

Human Action Prediction & Forecasting based on Visual Data

VICTORIA MANOUSAKI



UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF SCIENCES AND ENGINEERING

Report for the Fulfillment of the
Ph.D. Graduate Examinations Requirements

OCTOBER 2023



Operational Programme
Human Resources Development,
Education and Lifelong Learning

Co-financed by Greece and the European Union



The implementation of the doctoral thesis was co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the Act “Enhancing Human Resources Research Potential by undertaking a Doctoral Research” Sub-action 2: IKY Scholarship Programme for PhD candidates in the Greek Universities.

UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE

**Human Action Prediction & Forecasting
based on Visual Data**

Report for the Fulfillment of the
Ph.D. Graduate Examinations Requirements

Ph.D. Candidate:
Victoria Manousaki

Advisory Committee:

Supervisor: Antonis Argyros, Professor, Computer Science Department, University of Crete

Committee Member: Dimitrios Kosmopoulos, Associate Professor, Computer Engineering and Informatics Department, University of Patras

Committee Member: Anastasios Roussos, Principal Researcher, Institute of Computer Science, Foundation for Research and Technology

UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE
**Human Action Prediction & Forecasting
based on Visual Data**

PhD Dissertation Presented
by **Victoria Manousaki**

in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved by:



Author: Victoria Manousaki



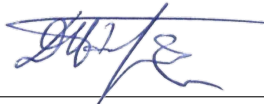
Supervisor: Antonis Argyros, Professor, University of Crete



Committee Member: Dimitrios Kosmopoulos, Associate Professor, University of Patras



Committee Member: Anastasios Roussos, Principal Researcher, Institute of Computer Science,
FORTH



Committee Member: Dimitrios Plexousakis, Professor, University of Crete



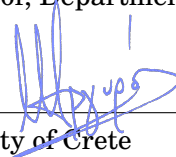
Committee Member: Panos Trahanias, Professor, University of Crete



Committee Member: Constantine Stephanidis, Professor, University of Crete



Committee Member: Costas Panagiotakis, Associate Professor, Department of Management
Science and Technology, Hellenic Mediterranean University



Department Chairman: Antonis Argyros, Professor, University of Crete

Heraklion, October 2023

Acknowledgments

I wish to extend my sincere gratitude to my supervisor, Prof. Antonis Argyros, for his invaluable guidance and collaboration throughout the course of my studies. His mentorship has been instrumental in completing this dissertation, and without his support, this PhD thesis would not have been feasible. Additionally, I would like to express my appreciation for his encouragement in facilitating my exploration of research topics that align with my academic interests.

I would also like to extend my appreciation to my supervisory committee, Prof. Kosmopoulos and Dr. Roussos, for their unwavering support and sustained interest in this thesis throughout the years. Furthermore, I express my gratitude to the examination committee: Prof. Plexousakis, Prof. Trahantias, Prof. Stephanidis, and Prof. Panagiotakis. Thank you for accepting the invitation to participate in the defense of this thesis, your interest in my work, and your devotion of the necessary time and effort, positively contributing to the examination process. I would also like to thank Konstantinos Papoutsakis for his support and guidance throughout this thesis.

I extend my sincere appreciation to my family for their unwavering support throughout my academic journey. Specifically, I wish to express profound gratitude to my parents, Lefteris and Konstantina, for their enduring support and for providing me with the opportunity to pursue my education. Furthermore, I want to thank Ioannis for his unconditional love and support throughout these years.

Abstract

The ability to observe human movements and predict their actions is a developmental skill acquired by humans early in life. When witnessing a person performing a task, we can easily forecast their subsequent actions based on contextual cues and past experiences. In this work, we aim at developing such abilities for machines, focusing on the tasks of vision-based action prediction, action anticipation and next-active-object prediction.

Action prediction is defined as the inference of an action label while the action is still ongoing. Such a capability is useful for early response and further action planning. We consider the problem of action prediction in scenarios involving humans interacting with objects. We formulate an approach that builds time series representations of the performance of the humans and the objects. Such a representation of an ongoing action is then compared to prototype actions. This is achieved by a Dynamic Time Warping (DTW)-based time series alignment framework which identifies the best match between the ongoing action and the prototype ones. We predict actions in trimmed and untrimmed action sequences with the use of the DTW algorithm. In the same vein, for the prediction of actions we propose two new alignment algorithms called OBE-S-DTW and OE-S-DTW that show superior results on the task of action prediction compared to DTW.

Following, we propose a graph-based methodology for the visual prediction of human-object interactions in videos. Rather than forecasting the human and object motion, we aim at predicting (a) the class of the on-going human-object interaction and (b) the class(es) of the next active object(s) (NAOs), i.e., the object(s) that will be involved in the interaction in the near future as well as the time the interaction will occur.

Finally, we address the problem of action anticipation by taking into consideration the history of all executed actions throughout long, procedural activities. A novel approach noted as Visual-Linguistic Modeling of Action History (VLMAH) is proposed that fuses the immediate past in the form of visual features as well as the distant past based on a cost-effective form of linguistic constructs (semantic labels of the nouns, verbs, or actions). Our approach generates accurate near-future action predictions during procedural activities by leveraging information on the long- and short-term past.

The proposed methods constitute solutions for the problems of action prediction and anticipation and next-active-object prediction. The aforementioned methodologies have been evaluated on challenging datasets and showcase results superior to the current state-of-art.

TABLE OF CONTENTS

	Page
List of Tables	vi
List of Figures	viii
1 Introduction	1
1.1 From Action Recognition to Action Prediction	3
1.2 Prediction	5
1.3 Our approach	8
1.4 Motivation	10
1.5 Organization of the thesis	12
2 Literature Review	13
2.1 Action & Activity Recognition	13
2.2 Action & Activity Prediction	14
2.3 Action Anticipation/Forecasting	16
2.4 Next Active Object Prediction	17
2.5 Temporal Alignment	20
2.6 Contribution & Open Challenges	21
3 DTW-based action prediction in trimmed sequences	23
3.1 Problem description	24
3.2 Feature Extraction	24
3.3 DTW-based Time Series Alignment	25
3.4 Early Fusion of Human and Object Representations	26
3.5 Experimental Evaluation	27

3.5.1	Datasets & feature representations:	27
3.5.2	Performance metrics:	29
3.5.3	Evaluation of DTW variants:	29
3.5.4	Evaluation of alternative representations:	30
3.5.5	Comparison to the state of the art:	31
3.6	Summary	32
4	Action Prediction in trimmed/untrimmed sequences	35
4.1	Problem description	36
4.2	Existing alignment methods	37
4.3	Proposed alignment methods	38
4.4	Alignment-based action prediction	40
4.5	Datasets & Metrics	42
4.5.1	Datasets	42
4.5.2	Performance metrics	45
4.5.3	Implementation issues	45
4.6	Experimental Evaluation	46
4.6.1	Evaluation in trimmed actions	46
4.6.2	Evaluation in untrimmed actions	48
4.6.3	Duration Prediction	52
4.7	Summary	53
5	Graph-based Action Prediction	55
5.1	Problem description	56
5.2	Graph-based video representation	58
5.3	Video comparison based on Graph Edit Distance	59
5.4	Datasets & Metrics	62
5.4.1	Datasets	62
5.4.2	Feature Extraction	62
5.4.3	Evaluation Metrics	62
5.5	Experimental Results	63
5.5.1	Activity Prediction/Early Recognition	63

5.5.2	The impact of parameter λ	65
5.5.3	Next-Active-Object Prediction	65
5.5.4	Next-Active-Object Time Prediction	66
5.5.5	Multiple Next-Active-Objects Prediction	67
5.5.6	Ablation Study	67
5.5.6.1	Temporal alignment	67
5.5.6.2	Duration Prognosis	68
5.6	Summary	71
6	Action Forecasting/Anticipation	73
6.1	Problem description	75
6.2	Visual-Linguistic Modeling of Action History framework	76
6.2.1	Visual Action Anticipation Module	76
6.2.2	Linguistic Action History Module	77
6.2.3	Visual Action Recognition Module	78
6.3	Experimental Setup	81
6.3.1	Datasets	81
6.3.2	Training, Testing & Input Configurations	83
6.4	Experimental Results	83
6.4.1	Action Anticipation	83
6.4.2	How much history is enough?	88
6.4.3	History-based action anticipation	89
6.4.4	Noise resistance	90
6.5	Summary	91
7	Conclusions	93
7.1	Contributions	93
7.2	Impact	95
7.3	Limitations	97
7.4	Directions for future work and research	98
7.5	Acknowledgements	99
	Bibliography	101

LIST OF TABLES

TABLE	Page
5.1 Next-active-object prediction accuracy for [2s, 1.75s, 1.5s, 1.25s, 1s, 0.75s, 0.5s, 0.25s] before the beginning of the next action for the CAD-120 dataset.	65
5.2 Time prediction error is the offset of the predicted time of the next-active-object use to the ground truth time of use compared to video length. Predictions are made from 0.25s to 2s prior to the start of the next action.	67
5.3 Accuracy for predicting multiple next-active-objects for different observation ratios. . .	67
6.1 Action anticipation accuracy for different timesteps (prior to the beginning of the next segment) for the Meccano dataset . $VLMAH_{GT}$ and $VMAH_{GT}$ represent the two variants of the proposed method when <i>ground truth annotations</i> are used as the linguistic action history. VLMAH makes use of the Linguistic Action History module while the action history is generated from the visual action recognition module. The comparison is between the [116] and the VLMAH methods.	85
6.2 Top-1/Top-5 accuracy results of [136] and the VLMAH variants on the Assembly-101 dataset for anticipation time $\tau_{ant} = 1s$, with or without the use of the linguistic action history module. TempAgg* denotes the single-task learning variant.	87
6.3 Top-1 accuracy results on the 50Salads dataset for the anticipation time $\tau_{ant} = 1s$. . .	87
6.4 The Top1 and Top5 accuracy scores achieved by the proposed framework using variable lengths of the linguistic action history on the Assembly-101 dataset . Zero percent (0%) is equivalent to the use of $VMAH_{GT}$ variant, while other action history percentage values refer to the use of the $VLMAH_{GT}$	89

6.5	Top-1 and Top-5 accuracy results for noun anticipation on the Meccano dataset of our VLMAH framework with different types of features and with the use of the visual action history (VMAH) module. The linguistic module contributes significantly to the anticipatory capability of the framework.	90
6.6	Evaluating the anticipatory capacity of the VLMAH framework on the Assembly-101 dataset for variable noise levels (disturbances) in the form of erroneous semantic labels in the action history.	91

LIST OF FIGURES

FIGURE	Page
1.1 Action recognition and action prediction presented on the same sequence [129]	4
2.1 Task of anticipating short-term object interactions. Models analyze video sequence V up to timestamp t , making predictions about the bounding box and class of upcoming active objects, the verb characterizing the impending interaction, a numeric value denoting when the interaction occurs ($t + \delta$), and a confidence score. Here, δ signifies the time gap between the final observable frame V_t and the frame of interaction at time ($t + \delta$). Image from [115].	18
3.1 Comparison of DTW_{oe} , SDTW and GAK on the MSR (upper left) CAD-120 (upper right) and MHAD (down) datasets.	30
3.2 Action prediction results for different representations. (Upper left) MSR Daily Activities Dataset, (Upper right) CAD-120 Dataset. (Down) Action prediction accuracy of our method in comparison to state of the art methods on the MSR Daily Activities Dataset.	31
4.1 Graphical illustration of the OE-S-DTW algorithm. On the horizontal axis we can observe a man performing an activity. On the vertical axis a woman is performing the same activity which is not yet completed. The light pink boxes represent the possible alignment paths while the black arrows represent a possible path. The two sequences share the same starting point but end at different points. The OE-S-DTW algorithm is able to match the partially observed activity with a part of the completely observed one.	39

4.2	Graphical illustration of the OBE-S-DTW algorithm. The activity illustrated at the left (rows) matches a part of the activity illustrated on the top (columns). At the top a zero-valued row is added. The light blue boxes represent all possible alignments while the black arrows show a possible warping path.	41
4.3	Action prediction accuracy in trimmed videos as a function of observation ratio involving skeletal features in the MHAD (Upper left), MSR (Upper right) and CAD-120 (Down) datasets. We compare the different alignment algorithms which are OBE-S-DTW (proposed), OBE-S-DTW (proposed), OE-DTW [143], OBE-DTW [143], OTAM [12] and SegmentalDTW [103].	48
4.4	Activity prediction results on the CAD120 dataset.	49
4.5	Action prediction accuracy in trimmed videos as a function of observation ratio involving VGG-16 features in the MHAD dataset.	49
4.6	Action prediction accuracy of our methods in comparison to state of the art methods on the MSR Daily Activities dataset.	50
4.7	Action prediction accuracy in untrimmed videos (video triplets) of the MHAD101 dataset as a function of observation ratio involving skeletal features (MHAD101-s, left) and VGG-16 features (MHAND101-v, right).	50
4.8	Performance metrics (F1-score, precision, recall and Intersection-Over-Union) for OBE-DTW and OBE-S-DTW on the MHAD101-s dataset. Prefix and postfix are denoted with the black vertical lines. Between these lines lies the action to be predicted which is observed in tenths.	51
4.9	Aligning unsegmented action sequences on the MHAD101-s/-v datasets using the OBE-S-DTW and OBE-DTW algorithms. Dark red and blue lines depict the accuracy of the alignment algorithms while observing the triplets from the prefix to the suffix. The light (red and blue) lines depict the observation of the triplet from the suffix to the prefix.	53
4.10	Percentage of the frames that are lost compared to the ground truth duration of the action.	54

5.1	By matching a partially executed and observed activity, to a prototype, fully observed one, we are able to infer correspondences of similar objects and human joints between the two videos. This, in turn, enables to perform activity and next-active-object prediction in the partially observed activity. The example in this figure refers to the “stacking objects” activity, which is performed with a different number and types of objects in the partially and the fully observed activities.	57
5.2	Graph matching of a complete video (reference) and an incomplete/partially observed (test) video. First, the fully connected graphs of each video are created based on the video entities. On the basis of these graphs, a bipartite graph between the action graphs is constructed. By calculating the GED, we are able to correspond nodes between the two original action graphs.	58
5.3	Observing the activity and making object predictions for [2s, 1.75s, 1.5s, 1.25s, 1s, 0.75s, 0.5s, 0.25s] before the beginning of the next action as in [35].	63
5.4	Activity prediction results for the (left) MSR Daily Activities and (right) CAD-120 datasets for different observation ratios.	64
5.5	Exploration of the user-defined λ parameter on the CAD-120 dataset. The values of the λ parameter are in the range [0, 1]. Some curves may be partially visible due to occlusions. Plots are separated in two figures to aid readability.	65
5.6	Activity prediction results on the CAD120 dataset using the GTF framework. The OBE-S-DTW and OE-S-DTW algorithms were used to quantify the motion dissimilarity of the entities involved in the considered activities.	68
5.7	End-frame prediction error calculated for all observation ratios of the middle actions of the triplets of MHAD101-s.	69
5.8	(Left) Action prediction results for the MSR Daily Activities dataset. (Right) Prognosis of the duration of the partially observed actions for the MSR Daily Activities dataset.	69
5.9	(Left) Activity prediction results for the CAD120 dataset. (Right) Prognosis of the duration of the partially observed activities of the CAD120 dataset.	70

6.1	We consider the problem of action anticipation in untrimmed videos of procedural activities. At a certain moment in time (decision point), the proposed framework (VLMAH) [88] anticipates the action (i.e., the unobserved action “take screw”) that is most likely to be performed after some anticipation time T_{ant} (depicted with orange color). This is performed on the basis of the history of all past actions up to the decision point (depicted with purple) which is modeled by integrating visual input regarding the immediate past and a linguistic description of the distant past.	74
6.2	The proposed VLMAH architecture. The Visual Action module and the Linguistic Action History module are presented. For the <i>Meccano</i> dataset, the encoders of the action module, generate Object, Hands, Gaze representations, whereas for the <i>Assembly-101</i> dataset, there is a single encoder network, TSM [75] while representations are split into 3 sub-sequences, as mentioned in Section 6.3.2. The detail level regarding the textual label descriptions is adaptable to the anticipation task at hand (action, motion motif (verb), or object (noun)). The final format also includes two special labels (START, END) that indicate the start and end of the action history sequence.	80

INTRODUCTION

The ability to observe human movements and predict their actions is a developmental skill acquired early in life. From a young age, children begin to understand and anticipate the actions of others. For example, when a baby sees someone raising their arms, they may predict that the person is about to pick them up. As children grow older, this skill becomes more refined, and they can make increasingly accurate predictions about the intentions of those around them.

When witnessing a person performing a task, we can easily forecast their subsequent actions based on contextual cues and past experiences. For instance, if we see someone walking towards a door while reaching for the doorknob, we can predict that they are about to open the door and step through it. This ability to predict actions is fundamental for social interactions, enabling us to respond appropriately to others and anticipate their needs. However, machines do not have this natural ability. Researchers and engineers use various techniques to develop computational models capable of predicting human actions. By feeding the system with large datasets containing action sequences and their corresponding outcomes, the machine learns to recognize patterns and make accurate predictions.

As the term "predicting actions" is quite broad, researchers often focus on specific areas to study and apply action prediction. For instance, action prediction can be applied in sports, where the system predicts the next move of players during a game [33]. In autonomous vehicles [27],

action prediction is crucial to anticipate the behavior and movement of pedestrians [138] and other vehicles on the road [34] in order to plan the route and avoid accidents. Human action prediction and/or forecasting refer to the anticipation of the forthcoming action label based on limited initial observations of the ongoing action. For example, consider a scenario where a person is cooking in a kitchen. Based on the initial frames of a video, an action prediction model should be able to predict whether the person is cutting vegetables, stirring a pot, or reaching for ingredients [19, 20]. This task has been explored by many researchers as it can provide information about people's interaction with objects and their intentions. Action prediction can help in the human-robot collaboration while following a recipe, in order to be able to provide assistance, prepare the tools, anticipate accidents, etc.

These actions may involve one or multiple subjects, making the task more complex. In a scene with multiple people, the system must identify and predict the actions of each individual correctly. Moreover, there can be one or more objects present in the scene that influence human actions. For instance, in a soccer match, the ball's position and trajectory will significantly impact players' actions, making action prediction more challenging and context-dependent.

Action prediction is not limited to specific environments; there are no restrictions on the scene type, allowing for actions to occur both indoors and outdoors. For instance, action prediction has been explored indoors in different contexts such as cooking [20] and assembling toys [116, 136] while it has also been explored outdoors in various scenarios such as pedestrian trajectory prediction [67] and assembling a tent [54]. The observations may encompass various modalities, such as RGB or grayscale images, RGB-D videos, and motion capture data. For example, in a gesture recognition system, motion capture data from a user's hand movements can be used to predict the intended action, like swiping or pointing.

Human action prediction has gained prominence in the field of Computational Vision due to its significance in early action recognition. In video surveillance, action prediction plays a vital role in proactively detecting potential threats and abnormal behavior. For instance, a security system can predict suspicious activities, such as someone leaving a package unattended in a public place.

Action prediction faces several challenges similar to those encountered in action recognition. In complex scenes, occlusions can obstruct the view of certain actions, making prediction more difficult. For example, in a crowded street, people walking in front of each other can create occlusion,

hindering the visibility of some actions. Lighting variations and scene background complexities can also impact action prediction accuracy. For instance, outdoor scenes with changing lighting conditions may introduce uncertainty in predicting actions due to altered appearances. Additionally, action prediction inherently involves ambiguity due to the limited number of observations, resulting in multiple potential predictions. In situations where the context is not clear, the system should provide multiple hypotheses, each with an associated confidence level, to account for uncertain predictions.

To address these challenges, a highly promising approach involves training computational systems with substantial datasets that encompass a wide range of observed action variations. Datasets with diverse scenarios, including various environmental conditions, action types, and object interactions, help improve the model's robustness and generalization capabilities.

Most existing works in the field have predominantly focused on utilizing deep neural networks for action prediction tasks. Deep learning models can learn hierarchical representations from data, making them capable of capturing complex spatio-temporal patterns and improving prediction accuracy. However, the success of deep neural networks relies on having a large number of training examples [4]. Therefore, the availability of extensive and diverse action prediction datasets is essential for building robust and effective prediction models.

In conclusion, human action prediction is a crucial area of research with various practical applications, ranging from autonomous systems to video surveillance. By leveraging advanced computational techniques and large datasets, researchers aim to develop accurate and efficient action prediction models capable of forecasting human actions in diverse and complex scenarios.

1.1 From Action Recognition to Action Prediction

In recent years, the domains of human motion analysis and action recognition have garnered significant attention due to their vital implications in various fields such as assisted living [7], surveillance [60], and human-computer/robot interaction [142, 24]. In the era of visual data-driven systems, there is a growing need for these systems to effectively communicate with the world, interact with other entities like robots, humans, and their surrounding environment [111, 146].

The concept of action recognition is central to this discussion. It refers to the ability to identify and classify actions once all the observations related to the action's execution have

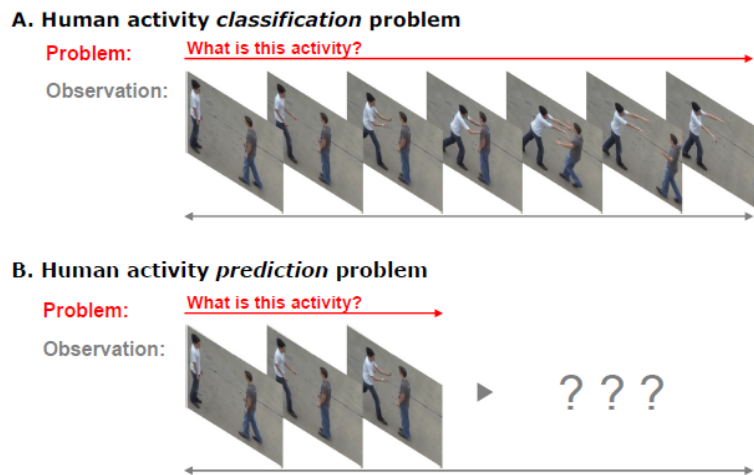


Figure 1.1: Action recognition and action prediction presented on the same sequence [129]

been captured. Over the years, action recognition has undergone extensive study and evolution, employing various techniques ranging from traditional machine learning algorithms to more recent advancements like deep neural networks. While action recognition remains a complex and ongoing area of research, notable progress has been made [48]. Early on, certain datasets captured under controlled laboratory conditions were successfully addressed [47], and there's been remarkable headway in addressing datasets captured in diverse real-world scenarios [164].

The domains of human action recognition and prediction face common challenges that are found in the field of Computer Vision, revolving around the observation and interpretation of human actions. These challenges encompass variations in execution styles, the physiological characteristics of individuals executing the actions, the appearance and clothing of the subjects, and scenarios involving multiple subjects. Furthermore, technical challenges emerge from factors such as varying camera viewpoints, diverse camera setups, changes in lighting conditions, complex backgrounds, and the presence of objects within the scene.

Intriguingly, human action recognition and prediction exhibit shared characteristics. Both tasks involve providing a label or classification based on observations of actions (see Fig. 1.1). The key distinction lies in the temporal aspect: action prediction involves generating a label with only a subset of initial observations, occurring at an earlier stage in the action's progression. The advantage of action prediction is its capacity to reduce response time, as the system can identify an action in its early stages by analyzing only a limited portion of the ongoing activity.

The advantages of action prediction have extensive implications. It allows systems to react swiftly to evolving situations, making it well-suited for scenarios where rapid responses are vital. Take, for example, a surveillance system that must foresee potentially harmful actions in real-time. By predicting actions early, the system can trigger timely alerts or interventions, enhancing security and safety measures.

In summary, the fields of human motion analysis, action recognition, and prediction have gained prominence due to their applications across diverse domains. The ability to recognize and predict actions from visual data holds immense potential for improving human-computer interactions, advancing surveillance systems, and enhancing overall situational awareness. While challenges persist, ongoing research and technological advancements continue to drive progress in these fields, shaping the future of predictive analysis and enhancing the capabilities of intelligent systems.

1.2 Prediction

The rise of Artificial Intelligence (AI) and the widespread availability of affordable visual recording devices have opened up exciting possibilities in predicting human actions. This convergence has led to an explosion in visual data—comprising images and videos—that is readily accessible online. Within this context, action prediction becomes an interesting area for improving recognition and prediction algorithms.

Within the scope of action prediction, the large amount of visual data holds immense promise for algorithmic training and refinement. The sheer volume of available data equips these algorithms with the ability to discern diverse variations in the execution of a single action. By immersing themselves in different contexts, settings, and styles of action execution, these algorithms can develop a comprehensive understanding of the intricate dynamics that define human actions.

Moreover, this plethora of visual data empowers prediction algorithms to anticipate forthcoming actions with heightened accuracy. A thorough analysis of historical action sequences allows these algorithms to unveil concealed patterns and correlations within the data. This predictive capability has extensive implications across domains such as video analysis, surveillance, and human-computer interaction. The power to proactively anticipate actions enables timely decision-making and the ability to take responsive actions.

In the midst of these advancements, challenges arise in the field of action prediction. The first challenge centers around the creation of algorithms capable of making accurate predictions based on limited observations. This task demands the development of complex model structures and training techniques that are quite advanced. The challenge here is to enable these algorithms to understand and make predictions using only a limited view of what's happening. This involves creating smart and sophisticated algorithms that can pick up subtle clues and patterns from the available data.

The second challenge focuses on finding ways to predict what actions will happen next, even when there is a significant amount of missing visual information about these actions. This challenge requires us to come up with innovative methods that can foresee future actions based on the context and previous patterns. It's like solving a puzzle without seeing all the pieces. This challenge encourages us to think creatively about how AI and computer vision can work together to fill in the gaps and predict actions that might not be directly visible.

These challenges serve as catalysts for creativity and progress in the realms of artificial intelligence and computational vision. They push researchers and scientists to explore new ideas, experiment with novel techniques, and develop tools that can push the boundaries of what is possible. The outcome of these efforts isn't just about better predictions; it is also about deepening our understanding of how AI can learn from incomplete information and make sense of complex situations. As a result, these challenges drive innovation, pushing the boundaries of what we can achieve in the exciting intersection of AI and computational vision.

The implications of effective action prediction are vast and transformative. In the context of human-robot collaboration, the ability to predict human actions can lead to safer and more efficient interactions. For instance, industrial robots can optimize their actions based on predictions, ensuring seamless collaboration and accident avoidance. Similarly, robotic agents that assist the elderly can anticipate their needs and even foresee potential dangers, enabling swift and proactive interventions.

In conclusion, the point where AI and the availability of visual data converge represents a critical moment for action prediction. This field has the potential to greatly improve the capabilities of recognition and prediction algorithms. By refining the accuracy and efficiency of action prediction, researchers and developers can advance various applications, spanning from industrial automation to personalized healthcare and more. As algorithms become better at

predicting human actions, they are on the verge of transforming how technology interacts with and enhances human activities.

The subcategories of the action prediction domain that are going to be analyzed in the following sections are:

- (a) **Action Prediction/Early Action Recognition [64]:** This involves predicting the label of the action being executed at the present moment, even when we have access to only a limited amount of observations.
- (b) **Action Anticipation [64]:** Focuses on anticipating the action that will happen in the immediate future, for which there are no direct observations.
- (c) **Next-active-object Prediction [37]:** This category revolves around predicting the object that will become active or engaged in the course of the activity.

Collectively, these subcategories enrich our comprehension of human action prediction. They empower algorithms to offer more accurate forecasts and project actions over time, thereby enhancing their utility across a diverse range of real-world scenarios.

1.3 Our approach

Our approach to effectively tackle the prediction challenge is carefully organized, creating various stages that systematically build upon one another. This framework starts off by addressing segmented actions and subsequently extends to anticipate actions embedded within extensive, unsegmented action sequences. Furthermore, our approach delves into the incorporation of objects, forecasting their role in driving future actions to completion. The following outlines the sequential progression of our methodology:

- **Action prediction:**
 - This phase initially revolves around trimmed single actions, where a fragment of the action is observed, and our aim is to predict the action label before its completion.
 - Progressing from trimmed single actions, our approach then extends to untrimmed action sequences, enabling us to predict action labels within the context of ongoing activities (multiple actions are executed in the progress of an activity).
- **Action anticipation:**
 - Building on the foundation laid in action prediction, we take a step further to anticipate actions that are about to unfold within more extended action sequences.
 - This aspect of our methodology showcases our approach to not only predict ongoing actions but also anticipate actions that are going to be executed in the future, demonstrating the predictive capabilities of our approach.
- **Next active object prediction:**
 - Our approach is not confined to predicting only actions; it also involves predicting the objects that will play an immediate role in the subsequent activities.
 - Going beyond singular object predictions, our framework extends its reach to anticipate all objects that will be utilized until the entire task reaches its completion. This holistic view of object prediction contributes to a better understanding of the task's dynamics.

In essence, our proposed approaches systematically address various levels of prediction difficulties and challenges. By moving from segmented actions to untrimmed sequences, we propose solutions for the problems of action prediction and forecasting as well as for the next active object prediction.

1.4 Motivation

Human action recognition has gathered the attention of the Computer Vision community for an extended period. Up until 2011, the primary focus had solely centered on recognizing human actions with full access to the execution details of these actions. In 2011, Ryoo introduced the concept of human action prediction for anticipating ongoing actions from streaming videos [128]. Ryoo aimed to identify actions with fewer observations, utilizing integral and dynamic histograms.

