PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

ACTIVITY RECOGNITION IN RGB-D VIDEOS USING HIERARCHICAL AND COMPOSITIONAL ENERGY-BASED MODELS

IVÁN ALBERTO LILLO VALLÉS

Thesis submitted to the Office of Graduate Studies
in partial fulfillment of the requirements for the Degree of
Doctor in Engineering Science

Advisor:
ÁLVARO SOTO

Santiago de Chile, October, 2018

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

ACTIVITY RECOGNITION IN RGB-D VIDEOS USING HIERARCHICAL AND COMPOSITIONAL ENERGY-BASED MODELS

IVÁN ALBERTO LILLO VALLÉS

Members of the Committee:
ÁLVARO SOTO
HANS LÖBEL
CRISTIÁN TEJOS
RENÉ VIDAL
JUAN CARLOS NIEBLES
GUSTAVO LAGOS

Thesis submitted to the Office of Graduate Studies
in partial fulfillment of the requirements for the Degree of
Doctor in Engineering Science

Santiago de Chile, October, 2018

To my lovely wife, Sandra, and
Isidora and Ignacio, our beautiful
children

ACKNOWLEDGMENTS

First, I sincerely thank my wife, Sandra. We start as a couple when I was starting as a PhD student. We lived together all the ups and downs, the moments of frustration and happiness. She was always supportive and patience. Now in this last step of the thesis, we have two beautiful children.

I also thank my advisor Professor Alvaro Soto, who always has the correct advice but let me run with my own ideas most of the time. He was a strong support along the complete thesis, and thanks to him, I am now a passionate of Artificial Intelligence. I would also like to thank Professor Juan Carlos Niebles, who co-authored all the papers related to this thesis. Without his great support, vision, advice and experience, it would not have been possible to obtain these results. I also thank Professor Domingo Mery, who during the first stage of this process was a co-advisor, and (now) Professor Hans Löbel, who became a PhD student six months before me and served me as a guide till now.

I thank my colleagues of the IA-Lab research group, orderless: Gabriel, Alejandro, Miguel, Julio y Felipe, for hours of prolific conversation, academic discussion and off-academic chattering, which were an important part of my life during these years.

This work was partially funded by FONDECYT grant 1120720. Ivan Lillo received funds from CONICYT-PCHA/Doctorado Nacional/2014-21120278.

Contents

List of Tables

List of Figures

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

RECONOCIMIENTO DE ACTIVIDADES EN VIDEOS RGB-D USANDO MODELOS JERARQUICOS COMPOSICIONALES BASADOS EN FUNCIONES DE ENERGIA

Tesis enviada a la Dirección de Postgrado en cumplimiento parcial de los requisitos para el grado de Doctor en Ciencias de la Ingeniería.

IVÁN ALBERTO LILLO VALLÉS

RESUMEN

El reconocimiento de actividades humanas en videos ha ganado gran interés en los últimos años. Varios métodos han sido propuestos, con diferente complejidad dependiendo del largo temporal de los videos, la modalidad de captura para adquirirlos, y el número de acciones ejecutadas por personas en una escena, entre otros. En este escenario, el reconocimiento de actividades complejas ha emergido como un tópico de activa investigación, ya que las personas pueden ejecutar múltiples acciones concurrentes tanto espacial como temporalmente en la misma escena.

Esta tesis se enfoca en el reconocimiento de actividades complejas usando cámaras RGB-D, las cuales poseen sensores de profundidad que permiten capturar video RGB (apariencia) e información de profundidad en tiempo real en ambientes de interior (indoor). La estimación de pose 3D de las articulaciones de un cuerpo humano (esqueleto) está incluido en el software proveído por estos dispositivos, lo que ha hecho aumentar la investigación basada en poses 3D de esqueletos.

Nuestro foco es el reconocimiento de actividades complejas, compuestas de acciones atómicas secuenciales y/o simultáneas, las que a su vez están compuestas por poses y movimientos de bajo nivel, enfocando el modelo en los movimientos de un sólo actor a la vez.

Nuestra contribución es la creación de un modelo jerárquico-composicional en tres niveles de abstracción. En el nivel inferior, características geométricas y de movimiento son usadas para aprender automáticamente un diccionario de poses, cuyas entradas son usadas para codificar segmentos temporales de acciones atómicas a nivel de cuadro de video. En el nivel intermedio, composiciones de elementos del diccionario de poses, por separado en cada región definida del cuerpo, son usadas para representar acciones atómicas, con una acción distinta para cada región, y donde además cada región se representa como una secuencia temporal de una o varias acciones atómicas. Finalmente, en el nivel superior, composiciones espaciales y temporales de acciones atómicas son ensambladas para representar actividades complejas, donde una actividad compleja es asignada a cada video.

El proceso de aprendizaje de los parámetros del modelo es planteado como una optimización de función de energía, usando una formulación de máximo margen, donde cada pose y acción atómica es modelada como un clasificador lineal.

Se presenta en esta tesis un modelo jerárquico base, el cual obtiene resultados satisfactorios en una base de datos de actividades complejas (Composable Activities Dataset). Adicionalmente, numerosas mejoras al modelo base son introducidas: (i) un cambio en representación de los clasificadores lineales de las acciones atómicas, que producen clasificadores ralos, donde las poses se especializan en pocas acciones atómicas; (ii) desde el video RGB, se extraen características de movimiento dentro de un pequeño lapso temporal, el cual se añade a las características geométricas del modelo base; (iii) se elabora una formulación alternativa más escalable, que no necesita de anotaciones espaciales de acciones atómicas, conservando sólo la supervisión temporal durante el entrenamiento; (iv) un modelo que incorpora flexibilidad de ejecución de poses y acciones atómicas, introduciendo *motion poselets* y *actionlets*; y (v) mecanismo para descartar poses no informativas, lo cual incrementa la robustez a errores comunes de estimación de pose.

Los experimentos realizados muestran los beneficios de usar un enfoque jerárquico que utiliza la composición de poses en acciones atómicas, y éstas en actividades complejas. En particular, el modelo resultante es capaz de identificar los intervalos temporales y las regiones espaciales donde ocurren las acciones atómicas, teniendo la interesante propiedad

de que la salida del modelo provee de información intermedia semántica, en conjunto con una clasificación de la actividad del video completo en el nivel superior.

El rendimiento de los métodos propuestos es evaluado usando múltiples bases de datos de reconocimiento de acciones. El modelo propuesto supera consistentemente modelos del estado del arte para reconocimiento de acciones complejas, mostrando cómo un modelo jerárquico y composicional es clave para inferir interacciones complejas usando representaciones semánticas simples como bloques constitutivos, que en nuestro caso son las poses inferidas y las acciones atómicas.

Palabras Claves: Reconocimiento de acciones, detección de acciones, modelo jerárquico, predicción estructural.

Miembros de la Comisión de Tesis Doctoral
ÁLVARO SOTO
HANS LÖBEL
CRISTIÁN TEJOS
RENÉ VIDAL
JUAN CARLOS NIEBLES
GUSTAVO LAGOS

Santiago de Chile, octubre, 2018

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
COLLEGE OF ENGINEERING

ACTIVITY RECOGNITION IN RGB-D VIDEOS USING HIERARCHICAL AND COMPOSITIONAL ENERGY-BASED MODELS

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the Degree of Doctor in Engineering Sciences by

IVÁN ALBERTO LILLO VALLÉS

ABSTRACT

Human activity recognition from videos is a very active research area in Computer Vision. Several approaches has been proposed, with different complexity according to the length of the videos, the capture modality used to acquire them, and the number of actions being performed by people in a single scene, among others. In this scenario, complex activity recognition has emerged as an active research topic, as people might perform several temporally and spatially concurrent actions in the same scene.

This thesis focuses on complex activity recognition using RGB-D cameras, which provide depth and appearance information in real time in indoor environments. The 3D pose estimation of body joints is included in the software provided in these devices, leveraging research based on 3D skeleton data.

Our focus is on recognizing complex activities composed of sequential or simultaneous atomic actions, which are also composed of low level body poses and motions of a single actor. This problem is tackled by introducing a hierarchical compositional model that operates at three levels of abstraction: activities (complex actions), atomic actions and poses.

Our contribution is a hierarchical compositional model that operates at three levels of abstraction. At the lower level, geometric and motion cues are used to automatically learn a dictionary of body poses to encode atomic action segments at a frame level. At the

intermediate level, compositions of learned body poses, split in body regions, are used to represent atomic human actions, with a single action per each body region, and where each region is composed of one or several sequential atomic actions. Finally, at the highest level, spatial and temporal compositions of atomic actions are assembled to represent complex human activities, using a single complex activity label for each video.

The model's parameters learning process is formulated as an energy minimization problem using a max-margin framework, where each body pose and atomic action is modeled by a linear classifier.

We first present a base hierarchical model which shows satisfactory performance when applied to a complex activity recognition benchmark (Composable Activities Dataset). Furthermore, a variety of additions to the base model are studied: (i) adding a sparsity term in atomic action classifiers, fostering the specialization of poses into a reduced set of actions; (ii) adding motion cues from RGB images, augmenting the geometric descriptor of the base model; (iii) a model trained with no spatial action supervision, automatically discovering active body parts from temporal action annotations only at training and testing time; (iv) a model incorporating flexible representations for *motion poselets* and *actionlets* that encodes the visual variability of body parts and atomic actions; and (v) a mechanism to discard idle or non-informative body regions which increases its robustness to common pose estimation errors like occlusions or poses that are not associated directly with the performed atomic actions.

Experimental results show the benefits of using a hierarchical model that exploits the sharing and composition of body poses into atomic actions, and atomic actions into activities. In particular, the resulting model is able to identify the temporal span of each atomic action as well as the body regions executing each action, having the appealing property to output mid-level semantic information in addition to high level activity classification.

The performance of the proposed method is evaluated using multiple action recognition benchmarks. The proposed model consistently outperforms baselines and state-of-the-art action recognition methods for complex action recognition, showing how hierarchical and compositional models are key to represent complex interactions using semantic low level building blocks, which are poses and atomic actions.

Keywords: Action recognition, action detection, hierarchical model, structural prediction.

Members of the Committee
ÁLVARO SOTO
HANS LÖBEL
CRISTIÁN TEJOS
RENÉ VIDAL
JUAN CARLOS NIEBLES
GUSTAVO LAGOS

Santiago de Chile, April, 2018

Chapter 1. INTRODUCTION

Human activity recognition is a key technology for the development of many important computer vision applications, such as surveillance, human-computer interaction, video annotation, video retrieval, among others. Consequently, it has received wide attention in the computer vision community (Aggarwal & Ryoo, 2011; Vishwakarma & Agrawal, 2013; Weinland, Ronfard, & Boyer, 2011) with a strong focus on recognition of atomic actions using short-length RGB videos. Several drawbacks emerge using only RGB videos, since the information acquired using these cameras as input suffer of problems like variable rotation, scaling, point-of-view, illumination or clutter, making the action inference difficult.

Recently, the emergence of capable and affordable RGB-D cameras has opened a new attractive scenario to recognize human activities, which emerge naturally when considering longer sequences of human actions and more natural scenarios. As a major advantage, RGB-D data facilitates the segmentation of the human body, as well as, the identification of relevant interest points, such as body joint positions (Shotton et al., 2011a), which are much more difficult to identify directly from RGB color images only.

As this area evolves, there has been an increasing interest to develop more flexible models that can extract useful knowledge from longer video sequences, featuring multiple concurrent or sequential actions, which are referred to as *complex actions* or *activities*, which are compositions of simpler atomic actions. These compositions can occur spatially and/or temporally, and they can involve interactions with the environment, other people, or specific objects. For instance, people can text while walking, or wave one hand while holding a phone to their ear with the other. It is noteworthy that different compositional arrangements of poses and atomic actions can yield different semantics at a higher level. Therefore, it is important for activity recognition systems to be aware of such compositional differences when discriminating activities at a higher level.

To facilitate tasks such as video tagging or retrieval, it is important to design models that can identify the spatial and temporal spans of each relevant action. As an example, Figure 1.1 illustrates a potential usage scenario, where an input video featuring a complex action is automatically annotated by identifying its underlying atomic actions and corresponding spatio-temporal spans.

Figure 1.1. Sample frames from a video sequence featuring a complex action. The video has a single complex action label (*Performing a reading session*) and several atomic action labels, where each sequence of atomic action labels are independent in each body region (arms, legs). Our method is able to identify the global complex action, as well as, the temporal and spatial span of meaningful atomic actions and local body part configurations.

A promising research direction for reasoning about complex human actions is to explicitly incorporate body pose representations. In effect, as noticed long ago, body poses are highly informative to discriminate among human actions (Johansson, 1973). Similarly, recent works have also demonstrated the relevance of explicitly incorporating body pose information in action recognition models (Jhuang, Gall, Zuffi, Schmid, & Black, 2013; C. Wang, Wang, & Yuille, 2013). While human body pose estimation from color images remains elusive, the emergence of accurate and cost-effective RGB-D cameras has enabled the development of robust techniques to identify body joint locations and to infer body poses (Shotton et al., 2011b).

