Time-slice Analysis of Dyadic Human Activity

Thèse

Maryam Ziaeefard

Doctorat en Génie Électrique
Philosophiæ doctor (Ph.D.)

Québec, Canada

Time-slice Analysis of Dyadic Human Activity

Thèse

Maryam Ziaeefard

Sous la direction de:

Robert Bergevin, directeur de recherche

Résumé

La reconnaissance d'activités humaines à partir de données vidéo est utilisée pour la surveillance ainsi que pour des applications d'interaction homme-machine. Le principal objectif est de classer les vidéos dans l'une des k classes d'actions à partir de vidéos entièrement observées. Cependant, de tout temps, les systèmes intelligents sont améliorés afin de prendre des décisions basées sur des incertitudes et ou des informations incomplètes. Ce besoin nous motive à introduire le problème de l'analyse de l'incertitude associée aux activités humaines et de pouvoir passer à un nouveau niveau de généralité lié aux problèmes d'analyse d'actions. Nous allons également présenter le problème de reconnaissance d'activités par intervalle de temps, qui vise à explorer l'activité humaine dans un intervalle de temps court. Il a été démontré que l'analyse par intervalle de temps est utile pour la caractérisation des mouvements et en général pour l'analyse de contenus vidéo. Ces études nous encouragent à utiliser ces intervalles de temps afin d'analyser l'incertitude associée aux activités humaines. Nous allons détailler à quel degré de certitude chaque activité se produit au cours de la vidéo.

Dans cette thèse, l'analyse par intervalle de temps d'activités humaines avec incertitudes sera structurée en 3 parties. i) Nous présentons une nouvelle famille de descripteurs spatiotemporels optimisés pour la prédiction précoce avec annotations d'intervalle de temps. Notre représentation prédictive du point d'intérêt spatiotemporel (Predict-STIP) est basée sur l'idée de la contingence entre intervalles de temps. ii) Nous exploitons des techniques de pointe pour extraire des points d'intérêts afin de représenter ces intervalles de temps. iii) Nous utilisons des relations (uniformes et par paires) basées sur les réseaux neuronaux convolutionnels entre les différentes parties du corps de l'individu dans chaque intervalle de temps. Les relations uniformes enregistrent l'apparence locale de la partie du corps tandis que les relations par paires captent les relations contextuelles locales entre les parties du corps. Nous extrayons les spécificités de chaque image dans l'intervalle de temps et examinons différentes façons de les agréger temporellement afin de générer un descripteur pour tout l'intervalle de temps.

En outre, nous créons une nouvelle base de données qui est annotée à de multiples intervalles de temps courts, permettant la modélisation de l'incertitude inhérente à la reconnaissance d'activités par intervalle de temps. Les résultats expérimentaux montrent l'efficience de notre stratégie dans l'analyse des mouvements humains avec incertitude.

Abstract

Recognizing human activities from video data is routinely leveraged for surveillance and human-computer interaction applications. The main focus has been classifying videos into one of k action classes from fully observed videos. However, intelligent systems must to make decisions under uncertainty, and based on incomplete information. This need motivates us to introduce the problem of analysing the uncertainty associated with human activities and move to a new level of generality in the action analysis problem. We also present the problem of time-slice activity recognition which aims to explore human activity at a small temporal granularity. Time-slice recognition is able to infer human behaviours from a short temporal window. It has been shown that temporal slice analysis is helpful for motion characterization and for video content representation in general. These studies motivate us to consider time-slices for analysing the uncertainty associated with human activities. We report to what degree of certainty each activity is occurring throughout the video from definitely not occurring to definitely occurring.

In this research, we propose three frameworks for time-slice analysis of dyadic human activity under uncertainty. i) We present a new family of spatio-temporal descriptors which are optimized for early prediction with time-slice action annotations. Our predictive spatiotemporal interest point (Predict-STIP) representation is based on the intuition of temporal contingency between time-slices. ii) we exploit state-of-the art techniques to extract interest points in order to represent time-slices. We also present an accumulative uncertainty to depict the uncertainty associated with partially observed videos for the task of early activity recognition. iii) we use Convolutional Neural Networks-based unary and pairwise relations between human body joints in each time-slice. The unary term captures the local appearance of the joints while the pairwise term captures the local contextual relations between the parts. We extract these features from each frame in a time-slice and examine different temporal aggregations to generate a descriptor for the whole time-slice.

Furthermore, we create a novel dataset which is annotated at multiple short temporal windows, allowing the modelling of the inherent uncertainty in time-slice activity recognition. All the three methods have been evaluated on TAP dataset. Experimental results demonstrate the effectiveness of our framework in the analysis of dyadic activities under uncertainty

Contents

List of Tables

List of Figures

To My Dear Parents,

Gone But Never Forgotten

Acknowledgement

I would like to thank my advisor Robert Bergevin for his help through my Ph.D. journey. I also thank Denis Laurendeau, the director of the Computer Vision and Systems Laboratory (CVSL), for providing equipments for my project.

A special thanks goes to JF Lalonde who kindly accepted revising this manuscript and helping me to prepare and submit this thesis. Thanks also to Paul Fortier for his support during the absence of my advisor.

I would also like to thank LP Morency for his advice and excellent guiding during my internship at University of Southern California. I had this opportunity to work under his supervision and make advantage of his knowledge in my project.

I thank my friends at CVSL lab and also Annette Schwerdtfeger for proofreading all my manuscripts. Last but not least, I am deeply grateful to my sister Saide for her endless encouragement. This work could not have been accomplished without her dedicated support and love.

Foreword

Four chapters of this thesis are composed of material already published or under review in technical conferences or journal papers. In this thesis, text and figures have been modified in order to be consistent with the rest of the document. Some material which did not find place in the original papers has been added to better clarify the ideas behind each method. Here, I detail my contributions to 4 research papers.

Paper 1: M. Ziaefard and R. Bergevin, "Semantic Activity Recognition: A Literature Review," Pattern Recognition Journal, Volume 48, Issue 8, August 2015, Pages 2329–2345. This paper reviews the state-of-the-art methods in activity recognition which use semantic features.

Paper 2: M. Ziaefard, R. Bergevin, and L.P. Morency, "Time-slice Prediction of Dyadic Human Activities," In British Machine Vision Conference (BMVC), Pages 167.1-167.13, BMVA Press, September 2015. In this paper, we introduce the problem of time-slice activity recognition which aims to explore human activity at a small temporal granularity (time-slice). Furthermore, we collect a new dataset which is annotated at multiple short temporal windows, allowing the modelling of the inherent uncertainty in time-slice activity recognition. The experiments were conducted at Institute for Creative Technologies (ICT) at University of Southern California while I was doing my internship under the supervision of Professor Morency.

Paper 3: M. Ziaefard and R. Bergevin, "Integration of Uncertainty in the Analysis of Dyadic Human Activities," In 13th Conference on Computer and Robot Vision (CRV), June 2016. The main focus of this paper is to analyse the uncertainty associated with dyadic human activities and move to a new level of generality in the action analysis problem. Analysing the uncertainty, here, refers to categorizing the likelihood of activities in the time-slices.

Paper 4: M. Ziaefard, R. Bergevin, and J-F Lalonde, "Deep Uncertainty Interpretation in Dyadic Human Activity Prediction," Submitted to 28th British Machine Vision Conference (BMVC 2017). In this paper, we proposed a method exploiting Convolutional Neural Networks (CNNs) to extract unary and pairwise probabilities of human body pose to analyse the uncertainty associated with human activities.

Introduction

0.1 The perspective

Analysing human activities from video data is leveraged for surveillance and human-computer interaction applications. In this context, there is a large body of work that analyse human activities from fully observed video sequences. However, analysing activities in shorter windows returns detailed information on what activities occur throughout the video. Therefore, in this thesis, we are more interested in exploring human activities at a small temporal granularity (time-slice).

Time-slice analysis infers human behaviour throughout the video. Considering a human action video starting with the initiation of an action, there is more confusion and uncertainty in the first few frames. As time passes by and informative frames are given, recognizing or predicting activities becomes easier. Humans can naturally model the uncertainty associated with with each activity. We do not need to see a full handshake video before being able to recognize it. This ability of humans to recognize an action before seeing it in full inspired us to introduce the time-slice activity analysis. We divide a video into several time fragments which are referred to as time-slice and analyse the possibility of the occurrence of different activities. In other words, instead of classifying a whole video to a single label, we analyse the possibility of occurrence of different activities and the uncertainty associated with activities in each time-slice.

Time-slice analysis is helpful for motion characterization and in general for video content representation. Extracting content of video automatically will extend the current scope of possibilities for video indexing and retrieval. For instance, we will be able to search for the parts of a video in which a certain activity is (not) occurring. An instance of real-world applications for the time-slice analysis is to automatically detect violent shots in movies in order to prevent children to watch them.

In the literature, human activities are categorized into atomic actions, people interactions, human-object interactions, and group activities in terms of the number of people involved in performing the activity. The focus of this thesis is analysing interactions between two people which is referred to as dyadic activities.

Conventional approaches recognize activities based on either the whole video sequence (holistic approach) or the early part of the video (early recognition). However, our time-slice study generalizes the part to any short-term observation anywhere in the video sequence. It is noteworthy to mention that the locations of the time-slices are unknown in both our learning and test phases. Another key difference of this work with the previous work is the integration of uncertainty in the activity analysis. That is, obtaining a measure to show how likely each possible activity occurs throughout the video which leads to a new level of generality in the action analysis problem.

The integration of uncertainty in the activity analysis helps to investigate this problem: to what extent the activity of interest is likely to occur during each time-slice throughout the video? Analysing the uncertainty associated with human activities has important applications for practical scenarios, where decisions have to be made even if the occurred activity cannot be predicted precisely.

Another outstanding aspect of this thesis is that it re-examines the problem of activity prediction with the state-of-the-art Convolutional Neural Network (CNN) models inspired by recent success of deep learning approaches in other computer vision domains. We propose a new framework for this problem and compare its performance with the results obtained by hand-crafted features.

0.2 Thesis outline

In the second chapter of this thesis, we present a comprehensive literature review and our motivations for analysing human activities in this thesis. Our contributions are the following:

- We present a detailed review on recent action recognition frameworks based on semantic information;
- We introduce a semantic space for feature descriptors;
- The performance of semantic and non-semantic methods is computed.

Chapter 3 is devoted to presenting a novel human activity recognition approach based on time-slices. In this chapter, we are interested in analysing our understanding of activities occurring within time-slice observations. Our contributions are threefold:

- We introduce the problem of time-slice activity recognition and compare time-slice recognition with the conventional approaches;
- We propose a new set of spatio-temporal features using time-slice action annotations that identify descriptors with broad temporal coverage;

– We collect a new dataset of time-slice annotations.

In Chapter 4, we introduce the problem of analysing the uncertainty associated with dyadic human activities in time-slices. Conventional methods have been classifying videos into certain action classes. In contrast, our focus in this chapter is to categorize the likelihood of activities occurring in each time-slice. Our contributions are the following:

– We propose a model to integrate uncertainty in the analysis of dyadic human activities in videos;

– We introduce a learning framework addressing this model;

– We compare different instantiations of the framework;

– We present a novel technique for evaluating the performance of early activity recognition methods based on the uncertainty associated with partially observed videos.

In Chapter 5, inspired by the strong performance of Convolutional Neural Networks (CNN) in other computer vision applications, we propose a CNN-based algorithm to analyse the uncertainty associated with dyadic activities. Deep learning techniques have offered a compelling alternative to hand-crafted features. We use a deep learning method to extract frame descriptors containing the probabilities of the presence of body joints and locations of other joints in the adjacency. Thus, we use deep learning along with human pose information to propose a method for time-slice activity analysis. We compare the proposed method with the results obtained by hand-crafted features. Our contributions in this chapter are:

– We introduce a novel way of exploiting CNN in time-slice analysis;

– We propose a technique to integrate unary and pairwise pose information to measure the uncertainty associated with activity recognition;

– We present a Single-Stream deep learning framework addressing this technique.

Finally, in Chapter 6, conclusions are drawn, and some of the possible applications based on the material developed in this dissertation are reviewed.

Chapter 1

Literature Review

The problem of analysing human activities in images and videos has received growing interest in the computer vision community. Analysing human activities refers to recognize or predict activities from single images or fully or partially observed videos. Given the significant literature in the area, we focus only on the most relevant works to our proposed algorithms in this chapter.

Semantic Human Activity Recognition: A Literature Review

Abstract

This paper presents an overview of state-of-the-art methods in activity recognition using semantic features. Unlike low-level features, semantic features describe inherent characteristics of activities. Therefore, semantics make the recognition task more reliable especially when the same actions look visually different due to the variety of action executions. We define a semantic space including the most popular semantic features of an action namely the human body (pose and poselet), attributes, related objects, and scene context. We present methods exploiting these semantic features to recognize activities from still images and video data as well as four groups of activities: atomic actions, people interactions, human-object interactions, and group activities. Furthermore, we provide potential applications of semantic approaches along with directions for future research.

1.1 Introduction

Human activity recognition is being leveraged for an increasingly wide variety of computer vision applications. What all of these works have in common is to study some aspects of human-computer interaction. Recognizing activities can range from a single person action to multi-people activity recognition. Generally, an action is defined as a single person activity but we use the terms action and activity interchangeably.

Figure 1.1: Kicking action. The same actions appear different due to different camera angles, clothes, body shapes, etc

A number of surveys have been published in activity recognition during the last decade. Most of the earlier reviews have focused on the introduction and general summarization of activity recognition methodologies (70; 5; 71). A study by (6) covered human activity recognition methods with a categorization based on the complexity of activities and recognition methodologies. Various challenges in action recognition were addressed and limitations of different approaches were discussed in (7). Recently, Aggarwal and Ryoo (73) conducted a survey emphasizing activity recognition methods for four groups of activities (atomic action, people interaction, human-object interaction, and group activity). They classified activity recognition methodologies into two categories: single-layered approaches and hierarchical approaches. Single-layered methods represent and recognize human activities directly based on sequences of images. On the other hand, hierarchical approaches describe high-level human activities by using simpler activities called sub-events which are suitable for the analysis of complex activities. Aggarwal and Ryoo (73) also mentioned a few semantic approaches without clearly explaining what semantics is and why it should be used. In this survey, we aim to cover the methods in the literature which address semantic activity understanding.

Human activity recognition methods can also be classified according to their input data. Traditional action recognition approaches used videos or image sequences while recent studies started to explore action recognition in still images. Compared to the video-based action recognition, still image-based action recognition has some special properties. For example, there is no motion in a still image, and thus many spatio-temporal features and methods that were developed for traditional video-based action recognition are not applicable to still images. A recent survey (74) presents a detailed overview of the existing approaches in still image-based action recognition and explains various features as well as related databases which have been used in analysing actions in still images.

Different levels of features have been used in activity recognition methods. Traditional action recognition methods rely mostly on tracking, and motion capture. Mid-level features such as spatio-temporal and bag-of-word features are used by recent approaches. Semantic features, meanwhile, are aimed to answer questions such as "what does it mean to perform an action?" or "How do we understand an action?". The term semantics refers to the study of meaning. For example, it is meaningful that a car and road appear in the same images, while a giraffe and a kitchen should not. A detailed definition of this term will be provided in Section 1.2.

Semantic features are useful to address the problem of intra-class variability. Intra-class variability refers to the differences in the same group of actions and how different instances of the same action resemble each other. As shown in Figure 1.1, people may perform the same action in different ways or even the same person may perform one action differently in different situations. In addition, humans vary significantly in appearance due to changes in clothing, body shape and viewpoint. Semantic features help to distinguish similar actions that differ visually but have common semantics.

Semantic approaches apply the human understanding of the activity. The human ability to recognize actions does not rely only on visual analysis of human body postures but also requires additional sources of information such as context or scene, knowledge about objects related to activities, or knowledge about the visual characteristics of activities. On the other hand, non-semantic approaches, here, refer to methods representing actions *only* in some form of low-level features such as silhouette, gradients, optical flow, etc. They do not incorporate human knowledge about activities. Non-semantic approaches capture the appearance and motion characteristics while semantic approaches describe inherent characteristics of activities. Non-semantic approaches are ideally appropriate for simple actions. However, they fail in complex situations due to the lack of semantics they represent.

To classify semantic approaches, we introduce a feature space called the "semantic space" which includes human knowledge about activities such as the body part (pose and poselet), object, scene, and attribute features. The semantic space is illustrated schematically in Figure 1.2. Based on exploiting these features, we categorize semantic methods into three categories: methods based on body parts, methods based on objects/scenes, and methods based on attributes.

The first feature of the semantic space is the body part. Neuropsychological studies indicate that semantic knowledge of human body parts might be distinct from knowledge of other object categories. Downing et al. (75) identified a subpart of the human extrastriate cortex involved in the visual processing of the human body and body parts, namely extrastriate body area or EBA. Their experimental results reveal that the EBA responds strongly and selectively to a variety of pictures of human bodies and body parts. The EBA may be crucial for perceiving the position and configuration of one's body, possibly as part of a general system for inferring the actions and intentions of others. Also, EBA may be involved in perceiving the configuration of one's own body. Peelen and Downing (76) and Schwarzlose et al. (77) worked also on the body selectivity of the brain. Methods for pose-based action recognition can either use pose estimation results as input for the action recognition step or address both pose estimation and action recognition concurrently. The latter approach has the advantage that errors due to inaccurate pose estimation will have less of an affect on the final quality of activity recognition. Semantics also captures salient body parts during an action which is referred to as a poselet. In 2D/3D images, a poselet is specified as a subset of the human

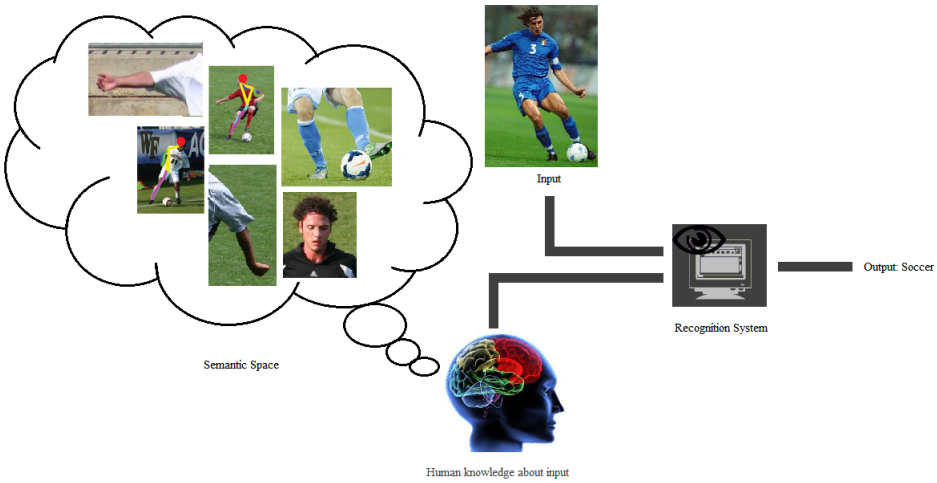


Figure 1.2: Semantic Space. Observing an action, e.g. "playing soccer", the human uses his knowledge to recognize the activity. We define a semantic space containing pose (specific body pose in soccer), poselet (extended right arm, straight left arm), object (soccer ball, interaction between one leg and a soccer ball), scene (soccer field), and attribute (looking-down head) which are illustrated in the figure.

pose. It usually refers to the union of adducent limbs, for example half of a torso and a left hand or a frontal view of both legs walking forward.

Semantic information can also be extracted from the scene where the action is performed. For instance, if an action scene context is recognized as a soccer field, it is more likely that the performed action is "playing soccer". In addition to the scene context, objects related to Human-Object Interaction (HOI) are used in semantic action recognition. Given a human action, there might be objects related to that action. Different actions are related to different objects. Knowing the related objects helps to recognize the corresponding actions. For example, a horse (with a human) is possibly related to the action of "riding a horse" while a phone (with a person) could be related to the action of "phoning".

Attributes are another important type of semantic feature which can be directly linked with the visual characteristics of actions. They can describe high-level activities better than the raw features (color, edge, etc.) extracted from images or videos. They use human knowledge to create descriptors that capture intrinsic properties of actions. Attributes describe spatial and temporal movements of the actor. For example, arm pendulum-like motion or the motion pattern of two legs, putting one foot in front of the other, are potential attributes for a walking action.

Despite the fact that some thinking about semantic activity recognition approaches has been provided in existing surveys, important issues still remain, such as how semantics is really useful and how reliable semantic features can be. The contributions of this paper are threefold.

First, this paper explains the term semantics including a review of semantics in neuroscience to show how action understanding relies on conceptual knowledge. It also introduces a semantic space to describe semantic features in detail and their use in other computer vision applications as well as their advantages and drawbacks. Secondly, it discusses and reviews semantic activity recognition methods in detail for both single images and video data considering four groups of activities. Finally, it compares the performance of semantic and non-semantic approaches to better understand the present level of semantic methods. Overall, this survey provides a comprehensive state-of-the-art review in semantic activity recognition. The purpose of this survey is to attempt to draw attention to high-level features and encourage researchers to propose semantic methods for action recognition. We show the taxonomy of semantic features and the publications that are reviewed in this study in Table 1.1.

The outline of the paper is as follows. The term semantics and a background of semantics in neuroscience are described in Section 1.2, followed by a presentation of semantic space in Section 1.3. Section 1.4 explains semantic approaches using body parts (pose and poselet). Section 1.5 and 1.6 describe methods based on objects/scenes and attributes respectively. The action recognition performance on the most popular datasets is presented in Section 1.7. Section 1.8 presents other semantic approaches and potential applications of semantic approaches. Finally, Section 1.9 concludes the paper and provides directions for future research.

1.2 What is Semantics?

It has been shown that semantics plays a key role in recognition in the human visual perception. Taking semantics into account to improve visual recognition has received considerable attention in the recent time. In this section, we first give a definition of semantics. Then, a review of semantics in neuroscience is studied to indicate how the human brain functions when it understands actions and how action understanding relies on semantic knowledge.

1.2.1 Definition of semantics

Generally, semantics refers to what the sender and receiver of a message mean and how they infer the context of the message. Semantics is the study of meaning. *Meaning* originates from the language spoken by the Angles and Saxons (or Old English) and is still related today with the German verb "meinen", i.e. to think or intend (78).

In action recognition, the semantic understanding enables users to apply prior knowledge to the recognition process. Semantics interprets an action as a relation between its features (e.g. body parts, corresponding objects, scene, etc.). The meaning of each action generally can be decomposed into the meanings of its features. For example, handshaking action can be interpreted as the certain movement of two people's hands.

Table 1.1: Taxonomy of mentioned methods based on types of features and input data on human activity recognition

	Video	**Single Image**
Pose	Park and Aggarwal (125) Lv and Nevatia (124) Eweiwi et al. (119) Cheema et al. (122) Vahdat et al. (123) C. Wang et al. (117) Chaaraoui et al. (121) Mukherjee et al. (115)	Raja et al. (114) Yao and Fei-Fei (116) Meng et al. (118) Yukita (85) Khan et al. (126)
Poselet	Raptis and Sigal (1) Nabi et al. (129) J. Wang et al. (130)	Yang et al. (127) Maji et al. (94) Zheng et al. (128) Chen and Grauman (131)
Scene	Jones and Shao (136) Ullah et al. (135) Zhang et al. (133) Liu et al. (134) Ikizler-Cinbis and Sclaroff (147) Han et al. (148) Marszalek et al. (132)	Li and Fei-Fei (146)
HOI	Ikizler-Cinbis and Sclaroff (147) Han et al. (148) Gupta et al. (138) Filipovych and Ribeiro (140) Wu et al. (139) Gupta and Davis (137) Kuniyoshi and Shimozaki (141)	Delaitre et al. (142) Desai et al. (143) Yao and Fei-Fei (144) Yao and Fei-Fei (145) Gupta et al. (138) Li and Fei-Fei (146)
Attribute	Liu et al. (104) Qiu et al. (151) Rohrbach et al. (153) Cheng et al. (175) Zhang et al. (152)	B. Yao et al. (154) Sharma et al. (149)
Linguistic descriptions	Park and Aggarwal (125) Rohrbach et al. (153) Motwani and Mooney (167) Guadarrama et al. (166)	

1.2.2 Semantics in neuroscience

Understanding the semantics and perception of actions in humans is a challenging task. It is an interdisciplinary research combining several scientific programs from computer science to brain science and psychology. To further clarify semantic action recognition, we review action understanding in neuroscience and present how the brain distinguishes activities.

Our understanding of the brain mechanisms for the recognition of actions has grown rapidly over the past decades. Early neuroscientific studies on monkeys (e.g. Gallese et al. (79)) have revealed that neurons of the rostral part of the inferior area, i.e. Mirror neurons, became active both when the monkey performed a given action and when it observed a similar action performed by the experimenter. The main property of the mirror neurons is to match observation of motor acts (goal-directed movements) with the execution of the same or similar motor acts. In other words, the mirror system provides a way to match observation and execution of events. A matching system, similar to that of mirror neurons in monkeys exists in humans and could be involved in the recognition of actions.

The semantic-level understanding of an action involves making a meaningful description of the action. It is more efficient for the brain to organize object and action categories into a continuous space that represents the semantic similarity between them. A study conducted by Huth et al. (80) showed that similar categories are located next to each other in the brain. The results of this research determined that brains of different people represent object and action categories in a common semantic space.

Based on neuropsychological evidence, humans recognize both the movement (physical) goals and action (semantic) goals of individuals with whom they are interacting. Physical movement goals are related to the kinematics of specific goal-directed movements (e.g., reaching towards the left), and semantic action goals show functional expectations that lead to movement execution (e.g., reaching towards a person to push him). Action goals rely on prior semantic knowledge and could be associated with expectation of multiple movement goals. The recognition of others' semantic action goals can be deliberate or spontaneous. For example, when you are having breakfast with your friend and reaching to the left for a knife, you are able to recognize your friend's physical goal towards a cup. You can also recognize his semantic goal that he is having coffee even though these goals are not directly related to your own behaviour.

1.3 Semantic space

Humans analyse the body posture along with the immediate physical and social setting in which an activity happens to recognize the activity. From the same point of view, we introduce a semantic space as a feature space including *only* features using human understanding to distinguish activities. Therefore, the elements of our semantic space are body parts (pose and poselet), objects related to actions, scenes, and attributes. These elements are described in the following subsections.

1.3.1 Pose

Park and Aggarwal (81) performed the first attempt in considering the human body in order to recognize activities. They estimated human poses using a stick figure model and recognized

interactions between two people. More similar to recent methods, Ikizler and Duygulu (82) discriminated actions according to the configuration of body parts. The body is represented by a set of oriented rectangles using the algorithm of Forsyth and Fleck (83). Rather than localizing the exact configuration of body parts, the distribution of the rectangular regions is extracted. Based on orientations and positions of these rectangles, a histogram is formed in order to define the pose descriptor for each frame. Then, four different methods are utilized to evaluate the performance of the pose descriptor: frame-by-frame voting, global histogramming, SVM classification, and dynamic time warping. Afterwards, human bodies are extracted from stick figures (87), 2D contours (88), or volumetric models such as cones, elliptical cylinders and spheres (89) based on the complexity required in applications. Stick figures, 2D contours, and volumetric models are illustrated in Figure 1.3(a), (b), and (c) respectively.

