LUKAS RYBOK

Unsupervised object candidate discovery for activity recognition

Unsupervised object candidate discovery for activity recognition

by
Lukas Rybok

Zur Erlangung des akademischen Grades
eines Doktors der Ingenieurwissenschaften
von der Fakultät für Informatik des
Karlsruher Instituts für Technologie (KIT)
genehmigte Dissertation von Lukas Rybok
Tag der mündlichen Prüfung: 30. Januar 2017
Referenten: Prof. Dr.-Ing. Rainer Stiefelhagen
Prof. Dr.-Ing. Tamim Asfour

Impressum
Verlag: Lukas Rybok, Rheinstraße 14b, 76185 Karlsruhe
Druck: epubli, ein Service der neopubli GmbH, Berlin

Biblografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

ISBN: 978-3-7450-1072-5

Abstract

Automatic interpretation of human motions from videos is an important component of many computer vision applications such as human-robot interaction, surveillance, and multimedia retrieval. While most existing approaches in this area are designed to classify simple actions such as `standing up` or `walking`, the scope of this work lies in the recognition of complex action sequences involving human-object interactions, also known as *activities*.

According to the action identification theory, an activity derives its meaning from the overall context and not from motion alone. Such contextual information may involve, among others, the sum of all previously performed actions, the location where the action in question is executed, as well as the objects that are manipulated by the actor. For instance, when only considering motion information while neglecting object knowledge, it is not possible to discern accurately whether a person raising his hand towards the mouth is `eating`, `drinking`, or `cleaning` his mouth.

Still, most works in action and activity recognition ignore any contextual cues and focus on the identification of activities based on motion alone. On the other hand, approaches that do incorporate object information usually depend on detectors that require supervised training. Since a substantial amount of manually annotated training data is needed to build the detectors, expanding such frameworks (*e.g.*, , adding new action classes) becomes the bottleneck for generalized tasks. Another disadvantage of obtaining object knowledge by relying on supervised detectors lies in the unreliability of state-of-the-art general purpose object detection approaches. Hence, the main goal of this work is to boost activity recognition performance by augmenting motion features with object information that can be obtained without any supervision.

Us humans have the remarkable ability to selectively attend to an area of the visual field while ignoring the surrounding regions. This process is known as attention and serves us as a selective gating mechanism that decides what will be processed at later stages (*e.g.*, object recognition).

The regions that draw our attention are also known as *proto-objects* and are defined as volatile units of visual information that may be validated as actual objects through focused attention. Or, to put it in other words, proto-objects are object- or object-part candidates that have been detected, but not yet identified. Motivated by the ability of humans to reliably determine such visually salient regions from the background, many approaches have been proposed in the literature to detect proto-objects with the least statistical knowledge of the objects themselves.

Proto-object features for activity recognition

Since visual attention and object recognition are tightly linked processes in the human visual system, there is an increasing interest in integrating both concepts to improve the performance of computer vision systems. In this work, we show that proto-object detection also allows us to find object candidate regions that can be used as an additional cue for motion-based activity recognition. To this end, we make use of a very fast visual saliency estimation method, that is based on quaternion DCT image signatures.

For the selection of a set of object candidates (*i.e.*, proto-objects) from the saliency maps, we propose an approach implementing the concept of inhibition of return. The extracted object-candidate features are further used in conjunction with state-of-the-art local spatio-temporal Bag-of-Words motion encoding methods to classify complex activities of daily living.

In an experimental evaluation on several widely used benchmark data sets, we demonstrate that proto-object based features yield a superior performance compared to only using motion information. Surprisingly, the proposed approach also outperforms methods relying on object knowledge from supervised detectors or manual annotations. Furthermore, the reported classification accuracy results in a clear improvement over current state-of-the-art methods for activity recognition.

KIT Robo-kitchen activity data set

Benchmark data sets are an important tool to assess the performance of computer vision approaches. Due to a lack of suitable benchmarks allowing an in-depth comparison of methods for the recognition of complex activities, part of this work consisted of the setup and collection of the now publicly available KIT Robo-kitchen data set. As the name suggests, a kitchen scenario has been chosen as the setting for the video recordings, since it provides a vast range of possible everyday activities involving human-object interactions. The participants were barely restricted in how to perform the activities which resulted in a collection of natural motions with much variation as opposed to most currently existing action and activity recognition data sets. Therefore the subjects only got brief information about what to do, such as where to find the required objects, or to perform the activity at a location of their choice at the table.

The resulting data set has since its publication served as a benchmark to evaluate the performance of several approaches aimed at the recognition of complex, long-lasting, quasi-periodic, and realistic human activities.

In summary, the main contributions of this work are:

- Recording and exploration of a novel video data set allowing the benchmarking of activity recognition approaches aiming at applications in the household robotics domain
- Introduction of proto-object based features as a contextual cue for activity recognition
- Experimental demonstration of the newly proposed features' superiority compared to state-of-the-art

Kurzzusammenfassung

Die automatische Interpretation menschlicher Bewegungsabläufe auf Basis von Videos ist ein wichtiger Bestandteil vieler Anwendungen im Bereich des Maschinellen Sehens, wie zum Beispiel Mensch-Roboter Interaktion, Videoüberwachung, und inhaltsbasierte Analyse von Multimedia Daten. Anders als die meisten Ansätze auf diesem Gebiet, die hauptsächlich auf die Klassifikation von einfachen Aktionen, wie `Aufstehen`, oder `Gehen` ausgerichtet sind, liegt der Schwerpunkt dieser Arbeit auf der Erkennung menschlicher Aktivitäten, d.h. komplexer Aktionssequenzen, die meist Interaktionen des Menschen mit Objekten beinhalten.

Gemäß der Aktionsidentifikationstheorie leiten menschliche Aktivitäten ihre Bedeutung nicht nur von den involvierten Bewegungsmustern ab, sondern vor allem vom generellen Kontext, in dem sie stattfinden. Zu diesen kontextuellen Informationen gehören unter anderem die Gesamtheit aller vorher furchgeführter Aktionen, der Ort an dem sich die aktive Person befindet, sowie die Menge der Objekte, die von ihr manipuliert werden. Es ist zum Beispiel nicht möglich auf alleiniger Basis von Bewegungsmustern und ohne jeglicher Miteinbeziehung von Objektwissen zu entschieden ob eine Person, die ihre Hand zum Mund führt gerade etwas isst oder trinkt, raucht, oder bloß die Lippen abwischt.

Die meisten Arbeiten auf dem Gebiet der computergestützten Aktions- und Aktivitätserkennung ignorieren allerdings jegliche durch den Kontext bedingte Informationen und beschränken sich auf die Identifikation menschlicher Aktivitäten auf Basis der beobachteten Bewegung. Wird jedoch Objektwissen für die Klassifikation miteinbezogen, so geschieht dies meist unter Zuhilfenahme von überwachten Detektoren, für deren Einrichtung widerum eine erhebliche Menge an Trainingsdaten erforderlich ist. Bedingt durch die hohen zeitlichen Kosten, die die Annotation dieser Trainingsdaten mit sich bringt, wird das Erweitern solcher Systeme, zum Beispiel durch das Hinzufügen neuer Typen von Aktionen, zum eigentlichen Flaschenhals. Ein weiterer Nachteil des Hinzuziehens von überwacht trainierten Objektdetektoren, ist deren Fehleranfälligkeit, selbst wenn die ver-

wendeten Algorithmen dem neuesten Stand der Technik entsprechen. Basierend auf dieser Beobachtung ist das Ziel dieser Arbeit die Leistungsfähigkeit computergestützter Aktivitätserkennung zu verbessern mit Hilfe der Hinzunahme von Objektwissen, welches im Gegensatz zu den bisherigen Ansätzen ohne überwachten Trainings gewonnen werden kann.
Wir Menschen haben die bemerkenswerte Fähigkeit selektiv die Aufmerksamkeit auf bestimmte Regionen im Blickfeld zu fokussieren und gleichzeitig nicht relevante Regionen auszublenden. Dieser kognitive Prozess erlaubt es uns unsere beschränkten Bewusstseinsressourcen unbewusst auf Inhalte zu richten, die anschließend durch das Gehirn ausgewertet werden. Zum Beispiel zur Interpretation visueller Muster als Objekte eines bestimmten Typs. Die Regionen im Blickfeld, die unsere Aufmerksamkeit unbewusst anziehen werden als *Proto-Objekte* bezeichnet. Sie sind definiert als unbestimmte Teile des visuellen Informationsspektrums, die zu einem späteren Zeitpunkt durch den Menschen als tatsächliche Objekte wahrgenommen werden können, wenn er seine Aufmerksamkeit auf diese richtet. Einfacher ausgedrückt: Proto-Objekte sind Kandidaten für Objekte, oder deren Bestandteile, die zwar lokalisiert aber noch nicht identifiziert wurden. Angeregt durch die menschliche Fähigkeit solche visuell hervorstechenden (salienten) Regionen zuverlässig vom Hintergrund zu unterscheiden, haben viele Wissenschaftler Methoden entwickelt, die es erlauben Proto-Objekte zu lokalisieren. Allen diesen Algorithmen ist gemein, dass möglichst wenig statistisches Wissens über tatsächliche Objekte vorausgesetzt wird.

Proto-object Merkmale für die Aktivitätserkennung

Visuelle Aufmerksamkeit und Objekterkennung sind sehr eng miteinander vernküpfte Prozesse im visuellen System des Menschen. Aus diesem Grund herrscht auf dem Gebiet des Maschinellen Sehens ein reges Interesse an der Integration beider Konzepte zur Erhöhung der Leistung aktueller Bilderkennungssysteme. Die im Rahmen dieser Arbeit entwickelten Methoden gehen in eine ähnliche Richtung: wir demonstrieren, dass die Lokalisation von Proto-Objekten es er-

laubt Objektkandidaten zu finden, die geeignet sind als zusätzliche Modalität zu dienen für die bewegungsbasierte Erkennung menschlicher Aktivitäten. Die Grundlage dieser Arbeit bildet dabei ein sehr effizienter Algorithmus, der die visuelle Salienz mit Hilfe von quaternionenbasierten DCT Bildsignaturen approximiert. Zur Extraktion einer Menge geeigneter Objektkandidaten (d.h. Proto-Objekten) aus den resultierenden Salienzkarten, haben wir eine Methode entwickelt, die den kognitiven Mechanismus des *Inhibition of Return* implementiert. Die auf diese Weise gewonnenen Objektkandidaten nutzen wir anschliessend in Kombination mit state-of-the-art Bag-of-Words Methoden zur Merkmalsbeschreibung von Bewegungsmustern um komplexe Aktivitäten des täglichen Lebens zu klassifizieren.
Wir evaluieren das im Rahmen dieser Arbeit entwickelte System auf diversen häufig genutzten Benchmark-Datensätzen und zeigen experimentell, dass das Miteinbeziehen von Proto-Objekten für die Aktivitätserkennung zu einer erheblichen Leistungssteigerung führt im Vergleich zu rein bewegungsbasierten Ansätzen. Zudem demonstrieren wir, dass das vorgestellte System bei der Erkennung menschlicher Aktivitäten deutlich weniger Fehler macht als eine Vielzahl von Methoden, die dem aktuellen Stand der Technik entsprechen. Überraschenderweise übertrifft unser System leistungsmäßig sogar Verfahren, die auf Objektwissen aufbauen, welches von überwacht trainierten Detektoren, oder manuell erstellten Annotationen stammt.

KIT Robo-kitchen activities Datensatz

Benchmark-Datensätze sind ein sehr wichtiges Mittel zum quantitativen Vergleich von computergestützten Mustererkennungsverfahren. Nach einer Überprüfung aller öffentlich verfügbaren, relevanten Benchmarks, haben wir jedoch festgestellt, dass keiner davon geeignet war für eine detaillierte Evaluation von Methoden zur Erkennung komplexer, menschlicher Aktivitäten. Aus diesem Grund bestand ein Teil dieser Arbeit aus der Konzeption und Aufnahme eines solchen Datensatzes, des *KIT Robo-kitchen* Benchmarks. Wie der Name vermuten lässt haben wir uns dabei für ein Küchenszenario entschieden, da es ermöglicht einen großen Umfang an Aktivitäten des

täglichen Lebens einzufangen, von denen viele Objektmanipulationen enthalten. Um eine möglichst umfangreiche Menge natürlicher Bewegungen zu erhalten, wurden die Teilnehmer während der Aufnahmen kaum eingeschränkt in der Art und Weise wie die diversen Aktivitäten auszuführen sind. Zu diesem Zweck haben wir den Probanden nur die Art der auszuführenden Aktivität mitgeteilt, sowie wo die benötigten Gegenstände zu finden sind, und ob die jeweilige Tätigkeit am Küchentisch oder auf der Arbeitsplatte auszuführen ist. Dies hebt KIT Robo-kitchen deutlich hervor gegenüber den meisten existierenden Datensätzen, die sehr unrealistisch gespielte Aktivitäten enthalten, welche unter Laborbedingungen aufgenommen wurden. Seit seiner Veröffentlichung wurde der resultierende Benchmark mehrfach verwendet zur Evaluation von Algorithmen, die darauf abzielen lang andauerne, realistische, komplexe, und quasiperiodische menschliche Aktivitäten zu erkennen.
Zusammenfassend betrachtet bestehen die Hauptbeiträge dieser Arbeit aus den folgenden Punkten:

- Erstellung und Exploration eines neuen Datensatzes, welcher das Benchmarking von Algorithmen zur automatischen Erkennung menschlicher Aktivitäten erlaubt, die vor allem auf Anwendungen im Bereich humanoider Haushaltsroboter ausgerichtet sind

- Einführung von auf Proto-Objekten basierenden Bildmerkmalen zur Beschreibung des Kontexts in welchem menschliche Bewegungen stattfinden und deren Verwendung zur automatischen Aktivitätserkennung

- Experimentelle Demonstration der Überlegenheit der im Rahmen dieser Arbeit entwickelten Bildmerkmale im Vergleich zu dem aktuellen Stand der Technik entsprechenden Methoden

Contents

Acronyms

ADL	Activities of Daily Living
AI	Artificial Intelligence
ANN	Artificial Neural Network
ASLAN	Action Similarity Challenge
BoW	Bag-of-Words Feature Encoding
CAD	Cornell Activity Data Set
CCTV	Close-circuit television
CIE L*A*B	Lightness, and Color Opponent Color Space defined by the Commission internationale de l'éclairage
CMU-MMAC	Carnegie Mellon University Multimodal Activity Database
ConvNet	Convolutional Neural Network
CRF	Conditional Random Field
DBN	Dynamic Bayes Network
DCT	Discrete Cosine Transform
DT	Dense Trajectory Features
EM	Expectation Maximization Algorithm
FFT	Fast Fourier Transform
FV	Fisher Vector Encoding
GMM	Gaussian Mixture Model
GPU	Graphical Processing Unit
HMAX	Hierarchical Model and X
HMDB	Human Motions Database
HMM	Hidden Markov Model
HOF	Histogram of Optical Flow
HOG	Histogram of Oriented Gradients
HOGHOF	Stacked Histogram of Oriented Gradients and Histogram of Optical Flow
HOHA	Hollywood Human Actions Data Set
iDT	Improved Dense Trajectory Features
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IXMAS	INRIA Xmas Motion Acquisition Sequences
KIT	Karlsruhe Institute of Technology
KLT	Kanade–Lucas–Tomasi Feature Tracker
KTH	Royal Institute of Technology (Swedish: Kungliga Tekniska Högskolan)
LBP	Local Binary Patterns Feature Descriptor
LDA	Latent Dirichlet Allocation
LIRIS	Laboratoire d'InfoRmatique en Image et Systèmes d'information

LLC	Locally Linear Coding
LSTM	Long Short Term Memory Network
MBH	Motion Boundary Histogram
MEA	Multi-Environment Action Data Set
MEI	Motion Energy Image
MEMM	Maximum Entropy HMM
MHI	Motion History Image
MINTA	Motion, Intention, and Activity Recognition Data Set
MPII	Max Planck Institute for Informatics
MRF	Markov Random Field
MSR	Microsoft Research
NLP	Natural Language Processing
OAC	Object Action Complex
ORGBD	MSR action recognition on online RGBD Data Set
PCA	Principal Component Analysis
pLSA	Probabilistic Latent Semantic Analysis
QDCT	Quaternion-based Discrete Cosine Transform
RANSAC	Random Sample Concensus
ReLU	Rectified Linear Unit
RFID	Radio-frequency Identification
RGB	The Red-Green-Blue Color Space
RGBD	The Red-Green-Blue-Depth Image Channels
RNN	Recurrent Neural Network
SIFT	Scale Invariant Feature Transform
SPM	Spatial Pyramid Matching
STIP	Space-Time Interest Point Features
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
TUM	Technical University Munich
UCF	University of Central Florida
URADL	University of Rochester Activity of Daily Living Data Set
VICON	Vantage Motion Capture Camera
VLAD	Vector of Locally Aggregated Descriptors
VQ	Vector Quantization

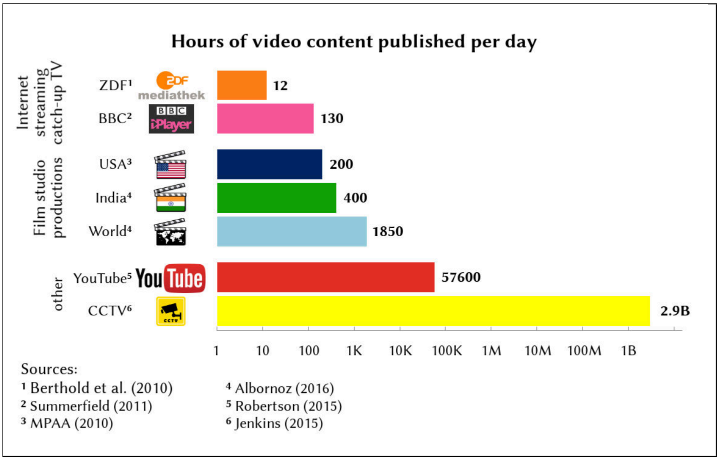
1 Introduction

Starting with the invention of motion picture cameras in the 1890s and the subsequent emergence of television stations in the late 1920s, film production companies were solely responsible for video publishing for nearly a century. Thanks to the technological advances of computers in the last decade, this situation has however drastically changed.

Faster Internet access and at the same time dramatically decreased costs for online storage space made it possible for everybody to publish videos on online streaming platforms like YouTube, Vimeo, or Dailymotion and share them with others. At the same time, video recording devices became constantly accessible to most people in the form of common consumer electronics hardware, like notebooks, mobile phones, and digital photo cameras further intensifying the shift towards user produced content.

To put the vast amount of video data that is made available each hour over various web streaming services into perspective, take a look at Fig. 1.1. According to the 2014 theatrical statistics report of the Motion Picture Association of America (*cf*., MPAA (2015)), each year around 800 feature length movies are being produced in the US (excluding adult video industry). Assuming an average movie length of 90 minutes, this results in 200 hours of content being produced in Hollywood each day, and subsequently made available online on movie streaming platforms like Netflix, Hulu, or Amazon Video.

India, the world's leading nation in cinematic productions even sur-



Sources:
[1] Berthold et al. (2010)
[2] Summerfield (2011)
[3] MPAA (2010)
[4] Albornoz (2016)
[5] Robertson (2015)
[6] Jenkins (2015)

Figure 1.1. *Each day an immense amount of video content is being produced all around the world. However, processing all of the data manually is a nearly impossible task. Thus, automatic video content analysis, including the recognition of human activities, is gaining an increased interest to serve as a viable alternative. Note, that a logarithmic scale is used to properly fit all data.*

passes Hollywood's output by a factor of two (*cf*., Albornoz (2016)). The various Internet streaming catch-up television services around the world are another major source of video data. For instance, 130 hours of original content is daily being added to the British BBC iPlayer (*cf*., Summerfield (2011)). Even though not impossible, tagging this constantly growing amount of videos manually with meta information that facilitates the retrieval of desired clips by the user would still be a very cost- and time-intensive task.

Managing the huge amounts of user generated content, that is daily uploaded to video-sharing websites is, however, a completely different story. As reported in 2015 by TubularInsights.com, the world's leading resource for analysis of the online video industries, more than 1.2 billion videos were made available on YouTube since its beginning in 2005 (*cf*., Robertson (2015)).

The current growth of the streaming platform even amounts to 33.3 % per year, corresponding to more than 57000 hours of content being uploaded each hour. This figures alone should make the necessity of automatic video processing tools self-evident, especially action and activity recognition software.

Still, one could argue that making the creator of the original content responsible for adding such meta information to each uploaded video clip. Nonetheless, many other application scenarios exist where this argument does not hold, such as video surveillance, home entertainment, or patient monitoring. According to a report published by IHS Technology (*cf.*, Jenkins (2015)), more than 245 million professionally installed CCTV devices have been operating worldwide in 2014. Driven by fear of terrorist attacks, burglary, dishonest employees, and vandalism, among others, a steady demand for more surveillance cameras is to be expected (*cf.*, Su et al. (2015)).

Due to the extremely large amount of data that is being recorded with such devices, the camera footage is typically assessed manually. In critical areas like public buildings, this task is carried out by human operators analyzing the output of dozens of cameras at the same time. However, in most cases, the recordings are only consulted after an incident has happened in order to help the authorities solving the cause (*cf.*, Arikuma and Mochizuki (2016)). Thus, putting any ethical controversies aside, surveillance would greatly benefit from automated video content analysis systems, *e.g.*, to sound an alarm once an unusual activity has been detected.

Action and activity recognition is also beneficial for many areas of our everyday life. Systems implementing it have actually already entered our homes in the form of consumer electronics devices that allow the users a motion based interaction with computers, such as the Microsoft XBox Kinect controller for video games, or the Leap Motion controller. Another application scenario is smart-homes, where activity recognition systems can be used for the prediction of the inhabitant's intention in order to pro-actively offer context-aware services.

With an aging demography of many western civilizations, ambient assisted living is yet another domain where human activity recognition is becoming a key component. It can be used to detect anomalies in patterns of activities of daily living in order to allow medical intervention before health problems occur. In nursery homes, video content analysis could be employed to replace traditional nursing alarms, which have a high false positive rate and thus often make the nurses prone to ignore the alarms (*cf*., Bell (2010)). Furthermore, it has been pointed out in clinical studies that the decline of elderly performance in activities of daily living has a potential to indicate emerging symptoms of dementia (*cf*., Laver et al. (2016)). Thus, recognizing these symptoms early enough would allow the patient to partake in treatments delaying further functional decline. Unfortunately, this skill assessment process is very time consuming, as well as error prone and thus could highly benefit from automatic video assessment technology (*cf*., Wilson et al. (2005)).

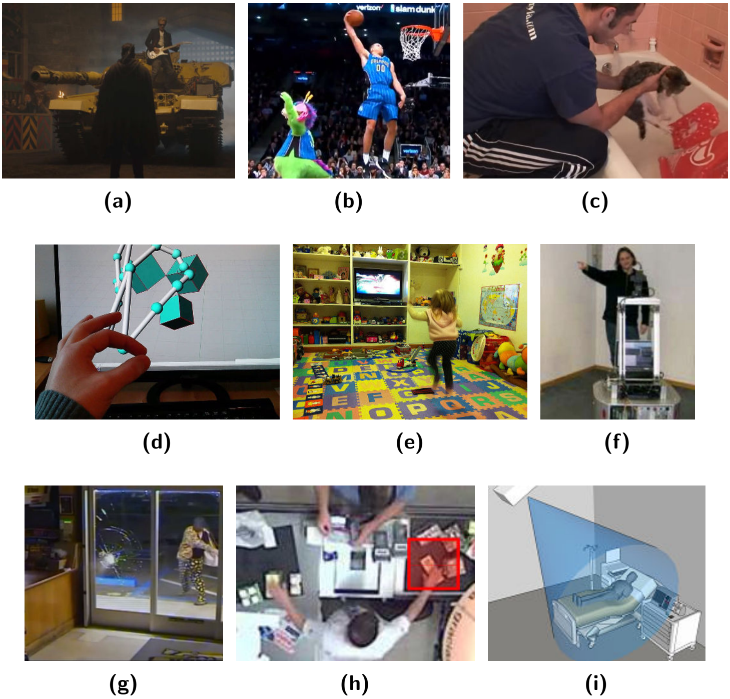
The application domain of video-based activity recognition that is most relevant to our work are however humanoid household robots, as shown in Fig. 1.5. Even though everybody having access to such service robots is still belonging to the realm of science-fiction, recent advances in Artificial Intelligence (AI), and mechanical technologies are promising rapid changes in the near future. In the first report of the One Hundred Year Study on Artificial Intelligence (AI100) released by the Standford University (*cf*., Stone et al. (2016)), a panel consisting of more than 20 world-renowned experts in AI has, among others, predicted how significantly service robots are going to influence our everyday life by the year 2030. Over the next fifteen years, the involved researchers expect an increasing focus on developing human-aware systems, that are specifically designed for the people they are meant to interact with.

Global Players, such as Amazon Robotics, Toyota, or Honda, as well as more than half a dozen startups around the world are currently developing robots for the home. Thus, it should not take much until the currently very slow growth in the diversity of robotic applications is going to start gaining momentum. Once the robots are deployed, cloud computing is going to enable sharing of data sets gathered at different homes to allow an incremental improvement of the systems. From the application centered standpoint, household robots can benefit from human activity understanding in many ways. Based on the way people are interacting with specific objects, the robot can learn their affordances, *i.e.*, intrinsic properties of objects allowing certain actions to be performed with it, while at the same time excluding others. Let's take an instance of the object `bottle` as an example; incorrectly assuming that it affords the action of `drinking`, instead of `greasing` can have fatal consequences.

Imitation learning (*cf.*, Billard et al. (2013)), also known as programming by demonstration, is another field where video-based activity recognition is applicable to robotics. The goal is to provide users without any technical background with a way to extend robot capabilities to novel situations. Just as human babies learn from adults, the robot should learn from the users, *e.g.*, that a jar of jam needs to be opened before its contents can be spread on bread.

As a final example of the many ways household robots can benefit from activity recognition, we want to name the understanding and anticipation of the human intentions. The robot should, for instance, decide what to do with a cup that is handed to him: placing it on a table, pouring some water in it, or putting it in the dishwasher. When considering the object alone, it is not possible to make any accurate decision. This changes however when also taking into account the activity that has been previously performed by the user. Likewise, even without any human intervention, the robot could offer pro-actively related services to the user based on the current situation.

To sum it up, whether it is content-based information retrieval of multimedia data, surveillance, home entertainment, ambient assisted living, or household robots, there exists a huge demand for automated video content analysis systems, especially ones implementing activity recognition. All of these examples also show that computer vision technologies have a high potential to influence all aspects of our everyday life in the near future.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 1.2. *Automatic action and activity recognition can be applied to many domains: (a)-(b) retrieval of specific events in movies and sports clips, (c) tagging and organizing of home videos, (d)-(f) gesture interpretation to enable an alternative communication channel with electronic devices (*e.g.*, computers, video game consoles, or service robots), (g)-(h) analysis of surveillance footage, or (i) patient monitoring in intensive care units.*

1.1 Background

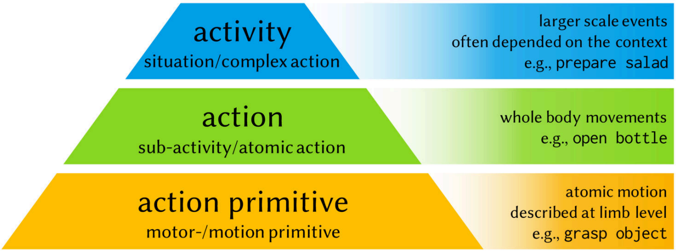


Figure 1.3. *Hierarchical motion event taxonomy introduced by Moeslund et al. (2006), that is used throughout this work. Entities at each level of complexity, consist of sequences of finer-grained events from lower levels.*

The goal of this work is to create a system for the automatic recognition of complex human activities from video material. Terms like actions, activities, or motion events are often used interchangeably in the literature.

There is, however, a general agreement that the taxonomies should be hierarchical, *i.e.*, entities at each level of complexity can be decomposed sequences of more fine-grained events found at lower levels. Nagel (1988) has for instance suggested using a hierarchy of change, event, verb, episode, and history to describe human motions at different semantic granularities. Another taxonomy has been proposed by Bobick (1997), where the terms movement (lowest level of complexity), activity, and action (highest level of complexity) are used. In this work, however, we are going to make use of the hierarchical taxonomy proposed by Moeslund et al. (2006) which is illustrated in Fig. 1.3. Human motions having the finest granularity are depicted as *action-primitives* which are atomic entities out of which actions are built, and usually describe movement at limb level. *Actions* describe in turn whole body movements which are often parts of *activities*. These are larger scale events that typically depend on the on the overall context, *e.g.*, manipulated objects, environment, of interacting humans.

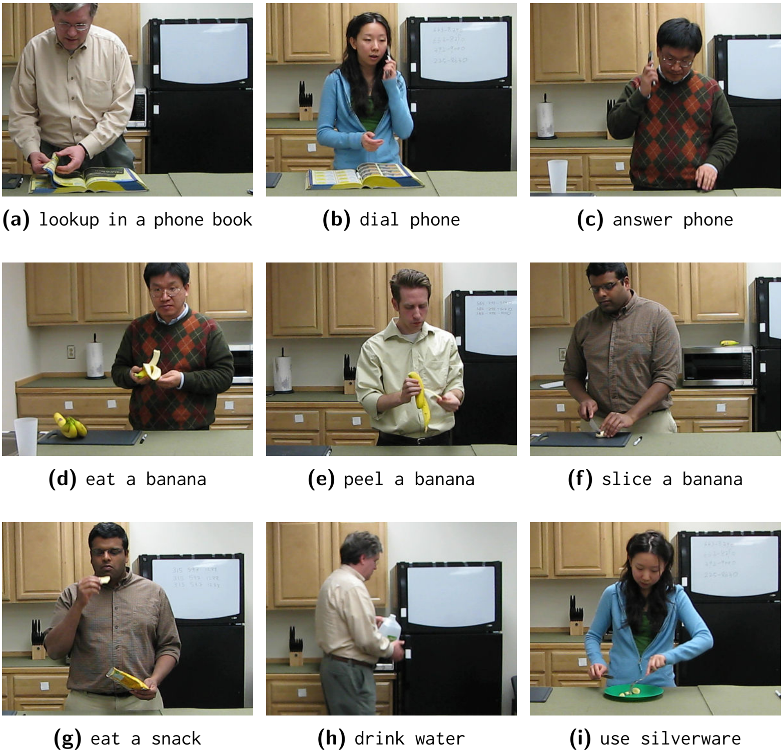
To further illustrate the differences between these three terms, let's consider the example given by Moeslund et al. (2006): playing tennis. Here, some exemplary action-primitives could be, *e.g.*, `forehand`, `backhand`, `run left`, and `run right`. Actions are sequences of action-primitives, *e.g.*, that are needed to `return a ball`. Note, that the action-primitives involved in an action depend on the overall context, *i.e.*, depending on the situation a `backhand`, `lob`, or `volley` may be required to return the ball. An activity in this example is then simply `playing tennis`; further example categories of activities can be found in Fig. 1.4.

The distinction between action-primitives, actions, and activities is however not always clear. For instance, the activity `sweep floor` might also be regarded as a periodic action due to its quite simple nature. Likewise, the action `jump` might as well be considered a action-primitive.

Finally, we would like to point out the importance of objects in the meaning of human motion events found at all levels of the hierarchy. While it is clear that many actions and activities exist that derive their meaning from object manipulations, it is interesting to note that even some action-primitives are intrinsically linked to objects as well.

Considered without any object information, the action-primitive `raising hand` is for instance not sufficient to be considered part of the action `drinking`, since it may as well belong to the actions of `smoking`, `eating`, or `cleaning mouth`. Instead, `raising bottle with hand` would be a more appropriate term to describe the action-primitive. Consequently, it is very important to also involve object knowledge, when developing motion, activity, and action recognition systems.

This observation that objects and actions are inseparably intertwined in cognitive processing was also one of the key ideas behind the PACO-PLUS research project (*cf.*, PACO-PLUS (2006)) funded by the European Commission in the years 2006-2010. As a universal representation of both concepts the project consortium has proposed the notion of object-action complexes (OACs, *cf.*, Geib et al. (2006) and Krueger et al. (2009)). Besides of allowing a formal description of object-action relationships, OACs also enable efficient planning and execution on robotic platforms. This makes OACs a powerful tool for high-level reasoning, which is, however, far beyond the scope of our work lying in sole activity classification from video data.



(a) `lookup in a phone book` **(b)** `dial phone` **(c)** `answer phone`

(d) `eat a banana` **(e)** `peel a banana` **(f)** `slice a banana`

(g) `eat a snack` **(h)** `drink water` **(i)** `use silverware`

Figure 1.4. *Example frames from nine (out of ten) activity categories of the URADL data set created by Messing et al. (2009). Note how motion, and object information complement each other when using both cues for activity recognition. For example, the classes* `eat snack` *and* `eat banana` *include similar motion patterns and thus can best be discriminated based on the manipulated objects.*

1.2 Problem Statement

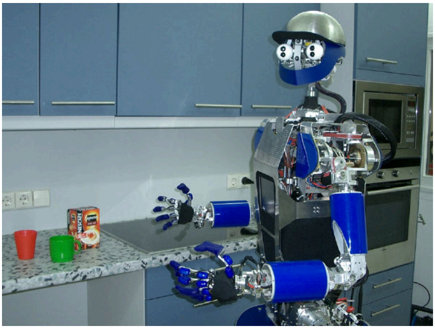


Figure 1.5. *The humanoid robot ARMAR III developed by Asfour et al. (2006) serves as the main target application of the activity recognition system developed in the context of this work. Goal of the proposed approach is to enable an implicit, non-verbal communication of a human with the robot through the recognition of activities based on the analysis of image sensor data (©2006 IEEE).*

This work has its roots in the Collaborative Research Center 588 - Humanoid Robots (*German*: Sonderforschungsbereich, SFB), of the German Research Foundation (*German*: Deutsche Forschungsgemeinschaft, DFG), which aimed at generating concepts for a humanoid robot that is able to share its activity space with a human partner. In this context, our application-driven goal was to enable an implicit, non-verbal communication channel between a human and a service robot based on the analysis of data from image sensors. As a typical application scenario, we envision the robot to take the role of a butler observing the scene from a point in the background and offering unsolicited help whenever he assesses that it might be required. Since gaining the understanding of what the observed person is doing is the best way to address this problem, the focus of this work lies in developing methods for automatic activity recognition from videos. Based on the observation that most complex activities involve some degree of human-object interaction, we put the emphasis of our work on exploiting object information to be used as a contextual cue for activity recognition. To illustrate this intuition, take a look at Fig. 1.4 showing frames selected from instances of nine activity categories of the popular URADL data set. In this example it can be clearly seen how motion and object information complement each other. For instance, it is very difficult for humans to discriminate between `eating` and `phoning` activities solely based on motion, since both mainly consist of the hand being moved in the direction of the face. However, when also considering the manipulated objects, the distinction between the classes is an easy task.

Our motivation for this research direction is further backed by the action identification theory from Vallacher and Wegner (1987), stating that actions (and thus activities) often derive their meaning from the context (*i.e.*, objects, or location). Most works in action and activity recognition, however, ignore contextual cues and focus on the identification of activities based on motion patterns alone. On the other hand, approaches that do incorporate object information usually depend on detectors that require supervised training. Since these detectors require a substantial amount of manually annotated training data, expanding such frameworks (*e.g.*, adding new action classes) becomes the bottleneck for generalized tasks. As an alternative, we investigate approaches exploiting context information without the need of additional supervised training, and propose proto-object features to serve this purpose.

The concept of proto-objects has been introduced as part the coherence theory of Rensink (2000b); Walther and Koch (2006), where they are defined as volatile units of visual information that may be validated as actual objects through focused attention. Basically, they serve the dual purpose of "being the highest-level output of low-level vision as well as the lowest-level operand on which high-level processes (such as object recognition or visual attention) can act" (*cf.*, Rensink (2000b)). In other words, proto-objects are object- or object-part candidates that have been detected, but not yet identified. This process allows us, humans, to selectively attend to an area of the visual field while ignoring the surrounding regions.

1.3 Main contributions and outline

The structure of this work is as follows:

Chapter 2: Related Work
The chapter begins with an overview of popular algorithms that can be applied to all areas of video content analysis. Next, we describe activity recognition approaches that are explicitly focusing on incorporating object information. To conclude this chapter, a discussion is given about the choices of algorithms we made to construct the motion recognition framework that is used as a baseline for the proposed object candidate features.

Chapter 3: Benchmark Data Sets
Benchmark data sets play a very important role when developing novel Computer Vision approaches. This chapter thus focuses on discussing publicly available action and activity recognition data sets, and selecting the most relevant ones to be used throughout this work. We further present the KIT Robo-kitchen activities data set (*cf*., Rybok et al. (2011)), which has been recorded as part of this work in order to capture the complexity of typical kitchen tasks and serves as a challenging benchmark for the evaluation of the proposed approach.

Chapter 4: Activity Recognition Framework
We have implemented several state-of-the-art local motion feature encoding methods to be used as a baseline, as well as in conjunction with the proposed object candidate features. In order to enable a fair comparison of this framework with the proposed approach, we evaluate it under different settings and select the strongest systems as a baseline for further experiments.

Chapter 5: Activity Recognition with Salient Object Candidates as Context
The presence or absence of certain objects often entails much information about the performed activities. However, most video-based activity recognition approaches either ignore such an important cue, or rely on supervised object detectors, which require much annotated

training data.
As an alternative, we introduce in this chapter a visual saliency based method to detect object candidate regions which we use as an additional cue for activity recognition (*cf*., Rybok et al. (2014)). Major advantages of these proto-object features are that they are fast to compute, do not require any additional manually annotated training data, and thus the approach can easily be applied to new domains. In an extensive experimental evaluation, we demonstrate the advantages of the proposed approach compared to pure motion-based algorithms, and also show its superiority to state-of-the-art.