In this thesis, we present an approach for human activity recognition that operates on body poses estimated from RGB-D videos to recognize human activities and to provide detailed information about complex human actions in RGB-D videos, with focus on activities that can be characterized by the body motions of a single actor. Specifically, given a video featuring a complex action, the model can identify the complex action occurring

in the video, in addition to the set of atomic actions that compose this complex action. Furthermore, for each atomic action, the model is also able to generate temporal annotations by estimating its starting and ending times (action temporal localization), and spatial annotations by inferring the body parts that are involved in the action execution.

A key aspect of the proposed method is a novel hierarchical and compositional model that operates at three semantic levels of abstraction. At the bottom level, our model learns a dictionary of representative body pose primitives, which encode the body joint geometry in a local spatio-temporal vicinity. At the mid-level, these body poses are combined to compose atomic actions, such as, walking or reading. Finally, at the top-level, atomic actions are combined to compose complex human activities, such as walking while talking on the phone and waving a hand. Additionally, our model also considers temporal relations among poses and atomic actions, which acts as inertial operators that biases the label inference of new videos to be consistent with the atomic actions and pose transitions seen during learning.

The use of intermediate abstraction levels that have a direct semantic interpretation, such as body poses or atomic actions, provides several advantages. At training time, it can take advantage of labeled data that can be available at intermediate abstraction levels. At test time, it enables the model to automatically identify useful information, such as the temporal span of atomic actions or the body regions invoved when executing the action. The inference of semantic information at the intermediate levels makes a notable difference with respect to blind compositional models based on deep learning techniques (Bengio, 2009). Additionally, the use of a compositional model can naturally handle scenarios of partial occlusions and pose estimation failures by inferring an appropriate spatial composition of visible and relevant body regions while dismissing the occluded or irrelevant ones.

Learning stage is formulated as an energy minimization problem, where structural hierarchical relations are modeled by sub-energy terms that constraint compositions among poses and actions, including the temporal relations of atomic actions and poses. Additionally, the energy function is complemented by regularization terms that foster the inference of a dictionary of body pose primitives that shares discriminative poses among action classes. This enables the model to use small dictionary sizes, to reduce over-fitting problems, and

to improve computational efficiency. As the poses are not known in advance, their labels are modeled as latent variables during training.

The development of the model is divided in three major sections. Firstly, a *core model* composed of a three-level hierarchy is presented (Lillo, Soto, & Niebles, 2014), highlighting the advantages of modeling complex actions using a hierarchical and compositional approach. In this case, a simple geometric vector computed from inferred body joints is used as feature, using fully-annotated complex actions and atomic actions as inputs for learning the model.

Secondly, an extension of the model is presented (Lillo, Niebles, & Soto, 2017), that seeks to improve the representation of poses and actions by incorporating three additions to the *core model*: (1) a more robust descriptor adding visual motion features extracted from RGB video frames, (2) a pose specialization method via sparse dictionary learning among the coefficients of the action classifiers, and (3) a *garbage collector* mechanism over the poses that are considered during learning. As the new descriptor incorporates geometry and motion, the pose dictionary primitives are called *motion poselets*. The learned *motion poselets* are shared between atomic actions, using sparsity in the components of the linear classifiers that represent the dictionary of atomic actions. This produces a highly specialized pose dictionary, with each entry only appearing in few atomic actions, helping in the semantic interpretation of poses. Furthermore, the extended model incorporates a *garbage collector* mechanism that identifies and discards idle or non-informative spatial areas of the input videos, providing an effective method to process long video sequences as it learns useful pose dictionaries. These extensions result in improved recognition performance and increased computational efficiency over the *core model*.

Lastly, a more general setup that builds upon the *core model* is presented (Lillo, Niebles, & Soto, 2016), looking for a model that improves generalization in terms of using only temporal video annotations and providing multi-modal representations of atomic actions. This model uses an initialization scheme for spatial atomic actions based on self-pace learning (Kumar, Packer, & Koller, 2010), providing an efficient and robust mechanism to infer, at test and training time, action labels for each detected *motion poselet*, as well as, their temporal and spatial span. Additionally, a multi-modal approach that trains a group of

actionlets for each atomic action is included. This provides a robust method to capture relevant intra-class variations in action execution. These improvements help the model to recognize several modalities of executing the same atomic action using different sets of poses. Furthermore, it is applicable to datasets which are not feasible to use with the previous approaches of (Lillo et al., 2014) and (Lillo et al., 2017).

This thesis is organized as follows. In Chapter 2, we review relevant works related to activity and action recognition in videos. In Chapter 3, we present the core hierarchical model for recognition of complex activities. In Chapter 4, we present an improvement over the *core model*, incorporating specialization of poses in atomic actions via a sparse regularizer, using a more robust motion descriptor in addition to the geometric descriptor and developing a *garbage collector* mechanism to detect non-informative poses. In Chapter 5, improvements in the generalization power over the *core model* are presented, allowing it to decrease the annotation level of the datasets for learning, and boosting the model to allow several modalities for the same atomic action. In Chapter 6, we present detailed experimentation results to the models presented. Finally, in Chapter 7 we conclude the thesis and guide sereval ways for future work.

Chapter 2. RELATED WORK

Visual recognition of human activities is a very active topic in the computer vision literature, as there are many potential applications that could benefit from reliable understanding of human behavior. Recent surveys show the breath of prior work (Aggarwal & Ryoo, 2011; Vishwakarma & Agrawal, 2013; Weinland et al., 2011), while also pointing at challenges and limitations of current methods. Here, we briefly survey some of the most relevant previous work that relates to the method proposed in this thesis.

The idea of using human body poses and configurations as an important cue for recognizing human actions has been explored recurrently, as poses provide strong cues on the actions being performed. Initially, most research focused on pose-based action recognition in color videos (Feng & Perona, 2002; Thurau & Hlavac, 2008), however, due to the development of pose estimation methods on depth images (Shotton et al., 2011b), there has been recent interest in pose-based action recognition from RGBD videos (Escorcia, Davila, Golparvar-Fard, & Niebles, 2012; J.-F. Hu, Zheng, Lai, & Zhang, 2015; Vemulapalli, Arrate, & Chellappa, 2014a). Some methods have tackled the problem of jointly recognizing actions and poses in videos (Nie, Xiong, & Zhu, 2015) and still images (Yao & Fei-Fei, 2010), with the hope to create positive feedback by solving both tasks simultaneously.

In terms of recognition of composable activities, a number of researchers have tackled this problem using composition of actions and low-level representations based on local interest points (Dollár, Rabaud, Cottrell, & Belongie, 2005; Laptev, 2005) by modeling their temporal arrangement. Some researchers have extended single image representations, such as correlatons (Savarese, Winn, & Criminisi, 2006) and spatial pyramids (Lazebnik, Schmid, & Ponce, 2006) to videos (Laptev, Marszalek, Schmid, & Rozenfeld, 2008), and have applied them to the problem of simple human action categorization. Others have proposed models for decomposing actions into short temporal motion segments (Gaidon, Harchaoui, & Schmid, 2013; Niebles, Chen, & Fei-Fei, 2010), but cannot capture spatial composition of actions. Recently, several graph-based models have been proposed to account for spatio-temporal composition of low-level features (Amer & Todorovic, 2012; Brendel & Todorovic, 2010, 2011).

Our work is also related to pose-based representation by first extracting information about the pose of the actor. There is a significant amount of pose-based action recognition methods in the literature. However, traditional methods have several limitations, for example silhouette based recognition methods assume a static camera (Bobick & Davis, 2001; Thurau & Hlavac, 2008). An alternative to avoid detailed and precise pose estimation is to use a coarser representation of human poses such as poseletes (Bourdev & Malik, 2009; Raptis & Sigal, 2013). Their representation relies on the construction of a large set of frequently occurring poses, which is used to represent the pose space in a quantized, compact and discriminative manner. Their approach has been applied to action recognition in still images (Maji, Bourdev, & Malik, 2011), as well as in videos (Tao & Vidal, 2015; L. Wang, Qiao, & Tang, 2014; Zanfir, Leordeanu, & Sminchisescu, 2013). However, even if accurate body pose estimation is available, these methods are tremendously affected by body occlusions and by unrelated limb postures and motions that are not involved in the action. In our case, we alleviate these problems using a hierarchical model that integrates the estimation of a dictionary of body poses with the inference of more elaborated levels of abstractions, such as atomic actions and composable activities. In particular, our current model includes the flexibility to filter-out body poses that are not involved in the current action.

Another line of work looks at annotating novel action videos by recognizing single actions (Ramanan & Forsyth, 2003), but ignoring the composition of those single actions into meaningful complex activities. As an example, in (Ikizler & Forsyth, 2008), the authors propose a model that composes actions using HMMs, but its application to activity classification is not discussed. In (Koppula, Gupta, & Saxena, 2013), a Markov Random Field is trained over small temporal segments. Their model includes the detection of objects and object affordance labels. In (Wei, Zheng, Zhao, & Zhu, 2013), wavelet features are computed over temporal segments in each body joint. Their resulting model infers the underlying temporal structure of sequences of actions.

Researchers have also explored the idea of fusing pose-based cues with other types of visual descriptors. For example, Cheron et al. (Chéron, Laptev, & Schmid, 2015) introduce P-CNN as a framework for incorporating pose-centered CNN features extracted from optical

flow and color. In the case of RGBD videos, researchers have proposed the fusion of depth and color features (J.-F. Hu et al., 2015; Kong & Fu, 2015). In general, the use of multiple types of features helps to disambiguate some of the most similar actions.

Also relevant to the proposed framework are hierarchical models for action recognition. In particular, the use of latent variables as an intermediary representation in the internal layers of the model can be a powerful tool to build discriminative models and meaningful representations (N. Hu, Englebienne, Lou, & Krose, 2014; Y. Wang & Mori, 2008). An alternative is to learn hierarchical models based on recurrent neural networks (Du, Wang, & Wang, 2015), but they tend to lack interpretability in their internal layers and require very large amounts of training data to achieve good generalization.

The goal of this research is to recognize composed activities from RGB-D videos. Recently, the interest in recognition of human actions and activities from RGB-D videos has increased rapidly (Aggarwal & Xia, 2014) mainly due to availability of capable and inexpensive new sensors. Some methods for low-level feature extraction have been proposed to leverage the 3D information available on RGB-D data (Oreifej & Liu, 2013; Wan, Ruan, Li, An, & Zhao, 2014; Luo, Wang, & Qi, 2014). Furthermore, the availability of fast algorithms for human pose estimation (Shotton et al., 2011a; Microsoft, 2012) from depth images helps to overcome the difficulty and high computational expense of human pose estimation from color images only. This has motivated a significant amount of methods that build representations on top of human poses (Sung, Ponce, Selman, & Saxena, 2012; Xia, Chen, & Aggarwal, 2012; Escorcia et al., 2012; Vemulapalli, Arrate, & Chellappa, 2014b). In addition to the use of body pose features, other researchers have also proposed fusing them with low-level features from color (Chaaraoui, Padilla-López, & Flórez-Revuelta, 2013; Shahroudy, Wang, & Ng, 2014; Zhu, Chen, & Guo, 2013) or depth (J. Wang, Liu, Wu, & Yuan, 2012). Unfortunately, these methods usually focus on categorizing simple and non-composed activities.

From a learning perspective, our work is related to methods for learning visual dictionaries from data. Early frameworks for dictionary learning focus on vector quantization, using k-means to cluster low-level keypoint descriptors (Csurka, Dance, Fan, Willamowski, & Bray, 2004). These approaches have spawned algorithmic variations that use alternative

quantization methods, discriminative dictionaries, or different pooling schemes (Jurie & Triggs, 2005; Lazebnik et al., 2006). Recently, sparse coding methods have also been used to obtain meaningful dictionaries that achieve low reconstruction error, high recognition rate, and attractive computational properties (Castrodad & Sapiro, 2012). Discriminatively trained sparse representations have also been proposed (Boureau, Bach, LeCun, & Ponce, 2010; Mairal, Bach, Ponce, Sapiro, & Zisserman, 2008), mostly building specific dictionaries for each target category. In contrast to our approach, these methods have mostly focused on non-hierarchical cases, where there is a weak connection between the construction of the mid-level dictionaries and the implementation of the top-level classifiers (J. Yang, Yu, Gong, & Huang, 2009).

