ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σύνθεση και ανίχνευση πλαστών (deep fake) βίντεο με αλγόριθμους βαθιάς μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΡΟΣΟΥ ΜΑΡΙΑ
Α.Μ. 151046

Επιβλέπων Καθηγητής: Αθανάσιος Βουλόδημος
Επίκουρος Καθηγητής ΠΑ.Δ.Α

Αθήνα, Ιούλιος 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σύνθεση και ανίχνευση πλαστών (deep fake) βίντεο με αλγόριθμους βαθιάς μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΡΟΣΟΥ ΜΑΡΙΑ
Α.Μ. 151046

Επιβλέπων καθηγητής: Αθανάσιος Βουλόδημος
Επίκουρος καθηγητής ΠΑ.Δ.Α

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Ιουλίου του 2021.

Athanasios Voulodimos Digitally signed by Athanasios Voulodimos Date: 2021.07.19 19:00:51 +03'00'	PARIS MASTOR OKOSTAS Digitally signed by PARIS MASTOROKOSTAS Date: 2021.07.19 19:22:40 +03'00'	Anastasios Kesidis Digitally signed by Anastasios Kesidis Date: 2021.07.19 22:31:31 +03'00'
.....	.....	.....
Αθανάσιος Βουλόδημος Επ. Καθηγητής ΠΑ.Δ.Α.	Πάρις Μαστοροκώστας Καθηγητής ΠΑ.Δ.Α.	Αναστάσιος Κεσίδης Αν. Καθηγητής ΠΑ.Δ.Α.

Αθήνα, Ιούλιος 2021

Δήλωση Συγγραφέα Διπλωματικής Εργασίας

Η κάτωθι υπογεγραμμένη Μαρία Δρόσου του Σπυρίδωνος, με αριθμό μητρώου 711151046 φοιτήτρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Η δηλούσα

Περίληψη

Η σύνθεση πλαστών (deep fake) βίντεο γίνεται με την εφαρμογή μηχανικής μάθησης και τεχνητής νοημοσύνης. Τα βίντεο αυτά δύναται να χρησιμοποιηθούν με καλές προθέσεις, όπως για παράδειγμα ως χιουμοριστικά βίντεο. Σε κάποιες περιπτώσεις όμως η χρήση τους μπορεί να είναι κακόβουλη, να έχουν δηλαδή στόχο την εξαπάτηση μέσω της προβολής τους ως δήθεν πραγματικά βίντεο. Λόγω της δυνητικά μεγάλης επιρροής που μπορούν να ασκήσουν τα βίντεο αυτά στη δημόσια σφαίρα, είναι αναγκαία η ανάπτυξη μοντέλων που μπορούν να ταυτοποιούν τέτοιες περιπτώσεις.

Στα πλαίσια της παρούσας εργασίας εισαγωγικά επισημαίνονται οι θετικοί και αρνητικοί τρόποι χρήσης τεχνολογιών σύνθεσης πλαστών βίντεο και εικόνων. Στη συνέχεια προσδιορίζονται οι κατηγορίες πλαστών βίντεο και παρουσιάζονται υπάρχουσες εφαρμογές σύνθεσής τους.

Ακολούθως αναφέρονται μέθοδοι που έχουν αναπτυχθεί ως τώρα με στόχο την ανίχνευσή πλαστών βίντεο. Οι μέθοδοι αυτοί κατηγοριοποιούνται σε μεθόδους που λαμβάνουν υπόψη την χρονική πληροφορία, δηλαδή την αλλαγή των χαρακτηριστικών μέσα σε μια αλληλουχία στιγμιότυπων του βίντεο, και σε μεθόδους που βασίζονται αποκλειστικά στη χωρική πληροφορία που εξάγεται από το κάθε στιγμιότυπο.

Τέλος, παρουσιάζεται η δομή τεσσάρων μοντέλων μηχανικής μάθησης, του R3D, του MC3, του R2Plus1D και του I3D. Στα πλαίσια της παρούσας εργασίας τα μοντέλα αυτά εκπαιδεύτηκαν στα δείγματα του συνόλου Celeb-DF-v2, με στόχο να ταξινομούν βίντεο σε πλαστά ή αυθεντικά. Τα αποτελέσματά τους παρουσιάζονται, αξιολογούνται και συγκρίνονται ως προς την ικανότητα ανίχνευσης πλαστών (deep fake) βίντεο.

Λέξεις κλειδιά

Όραση υπολογιστών, Σύνθεση πλαστών βίντεο, Ανίχνευση πλαστών βίντεο, γενετικά ανταγωνιστικά δίκτυα, αυτόματοι κωδικοποιητές, μηχανική μάθηση, συνελικτικά νευρωνικά δίκτυα

Abstract

Deep fake video generation uses machine learning and artificial intelligence. The synthesized videos can be used with good intentions, such as humorous videos. In some cases, however, their use can be malicious. That is when they aim to deceive through their promotion as supposedly real videos. Due to the potentially great influence that these videos can have on the public sphere, it is necessary to develop models that can identify such cases.

In the context of this paper, the positive and negative ways of using deepfake video and image generation technologies are pointed out. The categories of fake videos are then identified and existing deepfake video generation algorithms are presented.

After that, methods that have been developed for deepfake video detection are referenced. These techniques are categorized into methods that take into account temporal information, which is the change of features within a sequence of video frames, and into methods that rely solely on the spatial information extracted from each frame.

Finally, the architecture of four machine learning models is presented. These are the R3D, MC3, R2Plus1D and I3D models. In this present dissertation, these models were trained in the Celeb-DF-v2 dataset, with the aim of classifying videos as fake or authentic. Their results are presented, evaluated and compared in terms of the ability to detect deepfake videos.

Keywords

Computer Vision, Deepfake Video Generation, Deepfake Video Detection, Generative Adversarial Networks – GAN, Autoencoders, Machine Learning, Convolutional Neural Networks – CNN

Περιεχόμενα

Κατάλογος εικόνων

1. Εισαγωγή

Η σύνθεση πλαστών (deepfake) βίντεο γίνεται με τεχνικές μηχανικής μάθησης οι οποίες παραλλάσουν ένα βίντεο με σκοπό να φαίνεται ότι στο βίντεο συμβαίνει κάτι διαφορετικό από την πραγματικότητα. Σε κάποιες περιπτώσεις η χρήση τέτοιων τεχνικών γίνεται με καλές προθέσεις, όπως, για παράδειγμα, όταν χρησιμοποιούνται για την σύνθεση οπτικών εφέ, ψηφιακών μορφών (avatars) ή φίλτρων που χρησιμοποιούνται σε εφαρμογές ψυχαγωγίας, όπως το snapchat. Επιπλέον μια θετική εφαρμογή τους είναι η επεξεργασία σκηνών ταινιών για τυχόν διορθώσεις και αλλαγές, κάτι που γλιτώνει τους παραγωγούς ταινιών από το να γυρίσουν ξανά τις αντίστοιχες σκηνές [1].

Υπάρχουν όμως και πολλές περιπτώσεις κακόβουλης χρήσης αυτών των μεθόδων. Η πρώτη φορά που αυτό έγινε φανερό ήταν το 2017, όταν διαδόθηκε στο διαδίκτυο βίντεο με το πρόσωπο μιας διάσημης στη θέση του προσώπου μιας ηθοποιού ερωτικών ταινιών σε αντίστοιχο βίντεο. Άλλες αρνητικές εφαρμογές αυτών των τεχνικών είναι η παραγωγή βίντεο στα οποία ηγέτες, πρωθυπουργοί ή άλλα δημόσια πρόσωπα παριστάνονται να βγάζουν πλαστούς λόγους, οι οποίοι παρουσιάζονται ως πραγματικοί με στόχο την παραπλάνηση του κόσμου [2, 3, 4]. Με αυτόν τον τρόπο τα deepfake βίντεο μπορούν να προκαλέσουν πολιτικές και θρησκευτικές εντάσεις ή να εξαπατήσουν τον κόσμο και να επηρεάσουν τα εκλογικά αποτελέσματα [5].

Επιπλέον η διαδικασία που απαιτείται για να δημιουργηθεί ένα deepfake βίντεο γίνεται σταδιακά όλο και πιο απλή και αποτελεσματική, φτάνοντας στο σημείο να καθίσταται κάτι τέτοιο δυνατό ακόμη και με βάση μια μόνο εικόνα του ατόμου το οποίο πρόκειται να αποτελέσει το πρόσωπο ενός πλαστού βίντεο [6]. Λόγω αυτού και δεδομένων των κινδύνων που προαναφέρθηκαν καθίσταται αναγκαία η εύρεση μεθόδων ανίχνευσης πλαστών (deepfake) βίντεο. Συνεπώς τα τελευταία χρόνια έχουν αναπτυχθεί τεχνικές και μοντέλα μηχανικής μάθησης που εντοπίζουν ατέλειες των πλαστών βίντεο και χαρακτηριστικά τα οποία μπορούν να οδηγήσουν σε όσο το δυνατόν ασφαλές συμπέρασμα σχετικά με το εάν ένα βίντεο είναι πραγματικό ή όχι. Σύμφωνα με δεδομένα από την ιστοσελίδα https://app.dimensions.ai, φαίνεται ότι ο αριθμός των δημοσιεύσεων σχετικά με τον τομέα του deepfake έχει αυξηθεί ραγδαία τα τελευταία χρόνια (Εικόνα 1). Αν και λογικά οι δημοσιεύσεις είναι ακόμη περισσότερες, και πάλι είναι εμφανής η αυξανόμενη τάση που υπάρχει στις δημοσιεύσεις πάνω σε αυτό το θέμα [7].

Εικόνα 1: Αριθμός δημοσιεύσεων σε σχέση με τα deepfakes από το έτος 2016 ως το έτος 2020, όπως προκύπτουν από την ιστοσελίδα https://app.dimensions.ai, στην οποία έγινε αναζήτηση με το λήμμα «deepfake» στις 12 Μαΐου του 2021.

Προκειμένου να είναι εφικτή η ανάπτυξη μοντέλων μηχανικής μάθησης τα οποία εντοπίζουν τα πλαστά βίντεο, είναι χρήσιμη η ύπαρξη συνόλων δεδομένων που περιλαμβάνουν αρκετά βίντεο πλαστά και μη, με στόχο τη χρήση τους για την εκπαίδευση και την αξιολόγηση των μοντέλων. Μερικά γνωστά σύνολα δεδομένων είναι το FaceForencics++ [8] και το Celeb-DF [9].

1.1 Κίνητρο

Δεδομένου του κινδύνου από την κακόβουλη χρήση πλαστών (deepfake) βίντεο, είναι πολύ σημαντική η ανάπτυξη εφαρμογών που μπορούν να αξιολογήσουν με επιτυχία τη γνησιότητα ενός βίντεο. Στον τομέα αυτό έχουν υπάρξει διάφορες προσεγγίσεις. Παρ' όλα αυτά η συνεχής εξέλιξη και αναβάθμιση των μεθόδων σύνθεσης πλαστών (deepfake) βίντεο έχει ως συνέπεια τα ελαττώματα των αποτελεσμάτων τους να μειώνονται όλο και περισσότερο. Αυτό οδηγεί στην ανάγκη συνεχούς ανάπτυξης νέων μεθόδων ανίχνευσης πλαστών (deepfake) βίντεο που θα μπορούν να αντιμετωπίζουν νέες εξελιγμένες μορφές παραποιημένων βίντεο.

Η ανάγκη αυτή αποτέλεσε κίνητρο για την παρούσα εργασία, η οποία στόχο έχει να εκπαιδεύσει και να αξιολογήσει υπάρχουσες δομές συνελικτικών νευρωνικών δικτύων ως ταξινομητές που διαχωρίζουν τα πλαστά από τα γνήσια βίντεο.

1.2 Συμβολή

Το κύριο αντικείμενο της παρούσας διπλωματικής εργασίας είναι η ερευνητική ενασχόληση με τη σύνθεση και την ανίχνευση πλαστών (deepfake) βίντεο. Για τον σκοπό αυτό πραγματοποιήθηκαν τα ακόλουθα:

- Συγκέντρωση και παρουσίαση των πιο αξιοσημείωτων εφαρμογών σύνθεσης πλαστών (deepfake) βίντεο, αλλά και των αντίστοιχων μεθόδων ανίχνευσης πλαστών (deepfake) βίντεο, που έχουν αναπτυχθεί μέχρι σήμερα.

- Παραγωγή πηγαίου κώδικα σε γλώσσα Python για την εκπαίδευση των μοντέλων R3D [10], MC3 [11], R2Plus1D [11] και I3D [12] στα δεδομένα του συνόλου Celeb-DF-v2 [9], με στόχο την επιτυχή ταξινόμηση βίντεο σε πλαστά και σε αυθεντικά.

- Αξιολόγηση των παραπάνω μοντέλων με βάση τη συνολική ακρίβεια (Accuracy) και το εμβαδόν της καμπύλης λειτουργικών χαρακτηριστικών (Area Under the Curve of Receiver Characteristic Operator (AUC-ROC)).

1.3 Δομή εργασίας

Η παρούσα διπλωματική εργασία δομείται σε 5 κεφάλαια. Πέρα από το πρώτο κεφάλαιο που είναι η Εισαγωγή, το δεύτερο κεφάλαιο περιλαμβάνει το θεωρητικό υπόβαθρο για την σύνθεση πλαστών (deepfake) βίντεο, καθώς και τις υπάρχουσες εφαρμογές που έχουν αναπτυχθεί στο πεδίο αυτό. Το τρίτο κεφάλαιο περιλαμβάνει τις αντίστοιχες πληροφορίες σχετικά με τον τομέα της ανίχνευσης πλαστών (deepfake) βίντεο. Ακολούθως στο τέταρτο κεφάλαιο παρουσιάζεται το τεχνικό κομμάτι της εργασίας, δηλαδή η δομή των τεσσάρων μοντέλων που εκπαιδεύτηκαν και τα αποτελέσματά τους. Τέλος, το πέμπτο κεφάλαιο είναι ο επίλογος, όπου συνοψίζονται τα παραπάνω και προκύπτουν συμπεράσματα.

2. Σύνθεση πλαστών (deepfake) βίντεο

2.1 Κατηγορίες σύνθεσης πλαστών βίντεο

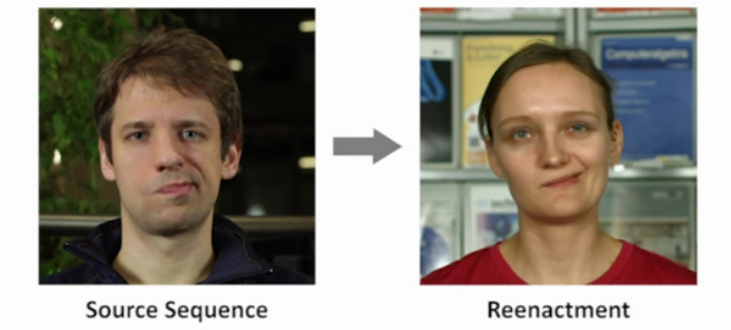
Η σύνθεση πλαστών βίντεο χωρίζεται σε τρεις κατηγορίες. Η πρώτη κατηγορία ονομάζεται face-swap και συνίσταται στην επικάλυψη του προσώπου σε ένα βίντεο από ένα άλλο πρόσωπο, έτσι ώστε το νέο πρόσωπο να φαίνεται ότι λέει και πράττει ό,τι έκανε το αρχικό πρόσωπο (Εικόνα 2).

Face-Swap

Εικόνα 2: Παράδειγμα σύνθεσης βίντεο με αντικατάσταση του προσώπου που παρουσιάζεται σε ένα βίντεο με ένα άλλο (face-swap). Πηγή εικόνας: https://electricalfundablog.com/deepfake-faceswap/ (Τελευταία πρόσβαση 19/5/2021)

Η δεύτερη κατηγορία είναι τα puppet-master. Στα βίντεο αυτά προσαρμόζεται η κινησιολογία του εικονιζόμενου προσώπου (puppet), με βάση τις κινήσεις του κεφαλιού, τις εκφράσεις του προσώπου και τις κινήσεις των ματιών ενός άλλου προσώπου (master), το οποίο κάθεται μπροστά σε μια κάμερα και φέρεται με τον τρόπο που επιθυμεί να φέρεται το αρχικό πρόσωπο στο βίντεο (Εικόνα 3).

Puppet-Master



Εικόνα 3: Παράδειγμα συγχρονισμού των κινήσεων του προσώπου στόχου (puppet) με βάση τις κινήσεις κάποιου άλλου προσώπου (master). Πηγή: https://medium.com/gradientcrescent/ai-truth-and-society-deepfakes-at-the-front-of-the-technological-cold-war-86c3b5103ce6 (Τελευταία πρόσβαση 19/5/2021)

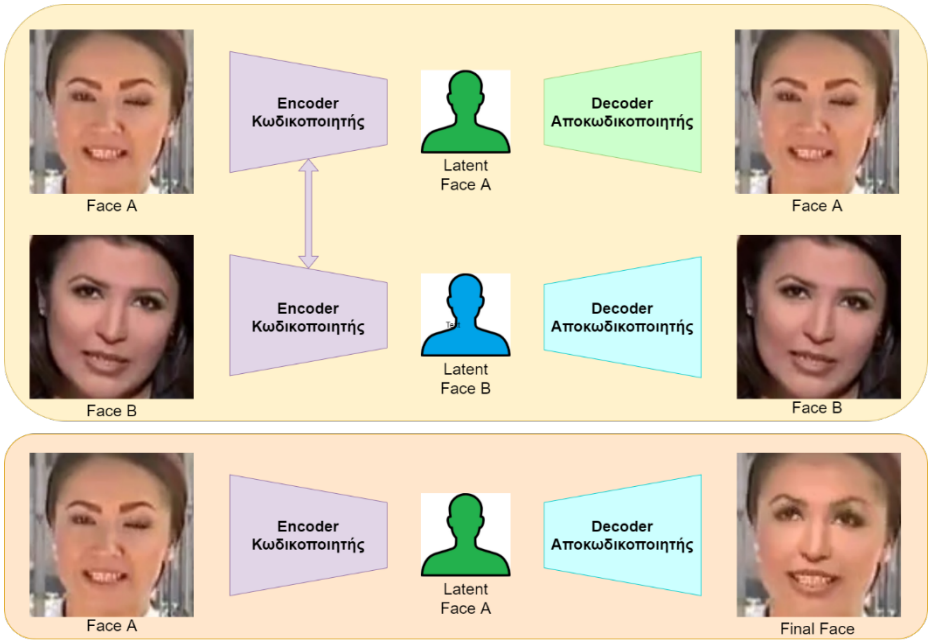
Η τρίτη κατηγορία σύνθεσης πλαστών βίντεο είναι τα lip-sync. Πρόκειται για βίντεο στα οποία έχει γίνει επεξεργασία, έτσι ώστε να συγχρονίζονται οι κινήσεις του στόματος ενός προσώπου σύμφωνα με συγκεκριμένο ηχητικό, για παράδειγμα μια καταγεγραμμένη ομιλία (Εικόνα 4). [13].

Εικόνα 4: Παράδειγμα σύνθεσης πλαστού βίντεο με συγχρονισμό των χειλιών με βάση ηχητικό ομιλίας (lip-sync). Πηγή: https://www.youtube.com/watch?v=Ic0TBhfuOrA&ab_channel=WhatMakeArt (Τελευταία πρόσβαση 18/5/2021)

2.3 Εφαρμογές σύνθεσης πλαστών βίντεο

Για τις παραπάνω μεθόδους παραγωγής πλαστών βίντεο χρησιμοποιούνται κυρίως δίκτυα βαθιάς μηχανικής μάθησης, όπως αυτόματοι κωδικοποιητές (autoencoders) [14] και γενετικά ανταγωνιστικά δίκτυα (generative adversarial networks – GANs), τα οποία μαθαίνουν την κινησιολογία ενός προσώπου και την αναπαράγουν σε ένα άλλο πρόσωπο [15]. Η πρώτη αξιοσημείωτη προσπάθεια σύνθεσης deepfake βίντεο ήταν το FakeApp [16]. Σε αυτή την περίπτωση χρησιμοποιήθηκε αυτόματος κωδικοποιητής με στόχο την αντικατάσταση ενός προσώπου από ένα άλλο σε μια εικόνα (FaceSwap) [17]. Η διαδικασία που ακολουθείται από αυτή τη μέθοδο είναι η ακόλουθη:

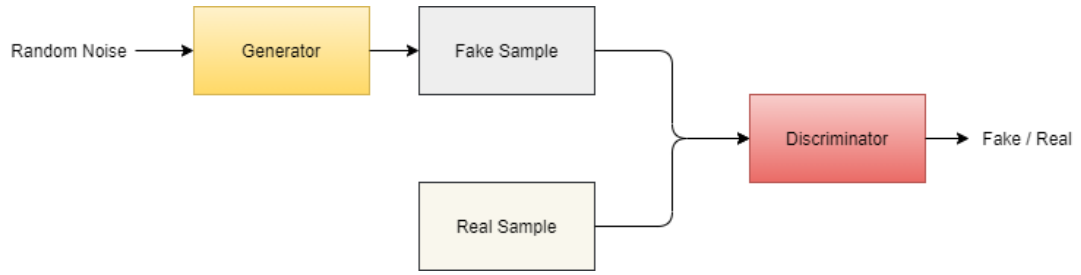
Αρχικά χρησιμοποιείται αυτόματος κωδικοποιητής για καθένα από τα δύο πρόσωπα. Ο αυτόματος κωδικοποιητής στο πρώτο μέρος του κωδικοποιεί το πρόσωπο σε μια μορφή που περιλαμβάνει πολύ λιγότερες πληροφορίες από την αρχική εικόνα του προσώπου, αλλά παρ' όλα αυτά οι πληροφορίες αυτές είναι αρκετές για την ανακατασκευή της αρχικής εικόνας στο δεύτερο κομμάτι του μοντέλου που γίνεται η αποκωδικοποίηση. Έτσι για κάθε πρόσωπο υπάρχει ο κωδικοποιητής που το μετατρέπει σε μια μορφή πιο αφηρημένη (latent face) και ένας αποκωδικοποιητής που παίρνει την αφηρημένη μορφή και αναπαράγει με βάση αυτή το πρόσωπο. Ο κωδικοποιητής των δύο προσώπων είναι κοινός, ενώ ο αποκωδικοποιητής ξεχωριστός για κάθε πρόσωπο. Η αλλαγή μεταξύ των δύο προσώπων γίνεται εφαρμόζοντας τον κοινό κωδικοποιητή στο πρώτο πρόσωπο, με αποτέλεσμα την παραγωγή της αφηρημένης εκδοχής αυτού του προσώπου, αλλά αποκωδικοποιώντας στη συνέχεια με τον αποκωδικοποιητή του δεύτερου προσώπου. Ο κοινός αυτόματος κωδικοποιητής έχει ως αποτέλεσμα να εντοπίζονται τα κοινά στοιχεία των δύο προσώπων, δηλαδή τα μάτια, η μύτη, το στόμα και γενικά στοιχεία που λίγο πολύ υπάρχουν σε όλα τα πρόσωπα. Αυτά τα κοινά στοιχεία αντιστοιχούνται ουσιαστικά από το ένα στο άλλο πρόσωπο μέσω της αποκωδικοποίησης του αφηρημένου προσώπου με τον αποκωδικοποιητή του άλλου προσώπου. Η διαδικασία αυτή φαίνεται παραστατικά στην Εικόνα 5.



Εικόνα 5: Εικονική αναπαράσταση της μεθόδου FaceSwap

Η παραπάνω προσέγγιση εφαρμόζεται και σε άλλες προσπάθειες σύνθεσης deepfake βίντεο, όπως για παράδειγμα στο DeepFaceLab [18, 19], το οποίο τη διευρύνει με νέα μοντέλα, όπως το H64, το H128, το LIAEF128 και το SAE και παράλληλα υποστηρίζει πολλαπλές μεθόδους εξαγωγής του προσώπου, όπως το S3FD, το MTCNN, και το dlib, αλλά και την επιλογή του προσώπου χειροκίνητα [20]. Επιπλέον η μέθοδος FaceSwap εφαρμόζεται και στα μοντέλα DFaker [21] και DeepFake_tf [22], το οποία είναι ουσιαστικά μεταξύ τους ίδια, απλά το πρώτο είναι υλοποιημένο σε keras και το δεύτερο σε tensorflow.

Προκειμένου να βελτιωθούν τα αποτελέσματα της παραπάνω μεθόδου ενσωματώθηκαν στην αρχιτεκτονική του προηγούμενου μοντέλου, ως συναρτήσεις σφάλματος, το σφάλμα ανταγωνισμού (adversarial loss) και το σφάλμα αντίληψης (perceptual loss) που υλοποιείται με το VGGFace [23]. Με αυτόν τον τρόπο προέκυψαν νέα μοντέλα βασισμένα στα γενετικά ανταγωνιστικά δίκτυα (Generative Adversarial Network GAN) [24]. Τα γεννητικά ανταγωνιστικά δίκτυα (generative adversarial networks - GAN) αποτελούνται από δύο διακριτά νευρωνικά δίκτυα τα οποία ανταγωνίζονται μεταξύ τους. Το ένα δίκτυο ονομάζεται γεννήτορας (generator) και αποτελεί το γεννητικό δίκτυο το οποίο είναι υπεύθυνο για τη σύνθεση νέων πλαστών δειγμάτων. Το άλλο δίκτυο ονομάζεται διευκρινιστής (discriminator) και αποτελεί το διευκρινιστικό δίκτυο, του οποίου σκοπός είναι να αναγνωρίζει αν ένα δείγμα είναι πλαστό, προέρχεται δηλαδή από το σύνολο των δειγμάτων που συνθέτει ο γεννήτορας ή αν είναι αυθεντικό, προέρχεται δηλαδή από το σύνολο των πραγματικών δεδομένων. Στην Εικόνα 6 απεικονίζεται διαγραμματικά το βασικό GAN όπως προτάθηκε από τον Ian Goodfellow και τους συνεργάτες του [24].



