Αναγνώριση σκηνών βίας με χρήση της ακουστικής πληροφορίας

Αριστείδου Ανδρέας[1]

Επιβλέποντες:
Θεοδωρίδης Σέργιος[1], Κοσμόπουλος Δημήτριος[2], Γιαννακόπουλος Θεόδωρος[1]

[1]. *Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών,*
aa662@cam.ac.uk, {stheodor, tyiannak}@di.uoa.gr
[2]. *Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών, ΕΚΕΦΕ "Δημόκριτος".*
dkosmo@iit.demokritos.gr

Περίληψη. Η εργασία αυτή επικεντρώνεται στη μελέτη των μεθόδων που αποσκοπούν στην αυτόματη και αξιόπιστη αναγνώριση των σκηνών βίας με χρήση της ακουστικής πληροφορίας του σήματος. Χρησιμοποιήθηκαν ηχητικά χαρακτηριστικά όπως η ενεργειακή εντροπία, η ενέργεια μικρής διάρκειας (short time energy), το ZCR, η φασματική διακύμανση (spectral flux). Στη συνέχεια, αναπτύξαμε ένα μηχανισμό ταξινόμησης SVM (Support Vector Machine), ο οποίος εκπαιδεύθηκε αξιοποιώντας τα ηχητικά χαρακτηριστικά γνωρίσματα της βίας. Τα αποτελέσματα της ταξινόμησης καταγράφονται, αξιολογούνται και αναφέρονται το ποσοστό επιτυχούς ταξινόμησης, το οποίο είναι καλύτερο από οποιαδήποτε άλλη προτεινόμενη μέθοδο.

Λέξεις Κλειδιά: Εντοπισμός βίας, Ταξινόμηση ήχου, Επεξεργασία Ήχου, Χαρακτηριστικά γνωρίσματα ήχου, Support Vector Machines (SVM).

1. Εισαγωγή

Στο προσεχές μέλλον, αναμένουμε μια τεράστια αύξηση των διαθέσιμων πολυμέσων, η οποία θα παρέχεται είτε δωρεάν, π.χ., στο World Wide Web, σε peer to peer δίκτυα κτλ., είτε μέσω ιδιωτικών δικτύων (δορυφορική τηλεόραση, δίκτυα κινητής τηλεφωνίας), σαν υπηρεσίες βίντεο κατόπιν απαιτήσεως (video-on-demand). Σχεδόν οποιοσδήποτε θα είναι ικανός να παρέχει πολυμέσα, στα οποία θα μπορεί να έχει πρόσβαση μεγάλος αριθμός ανθρώπων. Παρόλα αυτά, η διαθεσιμότητα αυτών των πολυμέσων δεν είναι πάντα ό,τι καλύτερο, λόγω των ιδιαιτεροτήτων της κάθε ηλικίας και των ευαισθησιών του κάθε ανθρώπου. Η τεράστια αύξηση στη χρήση ανάλογης τεχνολογίας από ευαίσθητα κοινωνικά σύνολα, όπως παιδιά, επιβάλλει την ανάγκη προστασίας τους από ενδεχόμενες αρνητικές επιπτώσεις.

Η εργασία αυτή αποτελεί μέρος μιας πολυμορφικής προσέγγισης, που στόχο έχει να συμβάλει στον αποτελεσματικό και έγκυρο εντοπισμό των σκηνών βίας. Η βία είναι ένα υποκειμενικό χαρακτηριστικό, κάτι που αποτελεί εμπόδιο στην προσπάθεια να δοθεί ένας

αξιόπιστος και με σαφήνεια ορισμός της. Ως βία μπορούμε να ορίσουμε οποιαδήποτε πράξη που μπορεί να προκαλέσει φυσικά ή και ψυχολογικά τραύματα σε ένα ή περισσότερα άτομα. Οπότε, βίαιη σκηνή σε ταινία μπορεί να χαρακτηριστεί οποιαδήποτε σκηνή στην οποία εμπεριέχεται μία ή περισσότερες πράξεις βίας ή όταν τα αποτελέσματα αυτής της πράξης καθιστούν εμφανή την πράξη που προηγήθηκε (π.χ. η κραυγή πόνου, η προβολή του τραυματισμένου ατόμου ή του πτώματος αλλά όχι απαραίτητα και της σκηνής που προηγήθηκε).

Σε αυτή την έρευνα επιδιώκουμε να αναλύσουμε και να ταξινομήσουμε τα ηχητικά γνωρίσματα των σκηνών βίας. Μερικές εφαρμογές αυτής της προσπάθειας αποτελούν η αυτόματη κατηγοριοποίηση των ταινιών σύμφωνα με το περιεχόμενο τους, η αυτόματη αναγνώριση των σκηνών από την ελεγκτική υπηρεσία και η επισήμανση τους ως ταινίες κατάλληλες για παιδιά, έφηβους, ενήλικες κ.τ.λ..