Our model also builds on ideas related to learning classifiers using a discriminative framework and latent variables. In particular, (Felzenszwalb, Mcallester, & Ramanan, 2008) uses a latent SVM scheme to develop an object recognition approach based on mixtures of multiscale deformable part models. This model is later extended to the case of action recognition (Niebles et al., 2010). In contrast to our approach, the model in (Niebles et al., 2010) is limited to binary classification problems. Recently, (Y. Wang & Mori, 2011) proposes a hierarchical latent variable approach to action recognition that directly considers the multiclass classification case. Their layered model incorporates information about patches, hidden-parts, and action class, where the meaning of the hidden layers is not clear. In contrast, our hierarchical model integrates semantically meaningful information at all layers: poses, actions, and activities. Unlike (Y. Wang & Mori, 2011), our model can account for compositions of actions into activities and, as a byproduct, outputs per-body-region and per-frame action classification, so it has the appealing property that mid-level semantics are produced in addition to the final activity classification decision. (J. Wang et al., 2012) proposes a model for action recognition in static images but it is not clear if an extension to spatio-temporal compositions is possible. Similarly to our approach, (Y. Wang & Mori, 2011) and (J. Wang et al., 2012) also use a latent SVM machinery for model learning and inference, but details of the formulations and regularization schemes are distinct to our framework. However, note that we focus the novelty of this thesis in the formulation of hierarchical model for recognition of composable activities, not the actual learning/inference algorithms.

In terms of hierarchical compositional models, our work is related to recent recognition approaches based on deep learning (DL) (Bengio, 2009; Krizhevsky, Sutskever, & Hinton, 2012). Most previous work on DL has a focus on analysis of static images. However, their training process has some similarities with our approach, since both incorporate joint hierarchical estimation of connected layers, spatial pooling schemes, and intermediate representations based on linear filters. DL is usually applied over the raw image representation using several layers of generic structures. As a consequence, DL architectures have a large number of parameters and they are usually difficult to train. In contrast, we embed semantic knowledge to our model by explicitly exploiting compositional relations among poses, actions, and activities. This leads to simpler architectures and enables incorporating labeled data at intermediate layers. Furthermore, our max-margin approach is based on a Hinge loss, and not quadratic or logistic functions commonly used to train DL architectures leading to different optimization problems.

Our method tackles some limitations of previous work with a new framework that models spatio-temporal compositions of activities using a hierarchy of three semantic levels. The compositional properties of our model enable it to provide meaningful annotations and to handle occlusions naturally.

Chapter 3. CORE HIERARCHICAL MODEL FOR COMPLEX ACTION RECOGNITION

Complex activities performed by humans usually can be decomposed in several atomic actions that are executed simultaneously or sequentially in a video. As shown in Figure 3.1, a single atomic action like *walking* can be present in many different activities. Furthermore, different regions of the body can execute a different atomic action at the same time, for example if a person is *walking* and *drinking* at the same time. The *core model* for complex action recognition presented in this chapter is organized according to a 3-level hierarchy of semantic information at different levels of abstraction: local body poses, atomic actions and activities. Many atomic actions are allowed in a single region, and only a single activity is defined for the whole video. This scheme allows to provide semantic and interpretable predictions of the body poses, atomic actions and activity that are present at each video frame. The first Section describes the input data and video representation used in the framework. Then, we present the main details of the energy functions that build the core model. Lastly, the learning and inference schemes for the *core model* are detailed.

3.1. Video representation

Starting from RGB-D videos of human body actions, the first step is to extract the (x, y, z) 3D joint poses in each frame using a skeleton tracker software, converting each frame into a feature vector representing body poses using the methods in (Shotton et al., 2011a). Specifically, given a video D with T frames, we extract a feature vector $X_D = \{x_1, \ldots, x_T\}$, where x_t is a set of pose features computed from the 3D body configuration estimated at frame t. The pose features x_t used as inputs of the *core model* are inspired by (Chen et al., 2010), which include angles between limbs and angles between limbs and planes spanned by body parts. Using angles as features allow us to tackle with the variability of point of view and scale of the 3D joint poses, producing a robust descriptor

Previous work (Escorcia et al., 2012; Sung et al., 2012) computes a global pose descriptor for the entire body. Instead, in the presented model the body pose is divided into R fixed spatial regions, using independent pose feature vectors computed for each region. Figure 3.2 illustrates the case when $R = 4$ that is used across all our experiments. The geometric

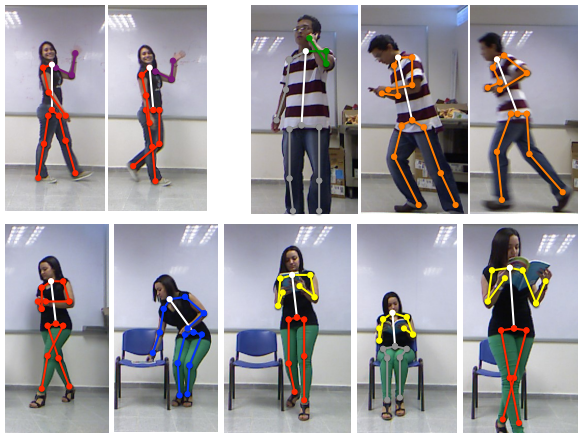

Figure 3.1. People perform complex activities that can be characterized as spatial and/or temporal compositions of simpler actions. *Top-right:* A person simultaneously *waves her hand* and *walks* by assigning subsets of body regions to different actions. *Top-right:* A person sequentially *talks on the phone* and, afterwards, *runs* away. *Bottom:* A person *walks* in a room, *picks a book up*, and *walks* while *reading a book*. We propose a novel formulation that uses RGB-D data to capture these spatio-temporal compositions of atomic actions in order to recognize human activities. In the *core model*, only sixteen joint poses are used as input for the activity recognition model.

descriptor is based on the spatial configuration of limbs and body joints. As illustrated in Figure 3.2, the pose features are computed starting with the positions of sixteen joints, which are grouped manually into $R = 4$ fixed regions. For each region, we compute relative angles among six selected segments, where each segment is a line that connects a pair of joints. We also compute relative angles between each segment and a plane spanned by a set of three joints in each region. This provides a total of 18 dimensions (angles) for the geometric descriptor associated to each body region, computed in each frame of the video. Table 3.1 shows the 6 segments and the plane that is defined for each body region.

3.2. Core Hierarchical Model

To recognize human activities and actions we propose a 3-level compositional hierarchical model. At the top level, our model assumes that each human activity, with a single activity label for the whole video, is composed by a temporal and spatial arrangement of

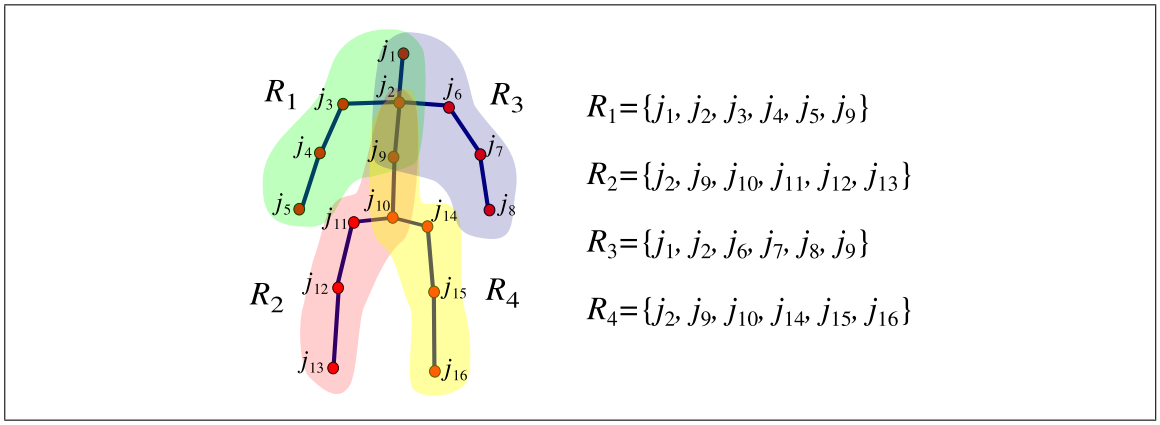

Figure 3.2. Skeleton representation used for splitting the human body into a set of four overlapping spatial regions.

Region	Segments	Plane defined by:
R_1, R_3	wrist → elbow elbow → shoulder shoulder → neck wrist → shoulder wrist → head neck → torso	shoulder elbow wrist
R_2, R_4	ankle → knee knee → hip hip → hip center ankle → hip ankle→ torso hip center → torso	hip knee ankle

Table 3.1. Segments and planes used by the geometric descriptor to define each body region.

atomic actions. At the intermediate level, our model assumes that each atomic action is composed by a temporal arrangement of body poses. Finally, at the bottom level of the hierarchy, our model assumes that each body pose is composed by a spatial arrangement of features derived from RGB-D data.

Given a video D, composed of T frames, where each frame t is described by a pose feature vector x_t, we define a video classification score, or energy function, for D as:

$$E(D) = E_{\text{activity}}(D) + E_{\text{actions}}(D) + E_{\text{poses}}(D) + E_{\text{action transition}}(D) + E_{\text{pose transition}}(D). \tag{3.1}$$

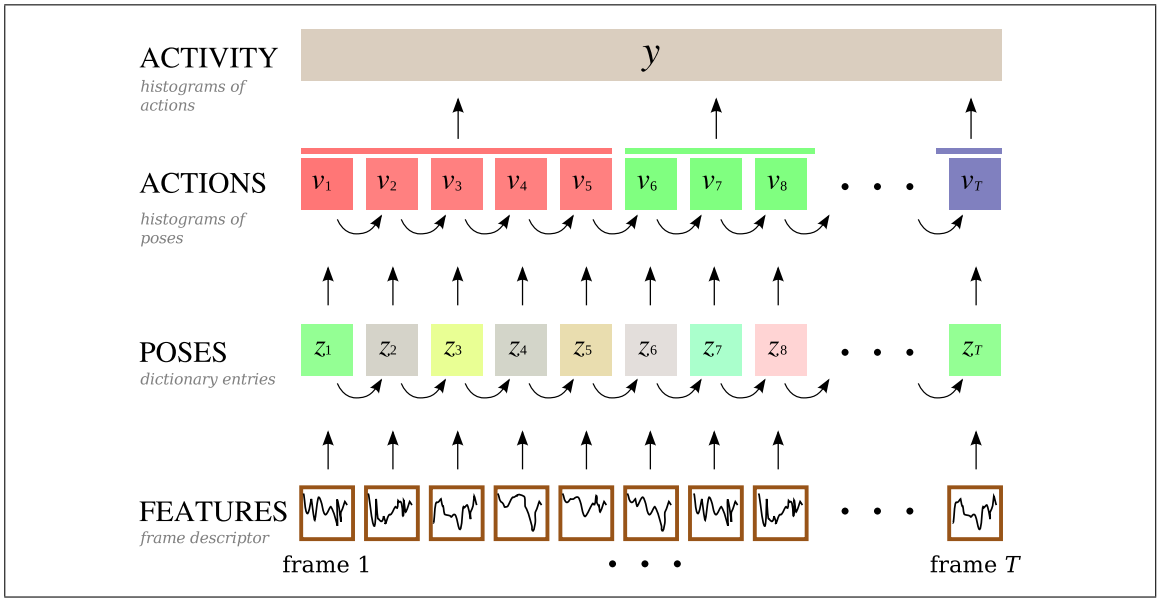

Figure 3.3. Pictorial representation of our discriminative hierarchical model for recognition of composable human activities, showing the case of $R = 1$ for simplicity. At the top level, activities are represented as compositions of atomic actions that are inferred at the intermediate level. These actions are in turn compositions of poses at the lower level, where the pose dictionary is learned from data. Our model also learn temporal transitions between consecutive poses and actions. Best viewed in color.

In Equation (3.1), energy $E(D)$ is expressed in terms of energy potentials associated to a single activity present in video D, as well as, its related sets of atomic actions and poses. We also consider two additional energy potentials that encode information related to temporal transitions between pairs of body poses ($E_{\text{pose transition}}$) and actions ($E_{\text{action transition}}$). Our goal is to find the spatial and temporal arrangement of body poses and atomic actions, as well as the underlying activity, that maximize $E(D)$.

In the following, we provide further details about our proposed energy function, and in the next section we present the corresponding learning and inference schemes. To simplify the notation, we first consider the case of representing the human body using only one spatial region ($R = 1$). Later, we extend the model to the case of $R > 1$ regions. In the equations, T_D represents the number of frames of video D.

$E_{\textbf{poses}}$**:** At the lowest level of the hierarchy, the main goal is to learn a dictionary of K body poses using pose feature vectors x_t, $t \in [1, \ldots, T_D]$. We encode each frame descriptor x_t, $t \in [1, \ldots, T_D]$, into one out of K body pose primitives. To achieve this we introduce a vector $Z = (z_1 \ldots z_t \ldots z_{T_D})$ of latent variables, where component z_t indicates the entry assigned to frame t from the dictionary of body poses, $z_i \in \{1, \ldots, K\}$. We obtain the dictionary of body poses by learning a set of K linear classifiers w_k that define the entries of the dictionary. In this way, the score of a candidate body pose k in frame t is given by the dot product between the pose frame descriptor x_t and the corresponding linear classifier $w_{z_t=k}$.