Εικόνα 6: Διαγραμματική απεικόνιση του βασικού GAN

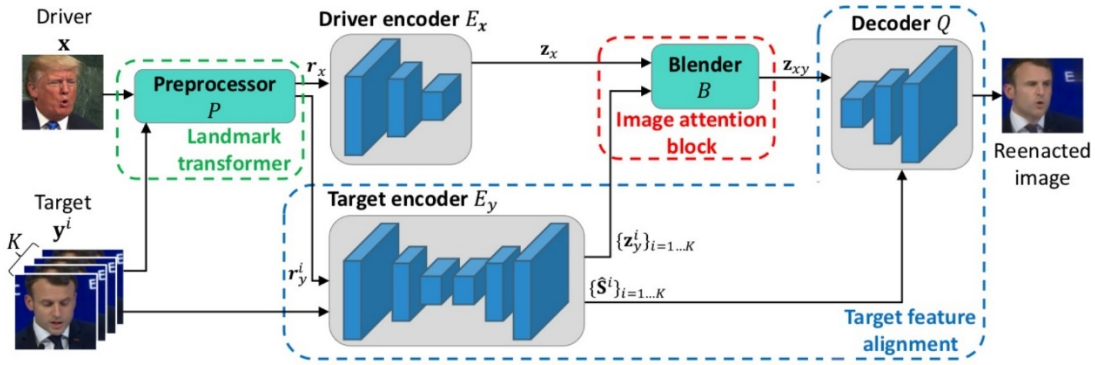
Στην παραπάνω αρχιτεκτονική έχουν γίνει πολλές διαφορετικές προσθήκες σε διάφορες προσεγγίσεις σύνθεσης πλαστών εικόνων και βίντεο, με στόχο καλύτερα αποτελέσματα. Ένα από τα μοντέλα που χρησιμοποιούν αυτή τη μέθοδο είναι το faceswap-GAN [25]. Στο μοντέλο αυτό με την προσθήκη του σφάλματος αντίληψης του VGGFace επιτυγχάνεται περισσότερη ρεαλιστικότητα στις κινήσεις των ματιών και ελάττωση των ατελειών στη συνοχή της εικόνας του προσώπου που προκύπτει. Με αυτό τον τρόπο το αποτέλεσμα είναι καλύτερο από αυτό του απλού ζεύγους κωδικοποιητή και αποκωδικοποιητή και παράλληλα υποστηρίζονται μεγαλύτερες αναλύσεις βίντεο.

Μια ακόμη αξιοσημείωτη εφαρμογή σύνθεσης πλαστών βίντεο είναι το Few-Shot Face Translation [26]. Η εφαρμογή αυτή εντάσσει στη διαδικασία τη χρήση προεκπαιδευμένου μοντέλου αναγνώρισης προσώπου, προκειμένου να εξάγει αφηρημένες (latent) εισόδους για επεξεργασία από το γενετικό ανταγωνιστικό δίκτυο (GAN). Επιπλέον ενσωματώνει σημασιολογικά δεδομένα που αντλούνται από δομές των AdaIN [27] και SPADE [28]. Με αυτόν τον τρόπο κατορθώνει ακόμη καλύτερα αποτελέσματα, χωρίς να χρειάζεται μεγάλο αριθμό διαθέσιμων εικόνων του προσώπου που πρόκειται να ανακατασκευαστεί.

Επίσης ενδιαφέρουσα είναι και η εφαρμογή AvatarMe [29], η οποία αναδημιουργεί πρόσωπα τριών διαστάσεων (3D) από τυχαίες εικόνες «στον έξω κόσμο». Η εφαρμογή αυτή έχει το πλεονέκτημα ότι μπορεί να ανακατασκευάσει αυθεντικά τρισδιάστατα πρόσωπα ανάλυσης 4K-6K από μία μόνο χαμηλής ανάλυσης εικόνα.

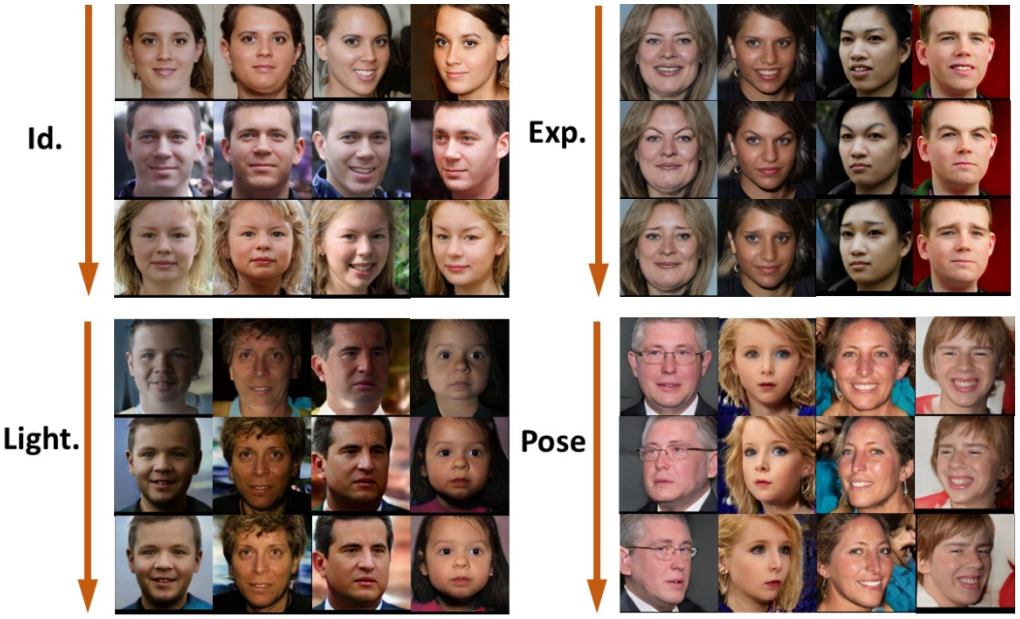
Επιπλέον μια μέθοδος σύνθεσης πλαστών βίντεο η οποία αξίζει να αναφερθεί είναι το MarioNETte [30]. Η προσέγγιση αυτή δημιουργεί αναπαράσταση προσώπου βασιζόμενη σε λίγες λήψεις, διατηρώντας όμως παράλληλα την ταυτότητα του προσώπου-στόχου. Το σημαντικότερο στοιχείο αυτής της εφαρμογής είναι ότι δεν χρειάζεται καμία φάση επιπλέον βελτίωσης προκειμένου να γίνει προσαρμογή της αναπαράστασης στην ταυτότητα του στόχου. Στην Εικόνα 7 φαίνεται το διάγραμμα του μοντέλου που χρησιμοποιήθηκε σε αυτή την προσέγγιση. Τα νέα στοιχεία της δομής του MarioNETte [30] είναι τρία. Το πρώτο είναι το επίπεδο landmark transformer (μετασχηματισμός σημείων αναφοράς), το οποίο διακρίνει τα χαρακτηριστικά που διαφοροποιούν τη δομή των δύο προσώπων σε δομικά χαρακτηριστικά-σημεία αναφοράς που σχετίζονται με τη γεωμετρία της ταυτότητας του προσώπου και σε αυτά που σχετίζονται με τη γεωμετρία της έκφρασης του προσώπου. Το δεύτερο είναι το επίπεδο image attention block (δομή ενημέρωσης της εικόνας), το οποίο είναι υπεύθυνο για την αποτελεσματική ανάμειξη των πληροφοριών στυλ που προέρχονται από τις πολλαπλές εικόνες του προσώπου στόχου. Τέλος το τρίτο στοιχείο είναι το επίπεδο target feature alignment (ευθυγράμμιση των χαρακτηριστικών του προσώπου στόχου), το οποίο επιτρέπει στο μοντέλο να εισάγει στην παραγόμενη εικόνα λεπτομερείς πληροφορίες στυλ των εικόνων του προσώπου στόχου. Με αυτές τις σημαντικές προσθήκες η εφαρμογή

MarioNETte [30] κατορθώνει τη διατήρηση της ταυτότητας του προσώπου στόχου στο παραγόμενο αποτέλεσμα, χρησιμοποιώντας λίγες μόνο λήψεις.



Εικόνα 7: Διαγραμματική απεικόνιση της εφαρμογής MarioNETte [30]

Μια από τις μεγαλύτερες δυσκολίες στη σύνθεση πλαστών βίντεο είναι η διαχείριση των διαφορετικών φωτισμών, καθώς και της ποικιλίας των στάσεων, των εκφράσεων και των στοιχείων της ταυτότητας του προσώπου. Αυτή τη δυσκολία αντιμετωπίζει ικανοποιητικά το μοντέλο DiscoFaceGAN [31], το οποίο συνθέτει πρόσωπα εικονικών ατόμων με ανεξάρτητες λανθάνουσες μεταβλητές ταυτότητας, έκφρασης, στάσης και φωτισμού (Εικόνα 8).



Εικόνα 8: Παραδείγματα εναλλαγής ταυτότητας, φωτισμού, έκφρασης και στάσης προσώπων με χρήση της εφαρμογής DiscoFaceGAN [31]

Το μοντέλο αυτό επιπλέον ενσωματώνει τα 3D προηγούμενα (priors) στην ανταγωνιστική μάθηση, ενώ η ρύθμιση του φωτισμού, της έκφρασης και της στάσης ενός προσώπου μπορεί

να γίνει σύμφωνα με τις αντίστοιχες μεταβλητές ενός άλλου προσώπου (face reenactment). Παραδείγματα τέτοιας μεταφοράς παρουσιάζονται στην Εικόνα 9.

Εικόνα 9: Παραδείγματα μεταφοράς έκφρασης, στάσης και φωτισμού από ένα πρόσωπο σε ένα άλλο με χρήση της εφαρμογής DiscoFaceGAN [31]

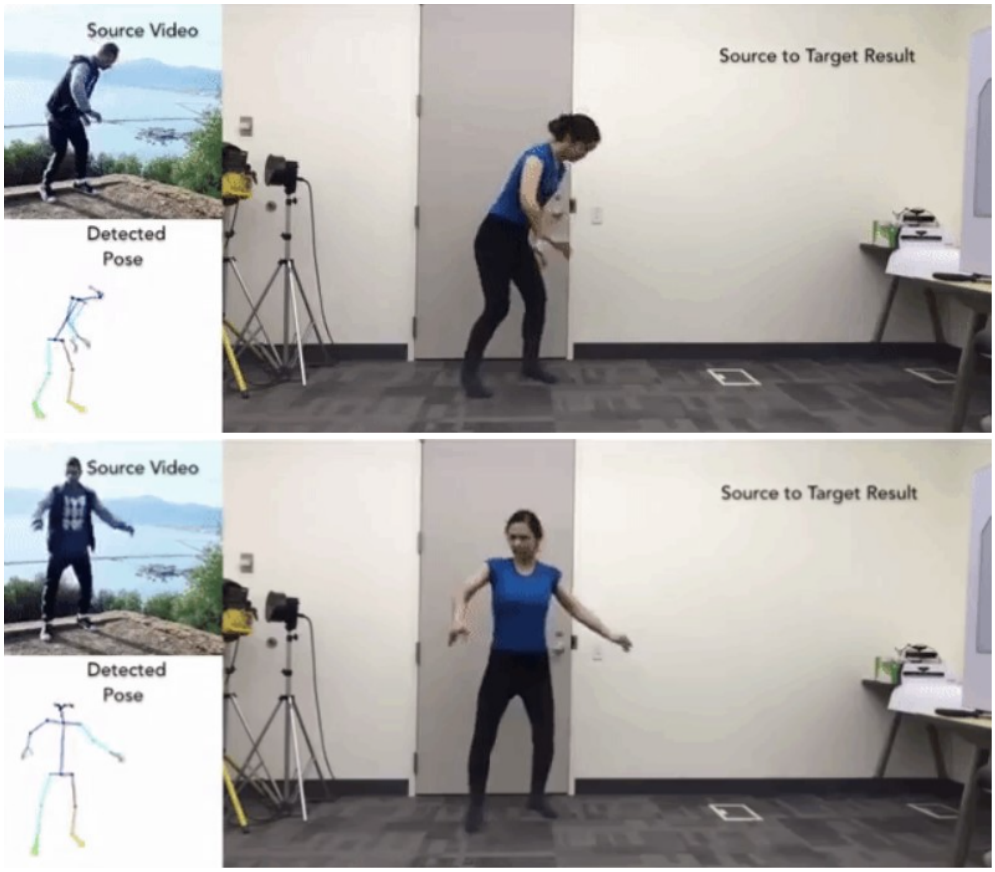
Επιπλέον μια μέθοδος βασισμένη σε γενετικά ανταγωνιστικά δίκτυα (GAN) είναι το StyleRig [32]. Η εφαρμογή αυτή δημιουργεί πορτρέτα προσώπων βασιζόμενη σε προεκπαιδευμένο και καθορισμένο StyleGAN στο οποίο προσθέτει την διατήρηση-έλεγχο των τρισδιάστατων σημασιολογικών παραμέτρων του προσώπου κατά την περιστροφή (rig-like control). Ένα επιπρόσθετο πλεονέκτημα του μοντέλου αυτού είναι ότι είναι αυτό-εκπαιδευόμενο χωρίς να υπάρχει ανάγκη για χειροκίνητες επεμβάσεις.

Ένα μειονέκτημα των περισσότερων μεθόδων ανταλλαγής προσώπων σε βίντεο είναι η ανάγκη εκπαίδευσης του εκάστοτε μοντέλου σε δεδομένα των προσώπων προς επεξεργασία, με σκοπό την ικανοποιητική ποιότητα ανακατασκευής του νέου προσώπου στη θέση του προηγούμενου προσώπου. Αυτό το επιλύει η εφαρμογή FaceShifter [33], η οποία μπορεί να εφαρμοστεί σε οποιαδήποτε νέα ζεύγη προσώπων χωρίς να απαιτεί ειδική εκπαίδευση πάνω σε αυτά. Η εφαρμογή αυτή ανταλλάζει πρόσωπα σε υψηλή ποιότητα εικόνας αξιοποιώντας και ενσωματώνοντας τα χαρακτηριστικά του προσώπου-στόχου.

Ακόμη ένα μοντέλο που επιτυγχάνει πολύ ικανοποιητικά αποτελέσματα χωρίς εκπαίδευση πάνω στο ζεύγος των προσώπων που πρόκειται να ανταλλάξει, είναι το FSGAN [34]. Το μοντέλο αυτό, παρά την έλλειψη εκπαίδευσης πάνω στα πρόσωπα, κατορθώνει να προσαρμόζει ικανοποιητικά τόσο τη στάση, όσο και τις εκφράσεις του προσώπου στόχου σύμφωνα με τις αντίστοιχες κινήσεις του προσώπου προς αντικατάσταση.

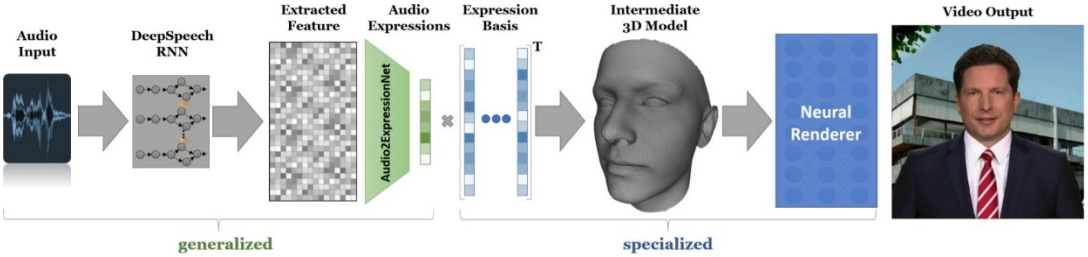
Πέρα από τη σύνθεση πλαστών βίντεο με παραποίηση μόνο της περιοχής του προσώπου, υπάρχουν και εφαρμογές που παραποιούν τη συνολική κίνηση του σώματος, όπως το “Do as I Do” Motion Transfer [35], το οποίο εντάσσεται στην κατηγορία της σύνθεσης πλαστών βίντεο τύπου Puppet-Master. Το μοντέλο αυτό μεταφέρει αυτόματα την κίνηση από ένα πρόσωπο-πηγή σε ένα πρόσωπο-στόχο, κάνοντας τη μεταφορά από ένα βίντεο σε ένα άλλο. Δηλαδή δεν παίρνει το πρόσωπο να το αντικαταστήσει στο βίντεο του άλλου προσώπου, αλλά προσαρμόζει την κίνηση του υποκειμένου ενός βίντεο σύμφωνα με

την κίνηση του υποκειμένου ενός άλλου βίντεο (Εικόνα 10). Μπορεί μάλιστα να δημιουργήσει ένα βίντεο συγχρονισμένων χορευτικών κινήσεων με πολλαπλά υποκείμενα.



Εικόνα 10: Παράδειγμα μεταφοράς της κίνησης ολόκληρου του σώματος από ένα βίντεο σε ένα άλλο με χρήση της εφαρμογής "Do as I Do" Motion Transfer [35]

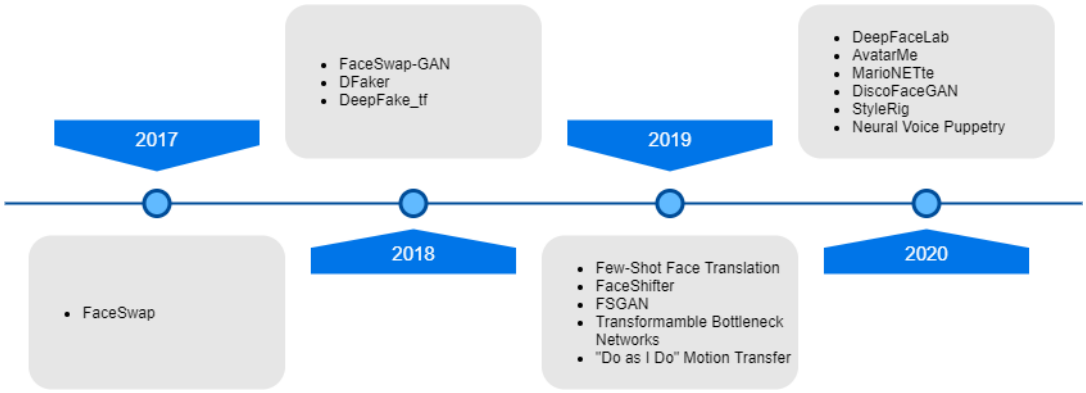
Τέλος, για την κατηγορία της σύνθεσης πλαστών βίντεο με βάση κάποιο ηχητικό απόσπασμα (Lip-Sync) μια αντιπροσωπευτική εφαρμογή είναι το Neural Voice Puppetry [36]. Η μέθοδος αυτή συνθέτει βίντεο ενός ομιλούμενου προσώπου από ένα ηχητικό απόσπασμα ενός άλλου προσώπου χρησιμοποιώντας τρισδιάστατη αναπαράσταση του προσώπου. Πιο συγκεκριμένα η διαδικασία που ακολουθείται είναι η εξής: Αρχικά το ηχητικό εισάγεται σε ένα μοντέλο DeepSpeech RNN το οποίο εξάγει κάποια χαρακτηριστικά. Τα χαρακτηριστικά αυτά τροφοδοτούνται σε ένα μικρό δίκτυο, το οποίο προβλέπει συντελεστές που αντιστοιχούν σε συγκεκριμένη έκφραση του προσώπου στόχου. Τέλος η έκφραση που προέκυψε προσαρμόζεται στο πρόσωπο-στόχο με χρήση ενός νευρωνικού δικτύου. Η παραπάνω ακολουθία διεργασιών παρουσιάζεται διαγραμματικά στην Εικόνα 11.



Εικόνα 11: Διαγραμματική απεικόνιση της ακολουθίας διεργασιών της εφαρμογής Neural Voice Puppetry [36]

Τα παραπάνω μοντέλα είναι αντιπροσωπευτικά δείγματα της προόδου που έχει γίνει ως τώρα στον τομέα της σύνθεσης πλαστών εικόνων και βίντεο. Δεδομένου ότι τα

περισσότερα από αυτά έχουν προταθεί μέσα στα τελευταία τρία χρόνια, είναι φανερό ότι η εξέλιξη στο συγκεκριμένο αντικείμενο είναι ραγδαία. Ακολούθως στην Εικόνα 12 παρουσιάζεται χρονολογικός πίνακας των πιο αξιοσημείωτων μεθόδων σύνθεσης πλαστών εικόνων και βίντεο που αναπτύχθηκαν την περίοδο 2017-2020.



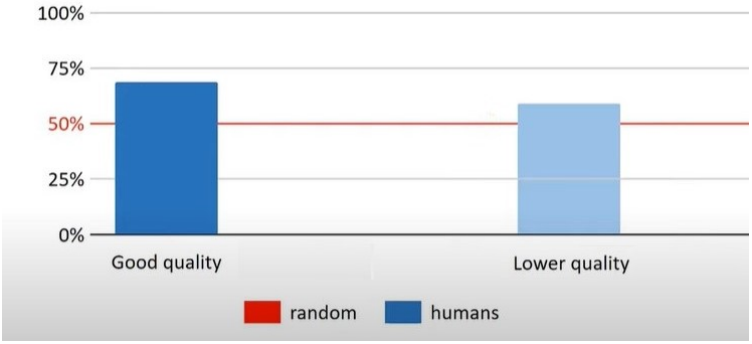
Εικόνα 12: Χρονολογικό διάγραμμα των βασικότερων μοντέλων σύνθεσης πλαστών βίντεο και εικόνων από το 2017 ως το 2020

3. Ανίχνευση πλαστών (deepfake) βίντεο

Σταδιακά τα deepfake βίντεο γίνονται όλο και πιο ρεαλιστικά, ενώ μειώνεται και η δυσκολία στη σύνθεσή τους και οι απαιτήσεις σε εξειδικευμένη γνώση από κάποιον που θέλει να δημιουργήσει ένα πλαστό βίντεο. Για τον λόγο αυτό ο κίνδυνος χρήσης deepfake τεχνικών για μη ηθικούς σκοπούς γίνεται όλο και πιο φανερός, κάτι που καθιστά αναγκαία την ανάπτυξη μεθόδων που επιτυγχάνουν την αποτελεσματική ανίχνευση πλαστών (deepfake) βίντεο.

Προς αυτή την κατεύθυνση, σε πρώτη φάση, εντοπίστηκαν κάποιες συχνές ατέλειες των πλαστών βίντεο, οι οποίες ήταν εφικτό να ανιχνευθούν με απλή οπτική παρατήρηση. Μερικές τέτοιες ατέλειες είναι η ασυνέχεια μεταξύ του προσώπου και των στοιχείων που το περιβάλλουν (μαλλιά, λαιμός κλπ.) και η τυχόν ασυμφωνία του προσώπου με τα υπόλοιπα στοιχεία του σώματος, ως προς το μέγεθος, το χρώμα κλπ.. Αυτού του τύπου οι ατέλειες προκύπτουν συνήθως σε περιπτώσεις χρήσης της μεθόδου αντικατάστασης προσώπου (face-swap). Άλλες αντίστοιχες ατέλειες είναι ότι πολλές φορές τα μάτια του προσώπου σε ένα πλαστό βίντεο δεν ανοιγοκλείνουν όπως στον φυσικό κόσμο και ότι η διάρκεια του βίντεο είναι συνήθως μικρή, καθώς είναι δύσκολο ένα μεγάλο πλαστό αληθοφανές βίντεο να παραχθεί σε εύλογο χρονικό διάστημα. Επιπλέον μια ατέλεια είναι συνήθως στο στόμα το κομμάτι των δοντιών, το οποίο συχνά δεν φαίνεται φυσιολογικό μορφολογικά [37].

Υπάρχουν και άλλες εμφανείς ατέλειες, όμως φαίνεται ότι σταδιακά αυτές μειώνονται όλο και περισσότερο με την περεταίρω ανάπτυξη των μεθόδων σύνθεσης πλαστών βίντεο. Χαρακτηριστικό είναι ότι, όπως φαίνεται στην Εικόνα 13, σε έρευνα με περισσότερους από 200 συμμετέχοντες, το μέσο ποσοστό των βίντεο που ταξινόμησαν σωστά οι ερωτηθέντες ως προς το αν πρόκειται για πλαστά ή πραγματικά βίντεο ήταν κάτω από 75% για βίντεο υψηλής ποιότητας που οι ατέλειες είναι συνήθως πιο εμφανείς και λίγο πάνω από το 50% για βίντεο χαμηλότερης ποιότητας. Δεδομένου ότι το 50% είναι ουσιαστικά αντίστοιχο της τυχαίας ταξινόμησης, φαίνεται ότι η οπτική παρατήρηση δεν είναι αρκετή για την ανίχνευση πλαστών (deepfake) βίντεο.