During that time, the scarcity of easily available training data hindered the development of a system capable of learning and adapting to the challenges inherent in prediction problems. The journey from 2011, marked by the application of traditional machine learning algorithms, to 2019, characterized by the adoption of deep neural networks, signifies an ongoing attempt to address human action prediction and forecasting challenges [145, 116, 43]. Over this span, the domain of action prediction has undergone a transformative evolution [63], progressing from predicting actions within milliseconds [15, 90] to forecasting actions that will happen in several minutes into the future [82, 30, 39, 133, 83]. Consequently, there exists a necessity to discover novel features, representations, and techniques that enable efficient recognition and prediction of ongoing actions in real-time.

Moreover, the implications of successful human action prediction extend to the scope of safety, particularly in the context of accident prevention. By forecasting potential actions and behaviors, computer vision systems could contribute to minimizing risks in various settings. In fields such as autonomous vehicles [91], where the movement of pedestrians and other vehicles is a pivotal concern, accurate prediction capabilities could enable vehicles to chart safer routes, make timely decisions, and avoid potential collisions. The capability to foresee the movements of pedestrians and other vehicles translates into safer navigation, as vehicles can proactively avoid potential hazards and calculate optimal paths. This not only safeguards lives but also builds trust in the transformative potential of self-driving technologies. This translates to improved road safety and a higher degree of confidence in the deployment of autonomous driving technologies.

Beyond the immediate applications, the ability to anticipate and predict future events has a profound impact on human cognition and interaction. Human beings inherently possess the capacity to foresee forthcoming outcomes, a skill that underpins strategic planning, decision-making, and harmonious engagement with the environment. Similarly, prediction plays a critical role in numerous technical domains. For example, the successful integration of home-assisting

robots hinges on their ability to accurately anticipate user actions and intentions. This capability not only enables them to assist users but also fosters a sense of adaptability and responsiveness that aligns with human expectations.

The primary area of interest in recent times has been human-object interactions in egocentric contexts, mainly because of their applicability in systems employing head-mounted cameras. These systems have led to the creation of new datasets, such as Meccano [116], Assembly101 [136], and Ego4D [43]. These datasets have spurred the introduction of fresh challenges aimed at addressing issues like action anticipation, short-term hand-object prediction, long-term activity prediction, and more in this domain in order to make advances in the prediction problems that arise.

In summary, the significance of human action prediction and anticipation resonates on multiple levels. From improving human-robot collaboration and safety to equipping systems with information about events that may unfold in the future, the advancement of this field holds the promise of reshaping our interaction with technology and our environment.

1.5 Organization of the thesis

The remainder of this thesis is structured as follows. Chapter 2 offers an examination of related scientific literature. Subsequent chapters provide insights into our primary methodology and the corresponding experimental evaluations conducted for each segment. Specifically, Chapter 3 delves into our approach for addressing action prediction in trimmed action sequences. Chapters 4 and 5 introduce two innovative techniques that address the challenge of action prediction in untrimmed action sequences. Additionally, Chapter 6 introduces a novel and efficient method tailored for action anticipation within lengthy untrimmed sequences. Concluding this thesis, Chapter 7 furnishes a comprehensive discussion regarding the impact and limitations inherent in the proposed framework. Furthermore, the chapter outlines research avenues for future exploration and articulates the objectives to be pursued.

LITERATURE REVIEW

2.1 Action & Activity Recognition

Action/Activity Recognition sets the thematic base upon which more fine-grained video understanding tasks, such as action detection, early action recognition, and action anticipation/prediction have been defined. In its most challenging form, it comprises the recognition of actions that involve human-object interactions, and action sets with high intra- and inter-class variability. With the advent of deep learning, video action recognition methods have become extremely efficient and effective in modeling short-range dependencies of actions with CNN-centered models [139, 13]. Moreover, the ability to model long-range dependencies of complex actions or long, composite activities has also been considerably improved using memorization layers, such as RNNs and their variants [162, 74], attention mechanisms [153, 5], and temporal frame dependency modeling at multiple time scales [31, 160].

The significant performance gains that have been witnessed in this field have also been fueled by the emergence of large-scale datasets [19, 94], that contain diverse action sets, viewing conditions (egocentric [19, 137, 43] or third-person [28, 71]) and videos in various contexts providing rich, multi-level annotation data and different information modalities. Such datasets enabled the design of multi-modal models that apart from appearance and motion, also exploit audio, gaze-related data, and most importantly language [56, 49]. In the concept of multi-modal action/activity modeling, the visual-linguistic fusion scheme is shown to be extremely effective at

representing the variability of complex actions and activities. This mainly relies on the action-related knowledge that is extracted using the lexical description of the action sequence and transitions, which is presented in the form of a simple text label or rich transcription/captions per action [56]. This information can be further processed using text statistics [125]. Recently, deep learning language models [141, 152], have also been proposed acting as a complimentary information source to the visual representation, expressed with handcrafted [124, 125] or deep learned [81, 68, 8] descriptors.

2.2 Action & Activity Prediction

Vision-based prediction is a rising topic in the field of computer vision [102]. From pedestrian trajectory prediction [118] to pose prediction [89] to accident anticipation [9], prediction has become the focus of several investigations [117]. Ryoo et al. [128] were the first to define the problem of vision-based action prediction as “an inference of unfinished activities given temporally incomplete videos”. Action prediction methods can be on trimmed or untrimmed videos [76].

Action prediction on trimmed videos focuses on recognizing the label of a video given incomplete observations at each point in time [38, 6]. Wang et al. [154] proposed a knowledge distillation framework for early action prediction. The framework employs a teacher model to recognize actions in complete videos and a student model to predict actions in partial videos. A teacher-student learning block is used to transfer knowledge from teacher to student. The framework uses mean square error (MSE) and maximum mean discrepancy (MMD) loss for distilling local and global distribution knowledge. The method in [3] uses a 3D convolutional neural network to extract spatio-temporal features and perform short-term action prediction by using multiple binary classifiers. Li et al. [72] explicitly focuses on mining and utilizing relationships and minor discrepancies within challenging pairs to build a discriminative model. It includes a Hard Instance-Interference Class (HI-IC) bank that dynamically records these challenging pairs during model learning. An adversarial learning scheme is proposed, using a feature generator to create perplexing features for challenging instances based on their similarity to interference classes. Additionally, a class discriminator is designed to distinguish these perplexing features, aiding in mining subtle differences within challenging pairs. This approach enhances the model’s ability to handle challenging pairs that are typically difficult for early activity prediction models. In [25] a probabilistic approach to predict everyday actions is presented by using motion trajectory

prediction and taking into account the objects and their affordances.

Action prediction in untrimmed videos is performed on action sequences and its goal is to recognize early (a) all the action labels that are present in a video and (b) their anticipated duration [58, 100, 93]. Since action boundaries are not known, Liu et al.[76] addresses the problem of online action prediction in streaming 3D skeleton sequences, where the goal is to recognize an ongoing activity based on only a partial observation. To tackle this, the authors introduce a dilated convolutional network that models motion dynamics in the temporal dimension using a sliding window approach. To handle temporal scale variations, a novel window scale selection method is proposed, allowing the network to focus on the performed part of the action while suppressing interference from previous actions. An activation-sharing scheme is also presented to improve computational efficiency by handling overlapping computations. Additionally, a hierarchy of dilated tree convolutions is designed to enhance performance by learning structured semantic representations over skeleton joints. The method presented in [28] predicted the future actions in two ways: one utilizing a Convolutional Neural Network (CNN) and the other a Recurrent Neural Network (RNN), to forecast a substantial sequence of future actions and their durations. These methods are trained to predict future video labels based on previously observed content. The results demonstrate that both approaches can accurately anticipate future actions, even in lengthy videos containing a wide range of different actions, and they exhibit robustness in handling noisy or erroneous input information. For this reason, Ke et al. [57] predicted actions in an one-shot fashion. The method presented in [40] recognizes the actions of the ongoing video along with their duration by inferring information about the verbs and the objects that are present in the scene through aligning video segments.

Action prediction aims to forecast the label of an action based on limited/partial observations. The majority of the proposed methods that tackle this problem consider (first person) egocentric videos [121, 163, 158, 134, 2], mainly due to the availability of large amounts of relevant video data and annotations [43, 18, 116]. An advantage of the work proposed by Furnari et al. [36, 35] is the ability to make predictions not only in first-person but also in third-person videos. Their work focuses on making predictions using multiple modalities such as RGB frames, optical flow and object-based features. Their architecture uses one Long Short-Term Memory (LSTM) for encoding the past time steps while the second LSTM makes predictions about the future. Wu et al. [155] opted to solve the problem of activity prediction by exploring spatio-temporal relations

between humans and objects. They used a graph-based neural network to encode the spatial relations between video entities at different time-scales.

2.3 Action Anticipation/Forecasting

This is defined as the task of predicting the class(es) of one or more future actions for which no observations are available at the decision time [64, 71]. The task has been well-explored for actions of various complexity that range from simple motion primitives of a single human action or a human-object interaction [64, 53] to long, composite, procedural or unconstrained activities [133, 83]. Anticipating the near-future actions is performed towards a limited set or even thousands of action categories [19, 136]. Forecasting of next actions is performed at “anticipation time” in video that can be set at variable time horizons ranging from short- to long-term predictions. Many existing approaches fix this important task parameter to 1 second prior to the start of the action of interest [77, 41], while others explore the predictability of actions for several seconds [113, 36, 79, 28, 59]. The problem was initially introduced in third-person videos [53, 28], but it has recently gained significant popularity in first-person (egocentric) videos [19, 43], too.

In [41], Video Transformers are proposed to accurately anticipate future actions. Without supervision the method learns to focus on the image areas where the hands and objects appear, while attends the most relevant frames for the prediction of the next action. Rodin et al. [122] tackles the problem of anticipation in untrimmed videos in an attempt to generalize and deal with unconstrained conditions in real world scenarios. The prominent method of Furnari et al. [35] explored the problem of action anticipation using “rolling-unrolling” LSTMs in order to summarize past actions and make predictions for the verb, noun and action of the next segment for multiple anticipation times. In [134] a multi-scale temporal model is proposed so that the past actions are aggregated for the future actions to be iteratively predicted. This framework performs predictions for the next action with an anticipation time of 1 second and is also capable of performing dense anticipation considering a large number of anticipated action classes. Our work complies with both methodologies so that the verb, noun and action predictions are made in the range of [0.25, 2] seconds with a step of 0.25 seconds.

Natural language processing (NLP) initially gained popularity in the cooking domain since recipes naturally contain a large variety of texts with instructions on food preparation. These large texts of instructions have attracted the interest for predictions of the next unobserved

steps of the recipe in natural language in the form of sentences. Sener et al. [135] created a hierarchical model for learning multi-step procedures of recipe datasets with text and visual context. Their zero-shot anticipation framework is able to transfer knowledge from large-scale text corpora to the visual domain for the prediction of coherent and plausible recipe instructions. The same authors improved their framework by integrating a temporal segment proposal method to the video encoder and additional losses at the recipe encoder to improve convergence [133]. By comparing to recipe generation networks they showed that this method can perform better even for unseen recipes and dishes. Contrary to methods [135, 133] that exploit text to provide information to the visual domain, Mahmud et al. [83] proposes a two-step approach where information on the visual spatio-temporal context of the observed actions and the linguistic labels of the anticipated actions along with scene context are incorporated for caption prediction. Text and/or captions of the observed actions are not utilized.

2.4 Next Active Object Prediction

The first approach to tackle the problem of next-active-object was Furnari et al. [37]. A sliding window was utilized in conjunction with an object detector in order to model each tracked trajectory and classify it as passive or active using random forests. The paper argues that the next-active-object can be distinguished from its frames immediately before it turns active. One very interesting characteristic of the method they propose is its ability to generalise to unseen object classes. However, their experiments show a loss of accuracy when dealing with the unseen object classes thus proposing to train the method with the object classes that will be present in the test set for better results.

The work of Dessalene et al. [22] employs graphs to predict the partially observed action and produce Contact Anticipation Maps which provide pixel-wise information of the anticipated time-to-contact involving one hand, either the left or the right. Also, they perform next-active-object segmentation by localizing candidate next active objects. These localizations are evaluated with the calculation of the Intersection over Union (IoU) value of the bounding boxes produced from the Faster-RCNN model. This work predicts the hand-object time-to-contact in egocentric videos but this does imply that this can be the next-active-object or that this object will be used immediately. Also, this is trained on annotated object classes of the dataset which implies that it cannot generalize to unseen object classes.

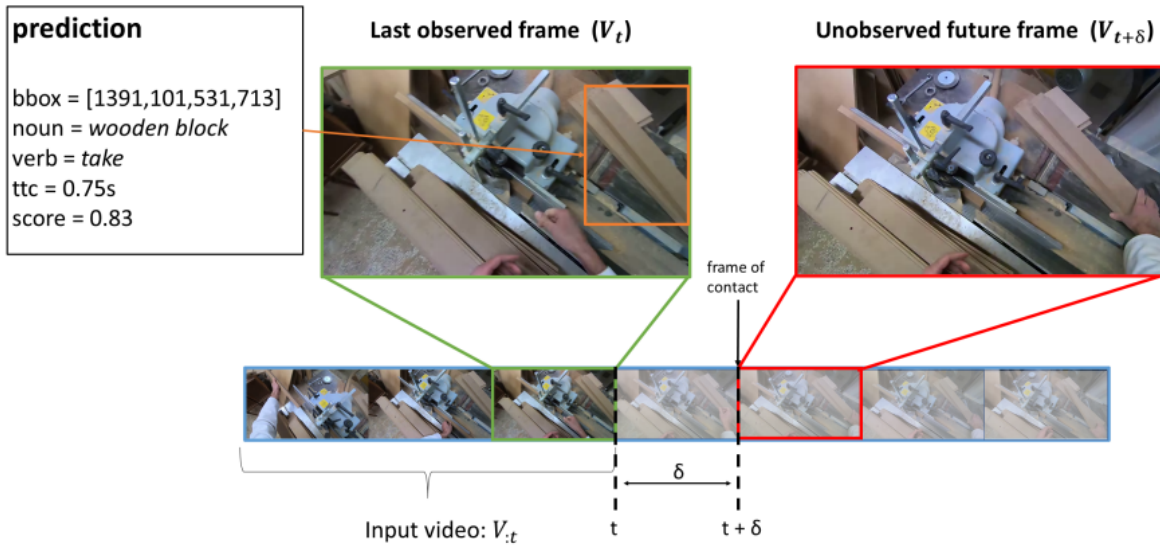


Figure 2.1: Task of anticipating short-term object interactions. Models analyze video sequence V up to timestamp t , making predictions about the bounding box and class of upcoming active objects, the verb characterizing the impending interaction, a numeric value denoting when the interaction occurs ($t + \delta$), and a confidence score. Here, δ signifies the time gap between the final observable frame V_t and the frame of interaction at time ($t + \delta$). Image from [115].

Jiang et al. [55] highlights the importance of predicting human intentions, particularly in human-robot interaction and rehabilitation robots. The study addresses the prediction of short-term next-active-objects in egocentric images, which represent objects humans will interact with soon, reflecting their intentions. Most existing methods focus on object-related cues, like appearance changes and trajectory shapes, for prediction. The paper introduces a deep neural network model that incorporates human-centered cues, specifically visual attention and hand positions, to enhance prediction. This involves constructing probability maps for attention and hand positions, leading to a probability distribution for the next-active-object.

In recent years, the nature of next-active-object prediction has evolved into a next-active-object detection task, encompassing the prediction of both the next-active-object label and its potential spatial location. The task is defined following the framework proposed in [43]. Given a video sequence V and a specific time point t , the objective involves processing the video up to time t . The goal is to generate a set of predictions regarding future object interactions that will occur after a specified time interval δ . Each prediction consists of the following components:

- A bounding box indicating the location of the next-active object that will be interacted with

in the future.

- A noun label describing the class of the identified object (e.g., "knife").
- A verb label describing the nature of the interaction that will transpire in the future (e.g., "take").
- A numerical value representing the "time to contact," which signifies the time in seconds between the current timestamp and the commencement of the interaction (e.g., 0.75 seconds).
- A confidence score is assigned to rank the predictions for assessment purposes.

The works of [43] and [116] proposed the Ego4D and Meccano datasets which have ground truth annotations that can accommodate the next-active-object detection task. They also provide baselines for the solution of this specific task.

Liu et al. [77] introduced a method for predicting future hand-object interactions in egocentric videos. Instead of predicting action labels or pixels, the approach directly forecasts the trajectory of hand motion and future contact points on the next active object (interaction hotspots). This compact representation offers a clear description of forthcoming interactions. To address this task, the authors develop an automated method to gather trajectory and hotspot labels using extensive data. They then utilize this data to train an Object-Centric Transformer (OCT) model for prediction. This model employs the self-attention mechanism of Transformers for reasoning about hand and object interactions. OCT also incorporates a probabilistic framework to sample future trajectories and hotspots, effectively handling prediction uncertainty.

In [109] the authors propose the TransFusion framework which is a multimodal architecture based on transformers. It utilizes language to summarize the action context, leveraging pre-trained image captioning and vision-language models to extract context from previous video frames. This context, along with the next video frame, is processed by a multimodal fusion module to predict the next object interaction. The approach enhances end-to-end learning efficiency and benefits from large pre-trained language models for commonsense and generalization.

Ragusa et al. [115] focus on short-term object interaction anticipation in the egocentric perspective. They propose an innovative end-to-end architecture named StillFast. This approach concurrently processes a static image and a video, detecting and localizing next-active objects,

forecasting the action that describes the future interaction, and determining the interaction’s timing. Through experiments on the extensive egocentric dataset EGO4D, the method demonstrates superior performance compared to existing approaches.

2.5 Temporal Alignment

The temporal alignment of sequences is a problem that has been explored for many years and remains of interest until today. A classical approach is the Dynamic Time Warping [130] algorithm which is capable of aligning segmented sequences by finding the minimum-cost warping path between them. The warping path is calculated upon the distance matrix which contains all the frame-wise distances between the two sequences to be aligned. The DTW score is based on the summation of all path-related values in the distance matrix. The DTW algorithm poses boundary constraints on the warping path which means that the sequences to be aligned must start and end at known frames i.e. the first frame of the first sequence will be matched to the first frame of the second sequence. DTW has been used in a variety of problems such as action cosegmentation [105], representation learning [45], etc.

The boundary constraints of DTW have been relaxed by the work of Tormene et al. [143] who proposed the Open-End DTW (OE-DTW) algorithm. The OE-DTW variant is capable of aligning sequences that have a known common start point but unknown endpoints. This relaxation of the endpoint constraint is useful when the sequences to be matched are partially observed or when other actions appear after the end of the sequence. The OE-DTW score is provided by the summation of all values of the minimum-cost alignment path. The difference to the DTW algorithm is that the alignment path that starts at the top-left point of the distance matrix should not necessarily end at the bottom-right cell of that matrix, thus permitting a certain sequence to match with a part of a reference one. OE-DTW has been used to compare motion curves for the rehabilitation of post-stroke patients [131] as well as for the evaluation of the user’s motion in visual observations of humans with Kinect [161].

Tormene et al. [143] also proposed the Open-Begin-End (OBE-DTW) [143] that aligns two unsegmented sequences, i.e., two sequences of unknown starting and ending points. The matching path defined by OBE-DTW does not necessarily have to start and end at the top-left and bottom-right cells of the distance matrix. OBE-DTW has been used in many contexts for unsegmented sequence alignment e.g., for the problem of classifying motion from depth cameras [61].

While the DTW algorithm considers the minimum-cost alignment path of the sequences, the Soft Dynamic Time Warping (S-DTW) [17] variant considers the soft-minimum of the distribution of all costs spanned by all possible alignments between two segmented sequences. This alignment score contains the summation of all path-based values. The S-DTW algorithm has been used by [46] as temporal alignment loss for training a neural network to learn better video representations. The differentiable alignment of S-DTW has also been used by Chang et al. [14] for the alignment and segmentation of actions by using the videos and the transcripts of the actions.

Segmental DTW [108] seeks for the minimum-cost sub-sequence alignment of pairs of unsegmented inputs. Segmental DTW decomposes the distance matrix in sets of overlapping areas and finds the local end-to-end alignments in these areas resulting in sub-sequence matching. Segmental DTW has been used in the context of action co-segmentation [103] in motion-capture data or video between pairs of actions for the detecting of commonalities of varying length, different actors, etc.

The Ordered Temporal Alignment Module (OTAM) [12] aligns segmented sequences of fixed length by using the soft-minimum operator and calculating all possible path-based alignments. The alignment score is given by aligning the sequences end-to-end using S-DTW, while the alignment path is retrieved by an OBE-DTW approximation. Cao et al. [12] used the OTAM alignment for few-shot video classification of fixed-length trimmed videos.

Finally, the Drop Dynamic Time Warping (Drop-DTW) [26] algorithm is a variant of DTW based on images where outliers are dropped during the alignment of sequences. Differently from OBE-DTW where the unrelated parts can be at the prefix or the suffix of an action, this DTW approximation is very useful in cases where the sequences to be aligned have unrelated parts anywhere inside the sequences. By eliminating all the irrelevant parts Drop-DTW results in more meaningful alignments.

2.6 Contribution & Open Challenges

The works discussed in Section 2 offer solutions for addressing the problems of action prediction, action forecasting, and prediction of the next active object. It is important to note that these problems are still far from being fully resolved. For this reason, there are still open challenges that the Computer Vision community addresses. The first challenge is improving the prediction accuracy for the problem of action prediction on segmented and unsegmented data. The limited

observations of partially executed actions impose a great challenge to our ability to predict the future with high accuracy. One open challenge associated with action anticipation is the ability to predict not only the next unobserved action but also the following actions after that (called long anticipation). We need the ability to learn from past actions and find correlations that will help us predict further into the future. Finally, an open challenge in the domain of prediction is the ability to foresee the label of the object that is going to be used in the immediate future. The objects that are present in the scene can provide important information about the actions that are going to take place in the future. So, our ability to predict the next object/s will enhance the predictive capability of the algorithms.

In the scope of this thesis, we address some of the aforementioned challenges. We propose algorithms for the prediction of ongoing actions in segmented and unsegmented action sequences in human-object interaction scenarios using a Dynamic Time Warping (DTW)-based framework, demonstrating improved accuracy over existing methods. We also proposed two new DTW-based alignment algorithms designed to align incomplete action sequences with complete ones, offering better alignment for unsegmented actions for the solution of the action prediction problem. Next, we addressed the problem of the next-active object prediction with the use of graphs and we proposed two new challenges which are the prediction of all the objects that will be used for the completion of an activity (multiple next-active objects prediction) and the prediction of the time that the next-active object will be used. In human-object interaction scenarios, all the information that we can retrieve from objects provides information that helps increase the predictive capabilities of the prediction algorithms. Finally, we propose an LSTM-based network that combines visual and linguistic information to enhance action anticipation, resulting in improved accuracy compared to other methods. This method can exploit information about all past actions in the compact way of linguistic features extracted from action labels instead of having to store and process the visual features of all past actions. This approach showcases state-of-art accuracy on the action anticipation problem while combining cost-effective linguistic features of all past actions with the visual features of the recent past.

DTW-BASED ACTION PREDICTION IN TRIMMED SEQUENCES

Action prediction is the art of making educated guesses about ongoing actions. Think of it like guessing what's happening in a scene while it is still unfolding. This skill is like having a crystal ball for quick responses and smart planning ahead. Our research zooms in on understanding actions when people are interacting with objects. We've crafted a method to track how people and objects are moving over time. It is like capturing a story of what is going on in an action that hasn't finished yet. Then, we take this ongoing action and match it up against actions we have in our database. This matching process is done using an alignment algorithm called Dynamic Time Warping (DTW), which helps us find the best fit between the action in progress and actions we know well.

Putting our method to the test, we examine three sets of data that are widely used in this field. The results of our experiments reveal something important: when we combine how people and objects are behaving (instead of using just the people or just the objects), our action predictions get a lot better. Our approach proves to be more accurate in predicting actions compared to other techniques.

We narrow our focus to a specific puzzle: predicting actions where people are doing things together with objects, using only what we can see. In this context, action prediction means figuring out what is going on even if the action isn't complete yet [128]. To crack this code, we use shortened video clips (trimmed video recordings) as input data, from which we can extract time

series of 3D skeletal data using various methods, such as those demonstrated by Qammaz et al. (2021) [112]. For the above reasons, in the current chapter we use the terms "video recordings" and "skeletal data" interchangeably.

The amount of action we can see depends on an observation ratio. This ratio can be any number between a tiny bit (more than zero) and the whole thing (100%). If we can see the entire action (100% observation ratio), it is like we've watched the entire story. In this case, predicting the action becomes as easy as knowing the full type of action that's happening.

3.1 Problem description

We assume that a video recording of an action (full or incomplete) can be represented as an N -dimensional time series through some appropriate feature extraction mechanism (see Sec. 3.2).

Let Q be such a time series representation of an incomplete action. We are interested in inferring the unknown action label $L(Q)$ of Q . We also consider a set of C time series P^i , $1 \leq i \leq C$, corresponding to known, prototype action executions with labels $L(P^i)$. Our approach compares Q with each of the time series P^i through Dynamic Time Warping-based alignment (DTW). Let $DTW(X, Y)$ denote the DTW alignment cost of time series X and Y . Then, action prediction can be formulated as:

$$(3.1) \quad L(Q) = L\left(\arg \min_{P^i, 1 \leq i \leq C} \left(DTW(Q, P^i)\right)\right).$$

Essentially, the proposed method predicts that the action label of Q is that of the prototype action P^i which can be aligned with Q at a minimum DTW-based alignment cost.

3.2 Feature Extraction

Given a video, we represent each of its frames as a multidimensional vector of action-related features. Depending on the employed scenario, such features encode the human body pose, the class and the pose of the involved object, or both. Section 3.5 presents different sets of extracted features for the standard benchmark datasets employed in this work.

3.3 DTW-based Time Series Alignment

Dynamic Time Warping (DTW): Let X and Y be two time series with $X = (x_1, \dots, x_l) \in \mathbb{R}^{n \times l}$ and $Y = (y_1, \dots, y_m) \in \mathbb{R}^{n \times m}$. We define the distance matrix $D(X, Y) = [d(x_i, y_j)]_{ij} \in \mathbb{R}^{l \times m}$, where $d(x, y)$ is the Euclidean distance between x and y . We also define Π , the set of all path-based alignments of X and Y , connecting the upper-left to the lower-right of the matrix D . Finally, let $\pi \in \Pi$ be one of all those alignments. The inner product $\langle \pi, D(X, Y) \rangle$ yields the alignment score associated with π .

DTW [130] is a dynamic programming algorithm that estimates the minimum-cost alignment of two time series. On the basis of the above notation, this is $DTW(X, Y) = \min_{\pi \in \Pi} D(X, Y)$. For the partial alignment, a variant of the original DTW [130] is employed, called open-end DTW [143].

Open-End Dynamic Time Warping (OE-DTW): The alignments can end at any point at the last column of the D matrix. The alignment score is normalized by the number of diagonal steps of the calculated optimal alignment path. Open-end DTW is defined as:

$$(3.2) \quad DTW_{oe}(X, Y) = \min_{j=1, \dots, m} DTW(X, Y_j).$$

Soft-DTW: Soft-DTW [17] builds upon the original and popular dynamic time warping (DTW) measure and considers a generalized soft minimum operator applied to the distribution of all costs spanned by all possible alignments between two time series of variable size. Given the following generalized minimum operator, subject to a smoothing parameter $\gamma \geq 0$,

$$(3.3) \quad \min \gamma(\pi_1, \dots, \pi_k) = \begin{cases} \min_{i \leq k} \pi_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^k e^{\pi_i/\gamma} & \gamma > 0, \end{cases}$$

the soft-DTW score is defined as:

$$(3.4) \quad SDTW_{\gamma}(X, Y) = \min^{\gamma} \{ \langle \pi, D(X, Y) \rangle, \pi \in \Pi \}.$$

The original DTW score is obtained by setting $\gamma = 0$.

Global Alignment Kernel: Global Alignment Kernel (GAK) [16] measures the similarity between two multidimensional time series X, Y . On top of $D(X, Y)$, the GAK is computed as:

$$(3.5) \quad GAK_{\gamma}(X, Y) = \sum_{\pi \in \Pi} \exp\left(\frac{-\langle \pi, D(X, Y) \rangle}{\gamma}\right).$$

In comparison to DTW, in order to find the alignment score of two time-series, instead of using the operators (min, +) on the $\langle \pi, D(x, y) \rangle$ GAK uses the (+,X) operators. According to [16], the GAK considers the full spectrum of the $\langle \pi, D(X, Y) \rangle, \pi \in \Pi$, while the DTW distance considers only the minimum score. In comparison to DTW and GAK, in order to find the alignment score of two time-series, instead of using the operators (min, +) (DTW - minimum score) or (+,X) (GAK - full spectrum of the $\langle \pi, D(X, Y) \rangle, \pi \in \Pi$), Soft-DTW unifies them with the use of the soft-minimum operator. More specifically, when $\gamma = 0$ we retrieve the DTW score while for $\gamma > 0$ we recover $SDTW_{\gamma} = -\gamma \log GAK_{\gamma}$.

A variety of alignment algorithms have been proposed in the literature for the alignment of time-series such as Connectionist Temporal Classification [44], Needleman–Wunsch algorithm [99], etc. DTW and its variants have been used extensively in recent years for alignment in works such as learning representations [46, 45] and view-invariant representations [159], action alignment & segmentation [14], video-text representation learning [62], etc. We opted to center our research around the DTW algorithm because it closely matched the focus of our study, and the community continues to advance this algorithm by incorporating it into neural networks.

