ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
”ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ”

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αναγνώριση Ανθρώπινης Συμπεριφοράς με Τεχνικές Βαθιάς Μάθησης

Αθανασία Χ. Τρανού

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Σταύρος Περαντώνης, Διευθυντής Ερευνών ΕΚΕΦΕ - Δημόκριτος

ΑΘΗΝΑ

Δεκέμβριος 2019

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
”INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY”

M.Sc THESIS

Human Action Recognition with Deep Learning Techniques

Athanasia C. Tranou

SUPERVISOR: Stavros Perantonis, Research Director National Center for Scientific Research “Demokritos”

ATHENS

December 2019

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αναγνώριση Ανθρώπινης Συμπεριφοράς με Τεχνικές Βαθιάς Μάθησης

Αθανασία Χ. Τρανού
ΑΜ: ΠΙΒ0190

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Σταύρος Περαντώνης, Διευθυντής Ερευνών ΕΚΕΦΕ - Δημόκριτος

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:

Σταύρος Περαντώνης, Διευθυντής Ερευνών ΕΚΕΦΕ - Δημόκριτος

Παναγιώτης Σταματόπουλος, Επίκουρος Καθηγητής Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Ευάγγελος Σπύρου, Συνεργαζόμενος Ερευνητής ΕΚΕΦΕ - Δημόκριτος

Ημερομηνία Εξέτασης: 12 12 2019

ΠΕΡΙΛΗΨΗ

Στην παρούσα διπλωματική εργασία μελετήθηκε η αυτόματη αναγνώριση ανθρώπινης συμπεριφοράς, η οποία μπορεί να εφαρμοστεί σε προγράμματα όπως είναι το 'Ambient Assisted Living'. Το πρόγραμμα αυτό ερευνά τρόπους βελτίωσης της ποιότητας ζωής των ατόμων τρίτης ηλικίας, με χρήση τεχνολογικών μέσων. Η υλοποίηση ενός τέτοιου αυτόματου συστήματος μπορεί να εξασφαλίσει την ανεξαρτησία του ατόμου, χωρίς να τεθεί η υγεία του σε κίνδυνο. Στην παρούσα εργασία, για την αναγνώριση ανθρώπινης συμπεριφοράς υλοποίηθηκαν και εφαρμόστηκαν αρχιτεκτονικές τεχνητών νευρωνικών δίκτυων. Επιπλέον, δόθηκε έμφαση στην αξιοποίηση των δεδομένων κατευθείαν από νευρωνικά δίκτυα, χωρίς προηγουμένως να έχουν υποστεί επεξεργασία, η οποία να τους έχει αλλάξει την μορφή τους.

Για την υλοποίηση της εκπαίδευσης και της αξιολόγησης των δικτύων χρησιμοποιήθηκε το NTU Dataset με 120 κατηγορίες, από τις οποίες επιλέχθηκαν οι 12 κατηγορίες που αφορούσαν ιατρικές καταστάσεις, καθώς κρίθηκαν περισσότερο σχετικές με το θέμα της εργασίας. Σε δεύτερη φάση, χρησιμοποιήθηκαν 40 κατηγορίες που σχετίζονται με καθημερινές δραστηριότητες. Τα δεδομένα, που αξιοποιηθηκαν, προέρχονταν από καταγραφή βίντεο και από αισθητήρα βάθους. Αρχικά, δοκιμάστηκε η αρχιτεκτονική που προτείνουν οι Zhu et al. στο [33] όμως κρίθηκε ακατάλληλη για το συγκεκριμένο πρόβλημα ταξινόμησης. Επομένως, υλοποιήθηκε και δοκιμάστηκε ένα απλό δίκτυο που συνδύαζε συνελικτικά 2D φίλτρα με LSTM και έδωσε ακρίβεια 75% στις εικόνες βάθους. Αντίθετα, στις έγχρωμες εικόνες (RGB), το συγκεκριμένο μοντέλο δεν μπόρεσε να εκπαιδευτεί ώστε να ξεχωρίσει τις κατηγορίες δραστηριοτήτων. Στη συνέχεια, δοκιμάστηκε η αρχιτεκτονική I3D στις εικόνες βάθους, η οποία πέτυχε ακρίβεια 82.4%, βάσει της αξιολόγησης cross-subject, ενώ στο cross-setup πέτυχε 80%. Όσον αφορά τις έγχρωμες εικόνες, η αρχιτεκτονική αυτή κατάφερε να φτάσει ακρίβεια 74.7% στο cross subject και 67.8% στο cross-setup. Σε επόμενο στάδιο, τα δίκτυα I3D, με εισόδους έχρωμες εικόνες και εικόνες βάθους, όταν συνενώθηκαν με ύστερη σύντηξη (late fusion) μέσου όρου, κατάφεραν να φτάσουν ακρίβεια 85.3% και 82.3% στο cross-subject και στο cross-setup αντίστοιχα. Τέλος, το I3D μοντέλο δοκιμάστηκε στις 40 κατηγορίες καθημερινών δραστηριοτήτων δίνοντας ακρίβειες 74.8% και 67.4% αντίστοιχα για τις εικόνες βάθους και τις έγχρωμες, βάσει της αξιολόγησης cross-setup. Κατόπιν εφαρμογής ύστερης σύντηξης των δύο δικτύων εκμετάλευσης διαφορετικού είδους πληροφορίας, η τελική ακρίβεια έφτασε στο 79.1%.