Chapter 6: Conclusions
We conclude this work with a summary of the main contributions of our research in the field of activity recognition and outline possible directions for future projects.

2 Related work

In recent years, action and activity recognition from image sequences has been extensively studied in the computer vision community. Therefore, we are only focus our literature overview on the works that are most related to our method. For a complete overview of the field, including methods for action recognition from still images (*e.g.*, Delaitre et al. (2011); Desai et al. (2010)), one- and zero-shot learning (*e.g.*, Al-Halah et al. (2016); Xu et al. (2015)), view-independent methods (*e.g.*, Cai et al. (2014); Junejo et al. (2010)), unusual event detection (*e.g.*, Kratz and Nishino (2009); Zhong et al. (2004)), we refer the reader to the list of comprehensive surveys that we have compiled in Tab. 2.1. Since one contribution of this work comprises the creation of a data set for the evaluation of human activity recognition algorithms, we are going to give an in depth review over existing benchmarks in Chapter 3.

Typically, action and activity recognition systems consist of a two stage process: motion feature extraction, followed by a video-wide holistic representation of these features which is used for classification. We commence our overview of related work with motion feature encodings in Sec. 2.1, which we divide into *holistic*, *human body-model based*, *local feature based*, and *biologically inspired* methods.

Next, we discuss in Sec.2.2 structured (*e.g.*, Hidden Markov models, Conditional Random Fields), and unstructured (*e.g.*, Bag-of-Words, topic models) video representation methods that are commonly employed for action and activity recognition. Activity recognition ap-

Reference	Description
Aggarwal and Park (2004)	human body modeling; action recognition
Moeslund et al. (2006)	human motion capture; body-model based action recognition
Poppe (2007)	marker-less human motion analysis
Krüger et al. (2007)	action recognition; imitation learning
Turaga et al. (2008)	action and activity recognition
Morris and Trivedi (2008)	trajectory based methods for action recognition in surveillance
Aggarwal and Ryoo (2010)	action, activity, human-object interaction, and group activity recognition; data sets
Poppe (2010)	action recognition and detection; data sets
Weinland et al. (2011)	action recognition; view-independence; data sets
Vishwakarma and Agrawal (2013)	action recognition for surveillance
Borges et al. (2013)	human detection; action, activity, and interaction recognition; data sets
Hassner (2013)	data sets for action recognition
Chaquet et al. (2013)	data sets for action recognition
Ye et al. (2013)	3D action, activity, hand pose, gesture, and facial feature analysis; data sets
Aggarwal and Xia (2014)	action, activity, interaction, and gesture recognition from 3D; data sets
Guo and Lai (2014)	action recognition from still images; data sets
Liang and Zheng (2015)	action recognition from 3D; data sets
Cheng et al. (2015)	action, and activity recognition; data sets
Vrigkas et al. (2015)	uni- and multi-modal action, and activity recognition; data sets
Onofri et al. (2016)	action and activity recognition leveraging context information; data sets
Herath et al. (2016)	deep learning architectures for action recognition; data sets
Zhu et al. (2016)	hand-crafted local features and deep learning based methods for action recognition

Table 2.1. *Overview of surveys published in the past decade focusing on approaches aimed at different fields of human motion analysis.*

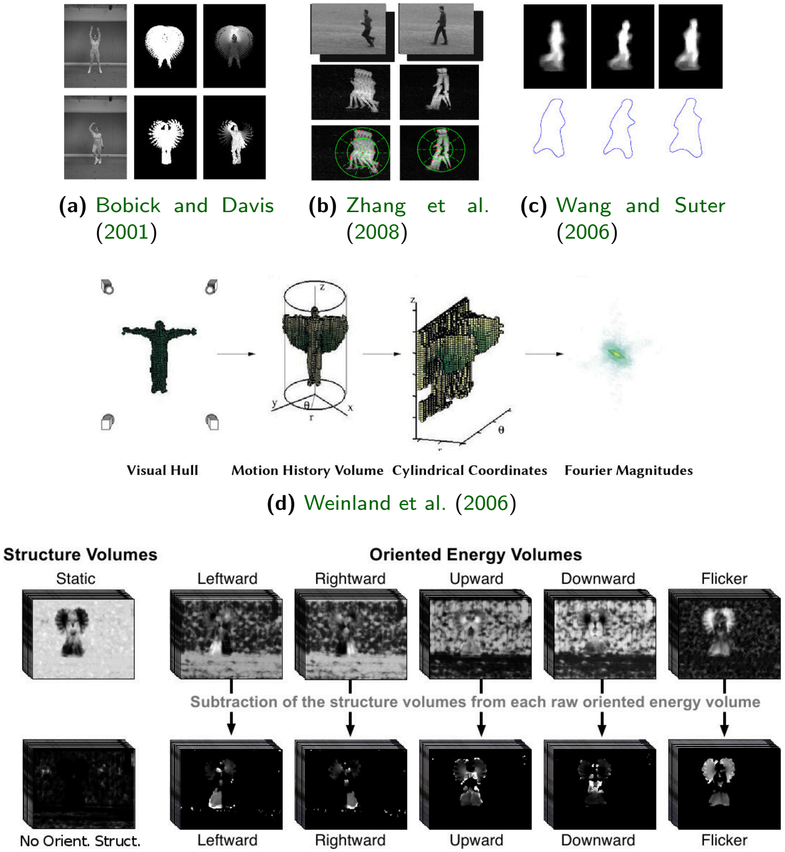
proaches laying the focus on explicitly involving object information are reviewed in Sec. 2.3. Finally, a discussion about our choices of algorithms employed in this work concludes this chapter in Sec. 2.4. Note that the chosen classification of the discussed approaches is not well defined. For example, the biologically inspired approaches discussed in Sec. 2.1.4 usually model an action as a whole and thus could be considered a member of the holistic approaches category. However, since they substantially differ in their concept from the discussed holistic methods, we decided to put them in a category of their own. The same applies to Bag-of-Words method aggregating local features for a global video description (see Sec. 2.2.1).

2.1 Motion representation

2.1.1. Holistic methods

Holistic (or global) methods model the observed actions as a whole, and thus do not require any information about body parts. Instead, only a global representation of the human body structure and motion is extracted directly from raw video sequences. Therefore, such methods are in general more robust and computationally efficient. This makes holistic methods especially interesting for real world applications, where body part localization is often difficult due to occlusions or background clutter.

Holistic action recognition methods can be divided into two major categories. The first one models motions in terms of the evolution of human silhouettes over time (*e.g.*, Bobick and Davis (2001); Sung et al. (2011); Wang and Suter (2006)). The silhouettes are obtained using either difference images, background-subtraction, or depth sensors. Methods belonging to the second class are mainly based on optical flow information, *e.g.*, Fathi and Mori (2008); Rodriguez et al. (2008); Schindler and van Gool (2008). This makes the feature calculation significantly slower, but more robust to self-occlusions. A detailed overview of methods belonging to each of the categories is given in the following.



(a) Bobick and Davis (2001)

(b) Zhang et al. (2008)

(c) Wang and Suter (2006)

(d) Weinland et al. (2006)

(e) Sadanand and Corso (2012)

Figure 2.1. *Examples of silhouette-based motion representations: (a) Motion Energy Images (center) and Motion History Images (right) (©2001 IEEE). (b) Motion Contexts (©2008 Springer). (c) Average Motion Energy (top) and Mean Motion Shape representations (bottom) of one action class (©2006 ACM). (d) Motion History Volume extending Motion History Images to 3D (©2006 Elsevier). (e) Spatio-temporal orientation energy features used in the action bank detectors (©2012 IEEE).*

Silhouette-based approaches

One of the first approaches that uses human silhouettes and its dynamics to represent actions has been proposed by Yamato et al. (1992). First, mesh features (*cf*., Umeda (1982)) are extracted from binarized images obtained through background subtraction. The features are then vector-quantized so that an action is represented by a sequence of discrete symbols, which is classified using HMMs.
Blank et al. (2005) stack silhouettes over an image sequence to obtain a 3D spatio-temporal volume, as depicted in Fig. 2.2(a). Then the solution of the Poisson equation is employed to derive local space-time saliency and orientation features. From this, global features are computed in the form of weighted moments.
Chen et al. (2007) encode the body as a star model of the extremities, which is obtained by fitting bounding a convex polygon to the silhouette. Actions are then modeled as sequences of the star figure's parameters which are represented as Gaussian mixture models. Example feature maps of samples belonging to one action category can be found in Fig. 2.2(b).
Weinland and Boyer (2011) base their approach on the idea of modeling actions by means of discriminative key-poses, as originally proposed by Carlsson and Sullivan (2001). Each action is thus represented as a set of key-pose exemplars which are directly mined from data through feature selection. Then body silhouettes extracted from each frame are matched against all exemplars belonging to one class and the resulting distances of best matches form the feature vector used for action classification.
One of the most popular uses of silhouette information for action recognition are the motion energy images (MEI) and motion history images (MHI) proposed by Bobick and Davis (2001). An MHI is generated by calculating the sum of all binary silhouette images weighted by a factor decaying back in time which makes recently moving pixels to have higher values. When thresholding an MHI above zero the MEI representation of the same image sequence can be obtained. Basically, MEI describe where an action occurs, and MHI how. Both motion representations have thus complementary

properties. An example can be found in Fig. 2.1(a): the MEI representations of different action categories look very similar, while both actions are easily distinguishable from each other based on the MHI. In their original approach, Bobick and Davis (2001) encode MHI and MEI with the seven Hu moments (*cf*., Hu (1962)) to achieve invariance to rotation, translation, and scale, and use them as temporal templates for action recognition. However, due to the simplicity of the MHI, as well as its low computational cost and descriptive power, various approaches have been subsequently adopted to improve it. Some of them are outlined in the following, while we refer the interested reader to the comprehensive survey from Ahad et al. (2010) for a broader overview.

Ahad et al. (2008) have identified the overwriting problem of the MHI, *i.e.*, self-occlusions that can occur when motions in opposing direction are performed within the time-frame of one MHI. As a solution, they propose to decompose the motion into four different directional components and describe each with a separate MHI.

Zhang et al. (2008) introduce the Motion Context (MC) descriptor, which can be seen as a hybrid of MHI and Shape Contexts (Belongie et al. (2000)) and offers a much richer descriptive capability than image moments. First, the action sequence is divided into non-overlapping groups of frames, which are each converted into an MHI variant. Each of those sets of subsequent frames is encoded as an MC and the sum of all MC descriptors is taken to represent a human action.

The Motion Context descriptor itself is a log-polar histogram of the pixel values of a single MHI (see Fig. 2.1(b) for an example). As the reference point for the histogram, the geometric center of the motion is taken.

The average motion energy (AME) and mean motion shape (MMS) descriptors proposed by Wang and Suter (2006) are other variants of the MHI. As the name suggests, AME descriptors are calculated as the mean of all binary silhouette images involving the human motions. In as similar fashion, MMS describe the silhouettes' mean shape, *i.e.*, its outer boundary. Instead of simple thresholding as in the case of MEI, Procrustes shape analysis (*cf*., Goodal (1991)) is

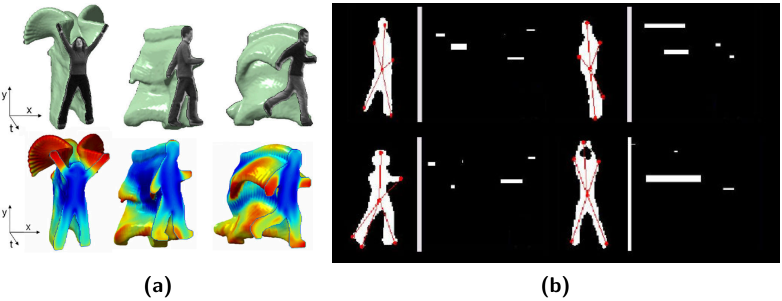


Figure 2.2. *(a) Space-time shapes for three different actions (top), and local space-time saliency features (bottom) as proposed by Blank et al. (2005). Feature values are encoded by the color spectrum (©2005 IEEE). (b) Star-figure silhouette representation, and the corresponding 2D template encoding of different actions as proposed by Chen et al. (2007) (©2007 IEEE).*

employed to obtain a descriptor that is invariant to translation, rotation, and scaling. Example images of both motion representations can be found in Fig. 2.1(c).

Motivated by the performance of the Object Bank framework (*cf*., Li et al. (2010a)) for object recognition, Sadanand and Corso (2012) propose the Action Bank representation. An Action Bank consists of a set of template-matching based action detectors trained on classes broadly sampled in semantic and viewpoint space. The detectors are used to extract mid-level features for the recognition of novel (*i.e.*, unseen) action classes by stacking their responses in one single vector. The features used in the Action Bank detectors itself are derived from seven raw spatio-temporal energy volumes, which are computed by applying 3D third derivate Gaussian filters to the input image volume. As depicted in Fig. 2.1(e), five pure energy volumes are then calculated by subtracting the structure volumes from raw volumes. Even though being conceptually simple, the semantic, mid-level Action Bank has been demonstrated to yield a performance superior to most, more complex local feature methods (*cf*., Sec. 2.1.3), and biologically inspired systems (*cf.*, Sec. 2.1.4) on realistic videos. Overall, silhouette-based approaches have proven to be quite successful for action recognition, especially due to their low computational costs. A major drawback are, however, self-occlusions, *i.e.*, when action relevant motion is performed in front of the body and thus ignored. This is especially harmful in the case of hands which play a very important role in most human actions. To combat these shortcomings, methods relying on other global information have been explored as well. Some of the most popular types are based on edge representation of the body (*e.g.*, Carlsson and Sullivan (2001)), depth data (*e.g.*, Li et al. (2010b); Oreifej and Liu (2013); Weinland et al. (2006)), or optical flow (*e.g.*, Ali and Shah (2010); Fathi and Mori (2008); Rodriguez et al. (2008)).

Depth volume based approaches

Depth volume based methods extend silhouette-based approaches to 3D. For instance, Weinland et al. (2006) propose with 3D Motion History Volumes (MHV) by replacing the pixels in the MHI calculation with voxels (see Fig. 2.1(d)). The voxels itself are obtained from the visual hull, which is the 3D counterpart of silhouettes (*cf.*, Laurentini (1994)). Alignment and comparison of MHV templates is achieved by using Fourier transform in cylindrical coordinates around the vertical axis.

With the advent of low-cost consumer electronic depth sensors, such as Microsoft Kinect, holistic space-time volume approaches based on depth maps have also been widely explored (*cf.*, Liang and Zheng (2015); Ye et al. (2013)). For instance, Yang et al. (2012) project the depth maps from the human body to three orthogonal planes and compute from each view a MEI. Then each MEI is encoded as a Histogram of Oriented Gradients and resulting feature vectors are concatenated and used for action recognition. Li et al. (2010b) also project the depth maps to the three orthogonal planes, however only sample the contour points of the resulting 2D silhouettes. Next, the silhouettes are modeled as bags of points by fitting Gaussian Mixtures on the contour points. These are finally used to represent

salient postures which correspond to nodes of an action graph (*cf*., Li et al. (2008)) capturing the dynamics of an action.
Instead of relying on 2D projections, Oreifej and Liu (2013) proposed a descriptor that captures motion and appearance in 4D spatio-temporal space. For the Histogram of Oriented 4D Normals (HON4D) features, a depth map sequence is treated as a 4D spatio-temporal shape from which a histogram of normal vectors is constructed. Since uniform quantization as usually employed to build histograms is often far from being optimal in 4D space, they also propose a non-uniform quantization technique.

Optical flow based approaches

Even though holistic, depth volume based approaches have proven to be much more robust compared to methods employing 2D body silhouettes, they have the disadvantage of relying on depth sensors. However, the majority of available videos (*e.g.*, from YouTube, TV channel archives) has been captured with conventional cameras making 2D methods that are robust to self-occlusions still necessary. Optical flow based approaches have this property, yet with the downside of a much higher computational cost due to flow estimation.
One of the first holistic action recognition methods exploiting optical flow of the observed human has been proposed by Polana and Nelson (1997). First, humans are being tracked in the scene, before an action representation is extracted using optical flow magnitude in a grid pattern centered on the tracked person. Next, a periodicity index is constructed and once the observed sequence is sufficiently periodic, it is segmented into individual cycles which are matched to other periodical actions.
Another early approach has been proposed by Efros et al. (2003) and encodes motion within a tracked rectangle in terms of the four directions of blurred optical flow. In order to classify an action, the sequence is frame-wise aligned to training data and the label of the sequence with the highest alignment score is taken. Fathi and Mori (2008) employ the same feature descriptor, however in conjunction with a more sophisticated classification method. The

spatio-temporal volume centered on the person is divided into a set of non-overlapping cuboids. Within each of them, the low-level optical flow features are used to train weak classifiers which are combined via an AdaBoost variant (*cf*., Schapire and Singer (1999)) to form informative mid-level features. These serve in turn as weak classifiers for a second, global AdaBoost layer. Multi-class action recognition is finally obtained by combining the binary classifiers using Hamming decoding (*cf*., Dietterich and Bakiri (1995)).

The action MACH (Maximum Average Correlation Height) framework from Rodriguez et al. (2008) is based on template-matching and extends MACH filters (*cf*., Hennings-Yeomans et al. (2007)) from spatial 2D to spatio-temporal 3D. MACH filters combine all instances of one class in one template which is matched in the frequency domain via a fast Fourier transform (FFT). Instead of using raw pixels within the spatio-temporal volume, the authors propose to use dense optical flow fields in order to better capture motion information. This leads, however, to vector valued data making it necessary to incorporate a generalized Fourier transform (*cf*., Ebling and Scheuermann (2005)) in the framework since FFT only operates on scalar values.

Ali and Shah (2010) proposed to describe motions with a set of kinematic features derived from optical flow. Example feature types are: vorticity measuring the local spin around the axis perpendicular to the plane of the flow field, and symmetric fields capturing the dynamics emphasizing the symmetry (or asymmetry) of a person around a diagonal axis. From each of the feature types, kinematic modes capturing representative dynamics of the motion are computed using Principal Component Analysis. These are finally used as action representation in a Multiple Instance Learning framework (*cf*., Chen et al. (2006)).

2.1.2. Human body model based methods

Johansson's psychophysical experiments with Moving Light Displays (*cf*., Johansson (1973)) have inspired many approaches in action recognition to use a similar motion representation of the human

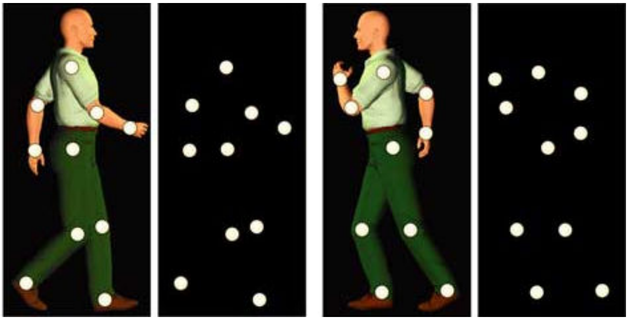


Figure 2.3. *Johansson (1973) has shown that humans can recognize actions from motions of a few light displays attached to an actor's body, but fail to realize any connection between the lights and a human body when no motion is perceived (reprinted from Giese and Poggio (2003), ©2003 Nature).*

body (see Fig. 2.3). For the experiments, bright light displays were attached to the main joints of an actor dressed in black and standing in front of a black background.

As long as the actor stood still, the lights bore no information to the observers in the sense that they could not even realize any connection between the static light displays and a human body. This changed however when the actor started to move. Not only allowed this the observers to recognize that the lights were actually attached to a human body, but also name the performed action, and even the actor's gender, as revealed in the study of Barclay et al. (1978). Overall, the interpretation of these experiments has led to two classes of methods for human motion interpretation based on a body model: *direct recognition* from motion in 2D, and *recognition by reconstruction* of the 3D body model (*cf*., Weinland et al. (2011)).

Direct action recognition approaches operate on anatomical landmarks or 2D body representations and can thus be applied to any image sequence (*e.g.*, Ali et al. (2007); Lv and Nevatia (2006)). Their main drawback is however that they usually are not invariant to the camera position relative to the filmed actor. On the other hand, view invariance can easily be achieved when operating on 3D body models, as in the recognition by reconstruction methods (*e.g.*, Ofli et al. (2014); Rohrbach et al. (2012b); Zhu et al. (2013)). Such approaches usually consist of two stages. First, the 3D body model needs to be estimated, and then actions can be recognized based on a body representation.

Localization and tracking of human body parts in 3D is a very challenging task and thus has attracted many computer vision researchers (*cf.*, the comprehensive survey by Moeslund et al. (2006)). Originally, human motion capture approaches estimated depth images with expensive multi-camera systems, or time-of-flight cameras. Furthermore, these algorithms usually have a very high computational complexity, all making action recognition by reconstruction very cumbersome. This has changed with the introduction of the Microsoft Kinect and similar low-cost consumer electronics depth sensors. Not only does the Kinect provide 3D depth data of a scene, but it also allows a fast and accurate estimation of the 3D position of skeletal joints using the method from Shotton et al. (2011). An overview of the approach including an example of a reconstructed body model from Kinect data can be found in Fig. 2.4.

These recent advances have lead to a renewed interest in human model based action recognition. Overall, one can distinguish three major classes of features used for model-based action recognition, which will be discussed in the following: features based on the joint location, relations between joints, and joint angles.

Joint location based features

One of the most straightforward skeleton representations for action recognition is the location of the joints. In order to achieve invariance to body size, location, and orientation, as well as camera position, the joint coordinates are usually first normalized. Viewpoint invariance can be addressed by centering the reference coordinate system for the joint locations on the subject and rotating it together with the body orientation (*cf.*, Xia et al. (2012)). Anthropometric differences between the human subjects can be for instance achieved by adjust-

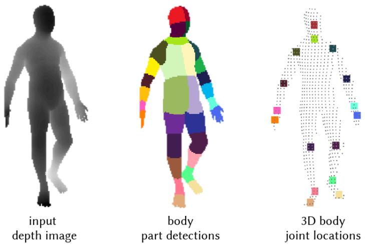
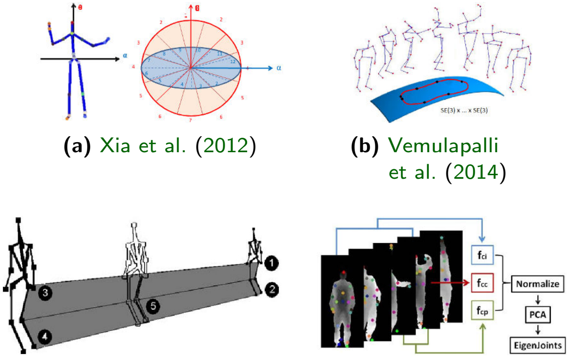


Figure 2.4. *Overview of the approach from Shotton et al. (2011) that is commonly used to reconstruct 3D human body models from data streams captured with Microsoft Kinect sensors: Each depth image is mapped to a per-pixel distribution of body parts through a randomized decision forest. These pixel labels are then used to infer the 3D body part locations (©2011 IEEE).*

ing the distances between connected body joints to match average segment lengths learned from training data, as has been done by Zanfir et al. (2013), and Seidenari et al. (2013).

Lv and Nevatia (2006) use similarly normalized 3D coordinates of different sets of joints together with HMMs as weak features for AdaBoost. Parameswaran and Chellappa (2003) achieve view and appearance invariance by projecting the 3D body joint locations to a 2D invariance space and model actions in terms of canonical body poses and 2D trajectories (see 2.5(c)). The approach by Vemulapalli et al. (2014) also maps the 3D skeleton to a different space and represents it as a point in a Lie group, which is a curved manifold in which actions can be modeled as curves (see 2.5(b)). Action recognition is then performed using a combination of dynamic time warping, Fourier temporal pyramids, and SVM classification.

Ali et al. (2007) use trajectories of selected landmarks on the body and represent body motion based on chaotic invariants. The approach by Bargi et al. (2012) allows a joint segmentation and classifications of actions and is based on an online hierarchical Dirichlet process HMM (HDP-HMM) and 3D joint positions expressed in a subject centered coordinate system as features. Likewise, Xia et al. (2012) also employ an HMM together with features based on subject centered



(a) Xia et al. (2012) **(b)** Vemulapalli et al. (2014)

(c) Parameswaran and Chellappa (2003) **(d)** Yang and Tian (2013)

Figure 2.5. *Examples of body model based motion representations: (a) Reference coordinates and spherical histogram of the HOJ3D features (©2012 IEEE). (b) Representation of an action as a curve in a Lie group (©2014 IEEE). (c) Geometrical invariants computed from five points lying on a plane (©2003 IEEE). (d) EigenJoint features (©2013 IEEE).*

joint locations. The skeleton is modeled with a spherical Histogram of joint locations (HOJ3D), centered on the hip center (see 2.5(a)). To make the descriptor scale-invariant, the radial distance of the joints is being discarded. The histogram is further compressed using Fisher's linear discriminant analysis, and vector quantized into prototypical postures that are used as features for a discrete HMM. Instead of relying on the classifier to achieve temporal modeling, Hussein et al. (2013) encode motion information directly in the body features. Their Covariance of 3D Joints (Cov3DJ) descriptor is based on the covariance matrix of the joint trajectories. Inspired by the idea of spatial pyramid matching (*cf*., Lazebnik et al. (2006)), long-term dynamics are further captured with a temporal pyramid.

The Moving Pose descriptor introduced by Zanfir et al. (2013) captured both pose and dynamics of the skeleton joints during an action. It consists of the normalized locations of each joint together with its velocity and acceleration. Each single-frame pose feature then votes for an action using a modified k-nearest neighbors class density estimator. Seidenari et al. (2013) go a slightly different way than the aforementioned approaches and proposed to encode the skeleton based on kinematic chains. Root of each chain is the torso, and the position of each joint is expressed relative to its parent joint. The frame descriptors are then directly used to compute a Video-to-Class distance in an extended Naïve Bayes nearest neighbor framework.

Pairwise joint relationship based features

Wang et al. (2012a) have demonstrated that using the pairwise relative positions between joints instead of the 3D joint coordinates results in more discriminative features. However, since considering all joint pairs leads to a redundant representation, Luo et al. (2013) use only the hip center as a reference point. Video sequences are then encoded as a Bag-of-Words using Sparse Coding together with a linear SVM for classification and it is shown that this approach outperforms the more complicated method proposed by Wang et al. (2012a).

Yang and Tian (2013) go in a different direction and propose to describe the skeleton with an even richer feature set than Wang et al. (2012a). Not only do they use the pairwise relative positions in the current frame c, but also between c and the previous frame, as well as between c and the initial frame, assuming that it approximates the neutral pose. This leads to a 2970-dimensional feature vector, which however contains a lot of redundancy. Therefore, Principal Component Analysis (PCA) is further employed to reduce the feature dimensionality resulting in the EignenJoints representation of the pose for each frame (see 2.5(d)). EigenJoints-like features have further been shown to yield very good action recognition performance when used with HDP-HMMs (Raman and Maybank (2015)), and Deep Belief Network HMMs (Wu et al. (2014)). A major disadvantage of body model based approaches is that sometimes the joints can be incorrectly detected or even completely lost, which dramatically affects action recognition accuracy. To overcome this drawback, Zhu

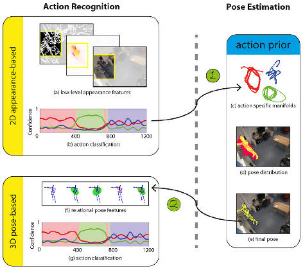


Figure 2.6. *Overview of the coupled action recognition and pose estimation approach from Yao et al. (2012): First actions are recognized based on low-level appearance features. The estimated actions are then used as prior distributions for the particle-based optimization of the 3D pose estimation system. Finally, relational features are extracted and used for action recognition (©2012 Springer).*

et al. (2013) proposed to fuse EigenJoints with local features (see Sec. 2.1.3) and showed experimentally that both feature types have complementary properties.
The coupled action recognition and pose estimation approach by Yao et al. (2012) goes one step further than all previously mentioned work (see Fig. 2.6). Knowing the performed action greatly simplifies the problem of reconstructing the body pose. It allows mapping the high-dimensional pose state-space to low-dimensional action specific manifolds, which are learned from motion capture data. Thus, first action recognition based on low-level appearance features is used as a prior to improve 3D pose estimation. From the pose, a set of relational features is then calculated, which in turn are used to improve action recognition. These features describe relations between pairs of joints, joints and a plane spanned by other joints, as well as the velocity of joints.

Joint angle based features

Instead of normalizing the joint locations to gain view- and anthropometric invariance, the same can be achieved by deriving body pose descriptors from joint angles. For instance, Ben-Arie et al. (2002) store a set of joint angles and angular velocity vectors of the major body parts in multidimensional hash tables. Actions are then recognized by indexing and sequencing a few pose feature vectors in the hash tables.

Gehrig and Schultz (2008) use joint angles obtained from a 3D marker-based motion capturing system as features for an HMM and increase action recognition robustness by means of feature selection. The sequence of most informative joints (SMIJ) introduced by Ofli et al. (2014) goes in a similar direction. At each time-step, a few joints which are assumed to be the most informative to infer the performed action are automatically selected by importance ranking based on entropy. The action is then partitioned into a set of temporal segments, each of which is represented as a time series of the aforementioned joints angles. The time series are then used for action recognition in conjunction with an SVM or Nearest-Neighbor classifier using a normalized edit distance as a similarity metric.

Sequences of joint angles are also used by Ohn-Bar and Trivedi (2013) to represent actions. In order to make the final feature vector be of a fixed dimension, as well as enrich its information, the time series data is further converted into a square matrix of similarities between all the sequences of joint angle values.

2.1.3. Local feature methods

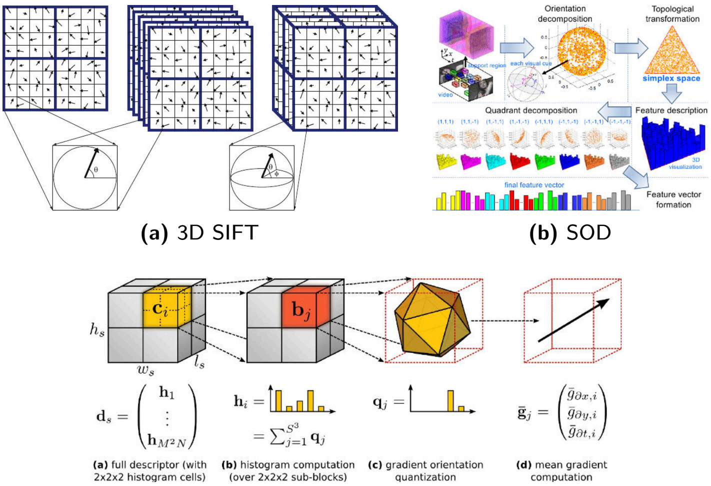
Local features (or interest points) are image patterns which strongly differ from their immediate neighborhood while being rich in information (*cf.*, Tuytelaars and Mikolajczyk (2007)). Such features are extracted from raw data in a two-stage process: detection, and descriptor calculation, *i.e.*, extraction of discriminative features from patches around interest points. As opposed to global approaches, which encode an action as a whole, local feature based methods describe the observation as a collection of local patches. During the past decade, such approaches have become incredibly popular in many fields of Computer Vision outperforming all other methods most of the time. Example applications are object-, and scene recognition (*e.g.*, Lazebnik et al. (2006); Lowe (1999); Mikolajczyk et al. (2005)), articulated pose estimation (*e.g.*, Andriluka et al. (2009)), object tracking (*e.g.*, Zhou et al. (2009)), video data mining (*e.g.*, Sivic and Zisserman (2003)), and wide baseline matching (*e.g.*, Tuytelaars and van Gool (2004)).

Motivated by the wide success of these approaches in the image domain, researchers have generalized 2D local features to the 3D spatio-temporal video domain. Their prime application was action recognition resulting in a huge increase in performance compared to holistic and body model based methods. Consequently, local features have been dominating the field of action recognition since their inception. Many space-time interest point (STIP) detectors and descriptors have been proposed which we are going to briefly discuss in the following. We refer the interested reader to Tab. 2.2 for an overview of works comparing different combinations of STIP detectors and descriptors on several benchmarks.

Feature Detectors

Space-time interest point (STIP) detectors usually select characteristic image volumes based on specific saliency criteria. They have been originally introduced by Laptev and Lindeberg (2003) as a spatio-temporal extension of the Harris-Laplace detector (*cf.*, Mikolajczyk and Schmid (2002)). A visualization of these Harris3D features can be found in Fig. 4.1. Note, that in the literature the terms Harris3D and STIP are used interchangeably, even though the latter encompasses the whole class of local 3D feature detectors.

In the following years, many other 2D feature detectors have been extended to spatio-temporal 3D. For instance, Oikonomopoulos et al. (2005) introduced a spatio-temporal version of the entropy-based saliency measure from Kadir and Brady (2003). An extension of the popular SIFT detector originally developed by Lowe (1999) has been proposed by Cheung and Hannarneh (2007). Willems et al. (2008) presented a spatio-temporal generalization of the saliency measure for blob detection from Beaudet (1978), which is based on the determinant of the Hessian.


θ
θ
ϕ
Orientation decomposition
Topological transformation
support region
each visual cue
video
simplex space
Quadrant decomposition
Feature description
3D visualization
final feature vector
Feature vector formation
h_s
w_s
l_s
$\mathbf{c}_i$
$\mathbf{b}_j$
$\mathbf{d}_s = \begin{pmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_{M^2N} \end{pmatrix}$
$\mathbf{h}_i =$
$= \sum_{j=1}^{S^3} \mathbf{q}_j$
$\mathbf{q}_j =$
$\bar{\mathbf{g}}_j = \begin{pmatrix} \bar{g}_{\partial x,i} \\ \bar{g}_{\partial y,i} \\ \bar{g}_{\partial t,i} \end{pmatrix}$
(a) full descriptor (with 2x2x2 histogram cells)
(b) histogram computation (over 2x2x2 sub-blocks)
(c) gradient orientation quantization
(d) mean gradient computation

(a) 3D SIFT

(b) SOD

(c) HOG3D

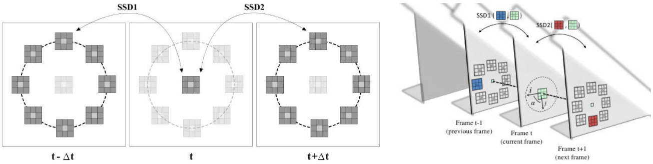
Figure 2.7. *Illustrations of different approaches generalizing the SIFT descriptor to 3D: (a) The 3D SIFT descriptor developed by Scovanner et al. (2007) (©2007 ACM). (b) Processing pipeline of the simplex-based orientation (SOD) descriptor from Zhang et al. (2014) (©2014 IEEE). (c) The HOG3D descriptor from Kläser et al. (2008) in which the 3D orientation is quantized based on regular polyhedrons.*

Dollár et al. (2006) argue that 3D counterparts to 2D interest point detectors are often inadequate for action recognition since they lead to very sparse results ignoring much informative motion. As a solution, they propose the Cuboid detector which treats spatial and temporal information separately. To this end, spatial 2D Gaussian kernels and temporal 1D Gabor filters are applied to the video data, and as with Harris3D, the local maxima of the filter responses are taken as interest points. Bregonzio et al. (2009) have identified several drawbacks of the Cuboid detector, such as false detections due to background noise, and its insensitivity to purely translational motion. To overcome these shortcomings, they proposed a two-stage process. First, frame differencing is applied in order to focus the attention of the detector to regions solely involving motion. Next, they apply 2D Gabor filters of several orientations to the regions of interest obtained from the first step.

STIP detectors are usually only applied to single channel grayscale data. It has however been shown in the 2D image domain, that using color information yields higher quality detections (*cf*., Burghouts and Geusebroek (2009); Gevers and Snoek (2010)). Everts et al. (2013) argue that using color information can make STIP detectors less sensitive to disturbing illumination conditions (*e.g*., shadows) while increasing their discriminative power. Therefore, they reformulate the Harris3D and Cuboids detectors to incorporate multiple photometric channels and reported a substantial increase in action recognition performance.

Usually, STIP detectors only use local information within a small region to determine salient points. This makes them however very susceptible to video noise, which is why Wong and Cipolla (2007) proposed an approach considering global information for local feature detection. To achieve this, an image sequence is decomposed into spatial and temporal components via non-negative matrix factorization (NNMF). From this, interest points are located using 2D spatial, and 1D temporal SIFT detectors.

Wang et al. (2009) observe that in the context of object recognition, dense sampling image patches often yields a superior performance to sparse interest points (*cf*., Jurie and Triggs (2005)). Therefore, they include this method in their large-scale evaluation of STIP detectors and descriptors for action recognition. Surprisingly, dense sampling at regular locations in space and time proved as well to outperform all tested STIP detectors on actions captured in a realistic setup. In follow-up works (Wang and Schmid (2013); Wang et al. (2011a)), they elaborate a more efficient way to extract spatio-temporal volumes for local feature approaches. Instead of sampling the patches on a regular grid in 3D space, Wang et al. (2011a) extract dense trajectories (DT) from video and use the spatio-temporal tubes in their neighborhood for STIP description. Wang and Schmid (2013) present an improved version of the DT detector by adding a pre-processing



(a) LTP **(b)** MIP

Figure 2.8. *Illustrations of two spatio-temporal extensions of the LBP texture descriptor: (a) The local trinary pattern (LTP) descriptor from Yeffet and Wolf (2009) considers only patch-wise SSD distances at the same spatial locations (©2009 IEEE). (b) The motion interchange pattern descriptor developed by Kliper-Gross et al. (2012b) generalizes LTP by also taking the SSD distances at different spatial locations into account. This results in a much richer but also higher dimensional descriptor than LTP (©2012 Springer).*

step to compensate camera motion. This is achieved by estimation the homography between consecutive frames based on local feature matching and RANSAC (*cf*., Szeliski (2006)), and filtering out human motion with a person detector. Since camera motion is often present in realistic videos, a substantial increase of the already good action recognition performance of DT has been achieved. Because of this, as well as the implementation being publicly available, iDT features have been employed in many approaches, each pushing the state-of-the-art forward (*e.g.*, Lan et al. (2015b); Sun et al. (2016); Tran et al. (2015); Wang et al. (2015a)).