Alignment Paths & Distance Matrix: While someone may argue that the alignment path is always towards the diagonal of the distance matrix this assumption is not always the case. This holds true for segmented actions that have been observed in their entirety and no unnecessary movements (such as e.g. standing still for a large amount of time) occur anywhere in the action. For all the other cases (such as ex. unsegmented actions, partially observed actions, etc.) the alignment path can occur anywhere in the distance matrix.

3.4 Early Fusion of Human and Object Representations

The aforementioned variants of DTW operate on the distance matrix $D(X, Y)$ of time series X and Y , which, for notational convenience, will be denoted with D . In the human-object interaction scenario we are considering, this distance matrix is defined as follows. First, we construct a distance matrix D_H which results from the frame-wise comparison of the part of

the representation that contains the information regarding the human. We also construct an analogous distance matrix D_O which results from the frame-wise comparison of the part of the representation that contains the information regarding the object with which the human interacts. This study is only evaluated using only rigid objects. Deformable objects have not been examined in the scope of this work. Then, the distance matrix D on which DTW operates is defined as:

$$(3.6) \quad D = \alpha_H D_H + \alpha_O D_O.$$

In the above equation, α_H and α_O are weighting factors that may be dataset-dependent, but have been defined experimentally and commonly for all employed datasets. If the two compared actions involve objects of the same class, then $\alpha_H = \alpha_O = 0.5$. If the two compared actions involve objects of different classes, then $\alpha_H = 0.7$ and $\alpha_O = 0.3$. The intuition behind this choice is that the same action can be performed by using different objects (e.g., reach, move, etc). Therefore, actions can still be compared, but with giving emphasis on the part of the representation concerning the humans rather than the objects. If no objects are present in the scene, then $\alpha_H = 1$ and $\alpha_O = 0$. Finally, in the case that one of the actions involve an object and the other does not, $\alpha_H = 3$ and $\alpha_O = 0$. Essentially, the two actions are again compared on the basis of the human performance, but the mismatch on the presence of objects is penalized by a large α_H value.

In the observed scene, several objects may be present. From those, we consider the one that is closest to and/or manipulated by the actor. One limitation of this choice is that we cannot take into account actions involving more than one object. However, ongoing research beyond the scope of this paper indicates that the extension of our approach towards handling more than one manipulated objects is feasible. Another limitation and future extension of our work lies in the need to know the start frame of an action. The generalization of our approach towards handling unsegmented actions is another topic of ongoing research.

3.5 Experimental Evaluation

3.5.1 Datasets & feature representations:

The proposed framework is evaluated on 3 standard datasets with different characteristics.

MHAD Dataset [101]: Contains 11 actions (jumping in place, jumping jacks, bending, punching,

waving one hand, waving two hands, clapping, throwing a ball, sit down and stand up, sit down, stand up) performed by 12 subjects. The majority of the actions do not involve objects (with the exception of the action “throwing a ball”). The database provides motion capture data containing the 3D positions of 43 LED markers, which have been processed to obtain 3D skeletal data of 30 joints. The standard evaluation split is used as in [101].

Features: The MHAD [101] dataset contains the 3D positions of skeletal joints. Based on these 3D positions, we build a human body representation as proposed in [120] and also used in [104, 84]. Specifically, a human pose is represented as a $30 + 30 + 4 = 64$ D vector. The first 30 dimensions encode angles of selected body parts with respect to a body-centered coordinate system. The next 30 dimensions encode the same angles in a camera-centered coordinate system. The representation is augmented with the 4 angles between the fore- and the back-arms as well as the angles between the upper- and lower legs.

MSR Daily Activity 3D Dataset [149]: Consists of 16 actions (drinking, eating, reading a book, speaking on cellphone, writing on paper, using a laptop, using a vacuum cleaner, cheering up, sitting still, tossing paper, playing a game, lie down on the sofa, walking, playing the guitar, standing up and sitting down) performed by 10 subjects. Every subject performs each action twice, once sitting on a sofa and once standing. We followed the experimental settings of [157, 119].

Features: The dataset contains the 3D skeletal joint positions for all the human joints. We consider only the 9 upper body joints due to the fact that the data for the lower body are quite noisy. The 3D upper skeletal joint positions are calculated to be invariant to the body center. The invariant 3D joint positions are concatenated with the 3D joint angles. The 3D joint angles are represented as a 30D vector. The 30 dimensions encode angles of selected body parts with respect to a body-centered coordinate system [120] but we are taking into account only the angles that correspond to the upper body. For the objects in this dataset, we employed the YoloV4 [11] trained on ImageNet in order to acquire fast and accurate labels and 2D positions of the objects in the scene. We densely annotated the training part of the MSR-Daily dataset and re-trained the YoloV4 [11] on the MSR Daily Activities dataset. The invariant 3D upper skeletal joint positions are 27D and the 3D joint angles that correspond to the upper body are 18D. The positions of the objects are 2D. Thus, each frame of a video is represented as a $27 + 18 + 2 = 47$ D vector. The joint

and object positions are divided with the torso-head distance.

CAD-120 Dataset [66]: Contains activities performed by 4 subjects, which can be subdivided into 10 sub-activities. The subjects perform the activities with different objects. Activities are observed from different viewpoints. The sub-activity labels are: reach, move, pour, eat, drink, open, place, close, clean, null. We are experimenting on the sub-activity labels using the standard 4-fold cross-validation as in [66]

Features: We used a set of features based on [66]. We employ ground truth annotations for the semantic labels, bounding box information, and 3D poses of the objects, as well as labels and 3D positions for tracked human body joints. To represent the trajectory of each object and body joint within a video, we encode their relative 3D positions in each frame, using the torso joint as a reference point. We divide the 3d joint and object positions with the torso-head distance. Specifically, to represent human motion we use the location of each of 8 joints (24D), the distance moved by each joint (8D) and the displacement of each joint (8D). For representing objects we used their 3D centroid location, the distance moved by the object’s centroid (1D), the displacement of the object’s centroid (1D) and the distance between each joint location and the object centroid (8D). In total, a frame of a sequence is represented as a 53D vector.

3.5.2 Performance metrics:

We measure the action prediction accuracy as a function of the observation ratio, i.e., the percentage of the part of the action that has been observed and compared to action prototypes. In our experiments, the observation ratio ranged from 10% to 100% in steps of 10%.

3.5.3 Evaluation of DTW variants:

We evaluated the three DTW variants presented in Sec. 3.3 with respect to their action prediction accuracy on all three datasets. During testing, every query sequence is compared to all sequences in the training set. We employed a publicly available [50] implementation of DTW_{oe} and the implementations of the DTW variants that reside in the Tslern toolkit [127]. The parameter γ was experimentally set equal to 0.1 for the (MHAD, MSR) datasets and to 0.01 for the CAD120 dataset. As it can be observed in Fig. 3.1, DTW_{oe} outperforms SDTW by a great margin in the MHAD and CAD120 datasets. In turn, SDTW clearly outperforms GAK. This holds true for all

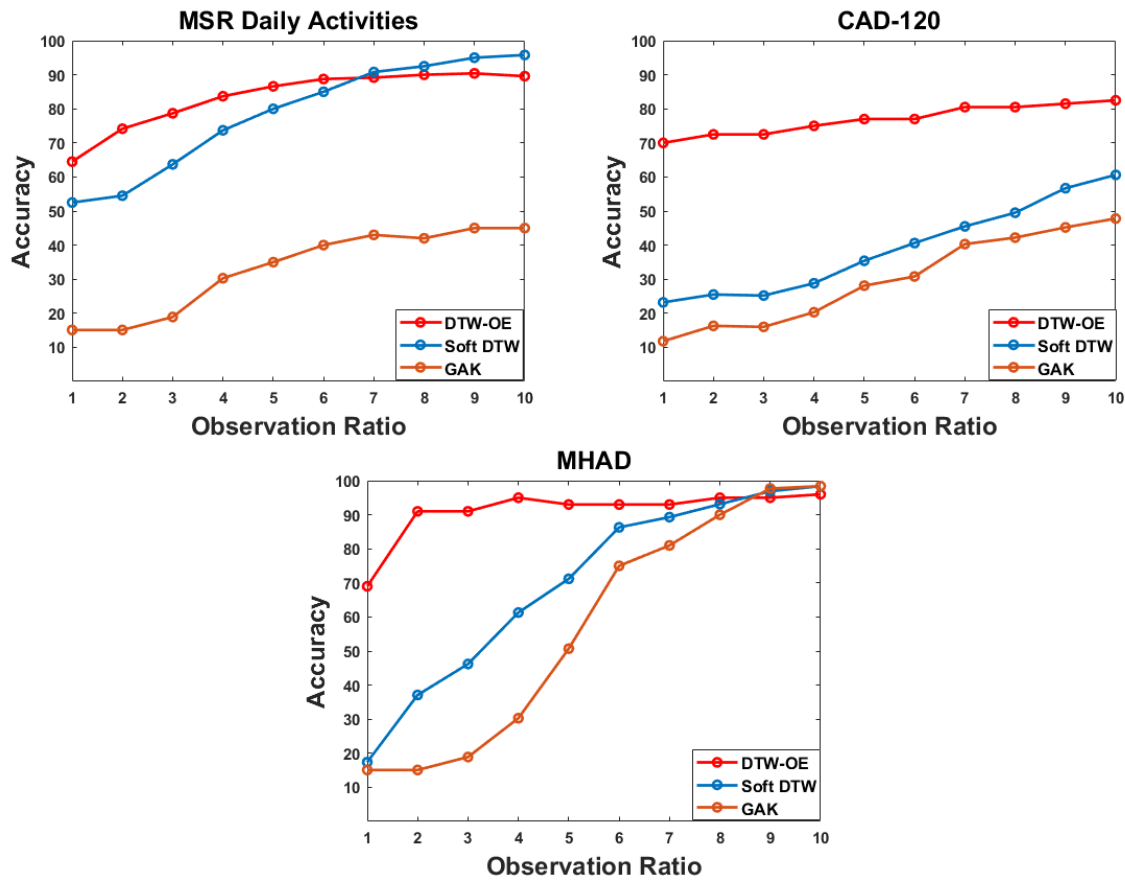


Figure 3.1: Comparison of DTW_{oe} , SDTW and GAK on the MSR (upper left) CAD-120 (upper right) and MHAD (down) datasets.

three datasets, regardless of whether they involve humans in interaction with objects (MSR, CAD120) or not (MHAD). Moreover, the superiority of DTW over the rest two variants is dominant especially in lower observation ratios. This shows the potential of the method for accurate and early action prediction.

3.5.4 Evaluation of alternative representations:

We evaluated the impact of action representations on action prediction. More specifically, we investigated three different experimental conditions, (a) representations that involve only the joints of the human actor (b) representations that involve only the class and the motion of the involved objects and (c) their early fusion, as presented in Sec. 3.4. Figure 3.2 (left, middle)

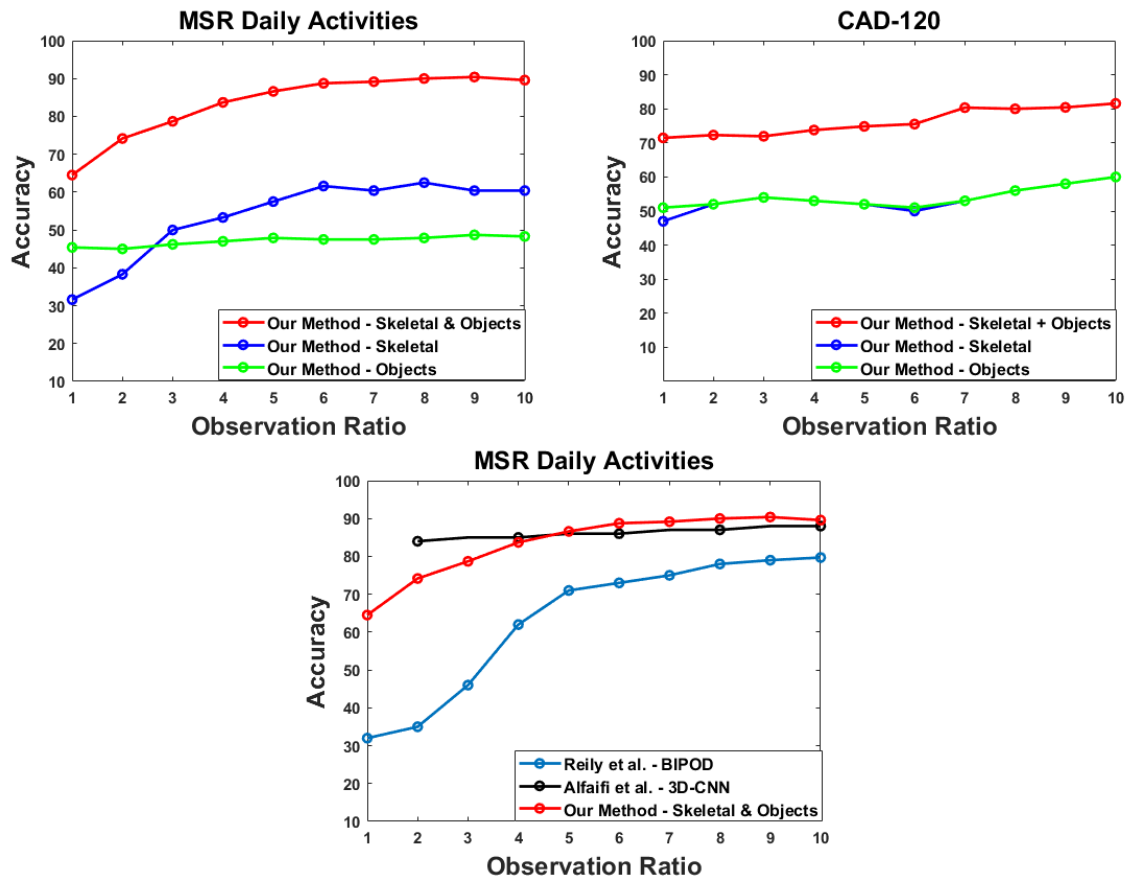


Figure 3.2: Action prediction results for different representations. (Upper left) MSR Daily Activities Dataset, (Upper right) CAD-120 Dataset. (Down) Action prediction accuracy of our method in comparison to state of the art methods on the MSR Daily Activities Dataset.

shows the results we obtained in the MSR and the CAD120 datasets¹, respectively. As it can be verified, the early fusion of the actor and object representations outperforms any of the individual representations in predictive power, by a vast margin (from a minimum of 10% to a maximum of 40%).

3.5.5 Comparison to the state of the art:

Figure 3.2 (right) presents a comparison of our approach to other competitive methods on the MSR dataset. Specifically, we are comparing to the work of Reily et al. [119] and to that of Alfaifi et al. [3]. As it can be observed, we outperform [119] at all observation ratios and [3] for all observation ratios greater than 40%.

¹MHAD is not included in this investigation as the vast majority of its actions do not involve human-object interactions.

To the best of our knowledge, there are no reported quantitative results for action prediction on the CAD120 dataset. We only report action classification results from the very recent method of Mavroudi et al. [92] that achieves an action classification accuracy of 90.4% which can be compared to the action prediction results of our method in the case of an observation ratio of 100%.

Similarly, there are no reported quantitative results for action prediction on the MHAD dataset. For action classification, the very recent method of Qin et al. [114] achieves an accuracy of 100%, compared to the action classification accuracy of 96% of our method, for an observation ratio of 100%. Interestingly, an action prediction accuracy of more than 90% is achieved by our method, even when a small portion of the activity has been observed (observation ratio of 20%).

3.6 Summary

Our approach to tackling the challenge of predicting human-object interactions revolves around the innovative concept of aligning fused, frame-based action representations of both human subjects and objects. We've ingeniously cast the complex task of action prediction into the realm of aligning these multidimensional time-series representations. Through this framework, we seamlessly integrate the dynamic interplay between human actions and object interactions, creating a holistic perspective that encapsulates the intricacies of the scenario.

This alignment process, pivotal to our approach, is executed through a Dynamic Time Warping (DTW)-based methodology. This technique effectively captures the temporal alignment and assesses the similarity between these fused action representations. Our study meticulously evaluates three distinct DTW variants within the context of predicting human-object interactions. To ensure the robustness and applicability of our framework, we conducted extensive evaluations on three widely recognized datasets, providing a comprehensive and reliable assessment of its performance.

In addition to scrutinizing the DTW variants, we dedicated a substantial effort to quantitatively analyze the significance of fusing human-based and object-based action representations. This investigation sheds light on the collaborative influence of these representations in the context of prediction accuracy. The results we obtained through rigorous experimentation present a compelling argument: the DTW_{oe} variant emerges as the superior performer among all tested

variations. Furthermore, the fusion of these representations has a remarkable impact on enhancing the predictive potential of our framework.

Our approach surpasses conventional methods and achieves notable success when compared to recently published competitive action prediction techniques. This success is not only attributed to the effectiveness of DTW_{oe} but also underscores the pivotal role of fusing human and object action representations. In essence, our work represents a significant leap forward in predicting human-object interactions, offering a powerful framework that harnesses both temporal alignment and comprehensive representation fusion to increase accuracy in predictive modeling.

ACTION PREDICTION IN TRIMMED/UNTRIMMED SEQUENCES

The task of aligning actions in videos has been addressed using algorithms like Dynamic Time Warping (DTW) and Soft Dynamic Time Warping (S-DTW). These methods excel at matching actions that are divided into segments. However, they face a difficulty when aligning actions that are continuous and unsegmented, occurring between other actions. To tackle this challenge, we employ two variations of DTW: Open-End DTW (OE-DTW) and Open-Begin-End DTW (OBE-DTW). OE-DTW aligns actions with known starting points but unknown endpoints, while OBE-DTW deals with actions that are completely unsegmented, having unknown start and end points.

In this study, we harness the strengths of S-DTW, OE-DTW, and OBE-DTW to introduce two novel DTW versions: Open-End Soft DTW (OE-S-DTW) and Open-Begin-End Soft DTW (OBE-S-DTW). These variants combine the flexibility of boundary constraints from S-DTW with the capacity of OE-DTW and OBE-DTW to handle unsegmented actions. This fusion results in more precise and differentiable alignment of ongoing actions within continuous, unsegmented videos.

To validate our approach, we subject our new algorithms to action prediction tasks using well-established datasets such as MHAD, MHAD101-v/-s, MSR Daily Activities, and CAD-120. The outcomes of our experiments demonstrate that the proposed algorithms outperform existing methods for video alignment.

In practical scenarios, cameras capture actions from diverse perspectives, speeds, and performers. Temporal video alignment algorithms are employed to synchronize these actions despite such variations. These algorithms find applications in areas like assessing action quality, co-segmenting actions, and retrieving specific frames.

The core of these alignment algorithms involves representing videos as time series, where each frame is depicted as a vector within a feature space. Our primary focus centers on aligning a test action’s time series with a reference action’s time series using DTW and S-DTW. While these techniques excel with segmented actions, actions are often not neatly segmented. We introduce OE-DTW and OBE-DTW to effectively manage actions with either known or unknown start and end points.

Building on this foundation, we propose OE-S-DTW and OBE-S-DTW, which meld the smoothness of S-DTW with the endpoint handling abilities of OE-DTW and OBE-DTW. These new variations find value in training neural networks to align partially or fully unsegmented input, representing an innovative use of soft DTW methods. These methods effectively isolate ongoing actions for alignment with reference actions.

Our proposed algorithms are employed to tackle the challenge of action prediction. We transform continuous videos into time series using features like skeletal data or deep features extracted from RGB video data. These unsegmented inputs are matched with reference action time series to predict ongoing actions before they conclude. Moreover, our methods can even forecast when these ongoing actions will come to an end.

In our experiments, OE-S-DTW showcases its effectiveness for segmented action sequences, replacing OE-DTW when differentiability is crucial. Meanwhile, OBE-S-DTW and OBE-DTW surpass other algorithms in segmented sequences and exhibit significant superiority in predicting actions in unsegmented sequences—a more complex and true-to-life scenario.

4.1 Problem description

The core issue addressed in this study is the prediction of actions, which is tackled using newly introduced segregational soft DTW variants. Specifically, the task involves converting continuous video data into time series featuring distinct attributes. This transformation is achieved using either skeletal data-derived features or deep features extracted from RGB video data through a VGG-16 network. The input, which can be partially or entirely unsegmented, is matched against

a collection of similarly represented reference action executions. The aim is to identify the best match, enabling predictions for ongoing actions even before their completion. Additionally, the anticipated conclusion time of the ongoing action can also be foreseen as a secondary outcome.

Let the test (query) action sequence X be represented as $X = (x_1, \dots, x_l) \in \mathbb{R}^{n \times l}$ and the reference video Y be represented as $Y = (y_1, \dots, y_m) \in \mathbb{R}^{n \times m}$. The Euclidean distance of frames x and y is defined as $d(x, y)$ and is used to create the distance matrix $D(X, Y) = [d(x_i, y_j)]_{i,j} \in \mathbb{R}^{l \times m}$ containing all pair-wise frame distances. The cumulative matrix that is based on D and represents all path-based alignments Π of X and Y , is denoted as $C(X, Y) = \{\langle \pi, D(X, Y) \rangle, \pi \in \Pi_{l,m}\}$ where Π represents all the alignments connecting the upper-left to the lower-right of the distance matrix. Given this notation, in the following sections we elaborate on the existing and proposed action sequence alignment algorithms.

4.2 Existing alignment methods

The minimum cost of aligning two time series in their entirety is given by DTW [130] at the last index of the cumulative matrix $C(X, Y)$. The alignment score is normalized with the size of the query. The DTW alignment cost is defined as:

$$(4.1) \quad DTW(X, Y) = \min_{\pi \in \Pi} C(X, Y).$$

Open-End Dynamic Time Warping (OE-DTW) [143]: is a variant of the original DTW [130]. This algorithm is useful when the query and reference sequences share the same start but since an action can be completely observed and the other partially observed they do not share the same endings. The cumulative matrix is calculated using the asymmetric pattern as follows: $C(x_i, y_j) = D(x_i, y_j) + \min(C(x_{i-1}, y_j), C(x_{i-1}, y_{j-1}), C(x_{i-1}, y_{j-2}))$. The alignment can end at any point at the last row of the cumulative matrix. The values are normalized by the size of the query. The alignment score is given by the minimum value in the last row. The alignment cost of OE-DTW is defined as:

$$(4.2) \quad OE - DTW(X, Y) = \min_{j=1, \dots, m} DTW(X, Y_j).$$

Open-Begin-End Dynamic Time Warping (OBE-DTW) [143]: OBE-DTW refers to the DTW variant which allows the matching of one sequence with a part anywhere inside the second sequence. To achieve this, a row with zero values is appended at the beginning of the distance matrix and the computations are performed as in OE-DTW. The new cumulative matrix is denoted as $C'(X, Y)$. The values are normalized by the length of the query. The back-tracing of the minimum-cost path starts from the minimum value of the last row and ends at the first zero-valued row. We introduce the zero-valued row at the end to facilitate the removal of irrelevant matches occurring at the beginning of the reference sequence. This row serves the purpose of easing the constraints related to the starting point of the alignment process. The alignment cost of OBE-DTW is defined as:

$$(4.3) \quad OBE - DTW(X, Y) = \min_{j=1, \dots, m} C'(X, Y_j).$$

Soft Dynamic Time Warping (S-DTW) [17]: S-DTW is an extension of the original DTW algorithm. In contrast to DTW which takes the minimum cost alignment path, S-DTW takes into account all possible alignments. The cumulative matrix $C(X, Y)$ is created by allowing horizontal, diagonal and vertical moves. The cumulative matrix is padded at the top with a row and at the left with a column so that $C_{i,0} = C_{0,j} = \infty$ for all $i, j \neq 0$ and $C_{0,0} = 0$.

The alignment cost of S-DTW is defined as:

$$(4.4) \quad SDTW_{\gamma}(X, Y) = \min_{\pi \in \Pi}^{\gamma} C(X, Y),$$

with

$$(4.5) \quad \min^{\gamma}(\pi_1, \dots, \pi_k) = \begin{cases} \min_{i \leq k} \pi_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^k e^{-\pi_i/\gamma} & \gamma > 0, \end{cases}$$

where $\gamma \geq 0$ is a smoothing hyper-parameter.

4.3 Proposed alignment methods

Open-End Soft DTW (OE-S-DTW): Based on the OE-DTW and the S-DTW, we propose OE-S-DTW, where instead of aligning two sequences to their entirety, we align them partially by having

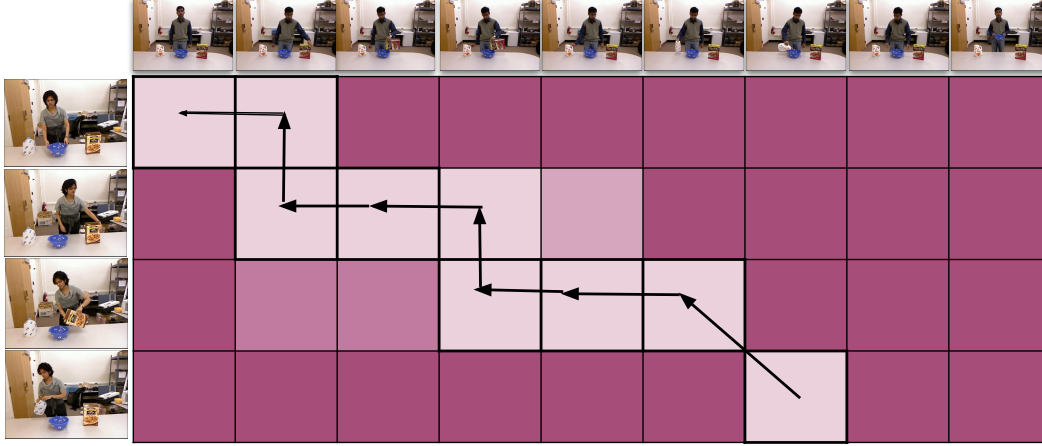


Figure 4.1: Graphical illustration of the OE-S-DTW algorithm. On the horizontal axis we can observe a man performing an activity. On the vertical axis a woman is performing the same activity which is not yet completed. The light pink boxes represent the possible alignment paths while the black arrows represent a possible path. The two sequences share the same starting point but end at different points. The OE-S-DTW algorithm is able to match the partially observed activity with a part of the completely observed one.

them anchored at the beginning, while their endpoints are free. We start by calculating the distance matrix which contains the pairwise distances of the sequences X and Y . The cumulative matrix is calculated using the \min^γ operator as follows:

$$(4.6) \quad C(x_i, y_j) = D(x_i, y_j) + \min^\gamma(C(x_{i-1}, y_j), C(x_{i-1}, y_{j-1}), C(x_i, y_{j-1})).$$

The alignment path can terminate at any point of the last row of the C matrix. The scores at the last row are normalized by the size of the query and the alignment value is the minimum of the last row. Then, the gradient is calculated from that point backwards to the common start point to find the alignment between the two time series. The final OE-S-DTW score is also normalised by the size of the matched reference. The alignment cost of OE-S-DTW is defined as:

$$(4.7) \quad OE-S-DTW(X, Y) = \min_{j=1, \dots, m}^\gamma SDTW_\gamma(X, Y_j).$$

A graphical illustration of the OE-S-DTW algorithm is presented in Fig. 4.1.

Open-Begin-End Soft Dynamic Time Warping (OBE-S-DTW): shares the same alternations as the OBE-DTW. Upon calculating the distance matrix $D(X, Y)$, a row of zero values is appended at the beginning of the distance matrix creating $D'(X, Y)$. Concluding the process at the zero-valued row enables us to filter out insignificant matches occurring at the beginning of the reference sequence. The inclusion of this row serves the purpose of loosening the constraints on the starting point for alignment. The cumulative matrix C' is calculated by using the \min^γ operator as follows:

$$(4.8) \quad C'(x_i, y_j) = D'(x_i, y_j) + \min^\gamma(C'(x_{i-1}, y_j), C'(x_{i-1}, y_{j-1}), C'(x_i, y_{j-1})).$$

The last row of C' is normalized by the size of the query. The alignment cost is the minimum value of the last row. Then, the gradient is computed from that point towards the zero-valued row and ends when it reaches it. The size of the matched reference corresponds to that range. The gradient gives as the alignment matrix, all possible alignments. Once the alignment path is obtained, we normalize the alignment cost with the size of the matching part of the reference sequence. The alignment cost of OBE-S-DTW is defined as:

$$(4.9) \quad OBE - S - DTW(X, Y) = \min_{j=1, \dots, m}^\gamma C'(X, Y_j).$$

An illustration of the OBE-S-DTW algorithm is provided in Fig. 4.2.

4.4 Alignment-based action prediction

Action prediction is defined as the problem of inferring the label of a partially executed action. We define the action observation ratio to be the percentage of the action that is already observed. In our experiments, the observation ratio varies in the range [10%, 100%]. When the observation ratio equals to 100% then the whole action has been observed. In this case, the problem of action prediction becomes identical to the problem of action classification.