As discussed earlier, pose is a high-level cue for activity recognition. Pose estimation, meanwhile, is a basic building block of many activity recognition algorithms. It refers to the process of estimating the configuration of fundamental parts or skeletal structure of a person. It has an effect on many tasks such as image understanding and gesture recognition. Recently, pose estimation has become more practical in real-time with the release of Kinect (90). Kinect has the ability to reliably estimate poses of the human user in real time using the output of a depth sensor. This system first determines to which body parts each pixel of the depth image belongs and then uses this information to localize different body joints. It is beyond the perspective of this survey to analyse all approaches in pose estimation. However, a review of pose estimation related to action recognition is covered in Section 1.4.1.

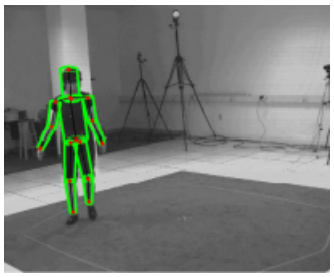
Should pose estimation and activity recognition be performed jointly? A. Yao et al. (91) discussed whether action recognition benefits from pose estimation. They showed that pose-based features outperform low-level appearance features even in the case of heavy noise. Adequate information is hidden in the pose of a human that is sufficient enough to distinguish actions. Pose-based methods are more robust to intra-class variety than appearance-based methods and are more meaningful and compact. However, it is not necessary to correctly extract a complete pose configuration in order to recognize actions.

1.3.2 Poselet

The notation of poselet was proposed first by Bourdev and Malik (92) and further developed by Bourdev et al. (93) for person detection and segmentation. Poselet is described as a specific part of the human pose. A poselet usually consists of more than one limb such as half of a torso and a left hand or a frontal view of two legs walking forward. The main advantage of poselets is that they capture the important parts of the body involved in actions. Therefore, it is possible to recognize actions as long as discriminative poses are detected even if other body parts are occluded or hard to localize. Figure 1.4 illustrates example poselets of the walking action. As we can see, each set of patches may look visually different, but they have a very

(a)

(b)



(c)

Figure 1.3: Pose representation of different models. a) Stick figures of two actions: baseball and badminton (84) b) 2D contours (rectangles) (85) c) Volumetric model (86).

similar semantic meaning.

Poselet was introduced since humans do not always need to see the whole person to make inferences about their activities. Given 3D annotation, the original poselet (92) was proposed based on two criteria to localize people, torso bounds, and key points: 1) configuration space and 2) appearance space. The former space refers to 3D coordinates of body joints and the latter shows pixel values. A poselet should be clustered easily in both spaces. Poselet classifiers were trained to distinguish the visual variation with common semantics. Given human annotation patches, poselet candidates were found by searching and finding the closest patches in a training set. Then, histograms of oriented gradients (HOG) were extracted from these poselets to train linear support vector machines (SVMs). Some poselets were more discriminative than others. Therefore, a Max Margin Hough Transform (95) was used to weight poselets. Some poselets were discarded because they were too close to each other (redundant examples) or there were few examples of them (rare examples) or the trained SVM scored lower than a threshold.

Apart from 3D information, poselets can be extracted from 2D annotations. Not surprisingly, 2D annotation makes the task much simpler than 3D annotation although it carries less information. Bourdev et al. (93) developed an algorithm for detecting and segmenting people using poselets from 2D annotations. It considers the empirical spatial distribution of key-points to cluster poselet activations for the detection of people. They used the spatial distribution of key-points because two consistent poselet activations will make a similar prediction of the position of the person's key-points.

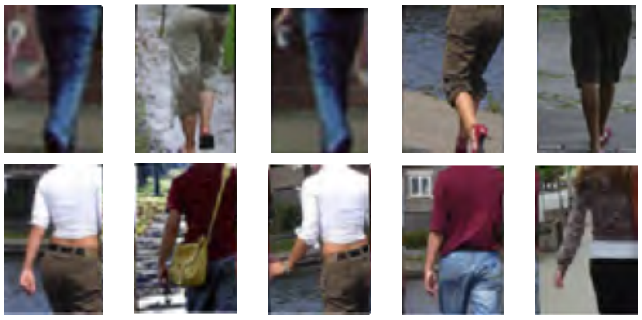


Figure 1.4: Example poselets of the walking action. The top row shows windows that capture legs while images of the bottom row show torsos and limbs. Each set may look visually different, but they have a very similar semantic meaning. The figure has been reproduced from (94).

In addition to people detection, human segmentation and pose estimation are two other applications of poselets. Wang et al. (96) introduced hierarchical poselets for parsing human pose. They proposed that rigid body parts are not necessarily the most salient features for visual recognition and might be confused with rectangles and parallel lines of objects in the background. They used 20 parts ranging from the configuration of the whole body to small rigid parts to represent humans. This representation captures more than one rigid part in addition to primitive parts (i.e. torso, head, limbs). Parameters of the model were learned in a max-margin fashion. Holt et al. (97) proposed another approach for static pose estimation using poselets. Poselets are extracted from training data. A multi-scale scanning window is applied over depth test images to detect poselets. Each window is evaluated by Random Decision Forest (98) as opposed to HOG to identify poselet activations. Body part locations are predicted by poselet activations. A combination of key-point prediction and a graphical model infer overall configuration. Pishchulin et al. (99) estimated human poses relying on poselet-based features in still images. They proposed a conditional model in which all parts are *a-priori* connected based on the pictorial structure (part-based estimation). They exploited poselet to capture complex dependencies between non-connected body parts. The limitation of this model is its dependency on torso detection. To form a feature vector, the position of the torso is first predicted in the test image and then the maximum poselet response in a region around the torso is computed.

The interest in using poselets has led to more research. Motivated by the success of poselets in human detection and pose estimation, recent work uses this context for human action recognition. Poselet-based methods are reviewed in Section 1.4.2. A list of approaches which have used poselets in people detection, pose estimation, and action recognition is shown in Table 1.2. From the table, one can see that poselet-based action recognition methods have a very short history although there are more publications in recent years.

Table 1.2: Summary of different approaches using poselet

First author	Year	People detection	Pose estimation	Action recognition
Bourdev	2009	√		
Bourdev	2010	√		
Yang	2010			√
Holt	2011		√	
Maji	2011		√	√
B. Yao	2011			√
Wang	2011		√	
Zheng	2012			√
Chen	2013			√
Nabi	2013			√
Pishchulin	2013		√	
Raptis	2013			√
Wang	2014			√

1.3.3 Object and scene

The background usually refers to the region without the foreground (human and/or object). It may be taken as the context or scene of a performed action. Background information can be extracted from the whole image, or only from the area of human bounding box, or a combination of both.

The semantic relationship between actions and background settings could be learned as a complementary concept for action recognition. Klaser et al. (100) showed that background suppression limits the classification performance and removes valuable context in realistic settings.

Although the action scene helps action recognition, it may also have negative effects when the scene is too noisy and cluttered. Moreover, one scene may include different actions and not provide helpful information to distinguish those actions.

In addition to the scene context, visual features extracted from a person interacting with a specific object in a specific manner are fundamental to human cognition for recognizing activities. Vaina and Jaulent (101) suggested that action comprehension requires understanding the goal of an action that is related to the compatibility of human movements and corresponding objects. In addition the co-occurrence of humans and objects, human-object configuration is also important. For example, the relative position of arms and a musical instrument distinguishes whether the person plays or simply holds the instrument.

Description

Indoor related:	Yes
Outdoor related:	Yes
Translation motion:	Yes
Arm pendulum-like motion:	Yes
Torso up-down motion:	No
Torso twist:	No
Having stick-like tool:	No

Naming: Walking

Description

Indoor related:	No
Outdoor related:	Yes
Translation motion:	No
Arm pendulum-like motion:	No
Torso up-down motion:	No
Torso twist:	Yes
Having stick-like tool:	Yes

Naming: Golf Swinging

Figure 1.5: Example attributes of walking and golf swinging. The figure was originally shown in (104).

1.3.4 Attributes

Attributes have been introduced as a type of semantic feature to assist in object recognition. Attributes are used by Farhadi et al. (102) to describe objects rather than simply name them, e.g. "metallic car" not just "car". This enables new objects to be recognized with few or no visual training examples. Object attributes can be semantic (metallic body) or discriminative (cars can have a metallic body but animals cannot). Semantic attributes describe parts (a car's wheels), shape (rectangular), and materials (metallic).

Apart from object classification, attributes have been used in describing people and activities. Attributes for person categories can be the gender, hairstyle, types of clothes, etc. Bourdev et al. (103) decomposed people images into a set of parts and poselets, each capturing a salient pattern corresponding to a given viewpoint. A separate attribute classifier is trained for each type of poselet based on the presence of a body part in the attribute. For example, a leg poselet is not used to train the "has-hat" attribute. Moreover, attributes have been considered for the task of activity recognition. Action attributes represent the visual characteristic of the actor or the scene wherein the action takes place. Examples of attributes in golf swinging vs. walking are illustrated in Figure 1.5. From the figure, it is clear that some attributes happen in walking, while some others occur in a golf swing that make it easier to distinguish the activities.

New types of objects/actions that have not been seen in training examples may appear in

the test set (Section 1.8.1). By learning the attributes, algorithms will be generalized to recognize unseen objects/actions. Even though an attribute-based algorithm may not name these new objects/actions, it is able to say something about them. Lampert et al. (105) used semantic attributes to detect new classes of objects. They employed the attributes to transfer knowledge between classes. Attributes are assigned to each class of objects in training and attribute values are predicted at test times and infer the output class label even for previously unseen classes.

However, an attribute-centric representation has the disadvantage of being sensitive to the process of selecting and assigning attributes to relevant action classes. One possible way to select attributes is to manually identify a set of high-level concepts that characterize action classes and choose appropriate attributes for each class of actions. But, a concern here is that if attributes are selected manually, it does not guarantee that all important patterns characterizing an action class are captured. Another issue is whether selected attributes capture intra-class variation. For instance, some examples of the golf swinging action may contain the attribute torso twist and some others may not. Potential approaches addressing these problems will be discussed further in Section 1.6.

1.4 Methods based on body parts

The human body can be detected automatically or labeled manually. Generally, the bounding box and human contour are used to show where the person is in the frame and determine the appropriate region for feature extraction. When performing different actions, body parts are in different poses. Poses can be extracted from the whole human body or some body parts (poselets). In this section, we focus on methods extracting the human body pose for action recognition.

1.4.1 Pose-based methods

In computer vision, human pose estimation refers to the localization of the multiple parts of a human body in an image. It is challenging to extract human pose especially in real world situations since poses are varied during actions and many degrees of freedom need to be estimated.

Analysis of human body parts is divided into different categories. Moeslund and Granum (106) proposed a general categorization for pose estimation. They separated pose estimation algorithms into three categories based on their use of a prior human model: model-free, indirect model, and direct model. Two well-established presentations i.e., part-based (107; 108) and exemplar-based approaches (109; 110) fall into the model-free category. Part-based models represent the human body as a set of rigid parts (e.g. torso, head, and limbs) while exemplar-based methods find images with close whole body configurations and assign poses of those

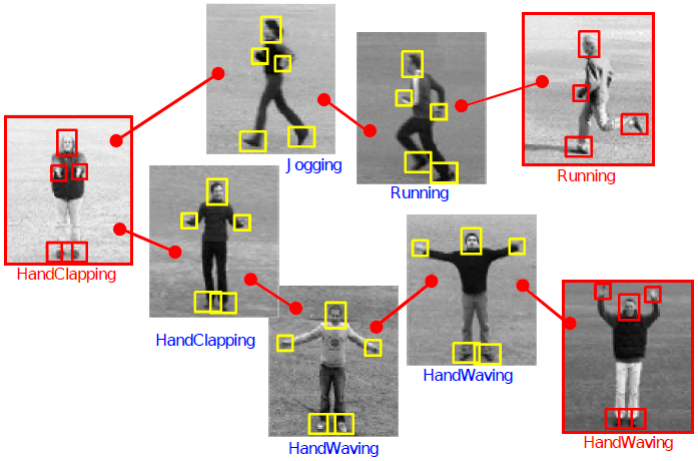


Figure 1.6: Joint pose estimation and action recognition originally shown in Raja et al. (114): Training images (red frames) are manually labeled. Part positions (yellow boxes) and action labels in test images are determined by optimizing the global graph energy.

well-matched training images to a test image. Both exemplar- and part-based approaches have advantages and disadvantages. Exemplar-based approaches are fast but limited in requiring good matching of the entire body. For example, exemplar-based approaches cannot perform well when a test image has common body parts with two different training images. In contrast, part-based approaches detect parts and assemble detected parts into a global configuration. These methods usually do not specify how the whole body looks in the image since the configuration of the body is typically defined as a pairwise relation between body parts.

Methods use *a-priori* model as a reference or look-up tables to interpret extracted data in indirect model classes (111). In this class, typical positions of the head, limbs or a bounding box of the entire human body represent the pose. Direct model class (112) uses 3D models representing poses by kinematic structures and requires more expensive computations during the matching procedure.

The majority of previous pose-based methods focused on fitting a body model to the image. The more complex the model, the better the results obtained. However, it requires more processing and training time.

Activity recognition and pose estimation are typically published as two separate research problems. However, we try to combine these two related research problems in this section. Holte et al. (113) reviewed both human pose estimation and activity recognition. They studied multi-view approaches for human 3D pose estimation and activity recognition. They compared several methods for multi-view human action recognition. Moreover, they discussed the future directions of 3D body pose estimation and human action recognition. Yukita (85) proposed a method that iteratively performs action classification and pose estimation. Initial action

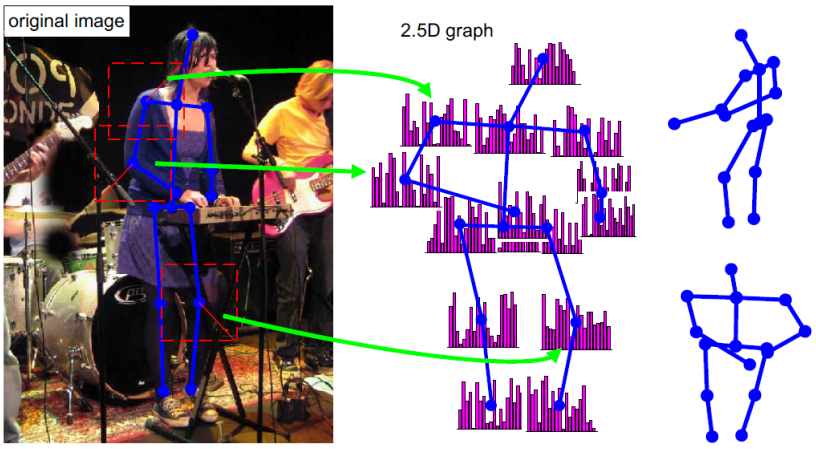


Figure 1.7: 2.5D graph representation, originally shown in Yao and Fei-Fei (116). The histograms represent appearance features extracted from the corresponding image regions. (b) The human body skeleton from the other views.

classification is achieved by employing sets of object detectors to represent scenes and objects as features and SVMs as classifiers. Next, pose estimation is performed using action-specific deformable part models. The estimated pose is then merged with the global features. The normalized relative positions of parts with respect to their centers are used for action re-classification. In another work, Raja et al. (114) connected similar images of actions into a graph as illustrated in Figure 1.6 to jointly address the problem of pose estimation and action recognition. The assumption behind the method is that regions with small distances in the image space will often have similar semantics. A graph is created with five body parts and a class label. There are constraints on the relative positions of body parts and their appearance.

Graphical models have been used widely for pose estimation. These models represent the connections and the relations between different body parts and performers. Poses are usually nodes in the graph and edges depict specific relationships between poses. Mukherjee et al. (115) proposed a graph theoretic approach to recognize interactions. They generated the dominating poses of each performer and used these as nodes of the graph. All possible combinations of dominating poses of two performers, doublets, are created for each interaction and ranked using a graph to produce dominating pose doublets. The distinctive set of dominating pose doublets is selected to represent the corresponding interaction. Another approach using a graphical model is a 2.5D graph representation proposed by Yao and Fei-Fei (116) to recognize actions from single images. It considers key joints of the body (graph nodes) along with spatial relation between key joints (graph edges). Each key joint is represented by 3D positions and 2D appearance features. An exemplar-based representation is used to classify actions. The similarity between actions is measured by matching their corresponding 2.5D graphs (Figure 1.7).

Many approaches model human poses by localizing body joints. A spatial configuration of body joints represents poses. C. Wang et al. (117) used body joints to model spatial pose structure as well as temporal pose evolution to recognize actions in a video. They estimated human joint locations and obtained the best estimated joints for each frame. Then they grouped estimated joints into five body parts and obtained sets of distinctive co-occurring pose sequences of body parts in spatial and temporal domains. For example, the "lifting" action involves the right and left arms moving up concurrently. In the test mode, histograms of detected part sets are created as inputs of SVM classifiers. Meng et al. (118) used the locations of the body joints to recognize human interactions. They called interactions between the parts of the same person and between the parts of different persons intra-person and inter-person interactions, respectively. Joint relative locations are represented as semantic spatial relation features to learn the model.

Some methods model actions as a sequence of key-poses using a compact representation instead of using body poses of all frames. For example in a pushing action, key-poses of the subject (who performs the action) are stepping forward, placing his hands in front, and pushing. For the object (which is pushed), the key-poses are a defensive pose, stepping backward, and falling back. Temporal key-poses reduce the intra-class variation within the same action and increase the inter-class variation between different actions. Eweiwi et al. (119) extracted spatial and temporal information (120) of each pose at each frame of the action sequence. To present the key-poses of each action, K-means clustering is used for each action separately. Chaaraoui et al. (121) represented a human body by contour points. To learn key-poses, they grouped all frames of the same action class into K clusters where the center of each cluster represents an initial key-pose. The process of clustering is repeated to avoid local minima. Results are generated by determining the Euclidean distance between training poses and initial key-poses. The best matches to initial key-poses are taken as the final key-poses. The key-pose learning process is repeated for the training samples of each action class. Cheema et al. (122) proposed a similar approach. However, they assigned rewards and penalties to key-poses. The key-poses which occur only within one action class and are distinguishable have higher weights. Vahdat et al. (123) proposed an algorithm to model interactions as sequences of key-poses. K-means is used to extract various human poses from given trajectories of people in the video. Initial key-pose candidates are provided by the nearest samples of the training set to the K-means centers. The parameters of the model are learned to find the key-poses in a test sequence and recognize the activity class.

Selecting key-poses follows certain rules such as taking into account the temporal order of an action. For example in "walking", crossing two legs should occur between left leg stepping and right-leg stepping. Furthermore, the transitions between different actions have a specific order; "sitting" cannot become "walking" without "standing up" in between. Lv and Nevatia (124) considered these rules and modeled actions from 3D positions of body joints. The difference

between joints of bodies in successive frames is computed and used to define key-poses.

Park and Aggarwal (125) studied the evolution of poses in activities to develop a method for human interaction recognition. The poses of tracked body parts are estimated at the low-level, and the overall body pose is estimated at the high level. The dynamics of the body pose changes during the interaction is analysed by a dynamic Bayesian network. Spatial and temporal constraints are defined to achieve interaction recognition. Spatial constraints are the relative position and orientation of the two persons' body parts. Temporal constraints were defined as causal and coincident relations of body pose changes. For instance, a kicking interaction contains two successive events:"a person moves forward with a stretched leg toward the second person" followed by "the backward movement of the second person" as a result of kicking.

Khan et al. (126) proposed a semantic pyramid approach based on extracted information from the full-body, upper-body, and face to recognize activities from still images. They used a set of pre-trained upper-body and face detectors to exploit semantic information automatically. The best candidate is selected from each body part detector for feature extraction. Semantic information from the full-body, upper-body, and face are combined into a single vector for classification.

1.4.2 Poselet-based methods

As aforementioned, conventional pose estimation methods deal with identifying full body parts and building human pose structure. The drawback of these approaches is that they fail in the case of severe occlusion and clutter. On the other hand, methods based on poselets are more reliable because the ongoing action can be recognized as long as important body parts in that action are visible. Although poselets of the same action might look different in appearance, they are similar semantically. In this section, we review methods that proposed algorithms for action recognition based on poselet descriptors.

Poselet is extracted from body parts and captures the salient body poses related to certain actions. Yang et al. (127) trained a system to localize important body parts for different actions in static images. It is the first work that uses the poselet context for activity recognition purposes. The method does not result in a perfect pose estimate but tries to detect discriminative poses for each action. The pose of the person is learned as a latent variable rather than using a pose estimation method. A four-part star-structured model is used to present the configuration of the body. The parameters of the model are learned to assign a class label to an unseen image. Similar to this approach, Maji et al. (94) conducted tasks of 3D pose estimation and action recognition. Poselet detectors are run in a scanning window over the image. The 3D orientation of heads and torsos of people is estimated using an activation vector. The activation vector is the sum of scores of detected poselets. For recognizing activities,

List of Poselets:
1.LegsExtended
2.HandDown
3.LegsStraight
4.BendHand
5.LegsOpen
6.Hand45°
7.Hand90°
8.LegHigh45°
9.HandExtension
⋮
23.Far Approaching Hands
24.Approaching Hands
25.Hands Contact
26.Person Approaching
27.Pushing Contact
28.LegHigh90°

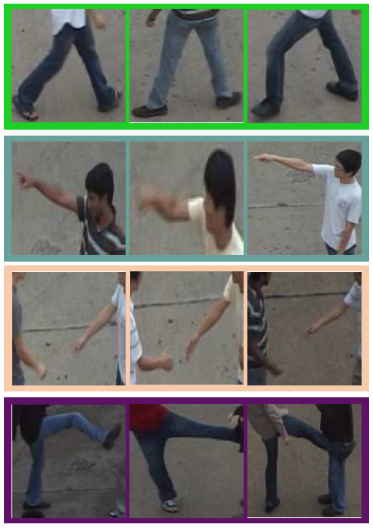


Figure 1.8: Samples of annotated poselets. The figure has been reproduced from (1).

poselet and appearance information are extracted from bounding boxes of people. They also considered the information extracted from the interaction with objects to distinguish actions.

There are important poses related to certain actions. Raptis and Sigal (1) trained a model to recognize actions in videos using key-frames as latent variables while the temporal order of key-frames is important. Key-frames are important poses of actors in certain actions that are learned in a max-margin discriminative framework. They annotated 28 types of poselets for different activities such as extended leg, approaching hand, pushing contact etc (see Figure 1.8 for examples of annotated poselets). Based on these pre-defined poselets and using HOG and BoW features (SIFT, Histogram of Optical Flow and Motion Boundaries), poselet classifiers are learned. The highest score from each classifier is collected and a poselet activation vector is built as the frame descriptor. A multi-class linear SVM is used to combine the scores obtained from each action model. This method is also applicable in the case of dropped frames or partially observed videos.

A poselet-based spatio-temporal method was proposed by Nabi et al. (129) to localize people and recognize multiple activities in a single video. This method models human interactions in crowded environments. A poselet activation pattern over time called TPOS (temporal poselet) was designed to extend the context of poselet. TPOS is obtained by concatenating all spatial poselet vectors of frames. Figure 1.9 illustrates TPOS representing the activations of three different poselets (face, torso, and legs) in time for a 10-frame video of a running person. The activation score of a bank of 150 poselet detectors is calculated. Temporal poselets with

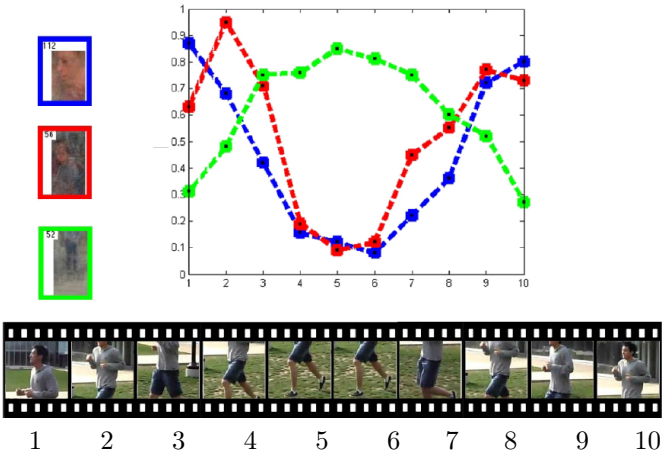


Figure 1.9: Temporal poselets. The correlation of three types of poselets (head, torso, and legs) is illustrated as blue, red, and green lines in a 10-frame video during the "running" activity (top-right). These types of poselets are displayed as profiles (top-left) and in the video (bottom). The figure has been reproduced from (129).

a score lower than a prefixed empirical threshold are removed and a K-means algorithm is used to cluster remaining poselets. Histograms representing the frequency of the K temporal poselets in each video are created. These histograms and corresponding activity labels train a SVM classifier to classify the activity in the input video.

Poselets have also been used to recognize actions from views that are unseen in the training videos. J. Wang et al. (130) took advantage of the Kinect skeleton data and 3D poselet detectors for cross-view action recognition from 2D video input. They designed a data mining method to find discriminative 3D poselets. The view-invariant 3D poselet detectors are trained and applied to all of the frames of the input video. Given an input video from a new viewpoint, the scores of all the 3D poselet detectors projected to all possible views are calculated. The highest detection score among all of the views is utilized as the 3D poselet detection score. Videos are divided into 3-scale pyramids in the spatio-temporal dimensions. The detection scores of the 3D poselet detectors at different scales are pooled and used as the features to train a linear SVM for action classification.

While most approaches use poselet to determine the class of activity, Chen and Grauman (131) proposed a poselet-based algorithm to expand training data from unlabeled videos to recognize activities. Padding training data helps to leverage prior knowledge of the system to be close to the human viewer. This method compares poses of a training set with unlabeled videos and enhances training poses with the matched pose and its neighbors. The system learns how human pose changes over time and uses this information to recognize activities in new images or videos. The poselet activation vector describes pose and domain adaptation, merging real and generated data to train action classifiers. This approach requires tracking information of actors of the unlabeled video. The algorithm is applied on both static images

and videos. Besides action recognition, this method is useful to augment benchmark dataset where data are sparse.

Using both pose and context information, Zheng et al. (128) proposed a method for static activity recognition. Poselet- and context-based classifiers are learned for each action. The former uses a poselet activation vector as features and the latter is obtained by sparse coding on the foreground and background. Given a test image, probabilities of classifiers are summed up. The classifier with the highest score is chosen as the predicted action label. Methods using context are reviewed in depth in the next section.

1.5 Methods based on objects and scenes

Given a human action, there may be objects related to that action. Different actions are related to different objects. Knowing the related objects helps to recognize the corresponding actions. For example, a horse (with a human) is possibly related to the action of "riding a horse" while a phone (with a person) could be related to the action of "phoning". In addition, some actions are executed in certain scenes, e.g. swimming in water and driving on the road. So extracting information from the action context or the whole scene is likely to be helpful for action analysis and recognition.