Εικόνα 13: Διάγραμμα που δείχνει το μέσο ποσοστό των βίντεο (υψηλής ή χαμηλότερης ποιότητας) που ταξινομήθηκαν ορθά ως πλαστά ή πραγματικά από ένα σύνολο περισσότερων από 200 ερωτηθέντων. Πηγή: https://www.youtube.com/watch?v=XMVmngZSvm0&ab_channel=MatthiasNiessner (Τελευταία πρόσβαση 19/5/2021)

Για τον λόγο αυτό αναπτύχθηκαν αλγόριθμοι που χρησιμοποιούν μηχανική μάθηση προκειμένου να διακρίνουν τα πλαστά από τα πραγματικά βίντεο. Οι αλγόριθμοι αυτοί χωρίζονται σε δύο βασικές κατηγορίες. Η πρώτη κατηγορία είναι τα μοντέλα μηχανικής μάθησης που ανιχνεύουν την πλαστή (deepfake) εικόνα και επομένως και στην περίπτωση του βίντεο δρουν με αντίστοιχο τρόπο, δηλαδή αποφασίζουν για κάθε στιγμιότυπο ξεχωριστά αν είναι πραγματικό ή όχι. Σε αυτές τις περιπτώσεις, αναλόγως τα κριτήρια που θέτει ο δημιουργός της αντίστοιχης εφαρμογής, προκύπτει συμπέρασμα και για το συνολικό βίντεο με βάση, για παράδειγμα, το αν το ποσοστό των στιγμιότυπων που είναι πλαστά

ξεπερνάει ένα προκαθορισμένο κατώφλι. Η δεύτερη κατηγορία είναι τα μοντέλα μηχανικής μάθησης που ανιχνεύουν αν ένα βίντεο είναι πλαστό με βάση μια σειρά στιγμιότυπων του βίντεο. Σε αυτή την περίπτωση υπάρχει η δυνατότητα το συμπέρασμα να βασιστεί όχι μόνο σε ατέλειες του κάθε στιγμιότυπου χωριστά, αλλά και σε ανωμαλίες που παρουσιάζονται στη μετάβαση από το ένα στιγμιότυπο στο άλλο.

3.1 Διαθέσιμα σύνολα δεδομένων

Προκειμένου να είναι δυνατό οι παραπάνω αλγόριθμοι να εκπαιδευτούν, ώστε να ξεχωρίζουν τα πλαστά βίντεο από τα πραγματικά απαραίτητη προϋπόθεση είναι η ύπαρξη συνόλου δεδομένων με ικανό αριθμό δειγμάτων πλαστών και πραγματικών βίντεο. Παρ' όλο που ο αριθμός των πλαστών βίντεο σταδιακά αυξάνεται, και πάλι δεν είναι αρκετός. Για τον λόγο αυτό αναπτύχθηκαν σύνολα δεδομένων ειδικά για αυτόν τον σκοπό. Τέτοια σύνολα είναι το FaceForencics++ [8], το DFDC [38], το DF-TIMIT [39], το UADFV [40] και το Celeb-DF [9].

Το σύνολο FaceForencics++ [8] έχει το πλεονέκτημα ότι περιλαμβάνει τέσσερις διαφορετικούς τύπους πλαστών βίντεο. Ο πρώτος τύπος είναι τα FaceSwap, που έχουν δημιουργηθεί με μια μέθοδο που δεν περιλαμβάνει χρήση μηχανικής μάθησης. Ο δεύτερος τύπος είναι τα DeepFakes που έχουν παραχθεί με χρήση του μοντέλου μηχανικής μάθησης που χρησιμοποιήθηκε και από την εφαρμογή FakeApp που προαναφέρθηκε. Ο τρίτος τύπος είναι τα Face2Face που δεν περιλαμβάνουν ανταλλαγή προσώπων όπως τα δύο προηγούμενα, αλλά αντίθετα πρόκειται για βίντεο παραποιημένα έτσι ώστε το πρόσωπο που δρα σε αυτά να κάνει κινήσεις που συμφωνούν με τις κινήσεις ενός άλλου προσώπου από άλλο βίντεο. Τέλος ο τέταρτος τύπος είναι τα NeuralTextures, τα οποία είναι βίντεο στα οποία έχει παραποιηθεί η περιοχή του στόματος του προσώπου προκειμένου να κινείται σύμφωνα με τον τρόπο που κινείται το στόμα κάποιου άλλου προσώπου από άλλο βίντεο. Η ποικιλία αυτή στις μεθόδους σύνθεσης των πλαστών βίντεο είναι πολύ σημαντική, γιατί επιτρέπει στα μοντέλα που εκπαιδεύονται με επιτυχία σε αυτό το σύνολο δεδομένων να είναι αποτελεσματικά για ποικίλους τύπους πλαστών βίντεο.

Από την άλλη πλευρά οι τεχνικές σύνθεσης βίντεο συνεχώς βελτιώνονται με αποτέλεσμα τα μειονεκτήματα και οι ατέλειες των αποτελεσμάτων τους να περιορίζονται. Για τον λόγο αυτό ένα μοντέλο που εκπαιδεύτηκε στα δεδομένα του συνόλου FaceForencics++ μπορεί να μην είναι τελικά αποτελεσματικό για δεδομένα άλλων νεότερων συνόλων που δημιουργήθηκαν με καινούριες μεθόδους σύνθεσης πλαστών βίντεο. Για παράδειγμα, το σύνολο δεδομένων Celeb-DF αποτελείται ναι μεν από βίντεο που έχουν παραποιηθεί με μία μόνο μέθοδο, όμως αυτή η μέθοδος είναι νεότερη και αποτελεσματικότερη παράγοντας πιο ρεαλιστικά και ποιοτικά ανώτερα βίντεο από τα αντίστοιχα του συνόλου FaceForencics++. Για τον λόγο αυτό μοντέλα που έχουν εκπαιδευτεί με επιτυχία στο σύνολο δεδομένων FaceForencics++ δεν έχουν ανάλογη επιτυχία στο σύνολο δεδομένων Celeb-DF. Επομένως, καθώς οι εξελίξεις στον χώρο της σύνθεσης πλαστών βίντεο προχωρούν, χρειάζεται συνεχώς να δημιουργούνται νέα σύνολα δεδομένων που καλύπτουν και τις νέες μεθόδους, ώστε να εκπαιδεύονται πάνω σε αυτά δίκτυα που στόχο έχουν την ταξινόμηση βίντεο σε παραποιημένα και σε αυθεντικά.

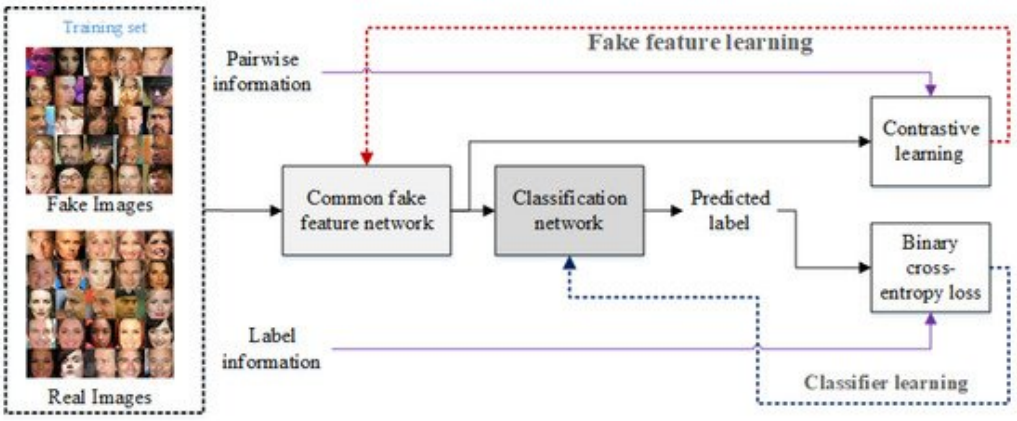
3.2 Ανίχνευση πλαστής εικόνας

Στον τομέα της ανίχνευσης πλαστής εικόνας ένα από τα μεγαλύτερα εμπόδια είναι η συνεχής εξέλιξη των συνελικτικών μοντέλων (CNN) και των γενετικών ανταγωνιστικών δικτύων (GAN), που οδηγεί σε όλο και πιο ρεαλιστικά αποτελέσματα, στα οποία η διαφορά μεταξύ πλαστού και πραγματικού είναι πολύ δύσκολο να βρεθεί. Σε μια πρόσφατη

προσπάθεια χρησιμοποιήθηκε η μέθοδος bag of words για την εξαγωγή χαρακτηριστικών από εικόνες πλαστές και μη. Τα χαρακτηριστικά αυτά τροφοδοτήθηκαν σε μοντέλα ταξινόμησης, όπως SVM, random forest και perceptrons πολλών επιπέδων (MLP), με στόχο τη διάκριση μεταξύ των πραγματικών και των πλαστών εικόνων προσώπου. Στην προσέγγιση αυτή παρατηρήθηκε ότι από το σύνολο των πλαστών εικόνων εκείνες που ήταν πιο δύσκολο να ανιχνευθούν ήταν αυτές που είχαν δημιουργηθεί με χρήση γενετικών ανταγωνιστικών δικτύων (GAN). Ο λόγος είναι ότι οι εικόνες αυτές είναι πιο αληθοφανείς και με υψηλότερη ποιότητα, κάτι που οφείλεται στην ικανότητα του γενετικού ανταγωνιστικού δικτύου να μαθαίνει την κατανομή πολύπλοκων δεδομένων και να παράγει νέα δεδομένα με παρόμοια κατανομή.

Επιπλέον, προκειμένου να εξαλειφθούν τυχόν ανωμαλίες της εξόδου των γενετικών ανταγωνιστικών δικτύων συχνά χρησιμοποιούνται τεχνικές όπως το φίλτρο θόλωσης Gaussian, οι οποίες έχουν το κόστος της χαμηλότερης ακρίβειας της εικόνας, αλλά αυξάνουν την στατιστική ομοιότητα μεταξύ πλαστών και αυθεντικών εικόνων σε επίπεδο εικονοστοχείου [41]. Βασιζόμενοι σε αυτό οι Agarwal και Varshney [42] χρησιμοποίησαν μια διαφορετική προσέγγιση για την ανίχνευση πλαστών εικόνων. Συγκεκριμένα στηρίχθηκαν στην ελάχιστη απόσταση μεταξύ των κατανομών αυθεντικών εικόνων και εικόνων που έχουν παραχθεί από ένα γενετικό ανταγωνιστικό δίκτυο (oracle error). Τα αποτελέσματα της έρευνάς τους έδειξαν ότι η απόσταση αυτή αυξάνεται όταν το γενετικό ανταγωνιστικό δίκτυο είναι λιγότερο ακριβές. Κατά αυτόν τον τρόπο είναι εφικτό να χρησιμοποιηθεί αυτή η απόσταση ως κριτήριο ταξινόμησης εικόνας σε πλαστή ή μη. Είναι βέβαια μια τεχνική, που όπως αναφέρθηκε λειτουργεί ικανοποιητικά για τις περιπτώσεις που τα πλαστά βίντεο δεν έχουν μεγάλη ακρίβεια.

Μια άλλη προσέγγιση για την ανίχνευση πλαστών εικόνων είναι αυτή που προτάθηκε από τον Hsu και τους συνεργάτες του [43]. Η μέθοδος αυτή αποτελείται από δύο φάσεις. Η πρώτη φάση συνίσταται στην εξαγωγή χαρακτηριστικών η οποία γίνεται με χρήση κοινού δικτύου εξαγωγής χαρακτηριστικών που διαφοροποιούν την πλαστή εικόνα από την πραγματική (Common Fake Feature Network – CFFN). Πρόκειται για ένα δίκτυο νευρώνων που δέχεται δύο εισόδους και χρησιμοποιώντας τα ίδια βάρη και για τις δύο, παράγει δύο συγκρίσιμες εξόδους. Η αρχιτεκτονική που χρησιμοποιήθηκε περιλαμβάνει πολλά πυκνά στρώματα νευρώνων (dense layers) με διαφορετικό αριθμό νευρώνων σε κάθε στρώμα. Ο αριθμός των στρωμάτων αυτών είναι τρία ή πέντε ανάλογα με το αν τα δεδομένα που εξετάζονται είναι εικόνες προσώπου ή αυθαίρετες εικόνες. Μέσω αυτού του μοντέλου εξάγονται πληροφορίες κατά ζεύγη πλαστών και πραγματικών εικόνων, οι οποίες αποτελούν τα χαρακτηριστικά της μεταξύ τους διαφοροποίησης. Στη δεύτερη φάση τα χαρακτηριστικά αυτά εισάγονται σε ένα μικρό συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network – CNN), το οποίο βασιζόμενο σε αυτά διακρίνει τις πλαστές εικόνες από τις πραγματικές. Οι δύο φάσεις παρουσιάζονται διαγραμματικά στην Εικόνα 14. Η διαδικασία αυτή επιτυγχάνει καλύτερα αποτελέσματα από τις προαναφερθείσες μεθόδους, ενώ λειτουργεί τόσο για ανίχνευση πλαστών εικόνων προσώπων όσο και για ανίχνευση πλαστών εικόνων γενικά. Επιπλέον τα δείγματα στα οποία εκπαιδεύτηκε το μοντέλο αυτό των δύο φάσεων προέρχονται από ποικίλα σύνολα δεδομένων και από διάφορες τεχνικές σύνθεσης πλαστών εικόνων βασισμένες σε διαφορετικά γενετικά ανταγωνιστικά δίκτυα, κάποια εκ των οποίων παράγουν αρκετά αληθοφανείς πλαστές εικόνες που σε μέγεθος φτάνουν το 128 x 128. Επομένως πρόκειται για μια αρκετά αξιόλογη προσέγγιση στον τομέα της ανίχνευσης πλαστών εικόνων.



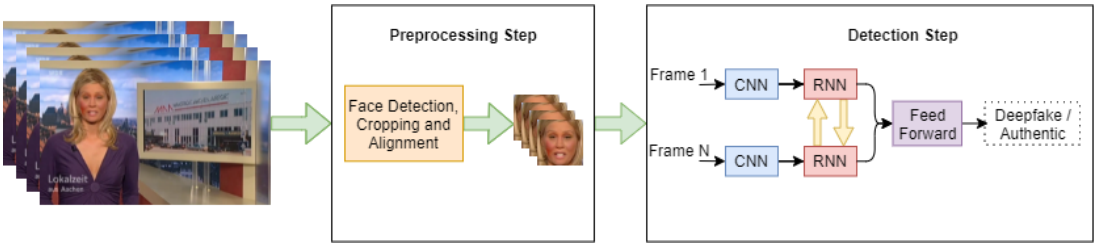
Εικόνα 14: Διαγραμματική απεικόνιση του συνδυασμού CFFN και CNN για την ανίχνευση πλαστών εικόνων [43]

3.3 Ανίχνευση πλαστού βίντεο

Οι περισσότερες μέθοδοι ανίχνευσης πλαστών εικόνων δε δύναται να χρησιμοποιηθούν για ανίχνευση πλαστών βίντεο, εξαιτίας του ότι τα στιγμιότυπα-εικόνες των βίντεο υποβαθμίζονται σε ποιότητα κατά τη συμπίεσή τους για την σύνθεση του βίντεο [44]. Οι εφαρμογές που έχουν αναπτυχθεί για την ανίχνευση πλαστών βίντεο χωρίζονται σε δύο διαφορετικές κατηγορίες. Η πρώτη περιλαμβάνει τεχνικές που εκμεταλλεύονται τα χρονικά χαρακτηριστικά του βίντεο, εντοπίζοντας ουσιαστικά ανωμαλίες στη χρονική αλληλουχία των χαρακτηριστικών των στιγμιότυπων. Η δεύτερη κατηγορία βασίζεται αποκλειστικά σε ατέλειες που υπάρχουν στο επίπεδο του κάθε στιγμιότυπου χωριστά, χωρίς να λαμβάνει υπόψη την πληροφορία από προηγούμενα ή επόμενα στιγμιότυπα.

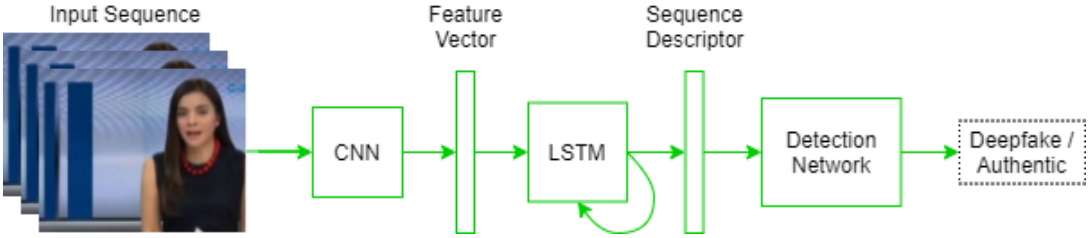
3.3.1 Τεχνικές βασισμένες στη χρονική πληροφορία του βίντεο

Ο Sabir και οι συνεργάτες του [45] στηρίχθηκαν στην παρατήρηση ότι στα πλαστά βίντεο συχνά υπάρχουν ατέλειες στη χρονική συνέχεια των κινήσεων των προσώπων. Οι ατέλειες αυτές οφείλονται στο γεγονός ότι τα βίντεο αυτά παράγονται από μεμονωμένη επεξεργασία κάθε στιγμιότυπου, χωρίς να λαμβάνεται υπόψη η πληροφορία των προηγούμενων και των επόμενων στιγμιότυπων. Η προσέγγιση του Sabir και των συνεργατών του [45] περιλαμβάνει την χρήση ενός αναδρομικού συνελικτικού μοντέλου (recurrent convolutional model - RCN) για τον εντοπισμό πλαστών (deepfake) βίντεο με βάση τα χωροχρονικά τους χαρακτηριστικά. Στο δίκτυο αυτό κάθε στιγμιότυπο εισάγεται αρχικά σε ένα συνελικτικό νευρωνικό δίκτυο (CNN), το οποίο έχει τη δομή του DenseNet [46]. Η έξοδος που προκύπτει για το αντίστοιχο στιγμιότυπο εισάγεται ακολούθως σε μια από τις μονάδες ενός αναδρομικού νευρωνικού δικτύου [47]. Το συνελικτικό νευρωνικό δίκτυο είναι υπεύθυνο για την εξαγωγή των χρήσιμων χαρακτηριστικών των στιγμιοτύπων, ενώ το αναδρομικό νευρωνικό δίκτυο είναι υπεύθυνο για την ανίχνευση ατελειών στη χρονική αλληλουχία αυτών των χαρακτηριστικών, ώστε να προκύψει τελικά συμπέρασμα για το αν το αντίστοιχο βίντεο είναι πλαστό ή όχι. Το συνολικό μοντέλο RCN ελέγχθηκε στο σύνολο δεδομένων FaceForencics++ [8], επιτυγχάνοντας πολύ ελπιδοφόρα αποτελέσματα. Η δομή του μοντέλου παρουσιάζεται διαγραμματικά στην Εικόνα 15.



Εικόνα 15: : Διαγραμματική απεικόνιση του μοντέλου RCN [45]

Αντίστοιχη προσπάθεια έγινε και από τους Guera και Delp [48], οι οποίοι υποστήριξαν ότι τα πλαστά βίντεο περιέχουν ασυνέπειες όχι μόνο εντός του κάθε στιγμιότυπου, αλλά και στη χρονική συνέχεια των στιγμιότυπων. Με βάση αυτή την παραδοχή πρότειναν μία μέθοδο ανίχνευσης πλαστών βίντεο η οποία περιλαμβάνει συνελικτικό νευρωνικό δίκτυο (CNN) ακολουθούμενο από αναδρομικό νευρωνικό δίκτυο τύπου LSTM (long short term memory). Όπως και στην προηγούμενη προσέγγιση, το CNN χρησιμοποιείται για την εξαγωγή χαρακτηριστικών σε επίπεδο στιγμιότυπου, ενώ το LSTM τροφοδοτείται με αυτά τα χαρακτηριστικά, προκειμένου να εξάγει μια περιγραφή της χρονικής ακολουθίας. Το αποτέλεσμα του LSTM εισάγεται τελικά σε ένα πλήρως συνδεδεμένο δίκτυο νευρώνων το οποίο αποφασίζει αν το βίντεο είναι αυθεντικό ή όχι. Η δομή αυτή παρουσιάζεται διαγραμματικά στην Εικόνα 16.



Εικόνα 16: Διαγραμματική απεικόνιση του συνδυασμού CNN και LSTM για ανίχνευση πλαστών βίντεο [48]

Μια διαφορετική προσέγγιση προτάθηκε στο [49]. Η προσέγγιση αυτή βασίστηκε στην παρατήρηση ότι τα μάτια των προσώπων στα πλαστά βίντεο ανοιγοκλείνουν λιγότερο συχνά ή και με μικρότερη διάρκεια από ότι τα μάτια των προσώπων στα αυθεντικά βίντεο. Για την ακρίβεια ένας μέσος ενήλικας υγιής συνήθως ανοιγοκλείνει τα μάτια του κάθε 2 με 10 δευτερόλεπτα και αυτό διαρκεί από 1 έως 4 δέκατα του δευτερολέπτου. Όμως οι μέθοδοι σύνθεσης πλαστών βίντεο συχνά βασίζονται σε φωτογραφίες του προσώπου οι οποίες έχουν βρεθεί στο διαδίκτυο και σπάνια εικονίζουν το πρόσωπο με κλειστά μάτια. Αυτό έχει ως συνέπεια οι αλγόριθμοι αυτοί να δημιουργούν πλαστά βίντεο που περιλαμβάνουν μικρότερο ποσοστό στιγμιότυπων με κλειστά μάτια από αυτό που θα έπρεπε προκειμένου τα πρόσωπα να ανοιγοκλείνουν τα μάτια τους με φυσιολογική συχνότητα και διάρκεια. Δεδομένου αυτού ο Li και οι συνεργάτες του, αφού περιέκοψαν την εικόνα του προσώπου από κάθε στιγμιότυπο, το ευθυγράμμισαν και ακολούθως εξήγαγαν την περιοχή των ματιών με βάση έξι σημεία αναφοράς. Όρισαν το μέγεθος του πλαισίου της περιοχής αυτής για κάθε στιγμιότυπο, έτσι ώστε τα πλαίσια αυτά να συνενωθούν για κάθε βίντεο δημιουργώντας μια ακολουθία στιγμιότυπων που περιλαμβάνουν μόνο την περιοχή των ματιών. Τα νέα αυτά στιγμιότυπα εισήχθησαν σε μακροπρόθεσμα αναδρομικά νευρωνικά δίκτυα (long-term recurrent convolutional networks – LRCN) [50]. Τα δίκτυα αυτά αποτελούνται από έναν συνελικτικό νευρωνικό δίκτυο (CNN) για την εξαγωγή χαρακτηριστικών, ένα LSTM για την περιγραφή της χρονικής ακολουθίας των χαρακτηριστικών αυτών και ένα πλήρως συνδεδεμένο στρώμα για την πρόβλεψη τις πιθανότητας ανοίγματος και κλεισίματος των ματιών. Το μάτι που ανοιγοκλείνει έχει ως αποτέλεσμα ισχυρές χρονικές εξαρτήσεις οι οποίες αποτυπώνονται μέσω του LSTM. Ο ρυθμός ανοίγματος και κλεισίματος των ματιών υπολογίζεται με βάση τα αποτελέσματα της πρόβλεψης, όπου το κλείσιμο και το άνοιγμα

των ματιών ορίζεται ως η μέγιστη τιμή πάνω από ένα κατώφλι 0.5 με διάρκεια μικρότερη από 7 στιγμιότυπα. Η μέθοδος αυτή αξιολογήθηκε σε δεδομένα που συλλέχθηκαν από το διαδίκτυο και περιλαμβάνουν 49 βίντεο συνεντεύξεων και διαλέξεων και τα αντίστοιχα πλαστά βίντεο που προήλθαν από τα προηγούμενα με χρήση αλγορίθμων σύνθεσης πλαστών (deepfake) βίντεο. Τα αποτελέσματα της μεθόδου ήταν πολύ ικανοποιητικά.

3.3.2 Τεχνικές βασισμένες στη χωρική πληροφορία εντός του στιγμιότυπου

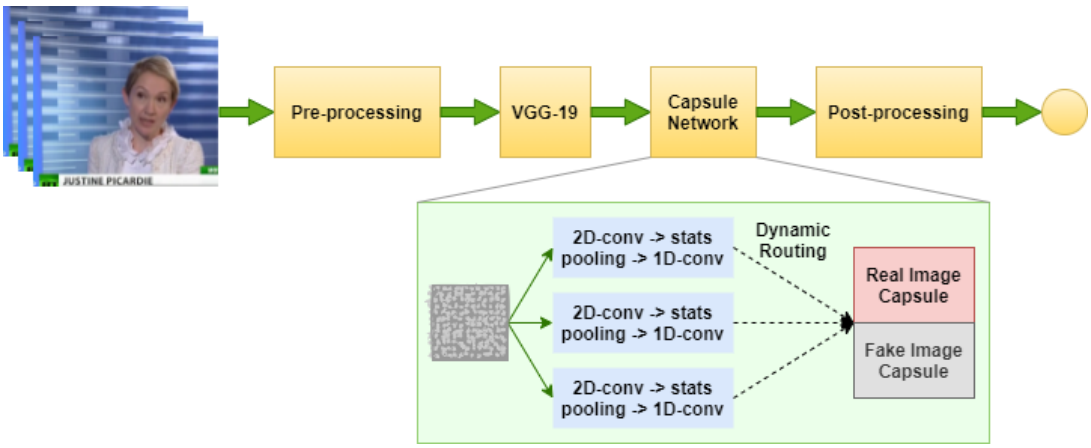
Όλες οι παραπάνω μέθοδοι ανίχνευσης πλαστών βίντεο αξιοποιούν χρονικά μοτίβα που υπάρχουν σε κάθε βίντεο, προκειμένου να καταλήξουν σε συμπέρασμα για την αυθεντικότητα ή μη του βίντεο. Υπάρχουν όμως και τεχνικές οι οποίες χωρίζουν το βίντεο σε στιγμιότυπα και εξετάζουν οπτικές ατέλειες μέσα στο καθένα από αυτά, έτσι ώστε να ξεχωρίσουν με βάση αυτές τα πλαστά από τα πραγματικά βίντεο. Από αυτές τις μεθόδους άλλες χρησιμοποιούν βαθιά (deep) νευρωνικά δίκτυα και άλλες ρηχά (shallow). Ακολούθως παρουσιάζονται αρχικά οι τεχνικές που χρησιμοποιούν βαθιά νευρωνικά δίκτυα.