Feature Descriptors

Once local features have been detected, it is necessary to establish a representation of its (spatio-temporal) neighborhood so that it can be matched with features extracted from other data samples. The simplest type of descriptor is a flattened vector of the raw pixel values within the interest point area. Its very high dimensionality would, however, result in a very high computational complexity for recognition. Such a descriptor would also lack other desirable traits, such as robustness to illumination and camera viewpoint changes, or invariance to local shape distortions. Ke and Sukthankar (2004) have found that all aforementioned properties can be achieved by simply applying PCA dimensionality reduction to the intensity data. A generalization of this PCA-SIFT descriptor to space-time volumes has been introduced by Dollár et al. (2006). Instead of only using (normalized) pixel-values, their Cuboid descriptor also captures motion information in form of optical flow and brightness gradients calculated at each spatio-temporal location inside the cuboid region. In object recognition, image descriptors based on histograms of oriented gradients (HOG) are particularly successful (*e.g.*, Dalal and Triggs (2005); Lowe (1999)). It is therefore not surprising that this concept has been as well applied to STIP descriptors.

Laptev and Pérez (2007) have extended the HOG descriptor from Dalal and Triggs (2005), by accumulating the 2D gradients within the space-time cuboids to histograms. However, 2D image gradients are not discriminative enough to represent motion. Therefore, they proposed to use HOG jointly with histograms of optical flow (HOF) giving rise to the popular HOGHOF descriptor.

Wang et al. (2011a) observe that the HOF descriptor is very prone to noise caused by camera motion since it is based on absolute motion. Therefore they propose to use, instead, the motion boundary histogram (MBH), that has been originally developed by Dalal et al. (2006) for human detection. It separates the optical flow field in its horizontal and vertical components, computes spatial derivatives for each of them, and finally quantizes the resulting orientation information into histograms.

Jain et al. (2013) argue that MBH only handles camera motion implicitly. As a better alternative, they propose to separate the optical flow into dominant and residual motion. Dominant motion is assumed to be caused by camera motion and thus discarded. Residual motion is however retained for action recognition, as it corresponds to motion happening in the individual scene. Additionally, a novel motion descriptor is introduced, the Divergence-Curl-Shear (DCS) descriptor. It is based on kinematic features, *i.e.*, first-order differential scalar quantities computed on the flow field, and captures more motion information than MBH.

All of these descriptors rely on optical flow estimation, which is very time-consuming due to its computational complexity. Kantorov and Laptev (2014) observe that video compression algorithms (*e.g.*, MPEG) heavily rely on motion estimation. Thus, they propose to approximate optical flow directly from the compressed video representation significantly speeding up descriptor computation at the cost of a small loss in action recognition performance.

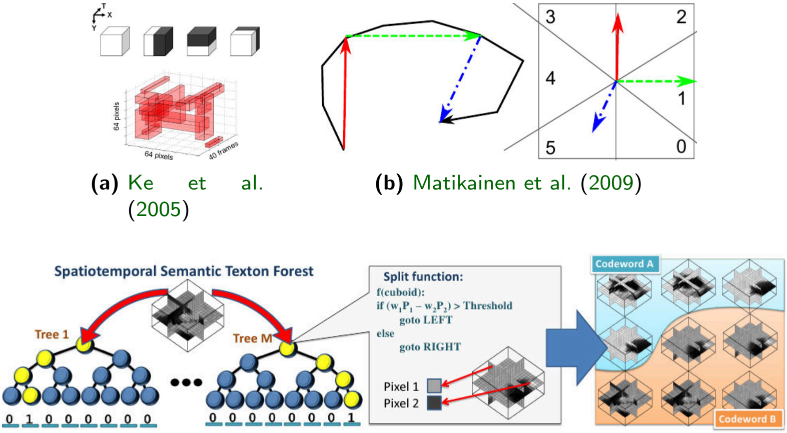
Several approaches to extend the SIFT image descriptor from Lowe (1999) to videos have been developed as well, all based on 3D gradients. Illustrations of the three most prominent spatio-temporal SIFT generalizations can be found in Fig. 2.7, all of which are be briefly described in the following.

Scovanner et al. (2007) proposed the 3D SIFT descriptor, for which gradients at randomly sampled positions vote into a 3D grid of histograms inside each STIP cuboid. To quantize the orientation, the gradients are represented in spherical coordinates.

Kläser et al. (2008) identify that such a quantization method leads to singularities at the poles since bins get significantly larger, the closer they get to the equator. As a solution, they propose a quantization scheme based on regular polyhedrons. This method suffers, however, of a limited discrimination power since only five regular polyhedrons exist resulting in a support of maximum 20 bins. Therefore, Zhang et al. (2014) propose to quantize and describe the gradients in the simplex topological vector space. To further increase the discriminative power of this simplex-based orientation decomposition (SOD) descriptor, a quadrant decomposition is additionally performed.

Shi et al. (2015) argue that gradient-based descriptors, such as HOG3D and 3D SIFT, suffer from a high dimensionality caused by 3D gradient computation. As an alternative, they propose the gradient boundary histogram (GBH) descriptor. It is based on time-derivatives of image gradients and thus emphasizes moving edge boundaries.

In 2D image analysis self-similarity based texture descriptors, like the local binary patterns (LBP) proposed by Ojala et al. (2002) are a



(a) Ke et al. (2005)

(b) Matikainen et al. (2009)

(c) Yu et al. (2010)

Figure 2.9. *Illustrations of several local space-time feature descriptors: (a) Extension of Haar-like features to capture motion information (©2005 IEEE). (b) Trajectory segment orientation histogram features (©2009 IEEE). (c) Spatio-temporal generalization of semantic texton forests.*

popular alternative to gradient-based encodings. LBP encodes each pixel with a binary code which is obtained by thresholding a neighborhood of pixels with the gray value of the center pixel. An image texture can then be described as a histogram of the LBP binary codes. Zhao and Pietikäinen (2007) have originally employed this concept to describe dynamic textures in videos for facial expression analysis. For their LBP-TOP descriptor, three histograms of LBP codes are concatenated, each computed from one orthogonal plane inside a spatio-temporal volume. The performance of LBP-TOP for the task of action recognition has been evaluated by Kellokumpu et al. (2011); Shao and Mattivi (2010).

The local trinary pattern (LTP) descriptor developed by Yeffet and Wolf (2009) is another generalization of LBP (see Fig. 2.8(a)). As the name suggests, each pixel is now encoded with an 8 trit value, which is computed from its 3D neighborhood in a similar fashion like the LBP code. Kliper-Gross et al. (2012b) further generalized LTP with the motion interchange pattern (MIP) descriptor, which considers more comparisons than LTP in each pixels neighborhood, as depicted in Fig. 2.8(b). This allows the descriptor to better encode motion direction at the cost of increasing the encoding length of one pixel by a factor of eight.

Yet another popular image descriptor type that has found its way to video analysis are the Haar-like features (*cf*., Viola and Jones (2004)). As with LBP, 3D extensions of it have been explored in form of volumetric features (*e.g*., Ke et al. (2005)) and three orthogonal planes (*e.g*., the eSURF descriptor from Willems et al. (2008)). Semantic texton forest descriptors (*cf*., Shotton et al. (2008)), which are the core of the human body pose reconstruction algorithm employed by the Microsoft Kinect, have also been generalized to spatio-temporal data by Yu et al. (2010).

So far, we have discussed local space-time features that are a direct extension of their 2D counterparts. They represent shape and motion by computing space-time signatures from neighboring pixels and aggregate them within video volumes centered at the location of STIPs resulting in a static descriptor. In contrast, trajectory features computed from tracked interest points (*e.g*., with the KLT tracker from Lucas and Kanade (1981)) capture long-term motion information.

For instance, Matikainen et al. (2009); Wang et al. (2011a) fix the length of the trajectories and describe their shape as a sequence of (normalized) displacement vectors (*i.e*., velocities). Messing et al. (2009) also represent trajectories as sequences of their velocity components, which are however log-polar quantized. Additionally to discretizing the displacement vectors, Bregonzio et al. (2010) encode the trajectory shape with its Fourier coefficients.

Sun et al. (2009a) also discretize the magnitude and orientation trajectory velocities in polar coordinates. However, they further map this representation to a fixed-length vector and model the intra-trajectory context by approximating a Markov stationary distribution. From this trajectory transition descriptor (TTD) a representation of the inter-trajectory context is computed. To this end, all TTD features are clustered and for all spatio-temporal volumes histograms of the TTD cluster indices are computed. These are then stacked to one occurrence matrix which is used to obtain the trajectory proximity descriptor (TPD) in form of a Markov stationary distribution vector. Inspired by the aforementioned approach, Matikainen et al. (2010) also employ polar discretization of the trajectory displacement vectors. Their spatial and temporal relationships are however modeled with relative space-time location probabilities estimated from training data.

2.1.4. Biologically inspired methods

Us humans would lose any competition versus a computer in tasks that can be directly translated into an algorithm. However, we still excel in areas like pattern recognition or language processing, where computers show at best a performance comparable to children. It is, therefore, no wonder that in the recent decades, much effort has been made to establish computational models mimicking the human brain (see Serre and Poggio (2010) for a recent review).

Handcrafted feature representations

One biologically inspired framework that had a significant impact on the Computer Vision community is the *HMAX* model (“Hierarchical Model and X“) proposed by Riesenhuber and Poggio (1999) for object recognition. HMAX is based on a simple hierarchical feed-forward architecture that models the ventral stream of the primary visual cortex, which is expected to be involved in shape representation (*cf.*, the two-streams hypothesis of the visual cortex from Goodale and Milner (1992)). It has originally been applied to simple object recognition tasks (Riesenhuber and Poggio (1999)) and subsequently demonstrated to outperform state-of-the-art in its updated form (Serre et al. (2007)).
Jhuang et al. (2007) have further extended the framework to model the dorsal stream, which is assumed, among others, to account for

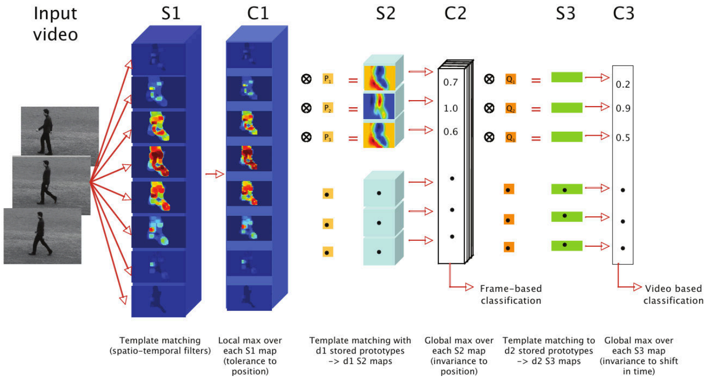

Input video
S1
C1
S2
C2
S3
C3
P1
P2
P3
0.7
1.0
0.6
Q1
Q2
Q3
0.2
0.9
0.5
Frame-based classification
Video based classification
Template matching (spatio-temporal filters)
Local max over each S1 map (tolerance to position)
Template matching with d1 stored prototypes -> d1 S2 maps
Global max over each S2 map (invariance to position)
Template matching to d2 stored prototypes -> d2 S3 maps
Global max over each S3 map (invariance to shift in time)

Figure 2.10. *Overview of the biologically inspired action recognition approach proposed by Jhuang et al. (2007). Similar to the human brain, the input video is processed in a hierarchical fashion, where each stage increases the model's robustness (©2007 IEEE).*

action perception of the brain. An overview of this approach is displayed in Fig. 2.10.

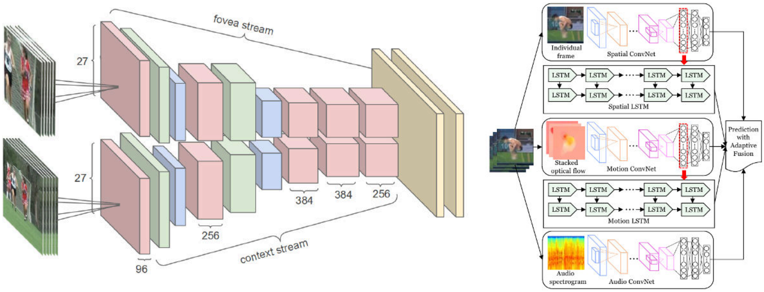
The basic features upon which the model builds are the S_1 units which are implemented as Gabor filters (*cf*., Gabor (1946)) that have been extended to capture the temporal dimension. Gabor filters have been shown to model simple cells in the visual cortex of mammalian brains (*cf*., Marcelja (1980)) and are thus employed in a wide range of biologically inspired systems (*e.g.*, Schindler and van Gool (2008); Shao et al. (2014)). The S_1 units are calculated on a dense spatio-temporal grid at different orientations and scales, in order to capture variations in size and rotation.

In the following C_1 stage, each S_1 type is locally max-pooled to incorporate some degree of invariance to small distortions. Then, S_2 feature maps are obtained by comparing the C_1 maps with templates learned from training data. This is followed by a global max-pooling over all feature maps obtained in the previous stage to obtain the spatially invariant C_2 feature representation. In order to add some temporal invariance to the C_2 features, another layer of template matching (S_3) and max-pooling (C_3) stages is added. Unlike in the

C_2 stage, where the units are pooled in each frame, C_3 features are calculated as the maximum response over the whole video.
The approach by Schindler and van Gool (2008) is similar in spirit to the spatio-temporal HMAX framework. However, besides of modeling the dorsal pathway of the visual cortex, it separately processes the ventral pathway as well, merging both before classification. Again, the simple perceptual units of the ventral stream (representing the form) are modeled using a bank of Gabor filters at different orientations and scales. The motion features, however, are implemented based on dense optical flow, which is converted to feature maps that are similar to the Gabor filter responses. The subsequent stages consist of local max-pooling and template matching, just as in HMAX. The final action representation is then obtained through a concatenation of similarities from both pathways.
The spatio-temporal Laplacian pyramid coding (STLPC) proposed by Shao et al. (2014) is another HMAX-like model. As the name suggests, the first layer consists of a Laplacian pyramid which has been found by Wilson and Bergen (1978) to resemble a multi-resolution technique employed by the human visual system. As in the SIFT descriptor (*cf*., Lowe (1999)), the Laplacian pyramid is approximated by differences of Gaussians. In order to extract edge and orientation features in the spatio-temporal domain, the image sequences are however convolved with a 3D kernel. In the next step, a 3D Gabor filter bank is applied to all levels of the pyramid intensifying the edge information. Finally, the filter responses are max-pooled between scales within a filter band, as well as over spatio-temporal neighborhoods making the descriptor scale-invariant and robust to position shifts. Since the resulting STLPC descriptor has a dimensionality of 5760, it is further compressed with a non-linear dimensionality reduction approach (*cf*., Zhang et al. (2009)).

Deep Learning based feature representations

A major disadvantage of the previously described biologically inspired methods is that all their parameters and connections are handcrafted. In contrast, artificial neural networks (ANNs, *cf*., Werbos (1982))



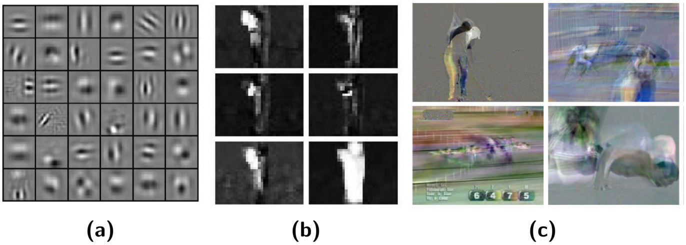
(a) Karpathy et al. (2014) **(b)** Wu et al. (2015)

Figure 2.11. *Examples of deep multi-stream network architectures for action recognition: (a) Multi-resolution ConvNet processing low-res images in a context stream and center cropped high-res images in a fovea stream. The streams consist of convolution (red), normalization (green), pooling (blue), and fully connected (yellow) layers (©2011 IEEE). (b) Multi-stream framework consisting of pre-trained ConvNets for feature extraction and separately trained LSTM paths for learning long-term dynamics.*

learn all the parameters directly from training data and are thus more flexible (note, that the network architecture still requires handcrafting).

Historically, ANNs gained much popularity in the 80's (*cf*., Schmidhuber (2015)) but subsequently were ousted in the late 90's by Support Vector Machines (*cf*., Cortes and Vapnik (1995)), and other, much simpler methods. While SVMs are simple and fast to setup and lead to outstanding pattern recognition results in all areas, ANNs suffered several problems making them nearly disappear.

Among the problems of ANNs was the false belief that gradient descent employed in the training would get trapped in local minima (*cf*., LeCun et al. (2015)). Furthermore, training large networks with many layers on conventional machines is very slow and can take up to several weeks of time (*e.g.*, Chatfield et al. (2014); He et al. (2015); Karpathy et al. (2014); Simonyan and Zisserman (2015)). The



(a) (b) (c)

Figure 2.12. *(a) Examples of layer 2 filters learned by the stacked convolutional ISA net from Le et al. (2011). Note the strong resemblance with Gabor filter responses (©2011 IEEE). (b) Subset of feature maps inferred from a KTH actions boxing clip using the unsupervised deep feature learning method proposed by Taylor et al. (2010) (©2010 Springer). (c) Dynamic images proposed by Bilen et al. (2016) summarizing action videos as 2D images that can be used as input for all conventional 2D ConvNets (©2016 IEEE).*

advent of fast GPUs significantly speeding up the training through massive parallelization, and the emergence of novel techniques (*e.g.*, ReLUs alleviating the vanishing gradient problem, or dropout to fight overfitting), allowed the creations of deep neural networks and has brought breakthroughs in many areas.

Some notable approaches in the field of Computer Vision are convolutional neural network (ConvNet) for object detection proposed by Erhan et al. (2014), Facebook's DeepFace system for face verification, as well as AlexNet (Krizhevsky et al. (2012)), GoogLeNet (Krizhevsky et al. (2012)), and the very deep VGGNets by Simonyan and Zisserman (2015), each drastically reducing the error on object recognition compared to shallow approaches. In few specialized tasks, that would require expert knowledge for humans to compete, deep learning methods have even achieved superhuman performance.

For instance, the traffic sign recognition system from Cireşan et al. (2012) not only achieves an error-rate that is six times lower than the best shallow (*i.e.*, non-deep learning based) method, but also beats humans in performing the same task. Likewise, deep learning approaches demonstrated superiority to humans in the tasks of fine-grained visual object recognition (He et al. (2015)), and identification of the geolocation of photos (Weyand et al. (2016)). In 2015, ANNs even enabled a computer program for the first time in history to beat a professional human Go player (Silver et al. (2016)).

It is, therefore, not surprising that researchers have started to explore the use of deep learning in action recognition as well (see Zhu et al. (2016) for a comprehensive survey). One of the first approaches applying deep learning to action recognition has been proposed by Ji et al. (2010) and extends 2D ConvNets to the spatio-temporal domain.

Taylor et al. (2010) simultaneously proposed another ConvNet architecture for action recognition, which is based on gated Restricted Boltzmann Machines. The network learns latent flow fields from pairs of temporally adjacent image frames in an unsupervised fashion (see Fig. 2.12(b)). The flow fields are then used as input features for a temporally extended ConvNet classifier that has originally been proposed for object recognition (Jarrett et al. (2009)).

Yet another unsupervised deep feature learning approach is the independent subspace analysis (ISA) based method proposed by Le et al. (2011). The receptive fields learned by ISA are similar to certain areas of the visual cortex resembling Gabor filter responses (*cf.*, Sec. 2.1.4, and Fig. 2.12(a)).

Motivated by the success of generic deep feature learning approaches in the image domain (*e.g.*, Chatfield et al. (2014); Jia et al. (2014); Sermanet et al. (2014)), Tran et al. (2015) introduced C3D, a generic 3D ConvNet for motion feature extraction. The employed ConvNet architecture is based on the VGGNet by Simonyan and Zisserman (2015), which uses very small convolutional kernels allowing rather deep models. One major advantage of C3D over other approaches is that the net only requires some fine-tuning to be applied to a new data set, instead of full re-training. Feature calculation is also very fast (*e.g.*, 91 times faster than improved Dense Trajectory features), while leading to a performance that is on par with local feature methods.

Just like the HMAX model, many deep learning methods for action recognition have been motivated by the two-stream hypothesis of the visual cortex, and thus process temporal and spatial information with separate nets (see Fig. 2.11 for example net architectures). Probably the first approach following this paradigm is the two-stream ConvNet proposed by Simonyan and Zisserman (2014). The spatial stream has a similar architecture to the ConvNet from Zeiler and Fergus (2014) and is pre-trained on a large still-image data set (ImageNet). The temporal stream is having the same structure but gets dense optical flow maps as input, and due to the lack of suitable data is fully trained from scratch. Fusion of both streams is performed at decision level via an SVM classifier trained on class score values from the softmax layers.

Wang et al. (2015b) employ the same two-stream framework as Simonyan and Zisserman (2014), yet in conjunction with two more recently proposed very deep models, namely GoogLeNet and VGGNet. Besides of streams for motion and appearance, Wu et al. (2015) have considered to also include acoustic information in a third stream (see Fig. 2.11(b)). An evaluation on UCF-101 and the Columbia Consumer Videos data set (Jiang et al. (2011b)) has, however, shown only a tiny improvement over using a two-stream architecture.

Karpathy et al. (2014) identify three reasons why action recognition has, so far, not benefited from deep learning as much as most other areas and propose ways to alleviate these problems, as outlined in the following. To cope with the necessity of large amounts of training data, the Sport-1M data set has been created, consisting of one million videos with 487 sports categories making it the currently largest action classification benchmark.

Since learning a deep ConvNet is very time-consuming, Karpathy et al. (2014) also propose a two-stream architecture speeding up training by a factor up to 4 without any sacrifice in accuracy (see Fig. 2.11(a)). The main idea is to process low-res images in a fovea stream while using center-cropped videos in the context stream which takes advantage of the camera bias present in most videos. Most importantly, an effective way is proposed to extend ConvNets from the 2D image domain to the 3D video domain while preserving action dynamics. Their Slow Pooling model first processes all video frames independently with a 2D ConvNet (AlexNet), and then hierarchically fuses frame level information over small temporal windows.

Ng et al. (2015) extend the work of Karpathy et al. (2014) by employing two-stream (raw images and optical flow) architectures and only a single max-pooling layer across all video frames. While using max-pooling in the image domain has many advantages, doing this in the time video domain results in a loss of dynamic information. Therefore they propose to use a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cells on top of the 2D ConvNet applied to each frame. Unlike plain feed-forward ANNs (*e.g.*, ConvNet), RNNs model dynamics with a feedback loop and are thus perfectly suited to process sequential data. RNNs suffer however from two major problems: vanishing (or exploding) gradient resulting in very slow training, and practical difficulty to capture long-term dependencies. To cope with these problems Hochreiter and Schmidhuber (1997) have designed LSTM cells, which are probably the most used RNN architectures today (*cf*., Baccouche et al. (2011); Donahue et al. (2015); Grushin et al. (2013); Li et al. (2016); Srivastava et al. (2015)). Nonetheless, LSTM architectures so far did not give any improvement in over feed-forward network models.

Classification of complex activities with ConvNets is addressed by Wang et al. (2014a). The framework uses raw depth data as input and incorporates a dynamically reconfigurable latent structure to decompose an activity into different length sub-actions.

Not only several ANN architectures for action recognition have been explored, but also approaches building around ANNs. Wang et al. (2015a) observe handcrafted local features to be complementary to deep learning based methods and thus propose a descriptor making advantage of both approaches. An overview of the calculation pipeline of their Trajectory-Pooled Deep Convolutional Descriptor (TDD) can be seen in Fig. 2.13. Basically, the TDD is a hybrid of the improved Dense Trajectories (iDT) descriptor proposed by Wang and Schmid (2013) (see Sec. 2.1.3) and the two-stream spatio-temporal ConvNet from Simonyan and Zisserman (2014). The approach starts

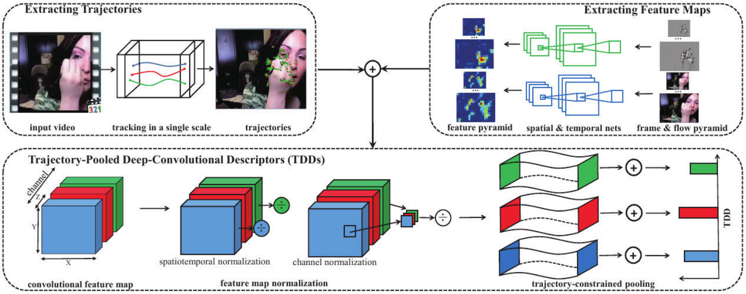

Extracting Trajectories
input video
tracking in a single scale
trajectories
Extracting Feature Maps
feature pyramid
spatial & temporal nets
frame & flow pyramid
Trajectory-Pooled Deep-Convolutional Descriptors (TDDs)
channel
Z
Y
X
convolutional feature map
spatiotemporal normalization
channel normalization
feature map normalization
trajectory-constrained pooling
TDD

Figure 2.13. *Extraction pipeline of Trajectory-Pooled Deep Convolutional Descriptor (TDD) proposed by Wang et al. (2015a). The algorithm consists of three steps: extracting trajectories, computing ConvNet-based feature maps, and pooling these feature maps within spatio-temporal volumes in the neighborhood of the trajectories (©2015 IEEE).*

off by computing normalized feature maps at different convolution layers of the two-stream ConvNet proposed by Simonyan and Zisserman (2014). Then, TDDs are obtained by sum-pooling the feature map contents within spatio-temporal volumes calculated in the neighborhood of tracked local features.

Fernando et al. (2016a) also pool deep feature maps using a method that has originally been proposed for handcrafted local descriptors, namely discriminative rank pooling (*cf*., Fernando et al. (2016b), and Sec. 2.1.3). Rank pooling models the evolution over time of motion and appearance in an image sequence by using the parameters of a linear ranking machine fitted to the data. To capture higher order dynamics, Fernando et al. (2016a) construct a hierarchical network of rank pooling layers that conceptually resembles a deep neural network.

Rank pooling is also employed by Bilen et al. (2016), however, to pre-process videos that serve as input to a ConvNet. The authors claim that one major disadvantage of ANNs is that their architecture needs to be handcrafted, which appears to be especially difficult when working on image sequences. Therefore, they summarize motion and appearance of an action in a single dynamic image by applying rank pooling to the video and use this representation in conjunction to a 2D ConvNet (VGGNet). Examples of dynamical images generated using the approach can be seen in Fig. 2.12(c). Surprisingly, such a simple and lossy video transformation leads to a very high classification rate, which is even on par with state-of-the-art when fusing deep feature maps with iDT descriptors.

Overall, it is left to say that just as in any other field of pattern recognition, deep learning methods are currently the most widely explored models for action recognition. However, unlike in many other fields, no huge leap in classification performance has been achieved yet for actions. The major challenges are the need of a suitably large training data set, the huge increase in the number of model parameters when moving from the 2D spatial to 3D spatio-temporal domain, and the difficulty to properly capture long-term dynamics with a network architecture. To put it into perspective, ImageNet (Deng et al. (2009)), a commonly used benchmark for object recognition, consists of about 15 million images belonging to all kinds of categories. Yet Sports-1M, the largest action recognition data set only consists of one million samples, all of which belong to only one field of actions. The current golden standard benchmarks are even smaller, with UCF-101 containing 9500 videos and HMDB-51 only 3700. In this context, one also needs to consider that actions are much more complex than objects due to their variations not only in appearance but also in dynamics. Spatio-temporal ANNs might, therefore, make a more complex network architecture necessary than their spatial counterparts. The resulting increase in training samples, as well as model parameters, would, however, inevitably lead to a huge increase in computational complexity. Nonetheless, deep learning methods are still a very promising direction for action recognition research.

2.2 Video representation

Once a representation of motions has been computed, the features still need to be mapped to the proper activity classes. This step usually involves a representation of the whole video, which can be structured, or unstructured.

Unstructured approaches, like the popular Bag-of-Words framework, disregard most spatio-temporal location information about the features and purely make the classification based on their presence (or absence). Even though this method may sound counter-productive for the recognition of activities due to their structured nature, it is employed successfully in most state-of-the-art action and activity recognition methods.

Structured methods, however, try to capture the underlying structure of a problem with (probabilistic) graphical models. They are probably the most popular approach for the recognition of very complex activities, yet often used in conjunction with unstructured approaches.

2.2.1. Unstructured models

The probably most widely employed unstructured model for image and video recognition task is the popular Bag-of-Words (BoW) representation. It originates from natural language processing and has been first applied to Computer Vision tasks by Csurka et al. (2004); Sivic and Zisserman (2003). In its most simple form, vector quantization (VQ), multi-dimensional feature vectors are first mapped to scalar dictionary indices of visual words. Next, a histogram of these indices is built, disregarding any spatial or temporal location of the original features. Typically, the dictionary is learned through k-means clustering (*cf.*, MacQueen (1967)), although more elaborate clustering methods are used as well, *e.g.*, Expectation Maximization (EM, *cf.*, Dempster et al. (1977)) to learn Gaussian Mixture Model (GMM) parameters for Fisher Vector encodings. Note, that unlike the other feature representations discussed in this section, BoW methods only combine local features into a global video descriptor, but do not implement any mapping of the descriptors to target categories. Thus, they are usually further employed as input features for classification algorithms, such as random decision forests (*cf.*, Ho (1995)), (naïve Bayes) nearest neighbors (*cf.*, Boiman et al. (2008)), logistic regres-

Reference	Detectors	Descriptors	BoW	Focus of experimental evaluation
Wang et al. (2009)	diverse	diverse	VQ	dense sampling parameters, computation time
Shao and Mattivi (2010)	diverse	diverse	VQ	codebook size, computation time
Bilinski and Bremond (2011)	Harris3D	STIP, HOG3D	VQ	codebook size
Wang et al. (2012c)	Harris3D	STIP	diverse	feature pooling and normalization, codebook size
Wu et al. (2014)	Harris3D, DT	STIP, DT	VLAD	BoW post-processing parameters
Peng et al. (2014a)	Harris3D, DT	STIP, DT	VLAD	codebook learning methods, computation time
Zhen and Shao (2016)	Cuboids	HOG3D	diverse	SVM kernel
Peng et al. (2016)	Harris3D, iDT	STIP, iDT	diverse	BoW post-processing parameters, codebook size, feature fusion, computation time

Table 2.2. *Overview of publications presenting a large-scale evaluation of different Bag-of-Words methods in conjunction with local, spatio-temporal features for action recognition. Note, that here the set of descriptors proposed by Laptev et al. (2008) is depicted as STIP, and for the set of descriptors employed by Wang and Schmid (2013), the term iDT is used.*

sion (*cf.*, Cox (1958)), or, most commonly, (soft-margin) Support Vector Machines (*cf.*, Cortes and Vapnik (1995)).

Shortly after they have been adapted to image classification tasks, BoW based approaches have been dominating this field of research for nearly one decade. For instance, during the whole duration (2005-2012) of the PASCAL Visual Object Categorization (VOC) project (*cf.*, Everingham et al. (2015)) and the early years of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC, *cf.*, Russakovsky et al. (2015)), both being the standard benchmarks for object detection and classification approaches, most submissions, as well as winning systems were based on variants of BoW. Consequently, BoW encodings have also been employed for action and activity recognition, starting with simple VQ (*e.g.*, Dollár et al. (2006); Kläser et al. (2008); Laptev et al. (2008); Scovanner et al. (2007); Wang et al. (2009)) which resulted in a huge improvement in performance over previous state-of-the-art. One major drawback of VQ methods is, however, that much information is lost in the discretization process of the high-dimensional raw feature vectors. Therefore many other BoW variants were proposed to compensate this loss (*e.g.*, the works of Cai et al. (2014); Jegou et al. (2012); Perronnin and Dance (2006); Wang et al. (2010)).

The locality-constrained linear coding (LLC) proposed by Wang et al. (2010) belongs to the category of reconstruction based encoding BoW approaches. These algorithms are designed from the perspective of the decoding process enforcing the codes to reconstruct the input descriptor. Works using this encoding for action recognition have been proposed by, *e.g.*, Peng et al. (2016); Rahmani et al. (2014). Sparse coding (*cf.*, Yang et al. (2009)) is another popular BoW encoding method belonging to the same category as LLC, that has widely been explored for action recognition (*e.g.*, Guha and Ward (2012); Luo et al. (2013); Yang and Tian (2014)).

The most successfully used BoW method is, however, the Fisher Vector (FV) encoding proposed by Perronnin and Dance (2006). Fisher Vectors are obtained by aggregating the first and second order statistics of local descriptors. Compared to VQ, FVs are very high dimensional, as their dimensionality depends both on the dictionary size, as well as the dimensionality of the local descriptors. Suggestions on how to improve the performance of FV-based methods by means of pre- and post-processing steps (*e.g.*, normalization, power transform) are discussed by Perronnin et al. (2010). Even though deep learning constitutes the, at the moment, probably most popular choice for action and activity recognition, Fisher Vector encoded STIPs are still among state-of-the-art on many action recognition benchmarks (*cf.*, Kantorov and Laptev (2014); Kuehne and Serre (2016); Oneata et al. (2013, 2014b); Rostamzadeh et al. (2013); Wang and Schmid (2013)). A multi-layer nested Fisher Vector encoding (SFV) for action recognition has been proposed by Peng et al. (2014c) demonstrating a performance improvement over the traditional FV. Since computation of the FV representation is comparably time-consuming, Jegou et al. (2012) proposed the vector of locally aggregated descriptors (VLAD) as an approximation of FV. Basically, VLAD can be viewed as a hard version of FV since it only keeps the first order statistics. This encoding is often applied as well for action recognition problems, *e.g.*, by Jain et al. (2013); Kantorov and Laptev (2014); Peng et al. (2014b). Besides of the aforementioned encodings, many other BoW variants have been proposed, *e.g.*, soft assignment coding (*cf.*, van Gemert et al. (2010)), orthogonal matching pursuit (*cf.*, Tropp and Gilbert (2007)), local coordinate coding (*cf.*, Yu et al. (2009)), multi-view super vector (*cf.*, Cai et al. (2014)). Since a complete discussion of these methods is beyond the scope of this literature review, we refer the interested reader to the references given in Tab. 2.2. All of them contain comparisons of most BoW methods evaluated on action recognition benchmarks under different settings.

Traditional BoW disregards all information about the spatial (and temporal) structure of features and is, therefore, incapable of capturing the shape of objects. To alleviate this shortcoming, several extensions of BoW have been developed that can be generally applied to all types of BoW methods. The spatial pyramid matching (SPM) scheme proposed by Lazebnik et al. (2006) partitions the input data (*i.e.*, images or videos) into a hierarchy of differently sized sub-regions and encodes each with a separate BoW encoding. Action recognition methods following this paradigm are, for example, proposed by Laptev et al. (2008), and other hierarchical BoW extensions by Niebles and Fei-Fei (2007); Sun et al. (2016). Spatio-temporal extensions of other methods successfully applied to many problems of image analysis, like the discriminatively trained deformable part model from Felzenszwalb et al. (2010), or the implicit shape model from Leibe et al. (2004) have been explored as well, *e.g.*, by Lan et al. (2011); Wang et al. (2014b).

Instead of generalizing methods originally proposed for 2D data, many other researchers developed BoW variants that explicitly take advantage of the spatio-temporal nature of actions. For example, Liu et al. (2014a) propose to segment action sequences into sub-actions and describe each as a separate BoW, while Karaman et al. (2014) focus on a spatio-temporal region segmentation. Another method to incorporate some degree of structural information in the BoW model is by regarding contextual statistics of neighboring interest points, as is done by *e.g.*, Bilinski and Bremond (2012); Wolf et al. (2014b); Zhang et al. (2012).

Two recent methods that deserve a more detailed look due to their impressive performance on current benchmarks are the multi-skip feature stacking (MIFS) technique proposed by Lan et al. (2015b), and rank pooling (or VideoDarwin) introduced by Fernando et al. (2016a). Lan et al. (2015b) observe that the core of the Gaussian pyramid employed in spatial pyramid matching consists of a convolutional smoothing operation making the approach incapable of generating new features at coarse scales. As a possible solution to this problem, MIFS is proposed, stacking features extracted with a family of differential filters parameterized with multiple time skips and encoding shift-invariance into the feature space. This allows the feature representation to capture actions at different speeds and ranges of motion.