We define define the energy potential E_{pose} associated to pose assignment Z, as the sum of the pose entry scores for all its frames in Equation (5.5) as:

$$E_{\text{poses}}(D) = \sum_{t=1}^{T_D} w_{z_t}^{\top} x_t = \sum_{t=1}^{T_D} \sum_{k=1}^{K} w_k^{\top} x_t \delta(z_t = k) \tag{3.2}$$

where $\delta(\ell) = 1$ if ℓ is true and $\delta(\ell) = 0$ if ℓ is false. Note that every frame t is associated with a single dictionary entry given by the corresponding pose entry z_t. Intuitively, a high energy score $E_{\text{poses}}(D)$ indicates that pose descriptors $\{x_1, \ldots, x_{T_D}\}$ are highly consistent with the assignment Z to the dictionary of body poses.

$E_{\textbf{actions}}$**:** At the second level of the hierarchy, we measure the compatibility between the inferred pose assignments Z and a set of A mid-level atomic actions. To do this, we introduce an assignment vector $V = (v_1 \ldots v_{T_D})$, where component $v_t \in \{1, \ldots, A\}$ indicates the atomic action label assigned to frame t. To aggregate evidence for an atomic action a, we build a histogram $h^a(Z, V)$ calculated over the body pose assignments Z that are associated by V to action a, in a way similar to using a bag-of-words (BoW) representation (average pooling) of body poses for action a. Specifically, for an action a in an input video, the k-th entry of its histogram $h^a(Z, V)$ of associated poses is given by:

$$h_k^a(Z, V) = \sum_{t=1}^{T_D} \delta(z_t = k)\delta(v_t = a). \tag{3.3}$$

To quantify the compatibility between each action a and the observed evidence $h^a(Z,V)$, we train linear classifiers to identify each action. Specifically, let $\beta_a = (\beta_{a,1} \dots \beta_{a,K})$ be the coefficients of the resulting linear classifier to identify atomic action $a \in \{1,\dots,A\}$. We access a compatibility score between action a and its corresponding histogram $h^a(Z,V)$ by computing the dot product $\beta_a^\top h^a(Z,V)$. By aggregating this score over all candidate actions in a given video D, we obtain the energy potential E_{actions} associated to an action assignment V and body pose assignment Z:

$$\begin{aligned} E_{\text{actions}}(D) &= \sum_{a=1}^{A} \beta_a^\top h^a(Z,V) \\ &= \sum_{t=1}^{T_D} \sum_{a=1}^{A} \sum_{k=1}^{K} \beta_{a,k} \delta(z_t = k) \delta(v_t = a). \end{aligned} \tag{3.4}$$

Intuitively, a high energy score in Equation (3.4) indicates that for the input video there is a high degree of consistency between the selected poses Z and action labels V. For the *core model* presented in this chapter, we assume that during training we have available the set of atomic action labels for each training video. Nevertheless, similar to Equation (3.2), it is possible to introduce latent variables and extend the model to the case that a dictionary of atomic actions needs to be learned. It is important to note that by calculating the histogram of body poses only over the time intervals where each action is executed, the resulting model is agnostic about when and how many times an action occurs during a video. As we will explain later, the action transition energy potential will constraint the action labeling by searching solutions that present smooth temporal action transitions between consecutive frames.

E_{activity}: At the third level of the hierarchy, we use the action vocabulary labels accumulated over all T_D frames to build a BoW representation for the underlying activity. Specifically, let $h^y(D)$ be the histogram corresponding to activity y in video D. Each entry a in $h^y(D)$ is given by:

$$h_a^y(D) = \sum_{t=1}^{T_D} \delta(v_t = a) \tag{3.5}$$

Using an histogram of actions representation, we compute linear activity potentials, learning vector α_y that acts as coefficients of a one-versus-all activity classifier. Specifically, the

energy associated with activities is given by

$$E_{\text{activity}}(D) = \alpha_y^\top h^y(D) = \sum_{a=1}^{A} \sum_{t=1}^{T_D} \alpha_{y,a} \delta(v_t = a) \tag{3.6}$$

Energy associated to temporal transitions: For the energy terms associated to action and pose transitions in Equation (3.1), we depart from BoW representations and we introduce energy potentials that take into account temporal co-occurrences between poses and actions in consecutive frames. Specifically, let coefficients $\gamma_{a,a'} \in \mathbb{R}$ and $\eta_{k,k'} \in \mathbb{R}$ quantify co-ocurrence strength between neighboring pair of actions (a, a') or pair of poses (k, k'), respectively. Action and pose transitions energy potentials in Equation (3.1) are given by:

$$\begin{aligned} E_{\text{action transition}}(D) &= \gamma^\top s(V) \\ &= \sum_{a=1}^{A} \sum_{a'=1}^{A} \gamma_{a,a'} \sum_{t=1}^{T_D-1} \delta(v_t = a)\delta(v_{t+1} = a') \end{aligned} \tag{3.7}$$

$$\begin{aligned} E_{\text{pose transition}}(D) &= \eta^\top p(Z) \\ &= \sum_{k=1}^{K} \sum_{k'=1}^{K} \eta_{k,k'} \sum_{t=1}^{T_D-1} \delta(z_t = k)\delta(z_{t+1} = k') \end{aligned} \tag{3.8}$$

At the frame level, the previous model relies only on global image representations extracted at each frame. However, several works have shown the relevance of including local spatial information to boost recognition results (Lazebnik et al., 2006), as shown in Figure 3.2, where we account for local information by dividing each body pose at frame t into $R = 4$ spatial regions. Consequently, Equations (3.2), (3.4), (3.6), (3.7), and (3.8) become, respectively:

$$E_{\text{poses}}(D) = \sum_{r=1}^{R} \sum_{t=1}^{T_D} {w_{z_t}^r}^\top x_{t,r} \tag{3.9}$$

$$E_{\text{actions}}(D) = \sum_{r=1}^{R} \sum_{a=1}^{A} \beta_a^{r\top} h^{a,r}(Z, V) \tag{3.10}$$

$$E_{\text{activity}}(D) = \sum_{r=1}^{R} \sum_{a=1}^{A} \sum_{t=1}^{T_D} \alpha_{y,a}^r \delta(v_{t,r} = a) \tag{3.11}$$

$$E_{\text{action transition}}(D) = \sum_{a=1}^{A}\sum_{a'=1}^{A}\sum_{r=1}^{R}\gamma_{a,a'}^{r}\sum_{t=1}^{T_D-1}\delta(v_{t,r}=a)\delta(v_{t+1,r}=a') \quad (3.12)$$

$$E_{\text{pose transition}}(D) = \sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{r=1}^{R}\eta_{k,k'}^{r}\sum_{t=1}^{T_D-1}\delta(z_{t-1,r}=k)\delta(z_{t,r}=k') \quad (3.13)$$

3.2.1. Learning

The goal of learning is to obtain the optimal parameters for our energy function in Equation (3.1) using a video training corpus, so that it can be used to correctly classify new activity videos. In particular, given a set of training videos with corresponding atomic action and activity labels, we look for pose assignments for every video i, Z_i, and energy parameters $[\alpha, \beta, w, \gamma, \eta]$ that maximize the energy function corresponding to the true assignment of action and activity labels in each training video.

We cast our problem using a max-margin formulation to learn all relevant parameters. In particular, rather than first learning a dictionary of body poses and then learning classifiers for actions and activities, our goal is to learn parameters simultaneously using a multiclass max-margin approach. The input to our training algorithm is a set of M video sequences, where each video i contains annotations of the activity y_i, which is a single activity label for the complete video, and atomic actions V_i, each region having its own atomic action annotations. The set of T_i video frames is described by the set of pose feature vectors $X_i = (x_1, \ldots, x_{T_i})$. Note that labels V_i of atomic actions are region dependent; for instance, the right arm could be executing the action "drinking", the legs "walking", while the left arm is at resting position or "idle". This setup enables the use of spatial and temporal compositions of atomic actions. We aim to find optimal values for parameter sets α, β, w, γ , and η, as well as, slack variables ξ and latent variables Z, by solving the following regularized max-margin learning problem:

$$\min_{W,\xi} \quad \Omega(W) + \frac{C}{M}\sum_{i=1}^{M}\xi_i, \quad (3.14)$$

where

$$W^\top = [\alpha^\top, \beta^\top, w^\top, \gamma^\top, \eta^\top], \tag{3.15}$$

and

$$\xi_i = \max_{Z,V,y}\{E(X_i, Z, V, y) + \Delta((y_i, V_i), (y, V)) - \max_{Z_i} E(X_i, Z_i, V_i, y_i)\}, \quad i \in [1, ...M]. \tag{3.16}$$

In Equation (3.14), $\Omega(\cdot)$ is a regularizer that encourages well behaved linear classifiers; in particular, for the presented *core model* the chosen regularizer is the L_2 norm. In Equation (3.16), each slack variable ξ_i quantifies the error of the inferred labeling for the corresponding video D_i.

The previous optimization problem searches for parameters that minimize the sum of the errors in the training set, by encouraging the correct labeling (y_i, V_i) to have higher energy compared to *every* other labeling. The formulation also enforces a margin between such labeling by introducing a loss function Δ, which penalizes incorrect labeling at the activity and atomic action levels. In our implementation, we favor predicting the correct labels as follows:

$$\Delta((y_i, V_i), (y, V)) = \lambda_1 \delta(y \neq y_i) + \frac{\lambda_2}{RT_i} \sum_{r=1}^{R} \sum_{t=1}^{T_i} \delta(v_{t,r} \neq v_{(t,r)_i}) \tag{3.17}$$

By selecting a large value of λ_1, we give a large penalty when the activity is not predicted correctly. The second term in Equation 3.17 adds a penalty proportional to the number of regions and frames that are not labeled with the correct atomic action according to V_i, weighted by λ_2.

The energy maximization for the correct labeling depends on the body pose labels Z_i, which are not known in advance. Given the loss function in Equation (3.17), the constrained optimization problem in Equation (3.14) is similar to a latent structural SVM case (Yu & Joachims, 2009a), therefore it can be solved using a Concave-Convex Procedure (CCCP) (Yuille & Rangarajan, 2003) which guarantees convergence to a local minimum or saddle point. The CCCP algorithm alternates between maximizing Equation (3.14) with respect to the latent variables, and solving a structural SVM optimization problem (Tsochantaridis, Hofmann, Joachims, & Altun, 2004) that treats latent variables as completely observed.

In the base formulation of our model, we use a quadratic regularizer $\Omega(W) = \frac{1}{2}||W||_2^2$. This regularizer has an effect to reduce overfitting by constraining the solution to output classifiers with small values on their coefficients. Using this kind of regularizer, the problem in Equation (3.14) can be stated directly as a Latent Structural SVM (LSSVM) problem. The first step is to compute the best label assignment Z_i^* for every video i:

$$Z_i^* = \max_Z E(X_i, Z, V_i, y_i) \tag{3.18}$$

The second step treats Z_i as known, and solve a standard structural SVM model:

$$\min_{\alpha,\beta,w,\gamma,\eta,\xi} \Omega(\alpha, \beta, w, \gamma, \eta) + \frac{C}{M} \sum_{i=1}^{M} \xi_i$$

subject to:

$$E(X_i, Z_i, V_i, y_i) - E(X_i, Z, V, y) \geq \Delta((y_i, V_i), (y, V)) - \xi_i, \ \forall y \in \mathcal{Y}, Z \in \mathcal{Z}, V \in \mathcal{V} \tag{3.19}$$

It is worth to highlight that given Z_i, the optimization problem in Equation (3.14) generates linear classifiers that maximize the energy function corresponding to the known activity and atomic actions labels annotated in the training set. In a test video, we do not know in advance the pose labels Z nor the action labels V, so both must be inferred. In both cases, each pose label z_t depends on the pose dictionary entry w_{z_t}, the feature descriptor of the frame x_t, the action label v_t, and the pose and action labels z_{t-1} and v_{t-1} associated to the previous frame. It is also important to highlight that the labeling of poses is integrated with the activity and actions of the complete video; we can interpret this behavior as a contextual priming for the poses.

An initial set of body pose labels is computed using k-means over the pose features computed in every frame. As we treat the assignments Z as latent variables, the choice of initialization is crucial to orient the optimization.

We describe now the two optimization steps in detail. We omit upper limits for summations and condense several summations using comma-separated indices where convenient. Also, we use the notation δ_j^i to refer to the function $\delta(i = j)$.