Τα πλαστά (deepfake) βίντεο συνήθως παράγονται σε περιορισμένη ανάλυση και γι' αυτό απαιτούν την κατάλληλη παραμόρφωση του προσώπου, όπως κλιμάκωση του μεγέθους του, περιστροφή και άλλα, προκειμένου να ταιριάζουν με τα αρχικά πραγματικά πρόσωπα. Εξαιτίας της ασυμφωνίας της ανάλυσης της περιοχής του παραποιημένου προσώπου με την ανάλυση του περιβάλλοντος χώρου, η παραπάνω διαδικασία προκαλεί ατέλειες οι οποίες μπορούν να ανιχνευθούν από συνελικτικά νευρωνικά δίκτυα (CNN), όπως το VGG16 [51], το ResNet50, το ResNet101 και το ResNet152 [52].

Μια μέθοδος βαθιάς μάθησης που εκμεταλλεύτηκε τις ατέλειες που προκύπτουν κατά την προσαρμογή του νέου προσώπου στο περιβάλλον του στιγμιότυπου είναι αυτή που προτάθηκε στο [53]. Η μέθοδος αυτή αξιολογήθηκε με βάση δύο σύνολα δεδομένων, το UAFV [54] και το DeepfakeTIMIT [39]. Το πρώτο περιέχει 49 πραγματικά και 49 πλαστά βίντεο με συνολικά 32752 στιγμιότυπα. Το δεύτερο περιλαμβάνει ένα σύνολο 320 βίντεο χαμηλής ποιότητας 64 x 64 και ένα σύνολο 320 βίντεο υψηλής ποιότητας 128 x 128 με συνολικά 10537 στιγμιότυπα αυθεντικών βίντεο και 34023 στιγμιότυπα παραποιημένων βίντεο. Η απόδοση της συγκεκριμένης μεθόδου συγκρίθηκε με άλλες μεθόδους. Κάποιες από αυτές ήταν το HeadPose [54], η μέθοδος ανίχνευσης αλλοίωσης του προσώπου με χρήση δύο ροών νευρωνικών δικτύων NN [55] και δύο είδη του MesoNet, δηλαδή το Meso4 και το MesoInception-4 [44]. Το πλεονέκτημα της προτεινόμενης μεθόδου είναι ότι δεν χρειάζεται να δημιουργεί πλαστά βίντεο ως αρνητικά παραδείγματα πριν την εκπαίδευση των μοντέλων ανίχνευσης. Αντίθετα τα αρνητικά παραδείγματα δημιουργούνται δυναμικά εξάγοντας την περιοχή του προσώπου από την πραγματική εικόνα και δημιουργώντας αντίγραφά της σε πολλαπλές κλίμακες πριν την εφαρμογή γκαουσιανού θολώματος (Gaussian blur) σε ένα τυχαία αντίγραφο από αυτά και την προσαρμογή του στην αρχική πραγματική εικόνα. Η πρακτική αυτή μειώνει τον απαιτούμενο χρόνο και τους υπολογιστικούς πόρους που χρειάζονται σε σύγκριση με άλλες μεθόδους, οι οποίες απαιτούν τη δημιουργία των πλαστών δειγμάτων εκ των προτέρων.

Μια άλλη προσέγγιση βαθιάς μάθησης είναι αυτή του Nguyen και των συνεργατών του [56], οι οποίοι προτείνουν τη χρήση δικτύων καψουλών (Capsule Networks) για την ανίχνευση πλαστών βίντεο. Τα δίκτυα αυτά αρχικά είχαν προταθεί για να αντιμετωπίσουν τους περιορισμούς που έχουν τα συνελικτικά νευρωνικά δίκτυα (CNN) όταν χρησιμοποιούνται προκειμένου να εντοπίσουν διαδικασίες γεωμορφολογικής επεξεργασίας που χρησιμοποιούνται για παραγωγή εικόνων της γης [57]. Η πρόσφατη ανάπτυξη ενός δικτύου καψουλών βασισμένου σε αλγόριθμο δυναμικής δρομολόγησης [58] ανέδειξε την ικανότητά του να περιγράφει ιεραρχικές σχέσεις της στάσης μεταξύ τμημάτων αντικειμένων. Η προσέγγιση αυτή χρησιμοποιήθηκε ως κομμάτι σε μια σειρά διεργασιών για την ανίχνευση

πλαστών εικόνων και βίντεο. Το συνολικό δίκτυο που προέκυψε παρουσιάζεται διαγραμματικά στην Εικόνα 17. Στο δίκτυο αυτό ένας δυναμικός αλγόριθμος δρομολόγησης οδηγεί τις εξόδους των τριών αρχικών καψουλών στις κάψουλες εξόδου μέσα από έναν αριθμό επαναλήψεων, προκειμένου να διαχωριστούν οι πραγματικές από τις πλαστές εικόνες. Η μέθοδος αυτή αξιολογήθηκε σε τέσσερα σύνολα δεδομένων, τα οποία καλύπτουν ένα ευρύ φάσμα πλαστών επιθέσεων με εικόνα και βίντεο. Συγκεκριμένα περιλαμβάνουν το σύνολο δεδομένων Idiap Research Institute replay-attack dataset [59], το σύνολο δεδομένων ανταλλαγής προσώπων που δημιουργήθηκε από τον Afchar και τους συνεργάτες του [44], το σύνολο δεδομένων FaceForensics [60], που παράχθηκε με τη μέθοδο Face2Face [61], και το σύνολο δεδομένων εικόνων δημιουργημένων μέσω υπολογιστή που παράχθηκε από τον Rahmouni και τους συνεργάτες του [62]. Τα αποτελέσματα της προτεινόμενης μεθόδου είναι καλύτερα σε σύγκριση με άλλες μεθόδους που έχουν αξιολογηθεί στα παραπάνω σύνολα δεδομένων, κάτι που δείχνει ότι τα δίκτυα καψουλών έχουν δυνατότητες να συμβάλουν στη δημιουργία μοντέλων ανίχνευσης πλαστών βίντεο που θα εντοπίζουν με επιτυχία πλαστά βίντεο και εικόνες.



Εικόνα 17: Μοντέλο ανίχνευσης πλαστών βίντεο βασισμένο σε δίκτυο κάψουλών με δυναμική δρομολόγηση [56]

Πέρα από τις μεθόδους που χρησιμοποιούν βαθιά νευρωνικά δίκτυα, υπάρχουν και προσεγγίσεις με ρηχά νευρωνικά δίκτυα ταξινόμησης. Ο Yang και οι συνεργάτες του [54] πρότειναν μια μέθοδο ανίχνευσης πλαστών εικόνων ή βίντεο, η οποία βασίζεται στη θέση και τον προσανατολισμό της τρισδιάστατης αναπαράστασης του κεφαλιού των εικονιζόμενων προσώπων και εξετάζει τις διαφορές που υπάρχουν μεταξύ πλαστών και πραγματικών εικόνων ή βίντεο ως προς τη στάση αυτή του τρισδιάστατου κεφαλιού. Η θέση και ο προσανατολισμός του κεφαλιού υπολογίζονται με βάση 68 σημεία αναφοράς της περιοχής του εικονιζόμενου προσώπου. Τα χαρακτηριστικά αυτά ακολούθως εισάγονται σε ένα μοντέλο ταξινόμησης SVM, το οποίο αποφασίζει αν το δείγμα είναι πλαστό ή αυθεντικό. Η παραπάνω προσέγγιση ελέγχθηκε σε δύο σύνολα δεδομένων, το UADFV [54] και ένα υποσύνολο των δεδομένων που χρησιμοποιούνται στο DARPA MediFor GAN Image / Video Challenge [63]. Το πρώτο σύνολο δεδομένων περιλαμβάνει 49 πλαστά (deepfake) βίντεο και τα αντίστοιχα πραγματικά βίντεο. Από το δεύτερο σύνολο συμπεριλήφθηκαν 241 πραγματικές εικόνες και 252 πλαστές εικόνες. Οι επιδόσεις της μεθόδου αυτής ήταν εξαιρετικές συγκριτικά με άλλες προσεγγίσεις.

Μια άλλη προσέγγιση με χρήση ρηχού δικτύου ταξινόμησης είναι αυτή που προτάθηκε στο [64]. Η μέθοδος αυτή, προκειμένου να ανιχνεύσει πλαστά βίντεο ή εικόνες, εκμεταλλεύεται τις ατέλειες που εντοπίζονται στην περιοχή των ματιών και των δοντιών, καθώς και στο περίγραμμα του προσώπου. Οι οπτικές αυτές ατέλειες προκύπτουν από την έλλειψη συνοχής των χαρακτηριστικών του προσώπου, τις ατέλειες ως προς τον φωτισμό και τις αλλοιώσεις του προσώπου λόγω μη ακριβούς εκτίμησης της γεωμετρίας του. Αυτά τα

προβλήματα έχουν ως αποτέλεσμα να είναι εφικτή η ανίχνευση των πλαστών βίντεο μέσω της έλλειψης αντανακλάσεων και λεπτομερειών στις περιοχές των ματιών και των δοντιών, καθώς και μέσω της εξαγωγής χαρακτηριστικών υφής της περιοχής του προσώπου. Στην προσέγγιση αυτή επομένως εξάγεται ένα διάνυσμα χαρακτηριστικών των ματιών, ένα διάνυσμα χαρακτηριστικών των δοντιών και ένα διάνυσμα χαρακτηριστικών του συνολικού προσώπου. Τα χαρακτηριστικά αυτά εξάγονται με χρήση σημείων αναφοράς. Ακολούθως εισάγονται τα χαρακτηριστικά αυτά σε δύο μοντέλα ταξινόμησης και συγκεκριμένα σε ένα μοντέλο λογιστικής παλινδρόμησης (logistic Regression) και σε ένα μικρό νευρωνικό δίκτυο. Τα δύο αυτά μοντέλα ταξινομούν τα βίντεο σε πλαστά ή σε αυθεντικά. Για τον έλεγχο της απόδοσης της μεθόδου χρησιμοποιήθηκε ένα σύνολο δεδομένων προερχόμενο από βίντεο του YouTube και τα αποτελέσματα ήταν ικανοποιητικά. Παρ' όλα αυτά η συγκεκριμένη μέθοδος έχει ένα μειονέκτημα. Απαιτεί εικόνες που καλύπτουν συγκεκριμένες προϋποθέσεις, όπως το να είναι τα μάτια ανοιχτά ή να φαίνονται τα δόντια του προσώπου.

Μια διαφορετική προσέγγιση προτάθηκε στο [65]. Η προσέγγιση αυτή εκμεταλλεύτηκε την έλλειψη ομοιομορφίας στην απόκριση της φωτογραφίας (photo response non uniformity – PRNU). Το PRNU είναι ένα στοιχείο θορύβου στο μοτίβο του αισθητήρα της κάμερας. Οφείλεται στην ατέλεια της κατασκευής των πλακιδίων πυριτίου και στην συνεπαγόμενη ασυνέπεια ως προς την ευαισθησία των εικονοστοιχείων (pixel) στο φως, εξαιτίας της ποικιλίας των φυσικών χαρακτηριστικών των πλακιδίων πυριτίου. Κατά τη λήψη μιας φωτογραφίας η ατέλεια του αισθητήρα εισάγεται στις ζώνες υψηλής συχνότητας του περιεχομένου της φωτογραφίας με τη μορφή αόρατου θορύβου. Επειδή η ατέλεια αυτή δεν είναι ομοιόμορφη σε ολόκληρη την γκοφρέτα πυριτίου (πολυκρυσταλικό πυρίτιο), οι αισθητήρες πυριτίου παράγουν μοναδικό PRNU. Για τον λόγο αυτό το PRNU θεωρείται κάτι σαν δακτυλικό αποτύπωμα, που αφήνουν οι ψηφιακές φωτογραφικές μηχανές στις φωτογραφίες που βγάζουν [66]. Δεδομένου ότι κατά την παραποίηση ενός βίντεο, αλλάζει σε κάθε στιγμιότυπο η περιοχή του προσώπου, είναι λογικό ότι θα αλλάζει λόγω αυτού και το μοτίβο PRNU του αντίστοιχου στιγμιότυπου. Στη μέθοδο που προτάθηκε αρχικά τα βίντεο χωρίζονται σε στιγμιότυπα και από τα στιγμιότυπα αυτά κρατιέται μόνο η περιοχή του προσώπου. Οι εικόνες που προκύπτουν ακολούθως διαχωρίζονται διαδοχικά σε οχτώ ομάδες για καθεμία από τις οποίες υπολογίζεται ένα μέσο πρότυπο PRNU. Ακολούθως υπολογίζονται κανονικοποιημένες επιδόσεις διασταυρούμενης συσχέτισης (normalized cross correlation scores) με σκοπό τη σύγκριση μεταξύ των προτύπων PRNU που προέκυψαν από τις οχτώ ομάδες. Οι συγγραφείς του [65] παρήγαγαν ένα σύνολο δεδομένων για τον έλεγχο της επιτυχίας της μεθόδου που πρότειναν. Το σύνολο αυτό αποτελείται από 10 αυθεντικά βίντεο και 16 παραποιημένα βίντεο, τα οποία δημιουργήθηκαν από τα αυθεντικά με χρήση του εργαλείου DeepFaceLab [18]. Η ανάλυση που έγινε σε αυτά τα δείγματα έδειξε ότι υπάρχει στατιστικά σημαντική διαφορά στη μέση κανονικοποιημένη επίδοση της διασταυρούμενης συσχέτισης (cross correlation score) μεταξύ των πλαστών βίντεο και των πραγματικών. Αν και το σύνολο δεδομένων στο οποίο έγινε η έρευνα είναι μικρό, και πάλι φαίνεται να ισχύει ότι το PRNU μπορεί να αποτελέσει κριτήριο για την ανίχνευση πλαστών βίντεο. Θα ήταν ενδιαφέρον να εξεταστεί η μέθοδος αυτή και σε μεγαλύτερα σύνολα δεδομένων, ώστε να είναι πιο σίγουρο το κατά πόσο είναι όντως αποτελεσματική.

Πέρα από τις ατέλειες που μπορεί να βρεθούν στα χαρακτηριστικά μιας εικόνας ή μιας σειράς στιγμιότυπων του βίντεο, υπάρχουν και άλλοι εύλογοι τρόποι για την ανίχνευση πλαστών βίντεο ή εικόνων. Η πιο αναμενόμενη μέθοδος για τον σκοπό αυτό είναι η αναζήτηση της προέλευσης της εικόνας, κάτι που βέβαια δεν υπάρχει σαν δυνατότητα προς το παρόν. Οι Hasan και Salah [67] πρότειναν ένα εργαλείο για αυτόν τον σκοπό, στηριζόμενοι στην υπόθεση ότι τα πραγματικά βίντεο προέρχονται από ανιχνεύσιμες πηγές σε αντίθεση με τα πλαστά. Συγκεκριμένα πρότειναν τη χρήση αλυσίδας συστοιχιών (blockchain) και έξυπνων συμβολαίων για να βοηθήσουν τους χρήστες να εντοπίσουν πλαστά βίντεο. Κάθε βίντεο σχετίζεται με ένα έξυπνο συμβόλαιο που συνδέεται με το γονικό του βίντεο και κάθε

γονικό βίντεο έχει έναν σύνδεσμο με το παιδί του σε μια ιεραρχική δομή. Μέσω αυτής της αλυσίδας οι χρήστες μπορούν να εντοπίσουν το αρχικό έξυπνο συμβόλαιο το οποίο συσχετίζεται με το γνήσιο βίντεο, ασχέτως αν το βίντεο αυτό έχει αντιγραφεί πολλές φορές. Ένα σημαντικό χαρακτηριστικό του έξυπνου συμβολαίου είναι τα μοναδικά κλειδιά (hashes) του παγκόσμιου συστήματος αρχείων, τα οποία χρησιμοποιούνται για την αποθήκευση βίντεο με τα μεταδεδομένα τους [68]. Έτσι κάθε βίντεο έχει ένα χαρακτηριστικό που το κάνει παγκοσμίως μοναδικό. Τα βασικά χαρακτηριστικά και οι λειτουργίες του έξυπνου συμβολαίου έχουν δοκιμαστεί με επιτυχία εναντίον πολλών κοινών προκλήσεων ασφαλείας. Επομένως πρόκειται για μια δοκιμασμένη πρακτική, η οποία μπορεί να επεκταθεί σε κάθε τύπο ψηφιακού περιεχομένου, όπως εικόνες, ηχητικά αρχεία ή και κείμενα.

4. Εκπαίδευση και αξιολόγηση μοντέλων ανίχνευσης πλαστών βίντεο

Στην παρούσα εργασία έγινε εκπαίδευση τεσσάρων μοντέλων μηχανικής μάθησης, με στόχο την ταξινόμηση βίντεο σε πλαστά ή πραγματικά. Τα μοντέλα που χρησιμοποιήθηκαν ήταν το R3D [10], το MC3 [11], το R2Plus1D [11] και το I3D [12]. Και τα τέσσερα αυτά μοντέλα έχουν χρησιμοποιηθεί με επιτυχία για αναγνώριση χωροχρονικών χαρακτηριστικών σε βίντεο με σκοπό τον προσδιορισμό της κατηγορίας της δραστηριότητας που παρουσιάζεται σε αυτό. Εξετάστηκε επομένως κατά πόσο αυτή η ικανότητα των μοντέλων στην κατανόηση και χωροχρονικής πληροφορίας μπορεί να τα καταστήσει αποτελεσματικά και στην ανίχνευση πλαστών βίντεο.

Ως σύνολο δεδομένων εκπαίδευσης και ελέγχου επιλέχθηκε το Celeb-DF-v2 [9]. Σε αυτό περιλαμβάνονται 590 πραγματικά βίντεο και 5639 πλαστά (deepfake) βίντεο. Τα πραγματικά βίντεο έχουν επιλεγεί από δημόσια διαθέσιμα βίντεο στο YouTube, που αντιστοιχούν σε συνεντεύξεις 59 διάσημων, σε καθένα από αυτά δηλαδή παριστάνεται κάποιος διάσημος να μιλάει. Τα βίντεο αυτά περιλαμβάνουν ποικιλία ως προς το φύλο, την ηλικία και την εθνικότητα των διάσημων προσώπων. Επιπλέον παρουσιάζουν μεταβολές στην ανάλυση της εικόνας του προσώπου, τον προσανατολισμό του, τον φωτισμό και το περιβάλλον του. Τα πλαστά βίντεο του συνόλου έχουν προέλθει από τα πραγματικά με ανταλλαγές στα πρόσωπα των ατόμων που παρουσιάζονται. Τα βίντεο αυτά, σύμφωνα με τους δημιουργούς του Celeb-DF-v2 [9], είναι πιο ρεαλιστικά και με λιγότερες ατέλειες σε σχέση με αντίστοιχα πλαστά βίντεο προηγούμενων γνωστών συνόλων δεδομένων. Η μέση διάρκειά τους είναι 13 δευτερόλεπτα με συχνότητα 30 καρέ ανά δευτερόλεπτο, ενώ η ανάλυσή της περιοχής του προσώπου είναι 256 x 256 pixels [9].

Ακολούθως παρουσιάζονται τα βήματα που έγιναν για την προεπεξεργασία των παραπάνω δεδομένων, η δομή των μοντέλων μηχανικής μάθησης που εφαρμόστηκαν και τα αποτελέσματά τους. Επιπλέον γίνεται σύγκριση των επιδόσεών των μοντέλων και αξιολόγησή τους.

4.1 Προεπεξεργασία δεδομένων

Τα βίντεο από το Celeb-DF-v2 [9] συμπεριλαμβάνουν πλεονάζουσα πληροφορία που δεν εξυπηρετεί στον εντοπισμό τυχόν ατελειών που θα οδηγήσουν σε συμπέρασμα για την αυθεντικότητά τους. Δεδομένου ότι πρόκειται για βίντεο στα οποία η αλλαγή έχει γίνει στην περιοχή του προσώπου, το μόνο ωφέλιμο κομμάτι για τα μοντέλα είναι αυτή η περιοχή. Επομένως σε πρώτη φάση χρησιμοποιήθηκε προεκπαιδευμένο μοντέλο αναγνώρισης προσώπου, με βάση το οποίο εντοπίστηκε σε κάθε στιγμιότυπο του εκάστοτε βίντεο η περιοχή του προσώπου. Για τον σκοπό αυτό δοκιμάστηκαν τα μοντέλα Haar Cascade [69] και RetinaFace [70].Το μοντέλο που προτιμήθηκε τελικά ήταν το RetinaFace, καθώς αυτό, παρά το γεγονός ότι ήταν πιο αργό, είχε καλύτερα αποτελέσματα σε αντίθεση με το πρώτο, το οποίο σε κάποια στιγμιότυπα αντί του προσώπου εντόπιζε κάποιο άλλο αντικείμενο. Επιπλέον το πλαίσιο της περιοχής του προσώπου ορίστηκε να είναι τετράγωνο με διαστάσεις σε pixels σταθερές για κάθε βίντεο ξεχωριστά, ακόμα κι αν το πρόσωπο μικραίνει, παραδείγματος χάρη λόγο απομάκρυνσής του από την κάμερα. Αυτό έγινε προκειμένου να αποφευχθεί η πρόκληση ανωμαλιών στη φυσική συνέχεια της κίνησης του προσώπου από στιγμιότυπο σε στιγμιότυπο του βίντεο. Συγκεκριμένα από κάθε βίντεο δημιουργήθηκε νέο βίντεο που δείχνει μόνο την περιοχή του προσώπου και έχει διατάσεις που αντιστοιχούν στις μέγιστες διατάσεις του προσώπου του αντίστοιχου αρχικού βίντεο.

Σε δεύτερη φάση τα παραπάνω βίντεο χωρίστηκαν σε δεδομένα εκπαίδευσης (64% του συνόλου), δεδομένα validation (16% του συνόλου) και δεδομένα test (20% του συνόλου). Καθένα από αυτά τα βίντεο χωρίστηκε σε στιγμιότυπα, δηλαδή σε εικόνες. Δεν κρατήθηκαν όλα τα στιγμιότυπα κάθε βίντεο, αλλά έγινε δειγματοληψία ανά 4 στιγμιότυπα, υπό τον όρο ότι με αυτή τη συχνότητα δειγματοληψίας πρόκυπταν τουλάχιστον 16 στιγμιότυπα για το εκάστοτε βίντεο. Για όσα βίντεο δεν πληρούσαν αυτόν τον όρο έγινε δειγματοληψία ανά 3 ή 2 ή 1 στιγμιότυπο, αναλόγως σε ποια από αυτές τις περιπτώσεις ίσχυε το κριτήριο των τουλάχιστον 16 στιγμιότυπων. Η διαδικασία αυτή έγινε δύο φορές. Τη μία φορά τα στιγμιότυπα που κρατήθηκαν ορίστηκε να έχουν μέγεθος 128 x 128 pixels, ενώ την άλλη φορά ορίστηκε το μέγεθος των στιγμιότυπων σε 256 x 256 pixels. Επιπλέον κατά την εκπαίδευση των μοντέλων από κάθε βίντεο επιλέγονταν τυχαία 16 στιγμιότυπα στη σειρά και γινόταν επίσης τυχαία περικοπή της εικόνας κάθε στιγμιότυπου σε 112 x 112 pixels για όσα μοντέλα εκπαιδεύονταν στα στιγμιότυπα των 128 x 128 pixels ή σε 224 x 224 pixels για όσα μοντέλα εκπαιδεύονταν σε στιγμιότυπα των 256 x 256 pixels. Η θέση του πλαισίου της τυχαίας περικοπής που εφαρμόστηκε ήταν κοινή και για τα 16 στιγμιότυπα που επιλέγονταν από κάθε βίντεο κατά την εκπαίδευση. Στην Εικόνα 18 φαίνεται στιγμιότυπο ενός από τα βίντεο του συνόλου δεδομένων, στο οποίο αποκόπτεται η περιοχή του προσώπου.