2. Σχετικές έρευνες

Η επεξεργασία του ήχου, ως προς το περιεχόμενο του, αποτελεί μέρος μιας ευρύτερης έρευνας, μια καλή επισκόπηση των οποίων δίνεται μέσα στο [1], η οποία στοχεύει στην ταξινόμηση σύμφωνα με το περιεχόμενο: (α) γενικής κατηγορίας (ύφος), με βάση δηλαδή το σκοπό για τον οποίο δημιουργήθηκε, π.χ., ειδήσεις, διαφημίσεις, αθλητικά, (β) σημασιολογικές ή λογικές μονάδες, π.χ., ο διάλογος σε μια ταινία, η πρόβλεψη του καιρού σε ένα δελτίο ειδήσεων, (γ) γεγονότα, π.χ., τέρμα κατά τη διάρκεια ενός αγώνα ποδοσφαίρου, έκρηξη σε μια ταινία.

Για το πρόβλημα εντοπισμού της βίας, οι προηγούμενες αναφορές είναι πολύ περιορισμένες και εστιάζονται μόνο στα οπτικά χαρακτηριστικά γνωρίσματα της βίας ([2], [3]). Επίσης, στην [4], τα ηχητικά χαρακτηριστικά χρησιμοποιούνται ως ένα πρόσθετο χαρακτηριστικό γνώρισμα που χρησιμοποιείται παράλληλα με την οπτική πληροφορία. Στο [4] δηλώνεται η χρήση ενός Gaussian μοντέλου στο οποίο υπολογίζεται ο λόγος πιθανοτήτων μεταξύ του ηχητικού σήματος εισόδου και των δύο ηχητικών κλάσεων, χρησιμοποιώντας την μέση τιμή και την απόκλιση για να καθοριστεί η κατηγορία (βία ή μη βία) στην οποία ανήκει ο συγκεκριμένος ήχος. Επιπλέον, οι συγγραφείς εντοπίζουν την απότομη αλλαγή στα ενεργειακά επίπεδα του ηχητικού σήματος και ορίζουν την ενεργειακή εντροπία ως ένα χαρακτηριστικό γνώρισμα που συνδέεται με τη χρήση βίας. Στην συνέχεια, ταξινομούν τις δύο κλάσεις με την χρήση ταξινομητή *k*-πλησιέστερων γειτόνων (*k*-*NN*, k=2).

Αυτό που μπορούμε να συμπεράνουμε από τις προηγούμενες έρευνες, είναι ότι αν και ο ήχος είναι μια πολύ χρήσιμη πηγή πληροφοριών, πολύ απλούστερη να επεξεργαστεί από ότι η οπτική, έχει μάλλον αγνοηθεί. Οπότε στόχος αυτής της εργασίας είναι η χρήση πρόσθετων χαρακτηριστικών γνωρισμάτων καθώς επίσης και καλύτερων μεθόδων ταξινόμησης που θα είναι σε θέση να παρέχουν καλύτερα αποτελέσματα.

3. Ηχητικά Χαρακτηριστικά

Η εργασία αυτή μελετά μεθόδους που αποσκοπούν στην αυτόματη και αξιόπιστη αναγνώριση των σκηνών βίας με χρήση της ακουστικής πληροφορίας του σήματος. Ουσιαστικά, εξάγονται τα ηχητικά χαρακτηριστικά που προσδιορίζουν τις σκηνές βίας, τα οποία εισάγονται σε ένα ταξινομητή Support Vector Machine (SVM) [5].

Για τον υπολογισμό των γνωρισμάτων αυτών υποθέτουμε πως το σήμα έχει ήδη κατατμηθεί και διαχωριστεί σε σκηνές σύμφωνα με το περιεχόμενο τους. Οι σκηνές χωρίζονται σε *W* χρονικά παράθυρα και στο καθ' ένα από αυτά υπολογίζονται τα γνωρίσματα του. Στη μελέτη αυτή χρησιμοποιούνται ηχητικά χαρακτηριστικά, τόσο από το πεδίο του χρόνου όσο και από το πεδίο των συχνοτήτων ([6]).

3.1 Χαρακτηριστικά γνωρίσματα στο πεδίο του χρόνου

3.1.1 Χαρακτηριστικά γνωρίσματα βασισμένα στην ενεργειακή εντροπία

Η ενεργειακή εντροπία είναι κατάλληλη για να εκφράσει απότομες αλλαγές στα ενεργειακά επίπεδα του ηχητικού σήματος. Η τιμή της για ένα πλαίσιο *j* μπορεί να υπολογιστεί με βάση την παρακάτω εξίσωση:

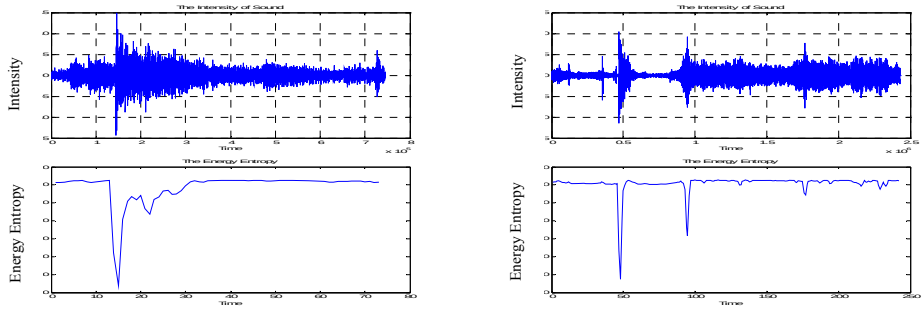
$$I_j = -\sum_{i=1..S} \sigma_i^2 \log_2 \sigma_i^2 \qquad \textbf{(1)}$$

όπου S είναι ο συνολικός αριθμός των τμημάτων ενός πλαισίου, $\sigma_i^2 = \frac{E_i}{E}$ είναι η κανονικοποιημένη ενέργεια του *i*-στου τμήματος ενός πλαισίου, *E,* η συνολική ενέργεια του σήματος και E_i , η ενέργεια του *i*-στου τμήματος.

Οι σκηνές βίας όπως πυροβολισμός, έκρηξη έχουν ιδιαίτερα ηχητικά χαρακτηριστικά. Συγκεκριμένα, συμβαίνουν συνήθως σε πολύ μικρά χρονικά διαστήματα και δημιουργούν μια απότομη αύξηση της ενέργειας του ήχου. Η διάρκεια που απαιτήθηκε για την αλλαγή του ενεργειακού επιπέδου των ηχητικών σημάτων, αποτελεί ένα αξιόπιστο κριτήριο διαχωρισμού της βίας. Για να μετρήσουμε αποτελεσματικά αυτό το χαρακτηριστικό γνώρισμα χρησιμοποιούμε το κριτήριο της *ενεργειακής εντροπίας,* στο οποίο, η τιμή είναι μεγάλη, για πλαίσια με μικρή ενεργειακή μεταβολή, ενώ είναι μικρή, στα πλαίσια όπου η ενέργεια μεταβάλλεται απότομα. Αυτό γίνετε εύκολα αντιληπτό στο σχήμα 1α όπου, λόγω της απότομης μεταβολής της ενέργειας του ήχου, το πλαίσιο της σκηνής του πυροβολισμού παρουσιάζει απότομη μείωση στην τιμή της ενεργειακής του εντροπίας.

Τα χαρακτηριστικά f_1 και f_2, τα οποία βασίζονται στην ενεργειακή εντροπία, είναι ο λόγος της μέγιστης ως προς την μέση και της μέγιστης ως προς την μεσαία (median) ενεργειακή εντροπία και υπολογίζονται ως εξής:

$$f_1 = \frac{\max_{j=1..W}(I_j)}{\frac{1}{W}\sum_{j=1..W} I_j} \qquad \textbf{(2)} \qquad f_2 = \frac{\max_{j=1..W}(I_j)}{\underset{j=1..W}{median}(I_j)} \qquad \textbf{(3)}$$



α. **β.**

Σχήμα 1: Ηχητικό σήμα με (α) πυροβολισμό (β) ξυλοδαρμό και η αντίστοιχη ενεργειακή εντροπία τους.

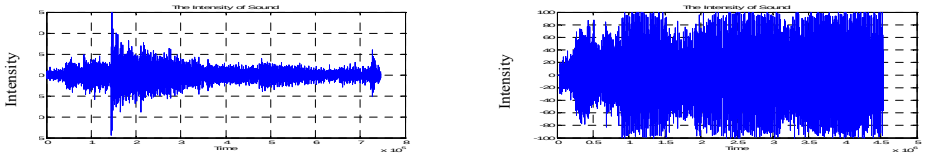
3.1.2 Χαρακτηριστικά γνωρίσματα βασισμένα στην ένταση του σήματος

Άλλο ένα χαρακτηριστικό που μπορεί να εντοπίσει απότομες μεταβολές της έντασης του σήματος είναι το κριτήριο f_3, το οποίο υπολογίζεται ως εξής:

$$f_3 = \frac{\max_{i=1..S'}(A_{0i})}{\frac{1}{S'}\sum_{i=1..S'} A_{0i}} \qquad \textbf{(4)}$$

όπου S' ο αριθμός των δειγμάτων του πλαισίου και A_{0i} η ένταση του i-στου δείγματος.

Σε αυτό το χαρακτηριστικό χρησιμοποιούμε τη μέγιστη τιμή της έντασης σε ένα ηχητικό δείγμα και διαιρώντας την με τη μέση τιμή της απόλυτης έντασης, προσδιορίζουμε σε ποια δείγματα υπάρχει έντονη μεταβολή, όπου ενδείκνυται να αποτελεί σκηνή βίας. Είναι εμφανές πως όταν υπάρχουν κορυφές στην ένταση του ηχητικού σήματος, οι τιμές του κριτηρίου θα είναι μεγάλες, ενώ στην αντίθετη περίπτωση η τιμές δεν θα ξεπερνούν ένα συγκεκριμένο κατώτατο όριο. Αυτό φαίνεται στο σχήμα 2, όπου η τιμές του δείγματος με σκηνή βίας είναι χαρακτηριστικά διαχωρίσιμες.