Learning methods using scene detectors have been used for activity recognition. Marszalek et al. (132) developed a joint scene-action SVM-based classifier by training several scene detectors. They used movie scripts to annotate videos and discover the re-occurring relation between scenes and actions. However, detector-based learning needs the prior knowledge about scene categories and usually depends on the dataset. Therefore, it is not generative and robust to dataset changing. It is also computationally expensive to collect annotated data for each scene category.

As a replacement for detector-based learning, Zhang et al. (133) proposed a generative learning method to recognize actions from scenes according to Multinomial and Dirichlet distributions. They segment each frame into person and background regions. Then, spatio-temporal interest points from the person region as well as color, shape, and local features from the background region are detected. These features are described in HOG, color histograms, Gist descriptors, and SIFT representing a bag-of-feature model.

The problem of local features is that they ignore temporal relation among features. In order to benefit from the dynamic property of actions, Liu et al. (134) proposed a trajectory-based method using scene context to infer activities. Also, Ullah et al. (135) used non-local cues available at the region-level of a video by capturing scene context. Videos are decomposed into region classes (e.g., road, side walk, and parking lot) augmenting local features to provide prior information for action recognition.

Jones and Shao (136) exploited the scene to improve unsupervised human action clustering. They proposed a dual assignment K-means clustering algorithm (DAKM) which captures the relationship between actions and scenes. The algorithm learns two clusterings of a dataset according to two views of the dataset. One view (extracted from motion features) is generated by the action of the video and the other view (extracted from static features) is generated by the scene of the video to improve both action and scene clustering.

Regarding object-based activity recognition, Gupta and Davis (137) conducted research for joint recognition of objects and actions based on shape and motion. They use the coherence between object type, action type, and object reaction to improve the recognition performance. Following this work, Gupta et al. (138) presented a Bayesian method integrating information from humans and objects. They combined contextual information and applied spatial and functional constraints to recognize human-object interactions in videos and static images. The semantic relationships between scene and scene objects are also considered in the static image setting. However, interactions here are limited to three sets of motions i.e., reaching, manipulation, and object reaction. Similarly, a dynamic Bayesian network model is applied in (139) to recognize activities. Radio-Frequency Identification Tags (RFID) and SIFT features are combined to categorize activities and corresponding objects. However, this method does not have a good performance if there is more than one action/object per video, more than one person per action, or occlusion.

Some methods use the fact that spatial configurations and motion patterns between actor and object become constrained by the target object. For example, actors may perform the "grasp a cup" activity at different speeds and with different configurations. However, at the moment of physical contact, actors' motions, appearances, and actor-object spatial configurations become constrained by the manipulated object. Filipovych and Ribeiro (140) showed that constrained motion and spatial configurations are descriptive of the specific actor-object interaction. They proposed a probabilistic framework that automatically learns models linking information about the interaction's dynamics, static appearances, and joint actor-object configurations. Likewise, Kuniyoshi and Shimozaki (141) considered spatial and temporal patterns constrained using neural network-based method.

As mentioned earlier, objects and human poses can serve as a mutual context for each other with respect to HOI activities. For example, knowing that the player is in the starting position with the golf club makes it easier to estimate the player's pose and having a player's pose helps to accurately detect the small golf ball. Models using this mutual context discover the relevant poses for each type of activity, and moreover the connectivity and relationships between the objects and body parts (142; 143; 144; 145).

Some works combine the advantages of both object and scene features in activity recognition. Li and Fei-Fei (146) proposed the first work on classifying events in static images by exploiting

scene and object categorizations. A generative graphical model using appearance and geometry information of local patches for the scene and objects is created to help recognition. Yet, they did not take into account the relationship among the objects and between objects and scenes. Ikizler-Cinbis and Sclaroff (147) combined the features of video elements i.e., objects, scene, and actions in a multiple instance-learning framework. However, they did not learn the relationship between these elements for action identification. They use motion information to extract object candidates along with Gist and color features as scene-centric features. Finally, Han et al. (148) proposed a Gaussian process classification approach for action recognition based on bag-of-detectors considering relations between object parts in the scene.

1.6 Methods based on attributes

As discussed earlier in Section 1.3, attributes are an element of our semantic space. Recent work has shown that attributes are effective features that describe a basic or an intrinsic characteristic of an activity. Occasionally, it is possible to assign an attribute to more than one action. For example, the "riding" attribute can be assigned to both "riding a bike" and "riding a horse". Therefore, it is important to select discriminative attributes which result in more accurate outputs. In this section, we explore several methods using attributes to model human actions.

Liu et al. (104) introduced action attributes to recognize human actions from videos. They modeled attributes as latent variables and formulated the classification problem using a latent linear SVM framework which selects the most discriminative and representative attributes for each action class. This method integrates manually specified and automatically extracted attributes. The absence or presence of each attribute forms a binary vector for the action class. To address the problem of action intra-class variability, they treated attributes as latent variables.

Sharma et al. (149) described activities with the expanded parts model (EPM) by modeling the appearance of humans. This model uses a collection of part templates to explain specific regions in images. The model selects discriminative parts of the action class and skips non-discriminative background regions. The immediate context around the person in the image e.g., the bike in "riding bike" (Figure 1.10) and the grass in "running" is also taken into account. The model scores a test image by representing it with the learned part templates. Parts compete to explain activities and only discriminative ones win. For instance, considering riding bike and riding horse, the person that has a similar pose is removed and the hair and helmet are taken as important parts to distinguish the actions.

An activity can be defined as a sequence of the action attributes. For example, the activity "retrieving an object from a box" is defined as the sequence of the action attributes "insert hand in box", "grab object", "remove hand from box", and "drop object". The modeling of attribute

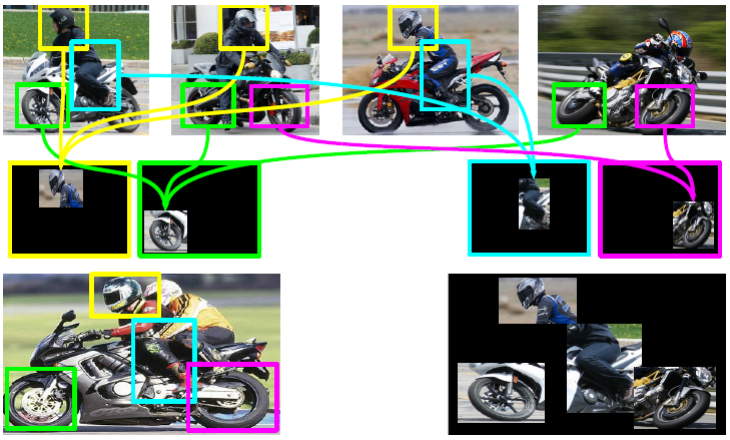


Figure 1.10: Expanded Parts Model, originally shown in Sharma et al. (149): in "riding bike". The model scores an image by representing it with the learned part templates.

dynamics is more discriminant than the modeling of attribute frequencies, i.e. simply recording the occurrence of the action attributes "remove", "grab", "insert", and "drop". In the absence of information regarding the sequence, attribute dynamics help to distinguish the activity "retrieving object from box" from the "storing object in box" which is defined as "remove hand from box", "grab object", "insert hand in box", and "drop object". While most of the attribute-based approaches represent actions as orderless attribute vectors, Li and Vasconcelos (150) proposed an algorithm to model dynamics of activity attributes. They used the binary dynamic system (BDS) to learn both the distribution and dynamics of different activities in attribute space. The BDS combines binary PCA and a least-square problem. A similarity measure between BDSs is introduced to design activity classifiers.

Qiu et al. (151) described an action video by a set of compact and discriminative action attributes. They proposed a dictionary-based method to learn human action attributes. Dictionary learning is an approach to learn dictionary items from a set of training samples. The dictionary items are considered as attributes in this algorithm. They exploited the appearance information between dictionary items and also the class label information associated with dictionary items. The mutual information for appearance information and class distributions is used to define the objective function. The objective function is optimized through a Gaussian process model as a sparse representation that speeds up the optimization process.

In the work of Zhang et al. (152) action classification and attribute classification were assumed to be two separate main and auxiliary tasks, respectively. They did not follow the conventional multi-task learning methods for joint class-attribute learning due to the fact that the attribute and class label contained different amounts of semantic information. They defined attribute regularization as a penalty term which gives a high penalty if outputs of the attribute classifiers are very different from human-defined attributes.

Many human activities are complex and composite such as assembling furniture or food preparation. In order to handle the diversity of composite activities, one efficient way is to represent the activities by shared and transferred contexts across activities and exploit their compositional nature. Rohrbach et al. (153) learned models for a large set of attributes shared across composite activity classes for the recognition of cooking activities. They took into account the co-occurrence and context of each activity and respective objects. Basic-level activities and participants are attributes of composite activities. For instance, pan and onion (participants) and fry (basic-level activity) are considered as attributes of the "preparing onion" composite action. They also used textual descriptions to connect a certain attribute to a specific composite activity.

Some methods use several semantic features to recognize activities. B. Yao et al. (154) exploited poselets, objects, and attributes to assist in action recognition in still images. They defined attributes as a verb related to the action such as sitting, biking, horse riding and parts consisting of related objects (helmet, bike...) and human poselets. For instance, the "riding" attribute is likely to co-occur with objects such as "horse" and "bike", but not "book" while the "pendulum-like arm" poselet is more likely to co-occur with the "walking" attribute. A sparse set of parts and attributes that are meaningful to the contents of each action are used to model actions. Attribute classifiers and part detectors are implemented on a test image. A normalized vector of obtained scores is used to represent the image and an SVM classifier is trained for action classification.

A system using several semantic features must recognize activities reliably and correctly even when one of those features fails. One solution is to design a framework that compensates for the failures of the features with the use of the decisions made by the other feature or recognition results of actions (155).

1.7 Semantic action recognition performance

To understand the present level of semantic action recognition performance, we compare action recognition accuracies obtained in previous approaches. We decided to present the reported action recognition results on the most popular datasets, namely UT-Interaction (156), Willow datasets (157), and Pascal VOC 2010 Action (158).

We report results of representative semantic and non-semantic methods on the three mentioned datasets in Table 1.3, Table 1.4, and Table 1.5, respectively. The performance of approaches is judged based on Average Precision (AP), which is the area under the Precision-Recall (PR) curve. AP has become a standard measure to validate algorithms in action recognition.

From Table 1.3, we can see that the semantic approaches are better than the non-semantic approaches for most cases in the UT-Interaction dataset, except for the work of Meng et al.

Table 1.3: Performance of different methods on the UT-Interaction dataset. The best results are marked in bold.

Methods	Semantic	Non-semantic	AP (%)
Ryoo and Aggarwal (160)		√	70.8
A. Yao et al. (159)		√	88
Delaitre et al. (157)		√	76.73
Matikainen et al. (161)		√	46.58
Ryoo (2) (best)		√	85
Vahdat et al. (123)	√		**93.3**
Kong et al. (162)	√		88.3
Meng et al. (118)	√		87.7
Rapti and Sigal (1)	√		**93.3**
Mukherjee et al. (115) (best)	√		86.67

Table 1.4: Performance of different methods on the Willow dataset. The best result is marked in bold.

Methods	Semantic	Non-semantic	AP (%)
Lazebnik et al. (163)		√	63.7
Delaitre et al. (157)	√		62.88
Maji et al. (94)	√		41
Delaitre et al. (142)	√		64.1
Zheng et al. (128)	√		65.4
Sharma et al. (149)	√		67.6
Khan et al. (164)		√	68
Khan et al. (126)	√		**72.1**

(118) and Mukherjee et al. (115) whose results are slightly lower than the best result in non-semantic methods, obtained by a Hough transform-based voting framework (159). Results are shown as the average of Set1 and Set2 in the UT-Interaction dataset. The UT-Interaction dataset contains a total of 20 video sequences for 6 classes of human-human interactions ("shake-hands", "point", "hug", "push", "kick", and "punch") in two sets taken on a parking lot and on a lawn.

The Willow action dataset is a challenging database for action recognition consisting of 7 action classes and 968 images downloaded from the internet. Action classes are "interacting with computer", "photographing", "playing instrument", "riding bike", "riding horse", "running", and "walking". Experiments show that the semantic method proposed by Khan et al. (126) achieves the best result methods with 72.1% AP in this dataset (Table 1.4).

The VOC 2010 action dataset contains 454 images and 9 types of actions: "phoning", "playing instrument", "reading", "riding bike", "riding horse", "running", "taking a photo", "using com-

Table 1.5: Performance of different methods on the PASCAL VOC 2010 dataset. The best result is marked in bold. SURREY-MK and UCLEAR-DOSP are the approaches presented in the PASCAL challenge (158).

Methods	**Semantic**	**Non-semantic**	**AP (%)**
SURREY-MK		√	62.2
UCLEAR-DOSP		√	61.1
Delaitre et al. (157)	√		52.9
B. Yao et al. (154)	√		65.1
Maji et al. (94)	√		59.7
Delaitre et al. (142)	√		60.66
Zheng et al. (128)	√		**68.8**
Khan et al. (165)		√	62.4
Khan et al. (126)	√		66.35

puter", and "walking". Using body parts only cannot achieve a very good performance on this dataset because there are different activities which share similar poses. These classes also have widely varying object types. Therefore, adding the object model boosts the performance of categories such as "riding bike" and "using computer" significantly. In addition, the context information improves the performance of activities such as "playing instrument" and "running" as these are often group activities. Table 1.5 shows a state-of-the-art comparison on the Pascal VOC 2010 dataset. The three best results belong to the semantic methods of Zheng et al. (128) combining human pose and context information, B. Yao et al. (154) based on semantic pyramids from body parts, and Khan et al. (126) using action attributes and parts with 68.8%, 65.1%, and 63.5% AP respectively.

Both groups of semantic and non-semantic methods have their expected advantages and drawbacks. In particular, body-part based methods rely heavily on the pose difference in actions. If poses of one action are not significantly different from those of other actions, these methods may fail, e.g. running versus walking. However, in the presence of distinguishable poses, they obtained the best performance. This is mainly because of semantic descriptors which act like human perception. On the other hand, non-semantic methods utilize low/mid-level features regardless of the body-parts and attributes and hence they are less sensitive to shared poses/attributes in different actions. However, low-level representations lack semantic interpretation since they usually disregard the context and extract information from the image locally. These descriptors are more sensitive to the appearance and unrelated information such as the background.

1.8 Applications and other semantic approaches

Recent approaches have aimed at structured representation of activities that go beyond the bags-of-words method. They have shown how semantic information is related to an action. Semantic features are not limited to what has been discussed here. We have reviewed the most important features in semantic space. One semantic feature that needs more research is linguistic description (125; 153; 166; 167). It learns relationships between subject/verb/object (S, V, O) of the action. For example, knowing "person" as the subject, "egg" as the object, and "make" as the predicted model, the most likely verb is "cook". Guadarrama et al. (166) mined (S, V, O) triplets from the natural language descriptions and built semantic hierarchies showing semantic relationships among parts of the triplets. Then, they learned a language model from a web-scale text dataset and used it as a prior on triplets to infer verbs.

In some other semantic approaches, a hierarchical representation and reasoning mechanism have been used to recognize activities. A set of rules that encode logical relationships among individual concepts is referred to as a reasoning mechanism. Real-world activity recognition systems typically follow a hierarchical approach. Modules such as background/foreground segmentation, tracking and object detection create the lower levels and action recognition modules create the mid-level. At the high-level, the reasoning engines encode the activity semantics based on the lower level. Thus, it is necessary to understand both hierarchical and reasoning mechanisms to deploy real-world applications. Chen et al (168) used a hierarchical framework for semantic understanding of activities. This method is a bottom-up process to recognize complex activities. First, lower-level actions are recognized by conventional machine learning methods (HMM classification and optical flow features). Afterwards, a resolution based reasoning method (169) is applied to recognize the composite activity using the recognized lower-level actions and logical rule representation. The logical rules reflect the semantic relationship between different actions. Therefore, the hierarchical structure is constructed based on low-level features, mid-level actions, and high-level activity. In other words, the recognition process has a hierarchical mechanism from low-level data to high-level semantic understanding.

In another work, Chen et al. (170) took uncertainty, temporal order, and spatial relationship into consideration and proposed another hierarchical approach based on the semantic understanding and logical uncertainty reasoning mechanism (171). The uncertainly reasoning mechanism indicates whether high-level logical rules hold or not. In a similar way, Ryoo and Aggarwal (172) represented a hierarchical approach composing complex activities into sub-events and specifying temporal, spatial, and logical relationships among sub-events. Recognition of human activities is performed by semantically matching constructed representations with actual observations. Ramirez-Amaro et al. (173) proposed a two-stage framework based on a reasoning mechanism to extract human activities from videos. In the first stage, they extract spatio-temporal features directly from video data to recognize general motions i.e.

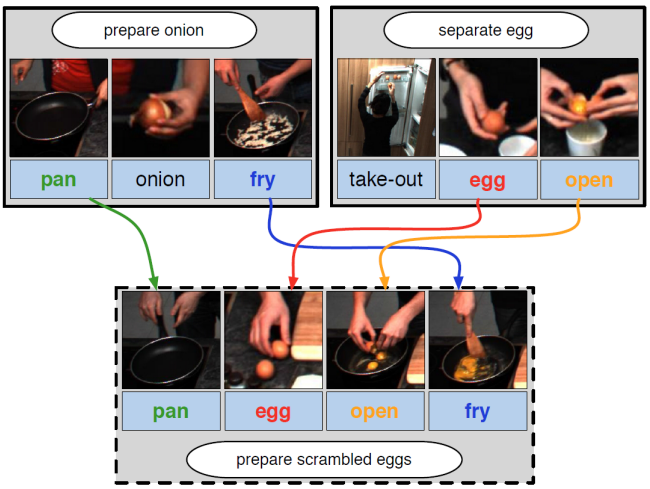


Figure 1.11: Zero-shot learning. Transferring knowledge from known actions (prepare onion and separate egg) to unknown action (prepare scrambled eggs). The figure was originally shown in (153).

moving, not moving or tool used. In the second stage, semantic rules are generated to reason about more specific activities such as reach, take, etc.

Thus far, we have discussed different semantic approaches and the features used in these approaches. Here, we present potential applications where semantic approaches may be of assistance.

1.8.1 Zero-shot learning

A semantic model is a powerful representation for recognizing action categories that have not been seen in the training phase. Semantic data allows the use of external knowledge to determine relevant information for a new activity. This type of recognition is referred to as zero-shot learning (174). It is based on transferring knowledge from known classes (with training samples) to unknown classes (without training samples).

Zero-shot learning represents new actions by incorporating human knowledge rather than training the system for the new input. For example, the new "preparing scrambled eggs" activity can be recognized by knowing activities such as "separate egg" and "preparing onion" as shown in Figure 1.11.

Many human activities share the same basic attributes. For example, the attribute "sitting" can be observed in the activities "watching television" and "working at the office". Cheng et al. (175) developed an attribute-based algorithm to recognize unseen activities. They used raw sensor data and extracted low-level signal features from the processed sensor data. Low-level features are transformed into a vector of semantic attributes and form an activity-attribute matrix. The activity-attribute matrix has the binary values indicating whether a certain attribute is

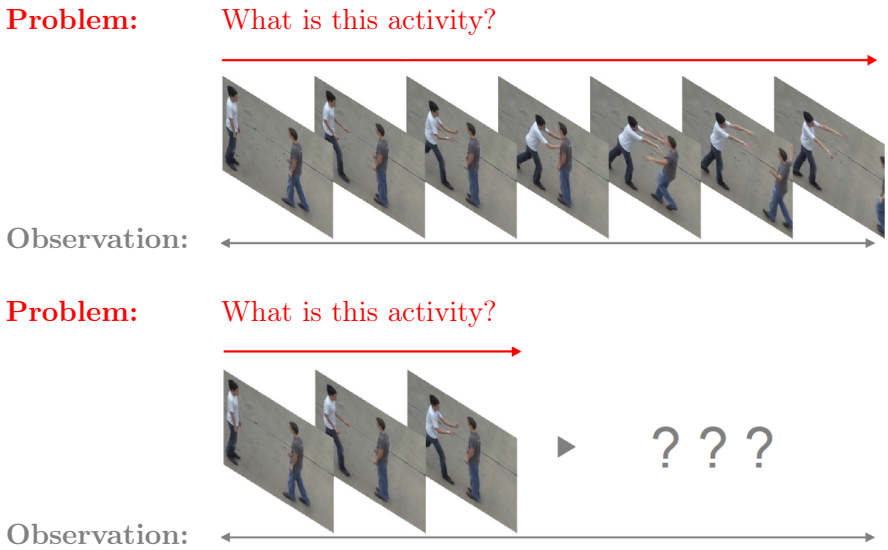


Figure 1.12: Early activity recognition. Top row shows activity classification from fully observed video and the bottom row shows the early activity recognition problem. The figure is reproduced from (2).

associated with an action. Given the activity-attribute matrix, attribute classification classifies the detected attribute as one of the activity classes. To reinforce activity recognition accuracy, the method uses an active learning parameter estimating the uncertainty of the recognition result. If the result is estimated to be highly uncertain, a user is asked to label it as a ground-truth label. Then using the labels, models for attribute detection and activity classification are re-trained and updated.

Algorithms proposed by Liu et al. (104); Qiu et al. (151); Rohrbach et al. (153) (Section 1.6) are zero-shot recognition methods that use attributes. These methods use attributes as a bridge to transfer knowledge from known classes to unknown classes.

Chen and Grauman (131) (Section 1.4.2) used the poselet activation vector to recognize activities from novel single images that were not observed in the set of training static images. They created a pool of plausible poses from training images and unlabeled videos to infer new actions.

1.8.2 Early activity recognition

Approaches recognizing activities from videos mostly classify activities after having fully observed videos. It would be more interesting if the system could recognize activities from partially observed videos and as early as possible. The overall goal of early activity recognition methods is illustrated in Figure 1.12 schematically.

Early recognition of ongoing activities has several practical applications. Early recognition

can help when the whole video stream is not available and activities are not recorded from the start to the end. It can also be useful in surveillance applications to identify thieves before a robbery occurs.

Although the bag-of-words feature is a popular method in human activity recognition, it is not an appropriate method for the task of early recognition since it ignores the temporal or sequential information of the activity. Furthermore, the initial part of the activity may not be sufficient to build an appropriate bag-of-words representation. Instead, semantic representations are useful for early recognition. Using semantic features allows an action to become distinguishable when we have access to only a fraction of frames since similar actions share the same meaning across frames. The semantic approach by (1) outperforms other existing approaches in early activity recognition. As mentioned before, (1) uses a poselet representation to extract key-frames and recognize activities.

Ryoo (2) performed the first attempt in early activity recognition. He represented an activity as an integral histogram of spatio-temporal features and modelled the distribution of features over time. His recognition methodology is named dynamic bag-of-words.

Li and Fu (8) proposed a method to model the temporal order of activities for early recognition using an autoregressive moving average model, ARMA. An ARMA model is a tool for understanding and predicting future observations of a time series. A combination of HMM and ARMA is used in this method.

In another work, (9) proposed an approach to train temporal event detectors for early event detection. The algorithm detects the temporal location and duration of an activity from a video. The performance of the method is evaluated on facial expressions, hand gestures and action datasets.

1.8.3 Gapped-video based activity recognition

Activity videos may include a temporal gap (e.g. dropped signal when making a video) that may occur any time and with any duration. If the video stream contains several activities sequentially, the temporal gap may divide the video into two different activity subsequences and make the recognition more challenging.

Cao et al. (176) conducted the only work in recognizing human activity from gapped videos using spatio-temporal features. Each activity is divided into temporal segments and sparse coding is applied to extract the activity likelihood for each segment. Likelihoods are combined for all segments to achieve a global posterior for the activity. The posterior is maximized to recognize the activity.

Learning semantic contexts enables an algorithm to describe the relation between the action category and each frame even in the presence of multiple actions in a gapped video. Therefore,

semantic approaches are likely to be helpful for gapped-videos.

1.8.4 Activity forecasting

Activity forecasting refers to predicting future unobserved actions and it is different from simply classifying a partial sequence as a given activity. Extracted semantic information such as scene understanding helps to forecast actions. Recent semantic scene labeling methods propose reliable ways of recognizing scene features such as pavement, building and car (177; 178). Having the knowledge of the human preference for using a certain physical environment (sidewalks, streets, etc.) allows us to perform higher levels of reasoning concerning future human actions. For instance, people desire to walk on sidewalks rather than streets. Therefore, when a person wants to approach his car which is parked further away, we predict that he will walk on the sidewalk as long as possible to reach his car. Furthermore, we can benefit from human knowledge of possible consequences of executing an action as semantic information for activity forecasting (179) e.g., what do I learn by doing this action? (immediate rewards), what will be the consequences of my actions in the future? (expected future rewards), and what do I intend to accomplish? (goals).

Kitani et al. (180) combined semantic scene understanding and noisy tracker observations to forecast activities. They predicted the walking path of a person in an urban environment such as road, sidewalk, and entrance based on historical data.

1.8.5 Activity analysis

Identification and description of activities are referred to as activity analysis. For example in some sports training applications, it is necessary to carefully analyse the video for effective training. Sports analysis can be performed by searching important poses of the athlete's body in the video. More generally, an activity can be analysed by extracting a few atomic key actions using semantic evidence.

Wu et al. (181) search key-frames in sport training videos based on poselet representation. An independent classifier is trained for each poselet. Classifiers are applied on test frames and poselets are detected in multi-scale scanning mode. The number of detected poselets by the first poselet classifier is computed for each frame. A frame with the maximum detection number is chosen as the key frame of the first poselet. Key-frames of other poselets are obtained in a similar way.

Some of the mentioned applications have already used semantic approaches, but they are still in their early stages. We believe this is just the beginning of using semantics and human knowledge in action recognition and it still requires more research when compared to the intensive studies in spatio-temporal based recognition.

1.9 Conclusion

Recent action recognition methods rely on low-level and mid-level features such as spatio-temporal interest points and trajectories. Although these methods provide reasonable results on several datasets, they fail with complex data due to the lack of semantics they represent. Several new approaches have aimed at structured representation of activities that go beyond low/mid- level features. We have focused on recent action recognition frameworks based on semantic information in this paper. We introduced a semantic space which mainly includes pose, poselet, object/scene context, and attributes. Linguistic descriptors and reasoning-based hierarchical semantic representation have also been used as semantic features. Different action recognition methods have been proposed using these meaningful features. We discussed that relying on low-level features is sub-optimal due to significant variation in scales, viewpoint, and pose in real-world data. Therefore, semantic descriptions that can capture meaningful information and are robust to visual variations are needed.

Experiments show that semantic methods outperform non-semantic based methods in most cases except when different activities share similar poses/attributes. In such cases, combining several features improves the performance.

