόμενα πλαίσια. Σήμα εισόδου (πάνω) και τρία διαδοχικά blocks δεδομένων με παραθύρωση Hamming και 50% επικάλυψη.

Η συνθήκη (55) δεν ισχύει για όλους τους συνδυασμούς παραθύρων και επικαλύψεων. Ένας από τους συνδυασμούς που ικανοποιούν τη συνθήκη είναι η χρήση περιοδικού παραθύρου Hamming άρτιου μήκους και 50% επικάλυψη [19]. Αυτή είναι και η ''παραθύρωση'' που θα κάνουμε εμείς.

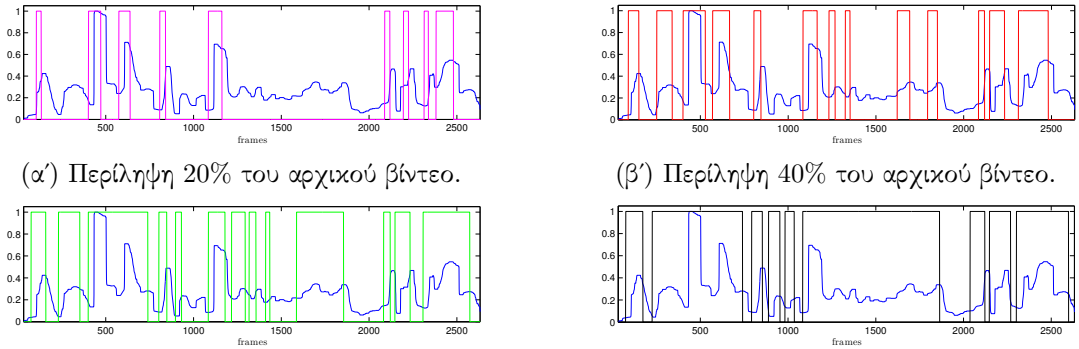
Επανερχόμενοι στο πρόβλημά μας, επιλέγουμε να κάνουμε την ανάλυση ανά blocks διάρκειας $T = 64$ frames. Για κάθε block παράγεται μία saliency curve η οποία δεν κανονικοποιείται και δεν φιλτράρεται από φίλτρο μέσης τιμής. Το σήμα αυτό πολλαπλασιάζεται εν συνεχεία με παράθυρο Hamming μήκους $L = 64$. Το κάθε νέο παραθυρωμένο σήμα συνενώνεται με το τρέχον ολικό saliency curve, σε σωστή θέση, τη στιγμή που ανάλυση γίνεται με 50% επικάλυψη, δηλαδή ανά 32 frames.

Εν συνεχεία, η ολική saliency curve φιλτράρεται με φίλτρο μέσης τιμής και κανονικοποιείται. Η υπόλοιπη ανάλυση που αφορά την τελική επιλογή των τμημάτων που θα περιληφθούν στο βίντεο έχει ήδη περιγραφεί. Με τη μέθοδο αυτή, όμως, τα πρώτα 32 frames αποκτούν πολύ μικρό saliency, καθώς ανήκουν μόνο στο πρώτο block, και άρα οι τιμές μικραίνουν λόγω της επίδρασης της παραθύρωσης Hamming. Προβληματικά επίσης είναι για τον ίδιο λόγο και τα τελευταία frames. Το πρόβλημα εκεί γίνεται ακόμα οξύτερο στην περίπτωση του $S_var$, διότι τα βάρη της εξίσωσης (49) σχετίζονται άμεσα με τον αριθμό των frames, που στο τελευταίο block δεδομένων μπορεί να μην είναι ακριβώς 64, αλλά μικρότερος.

Αντί να προβούμε σε ειδικές ενέργειες για την αντιμετώπιση των παραπάνω θεμάτων, επιλέγουμε απλά να αγνοήσουμε τα πρώτα 32 και τα τελευταία 32 frames. Έτσι κι αλλιώς, για ένα βίντεο 25frames/sec αυτό αντιστοιχεί σε $1.28sec$ στην αρχή του βίντεο και άλλα τόσα στο τέλος. Αντιλαμβανόμαστε ότι η πιθανότητα σε ένα μεγάλο βίντεο, του οποίου θέλουμε την περίληψη, να βρίσκεται πολύτιμη πληροφορία στην αρχή ή στο τέλος του είναι πρακτικά μηδενική.