α. **β.**

Σχήμα: 2 Ηχητικά δείγματα βασισμένα στο κριτήριο f_3 : (α) βία (f_3=19.12) (β) μουσική (f_3=5.47)

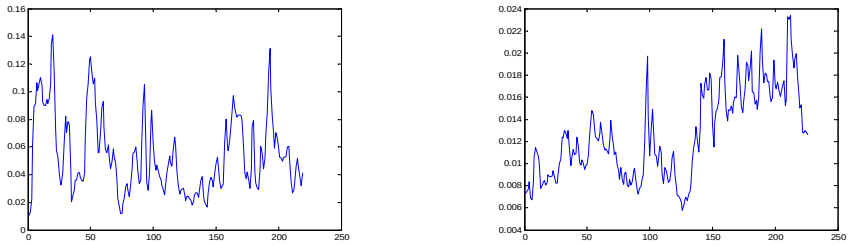
3.1.3 Χαρακτηριστικά γνωρίσματα βασισμένα στην ενέργεια μικρής διάρκειας

Λαμβάνοντας και πάλι υπόψη το γεγονός ότι οι σκηνές βίας, όπως κτύπημα, πυροβολισμός όπλου, έκρηξη, συμβαίνουν σε πολύ μικρά χρονικά διαστήματα και χαρακτηρίζονται από την ξαφνική μεταβολή του ήχου, εντάξαμε σαν κριτήριο και την

ενέργεια μικρής διάρκειας (Short time energy). Σε αυτό το κριτήριο υπολογίζουμε την ένταση του ηχητικού σήματος ανά μπλοκ. Ακολούθως μπορούμε να χρησιμοποιήσουμε τα χαρακτηριστικά f_4 και f_5 που αποτελούν τις τιμές της μέσης (mean) και σταθερής απόκλισης (variance) της συγκεκριμένης ενέργειας:

$$f_4 = \frac{1}{W}\sum_{j=1..W} N_j \qquad \textbf{(5)} \qquad\qquad f_5 = \frac{1}{W}\sum_{j=1..W}(N_j - f_4)^2 \qquad \textbf{(6)}$$

Παραδείγματα δύο διαφορετικών χαρακτηριστικών ακολουθιών, του κριτηρίου της ενέργειας μικρής διάρκειας, παρουσιάζονται στα σχήματα 3α και 3β. Είναι εμφανές πως, η ενεργειακή ακολουθία του 1ου ηχητικού δείγματος (πυροβολισμός) αποτελείται από "κορυφές" με πολύ υψηλές τιμές ενέργειας, σε αντίθεση με τις ενεργειακές μεταβάσεις του 2ου δείγματος, (μουσική) οι οποίες είναι μικρότερες. Όπως αναφέραμε και πριν, σε αυτή την εργασία χρησιμοποιούμε τις μέσες τιμές και την απόκλιση της ενέργειας ως διαχωριστικά χαρακτηριστικά.



α. **β.**

Σχήμα 3: Η ακολουθία Short time energy σε ηχητικά δείγματα **(a)** πυροβολισμός **(b)** μουσική

3.1.4 Γνωρίσματα βασισμένα στη διάσχιση του μηδενικού άξονα (ZCR)

Η διάσχιση του μηδενικού άξονα (Zero Crossing Rate, ZCR) αποτελεί μια από τις δημοφιλέστερες τεχνικές που χρησιμοποιούνται σήμερα για ταξινόμηση και κατηγοριοποίηση των ηχητικών δειγμάτων στο πεδίο του χρόνου. Με την μέθοδο αυτή μπορούμε να προσδιορίσουμε τη συχνότητα του σήματος. Ουσιαστικά, έχουμε Zero-Crossing (διάσχιση του μηδενικού άξονα) κάθε φορά που δύο διαδοχικές τιμές έχουν διαφορετικό πρόσημο. Το ZCR υπολογίζεται ως ο αριθμός των χρονικών διασχίσεων του μηδενικού άξονα (zero-crossings), διαιρεμένος με τον συνολικό αριθμό των δειγμάτων στο πλαίσιο.

$$Z_j = \frac{1}{2}\sum_{i=1..S} |\,\mathrm{sgn}(x_i) - \mathrm{sgn}(x_{i-1})\,| \cdot \omega(j-i) \qquad \textbf{(7)}$$

όπου: $\mathrm{sgn}(x_i) = \begin{cases} 1, x_i \geq 0 \\ -1, x_i < 0 \end{cases}$ και $\omega(n) = \begin{cases} 1, 0 \leq |\,n\,| \leq S-1 \\ 0, |\,n\,| \geq S \end{cases}$ (τετραγωνικό παράθυρο)

