ΠΡΟΣ

1) **Όλα τα μέλη ΔΕΠ του Τμήματος Επιστήμης Υπολογιστών**

2) **Τους εκπροσώπους των Μεταπτυχιακών φοιτητών του Τμήματος Επιστήμης Υπολογιστών**

3) **Την Επταμελή Εξεταστική Επιτροπή**

4) **Όλα τα μέλη της Πανεπιστημιακής Κοινότητας**

Πρόσκληση σε Δημόσια Παρουσίαση της Διδακτορικής Διατριβής του

κ. Παπουτσάκη Κωνσταντίνου

Doctoral Dissertation Defense

Mr. Konstantinos Papoutsakis

Την Παρασκευή, 06/12/2019 και ώρα 12:00 στην αίθουσα Τηλεδιάσκεψης Κ206 του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης στο Ηράκλειο, θα γίνει η δημόσια παρουσίαση και υποστήριξη της Διδακτορικής Διατριβής του υποψηφίου διδάκτορα του Τμήματος Επιστήμης Υπολογιστών κ. Παπουτσάκη Κωνσταντίνου με θέμα:

“Μη-εποπτευόμενη συν-τμηματοποίηση δράσεων σε ακολουθίες δεδομένων κίνησης και εικόνων”

“Unsupervised co-segmentation of actions in motion capture data and videos”

ΠΕΡΙΛΗΨΗ

Στην παρούσα διατριβή εστιάζουμε στο πρόβλημα της χρονικής συν-τμηματοποίησης δράσεων σε ακολουθίες πολυδιάστατων δεδομένων κίνησης (motion capture data) και σε ακολουθίες εικόνων (βίντεο). Δοσμένων δύο ακολουθιών δεδομένων που αναπαριστούν δράσεις/δραστηριότητες, στόχος είναι να εντοπίσουμε και να ορίσουμε τα χρονικά όρια για όλα τα ζεύγη υπο-ακολουθιών που αναπαριστούν μια κοινή δράση (common action or commonality), δηλαδή μια δράση που επαναλαμβάνεται ταυτόσημη ή με παρόμοιο τρόπο μεταξύ των ακολουθιών. Το εν λόγω πρόβλημα αποτελεί σημαντικό ερευνητικό θέμα στις περιοχές της Αναγνώρισης Προτύπων και της Υπολογιστικής Όρασης και παρά την ερευνητική προσπάθεια που έχει αφιερωθεί σε αυτό, δεν έχει επιλυθεί πλήρως.

Η παρούσα διατριβή περιγράφει μια αποδοτική, μη-εποπτευόμενη προσέγγιση η οποία δεν προϋποθέτει εκ των προτέρων γνώση και μοντέλα των δράσεων που εκτελούνται, ενώ υιοθετεί μια γενική και ευέλικτη μοντελοποίηση των δεδομένων εισόδου ως πολυδιάστατες χρονοσειρές. Θεωρούμε διαφορετικά σενάρια για τις ακολουθίες δράσεων που δημιουργούν ενδιαφέρουσες προκλήσεις ως προς την επίλυση του προβλήματος: (α) σε κάθε ακολουθία εμφανίζονται μία ή περισσότερες δράσεις που πραγματοποιούνται από ένα ή περισσότερα υποκείμενα (άτομα ή αντικείμενα), (β) στη γενική περίπτωση, ο αριθμός των κοινών δράσεων μεταξύ δύο ακολουθιών θεωρείται άγνωστος, (γ) μια κοινή δράση μπορεί να εντοπιστεί σε οποιοδήποτε χρονικό τμήμα μιας ακολουθίας, (δ) τα τμήματα μιας κοινής δράσης μεταξύ δύο ακολουθιών ενδέχεται να έχουν διαφορετική διάρκεια, να περιλαμβάνουν κινήσεις διαφορετικής ταχύτητας και τρόπου/τεχνικής εκτέλεσης, (ε) οι δράσεις που εμφανίζονται στις ακολουθίες ενδέχεται να αναπαριστούν φυσικές κινήσεις ενός ή περισσότερων ανθρώπων ή αντικειμένων, καθώς επίσης και περίπλοκες αλληλεπιδράσεις ανθρώπων με αντικείμενα.