Step 1: The first step is to infer latent variables $Z|\alpha, \beta, w, \gamma, \eta$:

$$Z_i^* = \max_Z E(X_i, Z, V_i, y_i) \tag{3.20}$$

$$= \max_Z \left(\sum_{r,a,t,k} \beta_{a,k}^r \delta_{z_{t,r}}^k \delta_{v_{t,r}}^a + \sum_{r,t,k} {w_k^r}^\top x_{t,r} \delta_{z_{t,r}}^k + \sum_{r,k,k'} \eta_{k',k}^r \sum_t \delta_{z_{t-1,r}}^{k'} \delta_{z_{t,r}}^k \right) \tag{3.21}$$

$$= \max_Z \left(\sum_{r,t,k} \left(\beta_{v_{t,r},k}^r + {w_k^r}^\top x_{t,r} + \sum_{k'} \eta_{k',k}^r \delta_{z_{t-1,r}}^{k'} \right) \delta_{z_{t,r}}^k \right) \tag{3.22}$$

Given V_i, we can solve Z_i^* separately for each region using dynamic programming. Omitting the video sequence index i for sake of notation simplicity and solving for the functions δ_j^i, for each region r we find Z_r^* solving:

$$Z_r^* = \underset{Z=(z_1,\ldots,z_T)}{\operatorname{argmax}} \sum_t \beta_{v_{t,r},z_t}^r + {w_{z_t}^r}^\top x_{t,r} + \eta_{z_{t-1},z_t}^r \tag{3.23}$$

We split the summation in Equation (3.23) in terms of pair-wise relations, and optimize it using a backward recursion:

$$f_t(z_t) = \max_{z_{t+1} \in \{1,\ldots,K\}} g(z_t, z_{t+1}) + f_{t+1}(z_{t+1}) \tag{3.24}$$

for all $z_t = 1, \ldots, K$, where $g(z_t, z_{t+1})$ defines the *benefit* of choosing the body pose label z_t given the pose label in the next frame z_{t+1}, and is given by

$$g(z_t, z_{t+1}) = \beta_{v_{t+1,r},z_{t+1}}^r + {w_{z_{t+1}}^r}^\top x_{t+1,r} + \eta_{z_t,z_{t+1}}^r. \tag{3.25}$$

Using dynamic programming, we find the optimal sequence Z_r^* for each region in $\mathcal{O}(TK^2)$ operations, whereas the exhaustive search approach is $\mathcal{O}(T^K)$. A greedy approach of selecting the optimal $z_{t,r}$ in every frame t using only the previous frame $t-1$ is $\mathcal{O}(TK)$, but it produces suboptimal assignments. It is noteworthy that the best body poses assignment do not depend on the activity label of the video given the atomic action labels.

Step 2: Get $\alpha, \beta, w, \gamma, \eta|Z$

In order to find the best set of parameters $W = (\alpha, \beta, w, \gamma, \eta)$, we solve the problem in Equation 3.19 using a Structrural SVM formulation, replacing the unknown body pose

asignments Z_i by the result of Step 1:

$$\min_{\alpha,\beta,w,\gamma,\eta,\xi} \Omega(\alpha,\beta,w,\gamma,\eta) + \frac{C}{M}\sum_{i=1}^{M}\xi_i$$

subject to:

$$E(X_i, Z_i^*, V_i, y_i) - E(X_i, Z, V, y) \geq \Delta((y_i, V_i), (y, V)) - \xi_i,\ \forall y \in \mathcal{Y}, Z \in \mathcal{Z}, V \in \mathcal{V}$$

$$\xi_i \geq 0 \tag{3.26}$$

We can solve the problem in Equation (3.26) by formulating the 1-*slack* version (Joachims, Finley, & Yu, 2009) of the problem:

$$\min_{\alpha,\beta,w,\gamma,\eta,\xi} \Omega(\alpha,\beta,w,\gamma,\eta) + C\xi$$

subject to:

$$\frac{1}{M}\sum_{i=1}^{M} E(X_i, Z_i^*, V_i, y_i) - E(X_i, Z, V, y) \geq \frac{1}{M}\sum_{i=1}^{M}\Delta((y_i, V_i), (y, V)) - \xi,\ \forall y \in \mathcal{Y}, Z \in \mathcal{Z}, V \in \mathcal{V}$$

$$\xi \geq 0 \tag{3.27}$$

which can be solved using the iterative cutting-plane algorithm (Joachims et al., 2009), finding in each iteration the most violated constraint given the solution for the parameters from the previous iteration. Using the property of *loss-augmented inference* of structural

SVM (Yu & Joachims, 2009b), finding the most violated constraint for a video i is given by

$$(\hat{y}, \hat{V}, \hat{Z}) = \underset{y,V,Z}{\text{argmax}} \quad \Delta((y_i, V_i), (y, V)) + E(X_i, Z, V, y) \tag{3.28}$$

$$\begin{aligned} = \underset{y,V,Z}{\text{argmax}} \quad & \lambda_1 \delta(y \neq y_i) + \frac{\lambda_2}{RT} \sum_{r,t} \delta(v_{t,r} \neq v_{(t,r)_i}) \\ & + \sum_{r,a,t} \alpha^r_{y,a} \delta^a_{v_{t,r}} + \sum_{r,a,t,k} \beta^r_{a,k} \delta^k_{z_{t,r}} \delta^a_{v_{t,r}} + \sum_{r,t,k} {w^r_k}^\top x_{t,r} \delta^k_{z_{t,r}} \\ & + \sum_{r,t,a,a'} \gamma^r_{a',a} \delta^{a'}_{v_{t-1,r}} \delta^a_{v_{t,r}} + \sum_{r,t,k,k'} \eta^r_{k',k} \delta^{k'}_{z_{t-1,r}} \delta^k_{z_{t,r}} \end{aligned} \tag{3.29}$$

$$\begin{aligned} = \underset{y,V,Z}{\text{argmax}} \quad & \lambda_1 \delta(y \neq y_i) + \sum_{r,t} \Bigg(\frac{\lambda_2}{RT} \delta(v_{t,r} \neq v_{(t,r)_i}) + \alpha^r_{y,v_{t,r}} + \beta^r_{v_{t,r},z_{t,r}} + {w^r_{z_{t,r}}}^\top x_{t,r} \\ & + \gamma^r_{v_{t-1,r},v_{t,r}} + \eta^r_{z_{t-1,r},z_{t,r}} \Bigg) \end{aligned} \tag{3.30}$$

It is important to note that the regularizer $\Omega(\cdot)$ does not play any role in solving Equation (3.28). We can solve (3.28) by exhaustively enumerating all values of activity y, and solving the following for every video i for a query class y:

$$\begin{aligned} \hat{V}_y, \hat{Z}_y = \underset{V,Z}{\text{argmax}} \quad \sum_{r,t} \Bigg(& \frac{\lambda_2}{RT} \delta(v_{t,r} \neq v_{(t,r)_i}) + \alpha^r_{y,v_{t,r}} + \beta^r_{v_{t,r},z_{t,r}} + {w^r_{z_{t,r}}}^\top x_{t,r} \\ & + \gamma^r_{v_{t-1,r},v_{t,r}} + \eta^r_{z_{t-1,r},z_{t,r}} \Bigg) . \end{aligned} \tag{3.31}$$

We can solve Equation (3.31) in every region using dynamic programming. As there is no terms related within regions, the labels of each region can be computed separately. We split the summation in Equation (3.31) in terms of pair-wise relations, and optimize it using a backward recursion:

$$F_t(v_t, z_t) = \max_{\substack{v_{t+1} \in \{1,\ldots,A\} \\ z_{t+1} \in \{1,\ldots,K\}}} G((v_t, z_t), (v_{t+1}, z_{t+1})) + F_{t+1}(v_{t+1}, z_{t+1}) \tag{3.32}$$

for all $v_t = 1, \ldots, A$ and $z_t = 1, \ldots, K$, where the function G defines the *benefit* of choosing the atomic action v_t and body pose label z_t given the atomic action label and body pose label in the next frame $t+1$. For each region r, G is given by

$$G^r((v_t, z_t), (v_{t+1}, z_{t+1})) = G^r_s((v_t, z_t), (v_{t+1}, z_{t+1})) + G^r_n(v_{t+1}, z_{t+1}) \tag{3.33}$$

$$G_s^r((v_t, z_t), (v_{t+1}, z_{t+1})) = \gamma^r_{v_{t,r}, v_{t+1,r}} + \eta^r_{z_{t,r}, z_{t+1,r}}$$
$$G_n^r(v_{t+1}, z_{t+1}) = \frac{\lambda_2}{RT}\delta(v_{t+1,r} \neq v_{(t+1,r)_i}) + \alpha^r_{y, v_{t+1,r}} + \beta^r_{v_{t+1,r}, z_{t+1,r}} + {w^r_{z_{t+1,r}}}^\top x_{t+1,r} \tag{3.34}$$

Using dynamic programming, we find the optimal sequence $\hat{V}_y$ and $\hat{Z}_y$ for each region in $\mathcal{O}(T(AK)^2)$ operations, whereas a suboptimal greedy approach is $\mathcal{O}(TAK)$. Despite using a greedy approach is tempting, the use of dynamic programming is mandatory for finding the most violated constraint and not a suboptimal approximation for every cutting plane step; otherwise, we have an early convergence to a suboptimal weight vector that can't reproduce the correct action labeling: while in learning we could get all constraints satisfied (because V_i is given), some sequences would have wrong assignments of activity and atomic actions at inference using the same training data.

In practice, we split G into a sequential term G_s and a non-sequential term G_n. Using dynamic programming "out of the box" for finding the best labeling for every video in every cutting-plane step is impractical when a high number of poses and actions is used, since we have AK (number of actions times number of poses) states spanning T frames, with AK in the order of thousands. To make the learning phase practical, for every frame we extract a subset of P pairs of states (v, z), selected according to the states that has higher G_n, the non-sequential terms for every frame. We add also the sequential terms that stays in the same state. Using a proper value of P in the order of hundreds, we are able to recover the optimal labeling for 99.98% of frames, reducing the order of the Dynamic Program to $\mathcal{O}(TP^2)$ for every video, which allows us to compute the most violated constraint in much lower time.

Algorithm 1 summarizes the iterative learning process of Step 2. We rewrite the energy terms in Equation (3.16) as a linear classifier over a structured vector constructed with the values of the input features and labels, i.e., $E(X, Z, V, y) = W^\top \psi(X, Z, V, y)$.

3.2.2. Inference

The input to the inference algorithm is a new video sequence with pose features X. The task is to infer the best activity label y^* and the best action labels V^*. Additionally,

Algorithm 1 Learning algorithm for W using one-slack formulation.

procedure LEARN W
  $Z \leftarrow Z_0$, $t \leftarrow 0$,
  repeat
    $Z^t \leftarrow Z$, $j \leftarrow 0$, $S = \varnothing$, $B = \varnothing$, $W \leftarrow W_0$
    repeat
      $W_j \leftarrow W$
      for $i = 1 : M$ **do**
        Find $\hat{y}_i, \hat{V}_i, \hat{Z}_i = \text{argmax}_{y,V,Z}\{{W_j}^\top \psi(X_i, y, V, Z) + \Delta(y_i, y, V_i, V)\}$
      end for
      Construct hyperplane $\bar{\Delta}\psi = \frac{1}{M}\sum_{i=1}^{M} \psi(X_i, y_i, V_i, Z_i^t) - \psi(X_i, \hat{y}_i, \hat{V}_i, \hat{Z}_i)$
      Construct margin $\bar{\delta} = \frac{1}{M}\sum_{i=1}^{M} \Delta(yi, \hat{y}_i, V_i, \hat{V}_i)$
      $S \leftarrow S \cup \bar{\Delta}\psi$
      $B \leftarrow B \cup \bar{\delta}$
      Find $W = \text{argmin}_w\{\Omega(w) + C \max_k\{B_k - w^\top S_k, 0\}\}$
      $j \leftarrow j + 1$
    until $[B_j - W^\top S_j] - [\max_k\{B_k - W^\top S_k\}] < \epsilon$
    for $i = 1 : M$ **do**
      Find $Z_i = \text{argmax}_z\{\psi(X_i, y_i, V_i, z)\}$
    end for
    $Z \leftarrow [\{Z_i\}]_{i=1}^{M}$
    $t \leftarrow t + 1$
  until assignments Z is (almost) the same as Z^t
end procedure

we also need to estimate latent variables Z^*.

$$y^*, V^*, Z^* = \underset{y,V,Z}{\text{argmax}}\, E(X, Z, V, y) \tag{3.35}$$

We can solve this by exhaustively enumerating all values of y, and solving the following at each step:

$$V_y^*, Z_y^* = \underset{V,Z}{\text{argmax}}\, E(X, Z, V, y) \tag{3.36}$$

Therefore, for each possible activity-class y, we must find V_y^* and Z_y^* frame-wise using:

$$\begin{aligned} v^*_{(t,r)_y}, z^*_{(t,r)_y} = \underset{v_t, z_t}{\text{argmax}}\, & \alpha_{y,r,v_t} + \beta_{v_t,r,z_t} + w^\top_{z_t,r} x_{t,r} \\ & + \gamma_{v_{t-1},v_t,r} + \eta_{z_{t-1},z_t,r} \end{aligned} \tag{3.37}$$

We can use the same Dynamic Program for Equation (3.31), setting $\lambda_1 = \lambda_2 = 0$. It is worth to note that the inference outputs for a new video are a single activity label and atomic action labels for every frame, obtaining semantic temporal and spatial annotations of atomic actions and poses via just maximizing the energy function, without any further reasoning or frame grouping.