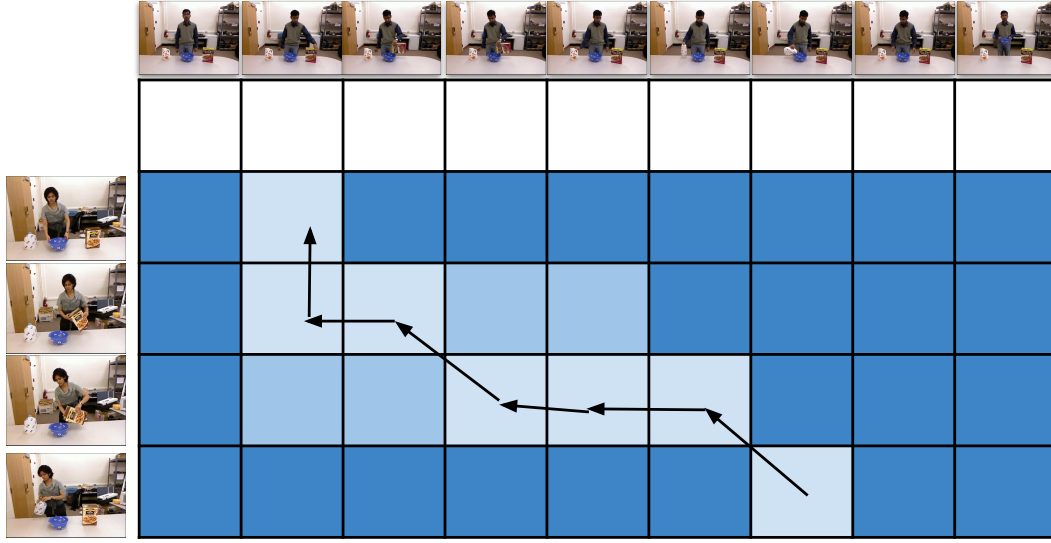


Figure 4.2: Graphical illustration of the OBE-S-DTW algorithm. The activity illustrated at the left (rows) matches a part of the activity illustrated on the top (columns). At the top a zero-valued row is added. The light blue boxes represent all possible alignments while the black arrows show a possible warping path.

To perform video-based action prediction, we represent the videos of executed actions as multidimensional time series. Given time series representations of several prototype executions of certain actions, and an incomplete video execution of one of these actions, we cast the problem of action prediction as a problem of aligning/matching the incomplete action execution to the prototype ones. The label of the closest matching prototype action is reported as the predicted label of the incomplete action.

In more detail, the unknown label $L(Q)$ of a time series representation of an incomplete action Q is inferred through the alignment/matching of incomplete and prototype actions. A set of K time series P^i , $1 \leq i \leq K$, corresponds to prototype videos with labels $L(S_i)$. The alignment cost of two time series X, Y is denoted as $Cost(X, Y)$. Thus:

$$(4.10) \quad L(Q) = L\left(\arg \min_{P^i, 1 \leq i \leq K} \left(Cost(Q, P^i)\right)\right).$$

The proposed methodology can infer the label of an incomplete/query video Q by determining which prototype/reference video S^i has the minimum alignment cost with Q . This is done through the proposed Dynamic Time Warping variants. The label of Q is set to $L(P^i)$.

4.5 Datasets & Metrics

Within this section, we will provide an account of the characteristics inherent to each dataset, outlining the specific features employed for analysis. Additionally, we will delve into the performance metrics utilized and address any implementation considerations.

4.5.1 Datasets

For the evaluation of the proposed methods we employ four standard benchmark datasets which contain trimmed and untrimmed action executions of humans interacting with objects. Action representations encode the human body/object pose and the class of the manipulated object. In general, we test our algorithms with skeletal (i.e., motion capture) 3-D data and features, but also with RGB-based features extracted by a VGG-16 neural network. We follow the approach of our work which was introduced in Chapter 3 regarding the fusion of human and object representations for the computation of the distance matrix of two action sequences. Specifically, the representation fusion is done by employing a weighted sum of the individual distance matrices of the human and object representations. The weights depend on the class of the manipulated objects. If no objects are present in the scene, we use only the human pose representations. In case there are several objects in the observed scene, following [86], we consider the object that is manipulated by and/or closest to the actor.

MHAD Dataset [101]: Contains trimmed executions of 11 human actions. Only one of them (“throwing a ball”) involves human-object interaction. The actions are performed by 5 female and 7 male subjects in different execution styles and speeds. The actions are: jumping in place, jumping jacks, bending, punching, waving one hand, waving two hands, clapping, throwing a ball, sit down and stand up, sit down, stand up. 3D skeletal data of 30 joints have been acquired from a motion capture system that provide 3D positions of 43 LED markers as well as RGB and depth frames. The first 7 subjects are used as the reference sequences while the last 5 subjects are used as the test sequences. The same evaluation split is used as in [101] and our previous work (see Chapter 3).

Skeletal features: Based on the 3D skeletal data of the 30 joints provided by the MHAD dataset, we employ the same human body representation as in [120, 104, 84]. Body-centered and camera-

centered features are employed resulting in a 60-dimensional vector. This vector is extended by 4 angles representing angles encoding the fore- and the back- arms and upper- and lower legs.

VGG features: For this type of data we opted to utilize the data provided in [8]. From a VGG-16 [139] network, 1-D feature vectors are extracted from the last fully-connected layer, resulting in a feature vector of 2048 dimensions for each frame of the sequence. For the RGB frames the network is not fine-tuned and the learned weights are maintained from the training on ImageNet [21]. In the case of optical flow the VGG-16 layers are fine-tuned starting from the last 2-D layer and above on optical flow data from the KTH dataset [132], freezing the rest with the weight values from ImageNet [21].

MHAD101 Dataset [104]: Contains concatenated actions from the MHAD dataset in order to form longer sequences of multiple actions. To alleviate possible ambiguities, the action labeled as sit down/stand up is excluded as it is a composition of the actions sit down and stand up. In the MHAD dataset each action is repeated five times by each subject but in this dataset only the first execution of an action by each subject is used. The skeletal data provided by the MHAD dataset are used and down-sampled to 30fps. The aforementioned actions (excluding the action sit-down/stand-up) are used for creating larger sequences of actions. The synthesised MHAD101 dataset contains 101 pairs of action sequences. In the first 50 paired sequences, each sequence consists of 3 concatenated action clips (triplets) and the paired sequences have exactly 1 in common. In the rest pairs of actions sequences, 4 to 7 actions are concatenated in a long sequence. In all the synthesised sequences the style and duration variability are promoted by using different subjects in forming different triplets. The lengths of the sequences range between 300 to 2150 frames.

Skeletal features: We use the MHAD101-s version of MHAD101 which contains skeletal features. We used only the first 50 pairs of action sequences. By splitting these 50 pairs, we extracted 100 action sequences where each of them contains 3 concatenated actions. These actions are synthesized from the same features that are described in Section 4.5.1.

VGG features: We use the MHAD101-v version of MHAD101 which contains the RGB videos of the same triplets as in MHAD101-s. We then extract features from the VGG-16 network as in [8].

We took into account all the available frames without down-sampling to 30fps.

MSR Daily Activity 3D Dataset [149]: Contains 16 trimmed executions of human-object interactions in two different settings, standing up and sitting on a sofa. The actions are: eating, speaking on cellphone, writing on paper, using a laptop, using a vacuum cleaner, cheering up, sitting still, tossing paper, playing a game, walking, lie down on the sofa, playing the guitar, reading a book, standing up, drinking and sitting down. The standard evaluation split is used as in [157, 119, 86].

Skeletal features: Following our previous work in Chapter 3, we represent the dataset with 3D joint angles and 3D skeletal joint positions. The 3D joint angles are based on the work of [120]. Due to the noisiness of the lower body data, only the upper body joints are used thus resulting in a 30-dimensional feature vector. The 3D joint angles are augmented with the 3D skeletal joint position of the upper body that are invariant to the body center resulting in a $27 + 18 = 45$ -dimensional vector. The object class and the 2D object position are acquired through the YoloV4 [11] algorithm as in our previous work in Chapter 3. The final feature vector per frame is 47-dimensional.

CAD-120 Dataset [66]: The CAD-120 dataset contains long and complex activities of human-object interactions. The activities are performed by male and female subjects and filmed from varying viewpoints. Moreover, the same actions are performed with different objects in order to induce variability in the executions. These activities can be trimmed in actions based on the provided ground truth data. The dataset provides annotations regarding the activity and action labels, object labels, affordance labels and temporal segmentation of activities. The action labels are: reach, move, pour, eat, drink, open, place, close, clean, null and the activities are: arranging objects, cleaning objects, having meal, making cereal, microwaving food, picking objects, stacking objects, taking food, taking medicine and un-stacking objects. The manipulated objects are: cup, box, bowl, plate, microwave, cloth, book, milk, remote, medicine box. The standard 4-fold cross validation split is used as in [66, 86, 155].

Skeletal features: We use ground truth annotations for the semantic labels, bounding box information, and 3D poses of the objects, and also labels and 3D positions for the upper body joints. To represent the trajectory of each object and body joint within a video, we encode their

relative 3D positions in each frame, using the torso joint as a reference point. The joint and object positions are divided by the torso-head distance. The feature vector representing the CAD-120 dataset contains the 3D location of 8 upper body joints, the distance moved by each joint and their displacement. The objects are represented using the 3D centroid location, the distance between the object centroid and each of the 8 human joints. Also, the distance moved by the object and the displacement of the object's centroid. These features are also employed in [66] and our work introduced in Chapter 3.

4.5.2 Performance metrics

The observation ratio of each video is ranging from 10% to 100% with step equal to 10%. The accuracy of the predicted action label is measured by comparing the partially observed video with the prototype videos. Action prediction is quantified using standard metrics such as F1-score, precision, recall and Intersection-Over-Union (IoU).

4.5.3 Implementation issues

For OE-DTW and OBE-DTW we employed a publicly available implementation¹. The implementations of OE-S-DTW and OBE-S-DTW were based on the S-DTW implementation in the Tslern toolkit [127]. The parameter γ was experimentally set equal to 1 for all datasets through evaluation in the range [0.001, 1].

The Segmental DTW implementation is provided by [103, 105]. The parameters of the Segmental DTW algorithm are set according to [103] where it is recommended the minimum length of a warping path to be half the length of the smallest action. Our experiments showed that if the minimum length is set too small, the algorithm ends up with smaller paths that do not represent an alignment. If the minimum length is very high, the algorithm is unable to align videos in the case of small observation ratios.

For our comparison with the work of Cao et al. [12] we need to stress the fact that we are not comparing directly with the full OTAM framework. The comparison is based on the alignment algorithm that creates the distance matrices and finds an alignment path between two sequences and classifies to the reference video that minimizes the alignment score. Since there is no code available for this method, we implemented the alignment component based on the details

¹<https://github.com/statefb/dtwalign>

provided in the paper. Also, the OTAM framework is only tested in trimmed action videos of fixed length. In [12] a cosine distance measure is proposed but in the data used in the current work the Euclidean distance yields the best results for this method. So, the reported results are based on the Euclidean distance and γ value equal to 1. The key differences are the padding with zeros at the start and end of the distance matrix and the different computation of the cumulative matrix which is

$$(4.11) \quad C(x_i, y_j) = D(x_i, y_j) + \min(C(x_{i-1}, y_{j-1}), C(x_i, y_{j-1})).$$

The alignment score is given at the last index of the cumulative matrix which denotes the alignment of the sequences in their entirety.

4.6 Experimental Evaluation

To evaluate the proposed OE-S-DTW and OBE-S-DTW algorithms on the task of human action prediction, we conduct three types of experiments. First, we use these algorithms to perform action prediction in trimmed action sequences containing one action. In this setting, the proposed algorithms are used to align and match an action of known start and variable observation ratio to a set of prototype actions (section 4.6.1). We also compare the performance of the proposed algorithms to a number of competing methods. In a second experiment, the input is a triplet of actions and the goal is to predict the label of the middle action under different observation ratios (section 4.6.2). In this untrimmed video/action setting, both the start and the end of the action are unknown to the algorithms. Finally, given the capability of the proposed algorithms to predict the label of the on-going action, we test how accurately they can predict the end-time of that action (section 4.6.3).

4.6.1 Evaluation in trimmed actions

In this set of experiments we used the trimmed video recordings of the MHAD, MSR Daily and CAD-120 datasets. In Fig. 4.3, we report results on experiments using skeletal features. We evaluate the proposed OE-S-DTW and OBE-S-DTW algorithms in comparison to OE-DTW [143], OBE-DTW [143], Segmental DTW [103] and OTAM [12]. As it can be verified, the performance of OE-DTW and OE-S-DTW is comparable and both achieve very high accuracy in all datasets.

Moreover, OBE-S-DTW outperforms OBE-DTW by a large margin in the MHAD and MSR Daily datasets. In the CAD-120 dataset the performance of OBE-DTW and OBE-S-DTW is practically the same. Note that OE-DTW and OE-S-DTW are aware of the common start of the actions while OBE-DTW and OBE-S-DTW, do not. Thus, OBE-DTW and OBE-S-DTW have to deal with a considerably less constrained/more difficult problem.

In the same figure, we can observe the performances of the two competitive methods, Segmental DTW and OTAM. The proposed algorithms outperform both Segmental DTW and OTAM by a great margin in all datasets.

For the MHAD dataset we also experimented with VGG features. Figure 4.5 shows the performance of all evaluated algorithms in such a setting. It can be observed that even with this type of features, the proposed OBE-S-DTW outperforms Segmental DTW and OTAM as well as OBE-DTW for observation ratio greater than 40%. OE-DTW performs comparably to our proposed OE-S-DTW variant.

For the MSR Daily Activities dataset, in Figure 4.6, we also compare our approach to the competitive methods of Reily et al. [119], Alfaifi et al. [3] and Manousaki et al. [86]. As it can be observed, we outperform [119] and Manousaki et al. [86] at all observation ratios and [3] for all observation ratios greater than 40%.

On top of the results presented above, we also evaluate the current framework on the activities of the CAD120 dataset. To do so, we compare the performance of the proposed OE-S-DTW and the OBE-S-DTW methods to several competitive methods [12, 3, 103, 119, 3] and the works of Wu et al. [155] and Manousaki et al. [87] that hold the state-of-art results. As it can be observed in Fig. 4.4 the OE-S-DTW alignment algorithm performs generally better than the OBE-S-DTW. This happens due to the fact that most of the activities start from the same pose so the OE-S-DTW that has the starting point constraint aligns the sequences more accurately. Generally, the alignment of activities using the current framework has low performance compared to the state-of-art. This happens because the activities are more complex compared to the actions. Additionally, this method has the drawback of taking into account only one object while in these activities several objects are used.

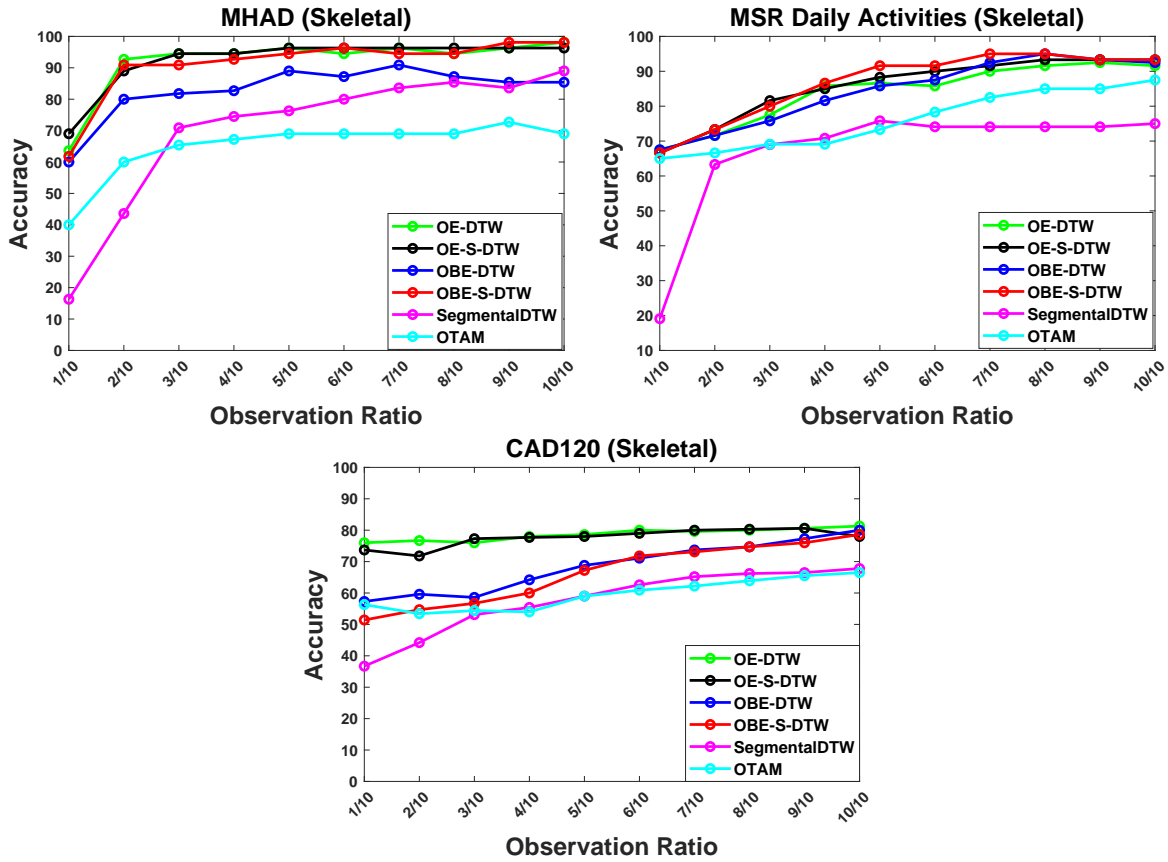


Figure 4.3: Action prediction accuracy in trimmed videos as a function of observation ratio involving skeletal features in the MHAD (Upper left), MSR (Upper right) and CAD-120 (Down) datasets. We compare the different alignment algorithms which are OBE-S-DTW (proposed), OBE-S-DTW (proposed), OE-DTW [143], OBE-DTW [143], OTAM [12] and SegmentalDTW [103].

4.6.2 Evaluation in untrimmed actions

The proposed methods are also evaluated in the untrimmed action sequences (action triplets) of MHAD101-s-v datasets. In this experiment we explore whether OBE-S-DTW can recognize an unsegmented action that appears between some other prefix and suffix actions. To do so, the algorithms progressively observe the whole triplet (3 actions in a row) and aim at segmenting and recognizing the middle action. To achieve this, in each triplet that is observed, we exclude the prefix/suffix actions from the set of the reference actions. The prefix and suffix actions are observed in thirds. The middle action is observed in tenths, so as to have a finer performance sampling relative to the observation ratio of the test action.

Figure 4.7 (left) shows the obtained results on the MHAD101-s dataset (skeletal features). In that plot, the two vertical black lines denote the ground-truth start and end of the middle

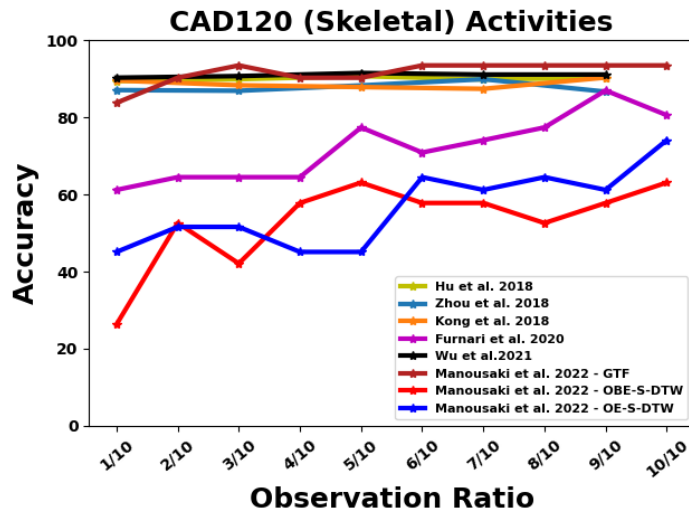


Figure 4.4: Activity prediction results on the CAD120 dataset.

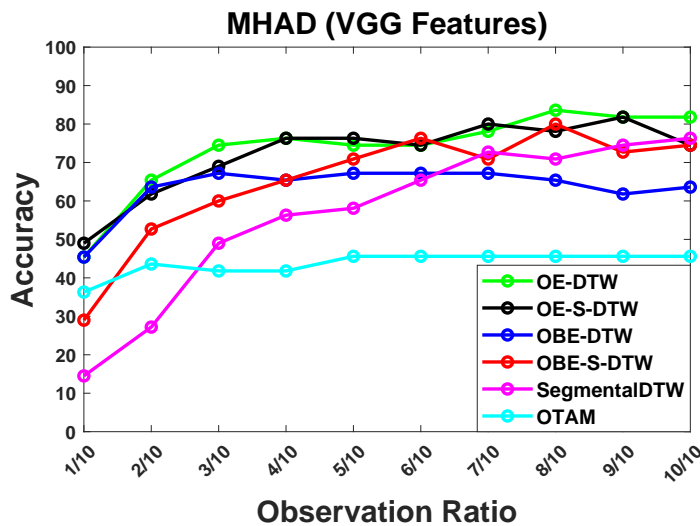


Figure 4.5: Action prediction accuracy in trimmed videos as a function of observation ratio involving VGG-16 features in the MHAD dataset.

action. High accuracy during the prefix denotes the ability of the algorithm to recognize that the algorithm correctly identifies that the sought action has not yet started. High accuracy during the suffix denotes the successful recognition of the middle action inside the triplet. As it can be observed, this experiment is not suited for the OE-DTW and OE-S-DTW variants which fail completely to segment and identify the middle action. OBE-S-DTW clearly outperforms OBE-DTW and all other evaluated algorithms by a large margin.

The same experiment is held using the MHAD101-v dataset using the VGG-16 features

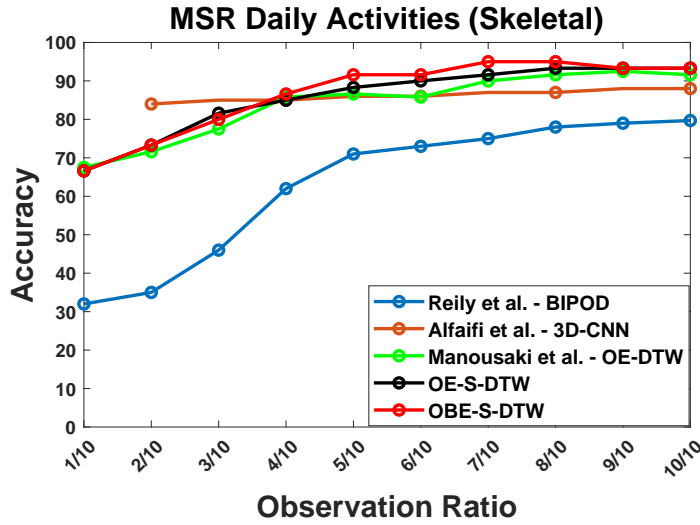


Figure 4.6: Action prediction accuracy of our methods in comparison to state of the art methods on the MSR Daily Activities dataset.

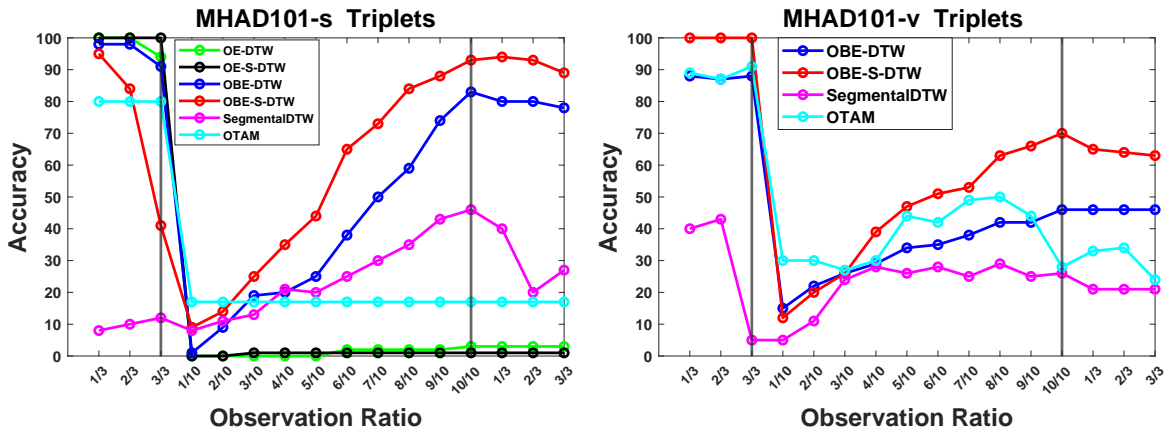


Figure 4.7: Action prediction accuracy in untrimmed videos (video triplets) of the MHAD101 dataset as a function of observation ratio involving skeletal features (MHAD101-s, left) and VGG-16 features (MHAD101-v, right).

(Figure 4.7, right). We observe that overall, the algorithms perform better with skeletal features than with VGG ones. However, their ranking and relative performance does not change. Thus, the superiority of the proposed OBE-S-DTW is independent of the type of the employed features.

In Figure 4.8 we also illustrate the precision, recall, F1-score and IoU for the two best performing algorithms, OBE-DTW and OBE-S-DTW. In all cases, OBE-S-DTW produces better action alignments and, thus, action predictions.

In an effort to gather further experimental evidence, we ran all the experiments reported in

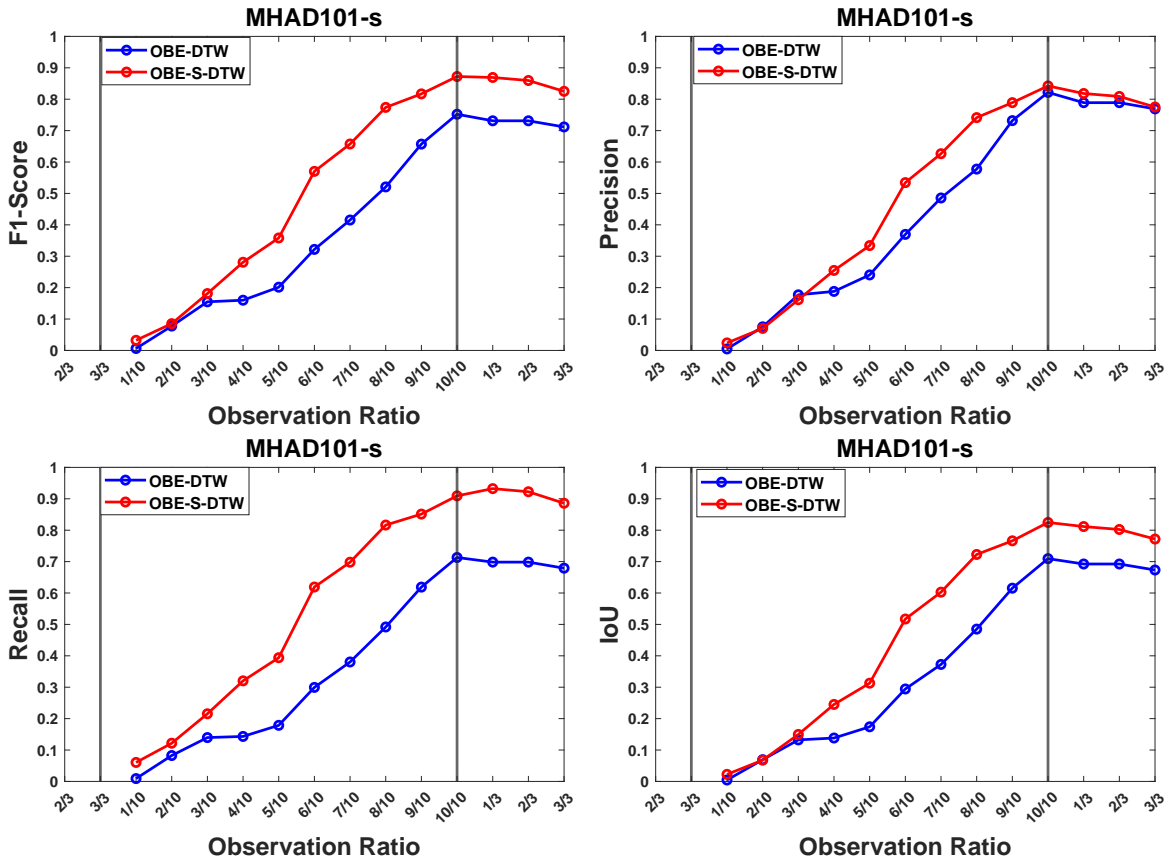


Figure 4.8: Performance metrics (F1-score, precision, recall and Intersection-Over-Union) for OBE-DTW and OBE-S-DTW on the MHAD101-s dataset. Prefix and postfix are denoted with the black vertical lines. Between these lines lies the action to be predicted which is observed in tenths.

this section by reversing the videos/action sequences of the evaluated datasets (i.e., observing them progressively from the end towards the start). Another reason for running this additional tests was to check whether the specific actions that appear as action prefixes or postfixes affect the performance of the evaluated algorithms. We report that this change in the observation order did not affect the performance of the evaluated algorithms and the conclusions of this study.

The OBE-S-DTW algorithm is capable of finding an action anywhere inside a long sequence of actions, as shown previously, for the MHAD101-s/-v datasets. This was achieved by checking whether the OBE-S-DTW can recognize an unsegmented action that appears between some other prefix and suffix actions. We complement those experiments here by observing action triplets not only from start to finish but also backwards (from the suffix to the prefix). It is noted that in each

observed triplet, the prefix/suffix actions are excluded from the set of the prototype actions.

Performance metrics: The prefix and suffix actions are progressively observed in thirds while the middle action is observed in tenths. This protocol is used to acquire finer performance measures for the action of interest. Accuracy is used as the metric for action prediction of the middle action. F1-score, precision, recall and Intersection-over-Union (IoU) are calculated for all observation ratios.