Προτείνουμε δύο καινοτόμες μεθόδους για την επίλυση του προβλήματος της χρονικής συν-τμηματοποίησης δράσεων σε ζεύγη ακολουθιών δεδομένων. Η πρώτη μέθοδος επιτυγχάνει την ανίχνευση και συν-τμηματοποίηση των Ν σημαντικότερων κοινών δράσεων μεταξύ των υπο σύγκριση ακολουθιών δεδομένων, βασιζόμενη στην ελαχιστοποίηση συνάρτησης κόστους που εκφράζει το κόστος μη-γραμμικής χρονικής στοίχισης των υπο-ακολουθιών των κοινών δράσεων, χρησιμοποιώντας την μέθοδο Dynamic Time Warping (DTW). Η διαδικασία ταυτόχρονης αναζήτησης λύσεων (κοινών δράσεων) και ελαχιστοποίησής της συνάρτησης κόστους μοντελοποιείται ως ένα στοχαστικό πρόβλημα βελτιστοποίησης, το οποίο λύνεται βάσει της εξελικτικής μεθόδου βελτιστοποίησης Canonical Particle Swarm Optimization (PSO). Η δεύτερη μέθοδος βασίζεται στην μοντελοποίηση του ίδιου προβλήματος, ως ένα πρόβλημα αναζήτησης σε γράφο. Ο γράφος ορίζεται ως ο πίνακας (μήτρα) που περιλαμβάνει τις Ευκλείδειες αποστάσεις όλων των δυνατών ζευγών καρέ των ακολουθιών εικόνων, καθένα από τα οποία αναπαρίσταται ως ένα διάνυσμα χαρακτηριστικών. Γίνεται χρήση του αλγορίθμου Johnson's για την αναζήτηση των συντομότερων μονοπατιών σε γράφο και κατ'επέκταση για την επίλυση του προβλήματος. Και οι δύο πρωτότυπες μεθοδολογίες υποβάλλονται σε εκτενείς πειραματικές διαδικασίες χρησιμοποιώντας πλήθος από ζεύγη ακολουθιών εικόνων (βίντεο) ή ακολουθιών που αναπαριστούν 3Δ δεδομένα καταγραφής κίνησης, αναδεικνύοντας την αποτελεσματικότητάς τους σε σύγκριση με άλλες υφιστάμενες αποδοτικές μεθόδους.

Επιπρόσθετα, βασιζόμενοι στην εύρωστη απόδοση των μεθόδων αυτών, αναπτύσσουμε μια νέα μέθοδο για την εκτίμηση της ομοιότητας μεταξύ δύο ακολουθιών δράσεων, που επίσης υποστηρίζει την εξαγωγή επιχειρημάτων που αιτιολογούν τον υπολογισμό αυτό. Η μέθοδος αυτή βασίζεται στην χρονική συν-τμηματοποίηση των ζευγών 3Δ τροχιών κίνησης των ανθρώπινων αρθρώσεων και των αντικειμένων που παρατηρούνται στις ακολουθίες, συνδυάζοντας επιπλέον εκ των προτέρων γνωστή σημασιολογική πληροφορία για τις κατηγορίες τους και

υπολογίζοντας την σημασιολογική τους ομοιότητα. Τα αποτελέσματα αυτής της διαδικασίας ανα ακολουθία μοντελοποιούνται ως ένας γράφος που αναπαριστά το περιεχόμενό της ακολουθίας ανά αντικείμενο. Συγκεκριμένα, κάθε αντικείμενο αντιστοιχεί σε ένα κόμβο του γράφου. Οι ακμές του γράφου μοντελοποιούν πληροφορία με βάση τα αποτελέσματα της χρονικής συν-τμηματοποίησης μεταξύ των αντικειμένων της ακολουθίας και της σημασιολογικής τους πληροφορίας, εφόσον αυτή είναι διαθέσιμη. Στη συνέχεια η ομοιότητα/απόσταση μεταξύ δύο ακολουθιών δράσεων βασίζεται στην απόσταση (Graph Edit Distance) μεταξύ των αντίστοιχων γράφων τους, και υπολογίζεται ως το κόστος μιας βέλτιστης λύσης αντιστοίχισης (bipartite graph matching) σε διμερή γράφο που συντίθεται απο τους δύο επιμέρους γράφων. Η προτεινόμενη μεθοδολογία αξιολογείται πειραματικά στα προβλήματα της κατηγοριοποίησης δράσεων, της αντιστοιχίας δράσεων (action matching) και στον υπολογισμό της σειράς κατάταξης μεταξύ ζευγών δράσεων με βάση την ομοιότητά τους (pairwise action ranking) ανάμεσα σε τριπλέτες ακολουθιών εικόνων. Τα αποτελέσματα οδηγούν στο συμπέρασμα ότι η προτεινόμενη μέθοδος έχει αξιόλογη απόδοση, συγκρίσιμη ή και καλύτερη αυτής των καλύτερων γνωστών σύγχρονων μεθόδων μη εποπτευόμενης και εποπτευόμενης μάθησης.