Για το μορφολογικό άνοιγμα και κλείσιμο επιλέγουμε $N = 20$ και $K = 10$. Το φίλτρο μέσης τιμής που χρησιμοποιείται είναι μήκους 41. Υλοποιούμε fade-in και fade-out για 10 frames.

Ένα παράδειγμα του τελικού αποτελέσματος επιλογής των τμημάτων έντονου ενδιαφέροντος δίνεται στο Σχήμα 19 για ένα σύντομο βίντεο (πρόκειται για το ίδιο βίντεο που χρησιμοποιήθηκε στο Σχήμα 13).




(α′) Περίληψη 20% του αρχικού βίντεο.

(β′) Περίληψη 40% του αρχικού βίντεο.

(γ′) Περίληψη 60% του αρχικού βίντεο.

(δ′) Περίληψη 80% του αρχικού βίντεο.

Σχήμα 19: Saliency curve για ένα βίντεο συνολικής διάρκειας 2664 frames με επισημειωμένα τα τμήματα που επιλέγονται να περιληφθούν στην περίληψη για διαφορετικά επιθυμητά ποσοστά summarization.

Στο σημείο αυτό, αξίζει να αναφερθούν λίγα πράγματα για την τεχνική fade-in και fade-out. Η τεχνική αυτή χρησιμοποιείται για να μειωθούν όσο γίνεται οι απότομες μεταβάσεις μεταξύ των διαφορετικών σκηνών που περιλαμβάνονται τελικά στην περίληψη και το αποτέλεσμα να είναι πιο ομαλό και ευχάριστο. Επιλέξαμε, όπως είπαμε, να κάνουμε fade-in στα πρώτα 10 frames κάθε τμήματος και fade-out στα τελευταία 10. Σημειώνουμε πως είμαστε βέβαιοι πως η επιλογή αυτή δεν είνα προβληματική (υπό την έννοια ότι δεν υπάρχει κίνδυνος να βρεθούν τμήματα όπου η τεχνική δεν είναι υλοποιήσιμη με τις συγκεκριμένες παραμέτρους), εφόσον στην περίληψη, κατόπιν του μορφολογικού ανοίγματος, έχουν εισαχθεί τμήματα μεγαλύτερα ή ίσα των 20 frames.

Με στόχο μία κατά το δυνατόν εύληπτη και αποδοτική υλοποίηση, προχωρούμε στα εξής βήματα. Κατ′ αρχήν, μετασχηματίζουμε την πληροφορία που περιέχεται στο διάνυσμα $I$, συμπιέζοντάς την με τον αλγόριθμο Run-Length Encoding (RLE), σύμφωνα με τον οποίο ακολουθίες διαδοχικών στοιχείων με την ίδια τιμή αποθηκεύονται ως το ζευγάρι της τιμής αυτής και του πλήθους των στοιχείων. Ένα απλό παράδειγμα ενός πιθανού διανύσματος $I$ (όπου δεν έχει γίνει κάποιο μορφολογικό κλείσιμο ή άνοιγμα σύμφωνα με τα όσα έχουν ειπωθεί), παρουσιάζεται στον Πίνακα 7.

Στη μετασχηματισμένη αυτή δομή βλέπουμε πως εμφανίζονται εναλλάξ μηδενικά και άσσοι, οπότε εναλλάξ τμήματα από frames περιλαμβάνονται ή όχι στην περίληψη. Για κάθε νέο τμήμα που εισέρχεται στην περίληψη, κάνουμε fade-in στα πρώτα 10 frames. Άμεσα, όπως βλέπουμε στον Πίνακα 7, γνωρίζουμε το μέγεθος του κάθε τμήματος, οπότε μπορούμε εύκολα να ξέρουμε ποια είναι τα τελευταία 10 frames του τμήματος για να εφαρμοστεί fade-out. Τονίζεται πως η

διαδικασία της ανάγνωσης του αρχικού αρχείου βίντεο και τελικά εγγραφής του νέου αρχείου, γίνεται frame-by-frame.

| Αρχική Μορφή της Πληροφορίας | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |

| Μετασχηματισμένη Μορφή της Πληροφορίας (με RLE) | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 5 | 2 | 2 | 4 | 4 | 3 | 5 |

Πίνακας 7: Παράδειγμα εφαρμογής του αλγορίθμου RLE.