Action prediction results: Figure 4.8 shows the F1-score, recall, precision and IoU scores for the OBE-DTW and OBE-S-DTW algorithms which are the better algorithms overall for aligning unsegmented sequences. From this plot we can observe that the OBE-S-DTW provides better alignments.

A comparison of the OBE-S-DTW and OE-S-DTW with the OE-DTW [143], OBE-DTW [143], SegmentalDTW [108] and OTAM [12] is provided in Fig. 4.7 (left) for the MHAD101-s and in Fig. 4.7 (right) for the MHAD101-v dataset. As it can be observed, the OBE-S-DTW has the best performance overall.

As mentioned earlier, on top of the results presented above, we evaluate the performance of the OE-DTW and OBE-S-DTW algorithms on reversed action triplets. In Fig. 4.9 (left) we observe how the algorithms performs for the MHAD101-s dataset while observing the triplet from start to end (deep red and deep blue lines) and how they perform while observing the triplet from the end to the start (light red and light blue lines). In Fig. 4.9, the two vertical black lines denote the ground-truth start and end of the middle action. High accuracy during the prefix denotes the ability of the algorithm to recognize that the algorithm correctly identifies that the sought action has not yet started. High accuracy during the suffix denotes the successful recognition of the middle action inside the triplet. We can observe a symmetrical effect in the results which means that observing the triplet from start to end or vice versa does not have a significant impact on the algorithms. Also, the OBE-S-DTW algorithm consistently outperforms the OBE-DTW algorithm. The same holds for the MHAD101-v dataset as it can be observed in Fig. 4.9 (right).

4.6.3 Duration Prediction

Being able to forecast the completion time of an ongoing action is an important piece of information in many vision applications. Our algorithms can derive this information by matching a partially

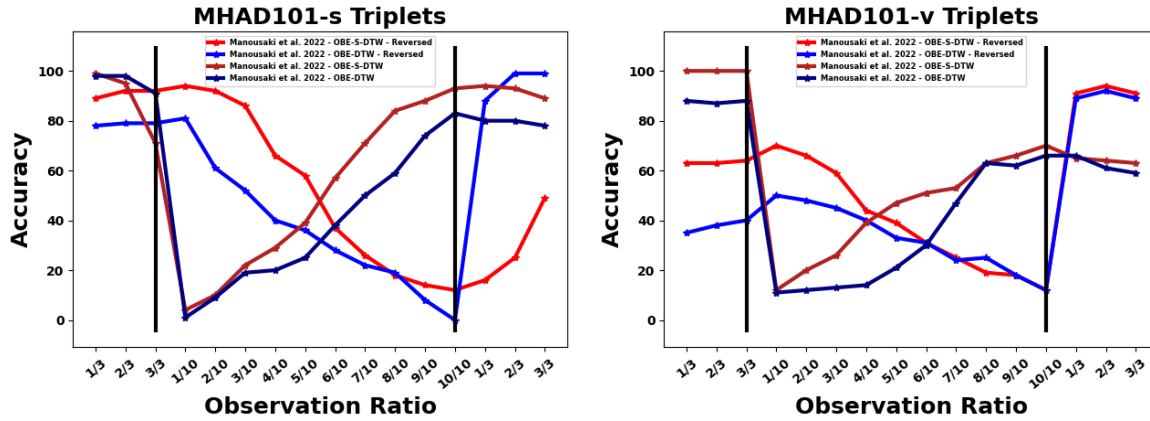


Figure 4.9: Aligning unsegmented action sequences on the MHAD101-s/v datasets using the OBE-S-DTW and OBE-DTW algorithms. Dark red and blue lines depict the accuracy of the alignment algorithms while observing the triplets from the prefix to the suffix. The light (red and blue) lines depict the observation of the triplet from the suffix to the prefix.

observed action to the reference ones and by assuming that this will have the duration of the closest match. In Figure 4.10 we can see the performance of OBE-DTW and OBE-S-DTW in this task. For a given observation ratio, we report the end-frame prediction error which is defined as the discrepancy of the estimated end of a certain action from its ground truth end, as a percentage of the test action length. When an action is wrongly classified by the algorithm, then a prediction error of 1.0 (100%) is added. We observe that as the algorithms see more of a certain action (larger observation ratio) they predict the action end more accurately. Moreover, the proposed OBE-S-DTW appears to outperform clearly OBE-DTW.

4.7 Summary

Within the scope of this research, we introduced not one, but two innovative temporal alignment algorithms designed specifically for matching incomplete videos characterized as multidimensional time series. These algorithms have demonstrated their superiority over existing Dynamic Time Warping (DTW) variations, particularly when it comes to the intricate task of predicting human actions within both trimmed and untrimmed video contexts.

Our comprehensive experimental analysis has yielded substantial evidence supporting the increase in performance of our proposed algorithms relative to the existing approaches. These algorithms exhibit a clear edge in terms of accuracy when compared to established DTW variants.

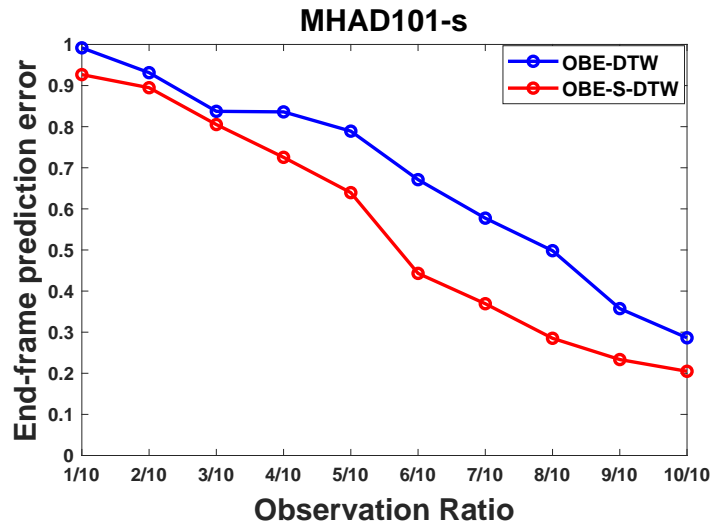


Figure 4.10: Percentage of the frames that are lost compared to the ground truth duration of the action.

This advancement is particularly pronounced within the domain of human action prediction, a challenge that requires nuanced temporal alignment to effectively anticipate actions in video sequences.

The impact of our approach is underscored by the results obtained from various datasets encompassing both skeletal and deep features. Our algorithms have consistently yielded substantial accuracy improvements in human action prediction scenarios across a range of datasets, including MHAD, MSR, CAD-120, MHAD101-s, and MHAD101-v. These datasets span diverse contexts and complexities, highlighting the versatility and efficacy of our algorithms in addressing the intricacies of human action prediction.

In essence, our work has introduced a transformative approach to temporal alignment, enhancing the accuracy of human action prediction within videos characterized by incompleteness. The algorithms' demonstrable superiority over established DTW variations, coupled with their success across multiple datasets, substantiates their potential to be used in numerous frameworks for the alignment of partially observed sequences.

GRAPH-BASED ACTION PREDICTION

Anticipatory prediction empowers intelligent agents to foresee potential outcomes or high-risk events in the future. This enables them to proactively plan responses for early intervention or corrective actions [52, 64, 102]. This skill is especially vital for various applications, including assistive robots in homes or industries [110], predicting pedestrian or obstacle trajectories for self-driving vehicles [148], and more. Our research zeroes in on the prediction of the meaning of an ongoing activity before it is done, as well as predicting the next active objects that will be involved to complete that activity.

Our proposed approach aims to capture the connections between humans and visible scene objects in order to predict the types of multiple next active objects the human will interact with to finish the ongoing activity. Traditional methods have been limited to predicting just one next active object [22, 37, 35]. Our approach is the first to simultaneously forecast both, the ongoing activity's label and all the next active objects that will be used for completion of the activity.

Furthermore, a crucial aspect for such prediction systems is estimating when these next active objects will come into play. Our method stands out as the pioneer in predicting next-active-objects alongside their anticipated involvement time in the activity.

Our contribution lies in jointly forecasting the activity and the objects that will partake in its execution until completion. Rather than focusing solely on predicting interaction points [78, 77, 96] of a next active object, we offer a comprehensive understanding of the activity within the

context of the human and scene objects. Our method revolves around calculating the dissimilarity between graphs that represent the components of the activity [107].

Specifically, we depict the human’s body joints and scene objects as nodes in a graph, with edges representing semantic and motion relations between them. The graph edit distance (GED) [1] is then utilized to calculate the dissimilarity between these graphs.

We validate our approach on video datasets portraying human-object interactions of varying complexity. The widely recognized MSR-Daily Activities dataset [149] covers activities involving none or one object handled by a single person. We further evaluate our method using the CAD-120 dataset [66], which features intricate and extended activities performed by different subjects using varying objects.

Our proposed method, Graphing The Future (GTF), jointly predicts activity labels and next-active-objects by assessing video dissimilarity through GED, while also estimating when these objects will be utilized in the ongoing activity. Our work pioneers the prediction of multiple next active objects in human-object interaction scenarios. GTF establishes pairwise correspondences between objects and human joints in comparing videos, based on semantic similarities and spatio-temporal relationships. This means our predictions remain possible even when interactions with specific objects have never been observed before.

5.1 Problem description

We introduce the GTF method that jointly tackles the tasks of activity prediction and of next active object(s) prediction in videos using graph-based representation of an activity and graph matching technique based on the Graph Edit Distance measure to compare pairs of videos.

The *activity prediction* task can be defined as the problem of inferring the label of an ongoing activity before its actual completion. Let an activity, noted as A , that starts at time t_s and ends at time t_e , thus has a duration $d = t_e - t_s$. Its observation time is defined in proportions of 10% of d . The goal is to predict the correct class as early as possible which implies access to fewer observations.

We also note the task of *next-active-object prediction* as the problem of the inference of the semantic label of an object that will be used in the progress of an activity. Multiple objects may be used in the progress of a given activity A . Related works [22, 35] predict the next-active-object



Figure 5.1: By matching a partially executed and observed activity, to a prototype, fully observed one, we are able to infer correspondences of similar objects and human joints between the two videos. This, in turn, enables to perform activity and next-active-object prediction in the partially observed activity. The example in this figure refers to the “stacking objects” activity, which is performed with a different number and types of objects in the partially and the fully observed activities.

in the segment preceding it’s use, i.e., an amount of time (measured in seconds) before the start of the action that involves the object of interest.

Our approach relies on a graph-based representation of an activity that is captured in video. The entities in a video regard the tracked human skeletal joints and the observable/visible objects. Each video entity is represented as a node of an undirected graph, which also models both semantic information (object label) and its motion (2D or 3D trajectory). Each graph edge connecting two nodes represents the semantic similarity and the spatio-temporal relationships of the interconnected video entities, as described in Section 5.3. Our goal is to devise a novel approach that is able to identify human joints and/or objects in two different videos, one fully and one partially observed video, that exhibit similar behaviors and interactions with other entities using bipartite graph-matching. As shown in Fig. 5.2 a fully and a partially observed video are

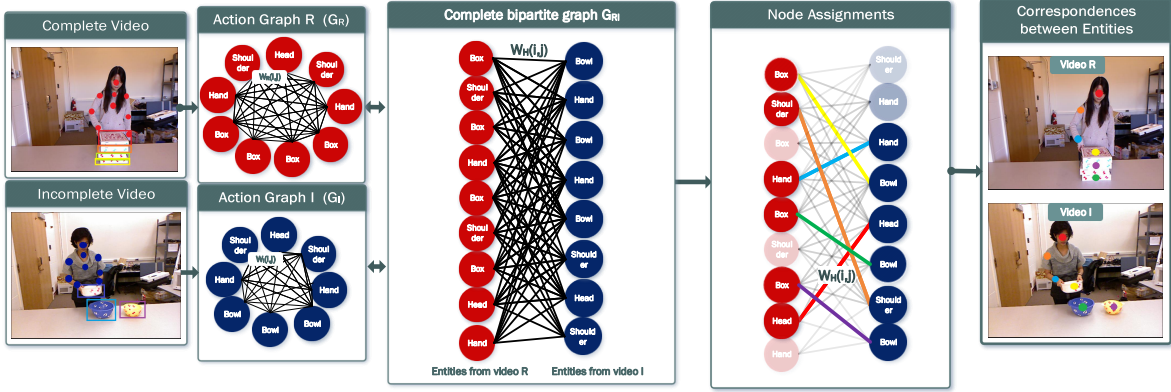


Figure 5.2: Graph matching of a complete video (reference) and an incomplete/partially observed (test) video. First, the fully connected graphs of each video are created based on the video entities. On the basis of these graphs, a bipartite graph between the action graphs is constructed. By calculating the GED, we are able to correspond nodes between the two original action graphs.

represented as two action graphs whose nodes represent the detected and tracked objects and human joints.

5.2 Graph-based video representation

Video Representation: Given a video of duration T frames, it can be seen, at an object-level, as a complete and undirected graph, noted as $G = (V, E)$. In the course of a video, entities such as human body joints and foreground objects are localized and tracked using 2D or 3D human body pose estimation and tracking as well as object detection methods, respectively. Each graph node is noted as $v \in V$ and graph edges are noted as $e_{ij} = (v_i, v_j) \in E$ between nodes $v_i, v_j \in V$, where $i \neq j$. The relations between the nodes describe their dissimilarity in the form of edge weights. The dissimilarity is described based on the semantic dissimilarity s_i and the motion dissimilarity m_i . The edge weight between two connected nodes is defined as the weighted sum of the semantic and motion dissimilarity as follows:

$$(5.1) \quad w_{ij} = (1 - \lambda) * m_{ij} + \lambda * s_{ij}.$$

The parameter $\lambda \in [0, 1]$ is user-defined and controls the contribution of the semantic and motion information. On the extremity of $\lambda = 0$, only motion information is considered while when $\lambda = 1$, only semantic information is used. In the experimental section of this paper,

we present an investigation of the effect of this parameter on the performance of the proposed method.

5.3 Video comparison based on Graph Edit Distance

Semantic Dissimilarity: The weights s_{ij} represent the semantic dissimilarity between the labels of the nodes v_i and v_j . The node labels are retrieved based on ground truth annotations or object recognition methods. The semantic similarity of nodes v_i and v_j with recognized labels l_i and l_j is described as $S(l_i, l_j)$ and is estimated using the WordNet [32] lexical database and the Natural Language Toolkit [80] to compute the path-based Wu-Palmer scaled metric [156]. The similarity is in the range (0, 1] with 1 identifying identical words so semantic weight is:

$$(5.2) \quad s_{ij} = 1 - S(l_i, l_j).$$

Motion Dissimilarity: Each node in the graph is described by a feature vector which can encode information such as the 2D/3D human joint location, the 2D/3D location of the object centroid or any other feature such as appearance, optical flow, etc. The extracted motion features for each dataset are described in section 5.4.2. The acquired 2D/3D skeletal-based pose features or the 2D/3D object-based pose features are described by a trajectory $t(v_i)$ encoding the movement of the video entity during the activity. A pair of trajectories $t(v_i)$ and $t(v_j)$ can be aligned temporally using the OBE-S-DTW algorithm (for details about the OBE-S-DTW algorithm see Chapter 4). The alignment cost of the trajectories $t(v_i)$ and $t(v_j)$ describes the motion dissimilarity of the graph nodes v_i and v_j and is divided by the summation of the length of the trajectory of the incomplete sequence $t(v_i)$ and the length of the trajectory of the reference sequence $t(v_j)$ that matched with $t(v_i)$ as proposed by the authors [85]. Thus, the weight m_{ij} of an edge connecting the graph nodes v_i and v_j is:

$$(5.3) \quad m_{i,j} = \frac{SSDTW(t(v_i), t(v_j))}{len(t(v_i)) + len(t(v_j))}.$$

Graph Operations: Having represented one partially observed and one complete video as graphs, we estimate their dissimilarity by using Graph Edit Distance (GED) [1]. GED is calculated by considering the edit operations (insertions, deletions and substitutions of nodes and/or edges) that

are needed in order to transform one graph into another with minimum cost. Our GTF approach is inspired by the approach of Papoutsakis et al. [107] which uses the GED in order to solve the problem of co-segmentation in triplets of videos. Different from [107] we propose to assess the GED between a pair of videos in order to perform activity prediction. Comparably to [107] our approach is based on semantic and motion similarity of the entities but instead of using the EVACO cosegmentation method [105] to compute the alignment cost of the co-segmented sub-sequences we employ the SSDTW algorithm [85] to align the trajectories between pairs of nodes. The SSDTW algorithm has been shown to have better performance in aligning incomplete/partially observed sequences for the task of action prediction.

We create a graph for each video G_I (Incomplete video) and G_R (Reference video) and assess their graph distance. W_I and W_R are the dissimilarity matrices of action graphs G_I and G_R with size $N_I \times N_I$ and $N_R \times N_R$, respectively, where N_I and N_R are the number of vertices of each graph. As seen in Fig. 5.2 the next step is to create the bipartite graph G_{IR} of the action graphs G_I and G_R . The edge weights W_H connecting the nodes of graph G_I to nodes of graph G_R are calculated using Equation (5.1). In order to calculate the GED on the bipartite graph we need to employ the Bipartite Graph Edit Distance (BP-GED) which solves an assignment problem on the complete bipartite graph using the Kuhn-Munkres algorithm [95]. The weights of the complete bipartite graph G_{IR} are: $W_{IR} = \begin{bmatrix} 0_{N_I, N_I} & W_H \\ W_H^T & 0_{N_R, N_R} \end{bmatrix}$ where $0_{x,y}$ stands for an $x \times y$ matrix of zeros. The solution of this assignment problem requires the definition of the graph edit operations and their associated costs.

Node operations: Consist of node insertions, deletions and substitutions. The cost of inserting and deleting a node v is:

$$(5.4) \quad nd_{in}(empty_node \rightarrow v_i) = \tau_v, \quad nd_{del}(v_i \rightarrow empty_node) = \tau_v$$

while the cost of substitution of node v with node u is:

$$(5.5) \quad nd_{sb}(v_i \rightarrow u_j) = \left[\frac{1}{2\tau_v} + \exp(-\alpha_v * W_H(i, j) + \sigma_v) \right]^{-1}.$$

The parameters of the cost operations for the nodes were set experimentally to $\tau_v = 0.4$, $\alpha_v = 0.1$

and $\sigma_v = 0.0$.

Edge operations: also consist of insertions, deletions and substitutions. The costs of inserting and deleting an edge from node n of graph G_I to node u of graph G_R is:

$$(5.6) \quad e_{in}(e_{ij}^{G_I} \rightarrow e_{mn}^{G_R}) = \tau_e, \quad e_{del}(e_{ij}^{G_I} \rightarrow e_{mn}^{G_R}) = \tau_e.$$

Finally, the cost of edge substitution is defined as:

$$(5.7) \quad e_{sb}(e_{ij}^{G_I} \rightarrow e_{mn}^{G_R}) = \left[\frac{1}{2\tau_e} + \exp\left(-\alpha_e \cdot \frac{W_I(i,j) + W_R(m,n)}{2} + \sigma_e\right) \right]^{-1}.$$

The parameters of the cost operations for the edges where set experimentally to $\tau_e = 0.3$, $\alpha_e = 0.1$ and $\sigma_e = 100$.

Action distance: The dissimilarity between a pair of graphs (G_I, G_R) is computed by the BP-GED which calculates the exact GED [1]. With GED the minimum edit operations are calculated for transforming graph G_I to graph G_R . The dissimilarity, denoted as BP-GED(G_I, G_R), in the work of [107] is normalized by the total number of objects. This normalization is effective when looking for commonalities between videos but is ineffective for activity prediction. In our work we need to be flexible in the number of objects that can be used during an activity while discarding irrelevant objects. In order to achieve this, we found that the best option is to normalize by the number of pairs of matched objects (MO). This helps us to assess our method on the objects that are important for the prediction and discard objects that may be present but with no use in the activity performed. Thus, the dissimilarity $D(G_I, G_R)$ of graphs G_I, G_R is defined as:

$$(5.8) \quad D(G_I, G_R) = \frac{BP-GED(G_I, G_R)}{MO}.$$

5.4 Datasets & Metrics

5.4.1 Datasets

MSR Daily Activity 3D Dataset [149]: The activities contained in this dataset involve human-object interactions in trimmed video executions. The dataset contains 16 activity classes the executions of which are performed by male and female subjects, the first time by standing up and the second by laying down. The dataset contains the 3D locations of the human body joints. The evaluation split of the related works [119, 86, 85] is used for a fair comparative evaluation.

CAD-120 Dataset [66]: Contains complex activities that represent human-object interactions performed by different subjects. The activities are performed using 10 different objects and are observed from varying viewpoints. Each of the 10 activities contains interactions with multiple object classes in different environments. The dataset provides annotations regarding the activity and sub-activity labels, object labels, affordance labels and temporal segmentation of activities. The split of the related work [155] is used for a fair comparative evaluation.

5.4.2 Feature Extraction

The employed datasets are recorded from a third-person viewpoint, therefore they provide information for the whole or upper body of the acting subjects. We decided to align with our previous work (Chapter 4) and consider only the upper body human joints for both datasets. For the MSR Daily Activity 3D Dataset the features used are the 3D joint angles and 3D skeletal joint positions [85]. Object classes and 2D object positions are obtained from YoloV4 [11]. For the CAD-120 Dataset the 3D location of the joints of the upper body are used. As for the objects, the ground truth labels are used along with their 3D centroid locations [86, 85]. For more details see Chapter 3.

5.4.3 Evaluation Metrics

Activity Prediction: Activities are observed in a range from 10% to 100% of their total duration with steps equal to 10%. At every step, the accuracy of the predicted activity label is evaluated

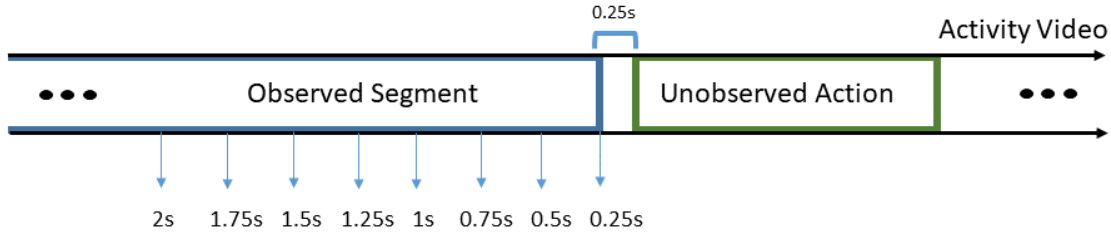


Figure 5.3: Observing the activity and making object predictions for [2s, 1.75s, 1.5s, 1.25s, 1s, 0.75s, 0.5s, 0.25s] before the beginning of the next action as in [35].

compared to the ground truth.

Next-Active-Object Prediction: At variable time steps before the start of the next segment (see Fig. 5.3) where the next-active-object will be used, we estimate the accuracy of the predicted object label compared to the ground truth label. Also, we calculate the time at which the next-active-object will be used in the activity. For the aforementioned time steps the prediction error is calculated as the difference of the predicted time of use and the ground truth time, divided by the length of the video.

5.5 Experimental Results

In this section, we will assess the predictive abilities of the GTF framework when applied to actions that are only partially observed, as well as its performance concerning the prediction of the next active object. Additionally, we will examine the influence of the λ parameter on determining the balance between semantic and motion information utilization. Lastly, we will extend our evaluation to encompass two novel tasks: forecasting the timing of use of the next active object and predicting multiple next active objects.

5.5.1 Activity Prediction/Early Recognition

Activity label prediction is performed by considering observation ratios in chunks of 10% until the end of the video. The label prediction at 100% can be regarded as activity recognition. The test video is compared with all the reference videos by calculating the GED and is assigned to the label of the minimum. In Fig. 5.4 (left) a comparison of our method against the competitive methods for the MSR dataset is shown. Our method outperforms the works of Cao et al. [12], Alfaifi et

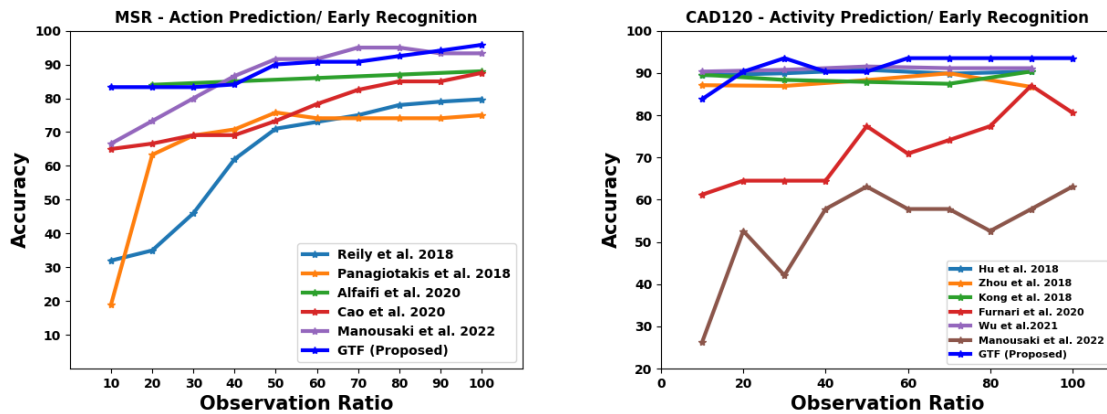


Figure 5.4: Activity prediction results for the (left) MSR Daily Activities and (right) CAD-120 datasets for different observation ratios.

al. [3] and others [103, 119, 3] by a large margin. Our work also outperforms our previous method presented in Chapter 4 (here mentioned as Manousaki et al. [85]) by a large margin at small observation ratios. Results of the competitive methods are taken as shown in [85].

CAD-120 is a challenging dataset due to the number of objects and their interchangeability in different executions of activities. In this dataset, our method outperforms the works of Manousaki et al. [85], Furnari et al. [35] and other competitive methods [65, 166, 51] by a large margin. It also outperforms the approach of Wu et al. [155] that holds the state-of-art performance, for all observation ratios greater than 20% (see Fig. 5.4, right). The results of the [65, 166, 51] and [155] methods are taken from the work of Wu et al. [155] while for our previous work (Manousaki et al. [85] - Chapter 4) we trained and tested using the activities (instead of actions) with the parameters mentioned in that paper.

In Figure 5.4 we can observe that the GTF method outperforms our previous work Manousaki et al. [85] - Chapter 4 on the CAD-120 dataset but the same does not apply to the MSR dataset for the majority of the observation ratios. The MSR dataset contains one object per action while the CAD-120 contains multiple objects per action. This difference is explained by the fact that the GTF method is designed to handle multiple objects while our previous work is designed to handle only one object (the closest to the user).

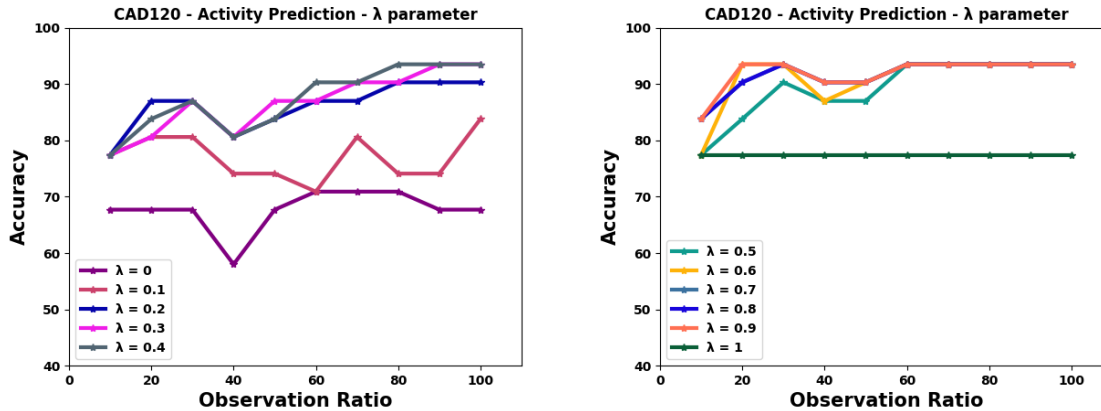


Figure 5.5: Exploration of the user-defined λ parameter on the CAD-120 dataset. The values of the λ parameter are in the range $[0, 1]$. Some curves may be partially visible due to occlusions. Plots are separated in two figures to aid readability.

	Next-Active-Object Prediction Accuracy							
Time	2.00s	1.75s	1.50s	1.25s	1.00s	0.75s	0.50s	0.25s
RULSTM [35]	18.6%	18.6%	18.0%	18.6%	18.6%	19.3%	20.0%	22.0%
GTF (Proposed)	87.0%	87.0%	86.6%	89.1%	90.0%	91.0%	95.0%	97.0%

Table 5.1: Next-active-object prediction accuracy for $[2s, 1.75s, 1.5s, 1.25s, 1s, 0.75s, 0.5s, 0.25s]$ before the beginning of the next action for the CAD-120 dataset.

5.5.2 The impact of parameter λ

Edge weights are determined based on the proportion of the semantic and motion information they convey. This proportion is quantified by the user-defined parameter λ (see Equation (5.1)). In Fig. 5.5 we present results that explore the impact of λ on the performance of our approach on the CAD-120 dataset. When $\lambda = 0$ (only motion features) and $\lambda = 1$ (only semantic features) the results are alike in terms of having the lowest ability to make accurate predictions. Their combination carries a lot more information and gives the best results. Some values are not visible in the plots because for different values of the λ parameter, accuracy values remain the same. After experimental evaluation, the best value across datasets is $\lambda = 0.8$.