In order to exploit the full potential of semantic methods, certain topics need further investigation. We have discussed applications of semantic approaches. One promising direction for exploiting semantic methods is recognizing new activities that have not been seen in training examples, referred to as zero-shot learning. Models using semantic information are able to incorporate human knowledge to describe new activities. Learning a model with human-specified knowledge is similar to the young infants learning process which enables them to interpret new observed actions. Semantic models, furthermore, allow an action to be distinguished even when we have access to only a fraction of frames. Semantic methods are useful in such a case since similar actions share the same meaning and characteristics across frames. Thus, extracting common characteristics enables the algorithm to recognize the ongoing activity although some frames are missing. There are other innovative topics in which semantic methods are likely to be helpful such as activity forecasting and activity analysis. What all of these topics have in common is the understanding of human activity as a core component of various kinds of context-aware and user-centric applications.

Semantic action recognition is a relatively new area. The existing work still requires more research when compared to the intensive studies in spatio-temporal based recognition. Future algorithms may better explore this aspect to connect algorithms to human knowledge. We hope this survey motivates future researchers to devote more attention to semantic information in modeling human actions.

Chapter 2

Time-slice Prediction of Dyadic Human Activities

Abstract

Recognizing human activities from video data is being leveraged for surveillance and human-computer interaction applications. In this paper, we introduce the problem of time-slice activity recognition which aims to explore human activity at a smaller temporal granularity. Time-slice recognition is able to infer human behaviours from a short temporal window. It has been shown that the temporal slice analysis is helpful for motion characterization and in general for video content representation. These studies motivate us to consider time-slices for activity recognition. To this intent, we propose a new family of spatio-temporal descriptors which are optimized for early prediction with time-slice action annotations. Our predictive spatio-temporal interest point (Predict-STIP) representation is based on the intuition of temporal contingency between time-slices. Furthermore, we introduce a new dataset which is annotated at multiple short temporal windows, allowing the modeling of the inherent uncertainty in time-slice activity recognition. Our experimental results show performance comparable to human annotations.

2.1 Introduction

Humans are good at anticipating and correctly predicting the activities of others during social interactions. For example, we do not need to see a full handshake before being able to recognize it. In fact, two people getting closer and lifting hands will most likely shake hands. Humans can naturally model the uncertainty associated with activity recognition. While great progress has been made in computer-based human activity recognition this past decade, computational algorithms are often lacking the predictive capabilities of humans. Also, most recent approaches are expecting a complete video with a large temporal window. Based on intuition from social psychology, we introduce a time-slice approach to human activity recognition which is based

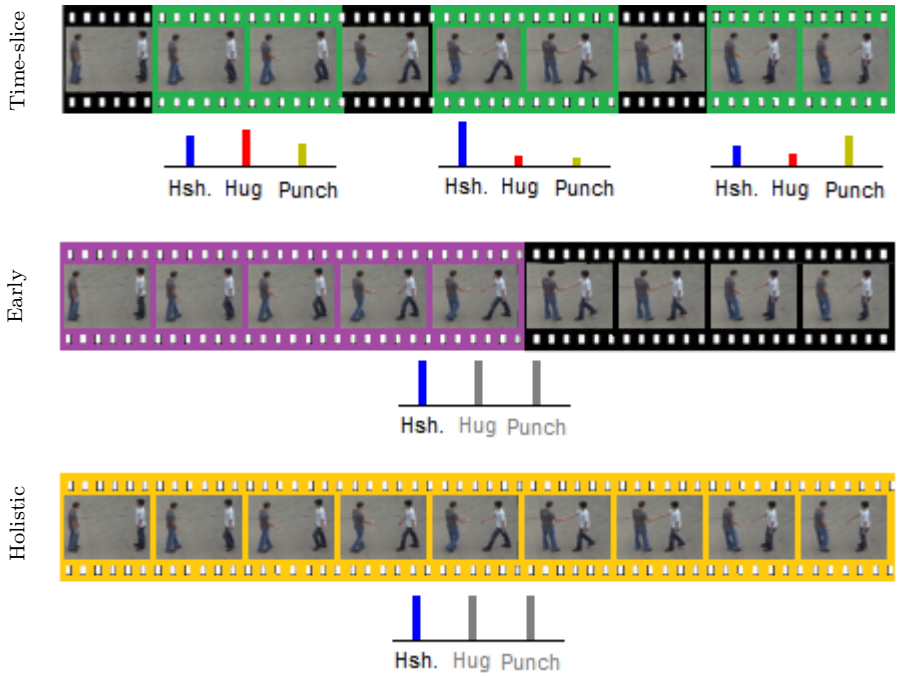


Figure 2.1: An illustration of human activity recognition problems. The first row illustrates "time-slice" recognition and the labels, i.e., Handshake (Hsh.), Hug, and Punch for different time-slices. The second and third rows show "early" recognition and "holistic" approaches where the label is the same for the whole sequence.

on short-term observations. We are interested in improving our understanding of the inherent uncertainty occurring with time-slice observations and building computational algorithms to properly model them. This work has several practical applications, outside the basic research question of better understanding human and computer perception of dyadic actions. It can be beneficial when the whole video stream is not available and activities are not recorded from the start to the end. It can also be useful in video indexing, retrieval, and analysis.

We present in Figure 2.1 an overview of our approach based on time-slice action prediction and contrast it with the conventional approaches which recognize actions based on either the whole video sequence (referred as "holistic" approach) or the first part of it (early recognition) (2). Our time-slice approach studies not only the beginning of the action sequence but generalizes this to any short-term observation anywhere in the video sequence. Another key novelty is in the explicit modeling of the uncertainty occurring when predicting actions based on time-slices.

In this paper, we propose a new set of spatio-temporal descriptors using time-slice action annotations for early activity prediction. We show our predictive spatio-temporal interest point (predict-STIP) representation is able to infer time slices of human activities based on discriminative descriptors. We select feature descriptors which are discriminative when an action is clearly occurring during a time-slice and is also visible outside on time-slices with

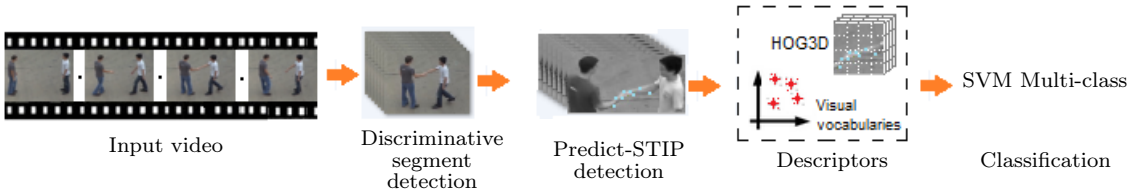


Figure 2.2: An overview of our method, Predict-STIP. Given an input video sequence, we first extract discriminative segments and then detect Predict-STIPs. HOG3D representation and BoW models are applied to prepare inputs for SVM classifiers.

uncertain action. Given their broader temporal range, we hypothesize that these descriptors can be better at prediction actions. An overview of our method "Predict-STIP" is illustrated in Figure 2.2. Our goal is to identify descriptors with broad temporal coverage. The details on how we do it can be described later. Our representation is amenable to early activity recognition. We show the comparison results of this work with the state-of-the-art in early activity recognition.

We introduce a new dataset, named Time-slice Action Prediction (TAP) dataset, to evaluate our proposed feature descriptors and enable future research on this topic. Our dataset could also be used for early activity recognition as well as holistic activity recognition. The dataset was created by extracting time-slices from existing public human action datasets and perform a perception study with multiple annotators giving continuous ratings for each action. The continuous ratings allow to represent the uncertainty in time-slice action prediction.

The outline of the paper is as follows. Section 2.2 provides an overview of the most relevant works to our paper in activity recognition. We present our new dataset in Section 2.3. Section 2.4 explains the methodology of our proposed method. Section 2.5 shows our experimental results, followed by conclusions in Section 2.6.

2.2 Related Work

A number of surveys have been published in activity recognition over the past decade (7; 6; 5). Given the significant literature review in this area, we focus only on the most relevant works.

Partially observed videos: Very few works have been devoted to recognizing activities from partially observed videos. Ryoo (2) performed the first attempt in early activity recognition and studied how feature distributions change over time. Li and Fu (8) used an autoregressive moving average model, ARMA, to model the temporal order of activities for early recognition. Raptis and Sigal (1) trained a model to recognize actions in videos using key-poselets as latent variables for partially observed activity recognition. Yu (10) trained a model using relative locations of space-time points to the center position. A Semantic framework was proposed

by Li et al. (11) for early recognition of long-duration complex activities by discovering the causal relationships between action units. Early event detection and recognition of human activity from gapped videos have also been studied in (9; 12) which used partially observed videos as input.

Space-time interest points: Recently, space-time interest points (STIPs) have received increasing interest due to their reasonable performance for activity recognition. STIP-based methods are invariant to geometric transformations which result in low variation by changes in scale, rotation, and viewpoint. Laptev and Lindeberg (13) proposed the notion of STIP built on the idea of the Harris and Stephens interest point operators (14). Several other methods have been reported (15; 16; 17) to improve STIP detection for human activity recognition. Chakraborty et al. (18) proposed a model for robust Selective STIP detection (S-STIPs) by applying background suppression as well as local and temporal constraints. This method outperforms existing STIP detection techniques and detects more stable and distinctive STIPs. We benefit from the advantages of S-STIPs to extract the initial interest points in our work. For exploring more approaches, we refer readers to a recent comprehensive survey of human action recognition with STIP detector by Das Dawn and Shaikh (19).

Key-components: The use of informative components (frames or time-slices) is in contrast to most research in video-based action recognition which often extracts features from much longer videos. Using a sparse set of frames allows the model to focus on the most discriminative parts of the action which are referred to as key-frames in literature review (20; 21; 22; 23). Key-frames are discrete sets of frames that capture discriminative parts of a video. On the other hand, time-slices are continuous sets of frames which represent temporal ordering and dynamic structure of the discriminative part of a video. This paper is the first effort to introduce time-slice for activity recognition.

Trajectory data: Among the local space-time features, tracking interest points through video sequences have been shown to be an efficient representation for action recognition (24; 25; 26). Shape, appearance, and motion descriptors are extracted from the trajectories of interest points to analyze detailed levels of human movements. Sun et al. (26) represented activities using trajectory transition and trajectory proximity descriptors. The trajectory extraction process is based on matching SIFT descriptors between two consecutive frames. The descriptors that are too far apart are discarded. Wang and Schmid (24) proposed a method using improved dense feature trajectories. They estimated the camera motion and removed it from the optical flow to have better motion-based descriptors. In this paper, we track specific spatiotemporal interest points backward and forward in time and extract predictive features based on the persistency of this trajectory data.

2.3 TAP Dataset

We are interested in social interactions, more specifically dyadic interactions. We use publicly available datasets so that people can replicate and extend our experiments. We focus on datasets with similar action labels in order to make the time-slice annotation task possible for crowd-sourcing.

We have extracted 2119 time-slices from 4 challenging datasets, i.e., UT-Interaction (segmented sets 1 and 2) (27), HMDB (28), TV Interaction (29), and Hollywood (30) datasets. Each time-slice contains one of seven interactions: handshake, high five, hug, kick, kiss, punch, and push. We performed a preliminary experiment to validate how many frames were necessary to have good agreement between annotators. We requested some annotators to recognize dyadic interaction examples with 5-, 10- and 15-frame time-slices. We decided to choose 10-frame time-slices for our work since 5-frame time-slices were too short and 15-frame time-slices were not fit to our goal which is studying the inherent uncertainty in activities. Our dataset is available as a public dataset to encourage researchers to continue this line of research. [1]

During our experiments, we grouped videos from constrained and unconstrained datasets. Constrained, here, refers to the restriction in the settings and activity execution. UT-interaction is our constrained dataset which contains acted interactions with a fixed background and profile viewpoint that are performed for research purpose. On the other hand, unconstrained datasets include activities which are taken in realistic settings, e.g. from TV shows. Unconstrained datasets are more challenging for activity recognition. HMDB, TV Interaction, and Hollywood are our unconstrained datasets. We selected videos of these datasets based on a camera angle ranging from -45 to +45 degree.

All time-slices were annotated by multiple online annotators (using the Crowdflower platform (31)). Three annotators rated each time-slice on how likely a specific action is occurring. For each time-slice and for each action, the annotator was asked to pick one of 5 likelihoods, i.e., definitely not occurring, unlikely to occur, neither likely nor unlikely to occur, likely to occur, and definitely occurring. Since time-slices are very short video clips (10 frames), annotators were allowed to replay each time-slice as many times as they wanted before making decisions. A sample design of our tasks in the Crowdflower platform along with some reports on contributors are illustrated in Figures 2.3 - 2.5.

Figure 2.6 illustrates how annotators rated two example videos. From the figure, one can see the confusion and uncertainty of annotators in the first time-slices of videos. As time passes and more information about the activity of interest is observed, annotators will better recognize the activity.

[1] http://vision.gel.ulaval.ca/users/maryam/TAP/

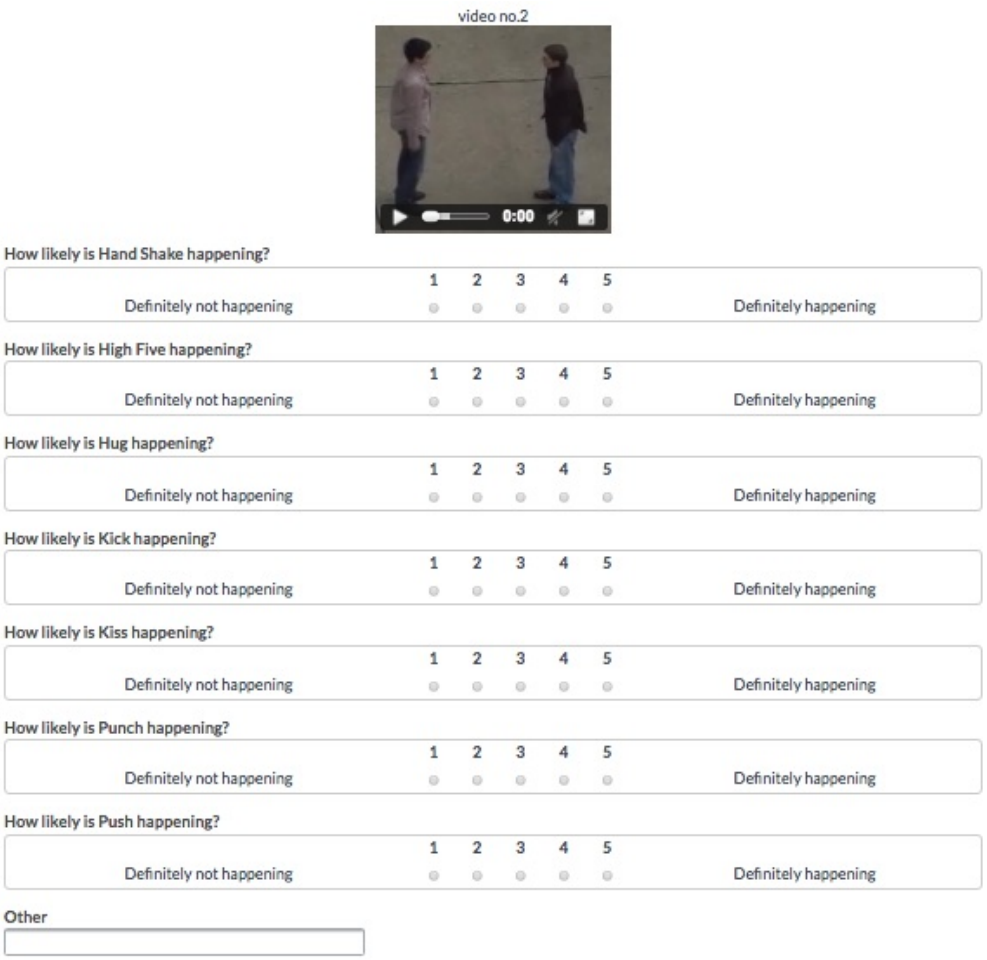
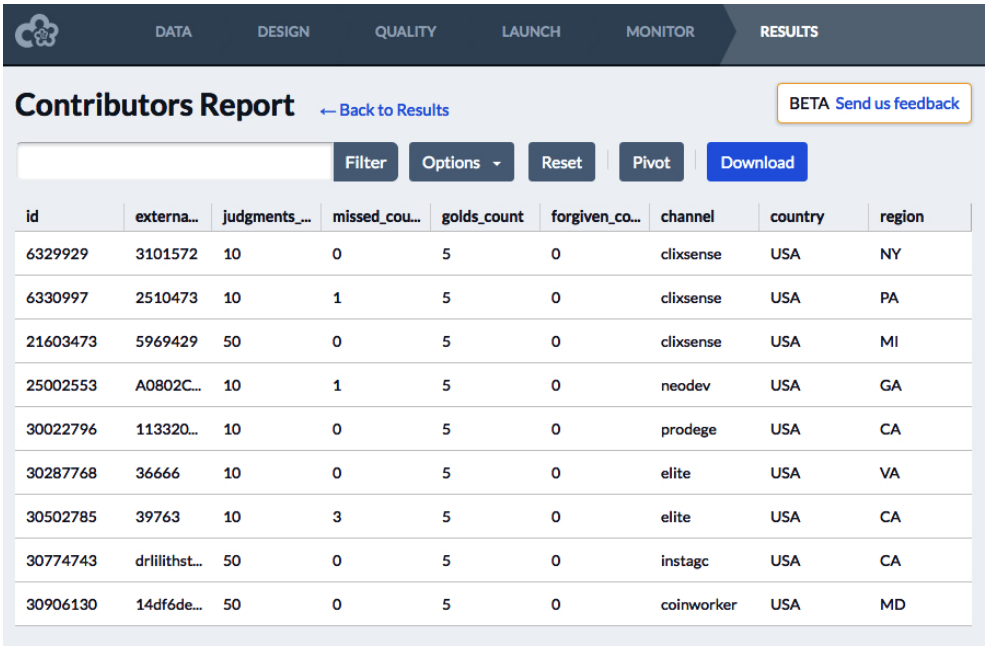


Figure 2.3: Sample design of our tasks in the Crowdflower platform. Choices 1 to 5 are 5 likelihoods which are defined in the instruction.

2.4 Methodology

Our approach to dyadic human activity recognition consists of three major contributions: i) a new learning approach in which discriminative video segments are used on the basis of human annotations and efficient spatio-temporal features (called predictive) are obtained on the basis of their persistence and ii) a more general definition of the activity recognition problem in which the uncertainty arising from observing a short time-slice from anywhere in the video sequence is explicitly taken into account, and iii) a demonstration that a baseline multi-label classification method can reproduce the features of the human annotation using the proposed learned model for this problem. We introduce discriminative segments since we need feature descriptors which are good when humans agree that an action is clearly occurring. We also require descriptors which have predictive powers when their broader temporal range is considered. We first determine discriminative segments of each video activity based on annotated data where all



id	externa...	judgments_...	missed_cou...	golds_count	forgiven_co...	channel	country	region
6329929	3101572	10	0	5	0	clixsense	USA	NY
6330997	2510473	10	1	5	0	clixsense	USA	PA
21603473	5969429	50	0	5	0	clixsense	USA	MI
25002553	A0802C...	10	1	5	0	neodev	USA	GA
30022796	113320...	10	0	5	0	prodege	USA	CA
30287768	36666	10	0	5	0	elite	USA	VA
30502785	39763	10	3	5	0	elite	USA	CA
30774743	drlilithst...	50	0	5	0	instagc	USA	CA
30906130	14df6de...	50	0	5	0	coinworker	USA	MD

Figure 2.4: Contributors report for a sample task in the Crowdflower platform.

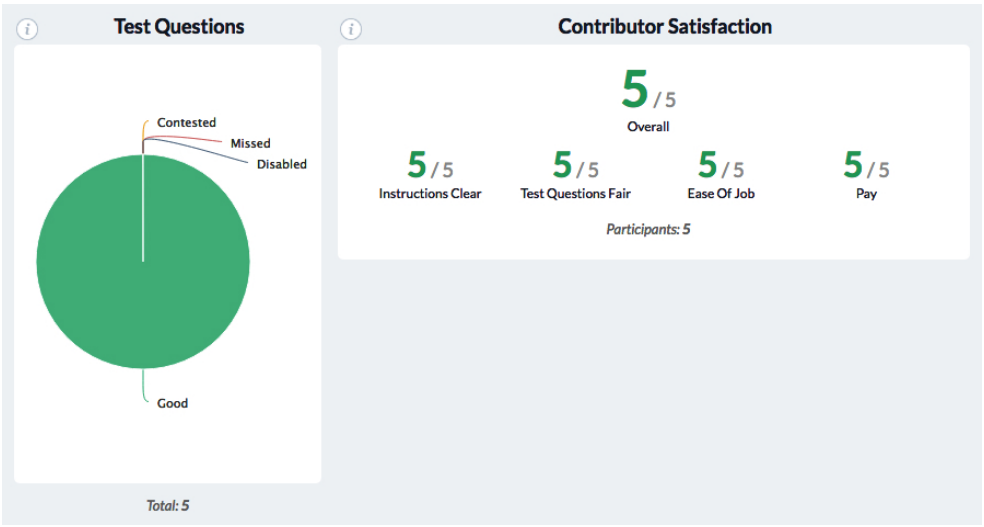


Figure 2.5: Contributor satisfaction for a sample task in the Crowdflower platform.

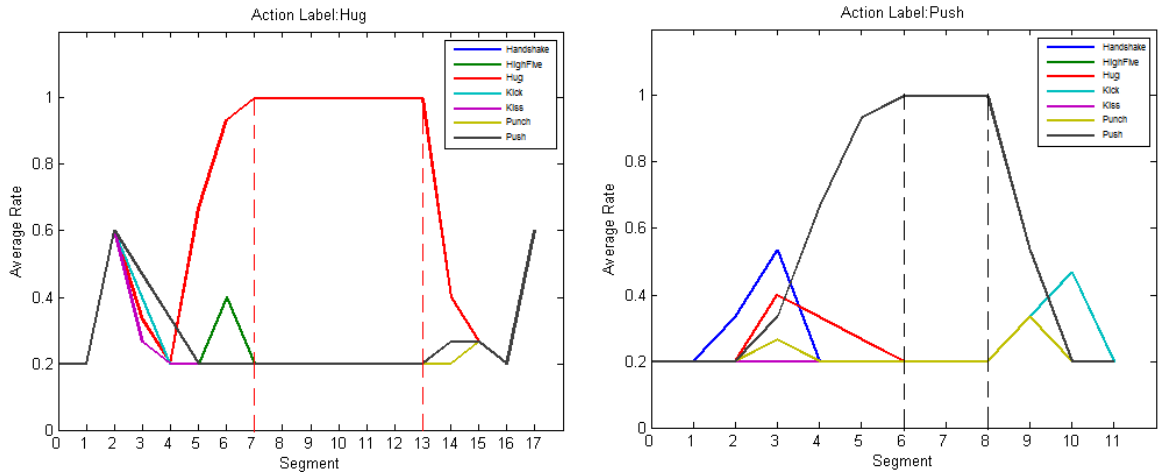


Figure 2.6: Human annotation. This figure shows the average rate of 3 annotators for two video examples: hug and push. For each possible activity and for each time-slice, the label provided by one annotator is first converted to a number on a linear scale from 0 to 1. The average of those numbers for more than one annotator (we used 3 here) is called the average rate annotation for the time-slice. This average rate is used to evaluate the performance of our method. Time-slices between dashed lines is the discriminative segment of the interaction given to the annotators.

annotators agreed an interaction of interest is occurring. We then use these segments to select predictive space-time interest points. Each predictive point is described by motion and appearance descriptors to learn the model. In the following subsections, the above steps are explained in more detail.

2.4.1 Discriminative segments

When analysing an interaction, we can definitely recognize the ongoing activity from specific time slices such as "two people are shaking each other's hands" slice in handshaking activity. These slices are referred to as discriminative segments in this paper. Discriminative segments, therefore, encode the most relevant slices of video to interested interaction. We define the temporal location of a discriminative segment based on annotated data. In preparing our dataset, we asked 3 users to annotate each time-slice as described in Section 2.3. To measure the reliability of agreement between annotators, we used Fleiss' kappa coefficient k (32) that assesses the agreement between more than two raters. This coefficient takes into consideration the agreement occurring by chance as shown in Equation 2.1. For each interaction video, time-slices where the annotators are in complete agreement, i.e. k=1, on definitely including the interaction of interest, are selected as discriminative segments.

$$k = \frac{P_l - \bar{P}_i}{1 - \bar{P}_i} \tag{2.1}$$

where $\bar{P}_i$ is the mean value to which annotators agreed for the certain interaction of interest and P_l is the summation over the square quantity of all time-slices assigned to certain likelihood categories. $1 - \bar{P}_i$ shows the degree of agreement that is attainable and $P_l - \bar{P}_i$ shows the degree of agreement that actually achieved. k between 0.81 and 1.00 shows an almost perfect agreement.

2.4.2 Predict-STIP

In this paper, we follow the recent progress in STIP-based recognition strategy. Existing STIP detectors are vulnerable to model the inherent uncertainty in partially observed action recognition and prediction, and therefore, are insufficient for time-slice recognition. To overcome this problem, we introduce a predictive representation which measures how long STIPs are observable in a video. STIPs which are active during the whole video are selected as Predict-STIP (P-STIP). In other words, P-STIPs are the STIPs that exist in first frames of the video and will still appear in upcoming frames.

Given a set of interaction video sequences $\{A_i \mid i = 1 : n\}$ and their associated discriminative segments $\{S_i \mid i = 1 : n\}$, our purpose is to detect P-STIPs P_i of each A_i. Our input variables are sequences of frames $A_i = \{f_i^1, ..., f_i^{e_i}\}$ and $S_i = \{s_i^1, ..., s_i^{N_i}\}$ where e_i and N_i are the length of the full video and the discriminative segment, respectively. To extract P-STIP, we first detect "$stip_{New}$" of s_i^1 as initial landmarks. We then track them backward and forward using Kanade-Lucas-Tomasi (KLT) algorithm (33; 34) to f_i^1 and $f_i^{e_i}$ and check whether or not they have existed during the whole video. We repeat these steps for all frames of S_i. Landmarks that are continuously observable are selected as P-STIPs P_i:

$$P_i = \{p_{(x_{j,t}, y_{j,t})} \in s_i^t, \quad t = 1, ..., N_i \mid \forall p \quad V_p = 1\} \tag{2.2}$$

where V is a validity matrix providing a logical array, indicating whether or not each point has existed during the whole video. $(x_{j,t}, y_{j,t})$ is the position of $stip_{New}$ p_j in the frame s_i^t.