Το χαρακτηριστικό που χρησιμοποιούμε σε αυτή την εργασία είναι ο λόγος της μέγιστης ως προς την μέση τιμή του ZCR.

$$f_6 = \frac{\max_{j=1..W} Z_j}{\frac{1}{W}\sum_{j=1..W} Z_j} \quad (8)$$

Οι σκηνές βίας έχουν σαφώς υψηλότερο λόγο από ότι οι σκηνές οι οποίες δεν εμπεριέχουν βία, κάτι που αποτελεί ένα αξιόπιστο διαχωριστικό κριτήριο. Στο σχήμα 4α, παρουσιάζεται μια ακολουθία τριών διαφορετικών ηχητικών σημάτων από ZCR τιμές που περιέχουν βία (πυροβολισμός), μουσική και ομιλία. Η γραφική παράσταση που ακολουθεί μπορεί να επιβεβαιώσει τα προαναφερθέντα για την επιλογή του κριτηρίου f_6.

3.2 Χαρακτηριστικά γνωρίσματα στο πεδίο των συχνοτήτων

Οι ακόλουθες μέθοδοι είναι πολύ δημοφιλείς στην ηχητική ταξινόμηση, παρέχοντας πληροφορίες για τις ιδιότητες του ήχου ([6]).

3.2.1 Χαρακτηριστικά γνωρίσματα βασισμένα στη Φασματική Διακύμανση

Η φασματική διακύμανση (Spectral flux) είναι ένας τρόπος υπολογισμού της τοπικής φασματικής μεταβολής μεταξύ διαδοχικών πλαισίων, και υπολογίζεται ως εξής:

$$F_j = \sum_{k=0..S-1} (N_{j,k} - N_{j-1,k})^2 \quad (9)$$

όπου $N_{j,k}$ είναι η φασματική ενέργεια του j-στου πλαισίου στο k-στο δείγμα.

Βασικά, η σκηνή βίας μπορεί να διαχωριστεί από τις υπόλοιπες σκηνές υπολογίζοντας τη μέγιστη τιμή της φασματικής διακύμανσης ως προς τον μέσο όρο των απολύτων τιμών της για κάθε δείγμα, όπως φαίνεται και στην εξίσωση που ακολουθεί:

$$f_7 = \frac{\max_{j=1..W} F_j}{\frac{1}{W}\sum_{j=1..W} F_j} \quad (10)$$

Η σκηνή βίας έχει σαφώς υψηλότερο λόγο από ότι οι σκηνές που δεν περιέχουν βία, καθώς η φασματική αλλαγή μεταξύ των διαδοχικών της πλαισίων είναι εντονότερη.

3.2.2 Χαρακτηριστικά γνωρίσματα βασισμένα στο Spectral Rolloff

Spectral rolloff είναι η συχνότητα $m_c^R(j)$ στην οποία το c είναι η ποσοστιαία κατανομή της έντασης του ήχου των συντελεστών του DFT (π.χ., c =70 ή 80), που συγκεντρώνεται στο πλαίσιο j. Το Spectral rolloff υπολογίζεται ως ακολούθως:

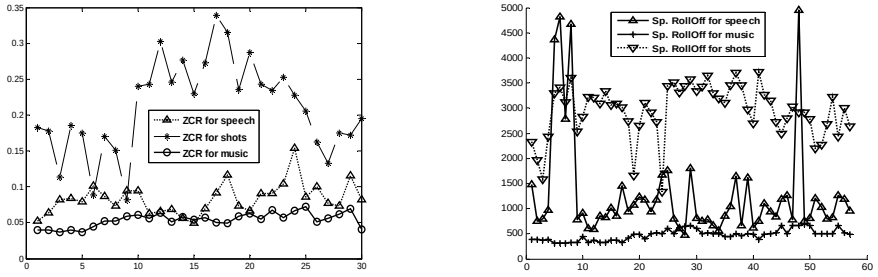
$$\sum_{k=0}^{m_c^R(j)} |X_{jk}| = \frac{c}{100} \sum_{k=0}^{S-1} |X_{jk}| \quad (11)$$

όπου x_{ji}, (i=0,1,…,S-1) είναι τα δείγματα του j-στου πλαισίου και X_{jk}, (k=0,1,…,S-1) οι αντίστοιχοι συντελεστές του DFT. Αυτό το χαρακτηριστικό αποτελεί μια εναλλακτική μέθοδο για τον υπολογισμό της φασματικής ενέργειας, όπου οι υψηλότεροι ήχοι δίνουν υψηλότερες τιμές.