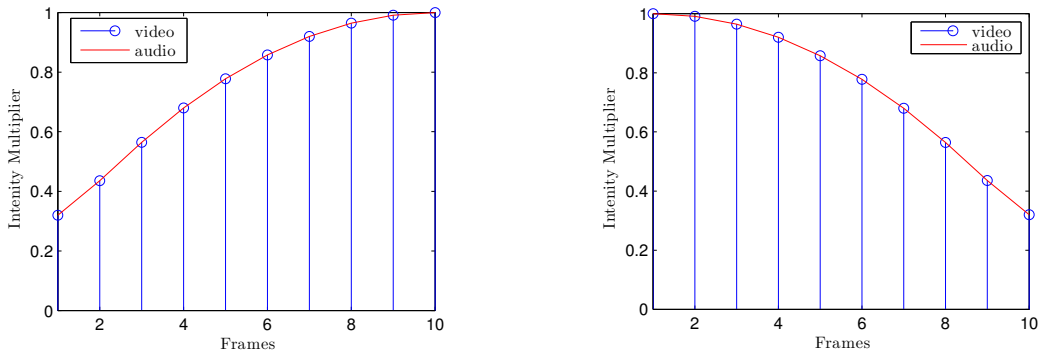
Σημειώνεται, ακόμη, πως, παρόλο που η εξαγωγή των σημείων ενδιαφέροντος γίνεται αποκλειστικά στη βάση των οπτικών χαρακτηριστικών του βίντεο, η τελικώς εξαγόμενη περίληψη είναι επιθυμητό να διατηρεί και την όποια ηχητική πληροφορία. Για το σκοπό αυτό, γίνεται σύγχρονη ανάγνωση (frame-by-frame) και εγγραφή και του ήχου. Στην ηχητική πληροφορία η μόνη επεξεργασία που λαμβάνει χώρα είναι η υλοποίηση και εδώ της τεχνικής fade-in, fade-out.

Σε κάθε περίπτωση η κλίμακα που χρησιμοποιείται για το fade-out είναι η αντίστροφη αυτής για το fade-in. Για το λόγο αυτό, αναλύεται στη συνέχεια μόνο το fade-in. Το πρόβλημα αρχικά προσεγγίστηκε με μία γραμμική κλίμακα όπου η ένταση και στα τρία χρωματικά κανάλια του βίντεο πολλαπλασιαζόταν σταδιακά με τιμές από το 0.1 μέχρι το 1.0. Ωστόσο, για βίντεο με FPS της τάξης των $25 frames/second$, το αποτέλεσμα δεν ήταν ευχάριστο στο μάτι, καθώς εμφανίζονταν αρκετά σχεδόν μαύρα frames που γίνονταν έντονα αντιληπτά.

Τελικά, κατόπιν λίγου πειραματισμού, καταλήξαμε σε μία σιγμοειδή κλίμακα (για την ακρίβεια, τμήμα σιγμοειδούς) για την ένταση των 10 frames, η οποία παρουσιάζεται, τόσο για το fade-in, όσο και για το fade-out, στο Σχήμα 20. Στο ίδιο Σχήμα παρουσιάζεται και η αντίστοιχη κλίμακα για τον ήχο. Η κλίμακα είναι η ίδια, με τη διαφορά, όμως, ότι κάθε frame δεν περιλαμβάνει ένα δείγμα ήχου (όπως συμβαίνει με την εικόνα), αλλά από περισσότερα. Για την ακρίβεια, ισχύει η εξίσωση (56).

$$Samples\ per\ Frame = \frac{Audio\ Frequency}{Frame\ Rate} \tag{56}$$

Για να επιτευχθεί, λοιπόν, μία ίδια κλιμάκωση του ήχου όπως της εικόνας, αρκεί στην ίδια δημιουργηθείσα κλίμακα να κάνουμε ένα interpolation, ώστε από μία κλίμακα 10 στοιχείων να περάσουμε σε μία κλίμακα $10 \cdot Samples\ per\ Frame$ στοιχείων.




(α′) Fade-in. (β′) Fade-out.

Σχήμα 20: Η κλίμακα που χρησιμοποιήθηκε για την υλοποίηση του fade-in και του fade-out.