As the name suggests, the VideoDarwin representation from Fernando et al. (2016a) captures the video-wide temporal evolution. First, all frames are separately encoded as a BoW and the results temporally smoothed, *e.g.*, by a moving average within a temporal sliding window. Then, the temporal evolution of a video is encoded

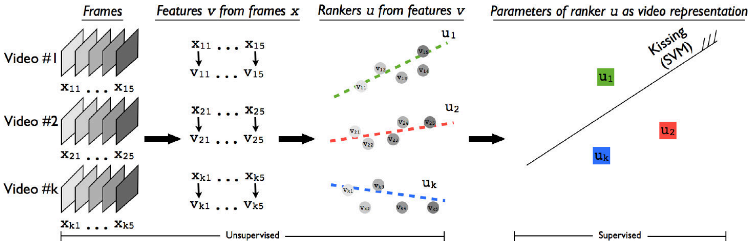


Figure 2.14. *Illustration of the processing pipeline used by the rank pooling algorithm for action recognition proposed by Fernando et al. (2016b). First, BoW features are computed for each frame t, using local descriptors extracted from all frames from the beginning up to frame t. Next, video representations u_i are learned from these features based on ranking machines. Finally, these video representations are employed as feature vectors for action classification (©2016 IEEE).*

in terms of the parameters of a linear ranking machine trained on all frame-wise feature representations. An illustration of the whole processing pipeline can be found in Fig. 2.14

Probabilistic topic models are another category of unstructured feature representations originally employed in NLP that have found their way to image and video processing. Topic models are statistical models that try to capture the latent topics that occur in a set of documents. Since they assume that the order of words in a document is not significant, topic models are often described as Bag-of-Words models. They are, however, conceptually different from the BoW models that we have previously discussed. One key difference lies in the diverse range of problems that latent topic models can be applied to, *e.g.*, for feature selection, clustering, and dimensionality reduction.

One of the first topic models that is still commonly used is the probabilistic latent semantic analysis (pLSA) introduced by Hofmann (1999). It models each word (*i.e.*, dictionary entry for local feature based methods) as a sample from a mixture model, where the mixture components can be viewed as representations of latent topics. Therefore each word is generated from a single topic and each document (*i.e.*, video in our case) is represented as a list of mixing proportions of this mixture components.

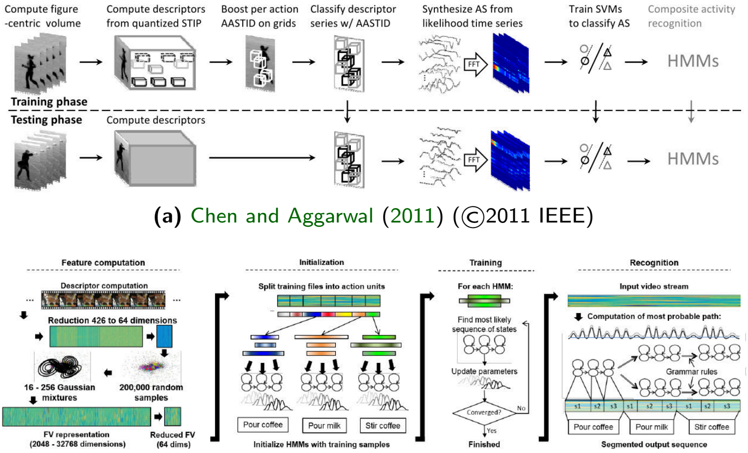
Example approaches employing pLSA for action recognition are the works of Wong et al. (2007); Zhang and Gong (2010a); Zhang et al. (2008). Note, that the latter two works also extend pLSA allowing it to make use of both semantic and structural information.

A theoretical shortcoming of pLSA that is observed by Blei et al. (2003) is its incompleteness in the sense that it provides no probabilistic modeling at document-level. As a solution, they proposed the latent Dirichlet allocation (LDA) which extends pLSA by assuming the topic distribution to have a Dirichlet prior. The application of LDA to action recognition has been explored by *e.g.*, Messing et al. (2009); Niebles et al. (2008). Variants of LDA developed for action recognition are the semi-latent Dirichlet allocation from Wang and Mori (2009), multi-class Δ LDA from Bregonzio et al. (2010), hierarchical variations of LDA, proposed by *e.g.*, Yang et al. (2014); Yin and Meng (2010).

Lastly, we should mention another very popular unstructured representation - boosting (*e.g.*, AdaBoost developed by Freund and Schapire (1995)). Boosting describes a family of machine learning algorithms that are based on the assumption that a combination of weak learners is capable of creating a single strong classifier. Weak learners are simple classifiers that are only slightly better than random guessing. In contrast, strong learners are classifiers that are well-correlated with the true classification.

Boosting has been successfully applied to many Computer Vision tasks, most notably it constitutes the core of the famous face-detection algorithm developed by Viola and Jones (2004). In the context of action recognition, many variants of boosting have been employed as well. For instance, Laptev and Pérez (2007); Liu et al. (2009) used AdaBoost to discriminate between different actions, Huang et al. (2011) proposed LatentBoost, a boosting variant incorporating latent variables for action recognition, and Fathi and Mori (2008); Kim and Cipolla (2009) use boosting for feature selection.

2.2.2. Structured Models


Compute figure -centric volume
Compute descriptors from quantized STIP
Boost per action AASTID on grids
Classify descriptor series w/ AASTID
Synthesize AS from likelihood time series
Train SVMs to classify AS
Composite activity recognition
FFT
HMMs
Training phase
Testing phase
Compute descriptors
FFT
HMMs
Feature computation
Descriptor computation
Reduction 426 to 64 dimensions
16 - 256 Gaussian mixtures
200,000 random samples
FV representation (2048 - 32768 dimensions)
Reduced FV (64 dims)
Initialization
Split training files into action units
Pour coffee
Pour milk
Stir coffee
Initialize HMMs with training samples
Training
For each HMM:
Find most likely sequence of states
Update parameters
Converged?
No
Yes
Finished
Recognition
Input video stream
Computation of most probable path:
Grammar rules
s1
s2
s3
s1
s2
s3
s1
s2
s3
Pour coffee
Pour milk
Stir coffee
Segmented output sequence

(a) Chen and Aggarwal (2011) (©2011 IEEE)

(b) Kuehne and Serre (2016) (©2016 IEEE)

Figure 2.15. *Illustrations of HMM architectures employed in two activity recognition methods that were inspired by automatic speech recognition systems.*

Even though unstructured models have been applied with great success to many action recognition tasks, the importance of temporal structure, especially for complex activities, has been widely studied as well. From all these probabilistic graphical models, the most prominent (and probably best studied) is certainly the hidden Markov model (HMM, *cf*., Baum and Petrie (1966)). Its great success for speech recognition and natural language processing made the HMM particularly famous, and thus it became a common tool used in time-series analysis.

An early (if not the first) attempt to employ HMMs for action recognition is undertaken by Yamato et al. (1992) for the classification of tennis strokes. Subsequently, many more researchers have studied the use of HMMs and their variants to model actions, and complex activities (*e.g.*, Ikizler and Forsyth (2008); Weinland et al. (2007a); Xia et al. (2012)). Chen and Aggarwal (2011); Kuehne and Serre

(2016); Kuehne et al. (2014) take inspiration from automatic speech recognition and model human activities as speech making use of HMMs. HMM architectures employed in these approaches to model complex activities are illustrated in Fig. 2.15. Example variants of HMMs that have been applied to action and activity recognition are conditional HMMs (*e.g.*, Glodek et al. (2012)), factorial HMMs (*e.g.*, Ramanan and Forsyth (2003)), variable duration HMMs (*e.g.*, Tang et al. (2012)), maximum entropy HMMs (MEMM, *e.g.*, Sung et al. (2011)), and HMMs with multiple independent observations (*e.g.*, Concha et al. (2011)).

Hybrid approaches, jointly leveraging the power of HMMs and other popular models have been explored as well. Bargi et al. (2012) proposed the hierarchical Dirichlet process HMM for a joint segmentation and classification of actions that allows for the discovery of new classes as they occur. Raman and Maybank (2015) also make use of HDPs (*cf.*, Teh et al. (2005)) to improve HMMs, however with the purpose to infer the number of hidden states automatically from training data instead of having to specify them a-priori. In order to combine the advantages of deep neural networks with HMMs, Wu et al. (2014) propose the use of deep ANNs to replace the Gaussian Mixture Models that are usually employed to model the underlying distribution of the HMM observation model.

A generalization of HMMs (and other linear state-space models) to arbitrary nonlinear and non-normal time-dependent domains has been established by Dagum et al. (1992) with the dynamic Bayes Networks (DBN). Their application to action recognition has been explored by *e.g.*, Gupta and Davis (2007); Laxton et al. (2007); Vo and Bobick (2014); Zeng and Ji (2010). In one of our earlier works (*cf.*, Gehrig et al. (2011)), we have proposed a hybrid DBN based approach (*cf.*, Schrempf et al. (2006)) to fuse higher-level dynamics, domain knowledge, and human motion estimates (*i.e.*, motion-primitives and actions) to classify complex high-level activities.

Another Bayes Network based multi-level system has been proposed by Park and Aggarwal (2004). It makes use of a Bayes Network architecture for body pose estimation. This pose model is subsequently transformed in a DBN for multi-human interaction by establishing temporal links between its hidden nodes.

A model very similar to a Bayesian Network in its representation dependencies is the Markov Random Fields (MRF) introduced by Kindermann and Snell (1980). The core difference between an MRF and a Bayesian Network is that the former is undirected and allows cycles in the graphical representation. This allows MRFs to model certain dependencies that cannot be established with a Bayes Network. Example approaches in the field of action and activity recognition that are based on MRFs have been developed by Choi and Savarese (2011); Koppula et al. (2013); Lu et al. (2015)

The aforementioned models (*i.e.*, with the exception of maximum entropy HMMs) have in common that they all belong to the category of generative machine learning algorithms. Generative algorithms try to model the underlying probability distribution from which the observed data samples were generated. Since the true nature of this distribution is usually not known, strong assumptions about it need to be made in order to achieve a good approximation. Consequently, this can lead to either very complex models or reduced performance. To counter these shortcomings Lafferty et al. (2001) have developed Conditional Random Fields (CRF), graphical models which are discriminative.

It is no surprise, that CRFs have found as well its way to action recognition. Sminchisescu et al. (2006) were the first to advocate the use of CRFs for human motion analysis. Since CRFs have the limitation of not being capable to properly capture intermediate structures, several extensions have been proposed, like hidden-state CRFs (*cf.*, Quattoni and Wang (2007)), or factorial CRFs (*cf.*, Sutton et al. (2007)), and, consequently, applied to recognize actions, *e.g.*, by Kjellström et al. (2008); Wang and Suter (2007); Wang and Mori (2008); Zhang and Gong (2010b).

In addition to graphical models, some researchers resorted to max-margin methods, formulating the learning problem using latent SVMs (*e.g.*, Yu and Joachims (2009)). Wang and Mori (2011) proposed the max-margin hidden-state CRF (MM-HCRF) and demonstrated its advantage over conventional HCRFs on several action recognition benchmarks. Similar to MM-HCRFs are latent structural SVMs (*cf*., Yu and Joachims (2009)), which found their application for action recognition in the works of *e.g.*, Liu et al. (2014b); Packer et al. (2012); Wu et al. (2013).

2.3 Activity recognition

As we have seen in the previous sections, there has been an impressive amount research conducted to automatically classify simple actions (*e.g.*, `standing up` or `smoking`) from video data. Even though recognition of atomic actions is interesting for multimedia retrieval or human surveillance tasks, many more real-world applications depend on the recognition of complex activities. Exemplary application areas are human-robot interaction, elderly care, and assistive technologies (*e.g.*, to monitor Alzheimer patients in order to remind them to perform forgotten everyday tasks). Therefore, with the increasing success of action recognition approaches, automatic understanding of activities has attracted the attention of many researchers as well. Since activities are sequences of fine-grained actions, methods discussed in Sec. 2.1.1 - 2.1.4 are also employed as building blocks of most activity recognition approaches. Due to the sequential nature of activities, probabilistic graphical models are particularly suited for recognition. For instance Kuehne and Serre (2016) encode local features (iDT) extracted from each video frame as separate Fisher Vectors which are used as mid-level features for HMMs to recognize atomic actions. These are then combined with a context-free grammar learned from training data to map the action sequences to activity labels.

In a similar fashion, Chen and Aggarwal (2011) model activities as speech. To this end, local spatio-temporal features encoding motion (HOF) and appearance (HOG) are densely sampled from video data and used as features for AdaBoost. Given the highest weighted weak classifiers, action spectrograms are synthesized from time-slices of the feature time-series via FFT. Next, linear SVMs are trained to classify actions from spectral data extracted from the time-slices.

For activity recognition, the SVMs are used to estimate the posterior probabilities in the observation model of HMMs.

Unstructured models have been (to a lesser extent) as well employed for activity recognition. Messing et al. (2009) capture long-term dynamics in their motion feature encoding (velocity history features) and use supervised latent Dirichlet allocation as a classifier. Ryoo and Aggarwal (2009) add structural relationships between space-time interest points to the Bag-of-Words model, by defining spatial and temporal predicates and binning the STIPs accordingly in 3D (feature × feature × relation) histograms. Other approaches to incorporate structural information in Bag-of-Words are the time-flexible kernel framework from Rodriguez et al. (2016) and Bag-of-Attribute-Dynamics model from Li et al. (2016). Methods based on an automatic decomposition of complex activities into atomic action segments have been explored as well, *e.g.*, by Wang et al. (2014b, 2016b).

All of the aforementioned activity recognition approaches have in common that they are solely based on low-level motion (and appearance) features and disregard any context information. Nonetheless, context, such as scene, or presence of certain objects, can often be used to constrain the search space of all possible activities to a small subset and should, therefore, improve the recognition performance. In fact, this claim is backed up by studies on human perception from the neurological (*e.g.*, Gallese et al. (1996); Nelissen et al. (2005)), as well psychological (*e.g.*, Bach et al. (2005); Bub and Masson (2006)) standpoint. It is, therefore, no wonder that joint modeling of scene, object, and action has recently become a hot topic of interest in the Computer Vision community.

In the following, we are going to give an overview of works using context information for activity recognition. Since our contribution is dealing with the incorporation of object information for activity recognition, we are going to restrict the literature review to works focusing on this field only.

2.3.1. Supervised object detection for activity recognition

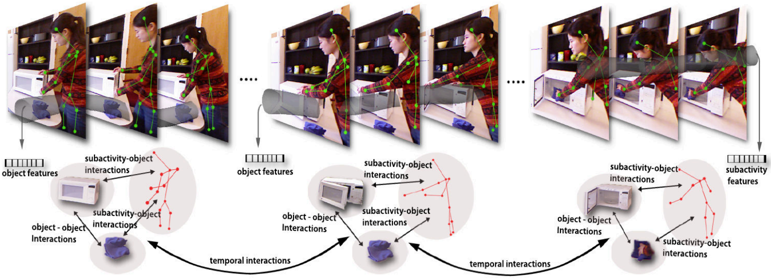

....
....
object features
subactivity-object interactions
subactivity-object interactions
object - object Interactions
temporal interactions
object features
subactivity-object interactions
object - object Interactions
subactivity-object interactions
temporal interactions
subactivity-object interactions
object - object Interactions
subactivity features
subactivity-object interactions

Figure 2.16. *Overview of the MRF-based approach proposed by Koppula et al. (2013) for joint action, object, and activity recognition from RGBD data.*

Some simple ways to incorporate object knowledge for activity recognition are by directly using ground-truth labels (*e.g.*, Hamid et al. (2009); Koppula and Saxena (2013a)), and possibly adding artificial noise to simulate imperfect detections (*e.g.*, Gehrig et al. (2011)), or by attaching RFID tags to all relevant objects (*e.g.*, Patterson et al. (2005); Wu et al. (2007)). The most common source of object knowledge for activity recognition is, however, supervised detectors. The renaissance of deep convolutional neural networks (*cf*., Sec. 2.1.4) made this method particularly attractive due to the high performance of ConvNet-based object detectors (*e.g.*, He et al. (2016)).

This progress also enabled Jain et al. (2015b) to investigate the extent of how 15000 object categories can help to classify actions. To this end, object categories with at least 100 samples were selected from ImageNet and used to train an AlexNet model (Krizhevsky et al. (2012)). An evaluation on several action recognition data sets revealed that when solely using the object detector responses as features already quite reasonable classification rates (*i.e.*, ˜20% (abs) lower than using motion features) can be achieved. Note, that the fusion of object and motion features only resulted in slight, yet significant performance improvement over motion features alone. Another finding was that actions have object preferences and thus, instead of using all object categories, selection can be beneficial when using general-purpose detectors.

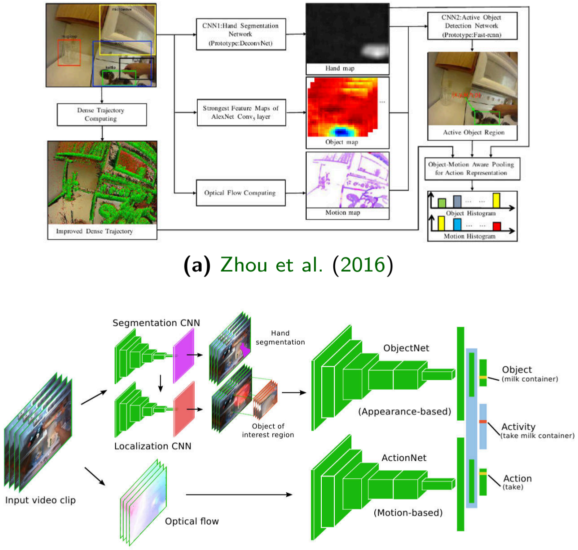
In a follow-up work, Jain et al. (2015a) propose object2action, a semantic embedding to classify actions without the need of video data for training (*i.e.*, zero-shot recognition). Instead, this action representation is solely based on object annotations, images, and textual descriptions.

Even though the aforementioned methods were only applied to the recognition of actions, the results should be also applicable for activities, which are nothing more than sequences of actions. In fact, Philipose et al. (2004) postulated the *invisible human hypothesis* stating that activities are well characterized by the objects that are manipulated during their performance. This hypothesis is supported by many works in the field of pervasive computing (*e.g.*, Patterson et al. (2005); Philipose et al. (2004); Wu et al. (2007)), where information about manipulated objects is obtained from RFID sensor-glove readings.

As far as unimodal activity recognition systems go that solely rely on video data, many researchers have explored the joint use of object and motion observations. Basically, three main types of approaches can be identified on how object knowledge is incorporated to aid activity recognition:

- object information is used as a separate cue (together with *e.g.*, motion features, or scene information) for activity recognition
- mutual relationship between objects and motions is modeled to improve activity recognition
- mutual relationship between objects and activities is leveraged to improve the performance of classifiers for both information sources

For example, the approach from Rohrbach et al. (2015) belongs to the first category. It is based on stacking temporally max-pooled responses of object and atomic action classifiers in a single vector which is mapped to the activity class by an SVM classifier. Several types


CNN1:Hand Segmentation Network (Prototype:DeconvNet)
Hand map
CNN2:Active Object Detection Network (Prototype:Fast-rcnn)
Dense Trajectory Computing
Strongest Feature Maps of AlexNet Conv5 layer
Object map
Active Object Region
Object-Motion Aware Pooling for Action Representation
Optical Flow Computing
Motion map
Object Histogram
Motion Histogram
Improved Dense Trajectory
Segmentation CNN
Hand segmentation
ObjectNet
Object (milk container)
(Appearance-based)
Localization CNN
Object of interest region
Activity (take milk container)
ActionNet
Action (take)
Input video clip
Optical flow
(Motion-based)

(a) Zhou et al. (2016)

(b) Ma et al. (2016)

Figure 2.17. *Overview of two example ConvNet architectures used for activity recognition with object knowledge: (a) Hybrid approach fusing iDT encoded motion information with ConvNet-based active object detections (©2016 IEEE). (b) Multi-stream approach where the second to last layers of object and action networks are fused for a joint inference of objects, actions, and activities (©2016 IEEE).*

of features are considered: hand centered motion and appearance features (iDT and color SIFT), body model features (joint velocity histories and trajectory FFT coefficients), holistic features (vector quantized iDT). All feature types are used separately for action, and object detection. In order to cope with the lack of training samples of complex activities, automatically mined script data is considered in the approach as well.

Two recent ConvNet-based approaches that belong to the same category as the aforementioned one have been proposed by Ni et al.

(2016); Zhou et al. (2016). The ConvNets are used for object detection and their responses are fused with Fisher Vector encoded iDT features for SVM classification. The main difference between both methods is that the former focuses on the distinction between active and inactive objects by means of hand segmentation and optical flow (see Fig. 2.17(a)). Yet the main contribution of the latter one is an LSTM-based object detection refinement between frames (*i.e.*, tracking).

The system proposed by Ni et al. (2014) is conceptually similar to the previously mentioned one, as it also obtains object knowledge through tracking by detection (however using shallow methods). Instead of fusing motion and object features by concatenation, the authors, however, opt to model the correlation between both cues with a CRF. Intuitively, exploiting the mutual relationship between the performed motion (or action) and observed objects has many advantages. Especially in the case of occlusions or miss-classifications, people's interaction with the objects can provide enough context information to compensate the object detection errors. Therefore, it is not surprising that most works making use of object detectors to improve activity recognition explicitly model the object-action relations.

A purely body model based approach has been proposed by Wei et al. (2016) with the 4D Human object interaction (4DHOI) model. The depth channel of a Kinect sensor is exploited to restrict the search space for object detection to non-void regions close to the human body. To encode body motion, the difference of joint coordinates in two successive frames is taken. The core of the algorithm is a stochastic hierarchical spatio-temporal graph representing 3D human-objects relations and temporal relations between sub-activities (*i.e.*, atomic actions). To learn the hierarchical structure of atomic actions, an Expectation Maximization step is employed.

The approach of Packer et al. (2012) also jointly models body pose trajectories and object manipulations for activity recognition. However, besides of skeleton based features the method also considers Cuboid features sampled along the pose trajectories. To speed up object detection, only regions that neither belong to the background nor can be explained as body parts are considered, and also tracked in subsequent frames. Additionally, hand regions are included in the set of object candidates, since the hand can easily occlude large parts of smaller objects. This way, most of the objects that are being manipulated by the observed person can be obtained.

Still, much helpful information is ignored when only focusing on foreground regions. Instead, landmarks are introduced to capture regions at which specific atomic actions occur (*e.g.*, a cutting board during a chopping action). These are modeled as latent variables in a latent structural SVM framework additionally to object and motion observation.

Koppula and Saxena (2013b) propose an automatic method for joint sub-activity (*i.e.*, action) and object affordance labeling (see Fig. 2.16). Relations between objects and actions are modeled as a Markov random field. A histogram of the inferred action and object affordance labels is finally used for high-level activity recognition. As with most methods of this kind, this approach relies on object annotations, since it depends on trained object classifiers.

Hu et al. (2015) introduce a human-object interaction descriptor (HOI) which relies on object- and human torso detection, as well as body pose annotations. The descriptor draws its power from the assumption that for different instances of an activity class, the manipulated object appears at a similar relative position to the human body. First atomic pose exemplar classifiers and object locations relative to the body are learned from training data. To compensate for inaccuracies of object detectors, object-location priors conditioned on the body pose are learned as well. Activities are then represented with such spatial pose-interaction exemplars which are probability density functions describing spatially how a person is interacting with a manipulated object (see Fig. 2.19(a) for examples).

Gupta et al. (2009) argue that it is often difficult to discriminate between objects based on their shape alone (*e.g.*, `spray can` vs. `drinking bottle`), yet knowledge about their functionality can provide necessary information for recognition. The same principle applies to actions that can often only be discerned through knowledge of the involved objects. Therefore, they present an approach to model such relationships for joint classification of activities and related objects. Activities are first classified with HMMs based on hand motion, and objects detected with a cascade of AdaBoost classifiers operating on HOG features. The joint relationship between objects and activities is established with a Bayes Network and it is demonstrated that indeed both cues can be used to improve each other for recognition. Likewise, Liu et al. (2014b) focus their work on inferring the best action, object, and scene combination for a test sample. The approach relies on pre-learned detectors for all contextual cues. A latent SVM is used to learn the co-occurrence relationship of object, scene, and action. Yao et al. (2011) explore with CRFs the use of another popular graphical model for simultaneous inference of activity, objects, and body part locations.

Deep Learning based approaches have been recently proposed as well for the joint recognition of objects, actions, and activities. The method proposed by Ma et al. (2016) consists of two ConvNet streams as shown in Fig. 2.17(b). One stream consists of a hand-segmentation net that is fine-tuned to localize manipulated objects, and a subsequent object recognition net operating on the object location heat map. The other stream is operating on dense optical flow and is trained for action recognition. In order to capture the co-relation of objects and actions, fusion is performed by concatenating the second last fully connected layers of both streams.

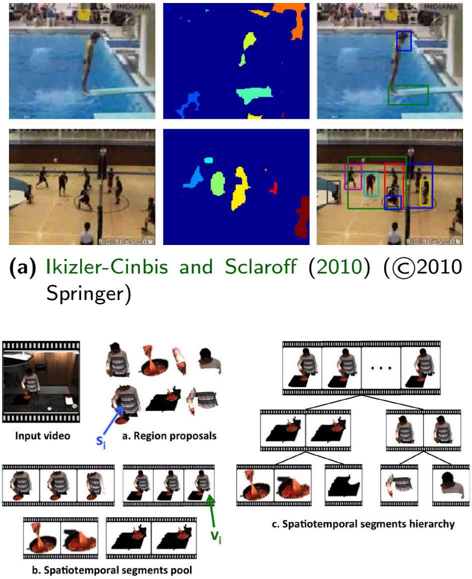
Unlike all previously mentioned approaches, Kjellström et al. (2008, 2011) focus their work on the simultaneous recognition of subactivities and manipulated objects, but not on activity recognition itself. Actions are represented by motion and appearance features of the hand and since only manipulated objects are considered, objects are represented by the same features as the hand shape. Detection of actions and objects is then performed with CRFs (and variants thereof), where both cues are jointly used as observed data.

2.3.2. Unsupervised object detection for activity recognition

As we have seen in the previous section, most activity recognition approaches are relying on supervised detectors to obtain object information. But building a robust detector handling all types of object classes is still challenging, despite the great advances that have been achieved thanks to deep learning methods. To put it into perspective, the winning system (*i.e.*, the deep residual net from He et al. (2016)) of the ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015, *cf*., Russakovsky et al. (2015)) achieved a classification error of around 20% in the task to localize 1000 object categories. This is an impressive development keeping in mind that the best shallow submission to same challenge happening in 2012 only achieved an average top-5 classification error of around 50% (reporting the top-1 error was not necessary at ILSVRC2012). Nonetheless, one should keep in mind that the real world contains much more objects than covered by ILSVRC, and it can still take years until we have accurate general purpose object detectors.

In the context of activity recognition, fine-grained knowledge about the observed objects(*e.g.*, `opened` vs. `closed` fridge) contains even more information than only knowing that the object is present. Such information can usually not be obtained from current object detectors. When knowing the target domain (*e.g.*, evaluation data set, or application scenario) of an activity recognition system, one could alternatively build a dedicated object detector using data from said domain. This, however, implies that besides of having to record training data for the activity recognition system, the much more time and cost expensive manual annotation of present objects would be required as well.

To circumvent such shortcomings, methods that can automatically extract potentially relevant image regions have been explored for activity recognition. In fact, one of the earliest attempts to jointly consider actions and objects for activity recognition (see Moore et al. (1999)), uses object information from both, supervised and unsupervised sources. The actions are inferred from hand trajectories using


Input video
a. Region proposals
s_i
v_i
b. Spatiotemporal segments pool
c. Spatiotemporal segments hierarchy

(a) Ikizler-Cinbis and Sclaroff (2010) (©2010 Springer)

(b) Lan et al. (2015a) (©2015 IEEE)

Figure 2.18. *Examples of object candidate regions detected in an unsupervised fashion and used for activity recognition. (all images are courtesy of the respective authors)*

HMMs, and the objects are recognized by means of template matching. Additionally, regions obtained from background subtraction that cannot be matched to known object categories are included in the template directory with the label `Unknown`. All three cues are finally joined in a Naïve Bayes classifier framework to recognize activities. Ikizler-Cinbis and Sclaroff (2010) integrate object, scene, and person information in a multiple instance framework for action recognition. The approach does not rely on any object annotations, but similar to the aforementioned approach assumes object candidates to be large moving regions. The major drawback of this assumption is that it only considers objects that are directly manipulated by the person. However, many background objects are often also relevant for the activity in questions, but those are neglected by this approach. Imagine for example a typical cooking activity; typically most of the ingredients are being static, while only a few are being handled (*e.g.*, cut, peeled, stirred, etc.) at any given time. Keep also in mind that during most of the actions the manipulated object may easily be occluded by the hands. Therefore, a more generic method for unsupervised objects region extraction would be desirable.

The easiest way to do so is by assuming all image segments to be viable object candidate regions, as has been done by Aksoy et al. (2011). Here, the image segments are computed with super-paramagnetic clustering in a spin-lattice model (*cf*., Dellen et al. (2009)) and their temporal coherence is ensured by incorporating in neighboring frames in the clustering process. Abramov et al. (2010) have further extended this approach to spatial 3D by means of stereo matching. Using all image segments as object candidates may be a suitable solution in controlled environments. However, realistic scenarios often contain significant amounts of image clutter resulting in the detection of too many image segments. This, in turn, deteriorates the quality (and thus discriminative power) of the object candidates. As a solution to this problem, we, therefore, propose in our work the usage of visual saliency to select the segments that resemble object, object parts the most (see Chapter 5).

An alternative approach has been proposed by Lan et al. (2015a) with the mid-level action elements (MAE) representation. It captures motion, objects, body parts, and their interactions in a supervised fashion.

The method uses the algorithm from Endres and Hoiem (2010) to mine regions of object-like appearance, and motion distinctive from the background. Next, discriminative (*i.e.*, semi-supervised) clustering is leveraged to discover the MAEs, spatio-temporal regions that are representative of the activities in the training set. Parameters of the activity recognition model are finally learned in a structured SVM framework.

Not directly related to activity recognition but still relevant to our work are approaches that infer actions from a single image based on human pose and the presence of objects. For example, Yao and

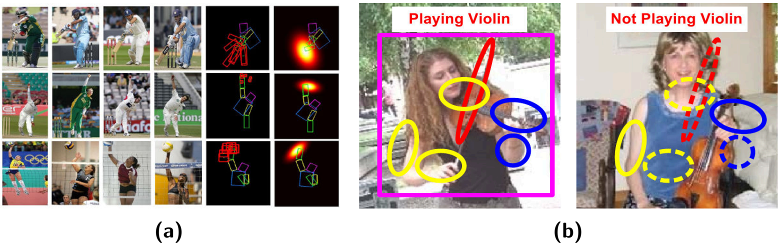


Figure 2.19. *Examples of spatial configurations between body parts and object regions learned to recognize actions from single images: (a) The human object-interaction descriptor proposed by Hu et al. (2015) (©2015 IEEE). (b) The grouplet representation from Yao and Fei-Fei (2010b) (©2010 IEEE).*

Fei-Fei (2010b) propose a method to encode structured information in images which is based on a data mining method incorporated with a parameter estimation step to discover discriminative groups of image patches, the grouplets. Example grouplets discriminating between the act of playing and holding a violin can be found in Fig. 2.19(b). Prest et al. (2012b) also model human actions in terms of spatial configurations between humans and objects. The approach employs, however, explicit human detection as well as an unsupervised objectness measure (*cf*., Alexe et al. (2012)) to determine the most action relevant region that is located close to the actor. The limitation of this type of approach is that manipulated objects are assumed always to be found at a specific location relative to the body. This assumption may hold true in the typical application scenarios of single-image based action recognition approaches, like in the context of playing musical instruments or performing sports. However, in the case of complex activities, it would often be violated when the spatial relations between objects and humans are somewhat arbitrary (*e.g.*, pick and place operations).

Another related, but slightly different, problem is formulated by Srikantha and Gall (2014): discovering objects from activities. To this end, human-object interaction video samples are treated as weak labels and used to infer the type and location of an object that is part of the interaction. First, spatio-temporal regions (tubes) are selected as object proposals, while the human body pose is assumed to be known. Then, an energy minimization algorithm is used to select one tube per video that most likely corresponds to the object in question. A combination of several quality measures for the objectness of a tube is considered jointly, such as appearance dissimilarity with the background, relation to the human body pose, similarity of the shape of tubes, or the correlation of the tube with human motion.

2.4 Discussion

Since the focus of the presented approach lies in mining object candidates for activity recognition and not in the motion representation itself, we want to make an existing framework the foundation of our work. To this end, we have reviewed at the beginning of this chapter different representations for action and activity recognition.

Holistic approaches, *i.e.*, methods that model the observed actions as a whole, are historically among the first representations that were used for action and activity recognition. They do not require the localization of any body parts, but are rather only based on either human silhouettes, or optical flow estimation. This makes such methods, in general, more robust and computationally efficient. Nonetheless, they suffer from many problems resulting in the Computer Vision research community nearly completely abandoning holistic representation in their original form. Example drawbacks of holistic methods are their general lack of invariance to camera view direction, heavy reliance on clear human silhouettes, and the difficulty to model humans appearing in different scales. Furthermore, they often fail to properly represent the complex structure of high-level activities.

In contrast, human body model based approaches suffer from none of the problems that are inherent to holistic methods. They rely, however, on a high-quality reconstruction of the human pose, which is difficult to obtain from 2D data in real-time. Of course, one can resort to consumer electronic depth sensors like the popular Microsoft

Kinect, which have brought a revival to body model based action and activity recognition. Yet, it is often not feasible to adopt such active sensors, which is, for instance, the case for the robotic platform at which our proposed approach is aimed for.

Biologically inspired methods have recently gained an especially high amount of attention in practically every field of computational pattern recognition. This development has been mainly caused by the progress of parallel graphical processing units making it computationally feasible to train deep artificial neural networks, and the resulting significant improvement over previous art when employing ConvNets. Since end-to-end deep learning methods represent methods as a whole, they can be seen as a special case of holistic representations, yet without having inherited their drawbacks. Unfortunately, the performance gain achieved with ConvNet-based approaches is only insignificant in comparison to state-of-the-art. Furthermore, the training process of ConvNets is very time consuming and thus it can take several weeks when using conventional hardware to learn an action recognition model from a suitable data set. Therefore we refrained from employing a ConvNet-based motion representation in favor of the previously popular local feature methods.

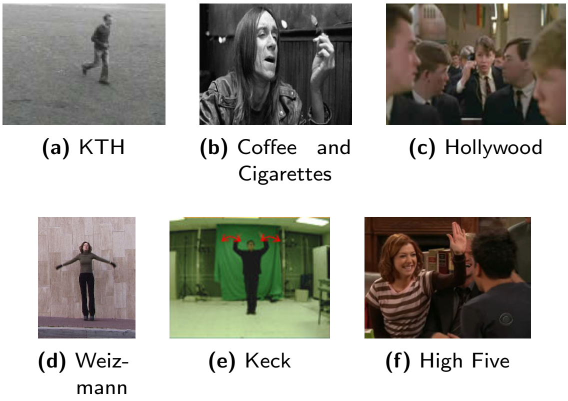
Local feature based representations share many properties with deep learning approaches (*e.g.*, pooling methods, similar feature representation). The main difference is, however, that they are fully handcrafted, as opposed to deep learning methods that learn the feature representation directly from training data. Even though they lost their status of being the core of most state-of-the-art algorithms to deep learning, they still yield a competitive performance in human motion analysis, but are much faster to train. For instance, the best pure ConvNet-based approach (*cf.*, Feichtenhofer et al. (2016)) achieves a classification accuracy of 92.5% and 65.4% on the two most commonly used action recognition benchmarks, UCF-101 and HMDB-51, respectively. This compares quite favorably with the pure local feature based system developed by de Souza et al. (2016), which uses iDT features encoded with augmented Fisher Vectors, and achieves a classification rate on same data sets of 90.6% and 67.8%, respectively.

For this reason, we have decided for a space-time interest point representation of human motions. More precisely, we investigate in this work the impact of the proposed object candidate features on the activity recognition results when using two of the most popular spatio-temporal local features, *i.e.*, HOGHOF encoded Harris3D interest points (from here on referred to as STIP) from Laptev and Lindeberg (2003), and improved Dense Trajectory features (iDT) developed by Wang and Schmid (2013).

Besides of motion representation, we have as well surveyed in this chapter methods that combine motion features for a video-wide representation of human activities. In general, these methods can be divided into two categories: structure preserving graphical models, and unstructured approaches. Since human activities have a highly structured nature (*i.e.*, in a sense that they are sequences of simple motions, and can be performed at different locations, as well as by different body parts), employing a structured model would be most intuitive. Nonetheless, approaches built upon unstructured methods, especially the Bag-of-Words model, have shown to yield very high action and activity recognition results and are still among the best on most benchmarks. Following the principle of Occam's razor (*cf.*, Gauch Jr. (2003)), we have opted for the much simpler of both video representation types, *i.e.*, the unstructured BoW model.

Bag-of-Words models have been greatly studied in conjunction with local spatio-temporal features for action recognition; usually together with STIP or iDT features to encode motion. A collection of these comparative analyses in the context of action recognition can be found in Tab. 2.2. Based on their characteristics, BoW encodings can be categorized into three groups, namely voting-based encodings (*e.g.*, VQ, soft assignment), reconstruction-based encodings (*e.g.*, sparse coding, LLC), and supervector-based encodings (*e.g.*, FV, VLAD). In order to allow a broader analysis of our contribution, we decided to select one representative from each category as our basic mid-level video representation.

3 Benchmark data sets



(a) KTH

(b) Coffee and Cigarettes

(c) Hollywood

(d) Weizmann

(e) Keck

(f) High Five

Figure 3.1. *Sample frames from benchmarks aimed at the recognition of simple actions: (a) KTH actions (Schüldt et al. (2004)). (b) Coffee and Cigarettes (Laptev and Pérez (2007)). (c) Hollywood actions (Laptev et al. (2008)). (d) Weizmann actions (Blank et al. (2005)). (e) Keck gestures (Lin et al. (2009)). (f) High Five (Patron-Perez et al. (2010)).*

The first step in developing any pattern recognition system is to acquire an adequate data set, that can be used for training purposes as well as a benchmark to compare different approaches to each other. Comprehensive surveys covering several aspects of data sets created

in the context of human action and activity recognition have been compiled by Chaquet et al. (2013); Hassner (2013). With the evolution of approaches aimed at this pattern recognition domain, these data sets have evolved as well. Starting with data sets containing only few, simple, and often staged actions that were recorded in controlled environments (*e.g.*, actions performed by members of a research lab), the development advanced to realistic data sets aimed at specific applications, *e.g.*, recognizing activities of daily living.
In the following, we give a brief overview of data sets compiled in mind with the comparison of recognition approaches (see Sec. 3.1), before we cover a selection of data sets that are mostly relevant for our work, *i.e.*, recognizing human-object interaction activities from 2D image sequences. Since none of the described benchmarks could fully satisfy the needs of our application scenario (*i.e.*, activity recognition for a humanoid household robot), we have recorded the KIT Robo-kitchen data set, which we describe in detail in Sec. 3.3. A discussion about the pros and cons of available activity recognition benchmarks, and consequent motivation behind our selection of data sets to evaluate the approach proposed in this work concludes this chapter.