Εικόνα 18: Απεικόνιση της διαδικασίας περικοπής της περιοχής του προσώπου σε κάθε στιγμιότυπο

4.2 Βασικά δομικά στοιχεία συνελικτικών δικτύων

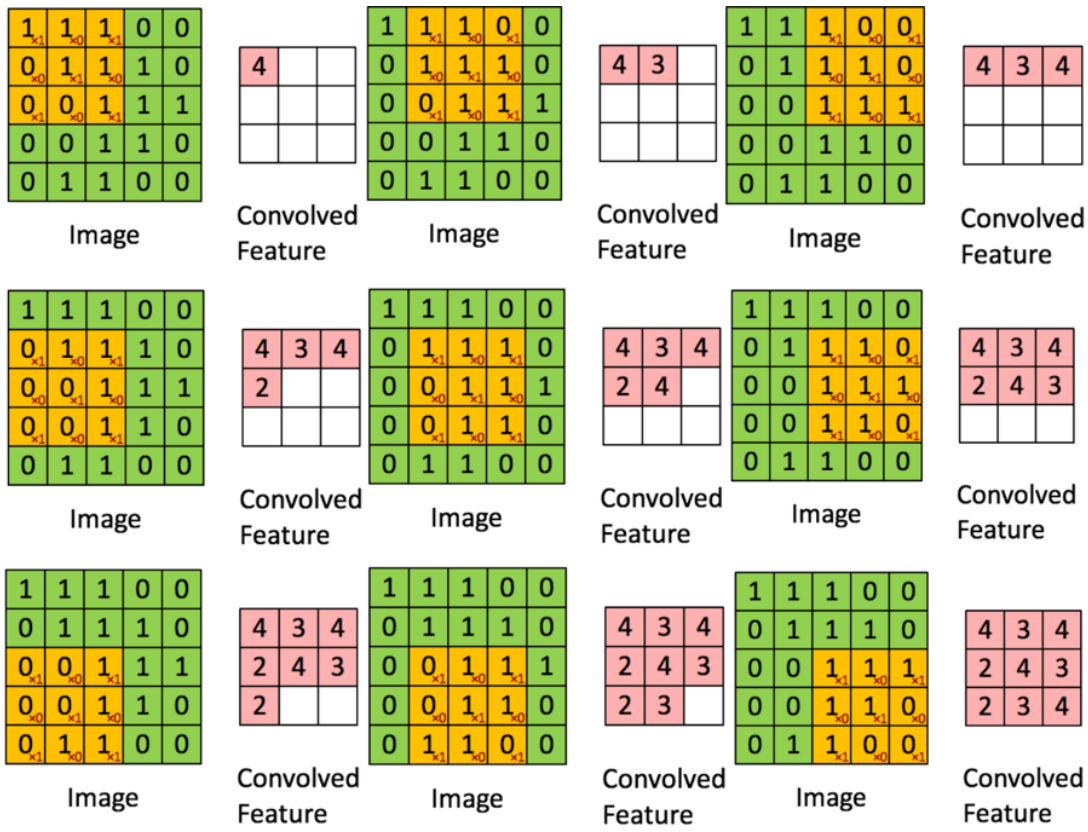
Τα μοντέλα που χρησιμοποιήθηκαν στην παρούσα εργασία ανήκουν στην κατηγορία των συνελικτικών νευρωνικών δικτύων και συγκεκριμένα στην υποκατηγορία των τρισδιάστατων συνελικτικών νευρωνικών δικτύων. Τα Συνελικτικά Νευρωνικά Δίκτυα περιλαμβάνουν ακολουθίες από συνελικτικά επίπεδα. Ένα συνελικτικό επίπεδο (convolutional layer) είναι ουσιαστικά ένα σύνολο από νευρώνες που εκτελούν συνέλιξη των φίλτρων που έχουν προκαθοριστεί, με τον τανυστή (συνήθως εικόνα ή σειρά από εικόνες) που δέχονται στην είσοδο. Κάθε επίπεδο μπορεί να περιλαμβάνει νευρώνες που εκτελούν συνέλιξη, διαδικασίες pooling, εισαγωγή μη γραμμικότητας ή ακόμη και κανονικοποίηση, ενώ έχει διακριτές εισόδους και εξόδους. Οι διαστάσεις των φίλτρων που περιλαμβάνουν, ο αριθμός τους και το βάθος τους μπορεί να διαφέρει σημαντικά ανάλογα με το πρόβλημα. Ακολούθως παρουσιάζονται και επεξηγούνται μερικά από τα βασικά δομικά στοιχεία της αρχιτεκτονικής μοντέλων συνέλιξης.

4.2.1 Είσοδος συνελικτικών δικτύων

Η είσοδος των συνελικτικών δικτύων είναι ένας τανυστής (tensor), δηλαδή μια δομή πολλών διαστάσεων. Δεδομένου ότι τα δίκτυα αυτά χρησιμοποιούνται κυρίως στο πεδίο της Όρασης Υπολογιστών, τις περισσότερες φορές η είσοδος που δέχονται είναι εικόνα ή ακολουθία εικόνων (βίντεο). Ο ρόλος τους είναι μέσω επαναλαμβανόμενων συνελίξεων και άλλων βοηθητικών στρωμάτων να συρρικνώσουν την είσοδο που δέχονται μετατρέποντάς την σε μια μορφή που κάνει την επεξεργασία της ευκολότερη, ενώ παράλληλα διατηρεί τα χαρακτηριστικά που είναι σημαντικά για την ορθή πρόβλεψη της εξόδου [71].

4.2.2 Συνελικτικό στρώμα

Το πιο σημαντικό δομικό στοιχείο των συνελικτικών μοντέλων είναι το στρώμα συνέλιξης. Η έξοδος αυτού του στρώματος είναι στην πραγματικότητα ένας χάρτης χαρακτηριστικών της εισόδου. Το κάθε συνελικτικό επίπεδο προσδιορίζεται από έναν αριθμό συνελικτικών φίλτρων, τις διαστάσεις αυτών των φίλτρων και το βήμα συνέλιξης (stride). Τα πιο συνηθισμένα συνελικτικά φίλτρα είναι δύο διαστάσεων n x m. Η διαδικασία της δυσδιάστατης συνέλιξης με αυτά τα φίλτρα συνίσταται στο πέρασμα του κέντρου τους από το κάθε εικονοστοιχείο (pixel) της εικόνας-εισόδου και την εφαρμογή συνέλιξης. Αυτό ισχύει βέβαια στην περίπτωση που το βήμα συνέλιξης είναι 1 x 1. Αν το βήμα είναι πάνω από 1 σε μία ή περισσότερες διαστάσεις (π.χ. 2 x 2 ή 1 x 2), τότε το κέντρο κάθε φίλτρου δεν περνά από όλα τα εικονοστοιχεία, αλλά προχωράει στην κάθε διάσταση της εισόδου ανά όσα εικονοστοιχεία-τιμές ορίζει το βήμα. Το αποτέλεσμα του εκάστοτε βήματος συνέλιξης είναι το άθροισμα των γινομένων που προκύπτουν από τους πολλαπλασιασμούς των παραμέτρων του φίλτρου με τις τιμές των n x m εικονοστοιχείων που καλύπτονται στο αντίστοιχο πέρασμα. Στην περίπτωση που υπάρχει και κατώφλι (bias) προστίθεται και αυτό στο τελικό αποτέλεσμα της συνέλιξης. Αυτή η διαδικασία επαναλαμβάνεται για όλες τις τιμές της εισόδου και από όλα τα συνελικτικά φίλτρα του στρώματος. Στην Εικόνα 19 απεικονίζεται η διαδικασία της δισδιάστατης συνέλιξης χωρίς πρόσθεση κατωφλίου για είσοδο 5 x 5, φίλτρο συνέλιξης 3 x 3 και βήμα 1 x 1 [71].



Εικόνα 19: Εικονική αναπαράσταση δυσδιάστατης συνέλιξης με φίλτρο 3 x 3

Όπως φαίνεται και από το παραπάνω παράδειγμα συνέλιξης το αποτέλεσμα της εξόδου έχει μικρότερες διαστάσεις από την είσοδο. Αυτό οφείλεται στο ότι το συνελικτικό φίλτρο είναι 3 x 3 και επομένως το κέντρο του δε μπορεί να περάσει από όλες τις τιμές τις εισόδου, γιατί θα προεξέχουν οι ακριανές παράμετροί του και δε θα καλύπτουν κάποια τιμή της εισόδου ώστε να πραγματοποιηθεί ο πολλαπλασιασμός τους με αυτή. Το πρόβλημα αυτό μπορεί να αποφευχθεί, εάν προστεθεί το ανάλογο εκτόπισμα (padding) στην εικόνα πριν την εφαρμογή φίλτρου, αν δηλαδή η εικόνα επεκταθεί κατά $\frac{n-1}{2}$ (εφόσον έχουμε φίλτρο n x n) προς κάθε διάσταση. Η εικόνα του παραδείγματος σε αυτή την περίπτωση θα επεκτεινόταν από 5 x 5 σε 6 x 6, αποκτώντας ένα περίβλημα πάχους ενός pixel το οποίο μπορεί να είχε π.χ. μηδενικές τιμές. Γενικά είναι συχνή πρακτική να επεκτείνονται οι τιμές της εισόδου μιας συνέλιξης προς κάθε διάσταση κατά τόσο όσο χρειάζεται για να μπορεί να περάσει το κέντρο του συνελικτικού φίλτρου από όλες τις αρχικές τιμές τις εισόδου.

Η συνέλιξη που παρουσιάστηκε στο παραπάνω παράδειγμα είναι με φίλτρο δύο διαστάσεων. Σε πολλές περιπτώσεις, όπως και στα μοντέλα που εκπαιδεύτηκαν στην παρούσα εργασία, χρησιμοποιούνται συνελικτικά φίλτρα τριών διαστάσεων (3D CNN) [11]. Σε αυτές τις περιπτώσεις η είσοδος συνήθως δεν είναι μία εικόνα, αλλά μια ακολουθία εικόνων και η συνέλιξη δε γίνεται μόνο ως προς τις διαστάσεις του χώρου, αλλά και ως προς τη διάσταση του χρόνου.

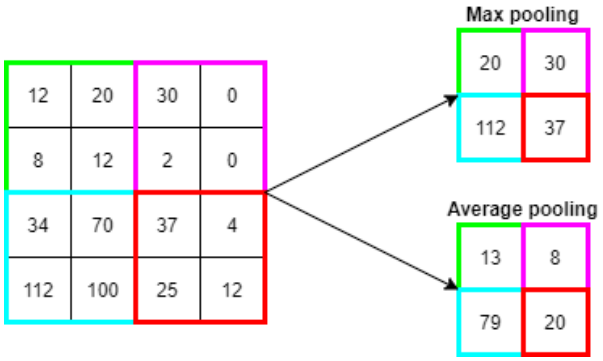
4.2.3 Κανονικοποίηση δέσμης δεδομένων (Batch normalization)

Η κανονικοποίηση δέσμης δεδομένων (batch normalization) [72] συνήθως τοποθετείται στην έξοδο του συνελικτικού στρώματος. Αντιμετωπίζει προβλήματα αστάθειας τα οποία προκύπτουν κατά την εκπαίδευση εξαιτίας της αλλαγής της κατανομής των δεδομένων που προκαλείται από τη διέλευσή τους από ένα στρώμα. Η κανονικοποίηση

πραγματοποιείται για κάθε δέσμη δεδομένων προσθέτοντας παραμέτρους κανονικοποίησης κατά την οπίσθια διάδοση του σφάλματος (back propagation of error). Με τη χρήση της έχει αποδειχτεί ότι τα νευρωνικά δίκτυα επιτυγχάνουν γρηγορότερη σύγκλιση, με μεγαλύτερη ακρίβεια.

4.2.4 Στρώμα συνένωσης (pooling layer)

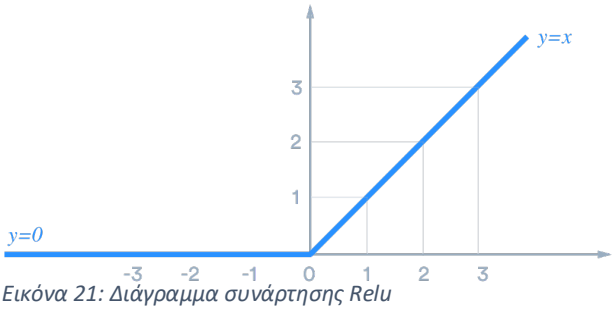
Όπως προαναφέρθηκε, στόχος των συνελικτικών δικτύων είναι να συρρικνώσουν την είσοδο που δέχονται μετατρέποντάς την σε μια μορφή που κάνει την επεξεργασία της ευκολότερη, ενώ παράλληλα διατηρεί τα χαρακτηριστικά που είναι σημαντικά για την ορθή πρόβλεψη της εξόδου [71]. Για τη μείωση του μεγέθους της εισόδου, ανάμεσα στα συνελικτικά δίκτυα χρησιμοποιούνται στρώματα συνένωσης (pooling layers). Υπάρχουν διάφοροι τύποι τέτοιων στρωμάτων ανάλογα με το κριτήριο που χρησιμοποιούν για τη συνένωση τιμών. Οι δύο πιο συνηθισμένες περιπτώσεις συνένωσης είναι η συνένωση μέσου όρου και η συνένωση μέγιστης τιμής. Ένα φίλτρο συνένωσης έχει τις διαστάσεις που ορίζονται από το μοντέλο και διαπερνά την είσοδο του στρώματός του με τον τρόπο που περνά και ένα φίλτρο συνέλιξης κάνοντας όμως άλλη διαδικασία από αυτή της συνέλιξης. Συγκεκριμένα με βάση κάποιο κριτήριο συνενώνει σε μία τιμή τις τιμές από τις οποίες περνά σε κάθε βήμα. Αν πρόκειται για φίλτρο μέγιστης τιμής η νέα τιμή που προκύπτει είναι η μέγιστη από τις τιμές που καλύπτει το φίλτρο αυτό κατά το συγκεκριμένο βήμα. Αν πρόκειται για φίλτρο μέσου όρου, η νέα τιμή είναι ο μέσος όρος των τιμών που καλύπτονται από το φίλτρο στο συγκεκριμένο βήμα. Η διαδικασία που ακολουθείται για τους δύο παραπάνω τύπους συνένωσης παρουσιάζεται με παράδειγμα στην Εικόνα 20.



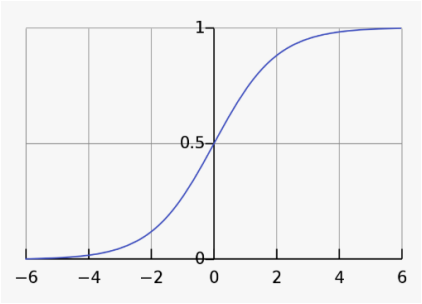
Εικόνα 20: Εικονική αναπαράσταση συνένωσης μέγιστης τιμής και συνένωσης μέσου όρου με φίλτρο 2 x 2

4.2.5 Συναρτήσεις ενεργοποίησης

Στην έξοδο κάθε στρώματος νευρώνων εφαρμόζεται συνήθως κάποια συνάρτηση ενεργοποίησης. Για τα συνελικτικά στρώματα η συνάρτηση που συνήθως εφαρμόζεται είναι η Relu. Η Relu επιτρέπει να περάσουν στην έξοδο του στρώματος όλες οι τιμές που είναι μεγαλύτερες ή ίσες με το 0 και μηδενίζει όλες τις αρνητικές τιμές, όπως φαίνεται και στο διάγραμμα που απεικονίζεται στην Εικόνα 21. Συνήθως στο τελευταίο στρώμα του νευρωνικού δικτύου η συνάρτηση ενεργοποίησης αντί για τη Relu είναι η Softmax, η οποία προσαρμόζει όλες τις τιμές εξόδου μεταξύ 0 και 1. Το διάγραμμά της απεικονίζεται στην Εικόνα 22.



Εικόνα 21: Διάγραμμα συνάρτησης Relu



Εικόνα 22: Διάγραμμα συνάρτησης Softmax

4.3 Μοντέλα που χρησιμοποιήθηκαν

4.3.1 Μοντέλο R3D

Η αρχιτεκτονική του μοντέλου R3D [10] που χρησιμοποιήθηκε βασίζεται σε τρισδιάστατα συνελικτικά δίκτυα (3D CNNs) [11]. Τα δίκτυα αυτά έχουν το προτέρημα ότι λαμβάνουν υπόψη όχι μόνο τις διαστάσεις του χώρου αλλά και τη διάσταση του χρόνου, δηλαδή μπορούν να βγάλουν συμπέρασμα για ένα βίντεο χωρίς να βασίζονται μόνο στην στατική πληροφορία κάθε στιγμιότυπου, αλλά συνυπολογίζοντας και την χρονική ακολουθία και μεταβολή αυτής της πληροφορίας σε μια σειρά στιγμιότυπων. Για τον λόγο αυτό τέτοια δίκτυα έχουν χρησιμοποιηθεί για αναγνώριση δραστηριότητας σε βίντεο, καθώς η δραστηριότητα συχνά ορίζεται από μια ακολουθία κινήσεων οι οποίες όμως αν γίνουν με διαφορετική σειρά ή αναμειχθούν με άλλες κινήσεις μπορεί να αποτελούν μια άλλη διαφορετική δραστηριότητα. Στην προκειμένη περίπτωση που μελετάται το αν κάποιο βίντεο είναι πλαστό ή όχι, η δυνατότητα αυτή του δικτύου είναι εξίσου χρήσιμη, επειδή θα μπορεί να συνυπολογίσει τυχόν ατέλειες που προκύπτουν στις μεταβάσεις μεταξύ των στιγμιότυπων σε πλαστά βίντεο.

Το μοντέλο R3D [10] που χρησιμοποιήθηκε δέχεται είσοδο διαστάσεων 3 x 16 x 112 x 112, όπου το 3 αντιστοιχεί στον αριθμό των RGB καναλιών της εικόνας-στιγμιότυπου, το 16 αντιστοιχεί στον αριθμό των στιγμιότυπων και το 112 x 112 είναι το ύψος επί το πλάτος της εικόνας. Αποτελείται συνολικά από 18 στρώματα. Το πρώτο στρώμα αντιστοιχεί σε τρισδιάστατη συνέλιξη της εισόδου με 64 συνελικτικά φίλτρα, των οποίων ο πυρήνας (kernel) είναι διαστάσεων 3 x 7 x 7 και το βήμα (stride) 1 x 2 x 2. Το εκτόπισμα (padding) είναι 1 x 3 x 3, ώστε να είναι αντίστοιχο με τον πυρήνα, να μπορεί δηλαδή το κέντρο του συνελικτικού φίλτρου να περάσει και από όλες τις ακριανές τιμές κάθε διάστασης. Η έξοδος αυτού του πρώτου συνελικτικού στρώματος είναι διαστάσεων 64 x 16 x 56 x 56. Ακολούθως στην έξοδο αυτή εφαρμόζεται κανονικοποίηση (batch normalization), καθώς και η συνάρτηση ενεργοποίησης Relu, η οποία μηδενίζει όλες τις αρνητικές τιμές.

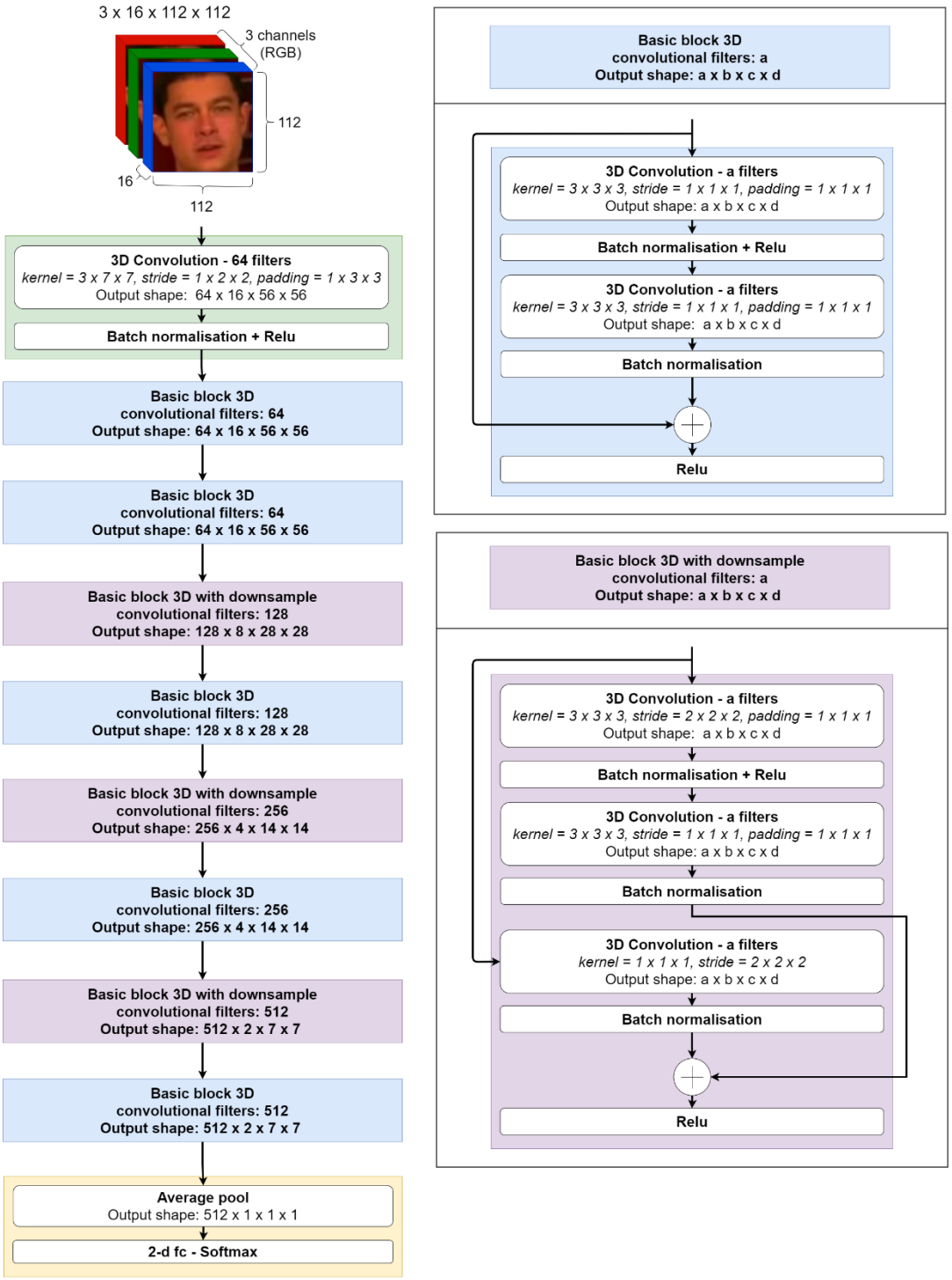
Ακολουθεί μια σειρά από 8 υπολειμματικά νευρωνικά δίκτυα (residual networks), τα οποία προσθέτουν στην έξοδο μιας ακολουθίας 2 συνελικτικών στρωμάτων την είσοδο αυτής της ακολουθίας. Πιο συγκεκριμένα υπάρχουν δύο τύπου δομές που εναλλάσσονται στα 8 αυτά επίπεδα. Η πρώτη περιλαμβάνει μια σειρά από 2 συνελίξεις ίδιου αριθμού φίλτρων με πυρήνα 3 x 3 x 3, βήμα συνέλιξης 1 x 1 x 1 και εκτόπισμα 1 x 1 x 1. Η έξοδος της πρώτης από τις δύο συνελίξεις κανονικοποιείται και περνά από τη συνάρτηση ενεργοποίησης Relu, ενώ η έξοδος της δεύτερης περνά μόνο από κανονικοποίηση σε πρώτη φάση. Ακολούθως προστίθεται σε αυτήν η είσοδος της προηγούμενης συνέλιξης και το τελικό αποτέλεσμα περνά από την Relu και έχει διαστάσεις ίδιες με την αρχική είσοδο της δομής αυτής. Ο δεύτερος τύπος υπολειματικής δομής είναι ίδιος, με τη διαφορά ότι

πραγματοποιείται συρρίκνωση του μεγέθους της εισόδου (downsample) και συγκεκριμένα υποδιπλασιασμός του μεγέθους των τριών διαστάσεών της. Για τον σκοπό αυτό στην πρώτη συνέλιξη αυτής της δομής το βήμα είναι διπλό, δηλαδή 2 x 2 x 2. Επιπλέον η έξοδος της δεύτερης συνέλιξης δεν προστίθεται απευθείας στην είσοδο της δομής, γιατί δεν έχει ίδιες διατάσεις, αφού κατά την πρώτη συνέλιξη μειώθηκε το μέγεθος των διαστάσεων εξαιτίας του διπλού βήματος συνέλιξης και αυτό το μέγεθος διατηρήθηκε στη δεύτερη συνέλιξη που είχε μονό βήμα. Για τον λόγο αυτό, πρώτα παραγματοποιείται στην είσοδο της δομής συρρίκνωση του μεγέθους της (downsample), ώστε να ταιριάζει με την έξοδο της δεύτερης συνέλιξης. Η συρρίκνωση αυτή γίνεται με χρήση συνέλιξης με πυρήνα 1 x 1 x 1 και βήμα 2 x 2 x 2. Τ ο αποτέλεσμα της συρρίκνωσης προστίθεται τελικά στην έξοδο της δεύτερης συνέλιξης και ακολουθεί η συνάρτηση ενεργοποίησης Relu.