5.5.3 Next-Active-Object Prediction

Our method is designed to accommodate videos captured from a third-person viewpoint as we need to have a view of the human joints and the surrounding objects. The most related work to ours is the work of Dessalene et al. [22] which is currently limited only to egocentric videos.

This does not allow for a comparison with that approach. We compare our method to the recent work of Furnari et al. [35]. This work performs on both egocentric and third-view datasets and is the method that [22] compares with. Their performance is comparable for the task of next-active-object prediction. However, instead of following their experimental scheme and evaluating only the accuracy of the prediction of the next-active-object, we also evaluate the accuracy of the prediction in relation to the time prior to the start of the action where the next-active-object will be used. Predictions are made in the range [2s, 1.75s, 1.5s, 1.25s, 1s, 0.75s, 0.5s, 0.25s] before the beginning of the action (see Fig. 5.3). As seen in Table 5.1 our method can correctly predict more objects as we move closer in time while [35] can predict less accurately the objects and is not affected by the time horizon. By comparing the graph of the partially observed video with those of the reference videos, the pair of graphs that have the smaller graph edit distance and object correspondences between the graphs are estimated (test and reference videos may have different number of objects). The work of Furnari et al. [35] is tested using the CAD120 dataset and the publicly available implementation. We extracted the 1024-dimensional features by using TSN [151] and calculated object features using the ground truth annotations. Their code accommodates the extraction of predictions at different seconds before the beginning of the action as described above.

5.5.4 Next-Active-Object Time Prediction

Another aspect of great importance is the ability to forecast the time at which the object will be used in the activity. With the use of the GTF method we are able to compare the partially observed video with the reference videos from the training. After finding the pair of graphs that have the smaller graph edit distance, we acquire the information about object correspondences. This ability to infer the object correspondences between the two videos allows us to have the same number of objects between the videos in order to perform video alignment with the use of SSDTW. The alignment provides the ability to find the point of the reference video that corresponds to the current point in time in the test video (matching point). This projection of time from the reference video to the test one, permits the forecasting of the time at which the next-active-objects will be engaged in the interaction. The prediction error is calculated as the offset of the predicted time of use from the ground truth time of use of the next-active-object compared to the duration of the video. The error is calculated upon the correct predictions of the next-active-object. In Table 5.2

CAD120	Next-Active-Object Time Prediction Error							
Time	2.00s	1.75s	1.50s	1.25s	1.00s	0.75s	0.50s	0.25s
GTF (Proposed)	0.471	0.463	0.46	0.457	0.443	0.405	0.36	0.325

Table 5.2: Time prediction error is the offset of the predicted time of the next-active-object use to the ground truth time of use compared to video length. Predictions are made from 0.25s to 2s prior to the start of the next action.

CAD120	Multiple Next-Active-Objects Prediction Accuracy									
Observation Ratio	10%	20%	30%	40%	50%	60%	70%	80%	90%	
GTF (Proposed)	41.7%	43.2%	45.6%	45.6%	47.1%	47.1%	48.6%	50%	55.9%	

Table 5.3: Accuracy for predicting multiple next-active-objects for different observation ratios.

we observe that this error is low, which means that we are able to accurately predict the time at which the next-active-object will be used in the activity.

5.5.5 Multiple Next-Active-Objects Prediction

Our method is capable of predicting not just one, but multiple next-active-objects. These predictions can be performed at different observation ratios from 10% to 90% (an observation ratio equal to 100% means that the whole video is observed, so next object prediction is not defined). The accuracy for each observation ratio for the predicted next-active-objects is presented in Table 5.3. The prediction is made through the correspondence of the objects between the reference and test graphs. By knowing the order in which the objects in the reference video are used, we can infer the order in which the objects of the test video will be used. After finding the matching point (see the previous section) we can infer the order of the matched objects from that point till the end. Prediction of multiple next-active-objects is challenging due to long time horizons involved and the related increased uncertainty.

5.5.6 Ablation Study

5.5.6.1 Temporal alignment

We evaluate the performance of OE-S-DTW and OBE-S-DTW on the problem of activity prediction when employed on the proposed graph-based framework Graphing The Future (GTF). As it can be observed in Fig. 5.6 the OBE-S-DTW algorithm helps the GTF method achieve much better results compared to the use of the OE-S-DTW algorithm. This method considers multiple objects

for the task of activity prediction. The motion weights with the use of OE-S-DTW and OBE-S-DTW are calculated for each pair of nodes in the action graphs and the bipartite graph. The boundary constraint of OE-S-DTW is not an advantage in this setting where objects can be used at different points in time and in mixed order. More specifically, the decrease in accuracy as we move towards bigger observation ratios is happening due to the OE-S-DTW not being able to match activities where there are motion differences in the executions of the actions due to the constraint in matching the first points of the two sequences.

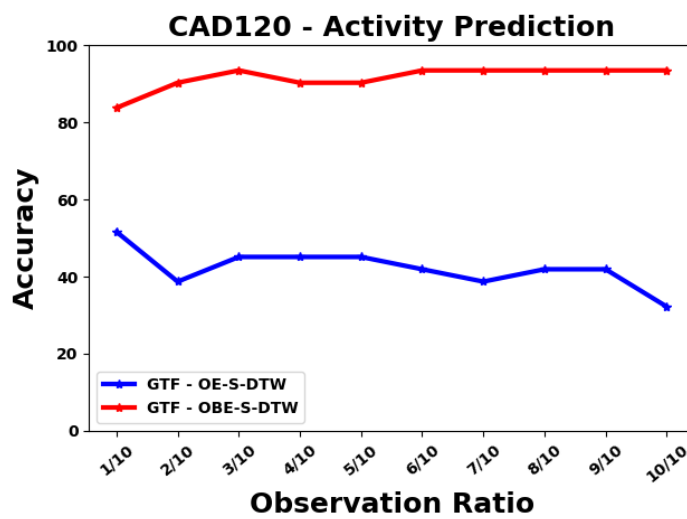


Figure 5.6: Activity prediction results on the CAD120 dataset using the GTF framework. The OBE-S-DTW and OE-S-DTW algorithms were used to quantify the motion dissimilarity of the entities involved in the considered activities.

5.5.6.2 Duration Prognosis

Knowing the label of an ongoing action before its completion is a very useful capability. In certain situations, it is equally important to be able to predict the time at which the currently observed action/activity will end. As proposed in the SSDTW (Sec. 4.6.3), action duration prognosis is defined as the prediction of the time remaining until the completion of the currently observed action. Currently, in the framework of SSDTW the duration prognosis has been evaluated only on the MHAD101-s dataset (see also Fig. 5.7). We extend this evaluation by performing duration prognosis of actions and activities on the MSR and CAD120 datasets, respectively.

Performance metrics: For a given observation ratio, we report the end-frame prediction error

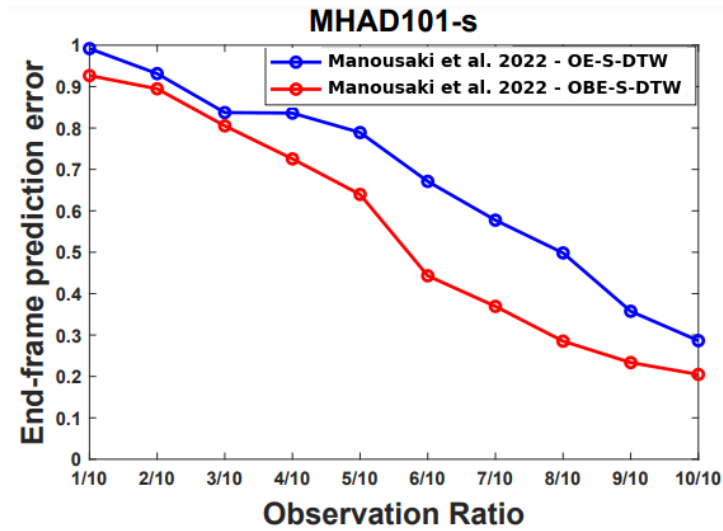


Figure 5.7: End-frame prediction error calculated for all observation ratios of the middle actions of the triplets of MHAD101-s.

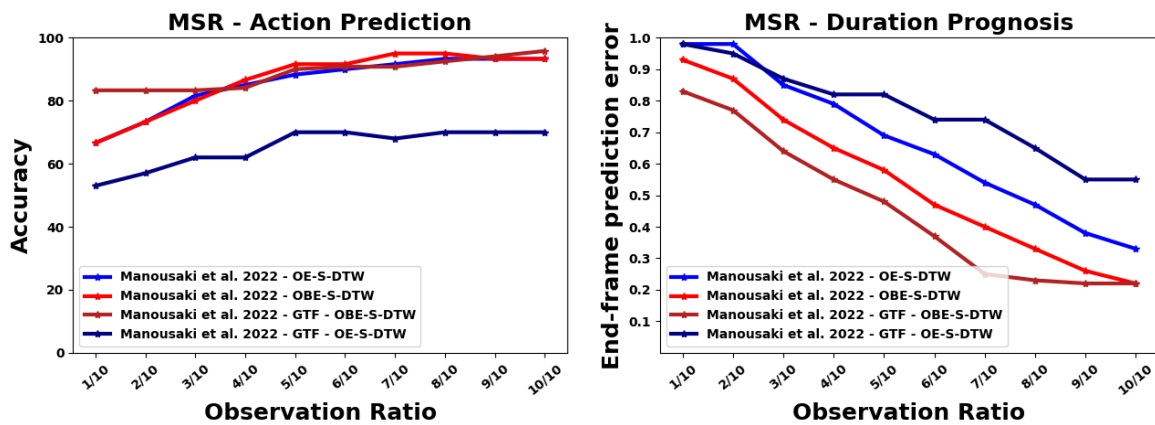


Figure 5.8: (Left) Action prediction results for the MSR Daily Activities dataset. (Right) Prognosis of the duration of the partially observed actions for the MSR Daily Activities dataset.

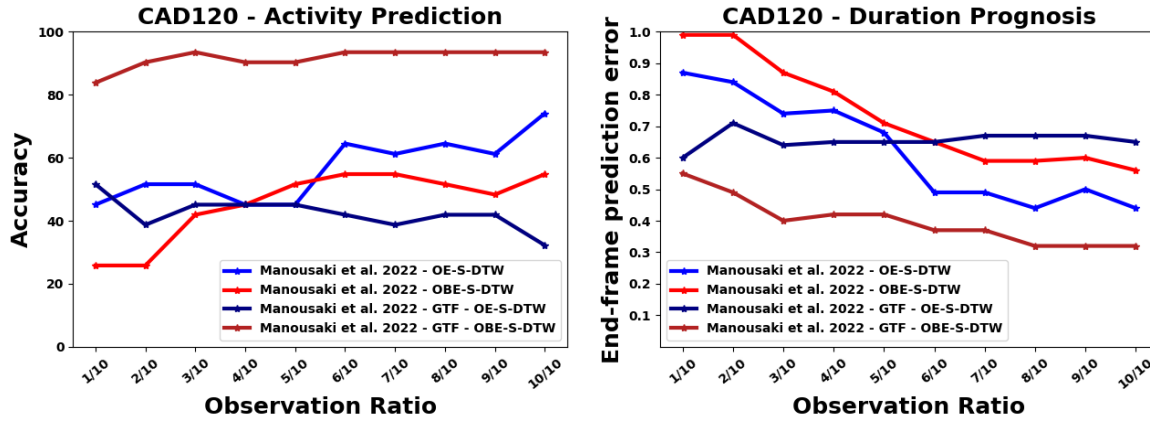


Figure 5.9: (Left) Activity prediction results for the CAD120 dataset. (Right) Prognosis of the duration of the partially observed activities of the CAD120 dataset.

which is defined as the discrepancy of the estimated end of a certain action/activity from its ground truth end, as a percentage of the test action length. When an action/activity is wrongly classified by the algorithm, then a prediction error of 1.0 (100%) is added.

Action duration prognosis results: Figure 5.7 shows the results of duration prognosis on the triplets of the MHAD101-s dataset, as presented in [85]. We can observe that the OBE-S-DTW has smaller error rates across all observation ratios compared to OBE-DTW.

We also report the evaluation of duration prognosis on the actions of the MSR dataset. Figure 5.8 (left) show the relevant action prediction results while in Fig. 5.8 (right) the results of the duration prognosis are presented. The OBE-S-DTW is the best choice for this dataset across the different frameworks. The higher the action prediction accuracy, the lower the duration prognosis error. As the observation ratios increase more meaningful alignments are established thus matching with actions of similar temporal duration. While the prediction accuracy of the different frameworks is similar, we observe differences in the end-frame prediction error. This happens because the test actions get classified to different reference actions thus having variability in their predicted temporal duration.

Activity duration prognosis results: Complementary to the duration prognosis for actions, we extend the experimental evaluation of duration prognosis on the activities of the CAD120 dataset. In Fig. 5.9 we present the results of activity prediction and duration prognosis of the SS-DTW and the GTF framework on the CAD120 dataset. As it can be observed, the OE-S-DTW performs

better than the OBE-S-DTW in the framework of SSDTW while the opposite holds true for the GTF framework. As explained earlier, these frameworks handle different numbers of objects. Thus, depending on the framework, the dataset and its characteristics, different alignment algorithms should be employed. Moreover, moving forward onto the timeline we can see that the end-frame prediction error is decreasing for both algorithms, as they observe a larger portion of the test activity.

5.6 Summary

In our research, we unveiled a novel methodology called Graphing The Future (GTF), which revolutionizes the way we match both complete and partially observed videos by leveraging a graph representation. This innovative approach capitalizes on Bipartite Graph matching techniques to establish connections between videos and exploit their inherent temporal dynamics. Central to our approach is the representation of human joints and objects as distinct nodes within the graph, while the edges between these nodes capture the intricate interplay of semantic and motion-based similarities.

The significance of this framework becomes evident as we showcase its capability to not only predict activities but also anticipate the next active object within a given context. By adopting this graph-centric formulation and computational process, we achieved groundbreaking advancements in the field of predictive modeling. Our results consistently outperformed existing benchmarks, underscoring the effectiveness of the GTF method in enhancing the accuracy and reliability of activity and object prediction tasks.

Furthermore, we tackled the challenging task of forecasting the timing at which the subsequent active object will come into play, along with the complex scenario of predicting multiple next active objects. This extension of our approach showcases its adaptability and versatility in addressing nuanced aspects of predictive modeling, which holds great promise for real-world applications.

The intrinsic ability of our approach to account for both partial observations and complete videos, coupled with its effectiveness in handling time-sensitive and multifaceted predictions, positions GTF as a trailblazing solution in the realm of video-based predictive modeling.

ACTION FORECASTING/ANTICIPATION

Anticipating future actions during an observed complex activity is a critical ability that enables humans to recognize intended goals and outcomes to proactively plan and engage in interactions with other humans and the environment in a timely, efficient, and safe manner. We accomplish this task naturally by perceiving visual information and learning from a few activities as well as based on self-experimentation; thus, it encompasses harnessing relevant kinematic and contextual knowledge rooted in perception, personal experience, and skills. These competencies are regarded as fundamental constituents of human intelligence.

Deriving effective solutions for similar competencies is also beneficial to AI-enabled agents and robots that operate in industrial and domestic environments in a multitude of real-world applications [64]. In particular, the anticipation of near or long-term future actions can efficiently be used to advance autonomous navigation or driver-assistance systems, leverage the ability of industrial or home/socially assistive robots towards fluent human-robot collaboration and interaction, drive optimization of industrial workflows and enhance human safety through real-time hazard/anomaly identification to preemptively signal alerts and aids [98].

To empower AI agents with such abilities, researchers have focused their efforts on video-based human understanding showing remarkable results on the tasks of recognition, detection, and short- or long-term prediction/anticipation of actions during long, composite activities [64]. Of these challenging tasks, the most recent and notable is action anticipation, which entails

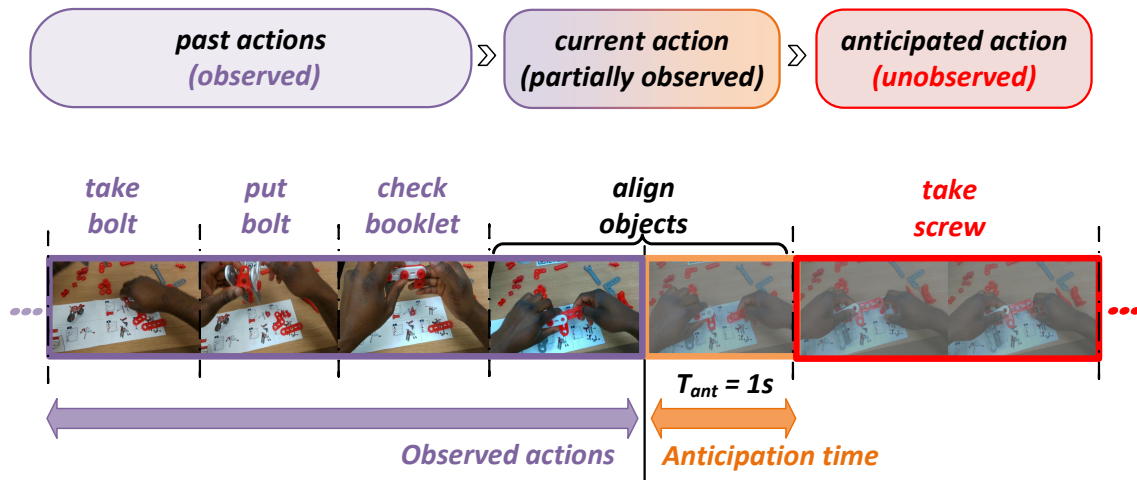


Figure 6.1: We consider the problem of action anticipation in untrimmed videos of procedural activities. At a certain moment in time (decision point), the proposed framework (VLMAH) [88] anticipates the action (i.e., the unobserved action “take screw”) that is most likely to be performed after some anticipation time T_{ant} (depicted with orange color). This is performed on the basis of the history of all past actions up to the decision point (depicted with purple) which is modeled by integrating visual input regarding the immediate past and a linguistic description of the distant past.

predicting the label of the following action(s) based on partial observation of the ongoing action and knowledge/observations of recent action history [121, 19], as shown in Figure 6.1. The ability to utilize information regarding the recent history of observed actions is essential for inferring potential action proposals for the anticipated (subsequent) action at the decision point T_{ant} , prior to its expected start time. This is referred to as anticipation time and is significant because it provides valuable information and a record of the temporal sequence of actions leading up to the anticipated action.

We identify the following questions towards this challenging task, which effective solutions have to deal with by assessing the best trade-off between the complexity of spatiotemporal visual feature modeling and the accuracy performance of action anticipation:

- How much of the action history should be considered to accurately predict future actions during complex activities?
- What is the most efficient way to model the temporal ordering of action history (past actions)?
- What information modalities could enhance action anticipation accuracy?

We address the challenging problem of anticipating the label of the next action/verb/noun and answer these important questions in the context of instructional activities by combining visual and linguistic information of actions observed during an ongoing activity considering both its recent and distant past. This is because retaining past information, thus modeling action history, is crucial for the predictability of the future. The visual features are very informative but tend to have high storage and computational cost, so retaining large amounts of visual features of past actions is not always feasible. On the other hand, language-based action descriptions are more coarse in nature but more efficient in terms of storage and processing time. Our method leans toward combining the best of both worlds by fusing high-cost visual features for the recent past and low-cost language features for the distant past.

We explore action anticipation in the context of procedural activities, where variations of the temporal ordering of actions are usually more constrained. Based on that, it is not surprising that the majority of existing works [41, 42, 158, 97, 79, 126, 165, 123, 53] aspire to tackle this problem using video datasets [19, 18, 73, 140, 69, 136, 116, 10] containing procedural activities. For instance, EpicKitchens [18] is one of the largest and most popular video datasets, among others [73, 140, 69, 135, 167], deals with the task of action anticipation featuring videos of cooking activities. Another popular domain of instructional activities that regard complex assembly activities [136, 116, 10, 54, 70, 106] in the context of industrial and non-industrial scenarios.

In particular, we focus on videos of assembly activities using the Meccano [116] and the Assembly-101 [136] video datasets. Those two can be considered complementary with respect to the types of the target activities, as participants in the former are provided with specific instructions to accomplish the assembly process of a toy vehicle, whereas in the latter participants were free to disassemble a fixed toy vehicle and then to assemble it from its parts, following a less constrained process.

6.1 Problem description

The task of predicting upcoming actions presents considerable complexity, with the structure and efficacy of prediction closely tied to the chosen time window for these anticipations. Specifically, the ability to predict actions is contingent upon the selection of decision points in time (referred to as timesteps) that occur before the initiation of the subsequent action segment.

To comprehensively evaluate the performance of our proposed framework, we adopt the evaluation protocol detailed in the research by Furnari et al. [35]. This protocol involves generating predictions at eight distinct anticipation timesteps preceding the onset of the near-future expected action. Referred to as τ_{ant} , this collection of anticipation times encompasses discrete values spanning from 2s to 0.25s, with each increment being 0.25s. Significantly, the upper threshold of this range, specifically 0.25s, closely aligns with the commencement of the predicted action. This evaluation setup ensures a thorough examination of the framework’s predictive capabilities across a diverse set of time horizons.

6.2 Visual-Linguistic Modeling of Action History framework

The proposed Visual-Linguistic Modeling of Action History framework, noted as VLMAH, is shown in Figure 6.2. It features a two-stream three branch deep neural network model design that comprises (a) a vision-based action anticipation sub-network, (b) an activity-level sub-network for temporal modeling based on natural language processing (NLP), and, (c) a vision-based action recognition sub-net. The action anticipation visual sub-net is able to estimate the next action given the visual representation of the current/ongoing action segment exploring the short- and long-term action dynamics. The action recognition sub-net exploits the same input to provide estimates for the current action class. Additionally, the NLP-driven activity-centric sub-net is responsible for the long-range temporal modeling of the relation of the current action to the previously observed actions to learn a stochastic model of the forthcoming action.

The last stage of the architecture combines the two representations (visual action anticipation sub-net & language modeling sub-net) to anticipate one of the following events, (a) the next action (fine-grained label description), (b) the active object of the next segment (noun), or (c) the next motion motif (verb).

6.2.1 Visual Action Anticipation Module

Given an input sequence x_t of the action y_t of an activity video sample $X_i = \{x_1, \dots, x_N\} \rightarrow Y$, the visual action-anticipation sub-net aims at learning the representation of the on-going action at a segment-wise level, that will enable the prediction of the forthcoming action y_{t+1} . To achieve this, the proposed module follows a multi-branch design that operates on an ensemble of different

vision-driven representations of the characteristics of the entire scene or of the key to the action scene elements, such as the actor’s body-part regions or the appearance states of the active-object.

On a technical basis, each branch of the proposed multi-branch design comprises a two-layer Bidirectional LSTM (BiLSTM) temporal encoder, followed by a Fully-Connected (FC) layer, that further encodes the representation into a $[1 \times 256]$ feature vector. Finally, the representations of all branches are fused via concatenation and forwarded to a FC layer that generates the final representation, which encodes the action segment into a $[1 \times 1024]$ feature vector. To form the inputs of this sub-net, we follow a sparse uniform sampling policy on the input sequence. Regarding the case of visual scene representation in the two datasets of interest, every single action of the action sequence that represents the activity has been encoded using a segment-wise temporal encoder network¹. Therefore, it corresponds to the feature-based representation of a segment formed based on the adjacent frames. This formulation scheme of the sub-net input enables the modeling of both short- and long-term appearance variations of the scene elements in the learning process.

6.2.2 Linguistic Action History Module

We argue that the knowledge of the preceding action occurrences, noted as action history, is important for learning to estimate at a certain time in the video, the label of the next-anticipated action (*action forecasting/anticipation*) or of the active object in that action, as it provides efficient, discriminative features to opt among potential candidate targets. We address this issue using a compact textual description of the preceding actions, in the compact form of action labels, compared to captions that feature extensive textual descriptions of actions. The sentence-based textual description of the preceding actions is processed using the NLP sub-network that comprises a Word Embedding layer followed by the same layer set as the branches of the action-centric visual module. This representation forms a $[1 \times 256]$ feature vector, which is concatenated with the representation of the action-centric module. The combined representation is then forwarded to a set of two FC layers to provide estimations on the next action/object class.

Delving into this representation of the action history, we restructure each label (length, semantic complexity, part-of-speech element position (verb, noun, adverb)), in a specific lexical format depending on the task at hand (action, motion verb, or noun anticipation), to facilitate the

¹For example, in Assembly-101 each action instance has been encoded using the effective Temporal Shift Module (TSM) [75]

learning process. Specifically, in the case of the verb (coarse motion motif) or noun (next-segment active-object) anticipation, we may have to deal with actions of a similar motion and object basis but of a different type of object upon which the action is performed. For example, consider the actions, *screw a screw with hands* and *screw a screw with screwdriver*. When asked to predict the key object(s)² of the next anticipated action, the action history module should maintain the key objects of the preceding action segments, and therefore the knowledge that the tool-medium is of no importance in this coarser anticipation problem. A similar convention is also considered for the task of predicting only the coarse motion motif label for the next action.

Under this premise, for the tasks of verb/noun next-segment prediction we restructure the available lexical information/labels of actions by discarding parts of the labels that refer to the usage of extra tools (annotated as nouns) to implement the corresponding action, i.e. the action labels are restructured to follow the format *action verb + noun*. In fact, this meta-processing of action labels that allows for a decoupled prediction of the next action verb or next action object(noun), is a common practice followed by the recent datasets targeting isolated motion motif or next-segment object prediction (e.g. Assembly-101 [136]). If such an action label format is available for the dataset in question, this label restructuring is skipped. The gain from such lexical decomposition is that the prediction task becomes simpler since the number of classes decreases, due to the fact that labels sharing the same action verb or action object (noun) are being merged, which allows for more samples to be associated with the specific motion motif or object state. Finally, in the case of the next action prediction (entire action context), we do not restructure the initial labels since the entire context of the preceding action labels is required to disambiguate between actions that share the same motion and object characteristics but differ on the execution medium.

6.2.3 Visual Action Recognition Module

The two aforementioned modules can be regarded as independent action anticipation models. In addition, a visual action recognition model is incorporated independently which during the inference stage operates on the same input sequence, denoted as x_t , as the action y_t . The purpose of this model is to provide estimates specifically for the current action y_t and fill the language-based action history.

²As key objects we refer to objects that affect the outcome of the activity, e.g. in a toy assembly activity on the parts that can alter the result.

Since the purpose of this model is to fill the action history, it remains independent from the action anticipation modules without any influence or connection, it can be trained separately and applied during the inference stage of the framework. In this work, instead of developing and training an action recognition module from scratch, we leverage the capabilities of state-of-the-art (SOTA) action recognition models that have been documented in the existing literature for each dataset. This approach is motivated by our objective to construct a visual-linguistic action anticipation framework, which can benefit from the advancements achieved by action recognition models specific to each dataset, thereby enhancing its overall performance.

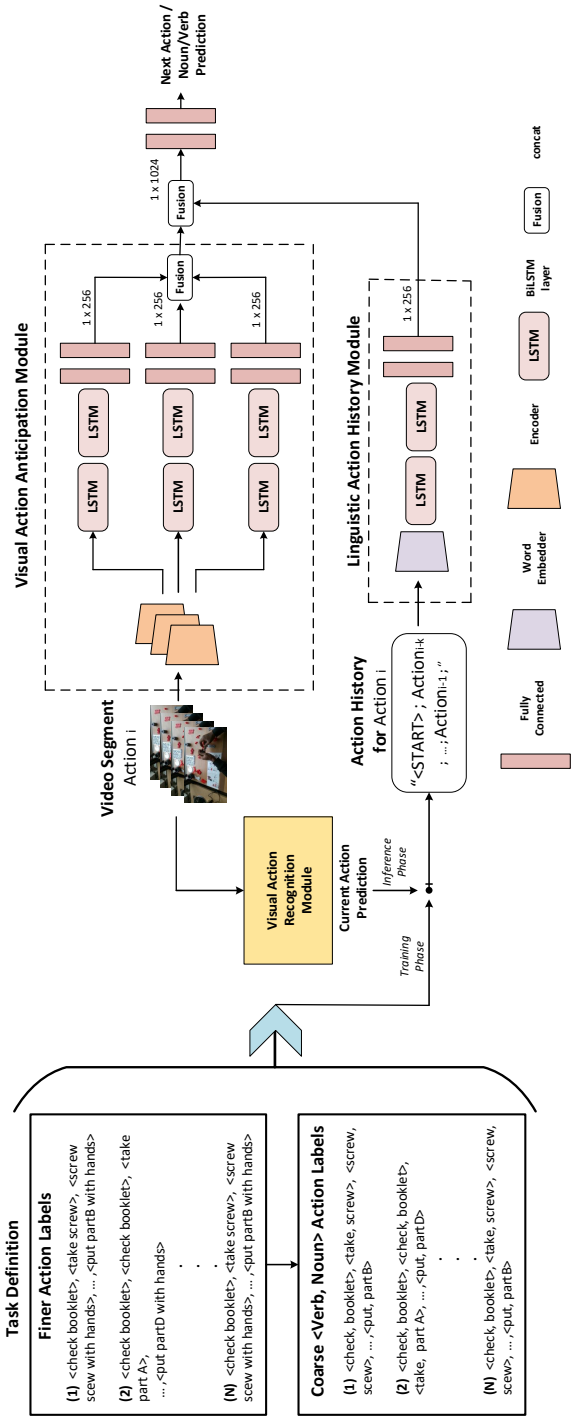


Figure 6.2: The proposed VLMAH architecture. The Visual Action module and the Linguistic Action History module are presented. For the *Meccano* dataset, the encoders of the action module, generate Object, Hands, Gaze representations, whereas for the *Assembly-101* dataset, there is a single encoder network, TSM [75] while representations are split into 3 sub-sequences, as mentioned in Section 6.3.2. The detail level regarding the textual label descriptions is adaptable to the anticipation task at hand (action, motion motif (verb), or object (noun)). The final format also includes two special labels (START, END) that indicate the start and end of the action history sequence.