Λέξεις κλειδιά: χρονική συν-τμηματοποίηση δράσεων/δραστηριοτήτων, ομοιότητα ακολουθιών εικόνων, ομοιότητα ακολουθών δεδομένων καταγραφής κίνησης, αναζήτηση/ανάκτηση ακολουθιών εικόνων, κατάταξη ομοιότητας δράσεων κατά ζεύγη, αντιστοίχιση/ταίριασμα δράσεων, αναγνώριση δράσεων/δραστηριοτήτων.

Επιβλέπων: Καθηγητής, Αντώνης Αργυρός

ABSTRACT

We focus on the problem of temporal co-segmentation of actions in sequences of 3D motion capture data and in image sequences (videos). Given two data sequences representing action relevant information, the goal is to detect and temporally co-segment all pairs of matching sub-sequences (temporal segments), where the segments of a pair represent a common (identical or similar) action or sub-action. This is an important and challenging problem in the research communities of Computer Vision, Pattern Recognition and Machine Learning, which despite the research efforts devoted to its solution, remains unsolved in its full generality.

We investigate the problem of interest by following a data-driven, unsupervised approach, where no a-priori models and labels of the actions represented in the

sequences are available. Various challenging scenarios and conditions are considered, i.e., (a) one or multiple actions are demonstrated by different subjects in each sequence, (b) the number of common actions between the sequences may be unknown, (c) the common actions may be located anywhere in the sequences, (d) instances of the common action or sub-action can be of variable duration and of different speed and execution style and (e) actions may involve a single or multiple humans, generic objects or even complex human-object interactions.

Two novel, efficient methodologies are proposed in this thesis to deal with this problem. They are based on a stochastic optimization approach and a deterministic, graph-based approach, respectively. Furthermore, we leverage the robust performance of the proposed temporal action co-segmentation strategies to develop a method that estimates the similarity of the original sequences and provides meaningful arguments supporting this estimation, making a step towards explainable assessment of video and action similarity.

Specifically, two novel methods are introduced to perform temporal co-segmentation between two sequences of motion capture data (3D/6D human skeletal data and/or object pose data) or of RGB images. Each data sequence is treated as a multivariate time-series for any of the data modalities. The first method discovers and co-segments the N best pairs of common sub-sequences (commonalities) between the compared time-series by minimizing a cost function that expresses their non-linear temporal alignment cost. The cost is quantified using the Dynamic Time Warping (DTW) method and its minimization is treated as a stochastic optimization problem that is solved using Canonical Particle Swarm Optimization (PSO). The PSO method relies on evolutionary search strategies to minimize the DTW-based cost function and is applied iteratively in order to discover the N best commonalities. The second method treats temporal action co-segmentation as a search problem on a graph defined on the matrix of the pair-wise Euclidean distances (EDM) of the frame-wise features between the two compared time-series. An efficient graph-based search algorithm is used for solving the problem of discovering N commonalities. The number of the N best commonalities to be discovered for two time-series may be unknown or given a-priori. Both methods have been extensively tested using pairs of image sequences (videos) or pairs of sequences containing 3D motion capture data. Various types of action scenarios have been considered such as physical exercises, daily living activities and human-object interaction, while quantitative experiments demonstrate the effectiveness of the proposed methods in comparison to existing, state-of-art approaches.

In addition, a novel method is proposed for fine-grained similarity assessment of two actions in videos that capitalizes on the effectiveness of temporal co-segmentation between the trajectories of the tracked human joints and/or the tracked objects and their semantic relatedness. A graph matching approach based on Graph Edit Distance is employed to combine the object-level features and semantic information, towards

computing spatio-temporal correspondences between objects across videos, if these objects are semantically related, if/when they interact similarly, or both.

The proposed framework aspires to take an important step towards explainable assessment of video and action similarity. It is evaluated on publicly available datasets on the tasks of action classification, action matching and action-based ranking in triplets of videos and is shown to compare favorably to state-of-the-art unsupervised and supervised learning methods.

Keywords: Temporal action co-segmentation, video similarity, temporal alignment, pairwise action ranking, action matching, action recognition, Graph Edit Distance, Particle Swarm Optimization.

Supervisor: Professor, Antonis Argyros