Οι δύο τύποι υπολειμματικών δομών που παρουσιάστηκαν εναλλάσσονται ως εξής: Ο πρώτος τύπος επαναλαμβάνεται δύο φορές μετά το αρχικό στρώμα και περιλαμβάνει 64 συνελικτικά φίλτρα. Ακολουθεί ο δεύτερος τύπος, που συμπεριλαμβάνει δηλαδή συρρίκνωση του δείγματος (downsample), με 128 φίλτρα. Μετά επαναλαμβάνεται ο πρώτος τύπος πάλι με 128 φίλτρα. Το ίδιο ζευγάρι δεύτερου τύπου ακολουθούμενο από τον πρώτο τύπο με ίσο αριθμό φίλτρων μεταξύ τους, επαναλαμβάνεται ακόμη 2 φορές, μία με 256 φίλτρα και μία με 512 φίλτρα. Η έξοδος που προκύπτει από αυτή τη σειρά υπολειμματικών νευρωνικών δικτύων περνάει στο επόμενο στρώμα το οποίο κάνει συνένωση μέσου όρου (average pooling) οδηγώντας σε διαστάσεις εξόδου 512 x 1 x 1 x 1. Οι 512 αυτές τιμές περνάνε στο τελικό πλήρως συνδεδεμένο στρώμα εξόδου, το οποίο έχει τόσους νευρώνες όσες οι κλάσεις, δηλαδή στην παρούσα υλοποίηση έχει δύο νευρώνες, έναν για τα πραγματικά βίντεο και ένα για τα πλαστά βίντεο. Η συνάρτηση ενεργοποίησης του τελικού αυτού στρώματος εξόδου είναι η Softmax.

Συνολικά το δίκτυο έχει 33.17 εκατομμύρια παραμέτρους για εκπαίδευση. Τα αρχικά βάρη που αντιστοιχούν σε αυτές τις παραμέτρους δεν ορίστηκαν τυχαία, αλλά το δίκτυο ξεκίνησε την εκπαίδευση όντας προεκπαιδευμένο στο σύνολο δεδομένων Kinetics. Το σύνολο αυτό έχει ως δεδομένα βίντεο από 400 διαφορετικές δραστηριότητες τις οποίες αναγνωρίζει το μοντέλο. Προκειμένου να χρησιμοποιηθεί η προϋπάρχουσα γνώση αυτού του δικτύου για το πρόβλημα της ταξινόμησης βίντεο σε πλαστά και μη, αρχικοποιήθηκε το μοντέλο με τα βάρη που προέκυψαν από την εκπαίδευσή του στο Kinetics [73] και έγινε αλλαγή στο τελευταίο στρώμα, ώστε να μην καταλήγει σε 400 νευρώνες εξόδου, αλλά σε 2, καθώς οι κλάσεις πλέον είναι δύο, τα πλαστά και τα πραγματικά βίντεο. Η δομή του μοντέλου φαίνεται αναλυτικά στην Εικόνα 23.

Μοντέλο R3D [10]

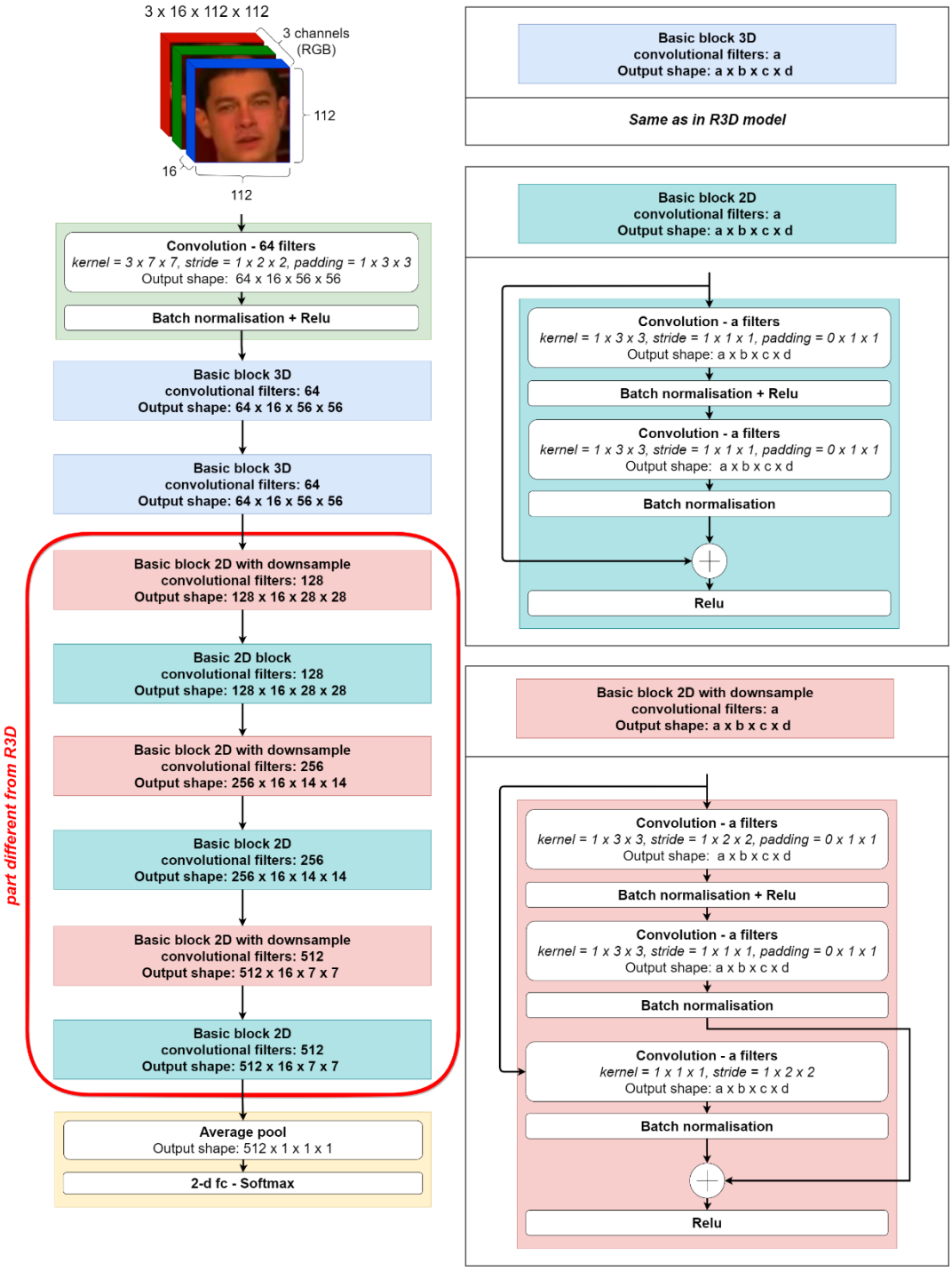


Εικόνα 23: Διαγραμματική απεικόνιση του μοντέλου R3D, που χρησιμοποιήθηκε στην παρούσα εργασία

4.3.2 Μοντέλο MC3

Το μοντέλο MC3 [11] έχει αντίστοιχη αρχιτεκτονική με το μοντέλο R3D με τη διαφορά ότι δεν περιλαμβάνουν όλα τα στρώματα τρισδιάστατες συνελίξεις. Για την ακρίβεια το μοντέλο αυτό είναι όμοιο με το R3D στο πρώτο στρώμα, αλλά και στα δύο επόμενα, δηλαδή στις δύο δομές υπολειμματικών νευρωνικών δικτύων με 64 φίλτρα χωρίς συρρίκνωση της εισόδου. Το μοντέλο MC3 ξεκινάει να διαφοροποιείται από την τρίτη δομή υπολειμματικού δικτύου, όπου αρχίζουν τα ζευγάρια υπολειμματικής δομής με συρρίκνωση της εισόδου (downsample) ακολουθούμενης από απλή υπολειμματική δομή με ίσο αριθμό φίλτρων. Η διαφορά του με το R3D είναι ότι στις δομές αυτές οι συνελίξεις που πραγματοποιούνται δεν είναι τρισδιάστατες, αλλά δισδιάστατες. Συγκεκριμένα έχουν πυρήνα 1 x 3 x 3, αντί για 3 x 3 x 3, ενώ το διπλό βήμα συνέλιξης στις περιπτώσεις συρρίκνωσης της εισόδου είναι 1 x 2 x 2, αντί για 2 x 2 x 2. Το τελικό στρώμα της συνένωσης μέσου όρου και της εξόδου παραμένει ίδιο με το αντίστοιχο στρώμα του R3D. Το μοντέλο MC3 στοχεύει στο να πετύχει τα αποτελέσματα του R3D με λιγότερες παραμέτρους εκπαίδευσης, δηλαδή με πιο απλό και γρήγορο τρόπο. Συνολικά οι παράμετροι εκπαίδευσής του στην παρούσα υλοποίηση είναι 11.49 εκατομμύρια, δηλαδή λιγότερες από τις μισές παραμέτρους του R3D. Επιπλέον και σε αυτή την περίπτωση το μοντέλο που χρησιμοποιήθηκε είχε προεκπαιδευτεί στο σύνολο δεδομένων Kinetics [73]. Η ακριβής δομή του μοντέλου MC3 φαίνεται στην Εικόνα 24, όπου με κόκκινο πλαίσιο περιβάλλεται το κομμάτι που διαφοροποιείται από το μοντέλο R3D.

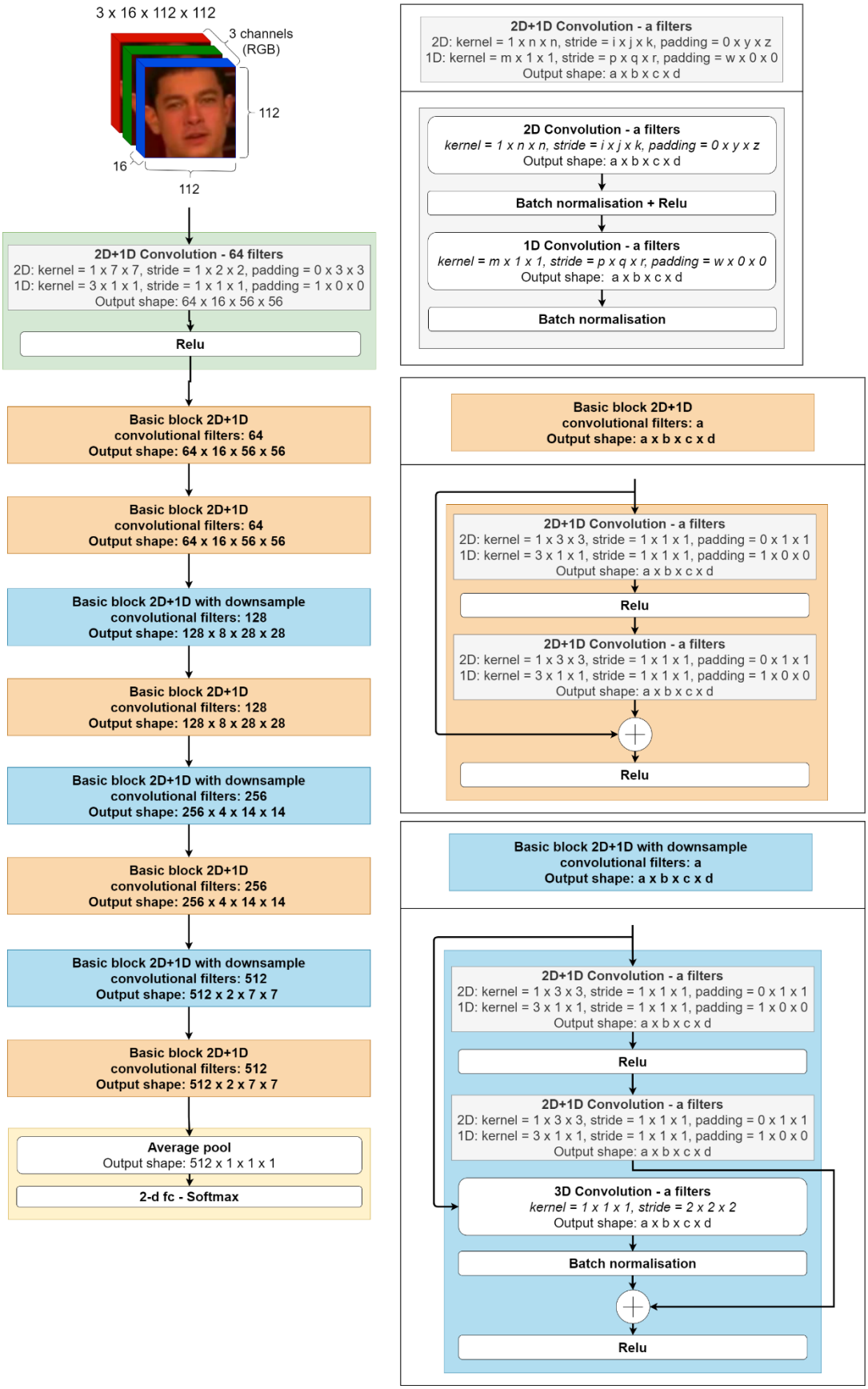
Μοντέλο MC3



Εικόνα 24: Διαγραμματική απεικόνιση του μοντέλου MC3, που χρησιμοποιήθηκε στην παρούσα εργασία

4.3.3 Μοντέλο R2Plus1D

Το μοντέλο R2Plus1D [11] βασίζεται επίσης στο μοντέλο R3D [10], αλλά αντί για τρισδιάστατη συνέλιξη, χρησιμοποιεί συνέλιξη δύο διαστάσεων ακολουθούμενη από συνέλιξη μίας διάστασης. Επομένως στο πρώτο στρώμα, αντί για 64 φίλτρα τρισδιάστατης συνέλιξης με πυρήνα 3 x 7 x 7, βήμα 1 x 2 x 2 και εκτόπισμα 1 x 1 x 1, έχει 64 φίλτρα δισδιάστατης συνέλιξης με πυρήνα 1 x 7 x 7, βήμα 1 x 2 x 2 και εκτόπισμα 0 x 1 x 1, ακολουθούμενα από 64 φίλτρα μονοδιάστατης συνέλιξης με πυρήνα 3 x 1 x 1, βήμα 1 x 1 x 1 και εκτόπισμα 1 x 0 x 0. Ουσιαστικά δηλαδή στο πρώτο βήμα κάνει συνέλιξη των δύο τελευταίων διαστάσεων και στο δεύτερο βήμα κάνει συνέλιξη της πρώτης από τις τρεις διαστάσεις. Κάθε συνέλιξη ακολουθείται από κανονικοποίηση (batch normalization) και τη συνάρτηση ενεργοποίησης Relu. Με την ίδια λογική μετατρέπονται και οι υπόλοιπες τρισδιάστατες συνελίξεις του R3D σε συνελίξεις 2D+1D. Επομένως στις συνελίξεις που το R3D είχε πυρήνα 3 x 3 x 3, βήμα 1 x 1 x 1 και εκτόπισμα 1 x 1 x 1, το R2Plus1D έχει μια δισδιάστατη συνέλιξη με πυρήνα 1 x 3 x 3, βήμα 1 x 1 x 1 και εκτόπισμα 0 x 1 x 1, ακολουθούμενη από μονοδιάστατη συνέλιξη με πυρήνα 3 x 1 x 1, βήμα 1 x 1 x 1 και εκτόπισμα 1 x 0 x 0. Αντίστοιχα όταν το βήμα της συνέλιξης διπλασιάζεται στις δομές υπολειμματικών δικτύων που περιλαμβάνουν συρρίκνωση της εισόδου, τότε για την πρώτη δισδιάστατη συνέλιξη το βήμα γίνεται 1 x 2 x 2 και για την δεύτερη μονοδιάστατη συνέλιξη γίνεται 2 x 1 x 1. Κατά τα άλλα το δίκτυο είναι ίδιο με το μοντέλο R3D. Ο αριθμός των παραμέτρων του είναι 31.30 εκατομμύρια, μικρότερος δηλαδή από του μοντέλου R3D κατά περίπου 10%. Τα αρχικά βάρη του μοντέλου προέρχονται, όπως και για τα δύο προηγούμενα μοντέλα, από προηγούμενη εκπαίδευσή του στο σύνολο δεδομένων Kinetics [73]. Αναλυτική και ακριβής απεικόνιση της δομής του μοντέλου R2Plus1D παρουσιάζεται στην Εικόνα 25.

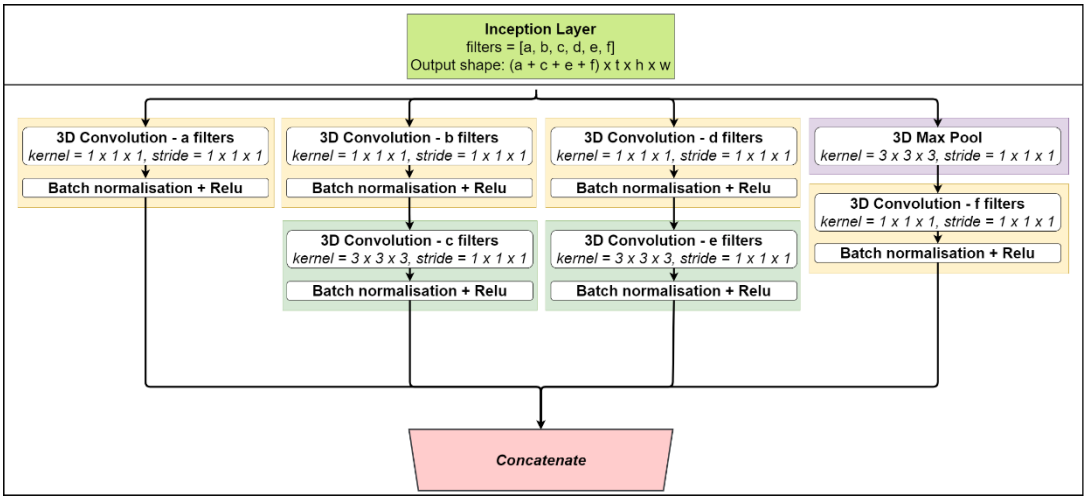


Εικόνα 25: Διαγραμματική απεικόνιση του R2Plus1D, που χρησιμοποιήθηκε στην παρούσα εργασία

4.3.4 Μοντέλο I3D

Το μοντέλο I3D [12] βασίζεται επίσης σε τρισδιάστατες συνελίξεις και είναι αρκετά αποτελεσματικό στην κατανόηση χωροχρονικής πληροφορίας. Η αρχιτεκτονική του είναι αρκετά διαφορετική από τα προηγούμενα μοντέλα. Βασικό δομικό του στοιχείο είναι το Inception, το οποίο είναι ένα δίκτυο συνελίξεων των οποίων τα αποτελέσματα συνενώνονται. Συγκεκριμένα η είσοδος αυτού του δικτύου περνά από μια συνέλιξη με πυρήνα 1 x 1 x 1 και βήμα 1 x 1 x 1. Η ίδια είσοδος περνά και από ένα ζευγάρι δύο συνελίξεων σε σειρά. Η πρώτη συνέλιξη είναι ίδια με την προαναφερθείσα συνέλιξη, ενώ η δεύτερη έχει τη διαφορά ότι ο πυρήνας της είναι 3 x 3 x 3. Η είσοδος επίσης περνά κι από ένα ακόμη ζευγάρι συνελίξεων ίδιο με το προηγούμενο. Τέλος η είσοδος περνά και από ένα στρώμα συνένωσης με βάση τη μέγιστη τιμή (max pool) ακολουθούμενο από συνέλιξη με πυρήνα 1 x 1 x 1 και βήμα 1 x 1 x 1. Το φίλτρο μέγιστης τιμής (max pool) είναι 3 x 3 x 3 και το βήμα του είναι 1 x 1 x 1. Από τις 4 παραπάνω δομές από της οποίες περνά η είσοδος προκύπτουν 4 αντίστοιχες έξοδοι. Οι έξοδοι αυτές συνενώνονται (concatenate) σε μία τελική έξοδο του τμήματος Inception. Στην Εικόνα 26 φαίνεται η αρχιτεκτονική του Inception διαγραμματικά. Οι συνελίξεις που περιλαμβάνει κάθε τμήμα Inception είναι συνολικά 6 (1 συνέλιξη στην πρώτη δομή που περνά η είσοδος, 4 συνελίξεις στα δύο επόμενα ζευγάρια συνελίξεων, στο καθένα από τα οποία εισέρχεται επίσης η είσοδος και άλλη μία μετά την συνένωση μέγιστης τιμής, από όπου πάλι περνά η είσοδος). Οι 6 αυτές συνελίξεις ποικίλουν σε κάθε τμήμα Inception ως προς τον αριθμό των φίλτρων που χρησιμοποιούν.

Δομή Inception που χρησιμοποιείται στο I3D



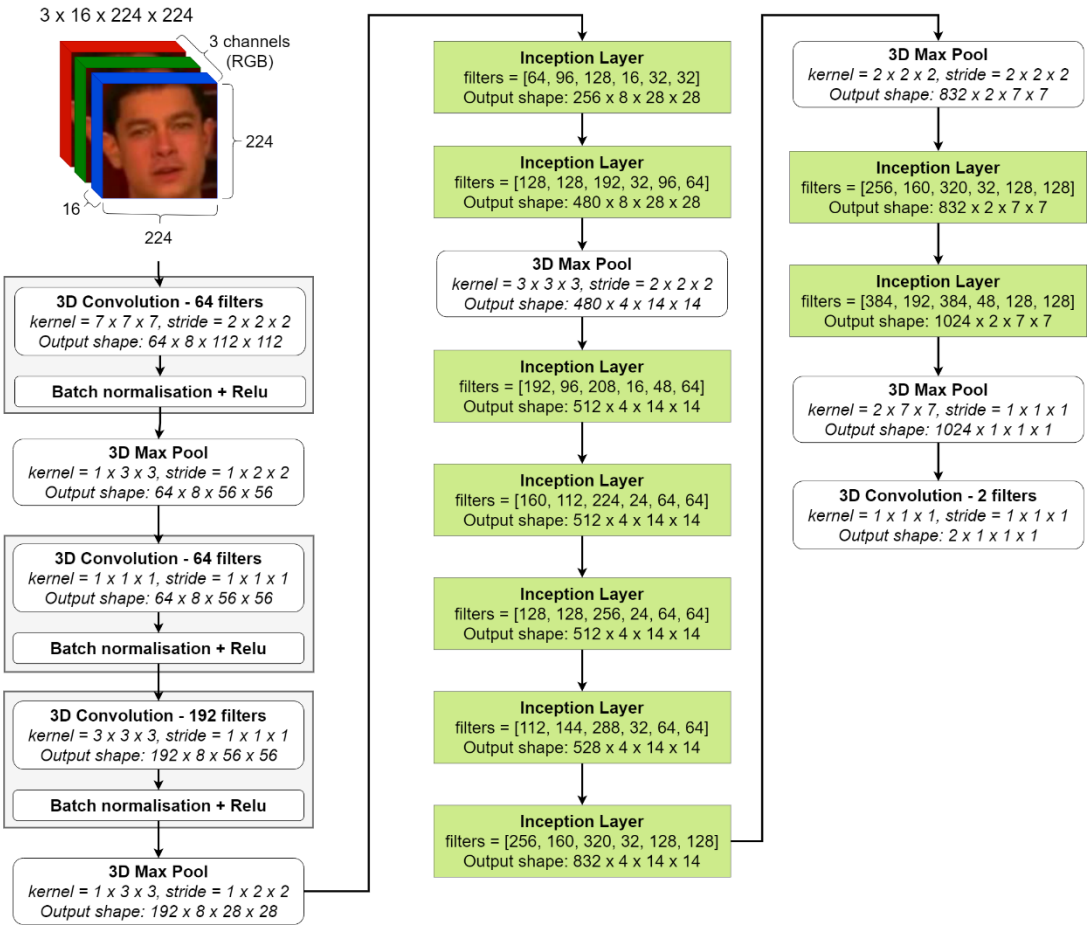
Εικόνα 26: Διαγραμματική απεικόνιση της αρχιτεκτονικής του δομικού στοιχείου Inception που χρησιμοποιείται στο μοντέλο I3D

Η συνολική αρχιτεκτονική του μοντέλου I3D χρησιμοποιεί σε διάφορα κομμάτια της την δομή Inception που αναφέρθηκε και διαμορφώνεται ως εξής: Αρχικά γίνεται συνέλιξη με πυρήνα 7 x 7 x 7 και βήμα 2 x 2 x 2. Η συνέλιξη αυτή συμπληρώνεται με κανονικοποίηση (batch normalization) και συνάρτηση ενεργοποίησης Relu, όπως και όλες οι υπόλοιπες συνελίξεις του δικτύου, εκτός της τελικής συνέλιξης του στρώματος εξόδου. Ακολουθεί συνένωση μέγιστης τιμής με πυρήνα 1 x 3 x 3 και βήμα 1 x 2 x 2. Στη συνέχεια γίνονται δύο συνελίξεις σε σειρά με βήμα 1 x 1 x 1 και πυρήνα για την πρώτη 1 x 1 x 1 και για τη δεύτερη 3 x 3 x 3. Στο επόμενο βήμα γίνεται πάλι συνένωση μέγιστης τιμής ίδια με την προηγούμενη. Ακολουθούν σε σειρά δύο δομές Inception. Ο αριθμός των φίλτρων των 6 συνελίξεών τους είναι για την πρώτη δομή 64, 96, 128, 16, 32 και 32 και για τη δεύτερη δομή 128, 128, 192, 32, 96 και 64. Ακολουθεί συνένωση μέγιστης τιμής με πυρήνα 3 x 3 x 3 και βήμα 2 x 2 x 2. Στη συνέχεια παρατάσσονται σε σειρά 5 δομές Inception. Ο αριθμός των φίλτρων των 6

συνελίξεών τους είναι για την πρώτη δομή 192, 96, 208, 16, 48 και 64, για τη δεύτερη δομή 160, 112, 224, 24, 64 και 64, για την τρίτη δομή 128, 128, 256, 24, 64 και 64, για την τέταρτη δομή 112, 144, 288, 32, 64 και 64 και για την πέμπτη δομή 256, 160, 320, 32, 128 και 128. Μετά τα 5 αυτά επίπεδα Inception ακολουθεί συνένωση μέγιστης τιμής με πυρήνα 2 x 2 x 2 και βήμα 2 x 2 x 2. Στη συνέχεια βρίσκονται σε σειρά δύο ακόμη δομές Inception. Ο αριθμός των συνελικτικών φίλτρων της πρώτης είναι 256, 160, 320, 32, 128 και 128, ενώ της δεύτερης είναι 384, 192, 384, 48, 128 και 128. Ακολούθως γίνεται μία τελευταία συνένωση μέγιστης τιμής με πυρήνα 2 x 7 x 7 και βήμα 1 x 1 x 1 και τέλος το στρώμα εξόδου είναι συνέλιξη με τόσα φίλτρα όσες και οι κλάσεις, τα οποία έχουν πυρήνα και βήμα διαστάσεων 1 x 1 x 1.