Το σχετικό χαρακτηριστικό γνώρισμα που προκύπτει είναι ο λόγος της μεγίστης τιμής ως προς τον μέσο όρο των απολύτων τιμών της *Spectral Rolloff* για κάθε δείγμα.

$$f_8 = \frac{\max_{j=1..W} m_c^R}{\frac{1}{W} \sum_{j=1..W} F_j} \quad (12)$$

Η σκηνή βίας έχει σαφώς υψηλότερο λόγο από ότι οι σκηνές στις οποίες δεν εμπεριέχεται βία, καθώς η ποσοστιαία αναλογία των συντελεστών του FFT που περιέχουν C (C=70% της ενέργειας) είναι μεγαλύτερη. Στο ακόλουθο σχήμα παρουσιάζεται ένα παράδειγμα εφαρμογής του spectral roll-off με τρία διαφορετικά ηχητικά δείγματα (πυροβολισμός, μουσική και ομιλία). Τα τρία δείγματα, που φαίνονται στο παράδειγμα μας, διαφέρουν στην μέση τιμή καθώς και στην κατανομή των ‘‘κορυφών’’ κατά τη διάρκεια του χρόνου.



α. **β.**

Σχήμα 4: (α) ZCR και (β) Spectral roll-off τριών διαφορετικών ηχητικών σημάτων που περιέχουν βία (πυροβολισμός), μουσική και ομιλία

4. Ταξινόμηση

Η διαδικασία διαχωρισμού και ταξινόμησης των σκηνών βίας, υλοποιείται μέσα από ένα μηχανισμό ταξινόμησης SVM (Support Vector Machines) [6]. Είναι γνωστό πως η ιδιαιτερότητα των ταξινομητών SVM οφείλεται στην ιδιότητα τους να υπολογίζουν και να διαχωρίζουν πολυδιάστατες κλάσεις είτε αυτές είναι γραμμικά διαχωρίσιμες είτε όχι ([7]).

Η διαδικασία της ταξινόμησης είναι βασισμένη σε δύο στάδια: (Α) εκπαίδευση (training), και (Β) έλεγχος (testing).

Στο στάδιο της εκπαίδευσης δίνουμε στον ταξινομητή σαν είσοδο ένα ηχητικό σήμα και καθορίζουμε το μήκος του παραθύρου που επεξεργαζόμαστε. Σαν πρώτη φάση θα πάρουμε διάφορες διακριτές τιμές X_i. Στην συνέχεια, αυτές οι τιμές των γνωρισμάτων, θα περάσουν ως είσοδος στον ταξινομητή SVM μαζί με την παράμετρο:

$$Y_i = \begin{cases} +1, & \beta\acute{\iota}\alpha \\ -1, & \mu\eta\ \beta\acute{\iota}\alpha \end{cases}$$

που θα δηλώνει κατά πόσο αποτελούν σήμα βίας ή μη βίας. Ακολούθως, ο ταξινομητής SVM έχει σαν έξοδο ένα αριθμό από παραμέτρους (support vectors) οι οποίοι στην συνέχεια θα χρησιμοποιηθούν για την ταξινόμηση ενός άγνωστου ηχητικού σήματος.

Στο στάδιο ελέγχου, αφού πρώτα οριστεί το μήκος του παραθύρου επεξεργασίας, ο ταξινομητής δέχεται σαν είσοδο ένα ηχητικό σήμα. Σαν πρώτη φάση, θα επιστρέψει διάφορες διακριτές τιμές X_i των χαρακτηριστικών γνωρισμάτων του σήματος. Αυτές οι διακριτές τιμές των γνωρισμάτων θα περάσουν σαν είσοδος στον ταξινομητή SVM μαζί με τις τιμές (support vectors) που είχαμε σαν έξοδο από το στάδιο της εκπαίδευσης. Χρησιμοποιώντας τις παραμέτρους αυτές, ο SVM θα μπορέσει να προσδιορίσει κατά πόσο αυτές οι σκηνές αποτελούν σκηνή βίας ή όχι σύμφωνα με το κριτήριο:

$$Y_i = \begin{cases} > 0, & \beta\acute{\iota}\alpha \\ < 0, & \mu\eta\ \beta\acute{\iota}\alpha \end{cases}$$

5. Πειραματικά αποτελέσματα

Στα πλαίσια αυτής της εργασίας, για σκοπούς εκπαίδευσης και ελέγχου απόδοσης του συστήματος ταξινόμησης, δημιουργήθηκε μια βάση δεδομένων η οποία αποτελείται από ηχητικά σήματα σκηνών βίας και μη σκηνών βίας. Ο ρυθμός δειγματοληψίας των ηχητικών δειγμάτων είναι 16*KHz* και η ανάλυση των δειγμάτων είναι 16 bits. Στη βάση δεδομένων περιλαμβάνονται σκηνές από πυροβολισμούς, εκρήξεις, ξυλοδαρμούς, πολεμικές τέχνες, κραυγές και μαχαιρώματα που αποτελούν τις σκηνές βίας. Απεναντίας, οι σκηνές μη βίας αποτελούνται από μουσικά σήματα, δείγματα από ειδήσεις, ομιλίες, ταινίες, διαφημίσεις καθώς και διάφορα άλλα σήματα που δεν εμπεριέχουν βία. Ωστόσο, ήχοι από πυροτεχνήματα, αεροπλάνα, μοτοσικλέτες, σπιναρίσματα αυτοκινήτων, ελικόπτερα έχουν υιοθετηθεί με σκοπό να χρησιμοποιηθούν (στην φάση εκπαίδευσης και ελέγχου) σαν ηχητικά δείγματα παραπλήσια στην βία.

