ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Νικοδήμου Βασίλειος- Κλείτος
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης
Επόπτης Μεταπτ. Εργασίας: Καθηγητής, Α. Αργυρός

Παρασκευή, 16/02/2018, 12:00
Αίθουσα Β108,Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

"Εκτίμηση της 3Δ Πόζας του Ανθρώπινου Χεριού από Μια Εικόνα Χρησιμοποιώντας Δίκτυα Συναρτήσεων Ακτινικής Βάσης Εκπαιδευμένα σε Συνθετικά Δεδομένα"

ΠΕΡΙΛΗΨΗ

Η παρακολούθηση και ανάλυση της ανθρώπινης κίνησης αποτελεί μια σημαντική κατηγορία προβλημάτων στον τομέα της Υπολογιστικής Όρασης. Μέσα σε αυτή την κατηγορία, τα προβλήματα που αφορούν στην εκτίμηση της 3Δ πόζας ενός ανθρώπινου χεριού είναι ιδιαίτερα ενδιαφέροντα. Αυτή η εργασία στοχεύει στην επίλυση του προβλήματος της εκτίμησης της πόζας ενός ανθρώπινου χεριού σε πραγματικό χρόνο, χρησιμοποιώντας μόνο οπτική πληροφορία. Πολλές προσεγγίσεις έχουν προταθεί για την επίλυση αυτού του προβλήματος, μεταξύ των οποίων και η εφαρμογή τεχνικών μηχανικής μάθησης. Η πρόσφατη επιτυχία των βαθέων νευρωνικών δικτύων σε προβλήματα υπολογιστικής όρασης έχει οδηγήσει σε σημαντική πρόοδο σε αυτόν τον τομέα. Ωστόσο, παρά τις έντονες προσπάθειες που έχουν αφιερωθεί στην επίλυση προβλημάτων αυτής της κατηγορίας, καμία μέθοδος δεν έχει καταφέρει να επιλύσει το πρόβλημα στη γενική του μορφή.

Οι τεχνικές μηχανικής μάθησης και ιδιαίτερα αυτές που βασίζονται σε βαθιά νευρωνικά δίκτυα απαιτούν μεγάλα επισημασμένα σύνολα δεδομένων για τη εκπαίδευσή τους. Η

επισήμανση σε πραγματικά δεδομένα είναι κοπιαστική και απαιτεί χρόνο και άλλους ανθρώπινους πόρους. Επομένως, προτιμάται ένας αυτόματος τρόπος δημιουργίας και επισημείωσης των δεδομένων εκπαίδευσης. Η χρήση συνθετικών δεδομένων παρέχει έναν εύκολο τρόπο για τη δημιουργία μεγάλου όγκου επισημασμένων δεδομένων υψηλής ακρίβειας. Ένα μειονέκτημα αυτής της προσέγγισης είναι ότι οι λεπτομέρειες των πραγματικών δεδομένων μπορεί να μην προσομοιωθούν με ικανοποιητική ακρίβεια κατά τη δημιουργία των συνθετικών δεδομένων. Οι υπάρχουσες τεχνικές μηχανικής μάθησης είναι ευαίσθητες στην κατανομή των δεδομένων εισόδου και συνεπώς, μπορεί να αποτύχουν να γενικεύσουν σε πραγματικά δεδομένα όταν εκπαιδεύονται σε συνθετικά δεδομένα.

Σε αυτή την εργασία παρουσιάζουμε μια νέα προσέγγιση για την εκτίμηση πόζας χεριού από δεδομένα βάθους από μία μόνο όψη. Πιο συγκεκριμένα, υποθέτουμε ότι η είσοδος είναι ένα μεμονωμένο καρέ δεδομένων βάθους, που απεικονίζει ένα απομονωμένο χέρι, δηλαδή ένα χέρι που δεν επικαλύπτεται από αντικείμενα στο περιβάλλον του. Το καρέ βάθους προσλαμβάνεται από έναν αισθητήρα βάθους και δεν χρησιμοποιούνται οπτικά βοηθήματα για να διευκολυνθεί η εργασία εντοπισμού του χεριού ή τμημάτων του. Η μέθοδος ακολουθεί μια στρατηγική εξειδίκευσης (coarse to fine) χρησιμοποιώντας δίκτυα συναρτήσεων ακτινικής βάσης (Radial Basis Function Networks, RBFNs) που εκπαιδεύονται σε ένα μεγάλο σύνολο συνθετικών δεδομένων.