# 4 Διεξαγωγή Πειραμάτων και Αξιολόγηση

Με βάση τη μέθοδο που έχει αναλυθεί παραπάνω, έγιναν οι περιλήψεις σε τρεις ταινίες μικρού μήκους, διάρκειας περίπου 10 λεπτών η καθεμία. Οι ταινίες που χρησιμοποιήθηκαν είναι οι εξής: The Most Beautiful Thing[3], Losses[4], This Way Up[5]. Πρόκειται για ταινίες διαφορετικής κατηγορίας η καθεμία. Συγκεκριμένα, η πρώτη είναι κοινωνική ταινία, η δεύτερη ταινία δράσης και η τρίτη μία animated κωμωδία.

Αρχικά, έγινε κάποια προεπεξεργασία, ώστε όλες να έχουν frame rate 25 $frames/sec$ και ανάλυση $320 \times 180$ $pixels$. Καμία επιπλέον επεξεργασία δεν έγινε ώστε να επιτευχθεί manually καλύτερο αποτέλεσμα. (Για παράδειγμα, θα μπορούσαμε εξ αρχής να αποκόψουμε όσα frames αντιστοιχούν σε τίτλους αρχής και τέλους, συντελεστές, κ.λπ.). Και οι τρεις ταινίες δόθηκαν ως είσοδοι στο πρόγραμμα σε container .avi.

Για κάθε μία ταινία, παράχθηκε ένα σύνολο περιλήψεων για διαφορετικά ποσοστά περίληψης $c$. Για καθένα από αυτά, παράχθηκαν τέσσερις διαφορετικές περιλήψεις που αντιστοιχούν στους τέσσερις διαφορετικούς τρόπους που παρουσιάστηκαν για τη συγχώνευση των χαρακτηριστικών της έντασης, του χρώματος και του προσανατολισμού (γραμμική - lin, βασισμένη στη διασπορά - var, μη-γραμμική με λογική μεγίστου - max, μη-γραμμική με λογική ελαχίστου - min).

Οι παραγόμενες περιλήψεις και πάλι αποθηκεύτηκαν σε container .avi με ίδια ανάλυση και frame rate. Παρόλο που χρησιμοποιήθηκε συμπίεση τόσο για το stream της εικόνας (MGPEG Compressor), όσο και για το stream του ήχου (Microsoft ADPCM), τα παραγόμενα αρχεία ήταν εξαιρετικά μεγάλα για την ανάλυση και τη διάρκειά τους[6]. Για το λόγο αυτό, έγινε εκ των υστέρων επεξεργασία των αρχείων βίντεο με σκοπό να μειωθεί το bitrate και άρα το μέγεθός τους. Η τελική μορφή των αρχείων ήταν σε container .mp4.

Από τις διαφορετικές περιλήψεις που δημιουργήθηκαν, θεωρήθηκε ότι ένα ικανοποιητικό ποσοστό περίληψης που συμβιβάζεται μεταξύ διατήρησης της απαραίτητης πληροφορίας και αρκετής ''συμπίεσης'' είναι της τάξης του 30%. Έτσι, λοιπόν, για έναν πρώτο πειραματισμό, ένα δείγμα ανθρώπων κλήθηκε να αξιολογήσει το αποτέλεσμα για τους διαφορετικούς τρόπους συγχώνευσης των χαρακτηριστικών, για ποσοστό περίληψης $c = 0.3$. Συγκεκριμένα, για τις τρεις διαφορετικές ταινίες και τους τέσσερις διαφορετικούς τρόπους ζητήθηκε να μπει ένας ακέραιος βαθμός στην κλίμακα $[1, 5]$, με 1 να δηλώνει το πολύ κακό και 5 το πολύ καλό. Για κάθε ταινία θα έπρεπε το κάθε υποκείμενο να βάλει τουλάχιστον ένα 5, ώστε όλα τα αποτελέσματα να είναι κατά κάποιο τρόπο κανονικοποιημένα ως προς την καλύτερη μέθοδο και να έχει νόημα η εξαγωγή ενός μέσου όρου για τη σύγκριση μεταξύ των διαφορετικών μεθόδων. Κριτήρια για τη βαθμολογία ήταν η ύπαρξη των φαινομενικά σημαντικών σκηνών της ταινίας στην περίληψη, η κατά το δυνατόν μείωση των artifacts (απότομες μεταβάσεις, κοψίματα σκηνών, κ.λπ.) και γενικότερα το κατά πόσο το αποτέλεσμα ήταν ευχάριστο για το θεατή.