Η είσοδος που πήρε το μοντέλο στην παρούσα υλοποίηση είναι 3 x 16 x 224 x 224, όπου το 3 αντιστοιχεί στα χρωματικά κανάλια (RGB) της εικόνας, το 16 αντιστοιχεί στο πλήθος των στιγμιότυπων και το 224 x 224 στις διαστάσεις των στιγμιότυπων. Επομένως το μοντέλο αυτό πήρε εικόνες διπλάσιου μεγέθους ως είσοδο, σε σχέση με τα τρία προηγούμενα μοντέλα που έπαιρναν στιγμιότυπα διαστάσεων 112 x 112, αντί για 224 x 224. Επιπλέον το μοντέλο που χρησιμοποιήθηκε ήταν προεκπαιδευμένο στο σύνολο δεδομένων Charades [74], το οποίο περιλαμβάνει βίντεο από 157 διαφορετικούς τύπους δραστηριοτήτων. Συνολικά οι παράμετροι εκπαίδευσης του μοντέλου στην παρούσα υλοποίηση ήταν 12.29 εκατομμύρια. Αναλυτικά η δομή του μοντέλου I3D που χρησιμοποιήθηκε παρουσιάζεται στην Εικόνα 27.

Μοντέλο I3D



Εικόνα 27: Διαγραμματική απεικόνιση του μοντέλου I3D, που χρησιμοποιήθηκε στην παρούσα εργασία

4.4 Παράμετροι εκπαίδευσης

Τα παραπάνω δίκτυα εκπαιδεύτηκαν για συνολικά 30 εποχές. Για κάθε δίκτυο σε κάθε εποχή αποθηκευόταν το μοντέλο με τα βάρη του, όπως είχαν διαμορφωθεί στο τέλος της εποχής. Από τα 30 μοντέλα επομένως που κρατήθηκαν για κάθε δίκτυο, επιλέχτηκε το μοντέλο με τις καλύτερες επιδόσεις στα δεδομένα ελέγχου (test), προκειμένου να συγκριθεί με τα αντίστοιχα καλύτερα μοντέλα των υπολοίπων δικτύων.

4.4.1 Συνάρτηση σφάλματος (loss function)

Ως συνάρτηση σφάλματος χρησιμοποιήθηκε η διασταυρούμενη εντροπία (cross entropy), η οποία ενδείκνυται όταν η έξοδος του δικτύου αναπαριστά κατανομή πιθανότητας. Η τιμή του σφάλματος για είσοδο x και έξοδο y υπολογίζεται από τον τύπο της Εξίσωσης [1].

$$CrossEntropyLoss(x, y) = -\sum_{i=1}^{n} y_i \times \log \frac{e^{y_{pred,i}}}{\sum_{j=1}^{n} e^{y_{pred,j}}} \quad [1]$$

όπου y_{pred} είναι ένα διάνυσμα από n προβλέψεις του δικτύου και y είναι ένα δυαδικό διάνυσμα που περιλαμβάνει τιμές 0 και 1 ανάλογα με την κλάση στην οποία ανήκει το καθένα από τα n δείγματα εισόδου.

Στη συνάρτηση αυτή εντάχθηκαν και συντελεστές βαρύτητας για κάθε κλάση, προκειμένου να αντιμετωπιστεί το πρόβλημα της ανισορροπίας του πλήθους των δειγμάτων μεταξύ των δύο κλάσεων. Συγκεκριμένα στο σύνολο δεδομένων Celeb-DF-v2 [9] τα πραγματικά βίντεο αντιπροσωπεύουν μόνο το 10% περίπου των δειγμάτων, ενώ τα υπόλοιπα είναι πλαστά. Αυτό επιφέρει τον κίνδυνο το μοντέλο κατά την εκπαίδευση να τείνει να ταξινομεί τα βίντεο πάντα ως πλαστά, καθώς κάτι τέτοιο οδηγεί σε μικρότερο ολικό σφάλμα, αφού τα περισσότερα βίντεο ανήκουν όντως σε αυτή την κλάση. Για τον λόγο αυτό επιλέχθηκαν συντελεστές βαρύτητας για τις δύο κλάσεις, ώστε με αυτόν τον τρόπο να έχει μεγαλύτερη βαρύτητα το σφάλμα που προκύπτει από λάθος ταξινομήσεις των δεδομένων της κλάσης με τη μικρότερη αντιπροσώπευση (πραγματικά βίντεο), σε σχέση με το σφάλμα από λάθος ταξινομήσεις της άλλης κλάσης (πλαστά βίντεο). Συγκεκριμένα ορίστηκε για την κλάση των πραγματικών βίντεο συντελεστής βαρύτητας (weight) $\frac{1}{375}$ και για την κλάση των πλαστών βίντεο συντελεστής $\frac{1}{4388}$.

4.4.2 Βελτιστοποιητής (optimizer) και ρυθμός μάθησης

Ως βελτιστοποιητής (optimizer) χρησιμοποιήθηκε ο αλγόριθμος στοχαστικής κατάβασης κλίσης (stochastic gradient descent - SGD). Στη μέθοδο αυτή τα βάρη του δικτύου ενημερώνονται για κάθε δείγμα εισόδου ξεχωριστά σύμφωνα με τον τύπο της Εξίσωσης [2].

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta, x_i, y_i) \quad [2]$$

όπου x_i το δείγμα εισόδου, y_i η αντίστοιχη ετικέτα, θ τα βάρη του δικτύου, η ο ρυθμός μάθησης και $\nabla_\theta J(\theta, x_i, y_i)$ η κλίση που προκύπτει από τη συνάρτηση σφάλματος J.

Ο SGD οδηγεί σε σύγκλιση στο ελάχιστο μιας καμπύλης του χώρου των παραμέτρων του δικτύου, ενώ παράλληλα, λόγω της υψηλής διακύμανσής των ανανεώσεών του, μπορεί να μεταπηδά σε νέα ενδεχομένως καλύτερα τοπικά ελάχιστα. Ωστόσο, οι έντονες αυτές εναλλαγές δυσκολεύουν τη σύγκλιση στο ακριβές ελάχιστο, καθώς ο SGD αναπηδά συνεχώς στον χώρο των παραμέτρων. Για τον λόγο αυτό ενδείκνυται ο ρυθμός μάθησης να μειώνεται σταδιακά κατά τη διάρκεια της εκπαίδευσης, ώστε με αυτόν τον τρόπο να μειώνεται και η

ένταση των αναπηδήσεων. Έτσι ο αλγόριθμος δύναται να συγκλίνει οριστικά σε ένα τοπικό ή ολικό ελάχιστο. Στην παρούσα υλοποίηση ο αρχικός ρυθμός μάθησης ορίστηκε σε 0.001 και ανά 10 εποχές ο ρυθμός αυτός διαιρούταν με το 10, ώστε να μειώνεται σταδιακά επιτρέποντας στον SGD να συγκλίνει τελικά σε ένα ελάχιστο.

4.4.3 Ορμή (momentum)

Ένα μειονέκτημα του SGD είναι ότι δυσκολεύεται να πλοηγηθεί σε χαράδρες, δηλαδή σε περιοχές όπου η επιφάνεια του χώρου των παραμέτρων εκπαίδευσης καμπυλώνει πολύ πιο απότομα σε μία διάσταση από ότι σε μια άλλη. Σε αυτές τις περιοχές ο αλγόριθμος ταλαντεύεται στης πλαγιές της χαράδρας προχωρώντας πολύ αργά προς το κάτω μέρος της χαράδρας. Για την επίλυση αυτού του προβλήματος χρησιμοποιείται η ορμή (momentum), η οποία βοηθά τον SGD να επιταχύνει προς την κατεύθυνση του τοπικού ελαχίστου και επιπλέον αποσβένει τις ταλαντώσεις [75]. Στην πραγματικότητα χρησιμοποιώντας την ορμή είναι σαν να έχουμε μια μπάλα που τη σπρώχνουμε από έναν λόφο και η ταχύτητά της αυξάνεται καθώς κυλάει προς τα κάτω. Το ίδιο ουσιαστικά συμβαίνει και κατά τις ενημερώσεις των παραμέτρων του δικτύου. Αυξάνεται δηλαδή η ορμή για τις διαστάσεις των οποίων οι κλίσεις δείχνουν προς τις ίδιες κατευθύνσεις που έδειχναν στην προηγούμενη ενημέρωση και μειώνεται για τις διαστάσεις των οποίων οι κλίσεις άλλαξαν κατεύθυνση. Αυτό επιτυγχάνεται με την πρόσθεση ενός ποσοστού γ του διανύσματος ενημέρωσης των παραμέτρων του προηγούμενου βήματος στο τρέχον διάνυσμα ενημέρωσης των παραμέτρων. Επομένως με την προσθήκη της ορμής ο τύπος του SGD αναδιαμορφώνεται με βάση τους τύπους των εξισώσεων [3] και [4].

$$\upsilon_t = \gamma \upsilon_{t-1} + \eta \cdot \nabla_\theta J(\theta) \quad [3]$$

$$\theta = \theta - \upsilon_t \quad [4]$$

Η παράμετρος γ της ορμής στην παρούσα υλοποίηση ορίστηκε σε 0.9.

4.4.4 Όρος κανονικοποίησης (weight decay)

Προκειμένου να αποφευχθεί η υπερπροσαρμογή των παραμέτρων του μοντέλου στα δεδομένα εκπαίδευσης, χρησιμοποιήθηκε επιπλέον όρος κανονικοποίησης (weight decay) [76]. Ο όρος αυτός πολλαπλασιάζεται με το άθροισμα των τετραγώνων των βαρών του δικτύου και προστίθεται στο σφάλμα. Έτσι κατά την προσαρμογή των βαρών σε κάθε βήμα εκπαίδευσης διατηρούνται μικρότερα τα βάρη αποτρέποντας την υπερεκπαίδευση. Ο όρος weight decay στην παρούσα υλοποίηση ορίστηκε σε 0.0005.

4.4.5 Οπίσθια διάδοση του σφάλματος (back propagation of error)

Για την προσαρμογή των παραμέτρων του δικτύου χρησιμοποιήθηκε η οπίσθια διάδοση του σφάλματος (backpropagation of error), η οποία αποτελεί μία συχνά χρησιμοποιούμενη μέθοδο εκπαίδευσης νευρωνικών δικτύων και συνδυάζεται συνήθως με μία μέθοδο βελτιστοποίησης, όπως ο SGD που παρουσιάστηκε παραπάνω. Στην οπίσθια διάδοση υπολογίζεται αρχικά η πρόβλεψη και το αντίστοιχο σφάλμα για κάθε δείγμα εισόδου. Ακολούθως αθροίζονται όλα τα σφάλματα για να υπολογιστεί το τελικό σφάλμα και στη συνέχεια διαδίδεται το τελικό σφάλμα προς τα πίσω στο δίκτυο, ώστε να υπολογιστεί η κλίση του κόστους-σφάλματος ως προς όλες τις παραμέτρους και να ανανεωθούν τα βάρη τους. Αρχικά για τις παραμέτρους των νευρώνων του στρώματος εξόδου οι κλίση προκύπτει από τη μερική παράγωγο του σφάλματος ως προς το αντίστοιχο βάρος του κάθε νευρώνα.

Για τον υπολογισμό της κλίσης στους νευρώνες και των υπόλοιπων στρωμάτων χρησιμοποιείται επαναλαμβανόμενα ο κανόνας της αλυσίδας [77].

4.5 Μετρικές αξιολόγησης

4.5.1 Συνολική ακρίβεια (accuracy)

Τα μοντέλα που εκπαιδεύτηκαν αξιολογήθηκαν με βάση κάποιες μετρικές απόδοσης. Η πρώτη μετρική είναι η συνολική ακρίβεια (accuracy), δηλαδή το ποσοστό των ορθών ταξινομήσεων στο σύνολο των ταξινομήσεων που έγιναν. Ο υπολογισμός της γίνεται σύμφωνα με την Εξίσωση [5]:

$$accuracy = \frac{number\ of\ correct\ predictions}{number\ of\ total\ predictions} \quad [5]$$

Η παραπάνω μετρική δίνει μια εικόνα για την ικανότητα του μοντέλου να ταξινομεί τα δεδομένα ορθά, όμως δεν είναι αρκετή, καθώς η κατανομή των δεδομένων στις δύο κλάσεις είναι άνιση. Συγκεκριμένα στο σύνολο δεδομένων Celeb-DF-v2 [9] τα δεδομένα της κλάσης των πλαστών βίντεο είναι περίπου το 90% των συνολικών δεδομένων, ενώ τα υπόλοιπο 10% περίπου είναι τα πραγματικά βίντεο. Αυτό καθιστά την συνολική ακρίβεια μη επαρκή μετρική επίδοσης, αφού ακόμα κι ένα μοντέλο που ταξινομεί όλα τα βίντεο ως πλαστά θα έχει συνολική ακρίβεια περίπου 90%, δηλαδή αρκετά υψηλή, παρά το γεγονός ότι στην πραγματικότητα δεν ξεχωρίζει τα πλαστά από τα πραγματικά βίντεο. Για τον λόγο αυτό χρησιμοποιήθηκαν επιπλέον μετρικές.

4.5.2 Εμβαδόν καμπύλης λειτουργικών χαρακτηριστικών (AUC-ROC)

Μια ακόμη μετρική που ενδείκνυται για προβλήματα δυαδικής ταξινόμησης, όπως εδώ, είναι το εμβαδόν της καμπύλης λειτουργικών χαρακτηριστικών (Area Under the Curve of Receiver Characteristic Operator (AUC-ROC)). Η καμπύλη αυτή προκύπτει με βάση δύο άλλες μετρικές. Αρχικά οι δύο κλάσεις χωρίζονται στην θετική (positive) και την αρνητική (negative) κλάση. Θετική (κλάση 1) για το παρόν πρόβλημα είναι τα πλαστά βίντεο (fake) και αρνητική (κλάση 0) είναι τα πραγματικά (real). Με βάση τις πραγματικές κλάσεις των δεδομένων (actual values) και τις προβλέψεις που προκύπτουν από το αντίστοιχο μοντέλο (predicted values) προκύπτουν τα πλήθη των αληθώς θετικών (True Positive – TP), των ψευδώς θετικών (False Positive - FP), των αληθώς αρνητικών (True Negative - TN) και των ψευδώς αρνητικών (False Negative – FN) προβλέψεων. Αληθώς θετικά (TP) είναι τα δεδομένα που ανήκουν στην κλάση 1 δηλαδή στα πλαστά βίντεο και ταξινομήθηκαν όντως ως πλαστά. Ψευδός θετικά (FN) είναι τα βίντεο που δεν ανήκουν στην κλάση των πλαστών βίντεο, αλλά παρ’ όλα αυτά ταξινομήθηκαν από το μοντέλο εσφαλμένα ως πλαστά. Αντίστοιχα αληθώς αρνητικά (TN) είναι τα βίντεο που ανήκουν στην κλάση των πραγματικών βίντεο και όντως ταξινομήθηκαν σε αυτή, ενώ ψευδώς αρνητικά (FN) είναι τα βίντεο που δεν ανήκουν στην κλάση των πραγματικών βίντεο, αλλά εσφαλμένα το μοντέλο τα ταξινόμησε σε αυτή.

Με βάση τα παραπάνω υπολογίζονται οι δύο μετρικές στις οποίες βασίζεται η μετρική ROC-AUC. Η πρώτη μετρική είναι η αναλογία των αληθώς θετικών (TPR) ή αλλιώς η ανάκληση (recall). Η μετρική αυτή εξετάζει τι ποσοστό από τα δεδομένα της θετικής κλάσης, δηλαδή από τα πλαστά βίντεο, ταξινομήθηκε ορθά. Υπολογίζεται σύμφωνα με τον τύπο της Εξίσωσης [6]:

$$TPR = \frac{TP}{TP + FN} \quad [6]$$

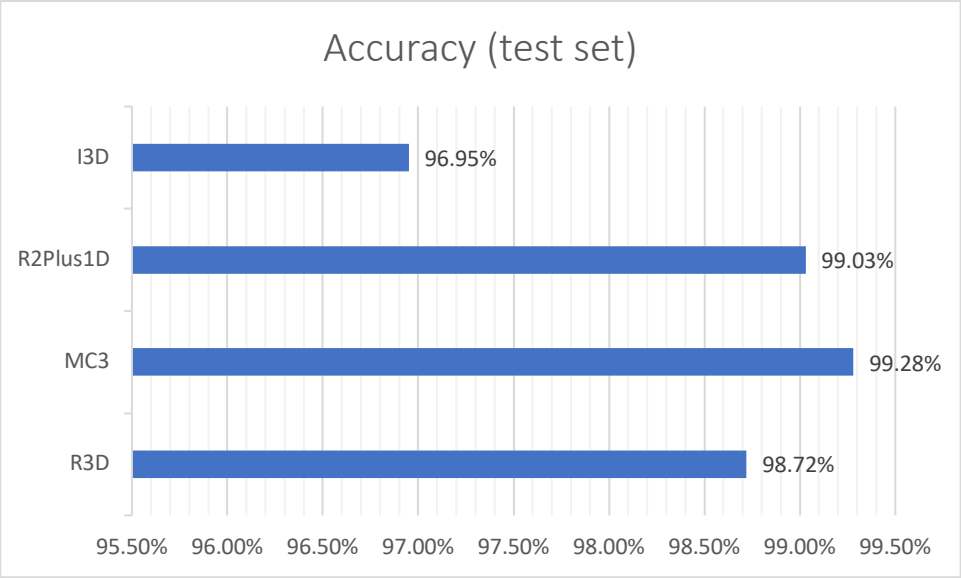
Η δεύτερη μετρική είναι η αναλογία των ψευδώς θετικών (FPR). Η μετρική αυτή εξετάζει τι ποσοστό των δεδομένων της αρνητικής κλάσης, δηλαδή των πραγματικών βίντεο, ταξινομήθηκε λανθασμένα στη θετική κλάση, δηλαδή στα πλαστά βίντεο. Για τον υπολογισμό της χρησιμοποιείται ο τύπος της Εξίσωσης [7]:

$$FPR = \frac{FP}{TN + FP} \quad [7]$$

Τα μοντέλα που χρησιμοποιήθηκαν ως αποτέλεσμα δεν έχουν τιμές 0 και 1, αλλά τιμές μεταξύ 0 και 1 και συγκεκριμένα έχουν δύο τιμές ως έξοδο, μία για την κλάση 1, των πλαστών βίντεο, και μία για την κλάση 0, των πραγματικών βίντεο. Αντί απευθείας να θεωρηθεί ως κλάση στην οποία ταξινομήθηκε το εκάστοτε δεδομένο η κλάση με τη μεγαλύτερη τιμή, μπορεί να γίνει μία διαφορετική ερμηνεία. Συγκεκριμένα κάθε μία από αυτές τις δύο τιμές μπορεί να ερμηνευθεί ως η πιθανότητα του εξεταζόμενου δεδομένου-βίντεο να ανήκει στην κάθε κλάση ξεχωριστά. Παίρνοντας ως έξοδο τη μία από αυτές τις δύο τιμές και συγκεκριμένα την πιθανότητα να ανήκει το δεδομένο στην θετική κλάση, δηλαδή στα πλαστά βίντεο, είναι εφικτό να προκύψουν διαφορετικά συμπεράσματα για το αποτέλεσμα ανάλογα με κάποιο κατώφλι (threshold) που θα ορίζει από ποια τιμή πιθανότητας και πάνω το δεδομένο θεωρείται ότι όντως ανήκει στη θετική κλάση των πλαστών βίντεο. Δοκιμάζοντας πολλά διαφορετικά κατώφλια μεταξύ 0 και 1 προκύπτουν διαφορετικά αποτελέσματα στο σύνολο δεδομένων test. Σε αυτά τα αποτελέσματα υπολογίζονται οι δύο μετρικές που προαναφέρθηκαν, δηλαδή η αναλογία των αληθώς θετικών (TPR) και η αναλογία των ψευδώς θετικών (FPR). Έτσι σχηματίζεται η καμπύλη ROC, η οποία απεικονίζει τη μετρική TPR έναντι της FPR για διάφορες τιμές κατωφλίου, διαχωρίζοντας ουσιαστικά το «σήμα» από τον «θόρυβο». Το εμβαδόν της περιοχής κάτω από την καμπύλη είναι μια μετρική που προσδιορίζει την ικανότητα του μοντέλου να διακρίνει τα δεδομένα μεταξύ των δύο κλάσεων. Η μέγιστη και βέλτιστη τιμή που μπορεί να έχει αυτό το εμβαδόν είναι 1 (ή 100%), ενώ η ελάχιστη είναι 0 (ή 0%) [78].

4.6 Αποτελέσματα

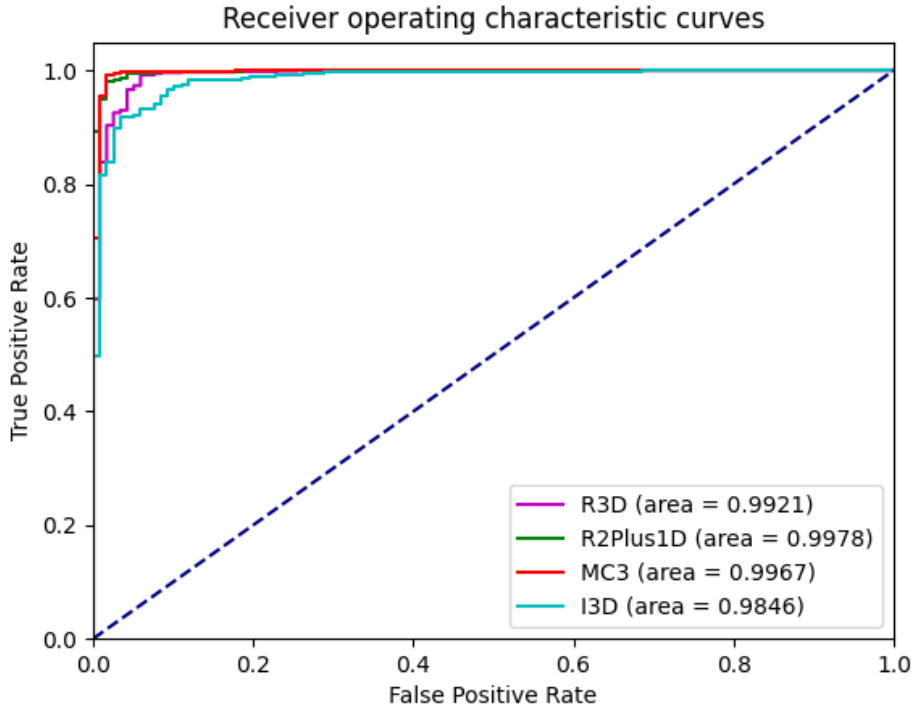
Τα καλύτερα μοντέλα που προέκυψαν από την εκπαίδευση κάθε δικτύου εξετάστηκαν στα δεδομένα test. Οι επιδόσεις τους στη μετρική της συνολικής ακρίβειας (accuracy) φαίνονται στην Εικόνα 28.



Εικόνα 28: Συγκριτικό διάγραμμα της συνολικής ακρίβειας (accuracy) των μοντέλων στα δεδομένα test

Παρατηρείται ότι όλα τα μοντέλα έχουν πολύ υψηλή ακρίβεια, άνω του 95% δηλαδή. Οι διαφορές μεταξύ τους είναι μικρές. Επιπλέον φαίνεται ότι τα τρία μοντέλα που βασίζονται σε υπολειμματικά νευρωνικά δίκτυα (residual networks), δηλαδή το R3D, το MC3 και το R2Plus1D, έχουν καλύτερη επίδοση από το μοντέλο I3D που βασίζεται σε δομές Inception. Επιπλέον το MC3, παρά το γεγονός ότι είναι το πιο ελαφρύ από τα μοντέλα υπολειμματικών νευρωνικών δικτύων, καθώς έχει τις λιγότερες παραμέτρους, επιτυγχάνει τελικά την καλύτερη επίδοση από όλα.

Τα αποτελέσματα των παραπάνω μοντέλων στη μετρική AUC-ROC φαίνονται συγκριτικά στην Εικόνα 29.



Εικόνα 29: Συγκριτικό διάγραμμα των επιδόσεων των μοντέλων στη μετρική AUC-ROC στα δεδομένα test

Από το διάγραμμα είναι εμφανές ότι όλα τα μοντέλα διακρίνουν εξαιρετικά καλά τα δεδομένα των δύο κλάσεων μεταξύ τους, καθώς οι τιμές των μοντέλων στη μετρική AUC-ROC είναι όλες πάνω από 0.98 (98%).Και στη μετρική αυτή βέβαια παρατηρείται καλύτερη επίδοση των μοντέλων που βασίζονται σε υπολειμματικά δίκτυα, σε αντίθεση με το I3D.