Η δημιουργία των συνθετικών δεδομένων που απαιτούνται για την εκπαίδευση των δικτύων ξεκινάει καταγράφοντας μια πραγματική ακολουθία ενός ανθρώπινου χεριού που εκτελεί διαφορετικές χειρονομίες. Στη συνέχεια εκτιμάται η πόζα του χεριού για κάθε καρέ της ακολουθίας χρησιμοποιώντας μια μέθοδο παρακολούθησης χεριών με υψηλό υπολογιστικό φόρτο, επιτυγχάνοντας ακριβείς εκτιμήσεις. Από αυτό το σύνολο όλων των ανακτημένων ποζών, επιλέγουμε αυτές που διαφέρουν περισσότερο μεταξύ τους. Χρησιμοποιούμε αυτό το αντιπροσωπευτικό σύνολο, μαζί με μια πυκνή δειγματοληψία όλων των πιθανών περιστροφών για να δημιουργήσουμε το συνθετικό σύνολο εκπαίδευσης.

Ένα δίκτυο αρχικοποίησης και πολλαπλά εξειδικευμένα δίκτυα εκπαιδεύονται σε τμήματα του συνόλου συνθετικών δεδομένων. Υπάρχουν δύο ειδών εξειδικευμένα δίκτυα. Η μία κατηγορία, περιλαμβάνει δίκτυα που είναι κατάλληλα εκπαιδευμένα για να ανακτήσουν τον προσανατολισμό των χεριών με δεδομένη την άρθρωση των δακτύλων. Η δεύτερη κατηγορία ανακτά την άρθρωση δεδομένου του προσανατολισμού του χεριού. Λαμβάνοντας σαν είσοδο ένα καρέ βάθους, χρησιμοποιούμε τα εκπαιδευμένα μοντέλα για να ανακτήσουμε την πόζα του χεριού. Για το σκοπό αυτό, το δίκτυο αρχικοποίησης χρησιμοποιείται για την εκτίμηση μιας αρχικής πόζας. Στη

συνέχεια, τα εξειδικευμένα δίκτυα χρησιμοποιούνται σε ένα επαναληπτικό σχήμα που έχει σκοπό να βελτιώσει την αρχική εκτίμηση. Αυτό το επαναληπτικό σχήμα βελτίωσης εκτελείται για έναν προκαθορισμένο αριθμό επαναλήψεων, μετά την ολοκλήρωση του οποίου ανακτάται η εκτίμηση πόζας.

Το συνολικό υπολογιστικό κόστος της προτεινόμενης προσέγγισης καθορίζεται από τις υπολογιστικές απαιτήσεις ενός μικρού πλήθους δικτύων RBFN, επιτυγχάνοντας επιδόσεις σχεδόν πραγματικού χρόνου. Επιπλέον, η προτεινόμενη μέθοδος επιδέχεται επιτάχυνσης μέσω παραλληλοποίησης, χάρη στον εγγενή παραλληλισμό των δικτύων RBFN. Η μέθοδος απαιτεί λίγα πραγματικά δεδομένα και σχεδόν καθόλου επισημείωση. Επιπλέον, έχει λίγες υπερ-παραμέτρους οι οποίες διερευνούνται πειραματικά για να προσδιοριστούν οι βέλτιστες τιμές τους. Η ποσοτική αξιολόγηση της μεθόδου βασίστηκε σε μια πραγματική ακολουθία για την οποία γνωρίζουμε τις πραγματικές πόζες του χεριού (ground truth). Επιπλέον, παρουσιάζουμε ποσοτικά αποτελέσματα σε ένα κοινά διαθέσιμο σύνολο δεδομένων (public dataset) που χρησιμοποιείται για την αξιολόγηση των μεθόδων εκτίμησης και παρακολούθησης πόζας χεριών. Ποιοτικά αποτελέσματα παρουσιάζονται και για τα δύο σύνολα δεδομένων τα οποία δείχνουν ότι η προσέγγισή μας επιτυγχάνει ικανοποιητικά αποτελέσματα σε όλες τις περιπτώσεις. Συμπερασματικά, η εργασία αυτή δείχνει ότι προτεινόμενη προσέγγιση για εκτίμηση της 3Δ πόζας του χεριού που βασίζεται σε δίκτυα RBFN μπορεί να γενικεύσει αρκετά καλά σε πραγματικά δεδομένα, ενώ έχει εκπαιδευθεί σε συνθετικά δεδομένα.