To speed up the tracking step and increase the efficiency of our algorithm, we select a new subset of S-STIPs (18), $stip_{New}$, instead of using all densely sampled S-STIPs from S_i. We initialize $stip_{New}$ as S-STIPs extracted from s_i^1 and track them. We then generate putative matches between previously tracked-$stip_{New}$, $stip_T$, and extracted S-STIPs, $stip_E$, of the current frame by finding points that have minimal differences in oriented phase data within windows surrounding each point (35). Only points that correlate most strongly with each other in both directions are returned as matched points, $stip_M$. Oriented phase data matcher performs better compared to normalized grayscale correlation. We also set a maximum search radius threshold for matching points to improve speed and accuracy since we do not want to match points, e.g. from an arm with points extracted from a leg. Consequently, only points whose Euclidean distance is below the threshold are considered for matching. Afterward, we

Algorithm 1 Predict-STIP detection from a discriminative segment

Input: Discriminative segment ($H \times W \times N$): S;
$S = \{s^i \mid i = 1 : N\}$ (contains all frames of a discriminative segment)
Definition:
f_1: The first frame of the full video
f_e: The last frame of the full video
V: Validity matrix provides a logical array, indicating whether or not each point has existed
Ensure: Predict-STIP: *PredectivePoints*

1. $N = size(S, 3)$; (Total no. of the discriminative segment's frames)
2. Initialize $stip_{New}$
3. Initialize *stip*
4. **for** $i = 1 \rightarrow N$ **do**
5. Track $stip_{New}$ backward to f_1 and forward to f_e and restore V matrixes
6. Let $stip_T$ be *stip* tracked from s^{i-1}
7. Let $stip_E$ be S-STIPs extracted from s^i
8. Match $stip_T$ with $stip_E$ and set $stip_M = stip_T \cap stip_E$
9. Update $stip_{New}$ via $stip_E \notin stip_M$
10. Update *stip* via $stip_{New} \cup stip_T$
11. **end for**
12. Check V matrixes
13. **Find** points where V_{points} are always equal to 1 and set as *PredectivePoints*
14. Return (*PredectivePoints*)

employ RANSAC algorithm (36) to exclude outliers and identify strong inliers. Figure 2.7.a illustrates the matching result of a sample frame. Finally, we update $stip_{New}$ via $stip_E$ that does not belong to $stip_M$. Therefore, the $stip_{New}$ is a new subset of S-STIPs that are not tracked from previous frames and still appear in each frame. The pseudo code for the full predictive feature detection is described in Algorithm 1. Figure 2.7.b also shows P-STIPs extracted by our approach compared with S-STIPs resulting from (18).

2.4.3 Descriptors and vocabulary building

Several local and global descriptors have been proposed in the past few years for STIP-based methods (37; 38; 39; 40). In this paper, we use HOG3D descriptors (40) to represent each interaction video. The HOG3D descriptor is based on histograms of 3D gradient orientations, where mean gradient vectors are computed using integral videos. With integral videos, 3D gradients can be efficiently calculated for any arbitrary point in a video. Given P-STIPs of each interaction video, we construct the HOG3D representation. Local regions are determined first by extracting P-STIPs and then histograms of gradient orientations are computed over a set of gradient vectors from the cuboid neighbourhood (4x4x4) around the P-STIPs. All histograms are concatenated to one descriptor vector for each video.

We compute the basic Bag-of-words model and quantize the descriptor vectors, HOG3D ex-

(a) (b)

Figure 2.7: **Predict-STIP** detection. The matching result of a sample "high five" action is shown in the left figure. The right figure displays S-STIP (18) and our predict-STIP extracted from the example.

tracted at P-STIPs, into 1000 bins associated with visual words using K-means clustering. BoW features are normalized so their L1 norm is 1.

2.4.4 Learning

The goal of our Predict-STIP method is to determine the interaction category of time-slices of video X among a set of classes $\{1, ..., K\}$. Therefore, our purpose is to learn a mapping $f(O) \rightarrow \{1, ..., K\}$ where $O \subset X$ refers to the time-slice observations and may occur at any time in the video. We present the videos with BoW descriptors obtained from P-STIPs. For each class of interaction, we learn a model with the corresponding BoW descriptors using multi-class SVM framework in the training phase.

At test time, a query video v_i which is a time-slice of a longer video is matched to the models according to the learned appearance and motion predictive features. To this intent, we extract S-STIPs (18) from v_i and match them to the pool of trained P-STIPs. S-STIPs of v_i that match to P-STIPs are selected as descriptors of P-STIPs of v_i (lookup table technique). HOG3D are applied on P-STIPS and BoW descriptors of v_i are extracted. Classification is made based on the score of interaction class-specific models applied on BoW descriptors.

2.5 Evaluation of predictive model

We present experimental results on two scenarios of our TAP dataset: constrained and unconstrained sets.

2.5.1 Constrained set

Samples in constrained set are time-slices of 5 interactions (handshake, hug, kick, punch, and push) collected from UT-Interaction dataset. To extract Predict-STIPs, we use a matching function with two adjustable parameters: matching window size and maximum search radius. We set the matching window size to 11 empirically and maximum search radius to 10 according to the resolution of images in the dataset. The number of P-STIPs is different from one action to another action and varies between 15-30.

We evaluate the time-slice recognition performance by using the standard "leave-one-out" method, one video is out each time, and fit the recognition problem in the context of multi-class classification. The average precision for all interactions (compared to human annotation) is given in the second column of Table 2.1. In order to visualize the performance of our method, we draw its average precision on a per time-slice basis and compare it to the average rate of human annotators (see Figure 2.8). Because the number of time-slices might be different in a few cases in our dataset, we compute the averages using video examples with the same number of time-slices. From the figure, we can see in some time-slices our approach outputs higher values than human annotators, e.g. time-slices 11 and 12 for the Hug action. In those cases, our method is thus better in recognizing the action from some limited time-slices.

Interestingly, since Predict-STIP is sufficient for holistic and early activity recognition, we can also compare it with the state-of-the-art on UT-Interaction dataset for those two different recognition context problems. Table 2.2 shows that our predictive representation outperforms all the state-of-the-art methods.

2.5.2 Unconstrained set

The unconstrained set is more challenging than constrained set in terms of background clutter, the number of people in the scene, the number of interactions, camera motion, and changes of viewpoints. This set includes time-slices of 6 realistic human interactions (handshake, high five, hug, kick, kiss, and punch) collected from HMDB, TV Interaction, and Hollywood TV show datasets.

The experimental setting of this set is similar to the constrained set. The performance of Predict-STIP on this set is also reported in Table 2.1. The results are obtained based on the number of correctly labeled time-slices compared to the human annotation. From the table, we can see that the results on the constrained set are better than the unconstrained set because the unconstrained set is more challenging. We can also see that the best results are obtained in handshake interaction for both datasets. High five interaction, meanwhile, has the minimum accuracy rate among 7 listed interactions.

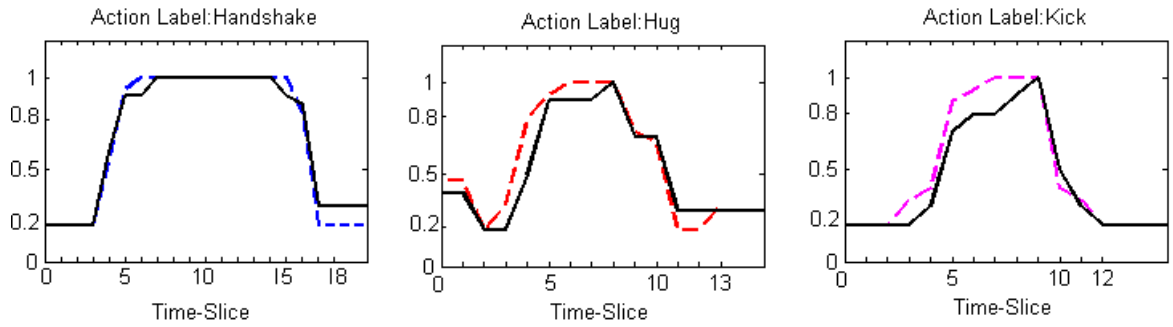


Figure 2.8: Comparison results of our method with the human annotation. Dashed and black lines show average rate by annotators and average precision of our method at test time, respectively.

Table 2.1: The average precision of Predict-STIP on constrained (UT-interaction dataset) and unconstrained sets (selected videos from HMDB, TV Interaction, and Hollywood TV show datasets).

	constrained set	unconstrained set
handshake	82%	76.3%
high five	–	61.4%
hug	81%	71%
kick	78%	73.7%
kiss	–	74%
punch	80%	76.2%
push	75%	–

Table 2.2: Performance comparison on the UT-Interaction Dataset. Early recognition and holistic recognition results are reported on the second and third columns, respectively.

Method	Accuracy with half observation	Accuracy with full observation
Our Model	**83%**	**95%**
Raptis and Sigal (1)	73.3%	93.3%
Yu et al. (10)	80%	91.7%
Ryoo (Best) (2)	70%	85%
Ryoo and Aggarwal (Best) (27)	31.7%	85%

2.6 Conclusions

In this paper, we introduced a predictive representation for a new problem of time-slice activity recognition. Time-slice activity recognition aims at exploring and recognizing an activity using a portion of the whole activity. We represented each video based on spatio-temporal descriptors of predictive features extracted from discriminative video segments. We also showed the effectiveness of our approach in a new dataset and compared it to the state-of-the-art. This dataset is available as a public dataset to encourage researchers to explore human activities at a smaller temporal granularity.

Chapter 3

Integration of Uncertainty in the Analysis of Dyadic Human Activities

Abstract

Action analysis from video data has been attracting more and more attention in computer vision over the past decade. The main focus has been classifying videos into one of k action classes from fully observed videos. However, intelligent systems are always enforced to make decisions under uncertainty and based on incomplete information. This need motivates us to introduce the problem of analysing the uncertainty associated with dyadic human activities and move to a new level of generality in the action analysis problem. Analysing the uncertainty, here, refers to categorizing the likelihood of activities in trimmed video clips called time-slices which are extracted from the full video. To this intent, we exploit the state-of-the-art methods to extract interest points in time-slices and represent them. We also present an accumulative uncertainty to depict the uncertainty associated with partially observed videos for the task of early activity recognition. The experiments demonstrate the effectiveness of our framework in analysing dyadic activities under uncertainty and in evaluating the performance of early activity recognition methods.

3.1 Introduction

Uncertainty is an essential and inevitable feature of daily life. A parent does not know exactly why a baby is crying. A speaker does not know exactly what the audience understands. An investor does not know exactly how the stock market changes.

Intelligent systems have to exploit whatever information they have when it comes to making decisions. There are many frameworks for solving problems involving a sequence of decisions with uncertain outcomes: Markov decision process, planning under uncertainty, model-free and model-based reinforcement learning, and batch reinforcement learning, to name a few.

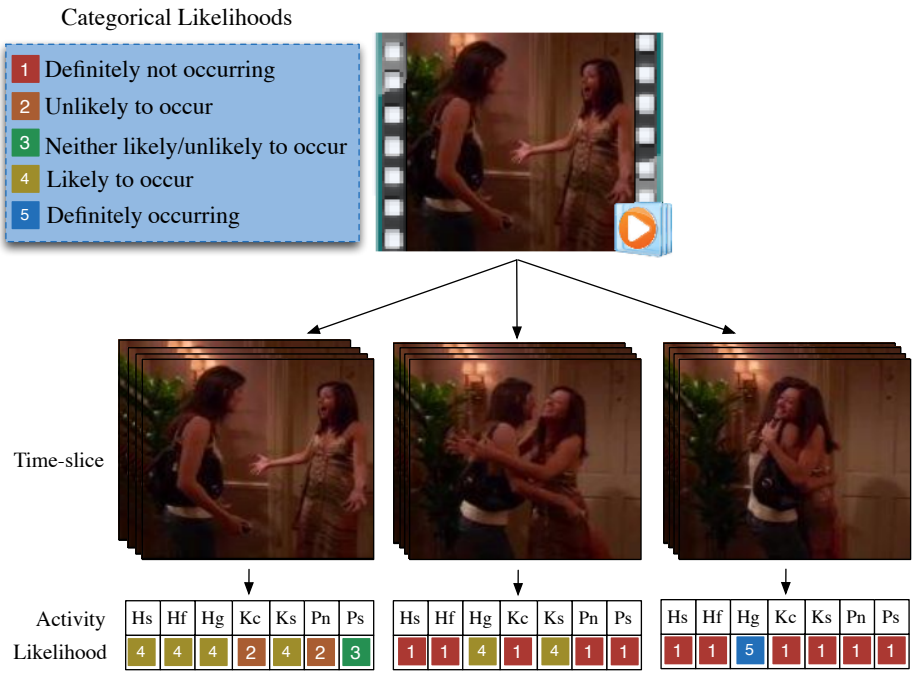


Figure 3.1: Integration of uncertainty in the analysis of dyadic human activities in videos. Full videos are divided into time-slices (we use time-slices of 10 consecutive frames) and human interactions (we use 7 interactions: handshake (Hs), high-five (Hf), hug (Hg), kick (Kc), kiss (Ks), punch (Pn), and push (Ps)) are analysed in these time-slices. The likelihood of each activity is reported from a list of categorical likelihoods in each time-slice.

In order to deal with uncertainty intelligently, we need to represent it and reason about it. There is an active domain called uncertainty reasoning in computer science (see (41) for more details). However, our interest here is not uncertainty reasoning per se but the representation of uncertainty associated with activity recognition.

Uncertainty abounds in every phase of computer vision. Recognizing objects that do not come from a database of geometrically precise models and scene understanding with either missing information or ambiguity in interpretation are examples of uncertainties in computer vision. Dominant uncertainties arise from lack of data (e.g. occlusions) versus lack of knowledge (weak models). The uncertainties that arise from a lack of data are the focus of this work.

With a similar purpose to this paper, i.e. analyzing the uncertainty in human activities, Schindler and Van Gool (42) investigated how many frames need to be accumulated over time to enable action classification. They used the entire video sequences as well as very short sub-sequences of videos, which are called snippets, and extracted local edges and optical flow to recognize single person actions. Recently, a few works have been proposed to recognize activities from partially observed videos ((1; 12)). Although these methods consider uncertainty,

they always analyse the beginning of the action sequence to classify videos into one of k action classes. In this context, the question arises as to whether it is feasible to obtain a measure of uncertainty associated with the labeled activity (or other possible activities) throughout the video. This paper tries to answer this particular question by the explicit modeling of the uncertainty occurring at a smaller temporal granularity referred to as time-slices which are a set of consecutive frames anywhere in the video.

Figure 3.1 illustrates the basic proposal behind our work. Different from conventional approaches in activity recognition, we do not aim to classify the video into one of k action classes. Instead, we want to analyse each time-slice extracted from the full video and classify the occurrence of activities in it into one of k categorical likelihoods. Categorical likelihoods are the following: definitely not occurring, unlikely to occur, neither likely nor unlikely to occur, likely to occur, and definitely occurring. Activities of interest in this paper are 7 human interactions: handshake, high-five, hug, kick, kiss, punch, and push.

There is a variety of work with promising results in the field of activity recognition. The majority of the works recognize an activity of interest occurring in a full video. However, one problem still remains: to what degree of certainty the activity of interest is occurring and other activities are not occurring during the video? To sidestep this problem, we propose a learning framework to integrate uncertainty in the analysis of human activities in videos. In particular, we explore the potential of different low-level feature detection and encoding techniques in capturing the uncertainty in activities and perform a quantitative comparison of these techniques. The details concerning this problem will be described later in Sections 3.3 and 3.4.

We quantify the performance of our proposed method using mean average precision (mAp) to evaluate classified categorical likelihoods as well as a metric to evaluate the uncertainty associated with each time-slice. We also report the uncertainty associated with time-slices in each activity. Besides reporting results on categorical likelihoods and uncertainty in time-slices, we show the effect of combining evidence from different time-slices and its effect on evaluating the performance of an early activity recognition method. Section 3.5 describes our experimental results in more detail.

In summary, our first contribution is a proposal to integrate uncertainty in the analysis of dyadic human activities in videos (Figure 3.1). It consists in categorizing the likelihood of activities, from a closed world, in each time-slice of a video. Methods developed under this proposal have important implications for practical scenarios, where decisions have to be made even if the occurring activity cannot be predicted precisely. Other application domains are content-based video search, indexing, retrieval, and summarization where the goal is to explore the content of the video and return time-slices where an activity of interest is (not) occurring with a specific degree of uncertainty.

Our paper offers three more contributions: i) a learning framework addressing this proposal (Figure 3.2), ii) a quantitative comparison of different instantiations of the framework using state-of-the-art techniques (Figure 3.3), and iii) a novel technique for evaluating the performance of early activity recognition methods based on the uncertainty associated with partially observed videos (Sections 3.3.4 and 3.5.3). The outline of the paper is as follows. In Section 3.2 we provide an overview of the most relevant work in activity analysis. In Section 3.3 we explain the framework of our proposed approach. In Section 3.4 we describe our time-slice representation. In Section 3.5 we report our experimental results, followed by conclusions in Section 3.6.

3.2 Related Work

Discriminative approaches have been widely used to recognize activities over the past decades. Most of these approaches make use of action descriptors with a bag-of-words (BoW) framework. Classical action descriptors include SIFT, SURF, MBH, and HoG which are computed in local space-time features. Among the local space-time features, space-time interest points (STIPs) (43) have shown a promising performance. Several methods have been reported (44; 18) to improve STIP detection for human activity recognition. For more details, we refer readers to a recent comprehensive survey of human action recognition with STIP detector by Das Dawn and Shaikh (19). Selection of encoding techniques is important to recognition performance in the BoW framework. Wang et al. (45) evaluated most of the encoding methods (vector quantization, soft-assignment encoding, sparse encoding, locality-constrained linear encoding, and Fisher kernel encoding) for action recognition and reported the Fisher coding method as the most effective method among them.

The time-slice representation has been used before in (44) for activity prediction. Discriminative segments of videos are extracted and predictive space-time interest points of these segments are detected. The HOG3D descriptor and BoW model are applied to the interest points preparing them for SVM classifiers. However, our goal is different from (44) which follows the conventional methods in terms of classifying videos into k action classes.

One way of seeing time-slice analysis of activity videos is as activity prediction. Li and Fu (11) proposed a model to depict predictability of long-duration activities. They represented activities as sequences of discrete action units and characterized the predictability using information entropy changes along action units. Yang et al. (46) exploited the shape and motion information to generate a temporal bag-of-words algorithm to predict daily activities. Huang and Kitani (47) showed how the actions of one person can be used to predict the actions of another person in dual-agent interaction. They modelled interaction as a Markov decision process where the actions of the initiating agent induce a cost topology over the reactive agent poses. Here, we use a simpler approach where the uncertainty for each time-slice is represented

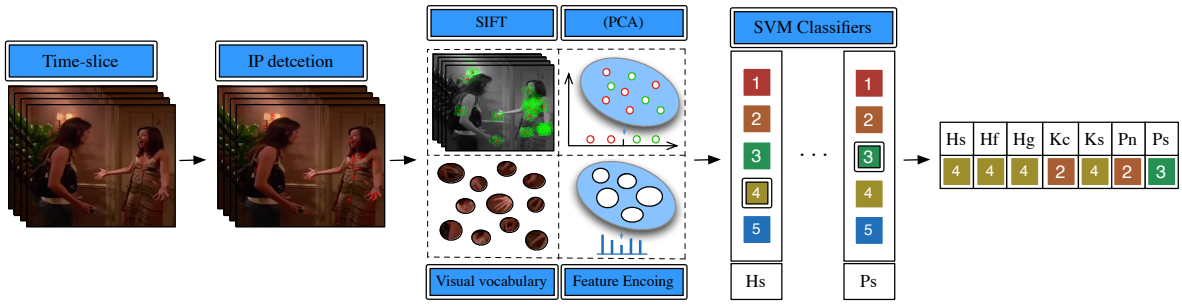


Figure 3.2: Method pipeline. Given a query time-slice, interest points (IPs) are detected (by SIFT/S-STIP/Predict-STIP). Detected IPs are described with SIFT features. Visual vocabularies are extracted from a learnt visual dictionary (by GMM/KNN). Features are encoded (by FV/BoW) and normalized to feed into pre-trained SVM classifiers (linear/non-linear with RBF kernel) to determine categorical likelihoods. In the case of FV encoding, features are sampled and their dimensionality is reduced by PCA. More details about the mentioned techniques are given in Section 3.4.

as categorical likelihoods which allows both activity prediction from any time-slice (versus the first time-slices of the video) and the uncertainty estimation.

3.3 Our Framework

Figure 3.2 shows the test pipeline of our framework. A training phase must precede the execution of this pipeline. The output of the pipeline is a categorical likelihood for each possible activity.

3.3.1 Categorical Likelihoods

In the training phase, we learn a mapping $f(T) \rightarrow \mathcal{L}^n$ where $T = \{t_1, ..., t_l\}$ denotes a time-slice with l frames. $\mathcal{L} = \{1, ..., k\}$ is a set of categorical likelihoods and $n = 7$ is the number of possible dyadic activities, i.e. handshake, high-five, hug, kick, kiss, punch, and push. We set $k = 5$ where the activity is definitely not occurring if $\mathcal{L} = 1$, unlikely to occur if $\mathcal{L} = 2$, neither likely nor unlikely to occur if $\mathcal{L} = 3$, likely to occur if $\mathcal{L} = 4$, and definitely occurring if $\mathcal{L} = 5$.

3.3.2 Training

Given a collection of time-slices of n activities in the training phase, we extract interest points (IPs) from them. We exploit local space-scale and general space-time features to extract IPs allowing us to capture distinctive information in space and time. We then describe extracted IPs with SIFT descriptors.

In order to build a compact representation of extracted features, we build a sequence of visual vocabularies on top of SIFT descriptors. We then use encoding algorithms on visual vocabularies and normalize the results to represent them as encoded features. Finally, the encoded features are fed into SVM classifiers to learn SVM parameters and train the model.

3.3.3 Testing

At test time, a query time-slice T is given to determine categorical likelihoods of n possible activities in it. From the given T, IPs are detected and described with SIFT features. Visual vocabularies of time-slice T are extracted from learnt visual vocabularies. Features are encoded and normalized to construct time-slice representations. These representations enter into SVM classifiers, which are pre-trained in the learning phase, to determine categorical likelihoods in T.

3.3.4 Uncertainty

In this paper, we analyze the uncertainty associated with a time-slice T with an uncertainty metric λ_T extracted from the uncertainty theory in mathematics (48). The uncertainty theory is used to indicate the belief degree that an uncertain event may occur or may have occurred. We cannot model the degree of belief with the probability theory since the probability theory models frequencies. Frequency is a factual property of an indeterminate event and does not vary with our state of knowledge and preference. In other words, the frequency exists in the long term and is relatively invariant, whether we can observe it or not, e.g. "the number 6 appearing when rolling a dice" event. On the other hand, in many cases, no samples are available to estimate a probability distribution of occurring events, e.g. "it will be sunny next week" event. In these cases, the uncertainty theory evaluates the belief degree that each event will occur.

Here, we are interested in investigating how uncertain/predictive a time-slice may be. We convert the output of the pipeline (the output in Figure 3.2) to one uncertainty measure by assigning a number to the uncertainty degree of each time-slice.

We first map the categorical likelihoods ($\mathcal{L} = \{1, 2, 3, 4, 5\}$) to a scale from 0 to 1 based on associated uncertainties and call them uncertainty values, i.e. $u = \{0, 0.5, 1, 0.5, 0\}$. We then define the uncertainty metric λ_T as the average of all u in a time-slice T:

$$\lambda_T = \frac{1}{n} \times \sum_{i=1}^{n} u_i \tag{3.1}$$

where n is the number of activities. A time-slice has no uncertainty if its uncertainty metric is 0 ($\lambda_T = 0$) because we then believe that the time-slice is a combination of $u = 0$, i.e. $\mathcal{L} = 1$ and $\mathcal{L} = 5$ where an activity is definitely not occurring or definitely occurring, respectively.

A time-slice is the most uncertain if its uncertainty metric is 1 ($\lambda_T = 1$) because both the occurring activity and not occurring activity may be regarded as equally likely. The closer λ_T is to 1, the more uncertain the time-slice is. Considering the output categorical likelihoods in Figure 3.2, λ_T is equal to 0.57 which shows the degree of uncertainty in this example.

We also report λ_A for each activity defined as the equation below to show the average uncertainty associated with the time-slices in each kind of activity.

$$\lambda_A = \frac{1}{t} \times \sum_{\tau=1}^{t} u_{A,\tau} \tag{3.2}$$

A is the activity of interest, t is the number of time-slices containing the activity of interest, and $u_{A,\tau}$ is the uncertainty value of activity A in time-slice τ.

Furthermore, we show the effect of combining evidence from different time-slices and its effect on evaluating the performance of an early activity recognition method. We define the accumulative uncertainty U_x as the average distance between λ_T and 1, over x time-slices as:

$$U_x = \frac{1}{x} \times \sum_{T=x_1}^{x_2} (1 - \lambda_T) \text{ where } x = x_2 - x_1 + 1 \tag{3.3}$$

We then compute the mean square error (MSE^*) based on U_x to evaluate the result of the early activity recognition method as follows:

$$MSE^* = \frac{1}{v} \times \sum_{i=1}^{v} U_x \times (\hat{y}_i - y_i)^2 \tag{3.4}$$

where v, $\hat{y}_i$, and y_i are the number of partially observed videos, the result activity label, and the ground-truth activity label respectively ($\hat{y}_i$ and $y_i \in \{0, 1\}$). U_x is the accumulative uncertainty of each partially observed video test. We assign a weight to a misclassified test video on the basis of its accumulative uncertainty to penalise the misclassification regards to the level of the uncertainty in the video. Regarding Equation 3.3 when $\lambda_T = 0$, U_x is equal to 1. In this case, the test video is certain; therefore with setting the weight to 1, we penalise the method more if it misclassified a non-uncertain video. In other cases if $U_x = 0$, there is the most uncertainty in all x time-slices and it means that all $\lambda_T = 1$ in the test video. This shows that when there is maximum uncertainty in all time-slices of a video test, it is hard to recognize the occurred activity and thus does not warrant a penalty.

3.4 Time-Slice Representation

In this section, we first present our feature extraction and encoding pipelines along with implementation details. We also present the dataset and discuss evaluation criteria.

3.4.1 Feature Extraction

SIFT features have proven to be extremely successful for a variety of datasets. These features are invariant to any scaling, rotation or translation of the image. The SIFT approach can be used for both interest point localization and description. SIFT densely detects interest point locations in the scale-space. However, we intend to describe time-slices with appearance features as well as motion features. Therefore, we use space time interest points, Selective STIP (S-STIP) (18) and Predictive STIP (Predict-STIP) (44), and compare the results of these three interest point detectors (See Figure 3.3).

S-STIP detection applies background suppression as well as local and temporal constraints to discard unwanted points. Predict-STIP measures how long STIPs are observable in a video and selects STIPs which are active during the whole video as Predict-STIPs. Default parameters are used for both methods.

We compute SIFT descriptors of detected interest points in each time-slice using a set of 16 histograms, aligned in a 4x4 grid, each with 8 orientation bins. These histograms result in a feature vector containing 128 elements which are used to characterize time-slices.