Επιπλέον συγκρίθηκαν οι χρόνοι που χρειάζεται το κάθε μοντέλο για την εξαγωγή αποτελεσμάτων για τα δεδομένα test και διαπιστώθηκε ότι τα πιο γρήγορα μοντέλα ήταν το MC3 και το I3D.

4.7 Συμπεράσματα και μελλοντικές κατευθύνσεις

Με βάση τα παραπάνω αποτελέσματα φαίνεται ότι τα τρισδιάστατα συνελικτικά νευρωνικά δίκτυα (3D CNN) αποτελούν μια πολύ αποτελεσματική μέθοδο για την ανίχνευση πλαστών βίντεο, καθώς έχουν τη δυνατότητα να συνυπολογίζουν τόσο τις χωρικές διαστάσεις όσο και τη διάσταση του χρόνου. Με αυτόν τον τρόπο καθίσταται εφικτός ο εντοπισμός όχι μόνο των ανωμαλιών που υπάρχουν στο εκάστοτε στιγμιότυπο ενός πλαστού βίντεο, αλλά και των ασυνεχειών που υπάρχουν στις μεταβάσεις μεταξύ στιγμιότυπων.

Επιπλέον φαίνεται ότι από τα μοντέλα που εξετάστηκαν καλύτερες επιδόσεις έχουν αυτά που βασίζονται σε υπολειμματικά νευρωνικά δίκτυα, που στην παρούσα εργασία είναι το R3D, το MC3 και το R2Plus1D. Το I3D που βασίζεται σε δομές Inception έχει επίσης ικανοποιητικές επιδόσεις, αλλά συγκριτικά με τα άλλα παραμένει ελαφρώς λιγότερο αποτελεσματικό.

Τα θετικά αποτελέσματα που προέκυψαν είναι πολύ ενθαρρυντικά, ιδίως αν ληφθεί υπόψη το επίπεδο δυσκολίας αναγνώρισης πλαστών βίντεο στο σύνολο δεδομένων Celeb-DF-v2 [9]. Παρ’ όλα αυτά το γεγονός ότι τα μοντέλα εκπαιδεύτηκαν σε αυτό μόνο το σύνολο, καθιστά πιθανό το να έχουν εξειδικευτεί στο να εντοπίζουν τις ατέλειες που προκύπτουν από τη μέθοδο σύνθεσης πλαστών βίντεο που χρησιμοποιήθηκε για τη δημιουργία των

πλαστών βίντεο του συγκεκριμένου συνόλου. Αυτό έχει ως συνέπεια το ενδεχόμενο να μην είναι αποτελεσματικά στην αναγνώριση πλαστών βίντεο που δημιουργήθηκαν με χρήση άλλων μεθόδων σύνθεσης. Για τον λόγο αυτό μελλοντικά θα μπορούσε να διερευνηθεί κατά πόσο τα μοντέλα αυτά έχουν επίσης ικανοποιητικά αποτελέσματα αν εκπαιδευτούν σε βίντεο που προέρχονται από πολλά διαφορετικά σύνολα δεδομένων στα οποία χρησιμοποιούνται ποικίλες μέθοδοι σύνθεσης πλαστών βίντεο.

5. Επίλογος

Η σύνθεση πλαστών (deepfake) βίντεο είναι ένας τομέας που αναπτύσσεται ραγδαία οδηγώντας σε όλο και πιο αληθοφανή αποτελέσματα παραποιημένων βίντεο. Πέρα από τις θετικές προεκτάσεις της τεχνολογίας αυτής, υπάρχει και ο κίνδυνος κακόβουλης χρήσης τους. Για τον λόγο αυτό είναι αναγκαία η ανάπτυξη μεθόδων ανίχνευσης πλαστών (deepfake) βίντεο. Αν και ήδη έχουν γίνει αντίστοιχες προσπάθειες, είναι φανερό ότι όσο αναπτύσσεται ο τομέας της σύνθεσης πλαστών βίντεο, θα πρέπει να ακολουθεί και να εξελίσσεται και ο τομέας της αναγνώρισης αντίστοιχων βίντεο.

Δεδομένων των παραπάνω, στην παρούσα διπλωματική εργασία επιδιώχθηκε η αντιμετώπιση του προβλήματος της κακόβουλης χρήσης πλαστών (deepfake) βίντεο μέσω της υλοποίησης μεθόδων ανίχνευσής τους. Για τον σκοπό αυτό αρχικά μελετήθηκαν τόσο οι υπάρχουσες μέθοδοι σύνθεσης τέτοιων βίντεο, όσο και οι αντίστοιχες μέθοδοι ανίχνευσής τους. Στη συνέχεια προτάθηκε ο εντοπισμός πλαστών βίντεο μέσω μοντέλων μηχανικής μάθησης βασιζόμενων σε τρισδιάστατα συνελικτικά νευρωνικά δίκτυα (3D CNN). Ο κώδικας που αναπτύχθηκε εκπαίδευσε τέσσερα διαφορετικά δίκτυα πάνω στο σύνολο δεδομένων Celeb-DF-v2 [9]. Το σύνολο αυτό περιλαμβάνει 590 πραγματικά βίντεο και 5639 πλαστά που έχουν προκύψει με ανταλλαγή προσώπων από τα πραγματικά. Ο λόγος επιλογής του συγκεκριμένου συνόλου δεδομένων ήταν ότι τα αποτελέσματα της μεθόδου που έχει χρησιμοποιηθεί για τη σύνθεση πλαστών βίντεο είναι εξαιρετικά αληθοφανή σε σύγκριση με άλλα πλαστά βίντεο αντίστοιχων συνόλων δεδομένων.

Τα δίκτυα που εκπαιδεύτηκαν ήταν το R3D [10], το MC3 [11], το R2Plus1D [11] και το I3D [12]. Τα τρία πρώτα βασίζονται σε δομές υπολειμματικών νευρωνικών δικτύων και η δομή τους είναι παρεμφερής, ενώ το τέταρτο είναι αρκετά διαφορετικό και βασίζεται σε δομές Inception. Το πλεονέκτημα και των τεσσάρων δικτύων είναι ότι έχουν την ικανότητα να κατανοούν χωροχρονική πληροφορία. Αυτό οφείλεται στο γεγονός ότι η αρχιτεκτονική τους στηρίζεται στη χρήση τρισδιάστατων συνελίξεων.

Ως τώρα τα δίκτυα αυτά έχουν φανεί πολύ αποτελεσματικά σε προβλήματα ταξινόμησης βίντεο σύμφωνα με τον τύπο δραστηριότητας που πραγματοποιείται. Η πρότερη γνώση που είχαν αυτά σε αυτό το είδος προβλήματος χρησιμοποιήθηκε στην παρούσα εργασία ως βάση. Έτσι τα τέσσερα μοντέλα εκπαιδεύτηκαν ξεκινώντας με τις τιμές παραμέτρων που είχαν προκύψει από προηγούμενη εκπαίδευσή τους σε σύνολα δεδομένων που περιλάμβαναν βίντεο διαφόρων τύπων δραστηριοτήτων. Στόχος τους πλέον όμως ήταν η διάκριση ανάμεσα σε πλαστά και πραγματικά βίντεο.

Τα αποτελέσματα ήταν πολύ ενθαρρυντικά. Για την ακρίβεια και τα τέσσερα μοντέλα ταξινόμησαν ορθά πάνω από το 96% των δεδομένων test. Για να επιβεβαιωθεί η επιτυχία των μοντέλων χρησιμοποιήθηκε ακόμη μία μετρική που ελέγχει την ικανότητα ενός μοντέλου να διαχωρίζει καλά τα δεδομένα κάθε κλάσης. Η μετρική αυτή είναι το εμβαδόν της καμπύλης λειτουργικών χαρακτηριστικών (Area Under the Curve of Receiver Characteristic Operator (AUC-ROC)). Και σε αυτή τη μετρική τα αποτελέσματα ήταν εξαιρετικά με τιμές άνω του 98%.

Συγκριτικά τα μοντέλα ήταν σχεδόν ισάξια. Τις χαμηλότερες επιδόσεις και στις δύο μετρικές είχε το I3D [12] (97% ακρίβεια και 98% AUC-ROC περίπου). Τα άλλα τρία μοντέλα (R3D [10], MC3 [11] και R2Plus1D [11]) δεν είχαν ουσιαστικές διαφορές μεταξύ τους, ενώ είχαν ελαφρώς καλύτερες επιδόσεις (99% ακρίβεια και 99% AUC-ROC περίπου).

Συμπερασματικά, με βάση τα παραπάνω, προέκυψε ότι η χρήση τρισδιάστατων συνελικτικών δικτύων είναι αρκετά αποτελεσματική για την ανίχνευση πλαστών (deepfake) βίντεο. Επιπλέον μέσω της σύγκρισης διαφορετικών μεθόδων φάνηκε ότι μοντέλα που περιλαμβάνουν υπολειμματικά νευρωνικά δίκτυα (R3D [10], MC3 [11] και R2Plus1D [11])

είναι πιο αποτελεσματικά από μοντέλα που βασίζονται σε δομές Inception (I3D [12]). Τα συμπεράσματα αυτά γίνεται να αξιοποιηθούν για περεταίρω έρευνα. Τα μοντέλα που χρησιμοποιήθηκαν, μπορούν μελλοντικά να εκπαιδευτούν σε ακόμη πιο δύσκολα σύνολα δεδομένων, τα οποία θα περιλαμβάνουν παραποιημένα βίντεο που έχουν προκύψει από διαφορετικές και ποικίλες μεθόδους σύνθεσης πλαστών βίντεο, πέρα από τη μία μόνο μέθοδο που χρησιμοποιείται στο σύνολο δεδομένων Celeb-DF-v2 [9].

Το πεδίο της αναγνώρισης πλαστών (deepfake) βίντεο έχει εξαιρετικό ενδιαφέρον και πολλά περιθώρια εξέλιξης. Καθώς οι τεχνολογίες σύνθεσης πλαστών βίντεο βελτιώνονται όλο και περισσότερο, προκύπτουν συνεχώς νέες προκλήσεις. Παρά το γεγονός ότι υπάρχουν ακόμη αρκετά περιθώρια βελτίωσης των μεθόδων ανίχνευσης παραποιημένων βίντεο, η μέχρι τώρα πρόοδος είναι πολύ σημαντική και βοηθητική για την περεταίρω εξέλιξη στον τομέα.

Βιβλιογραφία

[1] B. Marr, «The best (and scariest) examples of AI-enabled deepfakes,» 22 July 2019. [Ηλεκτρονικό]. Available: https://www.forbes.com/sites/bernardmarr/2019/07/22/the-best-and-scariest-examples-of-ai-enabled-deepfakes/. [Πρόσβαση 12 May 2021].

[2] Bloomberg, «How faking videos became easy and why that's so scary,» 29 August 2018. [Ηλεκτρονικό]. Available: https://fortune.com/2018/09/11/deep-fakes-obama-video/. [Πρόσβαση 12 May 2021].

[3] R. Chesney και D. Citron, «Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics,» *Foreign Affairs,* τόμ. 98, p. 147, 2019.

[4] T. Hwang, «Deepfakes: A Grounded Threat Assessment,» Center of Security and Emerging Technology (CSET), 2020.

[5] X. Zhou και R. Zafarani, «A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,» *ACM Computing Surveys (CSUR),* τόμ. 53, October 2020.

[6] E. Zakharov, A. Shysheya, E. Burkov και V. Lempitsky, Few-Shot Adversarial Learning of Realistic Neural Talking Head Models, arXiv:1905.08233v2, 2019.

[7] T. Nguyen, C. M. Nguyen, T. Nguyen, D. Nguyen και S. Nahavandi, « Deep Learning for Deepfakes Creation and Detection: A Survey,» 2019.

[8] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies και M. Nießner, «FaceForensics++: Learning to Detect Manipulated Facial Images,» σε *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, 2019.

[9] Y. Li, X. Yang, P. Sun, H. Qi και S. Lyu, «Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics,» σε *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] H. Kensho, K. Hirokatsu και S. Yutaka, «Learning Spatio-Temporal Features with 3D Residual Networks for Action,» *Proceedings of the IEEE InternationalConference on Computer Vision Workshops,* pp. 3154-3160, 2017.

[11] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun και M. Paluri, «A Closer Look at Spatiotemporal Convolutions for Action Recognition,» *Proceedings of theIEEE conference on Computer Vision and Pattern Recognition,* pp. 6450-6459, 2018.

[12] J. Carreira και A. Zisserman, «Quo vadis, actionrecognition? a new model and the kinetics dataset,» *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 6299-6308, 2017.

[13] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano και H. Li, «Protecting World Leaders Against Deep Fakes,» σε *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, California, 2019.

[14] A. Tewari, M. Zollhöfer, F. Bernard, P. Garrido, K. Hyeongwoo, P. Pérez και C. Theobalt, «High-Fidelity Monocular Face Reconstruction Based on an Unsupervised Model-Based Face Autoencoder,» *IEEE Transactions on Pattern Analysis and Machine Intelligence,* τόμ. 42, pp. 357-370, 1 February 2020.

[15] M. Y. Liu, X. Huang, J. Yu, T. C. Wang και A. Mallya, «Generative adversarial networks for image and video synthesis: Algorithms and applications,» *Proceedings of the IEEE,* τόμ. 109, pp. 839-862, May 2021.

[16] «FakeApp 2.2.0,» Malavida, [Ηλεκτρονικό]. Available: https://www.malavida.com/en/soft/fakeapp/. [Πρόσβαση 13 May 2021].

[17] «Faceswap: Deepfakes software for all,» [Ηλεκτρονικό]. Available: https://github.com/deepfakes/faceswap. [Πρόσβαση 13 May 2021].

[18] «DeepFaceLab,» [Ηλεκτρονικό]. Available: https://github.com/iperov/DeepFaceLab. [Πρόσβαση 13 May 2021].

[19] I. Petrov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, J. Jiang, L. RP, S. Zhang, P. Wu και W. Zhang, «DeepFaceLab: A simple, flexible and extensible face swapping framework,» *Computing Research Repository (CoRR),* May 2020.

[20] «DeepFaceLab: Explained and usage tutorial,» [Ηλεκτρονικό]. Available: https://mrdeepfakes.com/forums/thread-guide-deepfacelab-2-0-guide. [Πρόσβαση 2021 May 15].

[21] «DFaker,» [Ηλεκτρονικό]. Available: https://github.com/dfaker/df. [Πρόσβαση 13 May 2021].

[22] «DeepFake_tf: Deepfake based on tensorflow,» [Ηλεκτρονικό]. Available: https://github.com/StromWine/DeepFake_tf. [Πρόσβαση 13 May 2021].

[23] «Keras-VGGFace: VGGFace implementation with Keras framework,» [Ηλεκτρονικό]. Available: https://github.com/rcmalli/keras-vggface. [Πρόσβαση 14 May 2021].

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville και Y. Bengio, «Generative Adversarial Networks,» *Advances in Neural Information Processing Systems,* τόμ. 3, pp. 2672-2680, 10 6 2014.

[25] « Faceswap-GAN,» [Ηλεκτρονικό]. Available: https://github.com/shaoanlu/faceswap-GAN. [Πρόσβαση 14 May 2021].

[26] «Few-shot face translation,» [Ηλεκτρονικό]. Available: https://github.com/shaoanlu/fewshot-face-translation-GAN. [Πρόσβαση 5 June 2021].

[27] M. Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen και J. Kautz, «Few-Shot Unsupervised Image-to-Image Translation,» σε *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[28] T. Park, M. Y. Liu, T. C. Wang και J. Y. Zhu, «GauGAN: semantic image synthesis with spatially adaptive normalization,» σε *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[29] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh και S. Zafeiriou, «AvatarMe: Realistically Renderable 3D Facial Reconstruction "in-the-wild",» σε *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[30] S. Ha, M. Kersner, K. Beomsu, S. Seo και D. Kim, «MarioNETte: Few-Shot Face Reenactment Preserving Identity of Unseen Targets,» *Proceedings of the AAAI Conference on Artificial Intelligence,* τόμ. 34, pp. 10893-10900, April 2020.

[31] Y. Deng, J. Yang, D. Chen, F. Wen και X. Tong, «Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning,» *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 5154-5163, June 2020.

[32] A. Tawari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Perez, M. Zollhofer και C. Theobalt, «StyleRig: Rigging StyleGAN for 3D Control Over Portrait Images,» *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 6142-6151, June 2020.

[33] L. Li, J. Bao, H. Yang, D. Chen και F. Wen, «FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping,» arXiv preprint arXiv:1912.13457, 2019.

[34] Y. Nirkin, Y. Keller και T. Hassner, «FSGAN: Subject Agnostic Face Swapping and Reenactment,» *Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 7184-7193, 2019.

[35] C. Chan, S. Ginosar, T. Zhou και A. A. Efros, «Everybody Dance Now,» *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 5933-5942, October 2019.

[36] J. Thies, M. Elgharib, A. Tewari, C. Theobalt και M. Nie, « Neural Voice Puppetry: Audio-driven Facial Reenactment,» *European Conference on Computer Vision,* pp. 716-731, August 2020.

[37] A. Y. Xu, «AI, Truth, and Society: Deepfakes at the front of the Technological Cold War,» 2 July 2019. [Ηλεκτρονικό]. Available: https://medium.com/gradientcrescent/ai-truth-and-society-deepfakes-at-the-front-of-the-technological-cold-war-86c3b5103ce6. [Πρόσβαση 19 May 2021].

[38] B. Dolhansky, R. Howes, B. Pflaum, N. Baram και C. C. Ferrer, «The Deepfake Detection Challenge (DFDC) Preview Dataset,» *Computing Research Repository (CoRR),* 2019.

[39] P. Korshunov και S. Marcel, «DeepFakes: a New Threat to Face Recognition? Assessment and Detection,» *Computing Research Repository (CoRR),* 2018.

[40] X. Yang, Y. Li και S. Lyu, «Exposing Deep Fakes Using Inconsistent Head Poses,» *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP),* pp. 8261-8265, 2019.

[41] X. Xuan, B. Peng, W. Wang και J. Dong, «On the generalization of GAN image forensics,» *Computing Research Repository,* 2019.

[42] S. Agarwal και L. R. Varshney, «Limits of Deepfake Detection: A Robust Estimation Viewpoint,» *arXiv e-prints,* May 2019.

[43] C.-C. Hsu, Y.-X. Zhuang και C.-Y. Lee, «Deep Fake Image Detection Based on Pairwise Learning,» *Applied Sciences,* 2020.

[44] D. Archar, V. Nozick, J. Yamagishi και I. Echizen, «MesoNet: a Compact Facial Video Forgery Detection Network,» *2018 IEEE International Workshop on Information Forensics and Security (WIFS),* pp. 1-7, December 2018.

[45] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi και P. Natarajan, «Recurrent Convolutional Strategies for Face Manipulation Detection in Videos,» *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops,* pp. 80-87, 2019.

[46] G. Huang, Z. Liu, D. M. Van, W. Laurens και Q. Kilian, «Densely Connected Convolutional Networks,» *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2261-2269, 2017.

[47] K. Cho, B. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk και Y. Bengio, «Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,» *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 1724-1734, October 2014.

[48] D. Guera και E. Delp, «Deepfake Video Detection Using Recurrent Neural Networks,» *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS),* pp. 1-6, 2018.

[49] Y. Li, M.-C. Chang και S. Lyu, «In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,» *2018 IEEE International Workshop on Information Forensics and Security (WIFS),* pp. 1-7, December 2018.

[50] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko και T. Darrell, «Long-Term Recurrent Convolutional Networks for Visual Recognition and Description,» *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2625-2634, 2015.

[51] K. Simonyan και A. Zisserman, «Very Deep Convolutional Networks for Large-Scale Image Recognition,» *arXiv 1409.1556,* 2014.

[52] K. He, X. Zhang, S. Ren και J. Sun, «Deep Residual Learning for Image Recognition,» *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 770-778, 2016.

[53] Y. Li και S. Lyu, «Exposing DeepFake Videos By Detecting Face Warping Artifacts,» *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,* pp. 46-52, November 2018.

[54] X. Yang, Y. Li και S. Lyu, «Exposing Deep Fakes Using Inconsistent Head Poses,» *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 8261-8265, May 2019.

[55] P. Zhou, X. Han, V. I. Morariu και L. S. Davis, «Two-Stream Neural Networks for Tampered Face Detection,» *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops,* pp. 19-27, July 2017.

[56] H. H. Nguyen, J. Yamagshi και I. Echizen, «Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos,» *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 2307-2311, 17 May 2019.

[57] A. Krizheysky, S. D. Wang και G. Hinton, «Transforming Auto-encoders,» *International Conference on Artificial Neural Networks (ICANN),* pp. 44-51, June 2011.

[58] S. Sabour, N. Frosst και G. E. Hinton, «Dynamic Routing Between Capsules,» *Advances in Neural Information Processing Systems,* pp. 3856-3866, 2017.

[59] I. Chingovska, A. Anjos και S. Marcel, «On the effectiveness of local binary patterns in face anti-spoofing,» *Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG),* pp. 1-7, September 2012.

[60] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies και M. Nießner, «FaceForensics: A Large-scale Video Dataset for Forgery,» *arXiv:1803.09179v1,* 24 March 2018.

[61] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt και M. Nießner, «Face2Face: Real-time Face Capture and Reenactment of RGB Videos,» *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2387-2395, 2016.

[62] N. Rahmouni, V. Nozick, J. Yamagishi και I. Echizen, «Distinguishing computer graphics from natural images using convolution neural networks,» *2017 IEEE Workshop on Information Forensics and Security (WIFS),* pp. 1-6, December 2017.

[63] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates, A. P. Delgado, D. F. Zhou, T. N. Kheyrkhah, J. Smith και J. G. Fiscus, «MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation,» *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW),* pp. 63-72, January 2019.

[64] F. Matern, C. Riess και M. Stamminger, «Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations,» *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW),* pp. 83-92, January 2019.

[65] M. Koopman, A. M. Rodriguez και Z. Geradts, «Detection of Deepfake Video Manipulation,» *Proceedings of the 20th Irish Machine Vision and Image Processing conference (IMVIP),* pp. 133-136, 2018.

[66] K. Rosenfeld, T. Sencar και N. Memon, «A Study of the Robustness of PRNU-based Camera Identification,» *Media Forencics and Security,* February 2009.

[67] H. R. Hasan και K. Salah, «Combating Deepfake Videos Using Blockchain and Smart Contracts,» *IEEE Access,* pp. 41596-41606, January 2019.

[68] «IPFS powers the Distributed Web,» [Ηλεκτρονικό]. Available: https://ipfs.io/. [Πρόσβαση 10 June 2021].

[69] G. S. Behera, «Face Detection with Haar Cascade,» 24 December 2020. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/face-detection-with-haar-cascade-727f68dafd08. [Πρόσβαση 21 May 2021].

[70] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia και S. Zafeiriou, «RetinaFace: Single-stage Dense Face Localisation in the Wild,» *CoRR,* τόμ. abs/1905.00641, 2019.

[71] S. Sumit, «A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way,» 15 December 2018. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. [Πρόσβαση 5 June 2021].

[72] S. Ioffe και C. Szegedy, «Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,» *Proceedings of the 32nd International Conference on Machine Learning,* τόμ. 37, pp. 448-456, July 2015.

[73] DeepMind, «Kinetics,» 22 May 2017. [Ηλεκτρονικό]. Available: https://deepmind.com/research/open-source/kinetics. [Πρόσβαση 5 June 2021].

[74] «Perceptual Reasoning and Interaction Research - Charades,» June 2016. [Ηλεκτρονικό]. Available: https://prior.allenai.org/projects/charades. [Πρόσβαση 5 June 2021].

[75] S. Ruder, «An overview of gradient descent optimization algorithms,» 19 Januar 2016. [Ηλεκτρονικό]. Available: https://ruder.io/optimizing-gradient-descent. [Πρόσβαση 5 June 2021].

[76] D. Vasani, «This thing called Weight Decay,» 29 April 2019. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/this-thing-called-weight-decay-a7cd4bcfccab. [Πρόσβαση 5 June 2021].

[77] M. Mazur, «A Step by Step Backpropagation Example,» 17 March 2015. [Ηλεκτρονικό]. Available: https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example. [Πρόσβαση 5 June 2021].

[78] A. Bhandari, «AUC-ROC Curve in Machine Learning Clearly Explained,» 16 June 2020. [Ηλεκτρονικό]. Available: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/. [Πρόσβαση 4 June 2021].

[79] T. Karras, S. Laine και T. Aila, «A Style-Based Generator Architecture for Generative Adversarial Networks,» σε *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, California, 2019.

[80] T. Karras, T. Alia, S. Laine και J. Lehtinen, «Progressive growing of gans for improved quality, stability, and variation,» σε *International Conference on Learning Representations (ICLR)*, Vancouver, 2018.

[81] J. Lin, Y. Li και G. Yang, «FPGAN: Face de-identification method with generative adversarial networks for social robots,» *Neural Networks,* τόμ. 133, pp. 132-147, January 2021.

[82] K. Olszewski, S. Tulyakov, O. Woodford, H. Li και L. Luo, «Transformable Bottleneck Networks,» *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 7648-7657, October 2019.