Τα αποτελέσματα παρουσιάζονται στο Σχήμα 21.

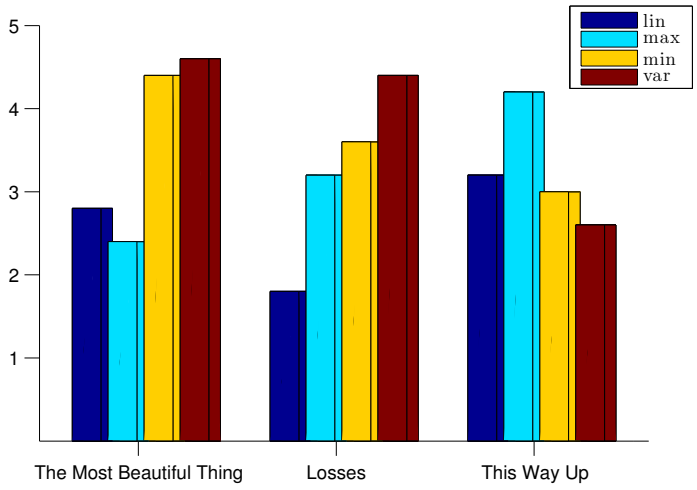
Για τις ταινίες The Most Beautiful Thing και Losses, βέλτιστη μέθοδος αναδεικνύεται η συγχώνευση που βασίζεται στη διασπορά, ενώ ακολουθεί η συγχώνευση που στηρίζεται στη λογική ελαχίστου. Ιδίως για την πρώτη ταινία, μάλιστα, οι δύο αυτές μέθοδοι φαίνεται να παρουσιάζουν μεγάλη ποιοτική διαφορά από τις υπόλοιπες δύο.

[3] https://www.youtube.com/watch?v=IP8psM4LWXk

[4] https://www.youtube.com/watch?v=BMhXexbDmv8

[5] https://www.youtube.com/watch?v=vmfC7VY5Ods

[6] Λόγω εγγενών περιορισμών της βιβλιοθήκης vision της MATLAB στις διαφορετικές της διανομές, η όλη διαδικασία που περιγράφεται όσον αφορά στην εγγραφή των περιλήψεων σε αρχείο είναι δυστυχώς εφικτή, με τον τρόπο που υλοποιήθηκε, μόνο σε συστήματα Windows.




Σχήμα 21: Οι διαφορετικές μέθοδοι συγχώνευσης για τις τρεις ταινίες, όπως αξιολογήθηκαν από ανεξάρτητους αξιολογητές.

Εντύπωση, ωστόσο, προξενούν τα αποτελέσματα που αφορούν στην ταινία This Way Up, όπου η κατάσταση φαίνεται να αντιστρέφεται, με τη συγχώνευση που βασίζεται σε λογική μεγίστου να παίρνει τα πρωτία και στην τελευταία θέση να βρίσκεται η μέθοδος που αναδεικνύεται καλύτερη στις άλλες δύο ταινίες. Υποθέτουμε πως ρόλο εδώ παίζει η ιδιαίτερη φύση της ταινίας. Όπως είπαμε, πρόκειται για μια animated ταινία (κινουμένων σχεδίων), γεγονός που τη διαφοροποιεί πολύ από τις άλλες δύο.

Ακόμη, δε θα πρέπει να ξεχνάμε πως το Σχημα 21 αντιπροσωπεύει βαθμολογίες όχι απόλυτες, αλλά σχετικές, ώστε να προκύψει μία ασφαλής σύγκριση των τεσσάρων μεθόδων. Όλοι οι αξιολογητές κλήθηκαν στη συνέχεια να δώσουν μία % εκτίμηση της ποιότητας της βέλτιστης κατά τη γνώμη τους περίληψης της κάθε ταινίας, δηλαδή αυτής στην οποία έβαλαν την υψηλότερη βαθμολογία. Παρόλο που δεν είχαμε στη διάθεσή μας περιλήψεις που έχουν εξαχθεί από άνθρωπο, οι αξιολογητές κλήθηκαν να δώσουν ένα μέτρο της ποιότητας σε σχέση με το αποτέλεσμα που θα περίμεναν από τη δουλειά ενός ειδικού, η οποία πιθανώς προορίζεται για εμπορική χρήση. Ο μέσος όρος των αποτελεσμάτων παρουσιάζεται στον Πίνακα 8.

| The Most Beautiful Thing | Losses | This Way Up |
|---|---|---|
| 79% | 69% | 64% |

Πίνακας 8: Μέση ποιότητα της βέλτιστης περίληψης για την κάθε ταινία, όπως εκτιμήθηκε από τους αξιολογητές.