Chapter 4. SPARSE FORMULATION OF CORE MODEL INCLUDING MOTION FEATURES AND FILTERING-OUT NON-INFORMATIVE POSES

The *core model* presented in previous Chapter assigns a body pose label to every region and frame in the video, while the body pose dictionary is shared by all the atomic action classifiers. However, not all poses should appear in every atomic action, and even some poses are not relevant at all since they do not carry enough discriminative information. Some examples are poses of the right arm when the left arm is executing an action (for example, is *waving hand*), or poses extracted from occluded regions. We improve the core model including the following main modifications: (i) introduce a model that accounts for pose specialization in every atomic action, (ii) include new visual features to describe poses and (iii) incorporate a mechanism to filter-out non-informative poses.

A summary of the main contributions added to the *core model* in this chapter is presented next.

i. Sparse regularizer fostering pose specialization in atomic actions: Our learning problem in Equation (3.14) incorporates a regularizer $\Omega(\cdot)$. In the *core model* formulation of Chapter 3, we use a quadratic regularizer $\Omega(W) = \frac{1}{2}||W||_2^2$. This regularizer has an effect to reduce overfitting by constraining the solution to output classifiers with small values on their coefficients. In this chapter we explore the use of a regularizer that fosters the generation of classifiers with a sparse set of values on their coefficients.

In particular, a useful sparsity constraint can be enforced on the coefficients of each atomic action classifiers β_a. Recall that atomic action classifiers β operates on histograms of pose assignments. Then, when β_a is sparse, the model encourages the corresponding atomic action classifier to be influenced by a small number of pose dictionary entries, and by propagation to the higher level, each activity is also influenced by a reduced number of poses. In this sense, we can see the sparsification of each β_a as a way to encourage the learning of poses that specialize to the identification of certain activities. As an example, if an action a is relevant to only one activity, a sparse β_a would make the poses used by action a to be highly discriminative to identify the corresponding action and activity. Otherwise,

if two or more activities share an atomic action, they will naturally share the set of most common poses used by these actions.

ii. Improved pose feature representation: we enhance our feature representation including RGB based motion cues. We refer to these new features as *motion features*. These are computed using a longer temporal span compared to geometric features which are calculated independently for every frame, as shown in Figure 4.1. As we augment the geometric descriptor with the motion feature, the same geometric pose configuration can be assigned to different dictionary entries according to the motion described in a vicinity of frames.

iii. Dealing with non-informative poses: it is common that during an action execution, several frames or frame regions are not directly related to the executed action, due to occluded parts or movements that do not participate directly in the discrimination of the actions. We propose a mechanism to deal with these non-informative elements via a *garbage collector* mechanism, grouping low-scored frames with respect to the pose dictionary. We learn a variable threshold θ^r per region that allows learning the pose dictionary with only *inlier* poses.

In this chapter, we describe the *sparse model* in detail. Part of the formulations presented in this chapter were presented in (Lillo et al., 2017).

4.1. Elastic Net Regularizer for sparse atomic action classifiers

Following the Elastic Net Regularizer (Zou & Hastie, 2005), we enforce sparsity on each atomic action classifier β_a by minimizing the L_1 and L_2 norms over its coefficients in the regularizer Ω. Additionally, we also introduce a positive constraint on each coefficient $\beta_{a,i} \geq 0$ to restrict the composition of atomic actions to poses that are "present" in the video sequence. This emulates a generative approach during the learning process. As a result, we use the following regularization constraint:

$$\begin{aligned} \Omega_s(W) &= \frac{1}{2}||W_{-\beta}||_2^2 + \frac{\mu}{2}||\beta||_2^2 + \rho||\beta||_1, \\ \beta_{a,k} &\geq 0, \;\; a \in \{1,\ldots,A\}, \; k \in \{1,\ldots,K\} \end{aligned} \tag{4.1}$$

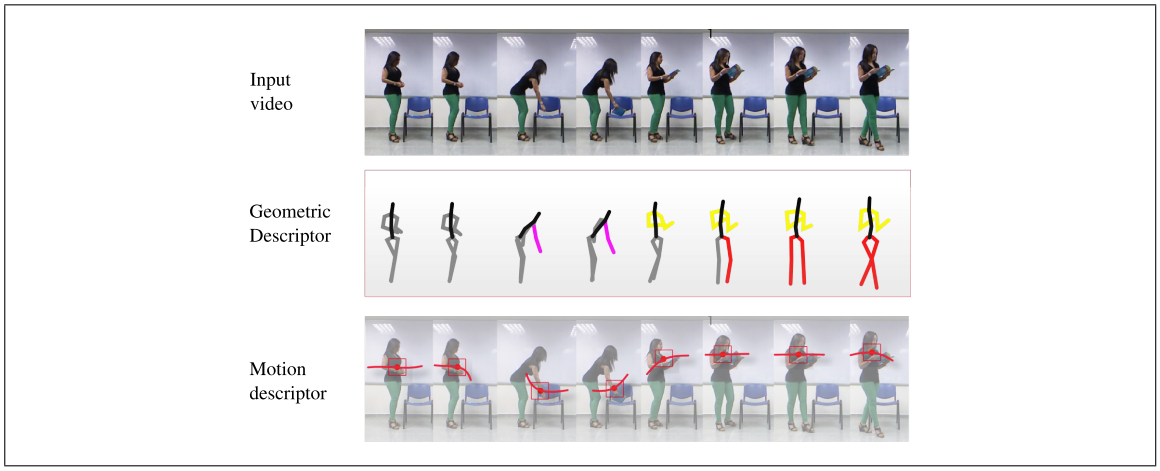

Figure 4.1. While geometric features are important for invariant video description, RGB videos also add relevant information for discriminating among activities and atomic actions. In a formulation using motion cues, every frame of the input video is described by a pose descriptor, built with a geometric descriptor, coding local geometry of the body from skeleton joints, and a motion descriptor, coding local motion during a short time interval, computed over estimated joints.

where $W_{-\beta}$ corresponds to the parameter vector W without the dimensions associated to the atomic action classifiers. Constants μ and ρ weight the influence of the L_1 and L_2 regularizers, respectively. The use of a L_1 and L_2 norms regularization makes the objective function of the dual problem differentiable, as shown in Section 4.5.2.

As we only are changing the regularizer Ω and not the energy function, the model is the same as in Equation (3.14), adding the constraint $\beta_{a,i} \geq 0$.

4.2. Enhanced pose descriptor using motion from RGB video

While the geometric descriptor captures body pose configurations, it misses to encode information about the dynamic of each body pose. We argue that motion cues are relevant to disambiguate poses that are similar in configuration but move differently. Our motion descriptor is based on the trajectory descriptor presented in (H. Wang, Klaser, Schmid, & Liu, 2011). While this descriptor also encodes appearance information, our experiments indicate that only the part encoding motion through a histogram of optical flow (HOF) helps to increase the recognition accuracy of our model. Furthermore, instead of calculating a dense descriptor as in (H. Wang et al., 2011), we only encode motion information around

each detected body joint location. Specifically, we compute at each joint location a HOF using RGB patches centered at the joint location for a temporal window of 15 frames. At each joint location, this produces a 108-dimensional descriptor, that we concatenate across all joints to obtain our motion descriptor. Finally, to reduce dimensionality to get a motion representation similar in number of dimensions with the geometric descriptor, we apply PCA to transform the concatenated descriptor into a 20-dimensional vector, keeping the dimensionality of our final descriptor relatively low. The final descriptor is the concatenation of the geometric and motion descriptors, $x_{t,r} = [x^g_{t,r}; x^m_{t,r}]$. We call *motion poselets* to the pose dictionary entries learned using this pose descriptor, since encodes pose and motion in a single descriptor.

4.3. A garbage collector for motion poselets

While poses are highly informative for action recognition, an input video might contain irrelevant or idle zones, where the underlying poses are noisy or non-discriminative to identify the actions being performed in the video. As a result, low-scoring motion poselets could degrade the pose classifiers during training, decreasing their performance. To deal with this problem, we include in our model a *garbage collector* mechanism for motion poselets. This mechanism operates by assigning all low-scoring motion poselets to a new $(K+1)$-th pose dictionary entry. These collected poses are associated with a learned score lower than θ^r, as in Equation (4.2). Our experiments show that this mechanism leads to learning more discriminative motion poselet classifiers, in a process that resembles RANSAC, selecting only *inlier* poses for training the motion poselet linear classifiers.

$$E_{\text{mot. poselet GC}}(Z, X) = \sum_{r,t} \left[\sum_k {w^r_k}^\top x_{t,r} \delta^k_{z_{(t,r)}} + \theta^r \delta^{K+1}_{z_{(t,r)}} \right] \tag{4.2}$$

The garbage collector mechanism integrates seamlessly with the code model, since few changes are needed. The coefficients θ^r are integrated in the vector of parameters W as $W^\top = [\alpha^\top, \beta^\top, w^\top, \gamma^\top, \eta^\top, \theta^\top]$. In terms of inferring latent variables or finding the most violated constraint, it is sufficient to compute the scores $w^\top_{k,r} x_{t,r}$ for $k = \{1, \ldots, K\}$, while θ^r will act as the score for the label $K+1$, and the dynamic program of Equation (3.31)

is solved for T frames, A atomic actions and $K+1$ poses. For initialization of the pose labels Z, we assign the $K+1$ label to the 20% of the farther frames with respect to the pose k-means cluster centers. In the experiments, we will show that the percentage of non-informative poses keeps relatively high when the model is trained, allowing the model to learn better body poses increasing the interpretability of the learned poses.

4.4. Measuring the influence of poses in activity classifiers

Using a sparse regularizer for action classifiers, only a few elements of the pose dictionary influences every action, and consequently few poses influence also activity classifiers. We measure the influence or importance of each pose entry in the dictionary for the recognition of each activity using the following metric. For every region r, we define the influence of pose k on activity y as

$$I^r(y,k) = \sum_{a=1}^{A} \alpha^r_{y,a}\beta^r_{a,k}\delta(\alpha^r_{y,a} > 0)\delta(\beta^r_{a,k} > 0) \tag{4.3}$$

4.5. Learning

The learning method is the same as the *core model* presented in Chapter 3, using a Latent Structured SVM framework, solving iteratively the best labeling of Z and optimal parameters for W. We assume that activity and atomic action annotations are given for training videos. In the following equations the energy terms in Equation (3.16) is written as a linear classifier over a structured vector constructed with the values of the input features and labels, i.e., $E(X,Z,V,y) = W^\top \psi(X,Z,V,y)$. The solution to this LSSVM problem implies the sequential iteration of two main steps. The first step consists of inferring for each input video the corresponding latent variables Z. To achieve this we solve:

$$Z_i^*|W = \max_Z \left\{ W^\top \psi(X_i, Z, V_i, y_i) \right\} \tag{4.4}$$

and then in a second step we use the estimated $Z_i,\ i \in \{1,\ldots,M\}$ to find the optimal values for W:

$$\min_{W,\xi} \quad \Omega_s(W) + \frac{C}{M}\sum_{i=1}^{M}\xi_i$$

subject to:

$$W^\top(\psi_i(X_i,Z_i,V_i,y_i) - \psi_i(X_i,Z,V,y)) \geq \Delta_i(y,V) - \xi_i \quad \forall y \in \mathcal{Y}, Z \in \mathcal{Z}, V \in \mathcal{V}, \xi_i \geq 0$$

$$\beta_{a,k} \geq 0, \quad a \in \{1,\ldots,A\},\ k \in \{1,\ldots,K\} \tag{4.5}$$

where $\Delta_i(y,v) = \Delta((y_i,V_i),(y,V))$.