3.4.2 Feature Encoding

Once the features are extracted, we use bag-of-words (BoW) and Fisher vector (FV) algorithms [1] to encode them. The FV is an extension of the BoW technique, which uses both the number of assigned visual vocabulary and their mean and variance using a Gaussian Mixture Model (GMM). Since the FV encodes more information, the amount of visual words in the FV is significantly lower than in the BoW and, therefore, the FV is faster to compute.

For the BoW features, we train a dictionary with K-means using 60,000 randomly sampled features, where the size of visual vocabulary (K) is set to 1000. An SVM classifier with a RBF kernel using the standard leave-one-out method is applied for classification.

Differently from the BoW encoding, we sample the features to reduce their dimensionality by PCA and compute a GMM. For a GMM of size d, we need about 1000*d training features. We set the GMM size to 64 and randomly sample a subset of 64,000 descriptors from the training set to compute PCA and estimate the GMM. As proposed in (49), we performed a preliminary experiment to represent FVs. We considered two strategies. First, we computed one FV per time-slice and then applied the normalization. Second, we computed and normalized one FV

[1] We use the VLFeat publicly available library for extracting SIFT and FVs.

per frame, and then averaged and renormalized the per-frame FVs. The results showed that the first option was more effective and it was used in all of our experiments. A linear SVM using the standard leave-one-out method is applied for classification with C=128 which has shown good results on validation data samples. Figure 3.3 indicates that the FV encoding method results in a better performance.

3.4.3 Dataset and Evaluation Criteria

The performance of our proposed method was quantified using the TAP dataset (44). This is the only dataset suitable for this task since it contains the annotations corresponding to categorical likelihoods of activities. The TAP dataset contains pre-temporally trimmed video clips (time-slices) of 7 dyadic human activities. As in (44), we use time-slices with 10 consecutive frames (without overlap) since it best shows the goal of studying the inherent uncertainty in activities. In (44), it was shown that 10-frame time-slices are more effective compared to 5- and 15-frame time-slices and that 5-option categorical likelihood best describes the occurrence of activities compared to 3- and 7-option categorical likelihoods. There are two sets in the TAP dataset: constrained and unconstrained sets. The constrained set includes 1129 time-slices extracted from staged scenarios of the UT-Interaction dataset (27). There are 990 time-slices in the unconstrained set which are extracted from a subsample of real-world scenarios of HMDB (28), TV Interaction (29), and Hollywood (30) datasets. This sub-sample is selected based on the camera angle ranging from -45 to +45 degrees. Each time-slice was annotated by 3 different annotators on how likely it is that each of the 7 dyadic activities are occurring. The annotators were requested to pick one of 5 categorical likelihoods from 1 (definitely not occurring) to 5 (definitely occurring). We set the medians of the 3 annotated values as the ground-truth class labels. Performance on the dataset is measured in terms of mean average precision (mAp) across the categories. This calculates the number of true labeled categorical likelihoods compared to ground-truth labels.

3.5 Experimental Results

In our experimental evaluation, the performance of our proposed framework was quantified as described above in turn for both constrained and unconstrained sets.

3.5.1 Comparison of different techniques

In our first experiment we compare SIFT, S-STIP, and Predict-STIP as interest point detectors and BoW and FV as encoding techniques. In Figure 3.3, we present the performance of different settings S1 (SIFT+BoW), S2 (S-STIP+ BoW), S3 (Predict-STIP+ BoW), S4 (SIFT+FV), S5 (S-STIP+ FV), S6 (Predict-STIP+ FV) in terms of mAp for constrained and unconstrained sets, respectively. The results are obtained based on the number of correctly classified categorical likelihoods in time-slices compared to the human annotation.

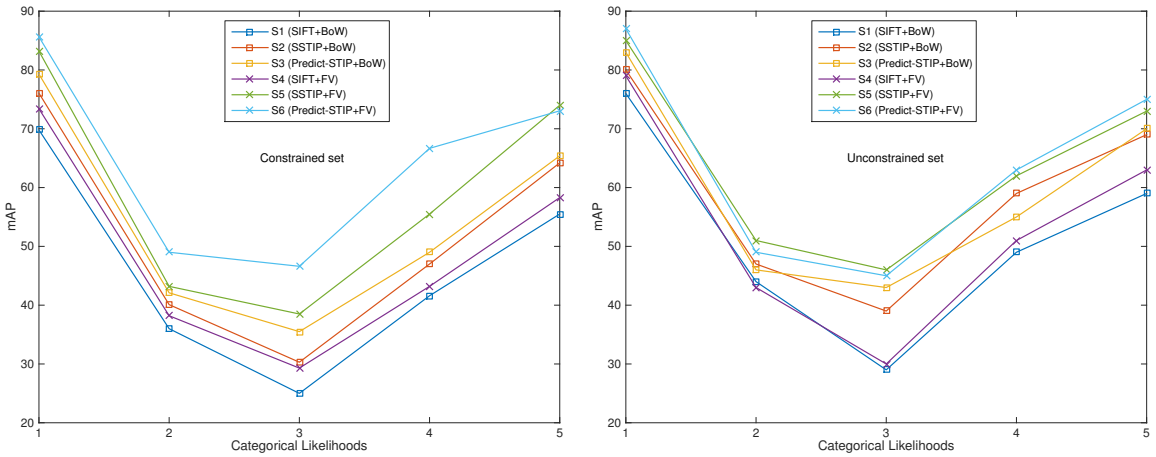


Figure 3.3: Performance comparison of different settings

As shown in Figure 3.3, the best results for classified categorical likelihoods belongs to $\mathcal{L} = 1$ and $\mathcal{L} = 5$. This is due to the fact that time-slices in these two categories contain more relevant motion or visual information. For instance, the "two people are shaking each other's hands" time-slice is easy to categorize as "push is definitely not occurring" ($\mathcal{L} = 1$) and "handshake is definitely occurring" ($\mathcal{L} = 5$). Also, we notice that $\mathcal{L} = 3$ which is the most uncertain category gives the worst results, probably because of the lack of information in time-slices of this category in either space or time.

Generally across all settings, FV with Predict-STIP leads to significantly better performance than others. Using the same interest point detector techniques, settings with the FV encoding method have better results than BoW. This confirms that encoding both first (number of assigned vocabularies) and second order (mean and variance) statistics is more effective. Settings with SIFT as the interest point detector result in lower mAps since SIFT does not consider the motion information carried in the video.

Using the best setting from these experiments, i.e. S6, the confusion matrix of our framework is illustrated for both sets in Figure 3.4. In both sets, $\mathcal{L} = 2$ and $\mathcal{L} = 3$ are the most misclassified categories. $\mathcal{L} = 2$ is confused with $\mathcal{L} = 1$ in most cases while $\mathcal{L} = 3$ is misclassified more as $\mathcal{L} = 4$ in the constrained set and $\mathcal{L} = 1$ in the unconstrained set.

We show some examples of the successes and failures of our framework using the S6 setting in Figure 3.5. Three matched categorical likelihoods between the result of our framework and the annotation is the threshold to determine success or failure. Our framework works well on time-slices with less uncertainty. In the example with the green bounding box including kiss and hug interactions, the annotation shows that kiss is definitely occurring while hug is definitely not occurring. However, our method shows that hug is definitely occurring as well which indicates the outperformance of our algorithm. Our framework tends to fail when the

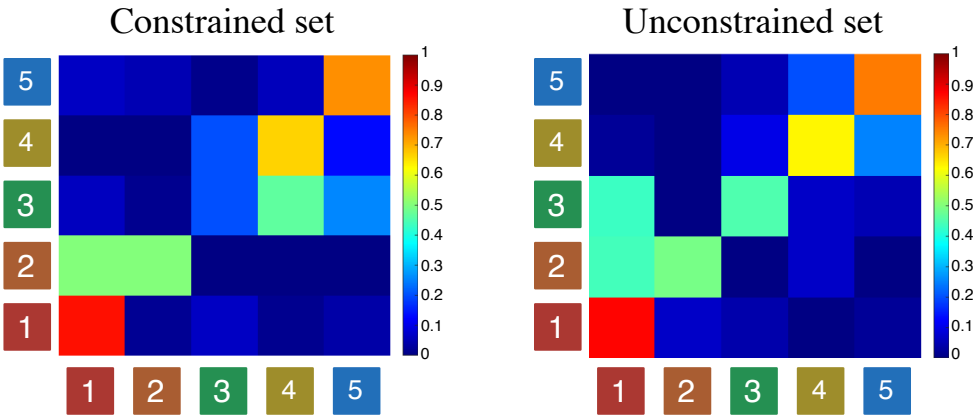


Figure 3.4: Confusion matrix of our second framework using S6.

time-slice does not contain enough information and even human annotators may fail. For instance, the time-slice with the red bounding box is labeled as *push* in the constrained set while human annotators stated that "*hug* is likely to happen" in this time-slice.

3.5.2 Time-slice uncertainty analysis

In our second experiment we compute the uncertainty associated with a time-slice. We assign λ_T to each time-slice on both the constrained and unconstrained sets using equation 3.1. λ_T shows how uncertain a time-slice is in classifying activities on average.

Figure 3.6 illustrates the histograms of λ_T for both sets. The closer λ_T is to 1 the more uncertain the time-slice is. We cannot compare the uncertainties of the two sets directly since the two sets do not contain the same number of time-slices. However, Figure 3.6 shows that the constrained set is more uncertain than the unconstrained set generally, although the constrained set looks visually simpler. This is due to the fact that the constrained set contains staged scenarios and actors perform activities at a slow pace. Therefore, 10-frame time-slices do not include enough information all the time while activities in the unconstrained set, which contains real-world scenarios from TV shows, are executed at a faster pace and result in more informative time-slices.

In another experiment, we report λ_A using Equation 3.2 for each activity of interest indicating the average uncertainty associated with the time-slices of the activity (See Table 3.1). For each activity of interest (A), the uncertainty which corresponds to that activity (u_A) is considered for all t time-slices containing the activity of interest. The average of all considered u_A is referred to as λ_A. Table 3.1 also shows that more uncertain and challenging activities (the closest λ_A to 1) are *punch* and *push* in the constrained set and *kiss* in the unconstrained set.

Figure 3.5: Successes (two top rows) and failures (two bottom rows) using S6 on both constrained and unconstrained sets. Frames are overlaid to show a complete time-slice as a single image.

We also notice that *high-five* and *kiss* in the constrained set and *push* in the unconstrained set have λ_A almost equal to 0. This is due to the fact that the constrained set does not contain *high-five* and *kiss* activities, and the unconstrained set does not include any *push* activity. Therefore, there is no uncertainty in the occurrence of these activities and the average uncertainties associated with them in corresponding sets are almost 0 (the activity is definitely not occurring).

3.5.3 Uncertainty fusion analysis

In our last experiment we combine uncertainty from different time-slices and show its effect on evaluating the performance of our method in the task of early activity recognition.

Table 3.1: λ_A for different activities.

	Hs	Hf	Hg	Kc	Ks	Pn	Ps
Constrained set	0.4	0.02	0.45	0.51	0.02	0.62	0.7
Unconstrained set	0.43	0.29	0.12	0.38	0.67	0.37	0

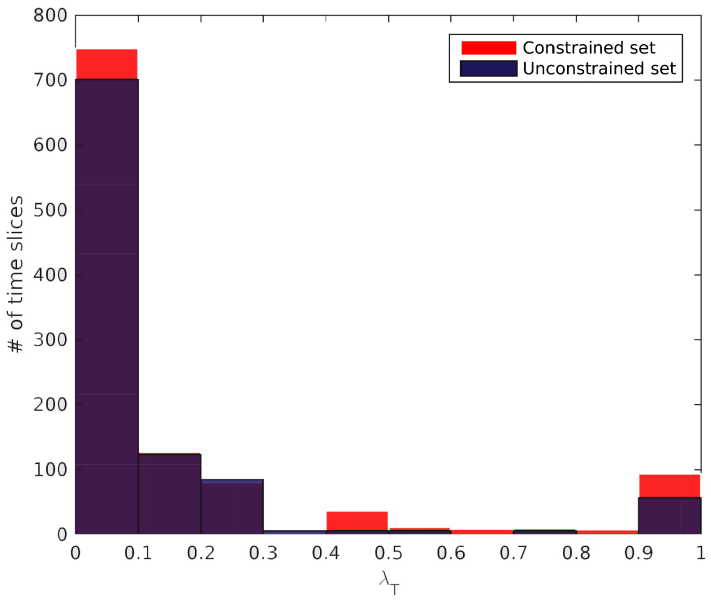


Figure 3.6: Time-slice uncertainty analysis.

In early activity recognition (e.g. recognizing activities by observing the first half of videos), all videos have the same weight. However, people perform the same activity in different ways and at different paces. Therefore, partially observed videos may be more or less informative, which affects the recognition results. Considering the uncertainty associated with partially observed videos helps to identify what confidence should be associated with the results.

We aggregate uncertainties from time-slices of the first half of a video and calculate the accumulative uncertainty U_x by Equation 3.3 where $x_1 = 1$ and $x_2 =$ the middle frame in each video. Since the constrained set contains more uncertainty and S6 results in a better performance, we use them to run this experiment. One video is put aside each time and the S6 training technique with linear SVM classifiers is applied on the rest to recognize activities from the half videos. To calculate the misclassification rate, we weigh misclassified query videos using Equation 3.4. In other words, the smaller error is assigned to misclassified videos that have more uncertainty compared to misclassified videos containing less uncertainty. The uncertainty-weighted misclassification rate obtained is improved with respect to the uniformly-weighted rate (12.3% versus 15%), which shows that classification errors are more frequent when the uncertainty is higher.

Table 3.2: Uncertainty measurement comparison for early activity recognition in the constrained set.

	Hs	Hf	Hg	Kc	Ks	Pn	Ps
Our method (U_x)	0.19	-	0.25	0.32	-	0.30	0.35
Baseline method	0.17	-	0.23	0.31	-	0.25	0.32

We compare our proposed uncertainty metric for early activity recognition (U_x) with a baseline uncertainty measurement grounded in the Platt scaling method (50) on the probabilistic interpretation of SVM scores. As the score of the classifier itself would also indicate the degree of confidence, we convert that score to a probability measure to serve as another kind of uncertainty to compare with U_x. The result of this comparison is given in Table 3.2 for each action where x is equal to the half length of video sequences. In other word, classification scores are transformed to class probabilities using Platt scaling. The difference between the average of these probabilities from 1 is referred to as a baseline uncertainty measurement for each video observation. For each activity, we compute the mean of all U_x as well as the mean of all baseline uncertainty measurements of corresponding videos and report them in Table 3.2. The baseline method computes the general uncertainty by fitting a logistic regression model to the classifier's scores while our method shows more detail by analysing uncertainty in every time-slices throughout the video. Therefore, U_x depicts more uncertainties which results in larger values in Table 3.2.

The accumulative uncertainty can also be of help in selecting training data. A subset of videos which has less uncertainty improves training models. In case of activity recognition, using time-slices to train a model in which U_x is greater so that an activity of interest is definitely occurring/likely to occur may lead to a better performance of an algorithm.

3.6 Conclusion

In this paper, we proposed a novel approach for integrating uncertainty in the analysis of dyadic human activities. The major contributions include classifying activity time-slices into one of 5 categorical likelihoods which show how likely each activity is occurring in the time-slice, a learning framework for this classification, a quantitative comparison of different instantiations of the framework using combinations of the state-of-the-art representations (Predict-STIP, S-STIP, FV, SIFT, and BoW), and a novel technique for evaluating the performance of early activity recognition methods. In our experimental results, we observed that across all of these representations, the combination of Predict-STIP and FV outperforms other settings. We also empirically showed that considering uncertainty is particularly beneficial for evaluating the performance of an algorithm in early activity recognition. The more uncertain a partially observed video is, the less likely it is to recognize the true activity. Uncertainty analysis is

also helpful to improve training data and to select segments of videos in which the activity of interest is occurring.

Chapter 4

Deep Uncertainty Interpretation in Dyadic Human Activity Prediction

Abstract

In this paper, we propose a deep learning framework to analyse the uncertainty associated with dyadic human activities at a small temporal granularity. Such time-slice analysis is able to infer human behaviours from short-term observations. This framework obtains a distribution over the likelihood of human activity categories that are to occur in each time-slice. Instead of classifying time-slices into k classes of activities, we report to what degree of certainty each activity is occurring throughout the time-slice from "definitely not occurring" to "definitely occurring". To this end, we extract CNN-based unary probabilities and pairwise relations between human body joints in each time-slice. The unary term gives cues on the local appearance of the part while the pairwise term captures the contextual relations between the parts. We extract these features from each frame in a time-slice and examine different temporal aggregation schemes to generate a descriptor for the whole time-slice. Evaluation is conducted on the TAP dataset which is well suited for time-slice activity analysis. Extensive experiments demonstrate the effectiveness of our approach for the task of uncertainty analysis in activity prediction.

4.1 Introduction

Uncertainty in videos comprise a large majority of the visual data in existence. Ambiguity in surveillance systems about the tracking, detection, and recognition of objects and humans are a few examples of uncertainty in real world. In computer vision and video analysis applications, complete knowledge of what is occurring in videos is impossible to acquire. Therefore, understanding the content of videos despite uncertainty is of paramount importance.

In activity recognition, the majority of previous work, either using hand-crafted or deep fea-

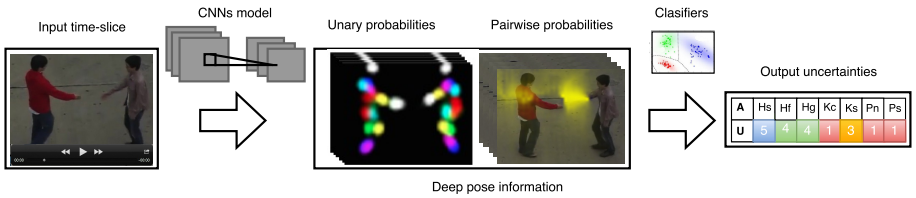


Figure 4.1: Method pipeline: In unary, color-coding corresponds to the probabilities of the presence of different body joints. Pairwise shows the probability of certain class c anywhere in the image given the fixed location of other body parts $c' \neq c$ for all cross-part pairs (for clarity, only pairwise probability of c =right wrist and c' =right elbow of one person is illustrated by yellow color).

tures, classify activities into one of k classes from a fully observed video. There are a few approaches that take into account some aspects of uncertainty such as recognizing activities from partially observed videos (42; 1; 12). However, these approaches still categorize activities into k classes. Based on intuition from social psychology, humans are good at anticipating and correctly predicting social interactions. But in a scenario with two people getting closer and lifting hands, what is the activity most likely to occur? Even humans may not be able to tell what activity is definitely occurring. They can only infer that handshake, hug, or high-five are examples of possible activities. Therefore, labelling this scenario to one of k activity classes is not suitable. In this context, the question arises as to whether it is feasible to measure the degree of certainty/uncertainty that some activities occur and other activities do not. This integration of uncertainty in activity recognition can be beneficial for practical scenarios such as content-based video search, indexing, retrieval, and summarization e.g. when the goal is to browse the content of videos and return sections of videos where an activity of interest is (not) occurring with some degree of uncertainty.

Following the work of Ziaeefard and Bergevin (55), we address time-slice analysis of dyadic human activities under uncertainty. Time-slice analysis aims to model the inherent uncertainty of activities occurring at a small temporal scale. The uncertainty is measured by classifying the occurrence of activities into one of k categorical likelihoods in time-slices (see Figure 4.1). Categorical likelihoods are the following: definitely not occurring (1), unlikely to occur (2), neither likely nor unlikely to occur (3), likely to occur (4), and definitely occurring (5). Our method is different from (55) in the sense that (55) used hand-crafted feature descriptors (e.g. SIFT and spatio-temporal interest points) in the context of Bag of Words and Fisher Vector models, while we propose a semantic deep learning framework which exploits explicit pose information.

We formulate our frame descriptor based on unary and pairwise pose information which are built on a deep Residual Network (ResNet) (56). Given a set of part candidates and a set of body part classes, e.g. head, shoulder, and knee, each part candidate has a unary score

for every body part class. Additionally, for every pair of distinct body part candidates and every two body part classes, the pairwise term is generated to predict the spatial relationships between joints. Therefore, our frame descriptor contains the probabilities of the presence of body joints and locations of other joints in the adjacency. Unary and pairwise terms help to extract consistent information on body joint configurations for different activities. In this way, similar features are returned from joints and joint relationships in intra-class likelihoods and activities. An illustration of unary and pairwise terms is shown in Figure 4.1. More details on unary and pairwise probabilities are given in Section 4.3.

As opposed to existing work which uses Two-Stream CNNs (one for spatial information and one for temporal features) (57; 58; 59; 60), we employ a One-Stream model. This has the benefit of capturing the configuration of body parts in each frame unary and pairwise which is computationally efficient. Tracking the change occurring in the configuration of body parts throughout the time-slice gives the required temporal information for uncertainty analysis.

In summary, our main contributions are three-fold: i) exploiting CNNs in time-slice analysis ii) a proposal to integrate unary and pairwise pose information to measure the uncertainty associated with activity recognition, and iii) a Single-Stream deep learning framework addressing this proposal (Figure 4.1).

4.2 Related work

Compared to single image analysis, there is relatively small number of work on applying CNNs to video classification and no work on time-slice analysis. Our interest here is not activity recognition per se, but we can relate our work to this area of research in the literature. In this section, we have a brief overview on CNNs [1] and then review the most relevant previous research to our framework.

4.2.1 A brief overview on CNNs

Conventional methods for solving computer vision problems and the degree of success they achieve have traditionally been due to carefully designed hand-crafted features. However, in the past few years, deep learning approaches have offered more successful alternatives which automatically extract features for a learning problem. Therefore, it is important to understand what kind of deep learning models are fit to a given problem.

Over the past few years, image classification (namely on the challenging ImageNet dataset (51)) using CNNs has influenced many works in computer vision. This work introduces a particular form of CNN models (also known as AlexNet in the honour of the first author, Alex Krizhevsky) which has since been widely used and updated by the computer vision community.

[1] This subsection is not part of the original paper. We add a brief overview on CNNs here to clarify the concept.

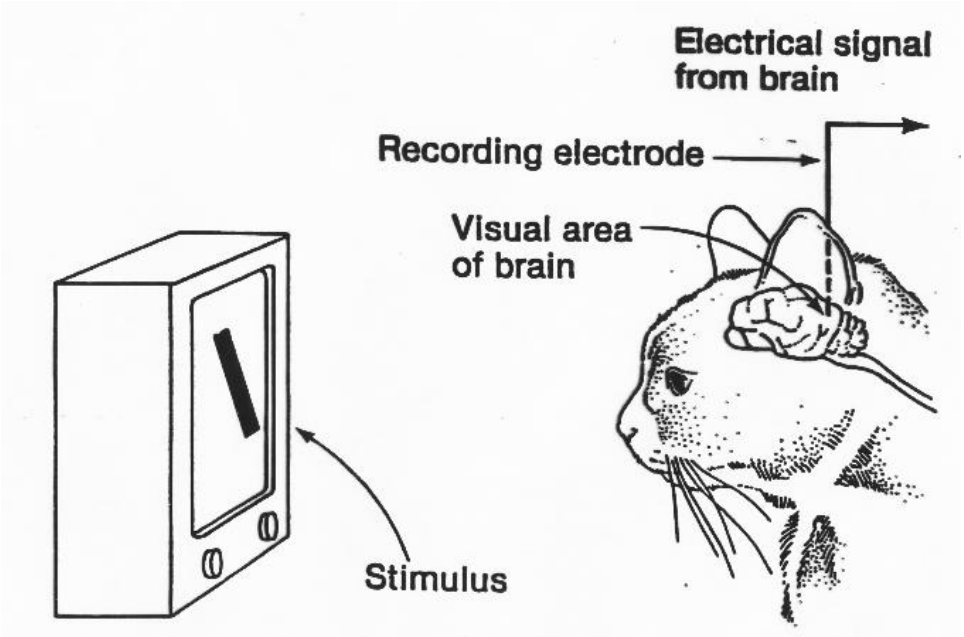


Figure 4.2: Hubel and Wiesel's experiment on understanding the biological vision system

CNNs are biologically-inspired by Hubel and Wiesel's early work on the visual cortex (52). Hubel and Wiesel performed experiments to find out how the vision works in our biological brain (Figure 4.2). They installed electrodes into the primary visual cortex area of the brain of a cat. They showed slides of different objects (stimulus) such as a mouse, flower, etc. to the cat and recorded the neural activities. They observed that fish and mouse slides did not excite the neurons, however the neurons fired every time they change slides. They realized the edges created by slides were changing and these moving edges drove the neurons. With this experiment, they recognized that neurons are organized in columns in our brain and every column of the neurons reacts to specific orientations of the stimulus. Therefore, the beginning phase of a visual processing is not analysing the holistic object but rather its simple structure, i.e. oriented edges.

Many neurally-inspired models have been proposed to solve problems in computer vision. Among them, CNNs have had a huge success in the literature. They are very similar to ordinary Neural Networks in terms of performing non-linearity computations and including a loss function on the last layer. However, they are less complex than fully connected networks since they have fewer parameters to train. CNNs are comprised of several layers stacked on top of each other to form a full architecture. There are three main types of layers: convolutional, pooling (sub-sampling layer), and fully-connected (FC) layers. An overview of CNN architecture is shown in Figure 4.3 and it will be discussed below in more detail.

The convolutional layer is the basic building block of a CNN model that does most of the

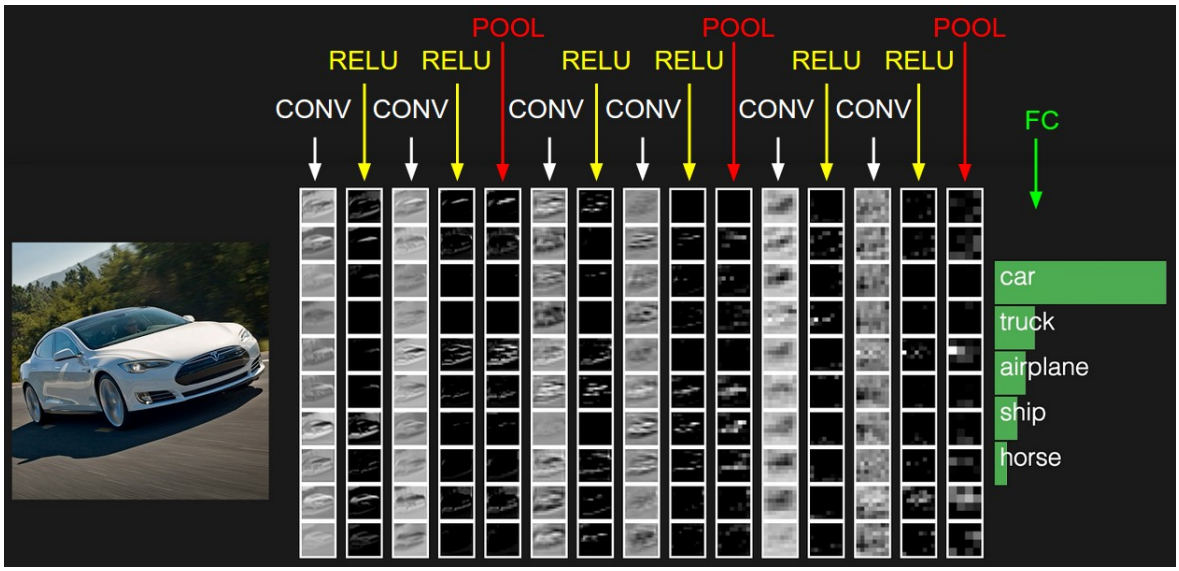


Figure 4.3: An example of CNN architecture (3).

computations. The input to a convolutional layer is a $m \times m \times c$ image where m is the height and width of the image and c is the number of channels. For instance, an RGB image has three channels, i.e. $c = 3$. The convolutional layer will have k filters of size $n \times n \times q$ where n is smaller than the dimension of the image and q can either be the same as the number of channels c or smaller and may vary for each filter. Filters are each convolved with the image to produce k feature maps of size $m - n + 1 \times m - n + 1$. Figure 4.4 shows feature map visualizations from Zeiler and Fergus model (4). It illustrates feature maps in different layers along with the corresponding image patches. In this Figure, we observe that Layer 1 and Layer 2 respond to low level information such as corners and other edge/color conjunctions. On the other hand, Layer 3 has more complex pattern and can capture similar textures. For instance, the sample in (Row 1, Column 1) shows mesh patterns and the sample in (Row 2, Column 4) extracts text features.