Σχετικά με τις κατηγορίες ιατρικών καταστάσεων, οι κατηγορίες έκτακτης ανάγκης, το 'staggering' και το 'falling down', αναγνωρίζονται με μεγάλη ακρίβεια από το μοντέλο. Ενώ, η κατηγορία 'sneeze/cough' είναι η λιγότερο αναγνωρίσιμη, καθώς έχει χαμηλά ποσοστά ακρίβειας από το μοντέλο ταξινόμησης. Όσο αφορά τις κατηγορίες καθημερινών δραστηριοτήτων, πάνω από τις μισές έχουν ποσοστό ακρίβειας μεγαλύτερο από 80%. Στις κατηγορίες που το μοντέλο δεν έδωσε καλά αποτελέσματα, παρατηρήθηκε ότι η χαμηλή ακρίβεια προέκυπτε επειδή υπήρχαν συγκεκριμένα ζευγάρια κλάσεων που, κυρίως, αυτά συγχέονταν.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Υπολογιστική Όραση, Μηχανική Μαθηση και Αναγνώριση Ανθρώπινων Κινήσεων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Αναγνώριση Δραστηριότητας, NTU 120 Dataset, Τεχνητά Νευρωνικά Δίκτυα, I3D

ABSTRACT

The present thesis examined the automatic recognition of human behavior, which can be applied to programs such as 'Ambient Assisted Living'. This project explores ways to improve the quality of life in older age, through the use of technological means. The achievement of such an automated system can guarantee one's independence without jeopardizing their health. In the present work, artificial neural network architectures were implemented and applied to identify human behavior. In addition, the emphasis was on utilizing the data directly from artificial neural networks, without having undergone preprocessing that changed their form.

The NTU Dataset of 120 categories was used to implement the training and the evaluation of the networks. From the whole dataset, the 12 categories, related to medical conditions, were selected as they were more relevant to the subject of the work. In the second phase, 40 categories, related to daily activities were used. The data that was utilized came from video capture and depth sensor. Initially, the architecture, proposed by Zhu et al. [33], was tested, but it was ultimately deemed inappropriate for this classification problem. Therefore, a simple network, combining 2D convolutional filters with LSTM, was implemented and tested. The accuracy of this network reached the 75 % in depth images. On contrast, when the input was the RGB data, this model failed to learn and generalize features. Subsequently, the I3D architecture was tested in depth data, which achieved an accuracy of 82.4% based on the cross-subject evaluation, while on cross-setup it achieved 80 %. As far as RGB data are concerned, this architecture managed 74.7 % accuracy in cross subject and 67.8 % in cross-setup. In the next step, when the two I3D networks, the one with Depth images input and the other with the RGB images, were conbined with average late fusion, they achieved 85.3 % and 82.3 % cross-subject and cross-setup accuracy respectively. Finally, the I3D model was tested in 40 daily activity categories giving 74.8 % and 67.4 % accuracy for Depth and RGB data, respectively, based on cross-setup evaluation. After the average late fusion of two networks, the final accuracy reached 79.1%.

Concerning the medical categories, emergency categories, such as 'staggering' and 'falling down', are accurately identified by the model. While the 'sneeze / cough' category is the least recognizable as it has the lowest accuracy. Regarding the categories of daily activities, more than half have an accuracy rate greater than 80%. In the categories where the model did not perform well, it was observed that the low accuracy was due to specific pairs of classes being mainly confused.

SUBJECT AREA: Computer Vision, Machine Learning and Human Action Recognition

KEYWORDS: Action Recognition, NTU 120 Dataset, Artificial Neural Networks, I3D