Όλα τα χαρακτηριστικά γνωρίσματα που έχουν αναφερθεί χρησιμοποιούν την τεχνική κατάτμησης του σήματος με χρήση παραθύρων. Προκειμένου να χωρίσουμε το σήμα σε διαδοχικά πλαίσια, ορίσαμε ένα παράθυρο (τετραγωνικό ή Hamming, ανάλογα με την περίπτωση), υπολογίζουμε τα χαρακτηριστικά γνωρίσματα του δείγματος του συγκεκριμένου πλαισίου και στην συνέχεια το ολισθήσαμε δεξιά κατά ένα βήμα.

Χρησιμοποιήσαμε διαφόρων μεγεθών παράθυρα για προεπεξεργασία των δειγμάτων, και εμπειρικά καταλήξαμε σε παράθυρα των 400*msec* και βήματος 200*msec* (50% επικάλυψη πλαισίου).

Στην συνέχεια, παρουσιάζονται τα αποτελέσματα της ταξινόμησης όταν τα χαρακτηριστικά γνωρίσματα χρησιμοποιούνται ξεχωριστά. Για αυτό, ο ταξινομητής SVM έχει εκπαιδευθεί και εξεταστεί 8 φορές, κάθε φορά βασισμένος σε ένα διαφορετικό χαρακτηριστικό γνώρισμα. Τα δείγματα εκπαίδευσης έχουν διάρκεια 10 λεπτών (περίπου 100 δείγματα κάθε περίπτωση) όσο περίπου και τα δείγματα ελέγχου. Τα αποτελέσματα της ταξινόμησης φαίνονται στον πίνακα 1. Εκτός από το μέσο ποσοστιαίο λάθος, παρουσιάζεται και το αρνητικό λάθος "false negative" (recall) καθώς και το θετικό λάθος "false positive" (precision). Είναι προφανές πως κανένα χαρακτηριστικό γνώρισμα από μόνο του δεν έχει γενική απόδοση καλύτερη του 80%. Ωστόσο, το ποσοστό λάθους δεν μπορεί να αποτελέσει κριτήριο κατά πόσο το συγκεκριμένο χαρακτηριστικό είναι αποτελεσματικό όταν χρησιμοποιηθεί σε συνδυασμό με άλλα ανεξάρτητα χαρακτηριστικά γνωρίσματα.

Τέλος, το σύστημα έχει εξεταστεί χρησιμοποιώντας και τα 8 χαρακτηριστικά γνωρίσματα για την εκπαίδευση του ταξινομητή (8-D). Σε αυτή την περίπτωση, το ποσοστό λάθους ταξινόμησης ανέρχεται στο *14.5%.* Ειδικότερα, το ποσοστό λάθους ταξινόμησης των σκηνών με περιεχόμενο βίας είναι σχεδόν 9.5%, ενώ το ποσοστό λάθους όταν το δείγμα εισαγωγής είναι σκηνή που δεν περιείχε βία (false alarm) κυμαίνεται στο 19.5% (λόγω παραπλήσιων σκηνών βίας). Με άλλα λόγια, η ακρίβεια (precision) του ταξινομητή, δηλαδή το ποσοστό των δειγμάτων που χαρακτηρίζονται από τον ταξινομητή ως βίαια και στην πραγματικότητα είναι βίαια, ανέρχεται στο 80.5%. Απεναντίας, η ανάκληση (recall) από τον τελικό ταξινομητή, δηλαδή το ποσοστό των βίαιων τμημάτων που ταξινομήθηκαν ως βία, είναι 90.5%. Επομένως, η μέση ακρίβεια του ταξινομητή υπολογίστηκε ως 85.5%.

Χαρακτηριστικό γνώρισμα	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	**8-D**
% False Negative	9	12.5	13.5	12	9	11	10	12.5	**4.75**
% False Positive	13.5	13	15	17	12.5	12.5	10.5	12	**9.75**
% Ποσοστό λάθους	22.5	25.5	28.5	29	21.5	23.5	20.5	24.5	**14.5**