A convolutional layer is followed by an additive bias and non-linearity called activation function. This activation function is applied to each element in the feature map. A sigmoid function was often used historically but has lost favor compared to ReLU (Rectified Linear Unit), which has been shown to work better and faster to train deep neural architectures on large and complex datasets.

Each feature map is then subsampled in a pooling layer usually with mean or max pooling over $p \times p$ windows where p varies between 2 for low dimension images (e.g. MNIST dataset (53)) and it is typically not more than 5 for larger images. The purpose of pooling layer is to progressively reduce the spatial size of the feature map to decrease the number of parameters and computation in the network and therefore to control the overfitting issue. Many

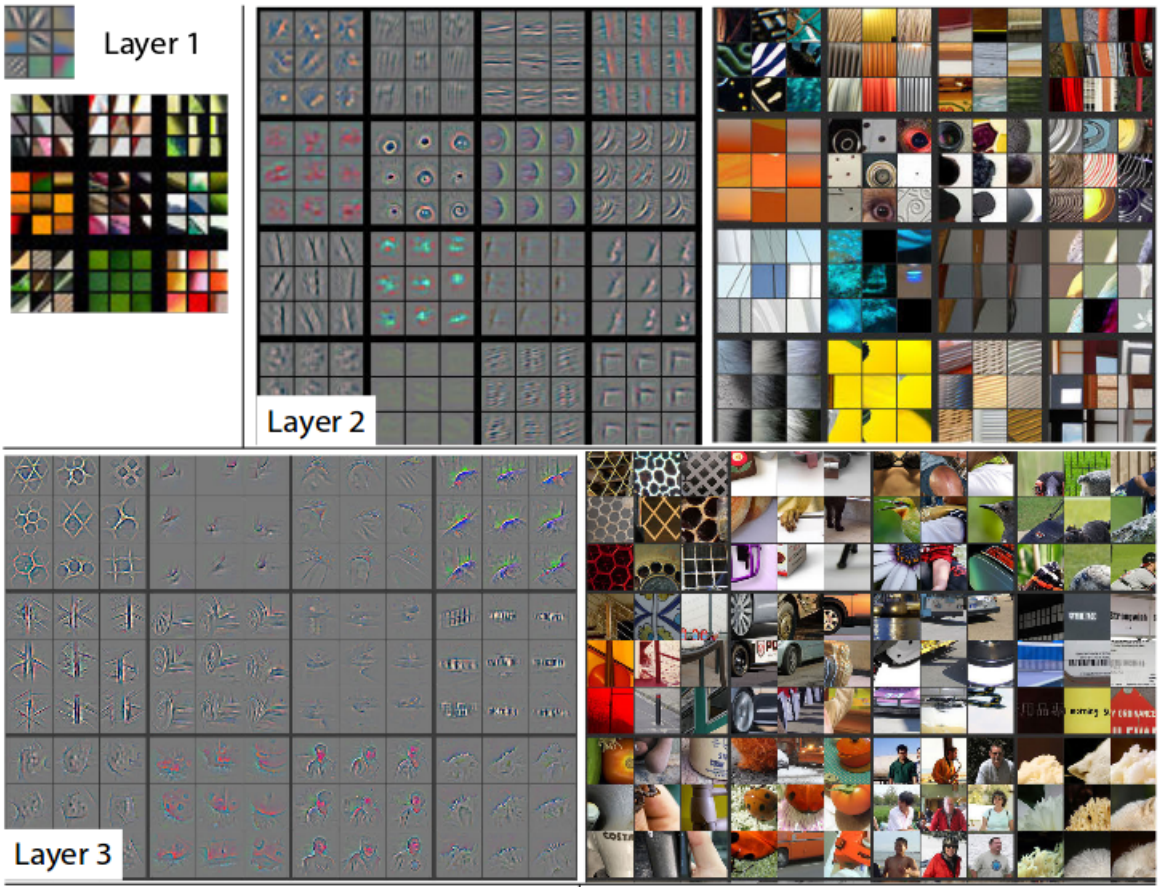


Figure 4.4: Visualization of feature maps in some sample layers of a CNN model (4).

researchers do not agree with the pooling operation in the network and intend to discard this layer. For example, Springenberg et al. (54) suggest to remove the pooling layer and repeat convolutional layers. They proposed to occasionally use a larger stride in the convolutional layer to decrease the size of the feature maps. It is possible that CNNs will consist of very few or even no pooling layers in future.

Typically, the last layer is fully-connected and its task is to compute the class scores which result in a vector of size $1 \times 1 \times l$. Each of the l numbers correspond to a class score and each neuron in this layer is connected to all the neurons in the previous layer. In this way, CNNs transform the input image layer by layer from the original pixel values to the final class scores. It is worthy to note that not all the layers contain parameters. In particular, convolutional and fully-connected layers include the weights and biases parameters of the neurons. The parameters in these layers are trained with the backpropagation algorithm. Hence, the class scores that the CNN model computes in fully-connected layer are consistent with the labels in the training set for each image.

4.2.2 Previous research

Pose information: Body poses are highly informative to discriminate between human activities. Deng et al. (61) trained CNN models to address the problem of group activity understanding in videos. They considered dependencies between individual actions, body poses, and group activities to predict class labels. Cheron et al. (62) proposed a Two-Stream pose-base CNN algorithm, where particular patches of appearance (RGB) and optical flow for human body parts are fed into CNN models for action recognition. Unlike their method, we do not use pose information to crop video frame and feed CNNs with the cropped patches. In our method, the deep descriptors are generated based on semantic relationships between body joints.

Deep activity recognition: Karpathy et al. (57) proposed multiresolution CNNs architectures to take advantage of both spatial and temporal information on the largescale Sports-1M dataset. Simonyan and Zisserman (58) trained two separate CNNs using the VGG model (63) to capture spatio-temporal features which are then combined by late fusion. Donahue et al. (64) combined CNNs and LSTM to introduce Long-term Recurrent Convolutional Networks for video activity recognition. Gkioxari and Malik (59) presented spatial- and motion-CNNs models operating on RGB and flow signals. Wang et al. (65) presented an action video representation combining the benefits of both hand-crafted and deep-learned features.

Uncertainty reasoning: Schindler and Van Gool (42) investigated the uncertainty in activity recognition with hand-crafted features to show how many frames need to be collected over time to enable action classification. In another category of work integrating uncertainty in activity recognition, there are a few methods proposed to recognize activities from partially observed videos ((1), (12)).

4.3 Deep pose information

We build our framework upon the recently proposed unary and pairwise terms for human body part detection conditioned directly on the image (66). This body part detection model named DeeperCut adapts the extremely deep Residual Network (ResNet). In the adapted ResNet, the final classification along with the average pooling layer are removed and the convolution layers are modified by adding holes (67). Deconvolutional layers for up-sampling (68) are also appended.

DeeperCut estimates poses of all people present in an image by minimizing a joint objective function. It starts from a set D of body part candidates which are selected by adapted Fast R-CNN (69), and a set C of body part classes. Each candidate $d \in D$ has a unary scoremap for every part class $c \in C$. The unary terms give the probability of part d at location l_i to

belong to class $c = j$.

$$U(l_i, c|I) = p(c = j|I(l_i)), \quad j \in C, \tag{4.1}$$

where I is the input image, and l_i is the center of the bounding box of the candidate part d. $C = 1, ..., K$ denotes the body part classes, e.g. head, shoulder, knee. The number of classes here is set to 14, i.e. $K = 14$.

The pairwise probabilities are also computed for every pair of distinct part candidates $d, d' \in D$ and every two part classes $c, c' \in C$. The pairwise probability is estimated with regressing from the location of a joint to the relative positions of all other joints. In each unary scoremap location $k = (x_k, y_k)$, the pairwise offset of c to c' is defined as a tuple:

$$P(cc')^k = (x_{c'} - x_k, y_{c'} - y_k), \tag{4.2}$$

where c is the the current joint and c' is each remaining joint.

We fine-tune the DeeperCut model on the TAP time-slice dataset (44) in order to extract consistent body part configurations from each frame. Our scoring function for each frame f is written as the concatenation of the unary and pairwise features:

$$S(l_i, cc'|f) = [U(l_i, c|f), P(cc')^k] \tag{4.3}$$

Given a human activity time-slice, we extract the unary and pairwise features f_u and f_p for all body joints of people appearing in each frame. In each frame, the dimensions of the unary array are $w \times h \times 14$ and the dimensions of the pairwise array are $w \times h \times 182 \times 2$. w and h are referred to as the width and height of an image patch, respectively. 14 denotes the number of the body part classes, 2 corresponds to x and y components of the pairwise offset between the two body parts, and 182 shows all cross-part pairs (directed, e.g. there are separate predictions from wrist to elbow and elbow to wrist), therefore it doesn't contain e.g. wrist-wrist relation.

The scoring functions of frames in a time-slice are aggregated over time to obtain a descriptor for the whole time-slice. The time-slice descriptors are exploited to train classifiers to categorize the likelihoods of activities. In Section 4.4, we give more details on aggregation strategies. We also evaluate the contribution of the unary and pairwise terms in our task of uncertainty interpretation in Section 4.5.

4.4 Learning

Given unary and pairwise features f_u and f_p for each frame t of the time-slice, we proceed with the aggregation of f_u and f_p over all frames to obtain a descriptor per time-slice. We con-

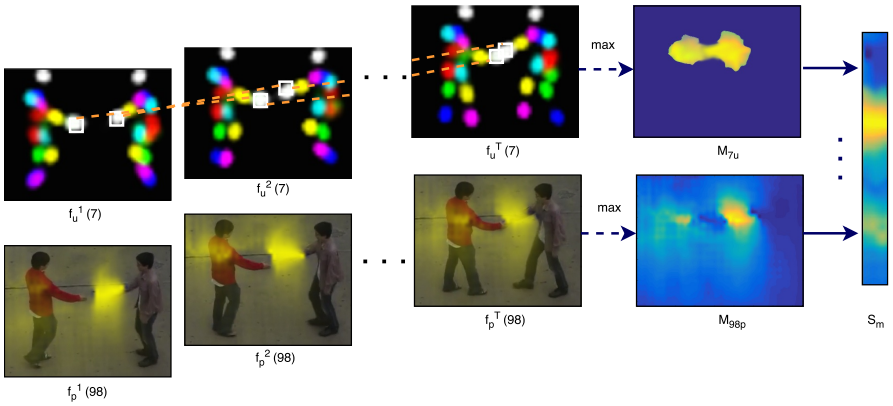


Figure 4.5: Aggregation framework. In this example, *max* aggregation, unary probabilities of right wrist ($i = 7$), and pairwise probabilities (c =right wrist, c' =right elbow), i.e., $j = 98$ are illustrated.

sider three different aggregation schemes to generate time-slice descriptors: simple aggregation (S_s), *max* aggregation (S_m), and $max - min$ aggregation (S_{mm}) by computing minimum and maximum values for each body joint i and cross-part pair j over T frames in each time-slice (See Figure 4.5)

$$\begin{aligned} m_{iu} &= \min_{1\leq t\leq T} f_u^t(i), \quad i = 1, ..., 14 \\ M_{iu} &= \max_{1\leq t\leq T} f_u^t(i), \quad i = 1, ..., 14 \\ m_{jp} &= \min_{1\leq t\leq T} f_p^t(j), \quad j = 1, ..., 128 \\ M_{jp} &= \max_{1\leq t\leq T} f_p^t(j), \quad j = 1, ..., 128 \end{aligned} \tag{4.4}$$

Following Eq. 4.3, the time-slice descriptor is defined by concatenating time-aggregated features as:

$$\begin{aligned} S_s &= [f_u^1, ..., f_u^T, f_p^1, ..., f_p^T] \\ S_m &= [M_{1u}, ..., M_{14u}, M_{1p}, ..., M_{128p}] \\ S_{mm} &= [m_{1u}, ..., m_{14u}, m_{1p}, ..., m_{128p}, S_m] \end{aligned} \tag{4.5}$$

max aggregation can be interpreted as the highest probability of presenting each body joint and the maximum relative distance of body joint pairs throughout the time-slice while *min* aggregation shows the lowest possibility of configuring body joints in certain ways. In Section 4.5 we evaluate the effect of different aggregation schemes. The three sets of descriptors are fed into SVM with an RBF kernel and KNN classifiers to cluster the uncertainty associated with dyadic human activities to 5 categorical likelihood $\mathcal{L} = 1, ..., 5$ i.e., the activity is definitely not occurring ($\mathcal{L} = 1$), unlikely to occur ($\mathcal{L} = 2$), neither likely nor unlikely to occur ($\mathcal{L} = 3$),

likely to occur ($\mathcal{L} = 4$), and definitely occurring ($\mathcal{L} = 5$). Dyadic activities are: handshake, high-five, hug, kick, kiss, punch, and push.

In a preliminary experiment, we applied a baseline uncertainty measurement on the interpretation of SVM scores on classifying 7 activity classes. We observed that the baseline method computes the general uncertainty by fitting a logistic regression model to the classifier's scores while explicit uncertainty classifiers show more detail by analysing uncertainty in every time-slice. This observation confirms the importance of applying explicit uncertainty classifiers instead of simply taking the SVM scores of activity classifiers into account.

4.5 Experimental results

In this section, we present the dataset used to evaluate the proposed method along with quantitative and qualitative results.

4.5.1 Dataset and evaluation criteria

In our experiments we use the TAP (Time-slice Activity Prediction) dataset (44) to evaluate the performance of our proposed method. TAP is the only dataset suitable for this task since it contains the annotations corresponding to categorical likelihoods of activities. The TAP dataset contains time-slices of 7 dyadic human activities. Each time-slice is a 10-frame video clip trimmed from a longer video with consecutive frames without overlap (i.e. $T = 10$ in Eq. 4.4 and 4.5). In (44), it was shown that 10-frame time-slices are more effective compared to 5- and 15-frame time-slices and that 5-option categorical likelihood best describes the occurrence of activities compared to 3- and 7-option categorical likelihoods. TAP has two subsets, i.e., **constrained** and **unconstrained** subsets. The constrained set includes 1129 time-slices extracted from staged scenarios of the UT-Interaction dataset (27), while the unconstrained set contains 990 time-slices extracted from a subsample of real-world scenarios of HMDB (28), TV Interaction (29), and Hollywood (30) datasets. This subsample is selected based on the camera angle ranging from -45 to +45 degrees. Each time-slice was annotated by 3 different annotators (using the Crowdflower platform (31)) on how likely each of the 7 dyadic activities are occurring. The annotators were requested to pick one of 5 categorical likelihoods from 1 (definitely not occurring) to 5 (definitely occurring). We set the medians of the 3 annotated values as the ground-truth class labels.

Since the original image data contains limited training images, we ran data augmentation algorithms to boost the performance of training. For this, we used horizontal flipping, translation, and fancy PCA (51). Therefore, we fine-tune the CNN model on 1129x3x10 (10 frames per time-slice) and 900x3x10 frames in constrained and unconstrained subsets, respectively. Performance on the dataset is measured in terms of mean average precision (mAp) across

Table 4.1: Effects of different versions of the proposed pipeline in uncertainty analysis on constrained and unconstrained sets

	Setting	$\mathcal{L}$=1	$\mathcal{L}$=2	$\mathcal{L}$=3	$\mathcal{L}$=4	$\mathcal{L}$=5	mAp
Const.	unary+SVM (S1)	0.811	0.554	0.451	0.639	0.712	0.633
	pairwise+SVM (S2)	0.852	0.573	0.472	0.654	0.736	0.657
	unary+pairwise+SVM (S3)	**0.921**	**0.631**	**0.511**	**0.794**	**0.801**	**0.732**
Unconst.	unary+SVM (S1)	0.798	0.580	0.412	0.661	0.683	0.627
	pairwise+SVM (S2)	0.852	0.615	0.458	0.692	0.723	0.667
	unary+pairwise+SVM (S3)	**0.913**	**0.652**	**0.522**	**0.756**	**0.802**	**0.729**

the 7 classes of activities. This computes the number of true labeled categorical likelihoods compared to ground-truth labels.

4.5.2 Evaluation of unary and pairwise probabilities

In this section, we evaluate the effects of unary and pairwise probabilities in uncertainty categorization of human activities. We extract deep pose information from each time-slice using unary only, pairwise only, and both unary-pairwise features. For simplicity, the features are aggregated using *max*-aggregation method over time-slice to generate a descriptor for the whole time-slice (for further evaluation on different aggregation approaches, refer to 4.5.3). The evaluation is performed by using the standard leave-one-out method, all time-slices trimmed from the same video are out each time, and by fitting the uncertainty interpretation in the context of multi-class classification with 5 categorical likelihoods for each class of activities. Classification is carried out with KNN and SVM for comparison.

Using KNN with unary probabilities (KNN+unary) decreases the performance by 2% compared to SVM+unary in constrained set. The performance also drops by 2% and 5% in KNN+pairwise and unary+pairwise+KNN compared to using SVM classifiers. Since KNN consistently performs worse than SVM, we exclude it from the reported results. In Table 4.1 we present the performance of settings S1 (unary+ SVM), S2 (pairwise+SVM), and S3 (unary+pairwise+SVM) in terms of mean average precision (mAp) on constrained and unconstrained sets. The results are obtained based on the number of correctly classified categorical likelihoods in time-slices compared to the human annotation. We also performed a preliminary experiment to evaluate 14 joint- and 12 joint- configurations for unary probabilities. In 12 joint configuration, we remove the information about the locations of head and neck joints since they vary slightly with the change of activities or uncertainty states. The results showed that the 12 joint option was more effective and it was used in all of our experiments.

The unary S1 setting achieves 63.3% and 62.7% mAp while using the pairwise S2 setting significantly improves performance achieving 65.7% and 66.7% mAp on constrained and unconstrained sets. This clearly shows the advantages of using pairwise to find the relation between body joints for multiple individuals and distinguish pose configurations in different

Table 4.2: Comparison of different aggregation schemes: simple, max, and $max - min$ aggregations on constrained and unconstrained sets

Setting	Constrained set (mAp)	Unconstrained set (mAp)
simple-aggr.	0.740	0.732
max-aggr.	0.732	0.729
$max - min$-aggr.	**0.746**	**0.740**

activities and uncertainty likelihoods. Using both unary and pairwise further boosts the performance (73.2% and 72.9% mAp), which can be attributed to better quality pose information since the probability of presenting body joints and dual relations between them are taken into consideration.

4.5.3 Aggregating pose information

Using the best setting from unary and pairwise evaluation experiment, i.e. S3, we compare different aggregation schemes explained in Section 4.4. We have evaluated all settings with different aggregations but we only report results with the best setting here.

Unary and pairwise features are first extracted for each frame and are concatenated over time-slice with three different aggregation methods to generate a descriptor for the whole time-slice. The three aggregation methods are: simple, max, and $max - min$ schemes.

Results of max aggregation for constrained and unconstrained sets are reported in Table 4.1 and compared with other aggregation schemes in Table 4.2. Table 4.2 shows the impact of combining min and max aggregations which leads to a 1.4% improvement over max only aggregation on constrained and 1.1% on unconstrained sets. The $max - min$ aggregation slightly improves the performance over simple aggregation (0.06% and 0.08%); however, it results into a 4-5x run-time [2] reduction since the dimension of feature vectors is dramatically reduced by 5 orders of magnitude. Overall, we observe performance improvement and dramatic reduction in run-time by applying $ma - min$ aggregation approach.

4.5.4 Time-slice uncertainty interpretation

In the remaining evaluation, we report results with the best version of our method, i.e., unary+pairwise+SVM+$max - min$-aggr.

Across all the experiments, we notice that the best results for classified categorical likelihoods belongs to $\mathcal{L} = 1$ and $\mathcal{L} = 5$. This is due to the fact that time-slices in these two categories contain more discriminative information in pose configuration to distinguish action classes. For instance, the "two people are shaking each other's hands" time-slice is easy to categorize as "push is definitely not occurring" ($\mathcal{L} = 1$) and "handshake is definitely occurring" ($\mathcal{L} = 5$).

[2] Run-time is measured on a single core Intel 2.60GHz

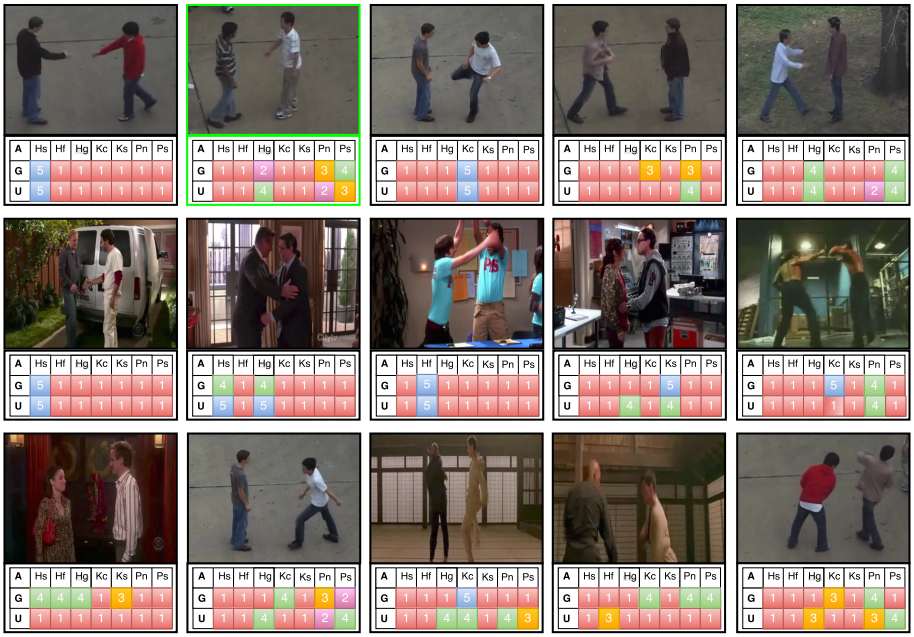


Figure 4.6: Successes (two top rows) and failures (bottom row) on both sets. Each time-slice sample is represented by its last frame. Notations G and U show ground-truth and our method uncertainty interpretation, respectively.

However, the "two people are approaching" time-slice in which push and handshaking are neither likely/nor unlikely to occur ($\mathcal{L} = 3$) is hard to categorize. Therefore, $\mathcal{L} = 3$ is the most uncertain category because of the lack of informative pose features to distinguish activities through the time-slice.

We show some qualitative results of the success and failure cases of our best framework in Figure 4.6. Four matched categorical likelihoods between the result of our framework and the ground-truth is the threshold to determine success or failure. Remarkably, the proposed approach correctly handles a case where the human annotator misclassified hug as punch which is shown in the example with the green bounding box. We also illustrate and analyse the failure cases of our proposed method at the bottom of Figure 4.6. The included examples further illustrate the difficulty of the task of uncertainty interpretation in time-slices. First and second examples show cases where the posture is not informative enough to distinguish activities. Another prominent error shown in the remaining examples corresponds to the cases where the activity is performed in a highly different way than other examples in the same category.

4.5.5 Comparison to the state of the art

We compare our proposed uncertainty interpretation method with the state of the art (55) on constrained and unconstrained sets. In the best-performing setting, (55) extracted hand-

Table 4.3: Performance comparison on constrained and unconstrained sets. The annotation (Best) and (Avg.) indicates the highest and the average performance that the particular method can achieve.

	Setting	$\mathcal{L}$=1	$\mathcal{L}$=2	$\mathcal{L}$=3	$\mathcal{L}$=4	$\mathcal{L}$=5	mAp
Const.	Ziaeefard and Bergevin (55) (Avg.)	0.781	0.418	0.361	0.502	0.635	0.539
	Ziaeefard and Bergevin (55) (Best)	0.881	0.502	0.484	0.683	0.737	0.657
	Our method (Avg.)	0.861	0.586	0.478	0.695	0.749	0.674
	Our method (Best)	**0.921**	**0.631**	**0.511**	**0.794**	**0.801**	**0.732**
Unonst.	Ziaeefard and Bergevin (55) (Avg.)	0.826	0.473	0.395	0.555	0.680	0.585
	Ziaeefard and Bergevin (55) (Best)	0.891	0.533	0.486	0.651	0.752	0.662
	Our method (Avg.)	0.854	0.615	0.464	0.703	0.736	0.674
	Our method (Best)	**0.913**	**0.652**	**0.522**	**0.756**	**0.802**	**0.729**

crafted space-time interest points called Predict-STIP for each time-slice and encoded them using Fisher Vectors (FV) method.

Table 4.3 shows that our best version of unary and pairwise features outperform state-of-the-art Predict-STIP+FV descriptors by a large margin (7.5% and 6.7% in constrained and unconstrained sets, respectively). To highlight improvements achieved by our proposed method, we refer to the best results for categorical likelihoods $\mathcal{L} = \{2, 4\}$ of our method over Predict-STIP+FV. It shows the significant improvements $\{12.9\%, 11.1\%\}$ in constrained and $\{11.9\%, 10.5\%\}$ in unconstrained sets. In particular, the $\mathcal{L} = 2$ and $\mathcal{L} = 4$ classes involve more uncertainty associated with human activities meaning smaller localized motion. This lack of discriminative motion information makes classification difficult for space-time Predict-STIP+FV features while our method benefits from the semantic pose information, the existence probabilities of body parts, and the relation between them throughout the time-slice.

The constrained set looks visually simpler. Therefore, extracted unary and pairwise probabilities are more accurate. On the other hand, this set is more uncertain since the constrained set contains staged scenarios and actors perform activities at a slow pace. Thus, 10-frame time-slices do not include enough information all the time. Activities in the unconstrained set are executed at a faster pace and result in more informative time-slices but more challenging for unary and pairwise features. It makes both sets challenging with comparable levels of difficulty and results in almost the same performance range in Table 4.3.