3.1 Action recognition data sets

Two very early action recognition data sets that have been used for a very long time as de-facto standard benchmarks are **KTH actions** from Schüldt et al. (2004), and the **Weizmann action** data set from Blank et al. (2005). Both contain only few and relatively simple, periodic actions, such as `running` or `boxing` that are performed in very constrained environments and do not contain much intra-class variation (see Fig. 3.1(d) and Fig. 3.1(a) for sample shots).
Another rather simple data set is **Keck gestures** created by Lin et al. (2009) (see Fig. 3.1(e)). It contains 14 categories of military signals and has been recorded in a lab environment. Sequences used for training were recorded with a static camera and a uniform background, while testing samples were collected in a more difficult

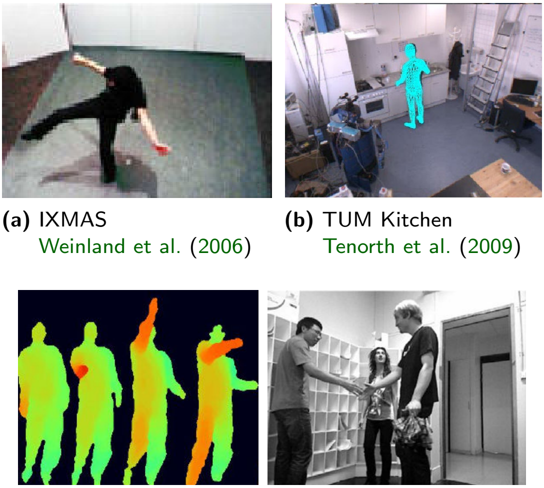
scenario containing background clutter and a moving camera.

The **IXMAS** (Weinland et al. (2006)) and **HumanEva** (Sigal et al. (2010)) data sets are conceptually similar to the aforementioned ones in the sense of a simplified setting. However, they were recorded using a multiple camera setup making the data also suitable to evaluate approaches aiming at view-independent action recognition, which is still a very challenging topic.

All of these data sets are of limited relevance to practical applications since the contained actions are composed of distinct movements often making them appear unnatural. Also, the recorded actions have a lack of variability in body postures when being compared to the same actions performed in the context of daily living activities. These shortcomings prompted the development of data sets containing more natural and complex actions which were recorded in a realistic environment.

Because it is difficult for people to act naturally when participating in an artificially set data collection, Laptev and Pérez (2007) proposed to collect more suitable data from movies instead. This development prompted in the creation of such data sets like **Coffee and Cigarettes** (*cf*., Laptev and Pérez (2007)), **Kissing/Slapping** (*cf*., Rodriguez et al. (2008)), and **High Five** (*cf*., Patron-Perez et al. (2010)), all allowing the evaluation of detectors discriminating between visually similar actions.

The **Hollywood human actions (HOHA)** data set is composed of movie scenes as well, however with the intention to evaluate approaches for action classification. The second version of HOHA, usually referred to as **Hollywood2** has been subsequently established by Duchenne et al. (2009) and is currently still commonly used to evaluate action classification approaches. The data set contains approximately 20 hours of video data collected from 69 movies with 12 categories of simple actions, from which around 800 sample clips were randomly chosen to constitute the training set and 800 taken from other movies for testing. Additionally, 800 action samples were mined automatically using video-to-data alignment as described by Laptev et al. (2008) and can be used as supplemental training data with noisy labels. Furthermore, Marszałek et al. (2009) observed the



(a) IXMAS
Weinland et al. (2006)

(b) TUM Kitchen
Tenorth et al. (2009)

(c) MSR Action 3D
Li et al. (2010b)

(d) LIRIS
Wolf et al. (2014a)

Figure 3.2. *Sample shots taken a selection of data sets allowing action recognition from depth data. The data sets depicted in the top row were recorded with a multiple camera setup, while the ones in the bottom row with a Microsoft Kinect.*

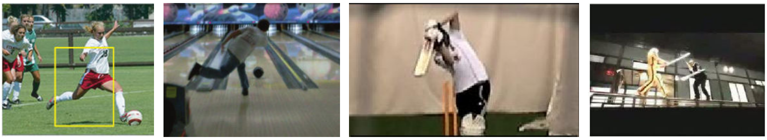
importance of context to discriminate actions and therefore provided scene annotations belonging to ten different categories. Example shots from the movie-based benchmarks can be found in Fig. 3.1.
Liu et al. (2009) propose to make use of a different source of videos to establish a set of realistic and very diverse action recognition samples recorded *in the wild*: home-videos published on YouTube. Their **YouTube action** data set (also referred to as UCF-11) contains interaction events between humans, actions involving object manipulation, and much variability with respect to viewing angles, lighting, background, and actors. In the following years, this benchmark has been twice extended in the form of UCF-50 (*cf*., Reddy and Shah (2013)), and UCF-101 (*cf*., Soomro et al. (2012)), spanning over 50 (and 101 respectively) action categories. At the moment of this

writing, **UCF-101** constitutes one of the two benchmarks being the golden standard to evaluate action recognition methods, with correct classification rates currently being in the range around 90% (*e.g.*, Feichtenhofer et al. (2016); Wang et al. (2015a)).

The other golden standard benchmark is the human motion database (**HMDB-51**) created by Kuehne et al. (2011). As the name suggests, it comprises 51 action categories, which all were collected from YouTube. Even though quite similar to UCF-101, HMDB-51 appears to pose a more difficult challenge as the highest reported correct classification rates are currently around 65% – 70% (*e.g.*, Feichtenhofer et al. (2016); Peng et al. (2014c)). Jhuang et al. (2013) selected a subset of 21 categories from HMDB-51 and annotated for each frame skeleton joints using a 2D articulated puppet model. The resulting data set has been released under the name joint-annotated HMDB (**J-HMDB**) with the intention to provide researchers with means for the understanding which parts of their algorithm affects action recognition performance the most.

All of the data sets described so far are only capturing the limiting set of simple repetitive (*e.g.*, `walking` or `waving`) and punctual actions (*e.g.*, `hugging` or `opening a door`). However, many interesting human actions are of a more complex nature. Some researchers resort therefore to collect samples from various sports actions featured on broadcast television channels. This resulted in the creation of the **UCF-sports** data set (*cf.*, Rodriguez et al. (2008)) including a total of 150 samples and 10 action categories, and the larger **Olympic sports** data set (*cf.*, Niebles et al. (2010)) covering 16 sports classes with 50 samples per class. The advent of deep learning methods also created a demand for a very large set of training samples resulting in the creation of the **Sports-1M** data set by Karpathy et al. (2014). It is so far the largest action recognition data set consisting of more than one million sports action samples from 487 categories.

The release of low-cost consumer electronics depth-sensors (*e.g.*, Microsoft Kinect) resulting in an increased research interest in action recognition from RGBD data made it necessary to create appropriate benchmark data sets as well. This gave rise to data sets like MSR Action 3D created by Li et al. (2010b), and LIRIS human activi-



(a) UCF Sports **(b)** Olympic Sports **(c)** UCF-101 **(d)** HMDB-51

Figure 3.3. *Sample frames from action recognition benchmarks consisting videos collected from YouTube, and real movies: (a) UCF Sports (Rodriguez et al. (2008)). (b) Olympic sports (Niebles et al. (2010)). (c) UCF-101 (Soomro et al. (2012)). (d) HMDB-51 (Kuehne et al. (2011)).*

ties from Wolf et al. (2014a) (see Fig. 3.2 for example shots). The action categories contained in the **MSR Action 3D** data set were chosen in the context of using actions to interact with video-game consoles and are, therefore, very simple (*e.g.*, `golf swing` or `forward punch`). In contrast, the **LIRIS human activities** data set has been specifically designed for the problem of recognizing complex actions (*e.g.*, interactions with several participants) in a realistic surveillance setting and in an office environment.

Other notable data sets aimed at the evaluation of different aspects of action recognition related systems are the TUM kitchen, and action similarity labeling (ASLAN) data sets. The **TUM kitchen** data set created by Tenorth et al. (2009) only consists of one high-level activity class (`setting a table`) aiming at the evaluation of fine-grained action detection systems. The videos were recorded in overhead views from four different angles (*cf.*, Fig. 3.2(b)), and data from other types of sensors (*e.g.*, RFID, magnetic reed sensors to detect when a door/drawer is opened) is provided as well. Additionally, human pose data obtained from a markerless body tracking system is provided as well. The **ASLAN** data set created by Kliper-Gross et al. (2012a) aims at a completely different task; deciding whether two given video samples belong to the same class or not.

(a) CMU-MMAC de la Torre et al. (2008)

(b) OPPORTUNITY Roggen et al. (2010)

(c) Poeticon Wallraven et al. (2011)

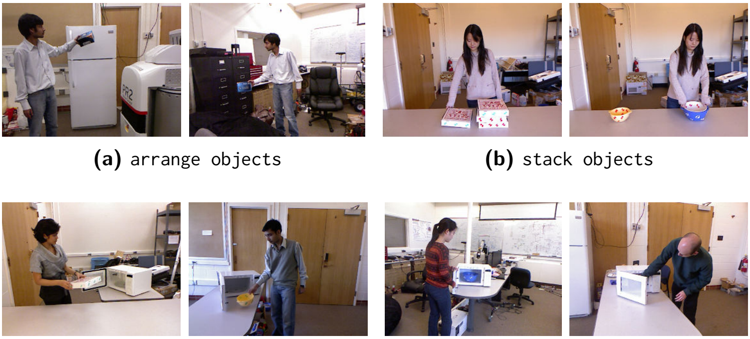
Figure 3.4. *Sample frames from data sets that are mainly aimed at using multiple (intrusive) sensor modalities for activity recognition.*

3.2 Activity recognition data sets

Among all possible applications of human activity analysis, the recognition of activities of daily living (ADL) has emerged as one predominant trend. Possible reasons for this development are the increasing interest in creating more natural human-machine interfaces, as well enable an automated monitoring of elderly people, or dementia patients.

Towards achieving this goal, many data sets have been established, each posing a different set of challenges. In the following, we present some of these activity recognition data sets, all of which we deem the most relevant to our work. A table summarizing some properties of the reviewed data sets can be found in Tab. 3.1.

As the name suggests, the **Objects in Action** data set (commonly referred to as **Gupta data set**) has been created by Gupta and Davis (2007) with the intention in mind to compare approaches explicitly incorporating object knowledge for activity recognition. Being the first of its kind, this data set is rather small (*e.g.*, consisting of only 54 samples of 6 different categories), the performed actions appear staged, and it has been recorded in a very controlled environment. Nonetheless, it poses some challenges, especially the discrimination of activity categories that are characterized by similar body motions, but consisting of the manipulation of different objects (*e.g.*, `pouring a can` vs. `using a flashlight`).



(a) `arrange objects` **(b)** `stack objects`

(c) `microwave food` **(d)** `take food`

Figure 3.5. *Sample frames from four (out of ten) activity categories of the CAD-120 data set (*cf.*, Koppula et al. (2013)). Note, how samples belonging to the same category were recorded in different environments, and may involve different objects (*e.g.*,* `stacking boxes`*, and* `stacking bowls`*). Furthermore, some activity categories share the same set of manipulated objects (*e.g.*, microwave).*

In contrast, the Carnegie Mellon University multimodal activity (**CMU-MMAC**) database has been created to capture human behavior in settings that are as natural as possible (*cf.*, de la Torre et al. (2008)). Other than cameras, a diverse set of sensors has been employed for the recordings, *e.g.*, accelerometers, microphones, marker-less motion capturing of one participant (VICON), RFID, skin temperature, galvanic skin response sensors. Employing this many sensors comes however with a price - most of them are intrusive, *i.e.*, they are attached to the human body and thus can easily obstruct the natural realization of some motions.
The setting of the recordings is a full kitchen setup, where each of the 39 subjects has been asked to prepare five different dishes, *e.g.*, `brownies`, `pizza`, or `sandwiches`. No further information has been provided in how to cook the dishes in order to make the subjects behave as natural as possible. Additionally to the high-level activity

category labels, Spriggs et al. (2009) provided for 16 subjects annotations of several involved sub-activities in the form of $<$*Verb*$>$ $<$*Object* $><$*Preposition*$><$*Object*$>$, *e.g.*, `take egg from fridge`, or `open can`.

The University of Rochester activities of daily living (**URADL**) data set from Messing et al. (2009) involves kitchen activities as well, however of much simpler nature. It consists of five subjects, each performing ten different activities three times in front of a high-resolution camera (1280x720 px) facing the subject. The categories were selected having an assisted cognition task in mind, and with the goal to capture activities that are difficult to separate based on a single cue. Example categories include `eating`, `peeling` or `cutting a banana`, as well as `dialing` or `answering a phone` (see Fig. 1.4 for more examples). Given its medium size (150 samples in total), the moderate level of challenges, and the complexity of the performed tasks being on the higher side, this data set has become one of the standard benchmarks for activity recognition systems.

Unlike any other data set described in this section, the focus of the **OPPORTUNITY** data set from Roggen et al. (2010) lies not in activity recognition from video, but rather from a set of 72 environmental, body, and object sensors of 10 modalities. Nonetheless, some image sequences have been captured during the recordings as well with the purpose to facilitate data annotation.

The data set contains around two hours of recordings per subject. Each of the twelve subjects performs a sequence of five different high-level ADL, namely `standing up`, `preparing breakfast`, `having breakfast`, `cleaning up`, and `having a rest`. Additional to the ADL runs, a drill run is provided where the participants were asked to repeat 20 times a sequence of simple actions, such as `opening and closing a door`, or `drinking`.

In one of our previous works, we have recorded the motion-primitive, intention, and activity recognition (**MINTA**) data set (*cf*., Gehrig et al. (2011)) aimed at the humanoid household robots application scenario. It contains recordings of six activity classes that are performed by each of the ten subjects ten times and also includes annotations of 60 temporally very fine-grained motion-primitives. It

(a) first person ADL **(b)** 50 salads **(c)** Breakfast **(d)** MPII Cooking 2

Figure 3.6. *Sample frames from activity recognition benchmarks recorded under realistic settings: (a) Fist person ADL (Pirsiavash and Ramanan (2012)). (b) 50 salads (Stein and McKenna (2013)). (c) Breakfast actions (Kuehne et al. (2014). (d) MPII Cooking Activities 2 (Rohrbach et al. (2015)).*

has however been set up in a very simplified way resulting in an unrealistic scenario.

The **Poeticon** enacted scenario corpus was created by Wallraven et al. (2011) with having in mind to provide a realistic data set of complex, long-lasting activity sequences that also include interactions between humans. It comprises of six everyday scenarios taking place in a kitchen/living-room setting which are simultaneously recorded from five different angles.

Each of the activities is performed three times by four different pairs of actors, and is based on a script which is rehearsed before the recordings. Since the room in which it all takes places also resembles more a theater stage than a real living room, the achieved level of realism is limited. Information about key objects and sub-activities, are additionally provided together with data from a VICON motion capturing system, as well as kinematic recordings obtained from an inertial sensor based motion-capture Moven suit from Xsens Technologies.

Activity recognition from first-person views achieved through wearable cameras has recently become a very active research area. Therefore, many data sets have been developed in this context as well, in order to provide the Computer Vision research community with challenging benchmarks (*e.g.*, Fathi et al. (2011); Hanheide et al.

(2006); Lee et al. (2012); Sun et al. (2009b); Sundaram and Cuevas (2009)). Since this topic is not directly related to our work, we only want to exemplary emphasize on one of the most popular first-person activity recognition data sets, activities of daily living (**ADL**) by Pirsiavash and Ramanan (2012).

The set of recorded activities has been selected based on medical literature on rehabilitation in order to capture as good as possible the basic movements a person is undergoing while performing everyday functions, such as eating, maintaining personal hygiene, or entertainment. In total, 18 unscripted activity categories are each performed by 20 different subjects in their home environment. Additionally to the video recordings, the data set contains annotations of 42 object categories in form of identity, bounding box, and information whether the object is currently being manipulated.

The MSR daily activity 3D data set (**MSRActivity3D**) has been captured by Wang et al. (2012b) with a Microsoft Kinect camera, and thus also contains depth information as well as the reconstructed 3D skeleton of the actors. The 16 activity types were chosen to capture typical human activities taking place in a living-room, and often involve object manipulations. Example activity categories are `drinking`, `eating`, `reading, a book`, `using a vacuum cleaner`, or `standing up`, all of which are rather simple and short, *i.e.*, around three to twelve seconds long. Furthermore, the recordings took place in a lab environment with simple and static backgrounds resulting in a comparatively easy benchmark.

The Cornell activity data set **CAD-120** created by Koppula et al. (2013) has also been recorded with a Kinect camera, and is a successor to **CAD-60** (*cf*., Sung et al. (2011)). It contains 124 activity sequences of ten different high-level activities, each performed three times by four subjects. The high-level activities are: `preparing cereals`, `cleaning objects`, `picking objects`, `taking food`, `having a meal`, `microwaving food`, `taking medicine`, `arranging objects`, `stacking objects`, and `unstacking objects`.

During the recordings, the subjects were only given a high-level description of the task, and were asked to perform the activities multiple times, each time with different objects. For example, the stacking and unstacking activities were performed with pizza boxes, plates, and bowls (see Fig. 3.5 for example shots).

Due to the brief description of the tasks, the sequences also vary significantly from subject to subject in terms of length and order of the involved sub-activities. Further challenges have been imposed by not always recording the activities in the same setting, and often having a very cluttered background.

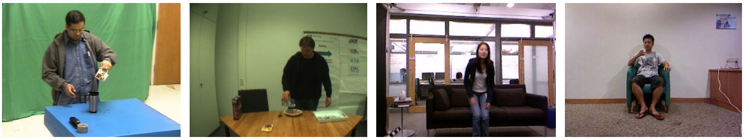
The **50 salads** data set has been created by Stein and McKenna (2013) and consists of 27 participants, each preparing a salad. These sequences can be further decomposed into three types of high-level activity classes: `preparing salad`, `preparing dressing`, `dressing and serving a salad`. Since the main task of this data set is providing data for fine-grained action detection, annotations of sub-activities are provided as well.

The recordings were performed with a top-down camera, as well as Kinect devices, and accelerometers that have been attached to the manipulated objects (see Fig. 3.6(b)). In order to increase the variance of the recorded data, participants were asked to follow certain steps in the salad preparation. They were, however, free to decide which objects to use, *e.g.*, whether the salad dressing should be prepared in a cup, or directly in the salad bowl.

Kuehne et al. (2014) have created with their **Breakfast** data set one of the currently largest fully annotated benchmarks for fine-grained activity recognition. The provided annotations thus not only describe the high-level activities, but also sub-activities, and even action-primitives. For instance, the action of `pouring milk` is further decomposed into finer chunks, like `grabbing milk`, `twisting the cap`, `opening the cap`, etc.

The recordings were performed by three to five cameras (depending on the location) and involved 52 participants, each conducting ten cooking activities in their home or office kitchens. The goal was to create a recording setup that closely reflects real-world conditions and, therefore, it not only took place in a natural environment but was fully unscripted. The only thing the participants were told was to prepare a certain dish, *e.g.*, cereals, coffee, tea, or a sandwich.

The MSR action recognition on online RGBD (**ORGBD**) action



(a) Objects in Action **(b)** MINTA **(c)** MSR Activity 3D **(d)** ORGBD

Figure 3.7. *Sample frames from activity recognition data sets, that were recorded under simple, and unrealistic conditions: (a) Objects in Action (Gupta and Davis (2007)). (b) Motion, Intention, and Activity data (Gehrig et al. (2011)). (c) MSR Activity 3D (Wang et al. (2012b)). (d) MSR Online RGBD Actions (Yu et al. (2015)).*

data set created by Yu et al. (2015) is the first benchmark for cross-environment and online activity recognition with depth sensors. It consists of three sets of depth sequences collected by using a Kinect device. The first one is designed for activity recognition in the same environment, and the second one recorded in a different setting from the first one is meant for cross-environment recognition. In the third set, each video contains multiple unsegmented activities.

Each set contains seven different categories of activities (plus one negative class consisting of random motions) that people usually perform in a living-room, like using a remote control, drinking, or picking up a phone. Since all of them involve manipulations of objects, object bounding box and identity labels are also provided in the training data. Having been created by the same research group that recorded MSRDailyActivity3D, it is set in a similar artificial environment and the activities appear unnatural as well.

Borreo et al. (2015) try to distinguish their multi-environment action data set (**MEA**) from other benchmarks by providing a multi-environment structure. To achieve this goal, five types of ADL are recorded in two different domestic environments, one resembling a kitchen, and the other a living-room. Unfortunately, the videos have been acquired with a camera embedded in a smartphone (iPhone 4) making it very difficult to reconstruct the full body pose. Without pose information, the motions recorded in the two settings differ too much from each other making the cross-environment recognition task close to impossible (*i.e.*, unless training data other than the provided is allowed to be used, *e.g.*, for zero-shot learning). This is reflected in the cross-environment activity recognition accuracy of the baseline system, which ranges around 20%, corresponding to random guessing of the five categories.

The **MPII Cooking Activities data set 2.0** has been introduced by Rohrbach et al. (2015) and is an extension of the **MPII Cooking Activities** (*cf.*, Rohrbach et al. (2012a)), and the **MPII Cooking Composite Activities** (*cf.*, Rohrbach et al. (2012a)) data sets. The underlying idea of its creation was to promote the development of approaches addressing the detection of fine-grained sub-activities and understanding how they are connected to high-level activities. It is set in a kitchen scenario and covers a range of typical kitchen activities which can be as simple as `sharpening a knife`, or as complex as `preparing a pizza`.

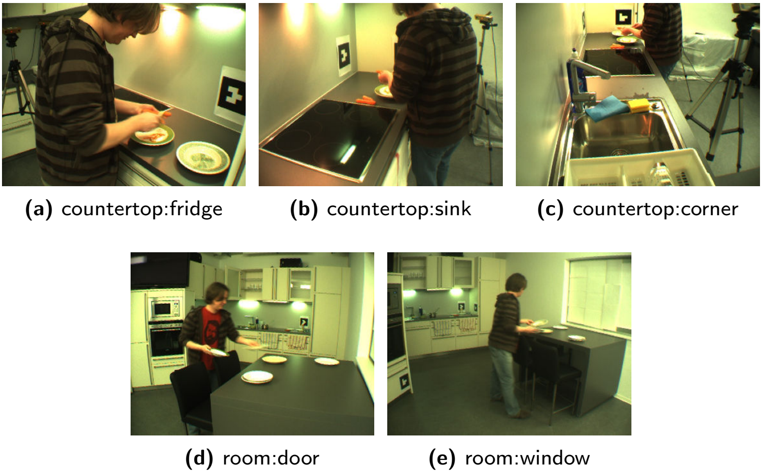
In total, it covers 59 activities which are performed by 30 different subjects resulting of 273 high-resolution (*i.e.*, 1624x1224 px) video sequences. Besides of activity annotations, 222 attribute labels for sub-activities and objects are provided. A particular challenge of this data set is that several activities are quite similar, like `preparing broccoli` vs. `preparing cauliflower`.

3.3 The KIT Robo-kitchen data set

After surveying all publicly available activity recognition benchmark data sets, we came to the conclusion that none of them fully satisfied our needs. Therefore, we have created the KIT Robo-kitchen data set (*cf.*, Rybok et al. (2011)) capturing the diverse challenges that can occur in the humanoid household robot domain. Our goal was to capture complex, long-lasting, quasi-periodic, and realistic kitchen activities, as opposed to data sets aimed at the high-level analysis of human motions.

Name	year	classes	subjects	repeats	clips	res (px)	fps	views	duration	3D data
Gupta	2007	6	10	1	54	640x480	15	1	~4-12s	none
CMU-MMAC	2008	5	39	1	870	1024x768	30	6	~3-7min	VICON
URADL	**2009**	**10**	**5**	**3**	**150**	**1280x720**	**30**	**1**	**~10-60s**	**none**
OPPORTUNITY	2010	5	12	1	60	N/A	N/A	3	N/A	none
MINTA	2011	7	10	10	688	640x480	30	1	~20-90s	none
Poeticon	2011	6	8	3	120	1960x1400	25	5	~2-7min	VICON
Robo-kitchen	**2011**	**12**	**27**	**1**	**540**	**640x480**	**15**	**2-3**	**~10-240s**	**stereo**
first person ADL	2012	18	20	1	360	1280x960	30	1	~9-350s	none
MSR Activity 3D	2012	16	10	2	320	640x480	15	1	~3-12s	Kinect
CAD-120	**2013**	**10**	**4**	**3**	**124**	**640x480**	**30**	**1**	**~5-40s**	**Kinect**
50 salads	2013	3	27	2	162	640x480	30	1	~0.5-4.5min	Kinect
Breakfast	2014	10	52	1	1721	320x240	15	3-5	~0.5-6min	stereo
ORGBD	2014	8	24	1-2	280	640x480	30	2	~9-115s	none
MEA	2015	5	17	1	124	1920x1080	30	1	$\sim 1min$	Kinect
MPII2	2015	59	30	N/A	273	1624x1224	29.4	1	~0.5-41min	none

Table 3.1. *Quantitative comparison of most publicly available activity recognition benchmark data sets, sorted by year of publication.Benchmarks highlighted in bold are used throughout this work for the experimental evaluations.*

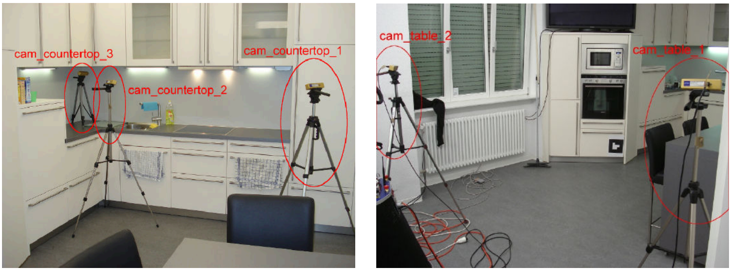


(a) countertop:fridge (b) countertop:sink (c) countertop:corner

(d) room:door (e) room:window

Figure 3.8. *Sample images taken from videos of the KIT Robo-kitchen data set recorded from all five viewpoints.*

Furthermore, the recording setup has been designed to resemble as closely as possible one of the household robot ARMAR III (*cf*., Asfour et al. (2006)), since the main motivation for this data set was driven by applications aimed at this specific robot. All of this poses many challenges for view-based activity recognition approaches, such as difficult lighting conditions, cluttered background, (self-) occlusions, different viewpoints, and a limited field of view. Most importantly, we barely restricted the way how the recorded subjects had to perform the activities resulting in a collection of natural motions with much variation as opposed to most currently publicly available data sets. Imitating humanoid robots in our setup also results in the use of stereo cameras (at a resolution of 640x480 px), which can be beneficial for activity recognition, since it allows for person tracking, and extraction of motion trajectories in 3D. It is also expected that the depth information will improve activity recognition, since it allows to infer the 3D position of people in the room, which is a strong prior on the likelihood of specific activities.



(a) counter-top camera setup

(b) room camera setup

Figure 3.9. *Locations of the cameras used for the recordings of the KIT Robo-kitchen data set. Since both setups were directed at different parts of the kitchen, not all captured activity categories are same for both setups.*

The cameras were positioned at different locations in the room that are easily accessible by a robot platform. The use of multiple viewpoints allows for the evaluation of activity recognition approaches aiming at achieving robustness to view changes.

Two different camera setups have been used as shown in Fig. 3.9, one focusing on activities performed on the counter-top, and the other capturing activities taking place in the whole room area. Our reasoning behind using two setups is application driven: when people occlude the area where the activity takes place with their body when viewed from the room setup, the robot should shift his location to a more suitable one. This is, for instance, the case when the cooking activities are performed at the counter-top. Example images captured with each of the cameras used in both setups can be seen in Fig. 3.8, and representative shots of all activity categories in Appendix A.

One of our main goals was that the activities were performed as natural as possible. Thus, we provided the participating subjects only with brief information about the recorded activities. Among the activity descriptions were explanations where to find the required objects, for how many people to set the table, and to perform some activities at a location of their choice at the table.

For the activity of setting the table, we also provided the participants with the sketch shown in Fig. 3.10 to give them an idea which items to use when setting the table. The intention was to motivate the participants to use a more complicated setup and therefore increasing the complexity of the involved motions.
Each activity has been performed once by 17 subjects of different age, gender, cultural background, and household skills in order to capture a high amount of variation, as opposed to having only a few actors repeating the activities several times. The duration of a video sequence varies between 10 seconds and 4 minutes, depending on the complexity of the activity, and the thoroughness of the subject.
Using the counter-top setup, we recorded seven different activities, which are described together with their canonical names in Tab. 3.2. All of the activities have been recorded from three different viewpoints at the same time, with the exception of `wash`, and `dry` because the camera in front of the sink had to be removed in order to allow access. It should be noted, that one of the cameras cannot be reached by a robot platform. However, since achieving robustness to view changes in activity recognition is an important, but still open topic, it has been added to the setup. Samples from the resulting views are given in Fig. 3.8 (a)-(c).

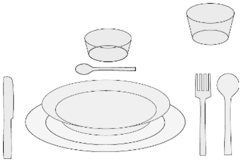


Figure 3.10. *A sketch outlining the setup of cups, plates, and silverware for the* `settable` *activity of the KIT Robo-kitchen data set. The sketch has been shown to all subjects prior to the recordings as a suggestion which objects to use while performing the activity in order to encourage them to perform more a more complex sequence of sub-activities while setting the table. Note, that this did not mean that the subjects were strictly following the provided setup.*

The recordings using the room setup are meant to model one of the primary applications of activity recognition for humanoid household robots. The key idea is that the robot takes the role of a servant observing the scene from a place where he has a good view over the room and offer his help proactively if he assesses it might be required. Situation understanding is also important for the robot when entering a room in search for a new task to be performed.
Note, that only two camera views were used for the room recordings, but the positions of both are easily reachable by a robot platform. Figures 3.8 (d)-(e) contain examples of the field of view of the cameras used in this setup, and Tab. 3.3 a list of the recorded activities. Many of the the room activities involve walking around the whole kitchen area and performing tasks at different locations of the kitchen. For example, the activity `set table` consists of opening/closing cupboards and drawers, and several repetitions of picking up objects, transporting them to the table, and placing them at the proper place.

Activity	Description	Seq. Length (s)	
		μ	σ
`peel`	Using a vegetable peeler	137	66
`cut`	Slicing vegetables with a knife	116	59
`fry`	Frying vegetables in a pan	75	17
`stir`	Stirring liquids in a pot on the stove	69	18
`wipe`	Wiping counter-top with a cloth	34	24
`wash`	Washing dishes in the sink	133	64
`dry`	Drying and stowing away dishes	86	44

Table 3.2. *Description of activities recorded using the "counter-top" setup, and statistical information about the sequences recorded in this setting.*

3.4 Discussion

In this chapter, we have given an overview of commonly used benchmarks for action and activity recognition. Furthermore, we have described the KIT Robo-kitchen data set which we have created in the course of this work. This data set is aimed at developing activity recognition systems for humanoid household robots. In order to best compare our approach with other methods under different challenges, we have selected several data sets other than KIT Robo-kitchen that are to be used throughout this work for the experimental evaluation. A quantitative summary of the properties of fifteen activity recognition benchmarks can be found in Tab. 3.1. Since not all have been recorded using sensor setups that are relevant to our work, we have narrowed the field down and assessed the remaining benchmarks based on quality criteria that we deemed most important. Because we want to apply our approach to real-world scenarios, the evaluation data sets should be realistic in terms of the way people behave as well as the setting.

Activity	Description	Seq. Length (s)	
		μ	σ
`peel`	Using a vegetable peeler	118	70
`cut`	Slicing vegetables with a knife	93	45
`wipe`	Wiping table with a cloth	90	19
`set table`	Setting table for three people	110	19
`clear table`	Putting dishes in a dishwasher	99	19
`empty dishwasher`	Stowing away cleaned dishes and cutlery from dishwasher	67	13
`sweep`	Sweeping floor with a broom	90	21
`coffee`	Reading newspaper at the table while drinking coffee	149	47
`pizza`	Eating pizza with cutlery	70	61
`soup`	Eating soup with a spoon	128	51

Table 3.3. *Description of activities recorded using the "room" setup, and statistical information about the sequences recorded in this setting.*

Furthermore, one of our goals is to model activities of a very high complexity (as opposed to simple actions of a very short length) and this should be reflected in the employed benchmark as well.

Most important are however the amount of contained samples, as well as the diversity of objects and activity categories. A proper benchmark should contain enough training samples to allow the employed machine learning algorithms to best capture the underlying structure. Also, the testing sample size should be large enough so that the experimental results can be of certain statistical relevance. A good benchmark should reflect the real-world as much as possible, and therefore the samples should have a big intra-class variance, as well as cover as many possible categories as possible. Since the benchmark should be challenging as well, it should contain activity categories that are very similar to each other, either based on the involved objects, or motions. A simple example are the activities of `eating a banana`, `drinking a cup of coffee`, and `picking up a phone`, all of which consist of the similar motion of moving one hand towards the face.

Using the previously discussed factors, we have created a qualitative rating of relevant activity recognition benchmarks which is presented in Tab. 3.4. It can be clearly seen that most data sets barely meet our quality criteria and are thus ruled out for the evaluation. For instance, the Gupta, MINTA, ORGDB, and MEA data sets have been recorded under settings that are too unrealistic. In contrast, CMU-MMAC, Poeticon, and 50 salads are all set in real-world environments, but contain too few high-level activity categories, and are therefore more suitable to assess fine-grained action recognition approaches.

Based on the quality criteria alone, the Breakfast and MPII2 data sets would be a perfect choice. Unfortunately, they were released too recently so that we could not consider them for this work. This leaves us with the CAD-120, URADL, and KIT Robo-kitchen data sets, which we are going to use throughout this work for a thorough evaluation of the presented approach.

Data set	sample size	diversity (obj)	diversity (act)	complexity	realism (env)	realism (act)
Gupta	- -	∘	- -	- -	- -	-
CMU-MMAC	+ +	+	- -	+ +	+	+
URADL	+	∘	+	∘	+ +	+
MINTA	+ +	-	∘	- -	-	- -
Poeticon	∘	∘	-	+ +	∘	∘
KIT (ours)	+ +	+	+ +	+	+ +	+ +
CAD-120	∘	+	+	-	-	+
50 salads	+	+	- -	+ +	+	+
Breakfast	+ +	+ +	+ +	+ +	+ +	+ +
ORGBD	+	- -	∘	∘	- -	∘
MEA	∘	-	-	+	∘	+
MPII2	+	+ +	+ +	+ +	+ +	+ +

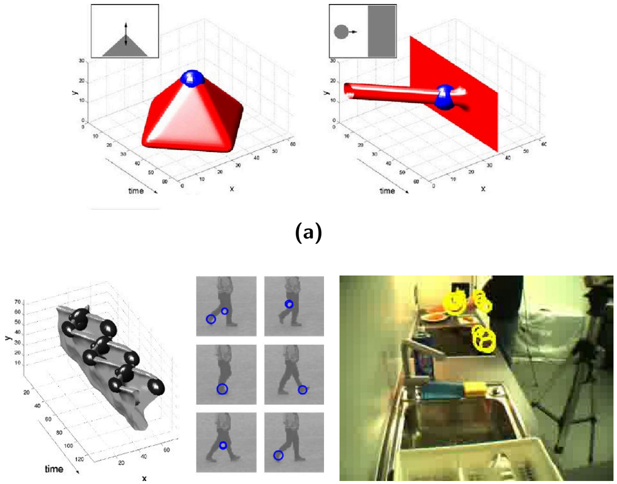
Table 3.4. *Qualitative assessment of publicly available activity recognition data sets that are most relevant to our work. Ratings range from - - (worst) to + + (best). The subjective criteria are: sample size,* i.e., *is the training set large enough for a proper evaluation; diversity (obj),* i.e., *are many different activity-relevant objects visible in the recordings; diversity (act),* i.e., *do the activity category samples differ much from each other; complexity,* i.e., *can the activities be decomposed in many actions; realism (env),* i.e., *how realistic is the setting, realism (act),* i.e., *are the subjects behaving in a realistic way or do the performed activities appear staged.*

4 Activity Recognition Framework: Evaluation and Analysis

The main contribution of this work are saliency-based object candidate region features that are to be used for activity recognition. Because object information is however not enough to properly discriminate between activities, we have created a pure motion-based activity recognition framework by implementing several state-of-the-art local spatio-temporal feature encodings, as described in this chapter. This framework serves as a baseline against which we compare the proto-object features, as well as the motion description that is used in conjunction with our approach. In this chapter, we also evaluate this pure motion-based framework under different settings and select the systems yielding the highest recognition rate to serve as our baseline.