Nikodemou Vassilis- Clitos
M.Sc. Thesis

Computer Science Department
University of Crete
Master's Thesis Supervisor: Professor A. Argyros

Friday, 16/02/2018, 12:00
Room B108, Computer Science Dept., University of Crete

“Single Shot 3D Hand Pose Estimation Using Radial Basis Function Networks Trained on Synthetic Data”

ABSTRACT

Human motion tracking and analysis forms an important category of problems in the field of Computer Vision. Within this category, the class of problems that deal with the

estimation of the full pose of a human hand are especially interesting. This thesis treats the problem of estimating in real time the full pose of a human hand, using only visual input. Many approaches have been proposed to solve this problem, including applying machine learning techniques. The recent success of deep neural networks for computer vision tasks has resulted in new advancements in this area. Despite the significant effort that has been devoted to the problem of 3D hand pose estimation, no method has succeeded to tackle the problem in its full generality.

Machine learning approaches and, in particular, deep learning ones require large annotated datasets for training. The annotation process in real-world data is human labor intensive and time consuming. Therefore, an automatic way of creating and annotating training data is preferable. The use of synthetic data provides an easy way to obtain large volumes of accurately annotated data. On the negative side of using synthetic data, details of the real data may not be accurately simulated. Existing machine learning techniques are sensitive to the distribution of input data and may fail to generalize to real-world data when trained on synthetic data.

In this thesis we present a novel framework to perform single shot 3D hand pose estimation from depth maps. More specifically, the input is assumed to be a single depth map, depicting a single hand in isolation, that is, not occluded by its surroundings. The depth map is acquired using a depth sensor, and no visual aids (e.g., markers) are used to facilitate the task of localizing the hand or parts of it. The method follows a coarse-to-fine strategy, employing Radial Basis Function Networks (RBFNs) that are trained on a large synthetic dataset.

In order to synthesize the dataset that is used to train the RBFNs, we capture a real-world sequence of a human hand performing a set of diverse hand gestures. We proceed to estimate the hand pose for each frame of the sequence using an offline hand tracking method with high computational budget, achieving accurate estimations. Given the set of all the recovered hand poses, we proceed to select the most diverse of them. We use this representative set, along with a dense sampling of all possible rotations as a seed to generate the synthetic training set.

An initialization RBFN and multiple specialized RBFNs are trained on parts of this large synthetic dataset. There are two classes of specialized RBFNs. One class is appropriately trained to recover the global hand rotation given the hand articulation and the second one to recover the global hand articulation given the hand rotation. Given an input depth map, we use the trained models to recover a hand pose. Towards this end, the initialization RBFN is used to provide a rough pose estimation. Subsequently, the specialized RBFNs are employed in an iterative refinement scheme in order to improve

the initial estimation. This iterative refinement scheme is repeated for a predetermined number of repetitions, after the completion of which the final estimation is retrieved.

The overall computational cost of the proposed approach is dominated by the computation of several RBFNs, yielding in practice a system that achieves close to real-time performance. Furthermore, the proposed method is parallelizable, taking advantage of the inherent data-parallelism of RBFNs. The method requires few real-world data and virtually no manual annotation, and it has few hyper-parameters that are experimentally investigated to identify their optimal values. We perform a quantitative evaluation of our method on a test sequence of our own. Additionally, we present quantitative results on a public dataset that is commonly used to evaluate hand pose estimation and tracking methods. Qualitative results are also presented for both datasets. We show that our approach achieves promising results in all cases. Conclusively, this work shows that the proposed RBFNs-based approach can generalize quite well when learning from synthetic data.