4.6 Conclusions

In this paper, we proposed a method using CNNs to extract unary and pairwise probabilities of human body pose to analyse the uncertainty associated with human activities. Our approach extracts semantic information about pose configuration, including the probability of presence of body joints, and it reasons about relations of a joint with other joints in each frame. These semantic features are aggregated over frames of a time-slice to generate one single

descriptor per time-slice. We feed the descriptors to SVM and KNN classifiers to categorize the likelihoods of activities occurring in time-slices. Our results show significant improvements over the current state of the art for uncertainty analysis in human activities.

Conclusion

The primary objective in this thesis was to present novel and efficient frameworks to analyse human activities from short temporal observations in videos. The problem of analysing human activities in videos has received growing interest in the computer vision community. Conventional methods analyse activities from fully observed videos. An interesting challenge is to explore activities throughout the video and analyse the uncertainty occurring at a small temporal observation windows, referred to as "time-slices". In this thesis, we improved our understanding of the inherent uncertainty occurring with time-slice observations and built computational algorithms to properly model them.

We introduced the new problem of time-slice activity analysis and proposed our first framework which fits in the standard activity recognition research topic. We then move forward toward a new level of generality in the second and third algorithms and explored the uncertainty associated with human activities in time-slices.

The activities of interest in this this thesis are dyadic human activities. Dyadic activities explored include handshake, high-five, hug, kick, kiss, punch, and push. Time-slice dyadic activity analysis has practical applications other than the basic research question of better understanding human-computer perception of dyadic actions. It can be beneficial when the whole video stream is not available and activities are not recorded from start to end. It can also be useful in content-based video search, indexing, retrieval, and summarization where the goal is to explore the content of the video and return time-slices where an activity of interest is occurring or not.

In Chapter 1, we reviewed the literature on recent action recognition frameworks based on semantic information. Recent action recognition methods rely on low-level and mid-level features such as spatio-temporal interest points and trajectories. Although these methods provide reasonable results, adding some semantic information will improve the performance. In this chapter, we introduced a semantic space which mainly includes pose, poselet, object/scene context, and attributes. Semantic descriptions capture meaningful information and are robust to visual variations. Afterward, we compared the performance of semantic and non-semantic methods. Experiments showed that semantic methods outperform non-semantic based methods in most cases except when different activities share similar poses/attributes. In such cases,

combining several features improves the performance.

In Chapter 2, we introduced a predictive representation for the new problem of time-slice activity recognition. We predicted the occurrence of an activity using a portion of the whole activity in time-slice activity recognition. We represented each video based on novel spatio-temporal descriptors extracted from discriminative video segments. Furthermore, we introduced a new dataset (TAP dataset) for time-slice activity recognition to evaluate the performance of our method and compare it to the state-of-the-art. The TAP dataset is publicly available to encourage researchers to explore human activities at a smaller temporal granularity.

In Chapter 3, we introduced the problem of analysing the uncertainty associated with human activities. In this chapter, a novel approach for uncertainty integration in the analysis of human activities is proposed. We classified time-slices into one of 5 categorical likelihoods from "Definitely Not Occurring" to "Definitely Occurring" which shows how likely each activity is occurring in the time-slice. We proposed a learning framework for this classification and evaluated the performance of this framework using combinations of state-of-the-art representations (Predict-STIP, SSTIP, FV, SIFT, and BoW). In our experimental results, we observed that across all of these representations, the combination of Predict-STIP and FV outperforms other settings. We also proposed a novel technique for evaluating the performance of early activity recognition. Besides, we explained how uncertainty analysis helps to improve training data. The most certain a partially observed video is, the more it helps to recognize the true activity which makes it a better candidate for training data.

In Chapter 4, following the success of deep learning frameworks in other computer vision applications, we introduced a CNNs-based model to categorize the uncertainty associated with human activities. We extracted unary and pairwise probabilities of human body pose to capture the local appearance of body parts as well as the contextual relations between the parts. Our method first extracts semantic information about pose configuration including the probability of presence of body joints and then reasons about relations of a joint with other joints in each frame. We concatenated unary and pairwise features and aggregated them over frames of a time-slice to generate one single descriptor per time-slice. The time-slice descriptors are fed to classifiers to categorize the likelihoods of activities occurring in time-slices.

Several possible real-world applications could exploit the material presented in this thesis. The presented methodology of time-slice activity analysis and integrating uncertainty in activity prediction can be optimized for surveillance systems. It can be helpful when the whole video stream is not available and activities are not recorded from start to end. It can also be beneficial for video retrieval to return video fragments where an activity of interest is (not) occurring with a specific degree of uncertainty.

We hope that this thesis will inspire other researchers to devote more attention to time-slice activity analysis under uncertainty.

The only thing that makes life possible is permanent, intolerable uncertainty; not knowing what comes next.

Ursula K. Le Guin

Bibliography

[1] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2650–2657, 2013.

[2] M. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1036–1043, 2011.

[3] F.-F. Li and A. Karpathy, "Cs231n: Convolutional neural networks for visual recognition," 2016.

[4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision (ECCV)*, pp. 818–833, 2014.

[5] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428 – 440, 1999.

[6] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, 2008.

[7] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976 – 990, 2010.

[8] K. Li and Y. Fu, "Arma-hmm: A new approach for early recognition of human activity," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 1779–1782, 2012.

[9] M. Hoai and F. De la Torre, "Max-margin early event detectors," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2863–2870, 2012.

[10] G. Yu, J. Yuan, and Z. Liu, "Predicting human activities using spatio-temporal structure of interest points," in *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 1049 – 1052, 2012.

[11] K. Li and Y. Fu, "Prediction of human activity by discovering temporal sequence patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1644–1657, 2014.

[12] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. Siskind, and S. Wang, "Recognize human activities from partially observed videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2658–2665, 2013.

[13] I. Laptev and T. Lindeberg, "Space-time interest points," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 432–439, 2003.

[14] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proceedings of Fourth Alvey Vision Conference*, pp. 147–151, 1988.

[15] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, 2005.

[16] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

[17] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 3, pp. 710–719, 2005.

[18] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzalez, "Selective spatio-temporal interest points," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 396 – 410, 2012. Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

[19] D. Das Dawn and S. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector," *The Visual Computer*, pp. 1–18, 2015.

[20] S. Carlsson and J. Sullivan, "Action recognition by shape matching to key frames," in *IEEE Computer Society Workshop on Models versus Exemplars in Computer Vision*, 2001.

[21] Z. Zhao and A. M. Elgammal, "Information theoretic key frame selection for action recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2008.

[22] Y. S. Sefidgar, A. Vahdat, S. Se, and G. Mori, "Discriminative key-component models for interaction detection and recognition," *Computer Vision and Image Understanding*, vol. 135, no. 0, pp. 16 – 30, 2015.

[23] L. Liu, L. Shao, and P. Rockett, "Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition," *Pattern Recognition*, vol. 46, no. 7, pp. 1810 – 1818, 2013.

[24] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 3551–3558, 2013.

[25] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 104–111, 2009.

[26] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2004–2011, 2009.

[27] M. S. Ryoo and J. K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)." http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.

[28] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[29] A. Patron, M. Marszalek, A. Zisserman, and I. Reid, "High five: Recognising human interactions in tv shows," in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 50.1–50.11, 2010.

[30] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

[31] www.crowdflower.com.

[32] J. Fleiss *et al.*, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[33] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, pp. 674–679, 1981.

[34] C. Tomasi and T. Kanade, "Detection and tracking of point features," tech. rep., International Journal of Computer Vision, 1991.

[35] P. D. Kovesi, "MATLAB and Octave functions for computer vision and image processing." Centre for Exploration Targeting, School of Earth

and Environment, The University of Western Australia. Available from: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>.

[36] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893 vol. 1, 2005.

[38] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision (ECCV)*, vol. 3952 of *Lecture Notes in Computer Science*, pp. 428–441, 2006.

[39] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European Conference on Computer Vision (ECCV)*, vol. 5303, pp. 650–663, 2008.

[40] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 995–1004, 2008.

[41] J. Y. Halpern, *Reasoning about uncertainty*. MIT Press, 2005.

[42] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, June.

[43] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[44] M. Ziaeefard, R. Bergevin, and L.-P. Morency, "Time-slice prediction of dyadic human activities," in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 167.1–167.13, 2015.

[45] X. Wang, L. Wang, and Y. Qiao, "A comparative study of encoding, pooling and normalization methods for action recognition," in *Proceedings of the 11th Asian Conference on Computer Vision*, pp. 572–585, 2013.

[46] H. L. Yang, A. S. Liu, and L. C. Fu, "Daily activity prediction based on spatial-temporal matrix for ongoing videos," in *Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pp. 258–263, 2015.

[47] D. A. Huang and K. M. Kitani, "Action-reaction: Forecasting the dynamics of human interaction," in *European Conference on Computer Vision (ECCV)*, 2014.

[48] B. Liu, "Uncertainty theory," in *Uncertainty Theory*, vol. 300, pp. 1–79, Springer Berlin Heidelberg, 2010.

[49] D. Oneata, J. Verbeek, and C. Schmid, "Action and Event Recognition with Fisher Vectors on a Compact Feature Set," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1817–1824, 2013.

[50] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classification*, pp. 61–74, MIT Press, 1999.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Conference on Neural Information Processing Systems (NIPS)*, pp. 1106–1114, 2012.

[52] D. Hubel and T. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574 – 591, 1959.

[53] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[54] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," *CoRR*, vol. abs/1412.6806, 2014.

[55] M. Ziaeefard and R. Bergevin, "Integration of uncertainty in the analysis of dyadic human activities," in *2016 13th Conference on Computer and Robot Vision (CRV)*, pp. 208–215, 2016.

[56] K. He, X. Zhang, S.Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[57] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1725–1732, 2014.

[58] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Conference on Neural Information Processing Systems (NIPS)*, pp. 568–576, 2014.

[59] G. Gkioxari and J. Malik, "Finding action tubes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[60] E. Ijjina and K. Chalavadi, "Human action recognition using genetic algorithms and convolutional neural networks," *Pattern Recognition*, vol. 59, no. C, pp. 199–212, 2016.

[61] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari, and G. Mori, "Deep structured models for group activity recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[62] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN Features for Action Recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[64] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[65] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4305–4314, 2015.

[66] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schieke, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision (ECCV)*, 2016.

[67] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," in *ICLR*, 2016.

[68] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015.

[69] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[70] C. Cedras and M. Shah, "Motion-based recognition a survey," *Image and Vision Computing*, vol. 13, pp. 129-155, 1995.

[71] D. Gavrila, "The visual analysis of human movement: a survey." *Computer Vision Image Understanding*, vol. 73, pp. 82-98, 1999.

[72] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *IEEE Transaction in Circuits System and Video Technology*, vol. 18, pp. 1473-1488, 2008.

[73] J.K. Aggarwal and M.S. Ryoo, "Human Activity Analysis: A Review," *ACM Computing Surveys*, vol. 43, pp. 1-43, 2011.

[74] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition*, vol. 47, pp. 3343-3361, 2014.

[75] P.E. Downing, Y. Jiang, M. Shuman, and N. Kanwisher, "A cortical area selective for visual processing of the human body," *Science*, vol. 293, pp. 2470-2473, 2001.

[76] M.V. Peelen and P.E. Downing, "Selectivity for the human body in the fusiform gyrus," *Journal of Neurophysiology*, vol. 93, pp. 603-608, 2005.

[77] R.F. Schwarzlose, C.I. Baker, and N.G. Kanwisher, "Separate face and body selectivity on the fusiform gyrus," *Journal of Neuroscience*, vol. 25, pp. 11055-11059, 2005.

[78] K. Allan, "Linguistic Meaning," *New York: Routledge Kegan Paul*, vol. 1. 1986.

[79] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex," *Brain journal*, vol. 2, pp. 593-609, 1996.

[80] A.G. Huth, S. Nishimoto, A.T. Vu, and J.L. Gallant, "A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain," *Neuron*, vol. 76, pp. 1210-1224, 2012.

[81] S. Park and J.K. Aggarwal, "Recognition of Human Interaction Using Multiple Features in Grayscale Images," *In Proceedings of International Conference on Pattern Recognition*, pp. 51-54, 2000.

[82] N. Ikizler and P. Duygulu,' 'Human Action Recognition Using Distribution of Oriented Rectangular Patches," *In Human Motion ICCV*, pp. 271-284, 2007.

[83] D. Forsyth and M. Fleck, "Body plans," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 678-683, 1997.

[84] Y. Yang and D. Ramanan, "Articulated Pose Estimation with Flexible Mixtures of Parts," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[85] N. Yukita, "Iterative Action and Pose Recognition Using Global-and-Pose Features and Action-Specific Models," *In IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 476-483, 2013.

[86] M. Andriluka, L. Sigal, and M.J. Black, "Benchmark datasets for pose estimation and tracking," *In Visual Analysis of Humans: Looking at People, Moesland, Hilton, Krüger and Sigal*, pp. 253-274, 2011.

[87] J.A. Webb and J.K. Aggarwal, "Visually interpreting the motion of objects in space," *IEEE Computer*, pp. 40-46, 1981.

[88] F.J. Perales and J. Torres, "A system for human motion matching between synthetic and real image based on a biomechanic graphical model," *In Proceedings of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 83-88, 1994.

[89] A. Shio and J. Sklansky, "Segmentation of people in motion," *In Proceedings of IEEE Workshop on Visual Motion*, pp. 325-332, 1991.

[90] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 1297-1304, 2011.

[91] A. Yao, J. Gall, G. Fanelli, and L.V. Gool, "Does Human Action Recognition Benefit from Pose Estimation?" *In Proceedings of the British Machine Vision Conference (BMVC)*. vol. 67, pp. 1-11, 2011.

[92] L. Bourdev, and J. Malik, "Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 1365-1372, 2009.

[93] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting People Using Mutually Consistent Poselet Activations," *European Conference on Computer Vision (ECCV)*, pp. 168-181, 2010.

[94] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3177-3184, 2011.

[95] S. Maji and J. Malik, "Object detection using a max-margin hough transform," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1038-1045, 2009.

[96] Y., Wang, D. Tran, and Z. Liao, "Learning Hierarchical Poselets for Human Parsing," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1705-1712, 2011.

[97] B., Holt, E.J. Ong, H. Cooper, and R. Bowden, "Putting the pieces together: Connected Poselets for human pose estimation," *International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1196-1201, 2011.

[98] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[99] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet Conditioned Pictorial Structures," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 588-595, 2013.

[100] A. Klaser , M. Marszałek , I. Laptev , and C. Schmid, "Will person detection help bag-of-features action recognition?" *Research Report*, no. RR-7373, 2010.

[101] L. Vaina and M. Jaulent, "Object structure and action requirements: A compatibility model for functional recognition," *International Journal of Intelligent Systems.* vol. 6, pp. 313–336, 1991.

[102] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing Objects by their Attributes," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1778-1785, 2009.

[103] L. Bourdev, S. Maji, and J. Malik, "Describing People: A Poselet-Based Approach to Attribute Classification," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 1543-1550, 2011.

[104] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3337-3344, 2011.

[105] C.H. Lampert, H. Nickisch, and S. Harmeling, "Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 951-958, 2009.

[106] T.B. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture," *Computer Vision and Image Understanding*, vol. 81, pp. 231-268, 2001.

[107] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision (IJCV)*, vol. 61, pp. 55-79, 2005.

[108] D. Ramanan, "Learning to parse images of articulated bodies," *In Neural Information Processing Systems (NIPS)*, pp. 1129-1136, 2006.

[109] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter sensitive hashing," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 750-757, 2011.

[110] J. Sullivan and S. Carlsson, "Recognizing and tracking human action," *In 7th European Conference on Computer Vision (ECCV)*, pp. 629-644, 2002.

[111] J. Starck and A. Hilton, "Spherical matching for temporal correspondence of non-rigid surfaces," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 1387-1394, 2005.

[112] R. Kehl, M. Bray, and L. VanGool, "Full body tracking from multiple views using stochastic sampling," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 129-136, 2005.

[113] M.B. Holte, C. Tran, M.M. Trivedi, and T.B. Moeslund, "Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments," *IEEE Journal of Selected Topics in Signal Proceedings*, vol. 6, pp. 538-552, 2011.

[114] K. Raja, I. Laptev, P. Perez, and L. Oisel, "Joint pose estimation and action recognition in image graphs," *In 18th IEEE Inter. Conference on Image Processing (ICIP)*, pp. 25-28, 2011.

[115] S. Mukherjee, S.K. Biswas, and D.P. Mukherjee, "Recognizing interactions between human performers by 'Dominating Pose Doublet," *Journal of Machine Vision and Applications*, vol. 25, pp. 1033-1052, 2014.

[116] B. Yao and L. Fei-Fei, "Action Recognition with Exemplar Based 2.5D Graph Matching," *In European Conference on Computer Vision (ECCV)*, pp. 173-186, 2012.

[117] C. Wang, Y. Wang, and A.L. Yuille, "An Approach to Pose-Based Action Recognition," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 915-922, 2013.

[118] L. Meng, L. Qing, P. Yang, J. Miao, X. Chen, and D.N. Metaxas, "Activity recognition based on semantic spatial relation," *In IEEE International Conference on Pattern Recognition (ICPR)*, pp. 609-612, 2012.

[119] A. Eweiwi, S. Cheema, C. Thurau, and C. Bauckhage, "Temporal key poses for human action recognition," *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1310-1317, 2011.

[120] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transaction Pattern recognition and Machine Intelligence*, vol. 23, pp. 257-267, 2001.

[121] A.A. Chaaraoui, P.C. Prez, and F.F. Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, pp. 1799-1807, 2013.

[122] S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage, "Action recognition by learning discriminative key poses," *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1302-1309, 2011.

[123] A. Vahdat, B. Gao, M. Ranjbar, and G. Mori, "A discriminative key pose sequence model for recognizing human interactions," *In IEEE International Workshop on Visual Surveillance*, pp. 1729-1736, 2011.

[124] F. Lv and R. Nevatia, "Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2007.

[125] S. Park and J.K. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," *Journal of Multimedia Systems*, vol. 10, pp. 164-179, 2004.

[126] F.S. Khan, J. van de Weijer, R.M. Anwer, M. Felsberg, and C. Gatta, "Semantic Pyramids for Gender and Action Recognition," *IEEE Transaction on Image Processing*, vol. 23, pp. 3633-3645, 2014.

[127] W. Yang, Y. Wang, and G. Mori, "Recognizing Human Actions from Still Images with Latent Poses," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2030-2037, 2010.

[128] Y. Zheng, Y.J. Zhang, X. Li, and B.D. Liu, "Action Recognition in Sill Images Using A Combination of human pose and context information," *In 19th IEEE International Conference on Image Processing (ICIP)*, pp. 785-788, 2012.

[129] M. Nabi, A.D. Bue, and V. Murino, "Temporal Poselets for Collective Activity Detection and Recognition," *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 500-507, 2013.

[130] J. Wang, X. Nie, Y. Xia, and Y. Wu, "Mining Discriminative 3D Poselet for Cross-view Action Recognition," *Applications of Computer Vision (WACV)*, pp. 634-639, 2014.

[131] C.Y. Chen and K. Grauman, "Watching Unlabeled Video Helps Learn New Human Actions from Very Few Labeled Snapshots," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 572-579, 2013.

[132] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2929-2936, 2009.

[133] Y. Zhang, W. Qu, and D. Wang, "Action-scene Model for Human Action Recognition from Videos," *2nd AASRI Conference on Computational Intelligence and Bioinformatics*, vol. 6, pp. 111–117, 2014,

[134] J. Liu, H. Xiang, Y. Shi, and D. Yu, "Action Recognition with Trajectory and Scene," *International Conference on Digital Home (ICDH)*. pp. 63-68, 2012.

[135] M.M. Ullah, S.N. Parizi, and I. Laptev, "Improving bag-of-features action recognition with non-local cues," *In Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1-11, 2010.

[136] S. Jones, and L. Shao, "Unsupervised Spectral Dual Assignment Clustering of Human Actions in Context," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 604-611, 2014.

[137] A. Gupta and L.S. Davis, "Objects in Action: An Approach for Combining Action Understanding and Object Perception," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2007.

[138] A. Gupta, A. Kembhavi, and L.S. Davis, "Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol 31, no. 10, pp. 1775-1789, 2009.

[139] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg, "A Scalable Approach to Activity Recognition Based on Object Use," *Proceedings In IEEE International Conference on Computer Vision (ICCV)*, pp. 1-8, 2007.

[140] R. Filipovych and E. Ribeiro, "Recognizing Primitive Interactions by Exploring Actor-Object States," *Proceedings IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, pp. 1-7, 2008.

[141] Y. Kuniyoshi and M. Shimozaki, "A Self-Organizing Neural Model for Context Based Action Recognition," *IEEE Engineering Medicine and Biology Society Conference on Neural Engineering*, pp. 442-445, 2003.

[142] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," *Neural Information Processing Systems (NIPS)*, pp. 1503-1511, 2011.

[143] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for static human-object interactions," *In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 9-16, 2010.

[144] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9-16, 2010.

[145] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17-24, 2010.

[146] L.J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 1-8, 2007.

[147] N. Ikizler-Cinbis and S, Sclaroff, "Object, scene and actions: combining multiple features for human action recognition," *In European Conference on Computer Vision (ECCV)*. vol. 6311, pp. 494-507, 2010.

[148] D. Han, L. Bo, and C. Sminchisescu, "Selection and context for action recognition," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 1933-1940, 2009.

[149] G. Sharma, F. Jurie, and C. Schmid, "Expanded Parts Model for Human Attribute and Action Recognition in Still Images," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 652-659, 2013.

[150] W. Li and N. Vasconcelos, "Recognizing Activities by Attribute Dynamics," *Advances in Neural Information Processing Systems (NIPS)*, pp. 1115-1123, 2012.

[151] Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 707-714, 2011.

[152] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Attribute Regularization Based Human Action Recognition," *IEEE Transaction on Information Forensics and Security*, pp. 1600-1609, 2013.

[153] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script Data for Attribute-based Recognition of Composite Activities," *European Conference on Computer Vision (ECCV)*, pp. 144-157, 2012.

[154] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L. Guibas, and L. Fei-Fei, "Human Action Recognition by Learning Bases of Action Attributes and Parts," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 1331-1338, 2011.

[155] M.S. Ryoo and J.K. Aggarwal, "Hierarchical Recognition of Human Activities Interacting with Objects," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8.

[156] M.S. Ryoo and J.K. Aggarwal, "UT-Interaction Dataset," *ICPR contest on Semantic Description of Human Activities (SDHA)*, 2010.

[157] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," *In Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1-11, 2010.

[158] M. Everingham, L.V. Gool, C.K.I. Williams, and J. Winn, A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*. vol. 88, pp. 303-338, 2010.

[159] A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2061-2068, 2010.

[160] M.S. Ryoo and J.K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," *In IEEE International Conference on Computer Vision (ICCV)*, pp. 1593-1600, 2009.

[161] P. Matikainen, M. Hebert, and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," *In European Conference on Computer Vision (ECCV)*, pp. 508-521, 2010.

[162] Y. Kong, Y. Jia, and Y. Fu, "Learning human interaction by interactive phrases," *In European Conference on Computer Vision (ECCV)*, pp. 300-313, 2012.

[163] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2169-2178, 2006.

[164] F.S. Khan, J. van de Weijer, A. Bagdanov, and M. Felsberg, "Scale coding bag-of-words for action recognition," *In IEEE International Conference on Pattern Recognition (ICPR)*, 2014.

[165] F.S. Khan, R.M. Anwer, J. van de Weijer, A. Bagdanov, A. Lopez, and M. Felsberg, "Coloring action recognition in still images," *IJCV*, vol. 105, pp. 205-221, 2013.

[166] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, "YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition," *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2712-2719, 2013.

[167] T.S. Motwani and R.J. Mooney, "Improving Video Activity Recognition using Object Recognition and Text Mining," *In Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, pp. 600-605, 2012.

[168] S. Chen, J. Liu, H. Wang, and J.C. Augusto, "A hierarchical human activity recognition framework based on automated reasoning," *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3495-3499, 2013.

[169] G.J. Wang and H.J. Zhou, "Introduction to Mathematical Logic and Resolution Principle. *2nd edition, Oxford: Alpha Science International Limited*, 2009.

[170] S. Chen, K. Clawson, M. Jing, J. Liu, H. Wang, and B. Scotney, "Uncertainty reasoning based formal framework for big video data understanding," *Proceedings of*

IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technologies, pp. 487-494, 2014.

[171] X. Liu, K. Clawson, H. Wang, B. Scotney, and J. Liu, "Complex event recognition with uncertainty reasoning," *Proceedings of International Conference on Machine Learning and Cybernetics*, pp. 1823-1828, 2013.

[172] M.S. Ryoo and J.K. Aggarwal, "Semantic representation and recognition of continued and recursive human activities," *International Journal of Computer Vision (IJCV)*, vol. 82, no. 1, pp. 1-24, 2009.

[173] K. Ramirez-Amaro, E.S. Kim, J. Kim, B.T. Zhang, M. Beetz, and G. Cheng, "Enhancing Human Action Recognition through Spatio-temporal Feature Learning and Semantic Rules," *In IEEE-RAS International Conference*, 2013.

[174] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell, "Zero-Shot Learning with Semantic Output Codes," *Neural Information Processing Systems (NIPS)*, pp. 1410-1418, 2009.

[175] H.T. Cheng, F.T. Sun, M.L. Griss, P. Davis, J. Li, and D. You, "NuActiv: recognizing unseen new activities using semantic attribute-based learning," *In Proceedings ACM Internat, Conference Mobile Systems, Applications, and Services (MobiSys)*, pp. 361-374, 2013.

[176] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, Y. Haonan, A. Michaux, L. Yuewei, S. Dickinson, J.M. Siskind, and W. Song, "Recognizing Human Activities from Partially Observed Videos," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2658-2665, 2013.

[177] D. Munoz, J.A. Bagnell, and M. Hebert, "Stacked hierarchical labeling," *In European Conference on Computer Vision (ECCV)*, pp. 57-70, 2010.

[178] D. Munoz, J.A. Bagnell, and M. Hebert, "Co-inference machines for multimodal scene analysis," *In European Conference on Computer Vision (ECCV)*, pp. 668-681, 2012.

[179] C. Baker, R. Saxe, and J. Tenenbaum, "Action understanding as inverse planning," *Cognition*. vol. 13, pp. 329-349, 2009.

[180] K.M. Kitani, B.D. Ziebart, J.A. Bagnell, and M. Hebert, "Activity Forecasting," *In European Conference on Computer Vision (ECCV)*, vol. 7575, pp. 201-214, 2012.

[181] L. Wu, J. Zhang, and F. Yan, "A poselet based key frame searching approach in sports training videos," *Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1-4, 2012.