4.1 Local feature extraction

As argued in Sec. 2.4, we decided to represent motion information by adapting local space-time feature based methods, since they are fast to train and yield a performance that is competitive with state-of-the-art. Specifically, we are employing the two most popular types of descriptors: space-time interest point features (STIP) from Laptev



(a)

(b)

(c)

Figure 4.1. *Visualizations of Harris3D feature detections obtained from different types of data: (a) When using synthetic data, example detections (illustrated as blue spheres) occur at the locations in front of a moving corner, and when a moving ball hits a wall. (b) Detections on real data: The 3D plot illustrates the thresholded level-surface of the leg data.(c) Detections on a sample from the KIT Robo-kitchen data set: The radius of the circles reflects the detection scale. ((a) and (b) are reprinted from Laptev and Lindeberg (2003), ©2003 IEEE)*

and Lindeberg (2003), and improved dense trajectory features (iDT) from Wang and Schmid (2013).

Both descriptor types represent different local feature localization schemes, namely sparse interest point detection, and dense sampling, and therefore may exhibit different properties with respect to the BoW variants that we employ as mid-level video representations. In the following, we give a brief introduction to both methods. Visualizations of STIP features detected in synthetic, as well as real data can be found in Fig. 4.1.

4.1.1. STIP features

Space-time interest points developed by Laptev and Lindeberg (2003) are a generalization of the Harris corner detector to video data. In order to allow the detection of these interest points on multiple scales, the image sequences are first convolved with a set of Gaussian kernels g, resulting in a linear scale-space representation of the input data (*cf.*, Witkin (1983)). Since spatial and temporal dimensions are treated independently and the same variance is used for both spatial dimensions, the Gaussian is characterized by only two hyper-parameters, *i.e.*, the spatial variance σ_l^2, and temporal variance τ_l^2, leading to

$$g(x, y, \sigma_l^2, \tau_l^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \cdot \exp -(\frac{x^2 + y^2}{2\sigma_l^2 - t^2/2\tau_l^2}). \tag{4.1}$$

The general idea of the Harris corner detector is to find spatial locations in an intensity image, where it has significant changes in both directions. The same applies when generalizing the detector to the 3D spatio-temporal space, but now instead of applying the operations to an image, we do it to an image sequence I. For a given scale (σ_l^2, τ_l^2), these interest points can be found by using a second-moment matrix μ integrated over a Gaussian window of spatial size σ_i^2 and temporal size τ_i^2:

$$\mu = g(x, y, \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}, \tag{4.2}$$

where the first order derivatives are defined as

$$L_{j \in \{x,y,t\}}(x, y, t, \sigma_l^2, \tau_l^2) = \partial_{j \in \{x,y,t\}}(g * I). \tag{4.3}$$

Note, that the integration scales σ_i^2 and τ_i^2 are related to the local scales σ_l^2 and τ_l^2 by a constant factor s, *i.e.*, $\sigma_i^2 = s\sigma_l^2$, and $\tau_i^2 = s\tau_l^2$. Harris corners can then be found at locations where the first two eigenvalues of μ are sufficiently large. Since exact computation of the eigenvalues is computationally expensive, Harris and Stephens (1988) suggested to compute the Harris corner measure H instead, where λ_1, λ_2, and λ_3 denote the eigenvalues:

$$\begin{aligned} H &= \det(\mu) - \text{trace}^3(\mu) \\ &= \lambda_1\lambda_2\lambda_3 - \kappa(\lambda_1 + \lambda_2 + \lambda_3)^3 \end{aligned} \tag{4.4}$$

Provided the tunable sensitivity parameter κ is sufficiently large, then positive local maxima of H correspond to space-time corners.

In the original Harris3D formulation, automatic scale selection is performed to determine the values of σ_l^2 and τ_l^2. Laptev et al. (2008) noted however that a more computationally efficient solution can be achieved, by omitting this step and instead detecting the interest points at multiple spatio-temporal scales, so that $\sigma_l^2 \in \{2^{(1+j)/2} | j = 1, .., 6\}$ and $\tau_l^2 \in \{2^{j/2} | j = 1, 2\}$. Since this choice of parameters has proven to yield good results, we employ that approach throughout this work as well.

Following interest point detection, histogram descriptors are computed within the spatio-temporal neighborhoods of the localized corners in order to characterize motion and appearance. Histogram based descriptors are robust to variations in rotation and translation, and therefore widely employed for image recognition tasks. The size $(\Delta_x, \Delta_y, \Delta_t)$ of a cuboid region around each interest points is a multiple of the scale parameters, *i.e.*, $\Delta_x = \Delta_y = 2k\sigma_l$, and $\Delta_t = 2k\tau_l$. Each cuboid is further divided into a (n_x, n_y, n_t) grid of sub-volumes before L_2 normalized histograms of oriented gradients (HOG) and histograms of optical flow (HOF) are computed for each sub-volume to describe the local structure.

HOG features are computed from gradient representations of the images that are created by the application of a Sobel filter. In order to achieve robust descriptors, the gradient orientation is coarsely discretized into four histogram bins, and magnitude information is discarded. In order to compute HOF features, first sparse optical flow is estimated using the KLT tracker developed by Lucas and Kanade (1981). Again, only the orientation of the optical flow vector is considered for the histogram descriptors, yet now five bins are used, four bins for direction, and one for no motion. As suggested by Laptev et al. (2008), we set the parameters to $k = 9$, $n_x = n_y = 3$, and $n_t = 2$, and thus a HOG descriptor has in our implementation a size of $\dim_{\text{HOG}} = 4 \cdot n_x \cdot n_y \cdot n_t = 72$, and a HOF descriptor of $\dim_{\text{HOF}} = 90$.

4.1.2. Improved Dense Trajectory features

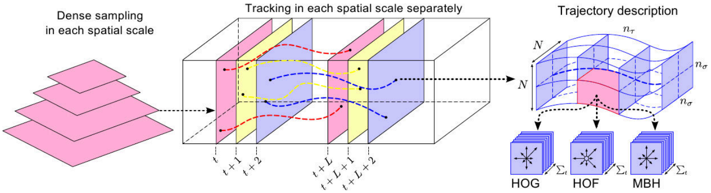


Figure 4.2. *Illustration of the (improved) Dense Trajectory descriptor (iDT) computation pipeline (cf., Wang et al. (2011a)). For each spatial scale, fixed-length feature trajectories are computed from dense optical flow. Local histogram descriptors (HOG, HOF, and MBH) are then extracted over spatio-temporal neighborhoods along these trajectories. To this end, the trajectory neighborhood is divided into a $n_x \times n_y \times n_z$ grid, and for each descriptor type, histogram features from all cells are stacked to form the final descriptor (©2011 IEEE).*

Instead of extracting descriptors only at sparse locations obtained from a spatio-temporal local feature detector, Wang et al. (2009) suggested to sample the cuboid volumes densely over the image sequence. Wang and Schmid (2013) have further elaborated on this idea and developed the improved Dense Trajectory (iDT) features, which are computed within space-time volumes around densely sampled local feature trajectories, as illustrated in Fig. 4.2. The feature trajectories are obtained by median-filtering a dense optical flow field, which is estimated using the OpenCV implementation of the approach from Farnebäck (2003). In order to prevent the tracked points from drifting too much during tracking from their initial position, the trajectory length is limited to a maximum of N_t frames. Further noise is removed by pruning feature points which are static,

originated from homogeneous regions, or show a displacement above a threshold, which can be attributed to tracking errors.
Similar to STIP features, local HOG and HOF descriptors are calculated to characterize the trajectory aligned 3D spatio-temporal volumes of size $N_x \times N_y$, which are further divided into a (n_x, n_y, n_t) grid of sub-volumes. For a fair comparison with other approaches based on iDT features, we follow the advice of Wang and Schmid (2013) and use a histogram resolution that is finer than the one used in STIP features. Thus, we now encode gradient orientation in an 8-bin HOG, and optical flow direction in a 9-bin HOF descriptor, in which one bin is, again, reserved to account for the lack of motion. Furthermore, instead of L_2 normalizing the histogram features, the RootSIFT normalization scheme (*cf*., Chatfield et al. (2011)) is applied, *i.e.*, the feature vectors are L_1 normalized before a square root operation is applied to each vector component.
Additional structural information is represented in the iDT descriptors with motion boundary histograms, and trajectory shape features. MBHs are gradient histograms of the optical flow field, which is separated into its horizontal and vertical components. The resulting $\mathrm{MBH_x}$ and $\mathrm{MBH_y}$ features use the same number of bins and normalization as HOG features. Further motion patterns are encoded in terms of displacement vectors $\Delta P_t = P_{t+1} - P_t$ of feature locations P_t along a trajectory. Thus, the normalized shape S of a trajectory of length N_t is represented as

$$S = \frac{(\Delta P_t, \ldots, \Delta P_{t+N_t-1})}{\sum_{j=t}^{t+N_t-1} \|\Delta P_j\|}. \tag{4.5}$$

As with STIP descriptors, we follow the suggestions of the reference implementations and set the parameters defining the space-time volumes $N_x = N_y = 32$ pixels, $N_t = 15$ frames, and $n_x = n_y = 2$, and $n_t = 3$. The final trajectory descriptors have thus a dimensionality of 30, HOG, MBH_x and MBH_y each of 96, HOF of 108, and the full iDT descriptor obtained by stacking all feature vectors has a dimensionality of 426.

4.2 Bag-of-Words representation

Once local descriptors have been computed for the video samples, they need to be combined to form a global video representation. To this end, we employ approaches belonging to the BoW family, which typically consist of two stages: codebook learning, and feature encoding. In the following, we describe in detail the employed BoW encodings, as well as codebook learning methods. An illustration of the whole BoW processing pipeline can be found in Fig. 4.3.

4.2.1. Codebook generation

Bag-of-Words type representations originate from natural language processing, where words are discrete members of a dictionary. In contrast, visual descriptors (*e.g.*, HOG, and HOF) are continuous, unbound, vector-valued variables. In order to represent them as Bags-of-Words, the feature space thus first needs to be discretized, which is achieved by clustering the descriptors into codebooks.

The most common approach to learn a BoW dictionary is the **k-means** clustering algorithm (*cf.*, MacQueen (1967)). Given a set of feature vectors $X = \{\mathrm{x}_n | n = 1, \ldots, N; \mathrm{x}_n \in \mathbb{R}^d\}$ the goal is to partition the feature set into K clusters $D = \{\mathrm{d}_k | k = 1, \ldots, K; \mathrm{d}_k \in \mathbb{R}^d\}$, *i.e.*, the visual dictionary. Each of the d_k is a prototype of the k-th cluster, *e.g.*, in form of the cluster mean, or median.

Let $R = \{r_{nk} | n = 1, \ldots, N; k = 1, \ldots, K; r_{nk} \in \{0, 1\}\}$ be a set of binary indicator variables for each feature x_n, so that $r_{nk} = 1$ if x_n belongs to cluster k, and $r_{nk} = 0$ otherwise. The objective function of the k-means algorithm can then be defined as

$$\min \mathcal{J}(\{r_{nk}, \mathrm{d}_k\}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \|\mathrm{x}_n - \mathrm{d}_k\|_2^2. \tag{4.6}$$

In order to find the values of r_{nk} and d_k that minimize $\mathcal{J}$, an EM-like iterative procedure is adapted. Each step consists of an optimization of $\mathcal{J}$ with respect to r_{nk}, followed by its optimization with respect to d_k. The algorithm is often initialized by setting r_{nk} to random values,

which given the nature of the algorithm to only guarantee local minima can lead to a sub-optimal partitioning. Therefore, we seed the locations of the cluster prototypes with the kmeans++ heuristic developed by Arthur and Vassilvirskii (2007), which has shown to yield much more stable results.
The major disadvantage of using a dictionary obtained by k-means or similar methods (*e.g.*, spectral clustering, affinity propagation) is that the algorithm performs a hard assignment of features to cluster prototypes and thus severely suffers from quantization errors. To minimize such information loss, some BoW approaches (*e.g.*, Fisher Vector encoding) learn Gaussian Mixture Models (**GMM**) instead, to represent the visual dictionary. A GMM is a generative model to describe a distribution over space:

$$p(\mathrm{x}; \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathrm{x}; \mu_k, \Sigma_k), \tag{4.7}$$

where K is the number of mixtures describing the codebook entries, $\mathcal{N}(\mathrm{x}; \mu_k, \Sigma_k)$ is an M-dimensional Normal distribution parametrized by a mean vector μ_k and covariance matrix Σ_k, π_k are the weights of the individual Gaussians, and $\theta = \{\pi_1, \mu_1, \Sigma_1, \ldots, \pi_K, \mu_K, \Sigma_K\}$ the model parameters. Given a feature set $X = \{\mathrm{x}_n | n = 1, \ldots, N; \mathrm{x}_n \in \mathbb{R}^d\}$, the parameters of a GMM are learned through a maximum likelihood estimation, *i.e.*,

$$\hat{\theta} = \arg\max_{\theta} \ln p(X; \theta). \tag{4.8}$$

The seemingly most popular approach to determine these mixture parameters is the EM algorithm (*cf.*, Dempster et al. (1977)), which we thus adapt in our approach as well.

4.2.2. Vector Quantization

The simplest form of BoW representations is Vector Quantization (VQ), and belongs to the category of voting-based BoW algorithms (*cf.*, Sivic and Zisserman (2003)). Given a codebook D of size K

learned from training data, the VQ voting value $\phi_i(\mathrm{x})$ of a local descriptor x to the i-th codebook entry d_i is calculated as

$$\phi_i(\mathrm{x}) = \begin{cases} 1 & \text{if } i = \arg\min_j \|\mathrm{x} - \mathrm{d}_j\|_2 \\ 0 & \text{otherwise} \end{cases} . \tag{4.9}$$

The VQ encoding for a single descriptor x is then defined as the binary indicator vector $\mathrm{s}_{\mathrm{VQ}} = (\phi_1(\mathrm{x}), \ldots, \phi_K(\mathrm{x}))^T$. In order to represent the set of all descriptors x extracted from a video sample as a BoW feature vector, the VQ encodings of all x are sum-pooled, *i.e.*, summed up to form one vector.

4.2.3. Locally Linear Coding

The Locally Linear Coding (LLC) algorithm introduced by Wang et al. (2010) belongs to the class of reconstruction-based BoW methods. In contrast to voting-based approaches, where each local descriptor is represented in terms of its voting value to one (or several) most similar dictionary word(s), reconstruction-based encoding methods are designed from the perspective of the decoding process. In other words, the codes s representing a descriptor x are enforced to reconstruct x. Usually, these encodings are formulated as a least-squares optimization problem with a regularization term

$$\mathrm{s}_j = \arg\min_{\mathrm{s}} \|\mathrm{x} - D\mathrm{s}\|_2^2 + \lambda\psi(\mathrm{s}), \tag{4.10}$$

where the least-squares term enforces a small reconstruction error, $\Psi(\mathrm{s})$ enforces certain properties of the code s, and λ is a weight factor. The basic idea behind Locally Linear Coding is to utilize a locality constraint Ψ_{LLC} and project each descriptor x into a local linear subspace spanned by $M \ll K$ codebook entries that are closest to x. The clear benefit of reconstructing x only in terms of its M most similar dictionary entries lies in computational efficiency, since such an approximation leads to a much smaller linear system to be solved in the least-squares term of Eq. 4.10. The locality constraint itself is defined as

$$\psi_{\mathrm{LLC}}(\mathrm{s}) = \|\mathrm{e} \odot \mathrm{s}\|_2^2, \quad \text{so that } 1^T\mathrm{s} = 1, \tag{4.11}$$

where $\odot$ denotes the element-wise multiplication of two vectors, and $\mathrm{e} \in \mathbb{R}^M$ is the locality adaptor that gives different freedom to each dictionary entry d_i proportional to the descriptor x. Specifically, it is defined as

$$\mathrm{e} = \exp\left(-\frac{\mathrm{dist}(\mathrm{x}, D)}{\sigma}\right), \tag{4.12}$$

where $\mathrm{dist}(\mathrm{x}, D) = (\mathrm{dist}(\mathrm{x}, d_1), \ldots, \mathrm{dist}(\mathrm{x}, d_M))^T$ for the M codebook words that are nearest to x. $\mathrm{dist}(\cdot)$ denotes the Euclidean distance, and σ is used to adjust the weight decay speed for the locality adaptor. The constraint $1^T\mathrm{s} = 1$ in Eq. 4.11 follows the shift-invariant requirements of the LLC encoding.

In our implementation of the LLC coding, we employ again the parameters suggested by the authors, and thus use $M = 5$ nearest neighbors to reconstruct x, and set the regularization weight to $\lambda = 10^{-4}$. Wang et al. (2010) suggested to employ the max operator to pool the feature codes, however, in our experiments, we made the experience, that sum-pooling yields much better results.

4.2.4. Fisher Vector encoding

The Fisher Vector (FV) encoding has been introduced by Perronnin and Dance (2006) for image classification and is based on the Fisher kernel (*cf*., Jaakola and Haussler (1999)). It captures the average first and second order differences between local feature descriptors and codebook entries and thus belongs to the category of supervector-based BoW encodings, which are in general very high dimensional. Unlike in the previously described BoW approaches, FV encodings start off with a GMM codebook, which can be thought of as a soft dictionary, since it also captures the shape of the clusters in terms of covariance matrices.

Given a codebook $D_{GMM} = \{(\pi_1, \mu_i, \Sigma_1), \ldots, (\pi_K, \mu_K, \Sigma_K)\}$ of size K, as described in Sec. 4.2.4, the membership of a local descriptor x to cluster k is expressed in terms of the two vectors:

$$\mathcal{G}^{\mathrm{x}}_{\mu,k} = \frac{1}{\sqrt{\pi_k}}\gamma_k\left(\frac{\mathrm{x}-\mu_k}{\sigma_k}\right), \text{ and } \mathcal{G}^{\mathrm{x}}_{\sigma,k} = \frac{1}{\sqrt{2\pi_k}}\gamma_k\left(\frac{(\mathrm{x}-\mu_k)^2}{\sigma_k^2} - 1\right), \tag{4.13}$$

where γ_k is the soft-assignment of x to the k-th Gaussian:

$$\gamma_k = \frac{\pi_k \mathcal{N}(\mathrm{x}; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathrm{x}; \mu_j, \Sigma_k)}. \tag{4.14}$$

The FV encoding s_{FV} of x is then obtained by concatenating the membership vectors $\mathcal{G}^{\mathrm{x}}_{\mu,k}$, and $\mathcal{G}^{\mathrm{x}}_{\mu,k}$, *i.e.*,

$$\mathrm{s}_{FV} = ((\mathcal{G}^{\mathrm{x}}_{\mu,1})^T, (\mathcal{G}^{\mathrm{x}}_{\sigma,1})^T, \ldots, (\mathcal{G}^{\mathrm{x}}_{\mu,K})^T, (\mathcal{G}^{\mathrm{x}}_{\sigma,K})^T)^T \tag{4.15}$$

Note, that the covariance matrices Σ_k are typically diagonal since computing full covariance matrices is too slow. As with the other BoW methods, the FV representation of a video sample is obtained by sum-pooling the FV encodings of all descriptors extracted from the sample.

Since the size of an FV encoding is $2DK$ and thus depends on the size of the local descriptors, the descriptors are typically first compressed via PCA. Furthermore, we adapt the suggestions from Perronnin et al. (2010) to improve the descriptive power of the FV encoding, and thus further apply an L_2 normalization, followed by an element-wise power transform, *i.e.*, apply the function $f(z) = \mathrm{sign}(z)|z|^{\alpha}$ to each vector component. The reasoning behind L_2 normalizing FV features is that this approximately cancels out the effect of sample-independent information from the encoding.

The motivation to also apply a power-transform is based on the observation that with an increasing size of the GMM, the FV representation gets sparser. However, the dot product on L_2 normalized vectors is equivalent to an L_2 distance, which is a poor similarity measure for sparse vectors. Because FV encodings are typically used in conjunction with linear SVMs for classification, that rely on the dot product, the sparsity of the FV should be first reduced, which is easily achieved by the power-transform. The optimal value of α for the power-transform depends on the number of Gaussians in the GMM. Since using a GMM of size $K = 256$ in conjunction with $\alpha = 0.5$ has shown to yield good results (*cf.*, Perronnin et al. (2010)), we follow these suggestions in our approach.

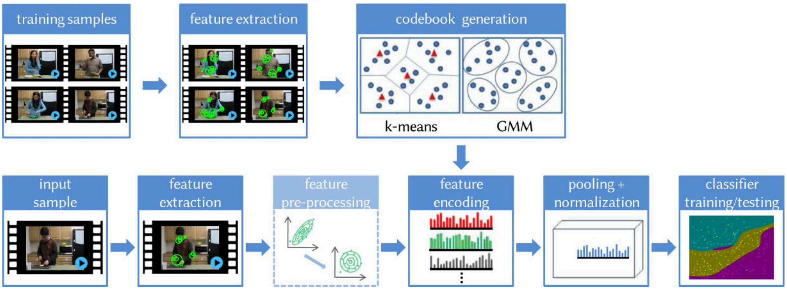


Figure 4.3. *Illustration of a typical BoW feature encoding pipeline: Local features extracted from training samples are first used to learn a codebook via k-means, or GMM clustering. Next, the codebook is used to compute BoW encodings of all local features. All BoW encodings from a single image-sequence are finally pooled and normalized to obtain a global representation of the video which can be used for classification. Typically, a set of linear SVMs is used as classifiers. After extraction, local feature descriptors can optionally be pre-processed,* e.g., *by reducing the dimensionality through PCA, or by applying a L_1, L_2, or rootSIFT feature normalization.*

4.3 Activity recognition

In order to represent motion information that is used in conjunction with the proposed object candidate regions to perform activity recognition, we make use of the BoW framework, as explained in the previous sections. More specifically, we employ STIP, and iDT features to represent motion, and encode each feature type with three BoW methods: VQ, LLC, and FV, all having different properties. The feature encoding pipeline is visualized in Fig. 4.3. The BoW representations are characterized by many parameters, which all have a direct impact on the activity recognition performance, most importantly the codebook size and type feature normalization. In the following, we want to determine experimentally good values for these parameters, in order to create a strong baseline for the proposed object candidate features.

4.3.1. Experimental setup

As argued in Sec. 3.4, we evaluate our approach on three activity recognition benchmarks, URADL, CAD-120, and KIT Robo-kitchen, since they best capture the challenges posed by a real-world environment. For the experiments on URADL and CAD-120, we follow the suggestions provided by their creators and employ a leave-one-subject-out evaluation protocol. Note, that unlike the other two benchmarks, KIT Robo-kitchen follows a slightly different experimental setup, where instead of using recordings of the whole duration of each video sample, all possible 150 frame long sub-sequences are taken for training/classification (*cf.*, Rybok et al. (2011)).

As a performance measure for the evaluated methods, we employ the correct classification rate averaged over all testing samples. We are mainly interested in creating a strong baseline against which we compare the proposed object-candidate features. Thus, the aim of this evaluation is to select values for some experimental settings, that have the strongest impact on the recognition performance. Based on the results, we also select the types of local feature descriptors that will be used in further experiments, since evaluating all combinations of motion- and object-candidate features is not feasible.

In order to map the BoW features to activity categories, we train linear SVMs, following a one-vs-all paradigm to allow multi-class classification. The free hyper-parameters of the SVMs are determined with a leave-one-subject-out cross-validation on the training data.

Since no clear guidelines are reported in related literature on which codebook-size to use for VQ, and LLC encodings, we emphasize this aspect in our evaluation. Feature normalization is another important factor that has often been reported to have a high impact on the final classification rate (*cf.*, Arandjelovic and Zisserman (2012); Chatfield et al. (2011); Peng et al. (2016); Ren and Ramanan (2013)). It is therefore addressed in this evaluation as well.

Regarding Fisher Vector encodings, related publications constantly report that using 256 GMM components suffice to achieve a good trade-off between computation time and classification performance (*e.g.*, Perronnin et al. (2010)). Furthermore, Perronnin et al. (2010)

suggest to apply an L_2 normalization to the Fisher Vectors, followed by an element-wise power transform in order to increase overall performance. Since using FVs for classification is very time-consuming (due to their very high dimensionality), we follow the aforementioned suggestions in our setup.

4.3.2. Effects of BoW normalization

We first want to determine an adequate type and order of preprocessing techniques that are applied to the feature vectors before the training/prediction step of the SVMs. This is more of a preliminary experiment, and thus we restrict this part of the evaluation to the CAD-120 data set, since it constitutes the best trade-off between size (and thus training time) and difficulty.

Typically, three categories of feature processing operations can be distinguished: feature scaling, feature normalization, and power-transform, all of which are being jointly considered in this experiment. Feature scaling is used to standardize the range of independent variables of the feature vector. Its purpose is to prevent features that have a broad range of values from dominating the similarity measure that is calculated by the classifier between all training-sample pairs. The simplest method is min-max normalization, *i.e.*, rescaling each feature to the range in $[0, 1]$ based on the extrema calculated from training data. Another common scaling technique is z-score scaling, where the features are standardized to zero-mean and unit-variance. Empirical studies have shown that SVMs usually work better if the data is properly normalized (*cf.*, Chatfield et al. (2011)); typically by applying L_1- or L_2-normalization. Element-wise power transform has also been pointed out to increase the discriminative power of a feature vector since it makes the distribution of the features more uniform (Arandjelovic and Zisserman (2012); Ren and Ramanan (2013)). It is implemented by raising each dimension of a vector to the power of α. We follow the suggestion of Ren and Ramanan (2013) and set $\alpha = 0.3$.

The full results of this experiment are reported in Appendix B, from which it is clearly visible that choosing the wrong feature nor-

Detector	Descriptor	VQ/codebook				LLC/codebook				FV
		1K	2K	3K	4K	1K	2K	3K	4K	
STIP	HOF	84.0	82.0	86.0	86.7	92.0	91.3	93.3	94.0	93.3
	HOGHOF	87.3	**90.7**	**90.7**	87.3	**94.7**	93.3	93.3	**94.7**	**96.7**
iDT	HOF	71.3	77.3	77.3	76.0	78.7	83.3	82.7	84.0	88.0
	MBH	78.7	82.7	82.7	83.3	82.7	84.7	85.3	**88.0**	**90.7**
	HOGHOF	78.0	83.3	85.3	**88.0**	84.7	84.7	87.3	86.0	87.3
	iDT	80.0	82.0	78.0	84.0	82.7	83.3	82.7	84.7	87.3

Table 4.1. *Activity recognition accuracy (in %) using different codebooks and motion feature encodings on the* **URADL** *data set.*

malization scheme can have a significantly negative impact on the classification rate. Since no clear trend can be observed from the results, we simply select for all following experiments the normalization method that on average yields the best results, *i.e.*, L_1 normalization followed by z-score scaling in the case of VQ encodings, and min-max normalization for LLC.

4.3.3. Effects of codebook and feature type

In the second set of experiments discussed in this chapter, we focus on evaluating all possible combinations of local feature descriptors and BoW representations, which are described in Sec. 4.1 and Sec. 4.2, respectively. Since we want our baseline system to be as strong as possible, we further investigate the impact of the BoW codebook on the recognition performance. Therefore, we run the experiments with different codebooks varying their size between 1000 and 4000. The results obtained from these experiments on the URADL data set can be found in Tab. 4.1. As expected, FV encoded features yield the highest recognition accuracy of 96.7%, which is already very close to the best performance reported outside of this work, ranging at 98.0% (*cf*., Escorcia and Niebles (2013); Yi and Lin (2013)). Nonetheless, the best results achieved when using VQ and LLC feature representations are very good as well, *i.e.*, 90.7% and 94.7%, yet still leave much room for improvement.

Detector	Descriptor	VQ/codebook				LLC/codebook				FV
		1K	2K	3K	4K	1K	2K	3K	4K	
STIP	HOF	67.7	72.6	72.6	71.8	71.0	75.0	75.8	73.4	75.0
	HOGHOF	75.8	75.0	76.6	**79.0**	75.8	79.0	79.0	**80.6**	**82.3**
iDT	HOF	58.1	62.1	62.9	63.7	59.7	59.7	62.9	66.1	66.9
	MBH	62.1	66.1	66.9	70.2	70.2	68.5	67.7	**74.2**	**75.0**
	HOGHOF	61.3	58.9	64.5	63.7	62.9	63.7	67.7	61.3	67.7
	iDT	66.1	**71.0**	69.4	**71.0**	68.5	71.0	72.6	73.4	**75.0**

Table 4.2. *Activity recognition accuracy (in %) using different codebooks and motion feature encodings on the* **CAD-120** *data set.*

Same experiments on the CAD-120 data set lead to similar results, as can be observed in Tab.. 4.2. Again, using Fisher Vectors results in the highest recognition rate of 82.3%, which is slightly lower than state-of-the-art, *i.e.*, 83.1% reported by Koppula and Saxena (2013a).

It is interesting to note, that contrary to the results reported in several large-scale evaluations of local spatio-temporal features for action recognition (*e.g.*, Peng et al. (2016); Wang et al. (2011a)), iDTs are clearly outperformed by STIP features on URADL, and CAD-120. This can probably be attributed to the sample size of URADL and CAD-120, which is much smaller than the number of samples contained in the benchmarks used for the large-scale evaluations, *e.g.*, KTH, Hollywood, or HMDB-51. This hypothesis is further backed up by the experimental results we have obtained on the much larger KIT Robo-kitchen data set (see Tab. 4.2). This time, iDT features are indeed superior to STIP in every feature encoding constellation used in this set of experiments.

4.3.4. Conclusion

Based on the experimental results discussed in the previous sections, we decided to use the following parameter settings for the motion-based baseline system that we employ in the evaluation of the object-candidate features (see Sec. 5.4). In order to increase the amount of

Detector	Descriptor	VQ/codebook				LLC/codebook				FV
		1K	2K	3K	4K	1K	2K	3K	4K	
STIP	HOF	85.7	87.4	88.3	89.0	77.1	82.6	83.2	84.2	**90.1**
	HOGHOF	86.0	88.2	**90.1**	88.2	80.7	83.1	84.4	**85.7**	88.2
iDT	HOF	85.7	87.7	89.6	88.8	72.9	76.3	79.3	79.8	**92.3**
	MBH	90.7	90.4	91.7	**92.3**	72.3	81.0	78.2	82.2	91.7
	HOGHOF	90.4	89.2	89.8	90.6	77.1	80.4	82.2	**82.6**	90.4
	iDT	89.6	89.5	91.4	91.5	78.3	75.6	77.3	81.4	**91.8**

Table 4.3. *Activity recognition accuracy (in %) using different codebooks and motion feature encodings on the* **KIT Robo-kitchen** *data set using* **150 frame** *long activity snippets.*

variation present in the baseline, for each detector type (*i.e.*, STIP and iDT), we select descriptors that consistently show the strongest (and weakest) performance, *i.e.*, HOF and the full descriptor vector. Since the baseline should also be as challenging as possible, we further select for the codebooks used in the BoW representations, the number of dictionary entries that yields the highest performance. For nearly all experiments regarding LLC encoded features, we will thus use 4000-word visual dictionaries, while for the VQ encodings the employed dictionary size varies between 3000 and 4000, depending on the benchmark, and feature detector/descriptor combination. Regarding post-processing, L_1 normalized and z-score scaled VQ encodings, and min-max normalized LLC features resulted in the best results, and are therefore used in further experiments.

5

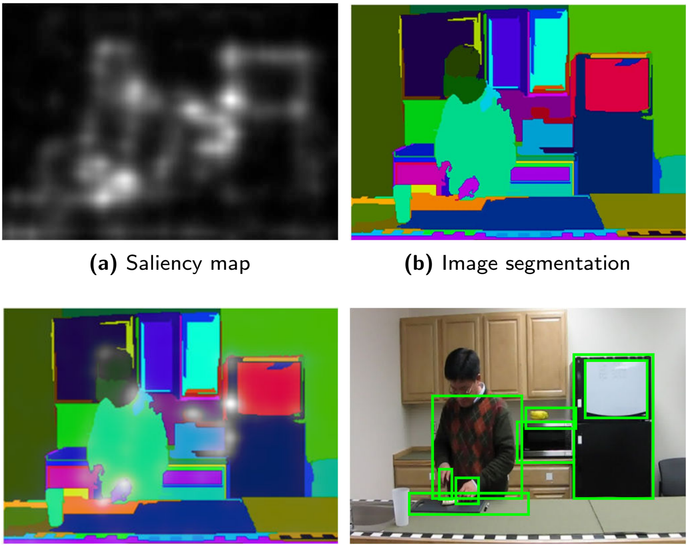
Unsupervised object candidate detection for activity recognition

According to action identification theory, actions and, as a consequence, activities are not only defined by motion patterns but derive their meaning from context (*cf*., Vallacher and Wegner (1987)). For example, the activities of `eating` and `using a phone` can look very similar in their motion patterns and thus be difficult to distinguish without incorporating the context in which they are performed.

Consequently, it may be necessary to also consider the manipulated objects in the process of recognizing activities. Most works in this field, however, either ignore any contextual knowledge, or rely on specifically trained detectors, which in turn require considerable amounts of training data making such approaches difficult to transfer to new domains.

Inspired by recent advances in computational modeling of visual attention, we propose to use salient proto-objects to detect object candidates that are potentially relevant for the activity. The major advantage of such an approach compared to supervised object detection is, that it does not require any additional object annotations. In the following, we describe the proposed proto-object based features and demonstrate experimentally that they allow the integration of contextual object knowledge into motion-based activity recognition.

5.1 Unsupervised discovery of object candidate regions



(a) Saliency map

(b) Image segmentation

(c) Saliency-weighted segments

(d) Proto-object locations

Figure 5.1. *Overview of the proposed unsupervised object-candidate detection approach. First, a QDCT-based saliency map and a graph-based image segmentation are calculated. Then, the segments with the highest saliency are selected as object candidates.*

We build our framework for the unsupervised discovery of object candidates upon the quaternion-based spectral saliency detection (QDCT) algorithm proposed by Schauerte and Stiefelhagen (2012b). Among the advantages of this approach are its simplicity, theoretical soundness, high accuracy in predicting foreground regions, and that it is fully unsupervised. The algorithm extends the image signature saliency descriptor proposed by Hou et al. (2011), by employing a quaternion representation of an image. This makes it possible to process all color channels simultaneously in a holistic fashion.

```
Input:
    $\theta$   max saliency threshold
    $K$   max number of segments
    $S$   set of (image-segments, saliency) pairs
Output:
    $O$   set of detected proto-objects
find max saliency value $s_{\max} = \max(S)$;
set $s' = s_{\max}$; $O = \{\}$;
while $s' > \theta \cdot s_{\max}$ $AND$ $|O| < K$ do
    set $s'$ to $\max(S)$;
    add image segment in $s'$ to $O$;
    remove image segment in $s'$ from $S$;
end
```

Algorithm 1: Extraction of the most salient proto-object regions from an image implementing attentional shifts and inhibition of return. Prior to this selection algorithm, each image-segment in S is assigned the highest saliency value within a saliency map's region it occupies.

These image signatures are defined as the signum function of the Discrete Cosine Transform (DCT) of an image I. A saliency map can be obtained by applying an inverse DCT to an image signature followed by smoothing with a Gaussian kernel g (*cf*., Hou et al. (2011)). More specifically, the QDCT based saliency map $S^{C}_{\mathrm{QDCT}}(I)$ is defined as:

$$S_{\mathrm{QDCT}}(I_Q) = g * \left[T(I_Q) \circ T(I_Q)\right] \quad \text{with} \tag{5.1}$$

$$T(I_Q) = \mathscr{D}_Q^{-1}(\mathrm{sgn}(\mathscr{D}_Q(I_Q))), \tag{5.2}$$

where I_Q is a quaternion representation of a multi-channel image, $\circ$ an element-wise multiplication, and $\mathscr{D}_Q$ the quaternion-based DCT. It has been demonstrated theoretically and experimentally by Hou et al. (2011) that such an approach concentrates the image energies on foreground regions and thus can be used to highlight object candidates. We calculate the saliency maps based on the CIE L*A*B color space since it has been shown by Schauerte and Stiefelhagen (2012b) to reliably yield better performance than most other color spaces. A saliency map obtained with the aforementioned approach can be found in Fig. 5.1(a).

5.2 Saliency-guided object candidate extraction

Peaks in a saliency map only indicate the positions of the proto-objects, however, the approximate spatial extent of each proto-object region still needs to be determined. One common approach is to operate on the saliency map itself, *e.g.*, by region growing or by thresholding (*cf*., Hou et al. (2011)). Yet, such a procedure is often highly sensitive to the choice of the saliency detection parameters which directly influences the size of the segmented proto-object regions. Instead, we use the saliency map to guide the proto-object selection directly in the image, as shown in Fig. 5.1.

First, we use the graph-based algorithm introduced by Felzenszwalb and Huttenlocher (2004) to segment each frame of a video sequence and use parameters yielding preferably large image segments (see Fig. 5.1(b)). In order to select a set of proto-objects, we then apply Algorithm 1, which implements attentional shifts and inhibition of return.

It iteratively selects the most salient segments, following the classical winner-take-all approach, and assigns to each segment the highest saliency value within the saliency map's region it occupies. This process is repeated until the saliency either gets below a threshold θ of the saliency maps's maximal value or the most K salient segments have been selected. Those segments form our set of proto-objects, *i.e.*, object candidate regions. In our experiments, we empirically determined the values of $\theta = 70\%$ and K=30.

To encode the appearance of the proto-object regions, we use the HOG features from Dalal and Triggs (2005), which proved, in preliminary experiments, to be superior to other popular feature descriptors, *e.g.*,

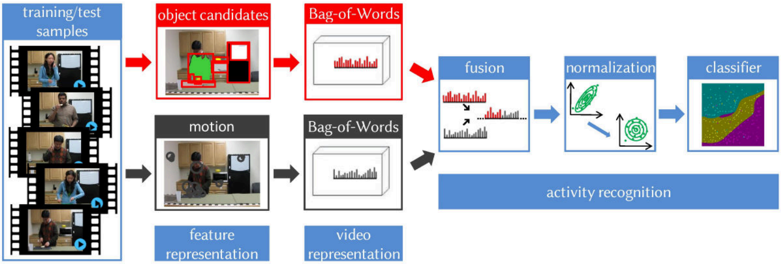


Figure 5.2. *Overview of the two-stream framework we employ to incorporate object candidate knowledge into motion-based activity recognition.*

SIFT, SURF, and ORB. Finally, we apply k-means clustering to obtain a set of object candidate prototypes which we use to represent object information for activity recognition. As can be observed in Fig. 5.3, many of the codewords correspond to real-world objects, or object parts, all of which are meaningful for activity recognition.