Παρατηρούμε πως η χαμηλότερη ποιότητα αποδόθηκε όπως περιμέναμε στην ταινία This Way Up, όπου παρατηρήθηκε και το ''παράδοξο'' που εξηγήθηκε προηγουμένως. Η ταινία Losses, αν και ταινία δράσης με έντονες αλλαγές και εμφανή σημεία και frames ενδιαφέροντος, πήρε χαμηλότερη απόλυτη βαθμολογία από το αναμενόμενο. Αποδίδουμε το γεγονός αυτό στο ότι στη συγκεκριμένη ταινία κόπηκαν στη μέση σκηνές διαλόγου, κάτι που έχει σαφώς αρνητικό αντίκτυπο στο θεατή. Ωστόσο, εφόσον η όλη επεξεργασία και ανάλυση έγινε μόνο στην πληροφορία της εικόνας και όχι του ήχου, αυτό ήταν κάτι που δεν μπορούσαμε να αποφύγουμε.

# Αναφορές

[1] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avtirhis, "Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual and Textual Attention", *IEEE Transactions on Multimedia*, vol. 10, pp 1-16, 2013.

[2] A. G. Money, H. Agius, "Video Summarisation: A conceptual framework and survey of the state of the art", *Journal of Visual Communication and Image Represantation*, vol. 19, pp 121-143, 2008.

[3] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp 1254-1259, 1998

[4] C. Koch, S. Ullman, "Shifts in Selective visual attention: towards the underlying neural circuity", *Human Neurobiology*, vol. 4, pp 219-227, 1985.

[5] L. Itti, C. Koch, "Computational Modelling of Visual Attention", *Nature Reviews Neuroscience*, vol. 2, pp 194-203, 2001.

[6] K. Rapantzikos, Y. Avrithis, S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition", *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[7] K. Koffka, *Principles Of Gestalt Psychology*, Routledge, 2013

[8] V. S. Ramachandran, *The Tell-Tale Brain: A Neuroscientist's Quest for What Makes Us Human*, W. W. Norton, 2012

[9] R. C. Gonzalez, R. E. Woods, *Ψηφιακή Επεξεργασία Εικόνας*, Εκδόσεις Τζιόλα, 2011

[10] E. H. Adelson, J. R. Bergen, "Spatiotemporal energy models for the perception of motion", *Journal of the Opticl Society of America A*, vol. 2, no. 2, pp 284-299, 1985

[11] W. T. Freeman, E. H. Adelson, "The Design and Use of Steerable Filters", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp 891-906, 1991

[12] Σ. Κόλλιας, *Ασκήσεις Ψηφιακής Επεξεργασίας και Ανάλυσης Εικόνων*, ΕΜΠ, 2005

[13] K. G. Derpanis, J. M. Gryn, "Three-dimensional n-th derivative of Gaussian separable steerable filters", *Proceedings, IEEE International Conference on Image Processing*, pp III–553–6, 2005

[14] K. G. Derpanis, J. M. Gryn, "Three-dimensional n-th derivative of Gaussian separable steerable filters", *Technical Report CS-2004-05*, York University, 2004

[15] Ν. Καδιανάκης, Σ. Καρανάσιος, *Γραμμική Άλγεβρα, Αναλυτική Γεωμετρία και Εφαρμογές*, Ν. Καδιανάκης, Σ. Καρανάσιος, 2008

[16] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, J. M. Ogden, "Pyramid methods in image processing", *RCA Engineer*, vol. 29, no. 6, pp 33-41, 1984

[17] P. J. Burt, E. H. Adelson, "The Laplacian Pyramid as a Compact Image Code", *IEEE Transactions on Communications*, vol. COM-31, no. 4, pp 532-540, 1983

[18] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, Y. Avrithis, "Video Event Detection and Summarization Using Audio, Visual and Text Saliency", *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp 3553-3556, 2009

[19] J. O. Smith III, *Spectral Audio Signal Processing*, W3K Publishing, 2011