We use cutting-plane algorithm (Joachims et al., 2009) to deal with the huge number of linear constraints. To do this, we reformulate the problem in Equation (4.5) as a 1-*slack* case (Joachims et al., 2009):

$$\min_{W_{-\beta},\beta,\xi} \frac{1}{2}||W_{-\beta}||_2^2 + \frac{\mu}{2}||\beta||_2^2 + \rho||\beta||_1 + C\xi$$

subject to:

$$\frac{1}{M}\sum_{i=1}^{M} W^\top(\psi_i(X_i,Z_i,V_i,y_i) - \psi_i(X_i,Z,V,y)) \geq \frac{1}{M}\sum_{i=1}^{M}\Delta_i(y,V) - \xi \tag{4.6}$$

$$\forall y \in \mathcal{Y}, Z \in \mathcal{Z}, V \in \mathcal{V}, \xi \geq 0$$

$$\beta_{a,k} \geq 0, \quad a \in \{1,\ldots,A\},\ k \in \{1,\ldots,K\}$$

Finding the best labeling (V,Z) of a video sequence given the parameters W is the same as in the *core model* of Chapter 3, since the labels do not depend on the regularizer. Using the Elastic Net regularizer and non-negativity constraints in some elements of W implies that the problem of Equation (4.6) can not be solved directly with a standard SVM solver, although a convenient dual problem can be stated. We now show the primal and dual formulations to solve Equation (4.6).

4.5.1. Primal formulation for Elastic Net regularizer

Recalling the 1-*slack* formulation of Equation (4.6), a single constrain is formed by finding the most violated constraint of each video i and then compute the average vector

of ψ and the average loss over all videos:

$$\bar{\psi} = \frac{1}{M}\sum_{i=1}^{M} \psi_i(X_i, Z_i, V_i, y_i) - \psi_i(X_i, \hat{Z}_i, \hat{V}_i, \hat{y}_i) \tag{4.7}$$

$$\bar{\Delta} = \frac{1}{M}\sum_{i=1}^{M} \Delta((y_i, V_i), (\hat{y}_i, \hat{V}_i)) \tag{4.8}$$

In each cutting-plane iteration, a new constraint is introduced into the working set of constraints, totalizing J constraints. The model using the working set of J constraints, and using the regularizer (4.1), can be stated as:

$$\begin{aligned} P1)\ \min_{W,\xi}\ & \textit{obj-primal}(W, \xi, \mu, \rho) = \frac{1}{2}||W_{-\beta}||_2^2 + \frac{\mu}{2}||\beta||_2^2 + \rho||\beta||_1 + C\xi \\ & \text{subject to:} \\ & \begin{bmatrix} -W^\top & -\xi \end{bmatrix} \begin{bmatrix} \bar{\psi}_j \\ 1 \end{bmatrix} + \bar{\Delta}_j \leq 0, j \in \{1, \ldots, J\} \\ & -\xi \leq 0 \end{aligned} \tag{4.9}$$

If we check the gradient of the objective function of $P1$, we have

$$\frac{\partial\ \textit{obj-primal}(W, \xi, \mu, \rho)}{\partial W_i} = \begin{cases} W_i & \text{if } W_i \in \{\alpha, w, \gamma, \eta, \theta\} \\ \mu W_i + \rho\frac{W_i}{|W_i|} & \text{if } W_i \in \beta \end{cases} \tag{4.10}$$

Note that the gradient is not differentiable for $\beta_i = 0$. If we add non-negativity constraints for β, the problem $P1$ becomes differentiable since the gradient is no longer discontinuous. The primal formulation could be solved using a standard non-linear optimization package that supports linear and bound constraints. Nevertheless, the number of dimensions of W is relatively high, and we add a new constraint in each cutting-plane iteration, so the cost of solving $P1$ in its primal form is high, specially when the number of constraints exceeds a few hundreds. It is important to note that each constraint added to the working set depends on W, which changes in every iteration. New values for W produces new *best* assignments. Using an alternative optimization method like augmented lagrangian or ADMM could seem a good choice at first; however, the structured vector ψ_i is produced by labels generated using the values of W, and not only by pairs of input-output values. In practice, an augmented lagrangian formulation for $P1$ produces stalling of W in a few set

of values. Due to these limitations, we decide in this thesis to use the dual formulation, as described next.

4.5.2. Dual formulation for Elastic Net regularizer

In Support Vector Machines, an interesting property of the dual formulation is that the number of dimensions of the dual problem is equal to the number of constraints, and involves solving a quadratic problem with a single linear constraint and bound constraints when 1-*slack* formulation is used. The same principle could be applied for our Elastic Net regularizer. We will see that including a mix of L_2 and L_1 norms, compared to use only L_1 as in a full sparse regularizer, carries at least two benefits: first, the objective function of the dual is differentiable, and also the regularization term appears in the objective function, and not as a set of non-linear constraints as usual when only L_1 norm is used. This nice properties allows us to use a general purpose optimization library in solving the dual problem.

We now derive the dual formulation of $P1$. Recalling from convex optimization theory, the dual formulation for the problem

$$\min_{x \in \mathrm{dom} f} f(x) \quad \text{subject to} \quad Qx - b \leq 0 \tag{4.11}$$

is given by

$$\max_{\phi \in \mathrm{dom} f^*} -f^*(-Q^\top \phi) - b^\top \phi \quad \text{subject to} \quad \phi \geq 0 \tag{4.12}$$

where f^* is the *convex conjugate* function of f, given by $f^*(y) = \sup_{x \in \mathrm{dom} f}\{y^\top x - f(x)\}$. The convex conjugate function is well defined only when it is bounded, and one of its nice properties is that the convex conjugate function of independent variables is the sum of the individual convex conjugate functions. With this in mind, we can formulate the dual of problem $P1$ using Equation (4.12) applied independently for $W_{-\beta}$, β and ξ. First, we reformulate the constraints of $P1$ in matrix form:

$$-\mathbf{Q} \begin{bmatrix} W \\ \xi \end{bmatrix} - \mathbf{b} \leq 0 \tag{4.13}$$

where

$$\mathbf{Q} = \begin{bmatrix} \bar{\Psi}_\alpha^\top & \bar{\Psi}_\beta^\top & \bar{\Psi}_w^\top & \bar{\Psi}_\gamma^\top & \bar{\Psi}_\eta^\top & \bar{\Psi}_\theta^\top & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 1 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} -\bar{\Delta}_1 \\ \vdots \\ -\bar{\Delta}_J \\ 0 \end{bmatrix}. \tag{4.14}$$

$\bar{\Psi}$ is the matrix of the stacked $\bar{\psi}_j,\ j = \{1, \ldots, J\}$, one column per constraint, and the subscript refers to the portion of indexes of the corresponding linear classifier.

Now, we can find the convex conjugate functions for $W_{-\beta}$, β and ξ:

Convex conjugate function for $f(x) = \frac{1}{2}||x||_2^2$.

It is easy to show that if $f(x) = \frac{1}{2}||x||_2^2$, then $f^*(y) = \frac{1}{2}||y||_2^2$.

Convex conjugate function for $f(x) = \frac{\mu}{2}||x||_2^2 + \rho||x||_1$.

From the definition of the complex conjugate function, we have

$$f^*(y) = \sup_{x \in \text{dom} f} \{y^\top x - \frac{\mu}{2}||x||_2^2 - \rho||x||_1\}. \tag{4.15}$$

As we are adding independent variables, we can formulate the convex conjugate function as

$$\begin{aligned} f^*(y) &= \sum_{i=1}^{M} f_i^*(y_i) \\ &= \sum_{i=1}^{M} \sup_{x_i \in \text{dom} f_i} \{y_i x_i - \mu x_i^2 - \rho|x_i|\}. \end{aligned} \tag{4.16}$$

With some manipulations, and assuming $\mu > 0$ and $\rho \geq 0$, we have

$$f_i^*(y_i) = \begin{cases} 0 & \text{for } |y_i| < \rho \\ \frac{1}{2\mu}(|y_i| - \rho)^2 & \text{for } |y_i| \geq \rho \end{cases}. \tag{4.17}$$

Then, adding the independent variables x_i, we have

$$f^*(y) = \frac{1}{2\mu} \sum_{i=1}^{M} \left[\max(0,\ |y_i| - \rho)\right]^2 . \tag{4.18}$$

Convex conjugate function for $f(x) = Cx$.

$$f(x) = Cx \Rightarrow f^*(y) = 0 \text{ only for } y = C \tag{4.19}$$

Using the matrix form of the constraints as in Equation (4.13), and considering the dual variables ϕ, the dual formulation can now be stated as:

$$\begin{aligned} &\min_{\phi} f^*(\mathbf{Q}^\top \phi) + \tilde{\mathbf{b}}^\top \phi \\ &\text{subject to} \quad \phi_i \geq 0. \end{aligned} \tag{4.20}$$

Finally, the dual formulation becomes

$$\begin{aligned} D1) \quad &\min_{\phi} \ \textit{obj-dual}(\bar{\Psi}, \phi) = \frac{1}{2}||\bar{\Psi}_{W_{-\beta}}\phi||_2^2 + \frac{1}{2\mu} \sum_{r,a,k} [\max(0, |\bar{\Psi}_{\beta_{r,a,k}}\phi| - \rho)]^2 + \mathbf{b}^\top \phi \\ &\text{subject to} \\ &\sum_{j=1}^{J} \phi_j \leq C, \quad \phi_j \geq 0 \ \ \forall\ j \in \{1, \ldots, J\}. \end{aligned} \tag{4.21}$$

The linear constraint $\sum_{i=0}^{M} \phi_{c_i} \leq C$ is produced by the linear term in the primal formulation, $C\xi$. Recall that $f^*(y) = 0$ only for $y = C$; in this case, $y = \sum_{i=1}^{M} \phi_{c_i} + \phi_\xi = C$. As ϕ_ξ must be always greater or equal to zero, it is enough to consider the constraint $\sum_{i=0}^{M} \phi_{c_i} \leq C$ and discard ϕ_ξ of the problem.

At this point there are some facts that is worth to highlight. First, for $\rho = 0$ and $\mu = 1$, we recover the original quadratic formulation (as a standard SSVM) in the dual problem. Second, for $\mu = 0$ and $\rho > 0$, the formulation is still useful, since the terms associated with $\bar{\Psi}_\beta$ disappear from the objective function, but it adds $R \times A \times K$ non-linear constraints of the form $|\bar{\Psi}_{\beta^r_{a,k}}\phi| \leq \rho$, so using only L_1 (a fully sparse regularizer) therefore is not the best idea since it force the dual model to use thousands of non-linear constraints, requiring an alternative solution method. And finally, the gradient of the objective function is continuous, as opposed to the gradient of the objective function of the primal problem,

and is given by

$$\frac{\partial\ obj\text{-}dual(\bar{\Psi},\phi)}{\partial\phi} = \bar{\Psi}_{W_{-\beta}}^{\top}\bar{\Psi}_{W_{-\beta}}\phi + \frac{1}{\mu}\sum_{r,a,k}\max(0,|\bar{\Psi}_{\beta_{r,a,k}}\phi| - \rho)\bar{\Psi}_{\beta_{r,a,k}}^{\top}\bar{\Psi}_{\beta_{r,a,k}}\phi + \mathbf{b}. \quad (4.22)$$

By using a combination of L_2 and L_1 norms in the dual formulation, the objective function is differentiable, as opposed to the primal formulation, so we can use any gradient-based solver for $D1$. Also, there is no extra constraints compared to using only L_1 norm, and the primal variables (i.e. linear classifiers) are recovered from the dual variables in a closed form.

To recover the primal variables from dual variables, we use the optimal x^* for $f^*(y)$. For quadratic terms (L_2 norm), we have $x^* = y = -Q^{\top}\phi$, so we have $\boldsymbol{\alpha} = \bar{\Psi}_{\alpha}\phi$, $\boldsymbol{w} = \bar{\Psi}_{w}\phi$, $\boldsymbol{\gamma} = \bar{\Psi}_{\gamma}\phi$, $\boldsymbol{\eta} = \bar{\Psi}_{\eta}\phi$ and $\boldsymbol{\theta} = \bar{\Psi}_{\theta}\phi$. For $\boldsymbol{\beta}$, we have the equation $y - \mu x^* - \rho\frac{x^*}{|x^*|} = 0$ for $|y| \geq \rho$, and $x^* = 0$ otherwise. Noting that x^* and y must have the same sign, we solve $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}_i = \begin{cases} 0 & \text{if } |\bar{\Psi}_{\beta_i}\phi| < \rho \\ \frac{\bar{\Psi}_{\beta_i}\phi}{\mu[1+\frac{\rho}{|\bar{\Psi}_{\beta_i}\phi|-\rho}]} & \text{if } |\bar{\Psi}_{\beta_i}\phi| \geq \rho \end{cases}. \quad (4.23)$$

From Equation (4.23), we can clearly see the effect of ρ: the bigger its value, the more coefficients of β will be zero.

As we will see in the experiments section, using a small value for μ and a relatively high value of ρ produce the coefficients of β to be sparse, making an efficient use of poses compared to using only a quadratic regularizer.

With some minor changes, we can include non-negativity constraints for the primal variables in the dual formulation. Specifically, we use non-negativity constraints in the classifiers β, with the objective that the sparse coefficients generated resemble a generative approach. Each non-negative constraint adds a new dual variable, so we will have new set of dual variables ϕ_{β}. In practice, only the dual variables associated to linear constraints are computed. ϕ_{β} can be computed in closed form, as they try to make the associated original weight to zero, so $\phi_{\beta} = \max(0, -\bar{\Psi}_{\beta}\phi_c)$, with ϕ_c the dual variables associated to the working set of constraints.