5.3 Activity recognition with object candidates

Since object knowledge alone is not enough information to discriminate activities, we also include motion information in order to recognize activities. To this end, we resort to a two-stream framework and process object candidates and motion independently as illustrated in Fig. 5.2. Once representations of the whole image sequence has been established, both information sources are being fused by feature vector concatenation. As argued in Sec. 4.3.2, we further perform feature normalization before classifier training/prediction, since this step increases the descriptive power of the features.

The attentive reader may have already noticed the resemblance of the employed two-stream approach compared to the biologically inspired action recognition methods presented in Sec. 2.1.4. In fact, the widely accepted two-streams hypothesis (*cf*., Goodale and Milner (1992))

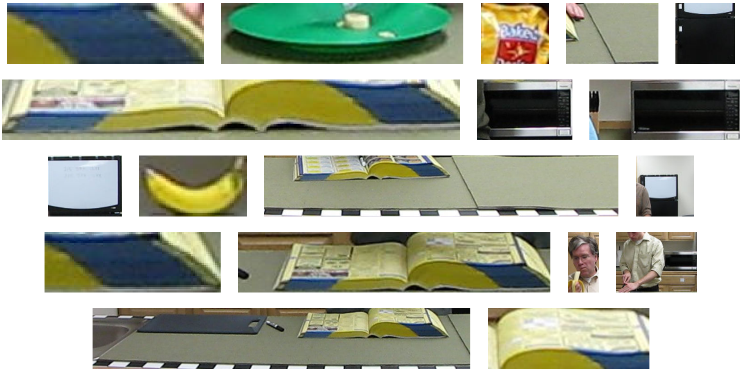


Figure 5.3. *Representatives of the first 18 proto-object feature codebook entries for subject 1 of the URADL data set. The codewords were selected based upon their Minimal-Redundancy-Maximal-Relevance score* (cf., *Peng et al.* (*2005*)).

states, among others, that motion and shape information is processed in the primary visual cortex separately as well. Nonetheless, unlike our work, biologically inspired methods usually utilize Gabor filters to model units at the lowest level of the visual cortex (*i.e.*, simple cells).

5.4 Experimental evaluation

We evaluate the proposed object candidate features on the same benchmarks that were used in the experiments regarding the motion-based baseline, namely on the URADL, CAD-120, and KIT Robo-kitchen data sets. Again, we report in all our experiments the correct classification rate averaged over the test samples. However, this time we focus on the aspect of how well the proto-objects perform alone, and in combination with motion features. Nonetheless, the same BoW encodings are used to describe motion as in Chapter 4, namely VQ, LLC, and FV representations of STIP and iDT features.

The BoW processing pipeline we use to separately represent motion-

Object source	Encoding	Codebook size				
		100	200	300	400	500
Segments	VQ	44.7	52.0	59.3	61.3	54.0
Proto-objects		70.7	72.7	**74.0**	72.7	69.3
Segments	LLC	63.3	62.0	60.7	61.3	54.0
proto-objects		68.0	68.7	70.0	**77.3**	70.7
Annotations	Histogram			90.0		
Detections				68.0		

Table 5.1. *Activity recognition accuracy (in %) when only using object feature encodings on the* **URADL** *data set. For each BoW encoding, the best results are highlighted.*

and object information is illustrated in Fig. 4.3. It consists of feature extraction and pre-processing, BoW encoding, and normalization. Prior to classification via linear SVMs, the motion- and object-candidate features are fused at feature-level. Based on the experimental results obtained in Sec. 4.3.2, we use an L_1 normalization followed by z-score scaling in the case of VQ encodings, and min-max normalization for LLC. For the FV encodings, we again adapt the suggestions from Perronnin et al. (2010). Thus, we first employ PCA to reduce the dimensionality of the local features, learn GMMs with 256 components, and post-process the FVs by first applying L_2 normalization, followed by an element-wise power transform.

To demonstrate the importance of saliency-driven object candidate selection, we also compare to the case where all image segments from the segmentation step are used and not only the most salient ones. If available, ground-truth object labels and object regions obtained from supervised detectors are also being compared against the proposed features. Independent of how the object-candidates are selected (*i.e.*, from proto-object extraction, supervised object detection, or image segmentation), all object regions are encoded with the HOG descriptors proposed by Dalal and Triggs (2005). Finally, we compare the proposed feature representation with state-of-the-art activity recognition approaches to demonstrate its effectiveness.

Object source	# Object candidates	STIP/codebook		iDT/codebook	
		HOF	HOGHOF	HOF	iDT
		4000	3000	3000	4000
None	0	86.7	90.7	77.3	84.0
Annotations	12	91.3	96.7	85.3	90.0
Detections	12	91.3	98.0	84.0	89.3
Segments	100	90.0	97.3	80.7	88.7
	200	90.7	97.3	79.3	88.0
	300	90.7	97.3	79.3	88.0
	400	90.7	98.0	80.0	88.7
	500	91.3	97.3	81.3	88.7
Proto-objects	100	92.0	97.3	86.0	90.0
	200	92.7	**100**	86.0	90.7
	300	**93.3**	**100**	**88.7**	90.7
	400	90.7	98.0	87.3	90.0
	500	92.7	98.7	88.0	**91.3**

Table 5.2. *Activity recognition accuracy (in %) using* **VQ** *encoded object features in conjunction with different motion features on the* **URADL** *data set. For each motion-feature type, the best results are highlighted.*

5.4.1. URADL data set

The University of Rochester Activities of Daily Living data set (URADL) from Messing et al. (2009) contains 150 high-resolution videos of ten activities which are often similar in motion and thus difficult to be separated without context knowledge. The ten activity categories are: `lookup in phonebook`, `dial phone`, `answer phone`, `eat a banana`, `peel a banana`, `slice a banana`, `eat a snack`, `drink water`, `use silverware`, and `write on whiteboard` (see also Fig.1.4). Each activity is performed three times by five different subjects and the evaluation is performed using leave-one-person-out cross-validation. To compare our method with approaches relying on object detections, we manually annotated all images of the data set with the

Object source	# Object candidates	STIP		iDT	
		HOF	HOGHOF	HOF	iDT
None	0	94.0	94.7	84.0	84.7
Annotations	12	**97.3**	**100**	93.3	94.0
Detections	12	**97.3**	99.3	93.3	93.3
Segments	100	96.0	98.0	92.0	92.7
	200	96.0	96.7	90.7	90.0
	300	**97.3**	96.0	90.7	90.7
	400	96.7	94.7	90.7	89.3
	500	96.7	96.0	92.7	90.0
Proto-objects	100	**97.3**	98.7	92.0	94.0
	200	96.7	98.7	94.0	94.7
	300	96.7	94.7	92.7	95.3
	400	**97.3**	94.7	94.0	**96.0**
	500	**97.3**	94.7	**94.7**	95.3

Table 5.3. *Activity recognition accuracy (in %) using* **LLC** *encoded object features in conjunction with different motion features (4K sized codebooks) on the* **URADL** *data set. For each motion-feature type, the best results are highlighted.*

location of the objects that we deemed the most relevant. The twelve labeled object categories are: `whiteboard`, `bottle`, `cup`, `plate`, `crisps`, `phone`, `knife-block`, `paper-roll`, `phonebook`, `peeled banana`, `banana`, and `knife` (see also Appendix C for sample images).

These labels were used to learn a set of state-of-the-art object detectors using the discriminatively trained part-based approach proposed by Felzenszwalb et al. (2010). When using these detectors on the test set, we have obtained an overall Mean Average Precision of 0.744.

We further employ the aforementioned ground-truth annotations to determine how well our approach performs compared to using perfect object knowledge. In order to integrate such object information into our classification framework, we simply treat the object classes as codebook entries and then calculate VQ-like histogram features.

In the first set of experiments, we analyze how well object features

Feature type	Object source	Encoding		
		VQ	LLC	FV
Objects	Annotations	90.0	-	-
	Detections	68.0	-	-
	Segments	61.3	63.3	-
	Proto-objects	74.0	77.3	-
Motions + Objects	None	90.7	94.7	96.7
	Annotations	96.7	**100**	**100**
	Detections	98.0	99.3	**100**
	Segments	98.0	98.0	98.0
	Proto-objects	**100**	98.7	**100**

Table 5.4. *Summary of the best activity recognition results (in %) obtained when evaluating different combinations of object- and motion features on the* **URADL** *data set.*

can be used to predict human activities, and report the results in Tab. 5.1. As expected, using object knowledge based on ground truth annotation results in the highest performance. It is however surprising that even though many activity categories are very similar to each other in terms of manipulated objects, still a very high recognition rate is achieved. For instance, when only using ground truth object labels, we obtained an accuracy of 90%, which is very close to the motion-only baseline ranging at 94.7%. Another interesting finding is that the proposed proto-object features outperform both the object-detector, and image-segmentation based baselines.

In the second part of the experiments, we jointly evaluate object- and motion features. Detailed results of this experiment are reported in Tab. 5.2 (VQ encoding), and Tab. 5.3 (LLC encoding). Due to the very high duration when training the classifiers using FV encodings, we restrict the corresponding set of experiments to an object-candidate codebook size of 300. The resulting accuracy is reported in Tab. 5.4 together with a summary of the best VQ-, and LLC-encoding based systems' performance.

Overall, these experiments suggest that proto-objects indeed have complementary properties to motion features which are beneficial for activity recognition. Surprisingly, integrating proto-objects with motion features performs as good or better than all object candidate selection baselines. A possible reason why the accuracy of proto-objects is comparable with ground-truth labels might be that the decision which object categories are relevant for the activities has been made by humans. In contrast, proto-objects selection is performed in a data-driven fashion free of annotator-bias, and thus better object-candidate regions might be selected.

5.4.2. CAD-120 data set

Object source	Encoding	Codebook size				
		100	200	300	400	500
Segments	VQ	41.9	47.6	50.8	**54.0**	45.2
Proto-objects		38.7	45.2	43.5	51.6	46.0
Segments	LLC	51.6	**54.8**	57.3	48.4	50.8
Proto-objects		40.3	53.2	43.5	48.4	50.8
Annotations	Histogram			74.2		

Table 5.5. *Activity recognition accuracy (in %) when only using object feature encodings on the* **CAD-120** *data set. For each BoW encoding, the best results are highlighted.*

In order to analyze the generalization ability of the proposed object-candidate features, we further evaluate our approach on the Cornel Activity Dataset-120 (CAD-120) created by Koppula et al. (2013). It contains 124 RGBD videos (we only used the color channels) of four subjects performing 10 activities (three repetitions, each time using different objects). The activity categories are: `preparing cereals`, `cleaning objects`, `stacking objects`, `taking food`, `having a meal`, `arranging objects`, `microwaving food`, `taking medicine`, `unstacking objects`, and `picking objects` (see also Fig. 3.5).

Some of the challenges of this benchmark are big variations of camera-view angles and recording locations within each activity class. For

Object source	# Object candidates	STIP/codebook		iDT/codebook	
		HOF	HOGHOF	HOF	iDT
		3000	4000	3000	4000
None	0	72.6	79.0	63.7	71.0
Annotations	10	**79.8**	**84.2**	70.2	77.4
Segments	100	72.6	78.2	67.7	69.4
	200	71.8	79.0	68.5	75.0
	300	73.4	78.2	64.5	74.2
	400	73.4	79.0	66.9	75.8
	500	71.8	79.8	65.3	72.6
Proto-objects	100	78.2	83.1	72.6	76.6
	200	77.4	83.1	71.8	76.6
	300	**79.8**	83.1	**74.2**	**79.8**
	400	79.0	82.3	73.4	78.2
	500	78.2	82.3	72.6	78.2

Table 5.6. *Activity recognition accuracy (in %) using* **VQ** *encoded object features in conjunction with different motion features on the* **CAD-120** *data set. For each motion-feature type, the best results are highlighted.*

comparison purposes, we use the same train-test split that is reported in related literature and follow a leave-one-person-out cross-validation protocol.

As in the previous section, the first set of experiments focuses on object features alone. Since ground-truth annotations of 10 objects have been provided by the authors of the data set, we also include them in the evaluation. As can be seen in Tab. 5.5, the essence of this experiment's results is comparable to the corresponding evaluation on URADL. Again, the best performance of 74.2% is achieved when using ground-truth object knowledge, which is comparable to the motion-based baseline and ∼ 20% (absolute) higher than the accuracy of all proto-object encodings. It should, however, be noted, that this time object-candidates obtained from image-segmentation constantly outperform proto-objects.

Object source	# Object candidates	STIP		iDT	
		HOF	HOGHOF	HOF	iDT
None	0	75.8	80.6	66.1	73.4
Annotations	10	86.3	87.1	79.0	83.1
Segments	100	74.2	75.0	66.9	72.6
	200	72.6	76.6	66.9	74.2
	300	68.5	74.2	71.8	66.9
	400	67.7	73.4	69.4	71.8
	500	72.6	74.2	66.1	68.5
Proto-objects	100	**83.9**	**86.3**	**75.0**	**80.6**
	200	81.5	85.5	73.4	79.0
	300	81.5	84.7	73.4	79.0
	400	80.6	85.5	73.4	78.2
	500	83.1	85.5	**75.0**	79.0

Table 5.7. *Activity recognition accuracy (in %) using* **LLC** *encoded object features in conjunction with different motion features (4K sized codebooks) on* **CAD-120***. For each motion-feature type, the best results are highlighted.*

In the second part of the evaluation, again, we analyze the impact of using object features as an additional cue to motion based activity recognition. A summary of the results can be found in Tab. 5.8, while details of the experiments with VQ and LLC encoded features are presented in Tab. 5.6 and Tab. 5.7, respectively. As in the experiments on the URADL data set, it can be observed that combining proto-objects with motion features clearly performs better than using motion features alone. Even though clearly outperformed by manual annotation based features, when using object features alone, proto-objects still yield a comparable performance to ground truth labels, when jointly using motion- and object features for the classification. Furthermore, using proto-objects constantly results in a better classification accuracy compared to using image segments as object-candidate cue in conjunction with motion features. In summary, this set of experiments again demonstrates the benefits of the proposed approach for activity recognition.

Feature type	Object source	Encoding		
		VQ	LLC	FV
Objects	Annotations	74.2	-	-
	Segments	54.0	54.8	-
	Proto-objects	51.6	53.2	-
Motions + Objects	None	79.0	80.6	82.3
	Annotations	**84.2**	**87.1**	**88.5**
	Segments	79.0	76.6	80.6
	Proto-objects	83.1	86.3	87.1

Table 5.8. *Summary of the best activity recognition performance results (in %) obtained when evaluating combinations of different object and motion features on* **CAD-120**.

5.4.3. KIT Robo-kitchen data set

Object source	Encoding	Codebook size				
		100	200	300	400	500
Segments	VQ	56.6	60.8	65.4	67.2	69.2
Proto-objects		56.1	67.8	70.4	**72.5**	71.8
Segments	LLC	48.2	56.0	64.9	62.7	68.0
Proto-objects		56.3	67.6	67.2	**68.2**	68.1

Table 5.9. *Activity recognition accuracy (in %) when only using object feature encodings on* **KIT Robo-kitchen**. *For each BoW encoding, the best results are highlighted.*

In the final set of experiments, we evaluate the proposed object-candidate features on data recorded in a scenario most closely resembling the application domain of this work, the KIT Robo-kitchen data set. It consists of videos of 14 activity categories, each performed by 17 persons of which ten are used as training data and the remaining seven serve as unseen data for testing.

A detailed description of this data set can be found in Sec. 3.3, explanations of the activity categories in Tab. 3.3, and representative shots of all categories in Appendix A.

Object source	# Object candidates	STIP/codebook		iDT/codebook	
		HOF 4000	HOGHOF 3000	HOF 3000	iDT 4000
None	0	89.0	90.1	89.6	91.5
Segments	100	89.3	91.0	89.7	91.3
	200	89.7	91.4	89.7	91.4
	300	90.3	91.6	89.6	91.2
	400	89.8	91.0	89.0	91.0
	500	90.8	91.6	90.1	91.3
Proto-objects	100	90.2	90.9	89.5	91.6
	200	**90.9**	**91.7**	**91.1**	**91.8**
	300	89.9	91.1	89.7	91.2
	400	90.4	91.4	90.4	91.1
	500	90.4	91.5	90.1	91.3

Table 5.10. *Activity recognition accuracy (in %) using* **VQ** *encoded object features in conjunction with different motion features on the* **KIT Robo-kitchen** *data set. For each motion-feature type, the best results are highlighted.*

Object source	# Object candidates	STIP		iDT	
		HOF	HOGHOF	HOF	iDT
None	0	84.2	85.7	79.8	81.4
Segments	100	87.8	87.1	84.9	88.3
	200	87.7	85.8	85.0	87.5
	300	88.3	86.9	85.6	87.9
	400	88.2	86.4	86.2	87.7
	500	87.1	85.6	**86.3**	87.7
Proto-objects	100	**89.4**	**88.9**	84.9	88.2
	200	87.7	88.6	85.2	**88.4**
	300	87.3	86.4	86.1	88.2
	400	87.4	86.5	**86.3**	88.3
	500	88.3	87.0	85.5	87.9

Table 5.11. *Activity recognition accuracy (in %) using* **LLC** *encoded object features in conjunction with different motion features (4K sized codebooks) on the* **KIT Robo-kitchen** *data set. For each motion-feature type, the best results are highlighted.*

Feature type	Object source	Encoding		
		VQ	LLC	FV
Objects	Segments	69.2	68.0	-
	Proto-objects	72.5	68.2	-
Motions + Objects	None	91.5	84.4	92.3
	Segments	91.6	88.3	90.7
	Proto-objects	**91.8**	**89.4**	**93.1**

Table 5.12. *Summary of the best activity recognition performance results (in %) obtained when evaluating combinations of different object and motion features on the* **KIT Robo-kitchen** *data set.*

As done in the evaluation of the motion-based baseline system which is described in Sec. 4.3, in these experiments we again only use data recorded from the *room:door* viewpoint, since it is the most challenging subset.

Unlike the other benchmarks used in this evaluation, one of the challenges of this data set is that the recognition is not based on clips spanning the whole activity, but rather of all possible 150 frame long sub-sequences of each video. The rationale behind this is application driven. The data set has been designed to model the household robot scenario, in which the robot should offer his services long before the user is finished with the current activity. Thus having to wait for a response until the observed person has already finished his task would be counter-productive.

Again, we begin with an isolated evaluation of the object candidate features, and report the results in Tab. 5.1. The classification accuracy when jointly using motion- and object-based features for activity recognition can be found in Fig. 5.12, while details of the corresponding experiments with VQ and LLC encoded features are given in Tab. 5.10 and Tab. 5.11, respectively. Similar to the experiments on the other two data sets, proto-objects outperform all baselines, which further backs up the usefulness of proto-object based features as an additional cue for activity recognition.

5.5 Discussion

In this chapter, we have introduced the idea of using proto-object based features to encode contextual information for activity recognition. The major advantage of such an approach is that it allows us to automatically extract object candidates from images without any need for annotated training data or motion information. In an experimental evaluation on three realistic data sets, we showed how well the proposed features complement motion information for activity recognition.

A comparison of the proposed approach with the state-of-the-art on all three benchmarks used in the evaluation, *i.e.*, URADL, CAD-120, and KIT Robo-kitchen, is presented in Tab. 5.13, Tab. 5.14, and Tab. 5.15, respectively. Even though we employ a simple feature-level fusion of motion- and object cues, our system outperforms all reported results on these benchmarks, most of which rely on a more complex modeling the relationship between motion and objects.

Interestingly, the works of Wang et al. (2014a) and Lin et al. (2016) are both building upon ConvNets, which applied to most problems usually yield a much better performance than BoW-based encodings. Nonetheless, the reported accuracy of the ConvNet-based approaches on CAD-120 is at least 5.9% (absolute) lower than the accuracy of our system. This observation should be, however, taken with a grain of salt: the CAD-120 data set is very small in terms of samples, and thus no general conclusions can be drawn from this experiment without further analysis.

Reference	Method	Accuracy
Matikainen et al. (2010)	pairwise feature relationships	70.0
Prest et al. (2012a)	human-object interaction features	92.0
Kuehne and Serre (2016)	iDT+FV+HMM	94.7
Yi and Lin (2013)	salient trajectory features+HOG3D	98.0
Escorcia and Niebles (2013)	human-object interaction features	98.0
Rostamzadeh et al. (2013)	2D body-model features+FV	98.7
Ours (best system)	proto-objects+VQ encoded STIP	**100**

Table 5.13. *Comparison of the activity recognition rate (in %) achieved with the proposed method and state-of-the-art approaches on the* **URADL** *data set.*

Reference	Method	Accuracy
Sung et al. (2012)	3D body-model features+MEMM	26.4
Lin et al. (2016)	ConvNet	74.7
Koppula et al. (2013)	object&sub-activity relations+MRF	75.0
Koppula et al. (2013)*	object&sub-activity relations+MRF	80.6
Wang et al. (2014a)	3D reconfigurable ConvNet	81.2
Devanne et al. (2017)	sub-activity segmentation+	82.3
Koppula and Saxena (2013a)*	object&sub-activity relations+CRF	83.1
Koperski and Bremond (2016)	2D body-model features+FV	85.5
Ours (best system)	proto-objects+FV encoded STIP	**87.1**

Table 5.14. *Comparison of the activity recognition rate (in %) achieved with the proposed method and state-of-the-art on* **CAD-120**. *Note, that works marked with (*) are relying on ground truth object labels.*

Reference	Method	Accuracy
Rybok et al. (2011)	VQ encoded STIP	84.9
Onofri et al. (2013)	multiple subsequence combination features	88.3
Ours (best system)	proto-objects+FV encoded STIP	**93.1**

Table 5.15. *Comparison of the activity recognition rate (in %) achieved with the proposed method and several state-of-the-art approaches on the on the room:door setup of the* **KIT Robo-kitchen** *data set.*

6 Conclusions

6.1 Summary

Object information is an important cue to discriminate between activities. However, most activity recognition approaches make use of only motion features alone, or rely on object detectors trained specifically for the target domain. Such object detectors require a significant amount of training data and complicate the transfer of the activity recognition framework to novel scenarios with different objects and object-action relationships. Motivated by recent advances in saliency detection, we have developed in this work a method to employ salient proto-objects for unsupervised discovery of object- and object-part candidates which we use as a contextual cue for activity recognition. In an experimental evaluation on three publicly available data sets, we demonstrated that the integration of proto-objects and simple motion features substantially improves recognition performance, and also outperforming the state-of-the-art.
The motivation behind our approach was driven by the goal to create a system allowing a humanoid service robot to understand typical household situations. As a possible application, we imagine the robot to take the role of a butler observing the scene from a point in the background and offering unsolicited help whenever he assesses that it might be required. Since none of the publicly available activity recognition benchmarks could be used to simulate the challenges posed by such a scenario, we have further created in the context of

this work a suitable data set, that we use among others to assess the proposed approach. This KIT Robo-kitchen data set consists of complex kitchen activities recorded in a realistic scenario.

6.2 Outlook

Even though the proposed approach to detect object-region candidates for activity recognition outperforms state-of-the-art approaches on several benchmarks, there is still room for future improvement.

Spatio-temporal proto-object segmentation
The probably most straightforward approach to increase the quality of the object candidate proposals is to enforce the extracted regions to be temporally coherent. Instead of treating the detections from neighboring frames independent from each other, it would be better to assign them to the same object region prototype. This could, for example, be achieved by incorporating temporal proximity in the distance function used to cluster the regions into prototypes. Another possible way to tackle this problem is by using a space-time supervoxel segmentation approach to select regions corresponding to salient proto-objects, *e.g.*, with the methods from Oneata et al. (2014a); Trichet and Nevatia (2013). Since proto-objects correspond to meaningful entities, another possibility to improve our approach is by employing a refinement step (*e.g.*, the techniques from Doersch et al. (2013); Singh et al. (2012)) after clustering in order to obtain object prototypes that are both discriminative and representative.

Discovering action-primitive candidates in an unsupervised fashion
So far, we have treated motion and object information independently from each other for activity recognition. Nonetheless, intuition, as well as related research, demonstrate that much better results can be achieved when explicitly modeling the co-relationship between both. This would, however, require some form of decomposition of the motions involved in the activities into meaningful parts, *i.e.*, action-primitives.

One possible way to tackle this problem would be by leveraging a similar framework as proposed in this work and explore (spatio-) temporal saliency for an unsupervised detection of action-primitive candidates. Alternatively, such a temporal segmentation could be achieved using a temporal clustering method (*e.g.*, Zhou et al. (2013)).
Furthermore, the act of moving an object is often directly corresponding to a single action-primitive. Thus, object candidate information should be emphasized in the segmentation process. The advantages of such an approach have been demonstrated by Wächter and Asfour (2015) in the context of imitation learning.

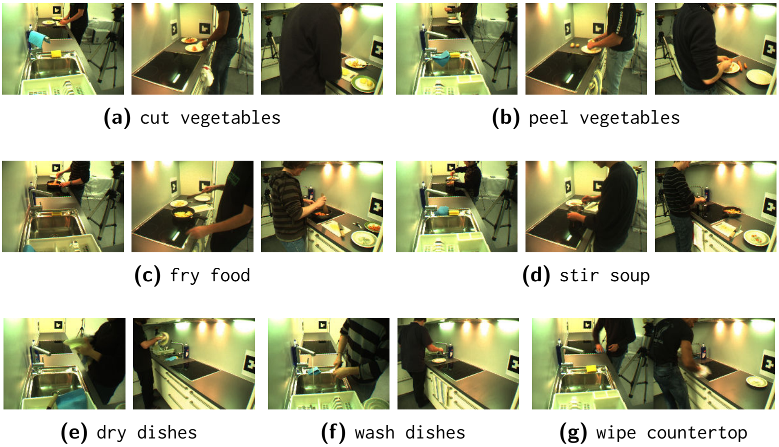
Explicitly modeling the co-relationship between candidate action-primitives and proto-objects
In one of our prior works (*cf*., Gehrig et al. (2011)), we have already explored the use of dynamic Bayesian networks to model the co-relationship between knowledge from different sources to improve activity recognition. However, we used object information directly from ground-truth annotations, and fine-grained action-primitive knowledge obtained from supervised classifiers.
Based on the results of this work, one possible direction for future research is, therefore, to explore the joint modeling of automatically mined object- and action-primitive candidates in a DBN framework. Since most current activity recognition data sets also contain depth information, the location of the observed person could be incorporated into the framework as well in order to restrict the set of possible activities. For instance, when the person is standing in front of a sink, it is more likely that he or she is doing the dishes than having a meal.

Automatic discovery and learning of previously unseen activity classes
Another direction for future research is driven by the application of activity recognition in the context of a household service robot - online learning. Typically, once an activity has been recognized, the robot would communicate with the user, and offer him his services. In the case of a miss-classification, the system could include the specific sample to the training base and thus improve future performance.

Likewise, if a novel activity category is detected, the system should be able to adapt and recognize this category in the future, *e.g.*, by means of zero-shot learning.
In previous works, we have already investigated zero-shot learning methods for action recognition (*cf.*, Al-Halah et al. (2014, 2016)). These were, however, based on attribute detections from supervised classifiers, and not an unsupervised attribute candidate proposal method.

A

Visual walk through the KIT Robo-kitchen data set



(a) `cut vegetables`

(b) `peel vegetables`

(c) `fry food`

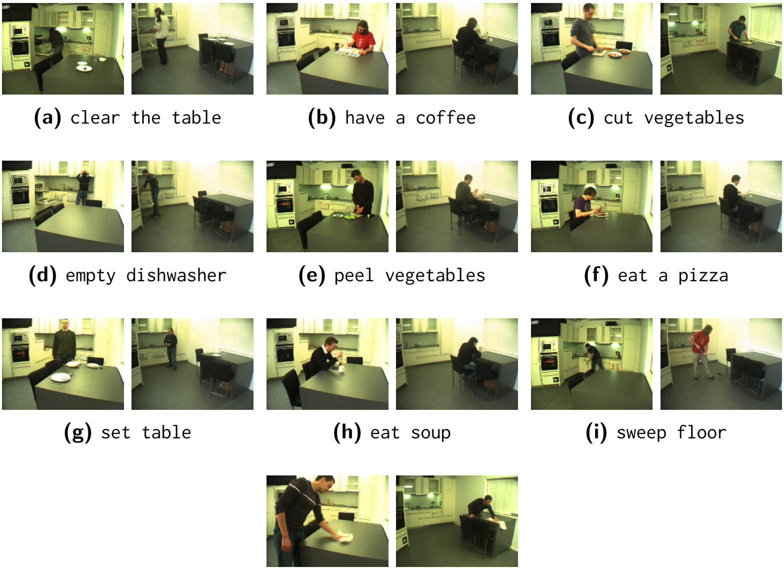
(d) `stir soup`

(e) `dry dishes`

(f) `wash dishes`

(g) `wipe countertop`

Figure A.1. *Sample frames of clips belonging to all seven activity categories of the KIT Robo-kitchen counter-top data subset. Note, that the activities* `dry`, `wash`, *and* `wipe` *could only be recorded with two cameras as the sink camera as was obstructing the actors in performing their task.*



(a) clear the table **(b)** have a coffee **(c)** cut vegetables

(d) empty dishwasher **(e)** peel vegetables **(f)** eat a pizza

(g) set table **(h)** eat soup **(i)** sweep floor

(j) wipe table

Figure A.2. *Sample frames selected from all ten activity categories of the KIT Robo-kitchen data set that were recorded with the room camera setup. All experimental evaluations of this work performed on the data set were only using image sequences recoded with the door camera, corresponding to the left image of each pair. Note, how the activities* eat pizza, eat soup, have coffee *are very similar to each other when regarding motion only, but can easily distinguished based on object knowledge. In contrast, the activities involving the preparation of vegetables,* i.e., peel, *and* cut *can best be distinguished from motion features.*

B

Feature Normalization

Normalization method			VQ		LLC	
first	second	third	STIP	iDT	STIP	iDT
-	-	-	75.0	73.3	71.7	69.4
L_1/sum	-	-	40.3	38.7	46.8	35.5
L_2	-	-	70.2	47.6	67.7	55.6
minmax	-	-	76.6	72.6	**80.6**	**73.4**
z-score	-	-	73.4	70.2	69.4	70.2
power	-	-	**79.0**	69.4	76.6	65.3
L_1/sum	minmax	-	74.2	71.8	71.0	68.5
L_1/sum	z-score	-	**79.0**	71.0	71.0	68.5
L_1/sum	power	-	78.2	66.9	79.0	66.9
L_2	minmax	-	75.0	70.2	75.8	71.8
L_2	z-score	-	75.0	68.5	76.6	67.7
L_2	power	-	77.4	67.7	77.4	65.3
minmax	L_1/sum	-	39.5	38.7	64.5	59.7
minmax	L_2	-	69.4	62.1	65.3	67.7
minmax	power	-	78.2	66.9	75.8	73.4
z-score	L_1/sum	-	59.7	58.9	67.7	61.3
z-score	L_2	-	72.6	65.3	70.2	64.5
z-score	power	-	76.6	69.4	77.4	65.3
L_1/sum	minmax	power	75.8	58.9	78.2	71.0
L_1/sum	z-score	power	67.7	72.6	78.2	71.0
L_1/sum	power	minmax	78.2	61.3	76.6	66.9
L_1/sum	power	z-score	71.8	64.5	66.9	66.1
L_2	minmax	power	74.2	64.5	66.9	66.1

Table B.1 – *Continued on next page*

Continued from previous page

Normalization method			VQ		LLC	
first	second	third	STIP	iDT	STIP	iDT
L_2	z-score	power	75.0	**73.4**	79.0	70.2
L_2	power	minmax	**79.0**	66.1	79.0	68.5
L_2	power	z-score	70.2	65.3	68.5	71.0
minmax	L_1/sum	power	74.2	68.5	75.8	67.7
minmax	L_2	power	77.4	67.7	74.2	69.4
minmax	power	L_1/sum	29.8	21.8	60.5	57.3
minmax	power	L_2	72.6	50.8	73.4	54.8
z-score	L_1/sum	power	75.8	70.2	78.2	69.4
z-score	L_2	power	76.6	67.7	75.8	66.9
z-score	power	L_1/sum	31.5	29.8	75.8	65.3
z-score	power	L_2	76.6	65.3	77.4	63.7

Table B.1. *Effects of different feature normalization methods on the activity recognition performance on the CAD-120 data set. STIP stands for HOG+HOF encoded Harris3D interest points, and iDT for improved Dense Trajectory features consisting of concatenated HOG, HOF, and MBH descriptors. Both BoW encodings (*i.e.,*VQ and LLC) use the same visual dictionary of size 4000. Note, that LLC encoded features may have negative values, and thus a sum normalization was used instead of L_1 in order to not lose any information.*

C

URADL object annotations

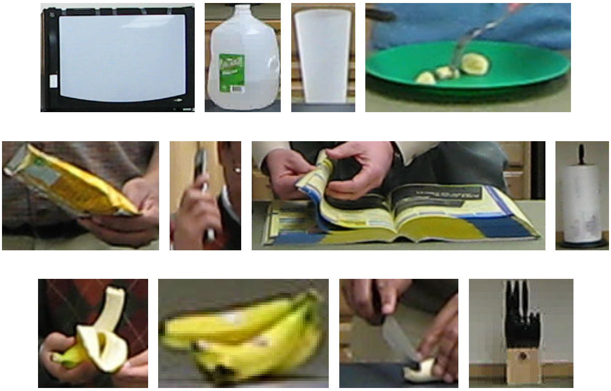


Figure C.1. *Example images of the 12 object categories from the URADL data set that we have manually annotated in order to compare the proposed proto-object based approach with systems relying on object detections. The object classes are:* whiteboard, bottle, cup, plate, crisps, phone, phone-book, paper-roll, peeled banana, bananas, knife, *and* knife-block. *Note, that some objects are not being interacted with during any of the activity categories,* e.g., paper-roll *and* banana. *Nonetheless, they were included in the label-set, since their presence or absence is directly correlated with some activities. For instance, the presence of a bunch of* bananas *indicates the activity of* cutting a banana.

Publications

Ziad Al-Halah, Lukas Rybok, and Rainer Stiefelhagen. Transfer metric learning for action similarity using high-level semantics. *Pattern Recognition Letters*, 72:82–90, 2016.

Manuel Martinez, Lukas Rybok, and Rainer Stiefelhagen. Action recognition in bed using BAMs for assisted living and elderly care. In *IAPR International Conference on Machine Vision Applications (MVA)*, 2015.

Lukas Rybok, Boris Schauerte, Ziad Al-Halah, and Rainer Stiefelhagen. "Important stuff, everywhere!" Activity recognition with salient proto-objects as context. In *Winter Conference on Computer Vision Applications (WACV)*, 2014.

Ziad Al-Halah, Lukas Rybok, and Rainer Stiefelhagen. What to transfer? High-level semantics in transfer metric learning for action similarity. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2014.

Dirk Gehrig, Peter Krauthausen, Lukas Rybok, Hildegard Kuehne, Uwe D. Hanebeck, Tanja Schultz, and Rainer Stiefelhagen. Combined intention, activity, and motion recognition for a humanoid household robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011.

Lukas Rybok, Simon Friedberger, Uwe D. Hanebeck, and Rainer Stiefelhagen. The KIT Robo-kitchen data set for the evaluation of view-based activity recognition systems. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2011.

Lukas Rybok, Michael Voit, Hazım Kemal Ekenel, and Rainer Stiefelhagen. Multi-view based estimation of human upper-body orientation. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2010.

Bibliography

A. Abramov, E. E. Aksoy, J. Dörr, F. Wörgötter, K. Pauwels, and B. Dellen. 3D semantic representation of actions from efficient stereo-image sequence segmentation on GPUs. In *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010. 70

J. K. Aggarwal and S. Park. Human motion: Modeling and recognition of actions and interactions. In *International Symposium on 3D Data Processing, Visualization and Transmission*, 2004. 16

J. K. Aggarwal and M. S. Ryoo. Human activity analysis. *ACM Computing Surveys*, 43(3), 2010. 16

J. K. Aggarwal and L. Xia. Human activity recognition from 3D data: A review. *Pattern Recognition Letters*, 48:70–80, 2014. 16

A. R. Ahad, T. Ogata, J. K. Tan, H. S. Kim, and S. Ishikawa. Motion recognition approach to solve overwriting in complex actions. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2008. 20

A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: Its variants and applications. *Machine Vision and Applications*, 23(2):255–281, 2010. 20

E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, and F. Wörgötter. Learning the semantics of object-action relations by observation. *International Journal of Robotics Research (IJRR)*, 30(10):1229–1249, 2011. 70

Z. Al-Halah, L. Rybok, and R. Stiefelhagen. What to transfer? High-level semantics in transfer metric learning for action similarity. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2014. 136

Z. Al-Halah, L. Rybok, and R. Stiefelhagen. Transfer metric learning for action similarity using high-level semantics. *Pattern Recognition Letters*, 72:82–90, 2016. 15, 136

L. A. Albornoz. Diversity and the film industry. Technical report, UNESCO Institute for Statistics, 2016. 2

B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 34(11):2189 – 2202, 2012. 71

S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 32(2): 288–303, 2010. 22, 24

S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2007. 25, 27

M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 32

R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 109, 110

T. Arikuma and Y. Mochizuki. Intelligent multimedia surveillance system for. *APSIPA Transactions on Signal and Information Processing (SIP)*, 5, 2016. 3

D. Arthur and S. Vassilvirskii. K-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007. 104

T. Asfour, K. Regenstein, P. Azad, J. Schroder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann. Armar-III: An integrated humanoid platform for sensory-motor control. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2006. 10, 90

M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep leraning for human action recognition. In *International Workshop on Human Behavior Understanding (HBU)*, 2011. 47

P. Bach, G. Knoblich, T. C. Gunter, A. D. Friederici, and W. Prinz. Action comprehension: Deriving spatial and functional relations. *Journal of Experimental Psychology: Human Perception and Performance*, 31(3):465–479, 2005. 61

C. D. Barclay, J. E. Cutting, and L. T. Kozlowski. Temporal and spatial factors in gait perception that influence gender recognition. *Perception & Psychophysics*, 23(2):145–152, 1978. 25

A. Bargi, R. Y. Da Xu, and M. Piccardi. An online HDP-HMM for joint action segmentation and classification in motion capture data. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012. 27, 58

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966. 57

P. R. Beaudet. Rotationally invariant image operators. In *IAPR International Conference on Pattern Recognition (ICPR)*, 1978. 32

L. Bell. Monitor alarm fatigue. *American Journal of Critical Care*, 19(1):38, 2010. 4

S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2000. 20

J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram. Human activity recognition using multidimensional indexing. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 24(8): 1091–1104, 2002. 31

A. Berthold, J. Schmidt, and R. Kirchknopf. EBU Technical Review - 2010 Q1. Technical report, European Broadcast-ing Union, 2010. URL https://tech.ebu.ch/docs/techreview/trev_2010-Q1_Mediathek.pdf.