6.3 Experimental Setup

We evaluate the proposed framework on three popular datasets of procedural activities. The main characteristics of the datasets are described in this section, such as the target activities, camera viewpoints, annotation data as well as multi-modal data and features provided (Section 6.3.1), followed by the evaluation protocols.

The experimental evaluation for the proposed framework follows a two-way narrative. Firstly, the generation of the action history module involves simulating the prediction scores of a realistic action recognition model on a given dataset. This step aims to showcase the performance of the proposed model in relation to the latest advancements for each dataset. Subsequently, the complete potential of the proposed model is presented by populating the action history module with past predictions obtained from an ideal visual action recognition model for each respective dataset. We should note that the realistic visual action recognizer performance follows the current SOTA action recognition scores reported for each examined dataset. Finally, we conduct experiments regarding different portions of the linguistic action history in order to assess the effect of the different sizes of the action history on the anticipation capabilities of the proposed framework.

6.3.1 Datasets

Meccano [116] is a multi-modal egocentric dataset created to study the interactions of humans and objects in industrial settings during instructional activities. Twenty different participants were requested to build a toy model of a motorbike. There exist 20 object classes, which include 16 classes that categorize 49 different toy components, 2 tool classes namely the screwdriver and the wrench, the instructions booklet, and a special class, noted as a partial model, for the under-construction toy object. Also, the dataset contains 12 verb classes and 61 action classes. In total, 20 videos are provided, 11 of which are used for training while the rest 9 videos are used for validation and testing.

The main goal for generating the dataset is to capture the human behavior from an egocentric perspective to gain knowledge about the interactions in industrial settings during instructional activities. The dataset provides multimodal data captured simultaneously with a custom headset that consist of RGB videos, depth maps and gaze signals and annotations tailored to the needs of five human behavior analysis tasks namely: action recognition, active objects detection and

recognition, egocentric human-objects interaction detection, action anticipation and next-active objects detection. Twenty (20) different participants were requested to build a toy model of a motorbike. There exist 20 object classes, which include 16 classes that categorize 49 different toy components, 2 tool classes namely the screwdriver and the wrench, the instructions booklet and a special class, noted as partial model, for the under-construction toy object. Also, the dataset contains 12 verb classes and 61 action classes. In total, 20 videos are provided, 11 of which are used for training while the rest 9 videos are used for validation and testing.

Meccano dataset provides gaze, object-centric features, and hands-centric features. The former type of features are computed based on the occurrences of the objects in each frame following the work of [36, 35]. Gaze features have been obtained by weighting the object-centric features with the distance between the center of objects bounding boxes and the gaze position in the image. The hand annotations of the dataset that contain the bounding boxes of both hands were used as hands-related features.

Assembly-101 [136] is a large-scale video dataset for the analysis and understanding of procedural activities regarding assembling and disassembling 101 "take-apart" toy vehicles captured from multiple viewpoints. In total 362 unique data sequences were captured synchronously by 4 egocentric and 8 static cameras and annotated with more than 100K coarse and 1M fine-grained action segments, targeting the challenging tasks of action recognition, action anticipation, temporal action segmentation, and mistake detection. Participants were instructed to disassemble and then assemble a toy vehicle without any instructions, which enhances the variability of the temporal ordering of actions performed by the participants during the procedural activities. A set of 90 object classes is considered that includes 5 tools together with the "hand". Also, 24 verbs are included along with the object classes form 1380 fine-grained action classes. A 60% of the available videos is used for training, while the rest 15% and 25% are utilized for validation and testing, respectively. Of the 101 toys, 25 of them are shared between all splits which sets the dataset even more challenging. For the RGB input, 2048-D frame-wise features are calculated using TSM [75] with an 8-frame input.

50Salads [140] is a multi-modal third-person instructional dataset of cooking-related activities. Each one of the twenty-five different participants is preparing two mixed salads. The dataset provides RGB videos, depth maps, accelerometer data, and high-level and low-level activity

annotations. The dataset consists of 17 action classes. We report top-1 accuracy averaged over the 5 pre-defined splits following the work of [121].

6.3.2 Training, Testing & Input Configurations

As noted in Section 6.2, the structure of the action-centered temporal modeling sub-network follows a three-branch design, that acquires three vision-centered input sequences.

For the Mecanno dataset [116], input refers to the available feature representations for a) *Gaze*, b) *Objects*, and c) *Hands*.

For the Assembly-101 [136], the available TSM [75] features for the RGB videos are utilized, which refer to frame-wise $[1 \times 2048]$ feature vectors. We restructure this representation to fit in the action-centric visual anticipation sub-net, as follows: a) split feature vectors into a set of two $[1 \times 1024]$ feature vectors to drive input to the first two branches and b) uniform sub-sampling is applied on the feature vector of the current frame of size $[1 \times 2048]$ into a $[1 \times 1024]$ and then calculate discrepancies between the sub-sampled feature representation of the previous frame to form the input feature vector for the third branch.

For 50Salads [140] we utilized pre-extracted I3D features from [29, 121], which correspond to frame-wise $[1 \times 2048]$ feature vectors, which were restructured in the form described for the ones of the Assembly dataset.

Regarding the training configurations, the batch size was set to 4 for all datasets. The loss minimization is performed using the Adam optimizer, with a learning rate of 0.001. The input sequence length was set to 8 frames, while a random clip cropping sampling scheme was utilized. During the inference phase, we simulated the performance of the realistic visual action recognizer, by exploiting the SOTA performance of SlowFast [31] for Meccano, with 49.66% top1 accuracy, of TSM [75] for Assembly101 with 39.2% top1 accuracy, and, of Therbligs [23] for 50Salads with 76.5% top1 accuracy.

6.4 Experimental Results

6.4.1 Action Anticipation

Predicting future actions is a challenging task, while modeling and performance greatly depend on the designated time horizon of the predictions. More specifically, predictions can be made

at different decision points in time (timesteps) prior to the start of the next segment. In order to establish an extensive performance assessment of the proposed framework, we adopt the evaluation protocol reported in Furnari et al. [35] where predictions are made at 8 different anticipation timesteps before the start of the near-future anticipated action. Noted as τ_{ant} , the set of anticipation time refers to discrete values in the range of $[2s, 0.25s]$ for a timestep of $0.25s$, while the upper limit of this interval, that is $0.25s$ is closest to the start of the anticipated action.

Meccano: For the prediction of each action, the input to our framework regards information originating from the selected anticipation time point and runs backward, toward the start of the video (see Figure 6.1). As described in the previous sections, we exploit visual information related to the recent past (visual-action module) for modeling the short-term action history and the long-term past based on the use of the linguistic action history module. We report the Top-1 and Top-5 accuracy of the predicted action of the next segment, according to the [116]. In this work, the authors proposed the RULSTM framework [35] to anticipate the next action. We employ the publicly available code³ of RULSTM for Meccano to replicate the experiments and also provide accurate results for the prediction of the noun and the verb of the next action-segment. We utilized a combination of information based on gaze, object-centric and hand-centric features that are provided by [116], as those are the most discriminative features according to their experimental evaluation.

³<https://github.com/fpv-ip1ab/MECCANO>

		Top-1 / Top-5 Accuracy% @ different T_{ant}									
		Timesteps									
Method		2s	1.75s	1.5s	1.25s	1s	0.75s	0.5s	0.25s		
Meccano [116]		30.89/65.14	30.50/65.11	30.99/66.17	30.85/65.92	30.53/66.49	31.10/67.06	31.10/67.84	31.24/70.00		
VLMah		33.12/77.85	32.12/77.78	31.48/78.49	32.33/80.41	31.25/76.30	32.17/82.39	34.07/78.58	38.34/79.19		
$VMAH_{GT}$	Noun	15.91/72.58	27.63/69.46	25.37/65.83	28.93/73.29	26.21/70.31	25.08/71.73	28.83/69.81	29.50/70.88		
$VLMah_{GT}$		37.57/79.40	41.33/82.88	35.09/80.75	35.65/79.33	39.35/82.31	40.55/84.94	39.55/81.24	40.63/80.54		
Meccano [116]		36.06/93.19	35.11/93.01	34.96/92.98	35.92/93.19	35.32/93.38	35.39/93.62	34.75/93.76	35.00/93.83		
VLMah		36.35/93.00	35.42/92.33	35.61/91.31	35.96/92.88	36.73/91.08	36.30/90.62	37.19/91.14	39.06/90.93		
$VMAH_{GT}$	Verb	25.71/87.85	29.75/87.64	25.71/88.06	29.11/89.48	27.48/87.99	25.92/85.51	25.78/86.57	31.25/84.30		
$VLMah_{GT}$		40.76/91.40	41.26/93.39	40.83/92.61	43.39/92.96	39.91/91.69	40.98/93.18	43.67/92.68	43.55/91.65		
Meccano [116]		23.37/54.65	23.48/55.99	23.30/56.56	23.97/57.73	24.08/58.23	24.50/59.96	25.60/61.31	28.87/63.40		
VLMah		24.75/54.23	24.35/55.16	24.22/53.09	22.79/53.98	28.90/58.13	25.29/53.16	26.47/56.71	29.12/58.01		
$VMAH_{GT}$	Action	27.20/49.08	28.91/51.63	26.99/48.57	28.98/52.20	28.62/50.49	26.99/49.94	27.77/49.86	28.03/51.70		
$VLMah_{GT}$		34.73/67.75	36.86/69.53	35.01/67.18	34.30/69.24	35.15/68.25	33.59/67.89	34.65/66.90	33.09/65.98		

Table 6.1: Action anticipation accuracy for different timesteps (prior to the beginning of the next segment) for the **Meccano dataset**. $VLMah_{GT}$ and $VMAH_{GT}$ represent the two variants of the proposed method when *ground truth annotations* are used as the linguistic action history. VLMah makes use of the Linguistic Action History module while the action history is generated from the visual action recognition module. The comparison is between the [116] and the VLMah methods.

We evaluate the proposed VLMAH framework for action forecasting using different anticipation timesteps (see Table 6.1), and under the use of a realistic and an ideal (oracle) action predictor (denoted as VLMAH and VLMAH_{GT} respectively) for past actions that populate the action history subnet. Under the use of a realistic visual action recognizer for past actions, our framework is compared to [116] which is the baseline and currently the SOTA method for the Meccano dataset. Our method outperforms the SOTA in Top-1 accuracy for the noun, verb and action scenarios for almost every anticipation time, by a considerable margin. We present to have a slight decrease in performance in the Top-5 accuracy for the verb and action scenarios. This happens due to the impact of the action recognizer in the linguistic action history from which we draw information for making predictions. Our accuracy margin increases significantly from 4.1% up to 9.05% if we consider an ideal (oracle-like) visual action recognizer that feeds the linguistic action history module with the true past action classes. This also means that any improvement in action recognition accuracy is expected to directly benefit action anticipation, too.

Assembly101: In [136] that have also introduced the Assembly-101 dataset, action anticipation is performed at the fixed timestep $\tau_{ant} = 1s$. To assess action anticipation performance in [136], the TempAgg [134] method is used⁴. Both the VLMAH and the TempAgg methods are trained to generate predictions at anticipation time $\tau_{ant} = 1s$ that are evaluated using the Top-1 and Top-5 accuracy measures. Since the test split of the dataset is not yet available, we train and test both methods on the training and validation splits, respectively, using the egocentric viewpoint and data captured by the *e4* camera which yields the best results according to the experiments reported in [136]. Both the proposed VLMAH and the TempAgg methods have been trained/tested on data captured by this specific viewpoint.

Table 6.2 presents the accuracy results at $\tau_{ant} = 1s$. We provide two results for our framework. We compare our work with the state-of-art on Assembly-101 dataset, the TempAgg [134] framework. Our work is a single-task learning framework so for a fair comparison we test TempAgg [134] under two learning settings, a multi-task and a single-task. The single-task setting is denoted with * in Table 6.2. The proposed approach outperforms state-of-the-art performance for the verb, noun, and action predictions by a large margin for this large and challenging dataset, even in the case that the linguistic action history module is not used. In particular, by using a

⁴Code provided online at <https://github.com/assembly-101>

Top-1/Top-5 Accuracy% @ $\tau_{ant} = 1s$			
Method	Noun	Verb	Action
TempAgg [136]	17.19 / 55.65	24.20 / 75.38	08.62 / 27.73
TempAgg [136]*	18.99 / 57.29	28.52 / 77.16	09.00 / 29.79
VLMAH	27.70 / 54.37	42.17 / 82.52	14.18 / 30.95
VMAH _{GT}	22.68 / 55.32	40.59 / 85.11	13.14 33.98
VLMAH _{GT}	55.27 / 83.89	61.12 / 93.03	34.26 58.89

Table 6.2: Top-1/Top-5 accuracy results of [136] and the VLMAH variants on the **Assembly-101 dataset** for anticipation time $\tau_{ant} = 1s$, with or without the use of the linguistic action history module. TempAgg* denotes the single-task learning variant.

Top-1 Acc% @ $\tau_{ant} = 1s$	
Method	Action
DMR [147]	06.20
RNN [28]	30.10
CNN [28]	29.80
TempAgg [134]	40.70
AVT [41]	48.00
VLMAH	<u>43.58</u>
VLMAH _{GT}	<u>55.49</u>

Table 6.3: Top-1 accuracy results on the 50Salads dataset for the anticipation time $\tau_{ant} = 1s$.

realistic visual action recognizer to populate the action history module, an increase in accuracy of 13.65% for the verb prediction, 8.71% for the noun prediction, and 5.18% for the action prediction for $\tau_{ant} = 1s$ was reported. Similarly to *Meccano*, the use of an oracle-like visual action recognizer to verify/correct past estimates in the history module further increases the action anticipation performance of the proposed method. Even if we use only the visual information (VMAH), we outperform the TempAgg* framework in general for a minimum of 4% up to 12%.

50-Salads: In Table 6.3 we present the accuracy scores at $\tau_{ant} = 1s$, and compare our proposed framework with recent works that tackle action anticipation in this dataset. We can observe that under the use of a realistic action, recognizer to validate/correct the past action estimates stored in the action history module, our method is only surpassed by AVT [41] ($\approx 4\%$), with our proposed action anticipation method however, having a vastly lower number of trainable parameters, and ease in adapting/incorporating the current action recognition advancements in each dataset.

6.4.2 How much history is enough?

In this study, we conducted ablation analyses to evaluate the performance of our proposed framework under various scenarios that pertain to the linguistic action history module’s role and the required amount of linguistic action history to enhance the predictability power of the framework. Despite the fact that action history can obtain long-term information faster and with less cost compared to the visual features one question to be answered is “*how much history is enough?*”. To answer this we evaluate our framework on the Assembly-101 dataset with different lengths of linguistic action history. From the previous sections, we have acquired the results of the evaluation of our framework with the full linguistic history of the observed actions⁵. In this experiment, we assess our framework by reducing the linguistic history to different percentages. The history percentages are in the range from 0% to 100%. Zero percent indicates the use of the VMAH_{GT} framework while all the other percentages imply the use of the VLMAH_{GT} framework with different percentages of action history. In this experiment, we use the VLMAH_{GT} instead of VLMAH in order to assess the effect of the available size of action history in case no errors from the Visual Action Recognition module are present in the action history. As seen in Table 6.4, the results differ between the action and the verb/noun predictions considering different amounts of observed history.

Initially, all experiments were performed using 100% of the textual action history, which referred to a memorization capacity of 854 actions (slowest assembler). Our experiments show that, for the task of fine-grained action anticipation (full label), considering the entire linguistic history was the best strategy since it allowed us to disambiguate between cases of candidate actions that exhibited high similarity in their preceding action history.

In contrast, for the prediction of the coarse-grained verb and noun classes our experiments indicate that considering a more recent history is the best strategy. We observe that considering a larger percentage of the action history on these cases introduces noise that results in a considerable decrease in prediction accuracy, potentially due to similarities in the sequence of verb/noun transitions between different assembly scenarios. This is a valid assumption since, as stated in Section 6.2.2, in these tasks the initial action labels were restructured into a two-part-of-speech label (*verb+noun*). This way, we discarded the fine-grained context of the label that

⁵A full history refers to the number of actions the slowest assembler from the training set performed to complete the assembling task.

Top-1/Top-5 Accuracy% @ $\tau_{ant} = 1s$			
History	Noun	Verb	Action
0%	22.68 / 55.32	40.59 / 85.11	13.14 / 33.98
1%	56.98 / 83.35	62.33 / 92.78	28.49 / 53.69
12.5%	56.86 / 84.08	62.83 / 93.40	28.20 / 51.38
25%	53.86 / 83.03	62.92 / 93.06	28.96 / 53.15
50%	56.92 / 84.54	63.99 / 93.16	27.13 / 51.02
75%	56.19 / 84.53	63.20 / 93.21	29.83 / 53.75
100%	52.16 / 83.81	61.12 / 93.03	34.51 / 58.44

Table 6.4: The Top1 and Top5 accuracy scores achieved by the proposed framework using variable lengths of the linguistic action history on the **Assembly-101 dataset**. Zero percent (0%) is equivalent to the use of VMAH_{GT} variant, while other action history percentage values refer to the use of the VLMAH_{GT}.

refers to the mediums (tools) utilized to perform the action. For example, in the case of the action label pair “*screw cabin with screwdriver*” and “*screw cabin with hands*”, which are two different action classes, the restructuring operation merged the two classes into the action “*screw cabin*”. We note that in Assembly-101 similar format is provided as annotation data.

6.4.3 History-based action anticipation

Our goal is to experimentally evaluate the impact that the linguistic action history contributes to our framework. So, for this experimental analysis, additionally to our VLMAH, we employ the VMAH variation which considers only video features of the current action as input. We assess all variations on the Meccano dataset for different types of features. The features used are RGB features, flow features, gaze features, object-centric and hand-centric features. The gaze features, object-centric and hand-centric features are described in detail in Section 6.3.1. Using the RGB modality we extracted features using the C3D [144] and TSN [151, 150] frameworks and we evaluate them separately. Also, flow features were extracted using the TSN framework. All TSN-based features have been extracted using the specifications of the authors in [35]. For this experiment, we used all the available visual and linguistic information of the observed actions. This implies that we take into consideration all the available frames prior to the start of the next segment **without** leaving an anticipation time $\tau_{ant} = 0.25s$.

As observed in Table 6.5, regardless of the type of features employed, the linguistic action history module contributes significantly to the predictability and overall performance of the method. In particular, for any given feature or combination of features, the consideration of the

Noun - Top-1 / Top-5 Accuracy% @ $\tau_{ant} = 0s$		
Method	Features	Top1 / Top5
VMAH	RGB-C3D	29.0 / 73.6
VLMAH	RGB-C3D/NLP	33.2 / 81.6
Improvement		4.2 / 8.0
VMAH	Obj	28.8 / 72.7
VLMAH	Obj/NLP	34.3 / 80.8
Improvement		5.5 / 8.1
VMAH	Hands	28.1 / 73.5
VLMAH	Hands/NLP	28.6 / 76.2
Improvement		0.5 / 2.7
VMAH	Gaze	22.5 / 72.3
VLMAH	Gaze/NLP	30.3 / 80.8
Improvement		7.8 / 8.5
VMAH	RGB-TSN	30.0 / 72.8
VLMAH	RGB-TSN/NLP	36.0 / 79.4
Improvement		6.0 / 6.6
VMAH	Flow-TSN	29.9 / 72.3
VLMAH	Flow-TSN/NLP	31.8 / 81.4
Improvement		1.9 / 9.1

Table 6.5: Top-1 and Top-5 accuracy results for **noun anticipation on the Meccano dataset** of our VLMAH framework with different types of features and with the use of the visual action history (VMAH) module. The linguistic module contributes significantly to the anticipatory capability of the framework.

linguistic action history always improves performance, considerably. The RGB information seems to be better encoded through the TSN framework and have the best performance overall. The object features and the RGB features extracted from C3D also exhibit good accuracy performance. Finally, the gaze features have the highest accuracy gain when combined with the linguistic action history.

6.4.4 Noise resistance

Our framework performs action forecasting making use of the ground truth semantic labels of the observed segments/actions. This experiment aims to explore the anticipatory capability of our framework in the presence of different levels of noise in the estimated labels throughout the action history, in other words given the expected errors in labels of the anticipated actions across the duration of a long, procedural activity, as if our method functions in an online manner. For this experiment, we consider the Assembly-101 dataset for anticipation time $\tau_{ant} = 1s$ setting the action history module operating at full capacity. The percentage values for the noise level added to the history labels are: 5%, 10%, and 20% respectively, indicating the number of observed

Top-1/Top-5 Accuracy% @ $\tau_{ant} = 1s$			
Noise	Noun	Verb	Action
0%	52.46 / 83.81	61.12 / 93.03	34.51 / 58.44
5%	51.40 / 80.95	60.38 / 92.15	33.15 / 56.73
10%	49.05 / 79.26	59.80 / 91.76	31.45 / 53.94
20%	44.69 / 75.96	54.98 / 89.61	27.77 / 50.18

Table 6.6: Evaluating the anticipatory capacity of the VLMAH framework on the **Assembly-101 dataset** for variable noise levels (disturbances) in the form of erroneous semantic labels in the action history.

segments that have been miss-classified with respect to the total number of segments.

Table 6.6 presents the findings obtained in the presence of three different levels of label noise in comparison with the results obtained when using ground truth labels (noise is equal to 0%). Our analysis revealed that in the presence of the highest noise level (20%), the accuracy of our framework decreased by less than 10% across all categories of verb, noun, and action labels. Despite these challenging conditions, our model demonstrates its robustness and effectiveness in handling label noise.

6.5 Summary

The primary objective of this study was to assess the impact of incorporating low-cost language features of all past actions with the vision features of the recent past. The language features act as supplementary information augmenting the network with the ability of memorizing past actions while enhancing the predictive capabilities beyond the visual cues of the current observed action. Through a comprehensive investigation, we delved into various facets concerning the effectiveness and resilience of this approach. We particularly focused on its performance in scenarios with varying degrees of past action misclassification rates and the influence of the length of encoded action history on the anticipation task, encompassing action, motion motif (verb), and object (noun) predictions.

Our rigorous experiments unveiled the manifold advantages inherent to this novel strategy, which has resulted in a significant advancement in comparison to existing methodologies. Notably, our approach demonstrated remarkable enhancements in achieving state-of-the-art scores across two complex video datasets featuring intricate procedural activities. The implications of these findings are profound, indicating the transformative potential of our strategy in revolutionizing

the field of action anticipation.

Furthermore, our investigation underscored the robustness of the language-driven history-logging mechanism, even when subjected to stringent constraints on memorization capacity and substantial rates of past action misclassifications. This resilience highlights the adaptability and effectiveness of the proposed approach in real-world scenarios where errors and uncertainties are inevitable.

In summary, our study has not only contributed to the expansion of knowledge in the domain of action anticipation but has also introduced a groundbreaking paradigm by integrating language-driven history-logging into the anticipation process. The substantial performance improvements observed across diverse datasets and challenging conditions validate the efficacy of this approach, emphasizing its potential to reshape the landscape of predictive modeling in action-based contexts.

CONCLUSIONS

7.1 Contributions

In the presented works we tackled the problems of action prediction, action anticipation, and next-active-object(s) prediction. Starting from the action prediction problem which entails deducing the ongoing action’s class label while it’s still unfolding. The study focuses on scenarios where humans interact with objects, utilizing trimmed video recordings as input to extract 3D skeletal data. The extent of observable action is determined by the observation ratio $[0, 100\%]$.

To tackle this challenge, our study presents an approach that constructs time series representations for both human and object interactions during ongoing actions. These representations are matched with prototype actions using a Dynamic Time Warping (DTW)-based framework for time series alignment. This DTW-driven alignment helps identify the best-matching relationship between the ongoing action and prototype actions. The approach is rigorously evaluated on three benchmark datasets, revealing the significance of blending human-centered and object-centered action representations for prediction accuracy. Additionally, the study demonstrates that their approach outperforms competing methods significantly, leading to notably improved action prediction accuracy.

Moving forward we tackled the problem of action prediction by aligning incomplete action sequences with complete ones and introduced innovative algorithms to address this issue. While established alignment methods like Dynamic Time Warping (DTW) and Soft Dynamic Time

Warping (S-DTW) excel at aligning complete sequences, they struggle with incomplete sequences. To tackle this, we introduced two new algorithms, Open-End Soft DTW (OE-S-DTW) and Open-Begin-End Soft DTW (OBE-S-DTW). These algorithms combine the partial alignment abilities of OE-DTW and OBE-DTW with the soft-minimum operator of S-DTW, resulting in improved and differentiable alignment, especially for continuous, unsegmented actions. These algorithms hold promise for action prediction, aligning ongoing incomplete actions with complete prototypes to predict subsequent actions.

The algorithms were tested on various datasets, including MHAD, MHAD101-v/-s, MSR Daily Activities, and CAD-120. Results indicate that the proposed algorithms surpass the performance of relevant state-of-the-art approaches, making them effective tools for addressing the challenge of action prediction involving incomplete sequences.

Our next approach towards the solution of the action prediction problem focuses on predicting the semantics of partially observed activities and the next active objects required to complete the ongoing activity. The proposed approach aims to model spatio-temporal relationships between humans and visible scene objects to predict the classes of multiple upcoming active objects involved in the activity's completion.

Existing methods can't predict more than one next active object, but the study introduces the first approach capable of jointly predicting the ongoing activity's semantics and multiple next active objects. Additionally, the study highlights the importance of forecasting the time at which these next active objects will be involved in the scenario. The proposed method not only predicts the next active objects but also estimates the time of their involvement in the activity, making it a novel and comprehensive approach in the field.

Finally, for the solution to the action anticipation problem, we introduced a novel approach called Visual-Linguistic Modeling of Action History (VLMAH). The aim is to improve action anticipation by considering the history of all executed actions in lengthy procedural activities. VLMAH combines visual features from the immediate past with linguistic constructs representing semantic labels of nouns, verbs, or actions from the distant past. This fusion of short- and long-term past information enables accurate predictions of near-future actions during procedural activities.

The approach is evaluated extensively on three challenging video datasets containing procedural activities: Meccano, Assembly-101, and 50Salads. The results demonstrate that utilizing

long-term action history leads to enhanced action anticipation and improves the Top-1 accuracy compared to state-of-the-art methods. In summary, the proposed VLMAH approach effectively addresses the task of action anticipation by leveraging both short- and long-term action history information for more accurate predictions during procedural activities. The code will be made available ¹.

7.2 Impact

The international research community has thoroughly assessed and reviewed the current study, documenting and publishing its findings through peer-reviewed papers in scientific journals and conference proceedings.

- Publications

- Thesis-related publications

- * **Manousaki Victoria**, Konstantinos Papoutsakis, and Antonis Argyros. "Action prediction during human-object interaction based on DTW and early fusion of human and object representations." International Conference on Computer Vision Systems. Cham: Springer International Publishing, 2021.
 - * **Manousaki Victoria**, and Antonis A. Argyros. "Segregational Soft Dynamic Time Warping and Its Application to Action Prediction." VISIGRAPP (5: VISAPP). 2022.
 - * **Manousaki Victoria**, and Antonis A. Argyros. "Partial Alignment of Time Series for Action and Activity Prediction", Springer Book of VISAPP 2022, (to appear), Springer, 2023.
 - * **Manousaki Victoria**, Konstantinos Papoutsakis, and Antonis Argyros. "Graphing the Future: Activity and Next Active Object Prediction using Graph-based Activity Representations." International Symposium on Visual Computing. Cham: Springer International Publishing, 2022.
 - * **Manousaki Victoria**, Bacharidis Konstantinos, Papoutsakis Konstantinos and Antonis A. Argyros ".VLMAH: Visual-Linguistic Modeling of Action History for

¹www.project.ics.gr/projects/cvrl/vlmah

Effective Action Anticipation". Eleventh International Workshop on Assistive Computer Vision and Robotics, ICCVW, 2023.

– Other publications

* Asvestopoulou Thomais, **Manousaki Victoria**, et al. "Towards a robust and accurate screening tool for dyslexia with data augmentation using GANs." 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2019.

* Asvestopoulou Thomais, **Manousaki Victoria**, et al. "Dyslexml: Screening tool for dyslexia using machine learning." arXiv preprint arXiv:1903.06274 (2019).

• Scholarships

– **IKY Scholarship Programme:** I am honored to have been selected as a recipient of the prestigious IKY Scholarship Programme for Ph.D. candidates studying in Greek Universities. This scholarship not only recognizes my academic dedication but also provides me with invaluable financial support to further pursue my research endeavors.

– **Hellenic Foundation for Research and Innovation (HFRI) PhD Fellowship:** Being awarded a scholarship from the Hellenic Foundation for Research and Innovation is a significant achievement in my academic journey. The HFRI PhD Fellowship, has a designated grant number 1592. This recognition from a reputable foundation motivated me to continue pushing the boundaries of research and innovation during my Ph.D. studies.