Πίνακας 1: Ποσοστό λάθους ταξινόμησης για κάθε χαρακτηριστικό γνώρισμα

6. Συμπεράσματα και μελλοντική εργασία

Σε αυτήν την εργασία, χρησιμοποιήθηκαν μερικά από τα δημοφιλέστερα ηχητικά χαρακτηριστικά πλαισίου (frame-level features) για την εξαγωγή των χαρακτηριστικών γνωρισμάτων της βίας, τα οποία εισήρθαν σε ένα ταξινομητή SVM για να δημιουργηθεί ένα ολοκληρωμένο και αξιόπιστο σύστημα εντοπισμού των σκηνών βίας. Κατά μέσο όρο, το 85.5% των ηχητικών δειγμάτων που εξετάστηκε έχει ταξινομηθεί σωστά, ενώ το ποσοστό λάθους του συστήματος ταξινόμησης, όταν το δείγμα εισαγωγής είναι σκηνή βίας, είναι μόλις 9.5%. Στη μελέτη αυτή, οι σκηνές βίας εντοπίστηκαν με χρήση μόνο ηχητικής

πληροφορίας, έχοντας ένα υψηλό ποσοστό ταξινόμησης. Αντιθέτως, έρευνες με ανάλογα ποσοστά ταξινόμησης (δες [2], [3], [4]), εστιάστηκαν κυρίως στην οπτική πληροφορία, χρησιμοποιώντας τα ηχητικά γνωρίσματα μόνο σαν πρόσθετη πληροφορία.

Η παρούσα μελέτη θα μπορούσε να αποτελέσει ισχυρή βάση για μελλοντική έρευνα. Καταρχάς, η προσθήκη περαιτέρω ηχητικών γνωρισμάτων δίνουν ένα πληρέστερο σύνολο δεδομένων που χαρακτηρίζουν τη βία. Σε αυτή την κατεύθυνση, το υπερεπίπεδο διαχωρισμού θα μπορούσε να εμπλουτιστεί με τα γνωρίσματα των Mel-frequency cepstral coefficients (MFCCs) και Linear Prediction Coefficients (LPCs). Μια άλλη προσέγγιση είναι η υιοθέτηση αποδοτικότερων αλγόριθμων ταξινόμησης, όπως Hidden Markov Models (HMMs). Επίσης, η εξειδικευμένη κατηγοριοποίηση και χαρακτηρισμός των ηχητικών δειγμάτων σαν έκρηξη, πυροβολισμός, κραυγή αντί ο απλός διαχωρισμός τους ως βία και μη βία αποτελεί μια ενδιαφέρουσα κατεύθυνση έρευνας. Παράλληλα, για τον εντοπισμό της βίας σε μεγαλύτερα ηχητικά δείγματα (π.χ. ηχητικό δείγμα μιας ολόκληρης ταινίας), απαιτείται να εφαρμοστεί ένας ισχυρός αλγόριθμος κατάτμησης, ο οποίος θα διαιρεί το ηχητικό σήμα σε πιο μικρά τμήματα ομοιογενών χαρακτηριστικών.

Το ποσοστό αναγνώρισης βελτιώνεται με την πολυμορφική επεξεργασία των δειγμάτων, όπου υπάρχει η δυνατότητα χρησιμοποίησης και της οπτικής πληροφορίας του σήματος. Τα χαρακτηριστικά γνωρίσματα που λαμβάνονται από έναν αλγόριθμο ανάλυσης των οπτικών πληροφοριών, όπως η λάμψη από πυροβολισμό ή έκρηξη, η σύντομη αύξηση των επιπέδων του κόκκινου χρώματος από την παρουσία αίματος σε συνδυασμό με τα ηχητικά γνωρίσματα, μπορούν να αποτελέσουν ένα αξιόπιστο σύστημα εντοπισμού της βίας. Συγκεκριμένα, η πολυμορφική επεξεργασία καθώς και η χρήση συνδυασμού ταξινομητών προτείνεται ως μια σημαντική μελλοντική ερευνητική κατεύθυνση.

Αναφορές – Βιβλιογραφία

1. Cees G.M. Snoek, Marcel Worring, Multimodal Video Indexing: A Review of the State-of-the-art. Multimedia Tools and Applications, 25(1), 5-35, 2005.
2. Vasconcelos N., Lippman A. Towards semantically meaningful feature spaces for the characterization of video content Image Processing, 1997. Proceedings., International Conference on Volume 1, Date: 26-29 Oct 1997, Pages: 25 - 28 vol.1
3. A. Datta, M. Shah, N. V. Lobo. "Person-on-Person Violence Detection in Video Data", IEEE International Conference on Pattern Recognition, Canada, 2002.
4. J. Nam, A.H. Tewfik, “Event-driven video abstraction and visualisation”, Multimedia Tools and Applications, 16(1-2), 55-77, 2002
5. George Tzanetakis, Georg Essl, and Perry Cook Automatic Musical Genre Classification of Audio Signals In. Proc. Int. Symposium on Music Information Retrieval (ISMIR), Bloomington, Indiana, 2001
6. Sergios Theodoridis, Konstantinos Koutroumbas, Pattern Recognition. Academic Press, 2005, 3rd Edtition.
7. N. Cristianini and J. Shawe-Taylor, "Support Vector Machines & other kernel-based learning methods", Cambridge University Press, ISBN 0-521-78019-5, 2000.