H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 44, 48

P. Bilinski and F. Bremond. Evaluation of local descriptors for action recognition in videos. In *International Conference on Computer Vision System (ICVS)*, 2011. 51

P. Bilinski and F. Bremond. Contextual statistics of space-time ordered features for human action recognition. In *IEEE International Conference on Advanced Video- and Signal-based Surveillance (AVSS)*, 2012. 54

A. Billard, S. Calinon, and R. Dillmann. Learning from human demonstration. In *Handbook of Robotics*. MIT Press, 2013. 5

M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *IEEE International Conference on Computer Vision (ICCV)*, 2005. 19, 21, 75, 76

D. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022, 2003. 56

A. F. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society of London*, 352:1257–1265, 1997. 7

A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 23(3):257–267, 2001. 17, 18, 19, 20

O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 50

P. V. K. Borges, N. Conci, and A. Cavallaro. Video-based human behavior understanding: A survey. *IEEE Transactions on Circuits* *and Systems for Video Technology (TCSVT)*, 23(11):1993–2008, 2013. 16

A. Borreo, L. Onofri, and P. Soda. A multi-environment dataset for activity of daily living recognition in video streams. In *International Conference of the IEEE Engeneering and Biology Society*, 2015. 87

M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 33

M. Bregonzio, J. Li, S. Gong, and T. Xiang. Discriminative topics modelling for action feature selection and recognition. In *British Machine Vision Conference (BMVC)*, 2010. 39, 56

D. N. Bub and M. E. J. Masson. Gestural knowledge evoked by objects as part of conceptual representations. *Aphasiology*, 20(9): 1112–1124, 2006. 61

G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding (CVIU)*, 113(1):48–62, 2009. 34

Z. Cai, L. Wang, X. Peng, and Y. Qiao. Multi-view super vector for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 15, 52, 53

S. Carlsson and J. Sullivan. Action recognition by shape matching to key frames. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2001. 19, 22

J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding (CVIU)*, 117(6):633–659, 2013. 16, 76

K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: An evaluation of recent feature encoding methods. In *British Machine Vision Conference (BMVC)*, 2011. 102, 109, 110

K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference (BMVC)*, 2014. 43, 45

C.-C. Chen and J. K. Aggarwal. Modeling human activities as speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 57, 60

D.-Y. Chen, S.-W. Shih, and H.-Y. M. Liao. Human action recognition using 2-d spatio-temporal templates. In *IEEE International Conference on Multimedia & Expo (ICME)*, 2007. 19, 21

Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 28(12):1931 – 1947, 2006. 24

G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles. Advances in human action recognition: A survey. *arXiv:1501.05964 [cs.CV]*, 2015. 16

W. Cheung and G. Hannarneh. N-SIFT: N-dimensional scale invariant feature transform for matching medical images. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2007. 32

W. Choi and S. Savarese. Learning context for collective activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 59

D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural Networks*, 32:333–338, 2012. 44

O. P. Concha, R. Y. Da Xu, Z. Moghaddam, and M. Piccardi. HMM-MIO: An enhanced hidden Markov model for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2011. 58

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 43, 52

D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2): 215–242, 1958. 52

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004. 50

P. Dagum, A. Galper, and E. Horvitz. Dynamic network models for forecasting. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 1992. 58

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 36, 118, 121

N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision (ECCV)*, 2006. 36

F. de la Torre, J. Hodgins, A. W. Bargteil, X. Martin, J. C. Macey, A. Collado, and P. Beltran. Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database. Technical report, Carnegie Mellon University, 2008. URL http://repository.cmu.edu/robotics/135/. 81, 82

R. de Souza, A. Gaidon, E. Vig, and A. M. Lopez. Sympathy for the details: Dense trajectories and hybrid classification architectures for action recognition. In *ECCV*, 2016. 73

V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2011. 15

B. Dellen, E. E. Aksoy, , and F. Wörgötter. Segment tracking via a spatiotemporal linking process in an n-d lattice model. *Sensors*, 9 (11):9355–9379, 2009. 70

A. P. Dempster, N. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. 50, 104

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 49

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)*, 2010. 15

M. Devanne, S. Berretti, P. Pala, H. Wannous, M. Daoudi, and A. del Bimbo. Motion segment decomposition of RGB-D sequences for human behavior understanding. *Pattern Recognition*, 61:223–233, 2017. 132

T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artifical Intelligence Research*, 2:263–286, 1995. 24

C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2013. 134

P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2006. 33, 36, 52

J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 47

O. Duchenne, I. Laptev, J. Sivic, and F. Bach. Automatic annotation of human actions in video. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 77

J. Ebling and G. Scheuermann. Clifford Fourier transform on vector fields. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 11(4):469–479, 2005. 24

A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision (ICCV)*, 2003. 23

I. Endres and D. Hoiem. Category independent object proposals. In *European Conference on Computer Vision (ECCV)*, 2010. 70

D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 44

V. Escorcia and J. C. Niebles. Spatio-temporal human-object interactions for action recognition in videos. *IEEE International Conference on Computer Vision (ICCV)*, 2013. 111, 132

M. Everingham, S. M. A. Eslami, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015. 52

I. Everts, J. C. van Gemert, and T. Gevers. Evaluation of color STIPs for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 34

G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*, 2003. 101

A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 17, 22, 23, 56

A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 84

C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 73, 79

P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, 59(2):167–181, 2004. 118

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 32(9):1627–1645, 2010. 54, 123

B. Fernando, P. Anderson, M. Hutter, and S. Gould. Discriminative hierarchical rank pooling for activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a. 48, 54

B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Rank pooling for action recognition. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 2016b. 48, 55

Y. Freund and R. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. *Computational Learning Theory*, 55(1):119–139, 1995. 56

D. Gabor. Theory of communication. *Journal of the Institute of Electrical Engineers - Part III*, 93(26):429–457, 1946. 41

V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti. Action recognition in the premotor cortex. *Brain: A Journal of Neurology*, 119: 593–609, 1996. 61

H. G. Gauch Jr. *Scientific Method in Practice*. Cambridge University Press, 2003. 74

D. Gehrig and T. Schultz. Selecting relevant features for human motion recognition. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2008. 31

D. Gehrig, P. Krauthausen, L. Rybok, H. Kuehne, U. D. Hanebeck, T. Schultz, and R. Stiefelhagen. Combined intention, activity, and motion recognition for a humanoid household robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011. 58, 62, 83, 87, 135

C. Geib, K. Mourao, R. Petrick, N. Pugeault, M. Steedman, N. Krueger, and F. Wörgötter. Object action complexes as an interface for planning and robot control. *IEEE-RAS International Conference on Humanoid Robots (Humanoids) Workshops*, 2006. 8

T. Gevers and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 32(9):1582 – 1596, 2010. 34

M. A. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nature Review Neuroscience*, 4(3):179–192, 2003. 25

M. Glodek, F. Schwenker, and G. Palm. Detecting actions by integrating sequential symbolic and sub-symbolic information in human activity recognition. In *International Conference on Machine Learning and Data Mining (MLDM)*, 2012. 58

C. Goodal. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):285–339, 1991. 20

M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. 40, 119

A. Grushin, D. D. Monner, J. A. Reggia, and A. Mishra. Robust human action recognition via long short-term memory. In *Inter-*

national Joint Conference on Neural Networks (IJCNN), 2013. 47

T. Guha and R. Ward. Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 34(8):1576–88, 2012. 52

G. Guo and A. Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014. 16

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)*, 2007. 58, 81, 87

A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 31(10):1775–1789, 2009. 66

R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C. Isbell. A novel sequence representation for unsupervised analysis of human activities. *Artificial Intelligence*, 173(14):1221–1244, 2009. 62

M. H. M. Hanheide, N. H. N. Hofemann, and G. S. G. Sagerer. Action recognition in a wearable assistance system. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2006. 84

C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference (AVC)*, 1988. 100

T. Hassner. A critical review of action recognition benchmarks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013. 16, 76

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human level performance on ImageNet classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 43, 45

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 62, 68

P. H. Hennings-Yeomans, B. V. K. Vijaya Kumar, and M. Savvides. Palmprint classification using multiple advanced correlation filters and palm-specific segmentation. *IEEE Transactions on Information Forensics and Security*, 2(3):613–622, 2007. 24

S. Herath, M. Harandi, and F. Porikli. Going deeper into action recognition: A survey. *arXiv:1605.04988 [cs.CV]*, 2016. 16

T. K. Ho. Random decision forests. In *International Conference on Document Analysis and Recognition (ICDAR)*, 1995. 50

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(6):1735–1780, 1997. 47

T. Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999. 55

X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 34(1):194–201, 2011. 116, 117, 118

J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang. Exemplar-based recognition of human-object interactions. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, PP(99), 2015. 66, 71

M.-K. Hu. Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, 8(2):179–187, 1962. 20

Z. F. Huang, W. Yang, Y. Wang, and G. Mori. Latent boosting for action recognition. In *British Machine Vision Conference (BMVC)*, 2011. 56

M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013. 28

N. Ikizler and D. A. Forsyth. Searching for complex human activities with no visual examples. *International Journal of Computer Vision (IJCV)*, 80(3):337–357, 2008. 57

N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *European Conference on Computer Vision (ECCV)*, 2010. 69

T. S. Jaakola and D. Haussler. Exploiting generative models in discriminative classifiers. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 1999. 106

M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 36, 53

M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *IEEE International Conference on Computer Vision (ICCV)*, 2015a. 63

M. Jain, J. C. van Gemert, and C. G. M. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015b. 62

K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 45

H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 34(9):1704–1716, 2012. 52, 53

N. Jenkins. Video surveillance camera installed base report. Technical report, IHS Markit, 2015. 3

H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *European Conference on Computer Vision (ECCV)*, 2007. 40, 41

H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 79

S. Ji, M. Yang, K. Yu, and W. Xu. 3D convolutional neural networks for human action recognition. In *International Conference on Machine Learning (ICML)*, 2010. 45

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia (ACM-MM)*, 2014. 45

Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2011b. 46

G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. 24, 25

I. Junejo, E. Dexter, I. Laptev, and P. Perez. View-independent action recognition from temporal self-similarities. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 33: 172–185, 2010. 15

F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2005. 34

T. Kadir and M. Brady. Scale saliency: A novel approach to salient feature and scale selection. In *International Conference on Visual Information Engineering (VIE)*, 2003. 32

V. Kantorov and I. Laptev. Efficient feature extraction, encoding, and classification for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 37, 53

S. Karaman, L. Seidenari, S. Ma, A. del Bimbo, and S. Scarloff. Adaptive structures pooling for action recognition. In *British Machine Vision Conference (BMVC)*, 2014. 54

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 43, 46, 47, 79

Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004. 36

Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *IEEE International Conference on Computer Vision (ICCV)*, 2005. 38, 39

V. Kellokumpu, G. Zhao, and M. Pietikäinen. Recognition of human actions using texture descriptors. *Machine Vision and Applications*, 22(5):767–780, 2011. 38

T.-k. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 31(8), 2009. 56

R. Kindermann and J. L. Snell. *Markov random fields and their applications*. American Mathematical Society, 1980. 59

H. Kjellström, J. Romero, D. Martínez, and D. Kragić. Simultaneous visual recognition of manipulation actions and manipulated objects. In *European Conference on Computer Vision (ECCV)*, 2008. 59, 67

H. Kjellström, J. Romero, and D. Kragić. Visual object-action recognition: Inferring object affordances from human demonstration.

Journal on Computer Vision and Image Understanding (CVIU), 115(1):81–90, 2011. 67

A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *British Machine Vision Conference (BMVC)*, 2008. 33, 37, 52

O. Kliper-Gross, T. Hassner, and L. Wolf. The action similarity labeling challenge. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 34(3):615–621, 2012a. 80

O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *European Conference on Computer Vision (ECCV)*, 2012b. 35, 39

M. Koperski and F. Bremond. Modeling spatial layout of features for real world scenario rgb-d action recognition. In *IEEE International Conference on Advanced Video- and Signal-based Surveillance (AVSS)*, 2016. 132

H. S. Koppula and A. Saxena. Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation. In *International Conference on Machine Learning (ICML)*, 2013a. 62, 112, 132

H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *Robotics: Science and Systems Conference (RSS)*, 2013b. 66

H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *International Journal of Robotics Research (IJRR)*, 32(8):951–970, 2013. 59, 62, 82, 85, 125, 132

L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 15

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2012. 44, 62

N. Krueger, J. Piater, F. Wörgötter, C. Geib, R. Petrick, M. Steedman, A. Ude, T. Asfour, D. Kraft, D. Omrcen, B. Hommel, A. Agostini, D. Kragic, J.-O. Eklundh, V. Krueger, C. Torras, and R. Dillmann. A formal definition of object-action complexes and examples at different levels of the processing hierarchy. Technical report, PACO-PLUS, 2009. URL http://www.paco-plus.org. 8

V. Krüger, D. Kragić, A. Ude, and C. Geib. The meaning of action: A review on action recognition. *Advanced Robotics*, 21(13):1473–1501, 2007. 16

H. Kuehne and T. Serre. An end-to-end generative framework for video segmentation and recognition. *IEEE Winter Conference on Computer Vision Applications (WACV)*, 2016. 53, 57, 60, 132

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. 79, 80

H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 58, 84, 86

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, 2001. 59

T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 54

T. Lan, Y. Zhu, A. R. Zamir, and S. Savarese. Action recognition by hierarchical mid-level action elements. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015a. 69, 70

Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj. Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015b. 35, 54

I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision (ICCV)*, 2003. 32, 74, 97, 98, 99

I. Laptev and P. Pérez. Retrieving actions in movies. In *IEEE International Conference on Computer Vision (ICCV)*, 2007. 36, 56, 75, 77

I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 51, 52, 54, 75, 77, 100, 101

A. Laurentini. The visual hull concept for silhouette based image understanding. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 16(2):150–162, 1994. 22

K. Laver, S. Dyer, C. Whitehead, L. Clemson, and M. Crotty. Interventions to delay functional decline in people with dementia: A systematic review of systematic reviews. *Geriatric medicine*, 6(4), 2016. 4

B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 58

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)*, 2006. 28, 32, 53

Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 44, 45

Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature Methods*, 13(1):35–35, 2015. 43

Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 85

B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004. 54

L.-j. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2010a. 21

W. Li, Z. Zhang, and Z. Liu. Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT*, 18(11):1499–1510, 2008. 23

W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010b. 22, 78, 79

Z. Li, E. Gavves, M. Jain, and C. G. M. Snoek. VideoLSTM convolves, attends and flows for action recognition. *arXiv:1607.01794 [cs.CV]*, 2016. 47, 61

B. Liang and L. Zheng. A survey on human action recognition using depth sensors. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2015. 16, 22

L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang. A deep structured model with radius–margin bound for 3D human activity recognition. *International Journal of Computer Vision (IJCV)*, 118(2):256–273, 2016. 131, 132

Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 75, 76

H. Liu, Q. Zhang, and Q. Sun. Human action classification based on sequential bag-of-words model. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2014a. 54

J. Liu, X. Wu, and Y. Feng. Modeling the relationship of action, object, and scene. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2014b. 60, 67

J. Liu, L. Jiebu, and M. Shah. Recognizing realistic actions from videos "in the wild". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 56, 78

D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, 1999. 32, 36, 37, 42

J. Lu, R. Xu, and J. J. Corso. Human action segmentation with hierarchical supervoxel consistency. In *CVPR*, 2015. 59

B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 3, 1981. 39, 100

J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 29, 52

F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using HMM and multi-class AdaBoost. In *European Conference on Computer Vision (ECCV)*, 2006. 25, 27

M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 64, 67

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1967. 50, 103

S. Marcelja. Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America*, 70(11): 1297–1300, 1980. 41

M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 77

P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *ICCV Workshop on Video-oriented Object and Event Classification*, 2009. 38, 39

P. Matikainen, M. Hebert, and R. Sukthankar. Representing pairwise spatial and temporal relations for action recognition. In *European Conference on Computer Vision (ECCV)*, 2010. 40, 132

R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 9, 39, 56, 61, 83, 122

K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision (ECCV)*, 2002. 32

K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2005. 32

T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):90–126, 2006. 7, 8, 16, 26

D. J. Moore, I. A. Essa, and M. H. Hayes. Exploiting human actions and object context for recognition tasks. In *IEEE International Conference on Computer Vision (ICCV)*, 1999. 68

B. T. Morris and M. M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 18(8):1114–1127, 2008. 16

MPAA. 2014 theatrical market statistics. Technical report, Motion Picture Association of America (MPAA), 2015. 1

H.-H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2):59–74, 1988. 7

K. Nelissen, G. Luppino, W. Vanduffel, G. Rizzalotti, and G. A. Orban. Observing others: Multiple action representation in the frontal lobe. *Science*, 310(5746):332–336, 2005. 61

J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 47

B. Ni, V. R. Paramathayalan, and P. Moulin. Multiple granularity analysis for fine-grained action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 65

B. Ni, X. Yang, and S. Gao. Progressively parsing interactional objects for fine grained action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 64

J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 54

J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision (IJCV)*, 79(3):299–318, 2008. 56

J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision (ECCV)*, 2010. 79, 80

F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 25(1):24–38, 2014. 26, 31

E. Ohn-Bar and M. M. Trivedi. Joint angles similarities and HOG2 for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013. 31

A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal saliency for human action recognition. In *IEEE International Conference on Multimedia & Expo (ICME)*, 2005. 32

T. Ojala, M. Pietikainen, and T. Mäenpää. Multiresolution grayscale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 24(7):971–987, 2002. 37

D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with Fisher vectors on a compact feature set. *IEEE International Conference on Computer Vision (ICCV)*, 2013. 53

D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection proposals. In *European Conference on Computer Vision (ECCV)*, 2014a. 134

D. Oneata, J. Verbeek, and C. Schmid. Efficient action localization with approximately normalized Fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014b. 53

L. Onofri, P. Soda, and G. Iannello. Multiple subsequence combination in human action recognition. *IET Computer Vision*, 2013. 132

L. Onofri, P. Soda, M. Pechenizkiy, and G. Iannello. A survey on using domain and contextual knowledge for human activity recognition in video streams. *Expert Systems with Applications*, 2016. 16

O. Oreifej and Z. Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 22, 23

B. Packer, K. Saenko, and D. Koller. A combined pose, object, and feature model for action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 60, 65

PACO-PLUS. PACO-PLUS: Perception, action and cognition through learning of object-action complexes, 2006. URL http://www.paco-plus.org. [Online; accessed 20-February-2017]. 8

V. Parameswaran and R. Chellappa. View invariants for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 27, 28

S. Park and J. K. Aggarwal. A hierarchical Bayesian network for event recognition of human actions and interactions. *Multimedia Systems*, 10(2):164–179, 2004. 58

A. Patron-Perez, M. Marszałek, A. Zisserman, and I. Reid. High five: Recognising human interactions in tv shows. In *British Machine Vision Conference (BMVC)*, 2010. 75, 77

D. J. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. In *International Symposium on Wearable Computers*, 2005. 62, 63

H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 27(8):1226–1238, 2005. 120

X. Peng, L. Wang, Y. Qiao, and Q. Peng. A joint evaluation of dictionary learning and feature encoding for action recognition. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2014a. 51

X. Peng, L. Wang, Y. Qiao, and Q. Peng. Boosting VLAD with supervised dictionary learning and high-order statistics. In *European Conference on Computer Vision (ECCV)*, 2014b. 53

X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked Fisher vectors. In *European Conference on Computer Vision (ECCV)*, 2014c. 53, 79

X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding (CVIU)*, pages 1–17, 2016. 51, 52, 109, 112

F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorizaton. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 52, 106

F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, 2010. 53, 107, 109, 121

M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Hähnel, D. Fox, and H. Kautz. Inferring ADLs from interactions with objects. *IEEE Pervasive Computing*, 3(4):50–57, 2004. 63

H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 84, 85

R. Polana and R. C. Nelson. Detection and recognition of periodic, nonrigid motion. *International Journal of Computer Vision (IJCV)*, 23(3):261–282, 1997. 23

R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding (CVIU)*, 108(1-2), 2007. 16

R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010. 16

A. Prest, V. Ferrari, and C. Schmid. Explicit modeling of human-object interactions in realistic videos. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 35(4): 835–848, 2012a. 132

A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 34(3): 601–614, 2012b. 71

A. Quattoni and S. Wang. Hidden conditional random fields. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 29(10):1848–1853, 2007. 59

H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Action classification with locality-constrained linear coding. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2014. 52

N. Raman and S. J. Maybank. Action classification using a discriminative multilevel HDP-HMM. *Neurocomputing*, 154:149–161, 2015. 29, 58

D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2003. 58

K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013. 78

X. Ren and D. Ramanan. Histograms of sparse codes for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 109, 110

R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1-3):17–42, 2000b. 11

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Reviews Neuroscience*, 2:1019–1025, 1999. 40

M. R. Robertson. 500 Hours of Video Uploaded To YouTube Every Minute [Forecast], 2015. URL http://tubularinsights.com/hours-minute-uploaded-youtube. [Online; accessed 20-November-2016]. 2

M. Rodriguez, C. Orrite, C. Medrano, and D. Makris. A time flexible kernel framework for video-based activity recognition. *Image and Vision Computing*, 48-49:26–36, 2016. 61

M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 17, 22, 24, 77, 79, 80

D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Prikl, F. Wagner, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, M. Creatura, and J. Millan. Walk-through the OPPORTUNITY dataset for activity recognition in sensor rich environments. In *International Conference on Pervasive Computing*, 2010. 81, 83

M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *European Conference on Computer Vision (ECCV)*, 2012a. 88

M. Rohrbach, A. Sikandar, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012b. 26

M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision (IJCV)*, 2015. 63, 84, 88

N. Rostamzadeh, G. Zen, I. Mironica, J. Uijlings, and N. Sebe. Daily living activities recognition via efficient high and low level cues combination and Fisher kernel representation. In *International Conference on Image Analysis and Processing (ICIAP)*, 2013. 53, 132

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 52, 68

L. Rybok, S. Friedberger, U. D. Hanebeck, and R. Stiefelhagen. The KIT Robo-kitchen data set for the evaluation of view-based activity recognition systems. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2011. 12, 88, 109, 132

L. Rybok, B. Schauerte, Z. Al-Halah, and R. Stiefelhagen. "Important stuff, everywhere!" Activity recognition with salient proto-objects as context. In *IEEE Winter Conference on Computer Vision Applications (WACV)*, 2014. 13

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)*, 2009. 61

S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *IEEE Conference on Computer Vision and*

Pattern Recognition (CVPR), 2012. 18, 21

R. Y. Schapire and Y. Singer. Improved boosting algorithm using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999. 24

B. Schauerte and R. Stiefelhagen. Quaternion-based spectral saliency detection for eye fixation prediction. In *European Conference on Computer Vision (ECCV)*, 2012b. 116, 118

K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 17, 41, 42

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. 43

O. C. Schrempf, A. Hanselmann, and U. D. Hanebeck. Efficient representation and fusion of hybrid joint densities for clusters in nonlinear hybrid Bayesian networks. In *International Conference on Information Fusion (FUSION)*, 2006. 58

C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2004. 75, 76

P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In *ACM Multimedia (ACM-MM)*, 2007. 33, 37, 52

L. Seidenari, V. Varano, S. Berretti, A. del Bimbo, and P. Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013. 27, 29

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2014. 45

T. Serre and T. Poggio. A neuromorphic approach to computer vision. *Communications of the ACM*, 53(10):54, 2010. 40

T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 29(3): 411–426, 2007. 40

L. Shao and R. Mattivi. Feature detector and descriptor evaluation in human action recognition. In *ACM International Conference on Image and Video Retrieval*, 2010. 38, 51

L. Shao, X. Zhen, D. Tao, and X. Li. Spatio-temporal Laplacian pyramid coding for action recognition. *IEEE Transactions on Cybernetics*, 44(6):817–827, 2014. 41, 42

F. Shi, R. Laganiere, and E. Petriu. Gradient boundary histograms for action recognition. In *IEEE Winter Conference on Computer Vision Applications (WACV)*, 2015. 37

J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 39

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 26, 27

L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1-2):4–27, 2010. 77

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 45

K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2014. 46, 47, 48

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 43, 44, 45

S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *European Conference on Computer Vision (ECCV)*, 2012. 134

J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2003. 32, 50, 104

C. Sminchisescu, A. Kanaujia, and D. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):210–220, 2006. 59

K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. Technical report, Center for Research in Computer Vision, 2012. URL http://arxiv.org/abs/1212.0402. 78, 80

E. H. Spriggs, F. de la Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2009. 83

A. Srikantha and J. Gall. Discovering object classes from activities. In *European Conference on Computer Vision (ECCV)*, 2014. 71

N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *British Machine Vision Conference (BMVC)*, 2015. 47

S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2013. 84, 86

P. Stone, R. Brook, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. Saxenian, J. Shah, M. Tambe, and A. Teller. Artificial intelligence and life in 2030. Technical report, Stanford University, 2016. 4

W. T. Su, Y. H. Lu, and A. S. Kaseb. Harvest the information from multimedia big data in global camera networks. In *IEEE International Conference on Multimedia Big Data (BigMM)*, 2015. 3

G. Summerfield. BBC iPlayer on TV in your living room: update, 2011. URL http://www.bbc.co.uk/blogs/bbcinternet/2011/04/bbc_iplayer_on_tv_update.html. [Online; accessed 20-November-2016]. 2

J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009a. 39

L. Sun, U. Klank, and M. Beetz. EYEWATCHME - 3D hand and object tracking for inside out activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2009b. 85

Q. Sun, H. Liu, L. Ma, and T. Zhang. A novel hierarchical bag-of-words model for compact action representation. *Neurocomputing*, 174:722–732, 2016. 35, 54

S. Sundaram and W. W. M. Cuevas. High level activity recognition using low resolution wearable vision. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2009. 85

J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from RGBD images. In *AAAI Workshop on Pattern, Activity and Intent Recognition*, 2011. 17, 58, 85

J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from RGBD images. In *IEEE-RAS International Conference on Robotics and Automation (ICRA)*, 2012. 132

C. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8:693–723, 2007. 59

R. Szeliski. Image alignment and stitching: A turorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2006. 35

K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 58

G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *European Conference on Computer Vision (ECCV)*, 2010. 44, 45

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2005. 58

M. Tenorth, J. Bandouch, and M. Beetz. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 78, 80

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 35, 45

R. Trichet and R. Nevatia. Video segmentation with spatio-temporal tubes. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2013. 134

J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007. 53

P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 18(11): 1473–1488, 2008. 16

T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: A survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2007. 31

T. Tuytelaars and L. van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision (IJCV)*, 59(1):61–85, 2004. 32

M. Umeda. Recognition of multi-font printed Chinese characters. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1982. 19

R. R. Vallacher and D. M. Wegner. What do people think they're doing? Action identification and human behavior. *Psychological Review*, 94(1):3–15, 1987. 11, 115

J. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 32(7):1271 – 1283, 2010. 53

R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3D skeletons as points in a Lie group. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 27, 28

P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2004. 39, 56

S. Vishwakarma and A. Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10):983–1009, 2013. 16

N. N. Vo and A. F. Bobick. From stochastic grammar to Bayes network: Probabilistic parsing of complex activity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 58

M. Vrigkas, C. Nikou, and I. Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2(28), 2015. 16

M. Wächter and T. Asfour. Hierarchical segmentation of manipulation action based on object relations and motion characteristics. In *IEEE International Conference on Advanced Robotics (ICAR)*, 2015. 135

C. Wallraven, M. Schultze, B. Mohler, A. Vatakis, and K. Pastra. The POETICON enacted scenario corpus - A tool for human and computational experiments on action understanding. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2011. 81, 84

D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural networks*, 19(9):1395–1407, 2006. 11

H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 34, 47, 51, 53, 74, 98, 101, 102

H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference (BMVC)*, 2009. 34, 51, 52, 101

H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011a. 34, 36, 39, 101, 112

J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012a. 29

J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012b. 85, 87

J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 52, 105, 106

K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo. 3D human activity recognition with reconfigurable convolutional neural networks. In *ACM Multimedia (ACM-MM)*, 2014a. 47, 131, 132

L. Wang and D. Suter. Informative shape representations for human action recognition. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2006. 17, 18, 20

L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 59

L. Wang, Y. Qiao, and X. Tang. Latent hierarchical model of temporal structure for complex activity classification. *IEEE Transactions on Image Processing*, 23(2):810–822, 2014b. 54, 61

L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015a. 35, 47, 48, 79

L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv:1507.02159 [cs.CV]*, 2015b. 46

L. Wang, Y. Qiao, and X. Tang. MoFap: A multi-level representation for action recognition. *International Journal of Computer Vision (IJCV)*, 119(3):254–271, 2016b. 61

X. Wang, L. Wang, and Y. Qiao. A comparative study of encoding, pooling and normalization methods for action recognition. In *IEEE Asian Conference on Computer Vision (ACCV)*, 2012c. 51

Y. Wang and G. Mori. Learning a discriminative hidden part model for human action recognition. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2008. 59

Y. Wang and G. Mori. Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 31(10):1762–1774, 2009. 56

Y. Wang and G. Mori. Hidden part models for human action recognition: Probabilistic vs. max-margin. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 33(7):1310–1323, 2011. 59

P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. Modeling 4D human-object interactions for event and object recognition. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 2016. 65

D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 19

D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):249–257, 2006. 18, 22, 77, 78

D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. In *IEEE International Conference on Computer Vision (ICCV)*, 2007a. 57

D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding (CVIU)*, 115(2):224–241, 2011. 16, 25

P. J. Werbos. Applications of advances in nonlinear sensitivity analysis. In *IFIP Conference on Systems Modeling and Optimization*, 1982. 42

T. Weyand, I. Kostrikov, and J. Philbin. PlaNet - Photo geolocation with convolutional neural networks. *arXiv:1602.05314 [cs.CV]*, 2016. 45

G. Willems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European Conference on Computer Vision (ECCV)*, 2008. 32, 39

D. H. Wilson, S. Consolvo, K. Fishkin, and M. Philipose. In-home assessment of the activities of daily living of the elderly. In *ACM Conference on Human Factors in Computing Systems (CHI) Workshops*, 2005. 4

H. R. Wilson and J. R. Bergen. A four mechanism model for threshold spatial vision. *Vision Research*, 19(1):19–32, 1978. 42

A. P. Witkin. Scale-space filtering. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1983. 99

C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C. E. Bichot, C. Garcia, and B. Sankur. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding (CVIU)*, 127:14–30, 2014a. 78, 80

L. Wolf, Y. Hanani, and T. Hassner. A piggyback representation for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014b. 54

S.-F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. In *IEEE International Conference on Computer Vision (ICCV)*, 2007. 34

S.-F. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 56

J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A scalable approach to activity recognition based on object use. In *IEEE International Conference on Computer Vision (ICCV)*, 2007. 62, 63

J. Wu, Y. Zhang, and W. Lin. Towards good practices for action video encoding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 29, 51, 58

X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia. Action recognition using multilevel features and latent structural SVM. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 23(8): 1422–1431, 2013. 60

Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, X. Xue, and J. Wang. Fusing multi-stream deep networks for video classification. *arXiv:1509.06086 [cs.CV]*, 2015. 43, 46

L. Xia, C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012. 26, 27, 28, 57

X. Xu, T. Hospedales, and S. Gong. Semantic embedding space for zero-shot action recognition. In *IEEE International Conference on Image Processing (ICIP)*, 2015. 15

J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1992. 19, 57

J. Yang, K. Yu, Y. Gong, and Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 52

S. Yang, C. Yuan, W. Hu, and X. Ding. A hierarchical model based on latent Dirichlet allocation for action recognition. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2014. 56

X. Yang and Y. Tian. Eigenjoints-based action recognition using naïve-Bayes-nearest-neighbor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 28, 29

X. Yang and Y. Tian. Action recognition using super sparse coding vector with spatio-temporal awareness. In *European Conference on Computer Vision (ECCV)*, 2014. 52

X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *ACM Multimedia (ACM-MM)*, 2012. 22

A. Yao, J. Gall, and L. van Gool. Coupled action recognition and pose estimation from multiple views. *International Journal of Computer Vision (IJCV)*, 100(1):16–37, 2012. 30

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)*, 2010b. 70, 71

B. Yao, A. Khosala, and L. Fei-Fei. Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. In *International Conference on Machine Learning (ICML)*, 2011. 67

M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall. A survey on human motion analysis from depth data. In *German Conference on Pattern Recognition (GCPR)*, 2013. 16, 22

L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 35, 38

Y. Yi and Y. Lin. Human action recognition with salient trajectories. *Signal Processing*, 93(11):2932–2941, 2013. 111, 132

J. Yin and Y. Meng. Human activity recognition in video using a hierarchical probabilistic latent model. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010. 56

C.-N. J. Yu and T. Joachims. Learning structural SVMs with latent variables. In *International Conference on Machine Learning (ICML)*, 2009. 59, 60

G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *IEEE Asian Conference on Computer Vision (ACCV)*, 2015. 87

K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *IEEE Conference on Neural Information Processing Systems (NIPS)*, 2009. 53

T.-H. Yu, T.-K. Kim, and R. Cipolla. Real-time action recognition by spatiotemporal semantic and structural forests. In *British Machine Vision Conference (BMVC)*, 2010. 38, 39

M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 27, 28

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014. 46

Z. Zeng and Q. Ji. Knowledge based activity recognition with dynamic Bayesian network. In *European Conference on Computer Vision (ECCV)*, 2010. 58

H. Zhang, W. Zhou, C. Reardon, and L. E. Parker. Simplex-based 3D spatio-temporal feature description for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 33, 37

J. Zhang and S. Gong. Action categorization by structural probabilistic latent semantic analysis. *Computer Vision and Image Understanding (CVIU)*, 114(8):857–864, 2010a. 56

J. Zhang and S. Gong. Action categorization with modified hidden conditional random field. *Pattern Recognition*, 43(1):197–203, 2010b. 59

T. Zhang, D. Tao, X. Li, and J. Yang. Marginal patch alignment for dimensionality reduction. *IEEE Transactions on Knowledge and Data Engeneering*, 21(9):1299–1313, 2009. 42

Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu. Contextual Fisher kernels for human action recognition. In *IAPR International Conference on Pattern Recognition (ICPR)*, 2012. 54

Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia. Motion context: A new representation for human action recognition. In *European Conference on Computer Vision (ECCV)*, 2008. 18, 20, 56

G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 29(6):915 – 928, 2007. 38

X. Zhen and L. Shao. Action recognition via spatio-temporal local features: A comprehensive study. *Image and Vision Computing*, 50, 2016. 51

H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004. 15

F. Zhou, F. de la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI)*, 35(3):582–596, 2013. 135

H. Zhou, Y. Yuan, and C. Shi. Object tracking using SIFT features and mean shift. *Computer Vision and Image Understanding (CVIU)*, 113(3):345–352, 2009. 32

Y. Zhou, B. Ni, R. Hong, X. Yang, and Q. Tian. Cascaded interactional targeting network for egocentric video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 64, 65

F. Zhu, L. Shao, J. Xie, and Y. Fang. From handcrafted to learned representations for human action recognition: A survey. *Image and Vision Computing*, 2016. 16, 45

Y. Zhu, W. Chen, and G. Guo. Fusing spatiotemporal features and joints for 3D action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013. 26, 29

D. Zongker. Chicken chicken chicken. *Annals of Improbable Research*, 12(5):16–21, 2006.

Object knowledge is an important cue to distinguish between human activities, but nevertheless usually disregarded in video-based activity recognition systems. In contrast, the aim of this work is to explore ways how to boost activity recognition performance by augmenting motion features with object information. Instead of relying on supervised detectors, the proposed object representation is motivated by a key mechanism of visual perception: saliency detection. Saliency detection serves as a gating mechanism selecting which information to process. It thus allows us, humans, to focus our visual attention on certain regions even before we identify them as actual objects. The proposed proto-object features are based on computational models implementing such an attentional process making the representation independent of statistical knowledge about objects. A major advantage of the present approach is, therefore, its ability to be transferred across domains without the explicit necessity of learning new object models.

0 0 1 2 8 2 1 6 8 3

ISBN 978-3-7450-1072-5

9 783745 010725