– **Computational Vision and Robotics Laboratory at FORTH:** I am grateful to the Computational Vision and Robotics Laboratory at the Foundation for Research and Technology - Hellas (FORTH) for granting me a scholarship. This opportunity not only provides me with financial assistance but also allows me to be a part of a dynamic research environment. The laboratory's expertise in cutting-edge technologies undoubtedly enriched my academic experience and contributed to the advancement of my research in the realms of computer vision and robotics.

These scholarships collectively acknowledge my dedication and contribution to the field of research in the Computer Vision area. I am excited to leverage these opportunities to further my studies, collaborate with experts, and make a meaningful impact in my chosen field.

7.3 Limitations

The proposed frameworks for action prediction are built upon aligning videos and sequences of motion capture data temporally. The extent to which the proposed methods can effectively scale in datasets containing a substantial number of activity categories (more than a hundred) requires further exploration.

Several parameters within the proposed methods are adjusted manually. Particularly, the λ parameter in the Equation 5.1, which signifies the contribution of semantic and motion-based data from tracked body joints and interacting objects, is defined by the user. Exploring an automated approach to tune or discover an optimal parameter configuration based on the specific activity category holds promise for enhancing the performance and user-friendliness of the proposed framework.

It is worth noting that the suggested method for evaluating video similarity through the bipartite Graph Edit Distance approach doesn't consistently generate object correspondences that carry semantic significance, aligning with interpretable action-relevant information between the compared videos. In simpler terms, while the GED-based scores between videos yield favorable outcomes for tasks like video matching, ranking, or classification, the clarity of results in terms of explicable object correspondences isn't always evident. Given the availability of one or more extensive video datasets with appropriately detailed ground truth data for training, adopting a supervised learning strategy becomes a promising avenue to fine-tune the weights and parameters of the proposed framework.

In essence, the introduced VLMAH framework, designed to tackle the challenges of action anticipation, showcases its capability to manage large datasets and forecast forthcoming segments encompassing verbs, nouns, and actions. Nonetheless, a constraint of this framework emerges from its dependence on the history of previously executed actions. The accumulation of errors and misclassifications from past actions holds the potential to significantly impact the accuracy of predictions, highlighting the importance of ongoing error management and correction mechanisms.

7.4 Directions for future work and research

In the next stages of our research, we will immerse ourselves in a concerted effort to assemble and meticulously curate more expansive and intricate datasets that capture the intricate dynamics of human-object interactions. These datasets will encompass a diverse array of objects and scenarios, ensuring that our analyses encapsulate a comprehensive spectrum of real-world complexities. By delving into these enriched datasets, we aim to unlock deeper insights into the nuances of action prediction within dynamic environments.

Furthermore, our exploration will extend to probing the profound impact of incorporating historical action sequences into our predictive framework. We will rigorously investigate how the integration of past actions can enhance the accuracy of long-range action anticipation. As part of this inquiry, we will meticulously examine the influence of the temporal positions of misclassifications, delving into both short-term and long-term past actions, and dissecting their repercussions on the precision of action anticipation. This comprehensive analysis will empower us to refine our methodology and devise strategies to mitigate the effects of misclassifications on predictive outcomes.

Excitingly, we are interested in integrating Large Language Models (LLMs) into our research framework in order to explore datasets that alongside the visual information provide texts that contain large amounts of information. This infusion of language understanding capabilities will empower our system to leverage linguistic context and semantics, opening new avenues for refining action anticipation accuracy and understanding the broader context of human-object interactions.

Moreover, our enthusiasm drives us to venture into a new and promising direction—the prediction of the future state of active objects. By extending our predictive capabilities beyond actions to the state of objects, we seek to unravel an additional layer of predictive insight. This novel trajectory holds the potential to revolutionize how we perceive and anticipate interactions in dynamic environments.

In summary, our future endeavors are deeply rooted in the pursuit of enriched datasets, enhanced historical integration, advanced language understanding through LLMs, and the pioneering exploration of predicting object states. These multifaceted research avenues collectively represent our commitment to advancing the frontiers of action prediction and understanding within complex interactive scenarios.



Operational Programme
**Human Resources Development,
Education and Lifelong Learning**
Co-financed by Greece and the European Union



7.5 Acknowledgements

- Partial funding for this work was provided by the Computational Vision and Robotics Laboratory at the Foundation for Research and Technology- Hellas (FORTH) (Period: 10/2017 - 09/2019)
- The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 1592.) (Period: 10/2019 - 02/2022)
- The implementation of the doctoral thesis was co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the Act “Enhancing Human Resources Research Potential by undertaking a Doctoral Research” Sub-action 2: IKY Scholarship Programme for PhD candidates in the Greek Universities (Period: 30/5/2022 - 30/9/2023)

BIBLIOGRAPHY

- [1] Zeina Abu-Aisheh et al. “An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems”. In: *ICPRAM*. 2015.
- [2] Yazan Abu Farha et al. “Long-term anticipation of activities with cycle consistency”. In: *DAGM German Conference on Pattern Recognition*. 2020.
- [3] R. Alfaifi and A. Artoli. “Human Action Prediction with 3D-CNN”. In: *SN Computer Science* (2020).
- [4] Md Zahangir Alom et al. “The history began from AlexNet: a comprehensive survey on deep learning approaches”. In: *arXiv preprint arXiv:1803.01164* (2018).
- [5] Anurag Arnab et al. “Vivit: A video vision transformer”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6836–6846.
- [6] M. Arzani et al. “Skeleton-based structured early activity prediction”. In: *Multimedia Tools and Applications* ().
- [7] Akin Avci et al. “Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey”. In: *23th International conference on architecture of computing systems 2010*. VDE. 2010, pp. 1–10.
- [8] Konstantinos Bacharidis and Antonis Argyros. “Improving Deep Learning Approaches for Human Activity Recognition based on Natural Language Processing of Action Labels”. In: *IJCNN*. IEEE. 2020.
- [9] W. Bao, Q. Yu, and Y. Kong. “Uncertainty-based Traffic Accident Anticipation with Spatio-Temporal Relational Learning”. In: *ACM Int’l Conf. on Multimedia*. 2020.
- [10] Yizhak Ben-Shabat et al. “The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 847–859.

- [11] A. Bochkovskiy, C.Y. Wang, and H.M. Liao. “YOLOv4: Optimal Speed and Accuracy of Object Detection”. In: *arXiv:2004.10934* (2020).
- [12] Kaidi Cao et al. “Few-shot video classification via temporal alignment”. In: *CVPR*. 2020.
- [13] Joao Carreira and Andrew Zisserman. “Quo vadis, action recognition? a new model and the kinetics dataset”. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.
- [14] Chien-Yi Chang et al. “D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation”. In: *CVPR*. 2019.
- [15] Yu-Wei Chao et al. “Forecasting Human Dynamics from Static Images.” In: *CVPR*. 2017, pp. 3643–3651.
- [16] M. Cuturi. “Fast global alignment kernels”. In: *ICML*. 2011.
- [17] M. Cuturi and M. Blondel. “Soft-DTW: a differentiable loss function for time-series”. In: *arXiv:1703.01541* (2017).
- [18] Dima Damen et al. “Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100”. In: *International Journal of Computer Vision* (), pp. 1–23.
- [19] Dima Damen et al. “Scaling egocentric vision: The epic-kitchens dataset”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 720–736.
- [20] Dima Damen et al. “The epic-kitchens dataset: Collection, challenges and baselines”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (2020), pp. 4125–4141.
- [21] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *IEEE CVPR*. 2009.
- [22] Eadom Dessalene et al. “Forecasting action through contact representations from first person video”. In: *IEEE PAMI* (2021).
- [23] Eadom Dessalene et al. “Therbligs in Action: Video Understanding through Motion Primitives”. In: *CVPR*. 2023.
- [24] Chhavi Dhiman and Dinesh Kumar Vishwakarma. “A review of state-of-the-art techniques for abnormal human activity recognition”. In: *Engineering Applications of Artificial Intelligence* 77 (2019), pp. 21–45.

- [25] Vibekananda Dutta and Teresa Zielinska. “Predicting human actions taking into account object affordances”. In: *Journal of Intelligent & Robotic Systems* (2019).
- [26] Nikita Dvornik et al. “Drop-DTW: Aligning Common Signal Between Sequences While Dropping Outliers”. In: *arXiv preprint arXiv:2108.11996* (2021).
- [27] Scott Ettinger et al. “Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset”. In: *Proceedings of the IEEE / CVF International Conference on Computer Vision*. 2021, pp. 9710–9719.
- [28] A. Farha, A. Richard, and J Gall. “When will you do what?-anticipating temporal occurrences of activities”. In: *IEEE CVPR*. 2018.
- [29] Yazan Abu Farha and Jurgen Gall. “Ms-tcn: Multi-stage temporal convolutional network for action segmentation”. In: *Proceedings of the IEEE / CVF conference on computer vision and pattern recognition*. 2019, pp. 3575–3584.
- [30] Yazan Abu Farha, Alexander Richard, and Juergen Gall. “When will you do what?-Anticipating Temporal Occurrences of Activities”. In: *arXiv preprint arXiv:1804.00892* (2018).
- [31] Christoph Feichtenhofer et al. “Slowfast networks for video recognition”. In: *Proceedings of the IEEE / CVF international conference on computer vision*. 2019, pp. 6202–6211.
- [32] Christiane Fellbaum. “WordNet and wordnets”. In: (2005).
- [33] Panna Felsen, Pulkit Agrawal, and Jitendra Malik. “What will happen next? forecasting player moves in sports videos”. In: *ICCV, Oct 1* (2017), p. 2.
- [34] Tharindu Fernando et al. “Deep inverse reinforcement learning for behavior prediction in autonomous driving: Accurate forecasts of vehicle motion”. In: *IEEE Signal Processing Magazine* 38.1 (2020), pp. 87–96.
- [35] A. Furnari and G. Farinella. “Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video”. In: *IEEE PAMI* (2020).
- [36] Antonino Furnari and Giovanni Maria Farinella. “What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention”. In: *Proceedings of the IEEE / CVF International Conference on Computer Vision*. 2019, pp. 6252–6261.

- [37] Antonino Furnari et al. “Next-active-object prediction from egocentric videos”. In: *JVCIR* (2017).
- [38] H. Gammulle et al. “Predicting the future: A jointly learnt model for action anticipation”. In: *IEEE ICCV*. 2019.
- [39] Harshala Gammulle et al. *Forecasting Future Action Sequences with Neural Memory Networks*. 2019. arXiv: 1909.09278 [cs.CV].
- [40] Reza Ghoddoosian, Saif Sayed, and Vassilis Athitsos. “Action Duration Prediction for Segment-Level Alignment of Weakly-Labeled Videos”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 2053–2062.
- [41] Rohit Girdhar and Kristen Grauman. “Anticipative video transformer”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 13505–13515.
- [42] Dayoung Gong et al. “Future transformer for long-term action anticipation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3052–3061.
- [43] Kristen Grauman et al. “Ego4d: Around the world in 3,000 hours of egocentric video”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18995–19012.
- [44] Alex Graves et al. “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 369–376.
- [45] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. “Representation Learning via Global Temporal Alignment and Cycle-Consistency”. In: *arXiv preprint arXiv:2105.05217* (2021).
- [46] Sanjay Haresh et al. “Learning by Aligning Videos in Time”. In: *arXiv preprint arXiv:2103.17260* (2021).
- [47] Tal Hassner. “A critical review of action recognition benchmarks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2013, pp. 245–250.

- [48] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. “Going deeper into action recognition: A survey”. In: *Image and vision computing* 60 (2017), pp. 4–21.
- [49] Jenhao Hsiao, Yikang Li, and Chiuman Ho. “Language-guided Multi-Modal Fusion for Video Action Recognition”. In: *Proceedings of the IEEE / CVF International Conference on Computer Vision*. 2021, pp. 3158–3162.
- [50] “<https://github.com/statefb/dtwalign>”. In: ().
- [51] Jian-Fang Hu et al. “Early action prediction by soft regression”. In: *PAMI* (2018).
- [52] Xuejiao Hu et al. “Online human action detection and anticipation in videos: A survey”. In: *Neurocomputing* (2022).
- [53] De-An Huang and Kris M. Kitani. “Action-Reaction: Forecasting the Dynamics of Human Interaction”. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 489–504. ISBN: 978-3-319-10584-0.
- [54] Youngkyoon Jang et al. “Epic-tent: An egocentric video dataset for camping tent assembly”. In: *Proceedings of the IEEE / CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0.
- [55] Jingjing Jiang et al. “Predicting short-term next-active-object through visual attention and hand position”. In: *Neurocomputing* 433 (2021), pp. 212–222.
- [56] Evangelos Kazakos et al. “With a little help from my temporal context: Multimodal egocentric action recognition”. In: *arXiv preprint arXiv:2111.01024* (2021).
- [57] Q. Ke, M. Fritz, and B. Schiele. “Time-conditioned action anticipation in one shot”. In: *IEEE CVPR*. 2019.
- [58] Q. Ke et al. “Learning latent global network for skeleton-based action prediction”. In: *IEEE Trans. on Image Processing* (2019).
- [59] Qihong Ke, Mario Fritz, and Bernt Schiele. “Time-Conditioned Action Anticipation in One Shot”. In: *2019 IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 9917–9926. DOI: 10.1109/CVPR.2019.01016.

- [60] Rajat Khurana and Alok Kumar Singh Kushwaha. “Deep Learning Approaches for Human Activity Recognition in Video Surveillance-A Survey”. In: *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. IEEE. 2019, pp. 542–544.
- [61] Dohyung Kim et al. “Classification of dance motions with depth cameras using subsequence dynamic time warping”. In: *SPPR*. IEEE. 2015.
- [62] Dohwan Ko et al. “Video-text representation learning via differentiable weak temporal alignment”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5016–5025.
- [63] Yu Kong and Yun Fu. “Human action recognition and prediction: A survey”. In: *arXiv preprint arXiv:1806.11230* (2018).
- [64] Yu Kong and Yun Fu. “Human action recognition and prediction: A survey”. In: *International Journal of Computer Vision* 130.5 (2022), pp. 1366–1401.
- [65] Yu Kong et al. “Action prediction from videos via memorizing hard-to-predict samples”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018.
- [66] H.S. Koppula, R. Gupta, and A. Saxena. “Learning human activities and object affordances from rgb-d videos”. In: *IJRR* (2013).
- [67] Parth Kothari, Sven Kreiss, and Alexandre Alahi. “Human trajectory forecasting in crowds: A deep learning perspective”. In: *IEEE Transactions on Intelligent Transportation Systems* 23.7 (2021), pp. 7386–7400.
- [68] Ranjay Krishna et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations”. In: *International Journal of Computer Vision* 123.1 (2017), pp. 32–73.
- [69] Hilde Kuehne, Ali Arslan, and Thomas Serre. “The language of actions: Recovering the syntax and semantics of goal-directed human activities”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 780–787.
- [70] Sateesh Kumar et al. “Unsupervised action segmentation by joint representation learning and online clustering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20174–20185.

- [71] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. “A Hierarchical Representation for Future Action Prediction”. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 689–704. ISBN: 978-3-319-10578-9.
- [72] T. Li et al. “Hard-net: hardness-aware discrimination network for 3D early activity prediction”. In: *ECCV*. 2020.
- [73] Yin Li, Miao Liu, and James M Rehg. “In the eye of beholder: Joint learning of gaze and actions in first person video”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 619–635.
- [74] Zhenyang Li et al. “Videolstm convolves, attends and flows for action recognition”. In: *Computer Vision and Image Understanding* 166 (2018), pp. 41–50.
- [75] Ji Lin, Chuang Gan, and Song Han. “Tsm: Temporal shift module for efficient video understanding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 7083–7093.
- [76] J. Liu et al. “Skeleton-based online action prediction using scale selection network”. In: *IEEE PAMI* (2019).
- [77] Miao Liu et al. “Forecasting human-object interaction: joint prediction of motor attention and actions in first person video”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer. 2020, pp. 704–721.
- [78] Shaowei Liu et al. “Joint Hand Motion and Interaction Hotspots Prediction from Egocentric Videos”. In: *IEEE CVPR*. 2022.
- [79] Tianshan Liu and Kin-Man Lam. “A Hybrid Egocentric Activity Anticipation Framework via Memory-Augmented Recurrent and One-shot Representation Forecasting”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13904–13913.
- [80] Edward Loper and Steven Bird. “Nltk: The natural language toolkit”. In: *arXiv preprint cs/0205028* (2002).
- [81] Cewu Lu et al. “Visual relationship detection with language priors”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 852–869.

- [82] Tahmida Mahmud, Mahmudul Hasan, and Amit K Roy-Chowdhury. “Joint prediction of activity labels and starting times in untrimmed videos”. In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE. 2017, pp. 5784–5793.
- [83] Tahmida Mahmud et al. “Prediction and description of near-future activities in video”. In: *Computer Vision and Image Understanding* 210 (2021), p. 103230.
- [84] V. Manousaki, K. Papoutsakis, and A. Argyros. “Evaluating Method Design Options for Action Classification based on Bags of Visual Words.” In: *VISAPP*. 2018.
- [85] Victoria Manousaki and Antonis A Argyros. “Segregational Soft Dynamic Time Warping and Its Application to Action Prediction.” In: *VISIGRAPP (5: VISAPP)*. 2022.
- [86] Victoria Manousaki, Konstantinos Papoutsakis, and Antonis Argyros. “Action Prediction During Human-Object Interaction Based on DTW and Early Fusion of Human and Object Representations”. In: *ICVS*. Springer. 2021.
- [87] Victoria Manousaki, Konstantinos Papoutsakis, and Antonis Argyros. “Graphing the Future: Activity and Next Active Object Prediction using Graph-based Activity Representations”. In: *International Symposium on Visual Computing*. Springer. 2022, pp. 299–312.
- [88] Victoria Manousaki et al. “VLMAH: Visual-Linguistic Modeling of Action History for Effective Action Anticipation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 1917–1927.
- [89] W. Mao, M. Liu, and M. Salzmann. “History repeats itself: Human motion prediction via motion attention”. In: *ECCV*. 2020.
- [90] Julieta Martinez, Michael J Black, and Javier Romero. “On human motion prediction using recurrent neural networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, pp. 4674–4683.
- [91] A. Mavrogiannis, R. Chandra, and D. Manocha. “B-GAP: Behavior-Guided Action Prediction for Autonomous Navigation”. In: *arXiv:2011.03748* (2020).
- [92] E. Mavroudi, B. Haro, and R Vidal. “Representation Learning on Visual-Symbolic Graphs for Video Understanding”. In: *ECCV*. 2020.

- [93] A. Miech et al. “Leveraging the present to anticipate the future in videos”. In: *IEEE CVPR Workshops*. 2019.
- [94] Antoine Miech et al. “Howto100m: Learning a text-video embedding by watching hundred million narrated video clips”. In: *Proceedings of the IEEE / CVF International Conference on Computer Vision*. 2019, pp. 2630–2640.
- [95] James Munkres. “Algorithms for the assignment and transportation problems”. In: *Journal of the society for industrial and applied mathematics* (1957).
- [96] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. “Grounded human-object interaction hotspots from video”. In: *IEEE ICCV*. 2019.
- [97] Megha Nawhal, Akash Abdu Jyothi, and Greg Mori. “Rethinking Learning Approaches for Long-Term Action Anticipation”. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*. Springer. 2022, pp. 558–576.
- [98] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. “A comprehensive review on deep learning-based methods for video anomaly detection”. In: *Image and Vision Computing* 106 (2021), p. 104078.
- [99] Saul B Needleman and Christian D Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of molecular biology* 48.3 (1970), pp. 443–453.
- [100] Y.B. Ng and F. Basura. “Forecasting future action sequences with attention: A new approach to weakly supervised action forecasting”. In: *IEEE Trans. on Image Processing* (2020).
- [101] F. Ofli et al. “Berkeley MHAD: A comprehensive multimodal human action database”. In: *IEEE WACV*. 2013.
- [102] S. Oprea et al. “A Review on Deep Learning Techniques for Video Prediction”. In: *IEEE PAMI* (2020).
- [103] Costas Panagiotakis, Konstantinos Papoutsakis, and Antonis Argyros. “A graph-based approach for detecting common actions in motion capture data and videos”. In: *In Pattern Recognition* (2018).

- [104] K. Papoutsakis, C. Panagiotakis, and A. Argyros. “Temporal Action Co-Segmentation in 3D Motion Capture Data and Videos”. In: *CVPR*. 2017.
- [105] Konstantinos Papoutsakis, Costas Panagiotakis, and Antonis A Argyros. “Temporal action co-segmentation in 3d motion capture data and videos”. In: *CVPR*. 2017.
- [106] Konstantinos Papoutsakis et al. “Detection of Physical Strain and Fatigue in Industrial Environments Using Visual and Non-Visual Low-Cost Sensors”. In: *Technologies* 10.2 (2022). ISSN: 2227-7080. DOI: 10.3390/technologies10020042. URL: <https://www.mdpi.com/2227-7080/10/2/42>.
- [107] Konstantinos E Papoutsakis and Antonis A Argyros. “Unsupervised and Explainable Assessment of Video Similarity.” In: *BMVC*. 2019.
- [108] Alex S Park and James R Glass. “Unsupervised pattern discovery in speech”. In: *IEEE Transactions on Audio, Speech, and Language Processing* (2007).
- [109] Razvan-George Pasca et al. “Summarize the Past to Predict the Future: Natural Language Descriptions of Context Boost Multimodal Object Interaction”. In: *arXiv preprint arXiv:2301.09209* (2023).
- [110] Tomislav Petković et al. “Human action prediction in collaborative environments based on shared-weight LSTMs with feature dimensionality reduction”. In: *Applied Soft Computing* (2022).
- [111] Ronald Poppe. “A survey on vision-based human action recognition”. In: *Image and vision computing* 28.6 (2010), pp. 976–990.
- [112] Ammar Qammar and Antonis Argyros. “Occlusion-tolerant and personalized 3D human pose estimation in RGB images”. In: *2020 ICPR*. IEEE. 2021.
- [113] Zhaobo Qi et al. “Self-regulated learning for egocentric video activity anticipation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [114] Y. Qin et al. “Skeleton-based action recognition by part-aware graph convolutional networks”. In: *The visual computer* (2020).
- [115] Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. “StillFast: An End-to-End Approach for Short-Term Object Interaction Anticipation”. In: *Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 3635–3644.

- [116] Francesco Ragusa et al. “The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain”. In: *Proceedings of the IEEE / CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1569–1578.
- [117] A. Rasouli. “Deep learning for vision-based prediction: A survey”. In: *arXiv:2007.00095* (2020).
- [118] A. Rasouli et al. “Multi-Modal Hybrid Architecture for Pedestrian Action Prediction”. In: *arXiv:2012.00514* (2020).
- [119] B. Reily et al. “Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction”. In: *Autonomous Robots* (2018).
- [120] I. Rius et al. “Action-specific motion prior for efficient Bayesian 3D human body tracking”. In: *Pattern Recognition* (2009).
- [121] Ivan Rodin et al. “Predicting the future from first person (egocentric) vision: A survey”. In: *Computer Vision and Image Understanding* 211 (2021), p. 103252.
- [122] Ivan Rodin et al. “Untrimmed Action Anticipation”. In: *International Conference on Image Analysis and Processing*. 2022.
- [123] Ivan Rodin et al. “Untrimmed Action Anticipation”. In: *Image Analysis and Processing – ICIAP 2022*. Ed. by Stan Sclaroff et al. Cham: Springer International Publishing, 2022, pp. 337–348. ISBN: 978-3-031-06433-3.
- [124] Marcus Rohrbach et al. “A database for fine grained activity detection of cooking activities”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 1194–1201.
- [125] Marcus Rohrbach et al. “Script data for attribute-based recognition of composite activities”. In: *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*. Springer. 2012, pp. 144–157.
- [126] Debaditya Roy and Basura Fernando. “Action anticipation using latent goal learning”. In: *Proceedings of the IEEE / CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 2745–2753.
- [127] R.Tavenard et al. “Tslearn, A Machine Learning Toolkit for Time Series Data”. In: *JMLR* (2020).

- [128] Michael S Ryoo. “Human activity prediction: Early recognition of ongoing activities from streaming videos”. In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, pp. 1036–1043.
- [129] MS Ryoo et al. “Robot-centric activity prediction from first-person videos: What will they do to me?” In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2015, pp. 295–302.
- [130] H. Sakoe and S. Chiba. “Dynamic programming algorithm optimization for spoken word recognition”. In: *IEEE ICASSP (1978)*.
- [131] Santiago Schez-Sobrino et al. “Automatic recognition of physical exercises performed by stroke survivors to improve remote rehabilitation”. In: *MAPR*. 2019.
- [132] Christian Schuldt, Ivan Laptev, and Barbara Caputo. “Recognizing human actions: a local SVM approach”. In: *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. IEEE. 2004, pp. 32–36.
- [133] Fadime Sener, Rishabh Saraf, and Angela Yao. “Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)*.
- [134] Fadime Sener, Dipika Singhania, and Angela Yao. “Temporal aggregate representations for long-range video understanding”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 154–171.
- [135] Fadime Sener and Angela Yao. “Zero-shot anticipation for instructional activities”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 862–871.
- [136] Fadime Sener et al. “Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 21096–21106.
- [137] Dandan Shan et al. “Understanding human hands in contact at internet scale”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9869–9878.

- [138] Neha Sharma, Chhavi Dhiman, and S Indu. “Pedestrian intention prediction for autonomous vehicles: A comprehensive survey”. In: *Neurocomputing* (2022).
- [139] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [140] Sebastian Stein and Stephen J McKenna. “Combining embedded accelerometers with computer vision for recognizing food preparation activities”. In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM. 2013, pp. 729–738.
- [141] Chen Sun et al. “Learning video representations using contrastive bidirectional transformer”. In: *arXiv preprint arXiv:1906.05743* (2019).
- [142] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. “Computational human-robot interaction”. In: *Foundations and Trends® in Robotics* 4.2-3 (2016), pp. 105–223.
- [143] Paolo Tormene et al. “Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation”. In: *Artificial intelligence in medicine* (2009).
- [144] Du Tran et al. “Learning spatiotemporal features with 3d convolutional networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4489–4497.
- [145] Nghia Pham Trong et al. “A comprehensive survey on human activity prediction”. In: *International Conference on Computational Science and Its Applications*. Springer. 2017, pp. 411–425.
- [146] Pavan Turaga et al. “Machine recognition of human activities: A survey”. In: *IEEE Transactions on Circuits and Systems for Video technology* 18.11 (2008), p. 1473.
- [147] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. “Anticipating visual representations from unlabeled video”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 98–106.
- [148] Chuhua Wang et al. “Stepwise goal-driven networks for trajectory prediction”. In: *IEEE Robotics and Automation Letters* (2022).
- [149] J. Wang et al. “Mining actionlet ensemble for action recognition with depth cameras”. In: *IEEE CVPR*. 2012.

- [150] Limin Wang et al. “Temporal segment networks for action recognition in videos”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.11 (2018), pp. 2740–2755.
- [151] Limin Wang et al. “Temporal segment networks: Towards good practices for deep action recognition”. In: *European conference on computer vision*. Springer. 2016, pp. 20–36.
- [152] Mengmeng Wang, Jiazheng Xing, and Yong Liu. “Actionclip: A new paradigm for video action recognition”. In: *arXiv preprint arXiv:2109.08472* (2021).
- [153] Xiaolong Wang et al. “Non-local neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7794–7803.
- [154] Xionghui Wang et al. “Progressive Teacher-Student Learning for Early Action Prediction”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3556–3565.
- [155] Xinxiao Wu et al. “Spatial-temporal relation reasoning for action prediction in videos”. In: *IJCV* (2021).
- [156] Zhibiao Wu and Martha Palmer. “Verb semantics and lexical selection”. In: *arXiv preprint cmp-lg/9406033* (1994).
- [157] L. Xia and JK Aggarwal. “Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera”. In: *IEEE CVPR*. 2013.
- [158] Xinyu Xu, Yong-Lu Li, and Cewu Lu. “Learning to anticipate future with dynamic context removal”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12734–12744.
- [159] Zihui Xue and Kristen Grauman. “Learning Fine-grained View-Invariant Representations from Unpaired Ego-Exo Videos via Temporal Alignment”. In: *arXiv preprint arXiv:2306.05526* (2023).
- [160] Ceyuan Yang et al. “Temporal pyramid network for action recognition”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 591–600.
- [161] Chuan-Kai Yang and Richard Tondowidjojo. “Kinect v2 Based Real-Time Motion Comparison with Re-targeting and Color Code Feedback”. In: *IEEE GCCE*. 2019.

- [162] Joe Yue-Hei Ng et al. “Beyond short snippets: Deep networks for video classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4694–4702.
- [163] Olga Zatsarynna, Yazan Abu Farha, and Juergen Gall. “Multi-modal temporal convolutional network for anticipating actions in egocentric videos”. In: *IEEE CVPR*. 2021.
- [164] Jing Zhang et al. “RGB-D-based action recognition datasets: A survey”. In: *Pattern Recognition* 60 (2016), pp. 86–105.
- [165] Zeyun Zhong et al. “Anticipative Feature Fusion Transformer for Multi-Modal Action Anticipation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 6068–6077.
- [166] Bolei Zhou et al. “Temporal relational reasoning in videos”. In: *ECCV*. 2018.
- [167] Luowei Zhou, Chenliang Xu, and Jason Corso. “Towards automatic learning of procedures from web instructional videos”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.