Using the sparse regularizer produces sparse coefficients for atomic action classifiers (β). However, as we are jointly learning the classifiers of poses (w), the sparseness of the pose assignments inside the actions is not completely reflected during training in the imputed latent pose assignments Z_i, since during training of W we assume that Z_i are true assignments so the learning process will try to overestimate the pose scores to mimic the imputed poses during inference. To compensate this effect, we enforce a grouping effect, seeking that few poses have influence in each activity, with the goal of improved pose specialization. In Equation (4.3), the marginalization over the set of positive atomic actions makes explicit the fact that pose entry k influences activity y only through its compositional role to generate an action a. After a few iterations of our algorithm, we enforce the assignments of Z_i to be one of the influential poses ($I^r(y,k) > 0$) using the known activity label y_i for every video, changing Equation (4.4) to

$$Z_i^{r*}|W = \max_{Z^r=\{z_j^r\}_{j=1}^T | I^r(y_i,z_j^r)>0} \left\{ W^\top \psi(X_i^r, Z^r, V_i^r, y_i) \right\} \tag{4.24}$$

where we explicitly divide Equation (4.4) into R subproblems, one for each body region. We emphasize that the forced grouping effect is applied only after few iterations are computed, since we need to learn good pose classifiers before some dictionary entries are turned off for each activity. It is important to note that this operation is only applied during training.

4.6. Inference

The inference of labels when a new video arrives is the same as the *core model* of Chapter 3, since it does not depend on the election of regularizer. Then, the inference is solved by exhaustively enumerating all values of y, and solving the following at each step:

$$\begin{aligned} v^*_{(t,r)_y}, z^*_{(t,r)_y} = \operatorname*{argmax}_{v_t,z_t} & \alpha_{y,r,v_t} + \beta_{v_t,r,z_t} + w^\top_{z_t,r} x_{t,r} \\ & + \gamma_{v_{t-1},v_t,r} + \eta_{z_{t-1},z_t,r} \end{aligned} \tag{4.25}$$

Chapter 5. MULTIMODAL REPRESENTATION OF ACTIONS USING ONLY TEMPORAL ACTION ANNOTATIONS

The *core model* has the appealing property of delivering semantically interpretable outputs for the activity in each input video, in addition to spatio-temporal annotations of atomic actions. This is achieved by using labeled data at two abstractions levels, requiring during training the activity label performed during the whole video, and atomic actions labels for each frame and region of the body. In this chapter, we explore two improvements over the *core model*: i) reducing the intermediate levels required at training time, and ii) increasing the intermediate representation of actions to model multimodal distributions.

i. Only temporal annotations for atomic actions: usually, activity datasets do not have the required level of spatial annotations as needed by the *core model* presented in Chapter 3, as it requires to access spatio-temporal annotations of atomic actions during training. Moreover, the granularity of spatial annotations must be linked to the election of regions on the human body. Furthermore, this level of annotations is costly and time consuming, and more important, it does not allow the model to be used in more general action situations. We propose to treat the spatial arrangement of atomic actions as latent variables, keeping the temporal information of actions as known during training. As with every latent variable model, a careful initialization of the spatial arrangement of atomic actions is necessary. We develop an initialization scheme of atomic actions based on self-pace learning (Kumar et al., 2010), requiring only temporal annotations of actions and inferring the best spatial arrangement of actions during training and testing. These setup makes the atomic action annotation independent of the predefined human body regions.

ii. Multimodality of atomic actions: we explore a change in the formulation that deals with multi-modal representations of atomic actions, that we call *actionlets*. Usually, humans execute a similar action using different body configurations; for instance, the action *waving hand* can be executed moving just the hand or moving a complete arm. To deal with these differences in action execution, several atomic action classifiers are grouped under the same *semantic action* label. Each of these classifiers encode what we refer as an *actionlet*. In this way, we express the intermediate representation of our model in terms of *motion poselets*,

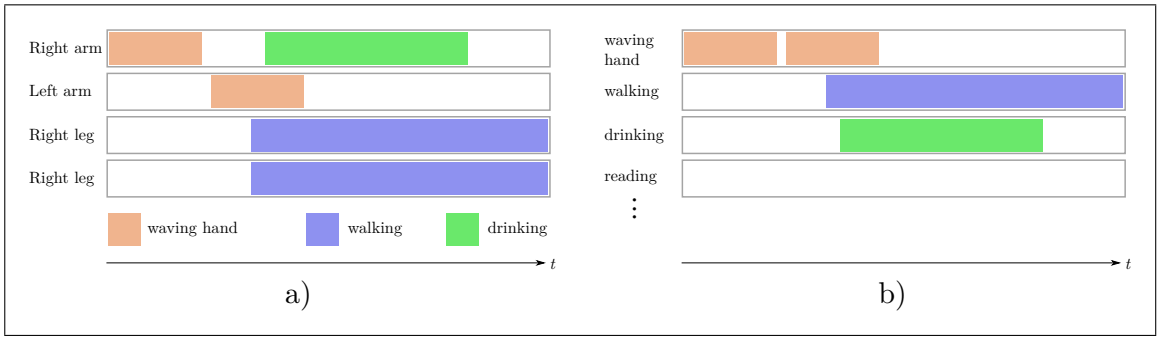

Figure 5.1. We discard spatial information of atomic action labels and keep only temporal annotations. **a)** In the *core model*, every region (arms and legs) have an independent set of temporal atomic action annotations. **b)** The spatial information in a) is discarded, keeping only the temporal information per atomic action in every video. The goal of the improved model is to infer the spatial arrangement of the actions, as well as, their temporal span.

which are poses that encode geometry and motion, and *actionlets*, which are temporal arrangements of *motion poselets* encoded by linear classifiers.

In the following, we detail the formulation for the improvements over the *core model*, which we call *multimodal model*. In Chapter 6 we provide empirical evidence indicating that the integration of the previous contributions in a single hierarchical model, generates a highly informative and accurate solution that outperforms state-of-the-art approaches.

5.1. Latent spatial assignment of atomic actions

The *core model* needs two kind of annotations for every video in the training stage: a global label indicating the activity or complex action executed in the whole video, and a set of atomic action labels that encode one or several time intervals where the action is executed, independent for every region r. This kind of atomic action labels are costly to obtain and do not scale well if a different splitting of the human body is used, for example, if using $R = 5$ considering the torso as a region by itself. With this in mind, we develop a model that treats spatial assignment of atomic actions to regions as latent variables. We create a mechanism to initialize the atomic action labels in every region using temporal annotations of initial and ending frames for the atomic actions present in the videos. This setup allows us to use any kind of spatial arrangement as only temporal information is provided. Figure 5.1 shows how the spatial information is discarded.

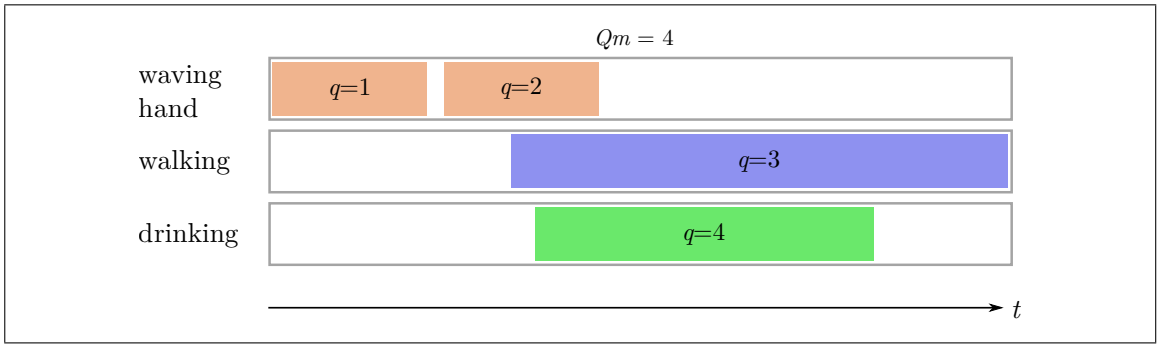

Figure 5.2. Construction of action intervals for a sample video. Every time an atomic action starts or ends a new boundary is defined. Action intervals q are defined as non-overlapping time segments between the defined boundaries where at least one atomic action is active, and each action interval corresponds to a single atomic action. In our formulation, every region can be assigned all the action intervals as long they do not overlap in time.

The *core model* needs the spatio-temporal annotations in order to learn parameters that allows us to infer the region of the body that executes the actions at test time. For this reason, we propose an extended model that treats the spatial arrangement of atomic actions as latent variables.

An important step in the training process is the initialization of latent variables. This is a challenging task due to the lack of spatial supervision. At each time instance, the available atomic actions can be associated with any of the R body regions. We adopt the machinery of self-paced learning (Kumar et al., 2010) to provide a suitable solution and formulate the initial association between atomic actions and body regions as an optimization problem. First, we split the video m into Q_m action intervals, where each action interval boundary is defined when an atomic action starts or ends, as Figure 5.2 shows. The action intervals are only defined when at least one atomic action is active in that time interval, and each action interval corresponds to a single atomic action. The goal is to assign a single action for every action interval to all defined regions, inferring the most likely atomic action annotation given geometric and motion features inside the action interval. We also want to assign no action labels to time intervals in regions that do not have enough match with any atomic action.

We define $b^m_{r,q} = 1$ when action interval $q \in \{1, \ldots, Q_m\}$ is active in region r of video m; and $b^m_{r,q} = 0$ otherwise. Each action interval q is associated with a single known atomic

action a_q. We assume that we have initial motion poselet labels $z_{t,r}$ in each frame and region, given by a initial k-means partitioning as in the *core model*. We describe the action interval q and region r using the histogram $h^m_{r,q}$ of initial motion poselet labels. We can find the correspondence between action intervals and regions using a formulation that resembles the operation of k-means, with the following additional structural constraints:

i Atomic actions intervals must not overlap in the same body region.

ii A labeled atomic action must be present at least in one region. This constraint makes the temporal labeling consistent.

We formulate the labeling process as a binary Integer Linear Programming (ILP) problem using M videos, R body regions and Q_m action intervals as:

$$
\begin{aligned}
P\text{q}) \quad & \min_{b,\mu} \sum_{m=1}^{M} \sum_{r=1}^{R} \sum_{q=1}^{Q_m} b^m_{r,q} d(h^m_{r,q} - \mu^r_{a_q}) - \frac{1}{\lambda_3} b^m_{r,q} \\
& \text{s.t.} \quad \sum_{r=1}^{R} b^m_{r,q} \geq 1, \, \forall q, \, \forall m \\
& \qquad b^m_{r,q_1} + b^m_{r,q_2} \leq 1 \text{ if } q_1 \cap q_2 \neq \emptyset, \, \forall r, \, \forall m \\
& \qquad b^m_{r,q} \in \{0,1\}, \, \forall q, \, \forall r, \, \forall m
\end{aligned}
\tag{5.1}
$$

with

$$
d(h^m_{r,q} - \mu^r_{a_q}) = \sum_{k=1}^{K} (h^m_{r,q}[k] - \mu^r_{a_q}[k])^2 / (h^m_{r,q}[k] + \mu^r_{a_q}[k]). \tag{5.2}
$$

In problem Pq, $\mu^r_{a_q}$ are the means of the descriptors with action label a_q within region r. The condition $q_1 \cap q_2 \neq \emptyset$ refers to temporal intersection of two action intervals. The distance metric between the action interval descriptors $h^m_{r,q}$ and the mean of action interval descriptors assigned to an action a and region r, $\mu^r_{a_q}$, is selected to be the Chi-Squared distance, as shown in Equation (5.2). We solve Pq iteratively using a block coordinate descending scheme, alternating between solving the assignments $b^m_{r,q}$ with $\mu^r_{a_q}$ fixed, and then fixing $\mu^r_{a_q}$ to solve $b^m_{r,q}$, relaxing Pq to solve a linear program. The relaxation involves constraining $b^w_{r,q}$ to lie in the real number interval $[0, 1]$. In practice, almost all components of $b^w_{r,q}$ take the values 0 or 1. Note that the second term of the objective function in Pq resembles the objective function of *self-paced* learning (Kumar et al., 2010), managing the balance between assigning a single region to every action interval, or assigning all possible regions to the respective action interval, always satisfying the structural constraints. This

is important since the objective function of Pq will try to assign the least number of regions subject to the constraints. Then, actions like *walking* would always be assigned to a single leg. Using a proper value for λ_3, the single region solution for actions like *walking* should be suboptimal with respect to selecting two regions (both legs).

In terms of changes in the formulation with respect to the *core model*, there are two main modifications. First, as the spatial information of the atomic actions is not known in advance, we can not use this information directly in the loss function of the *core model* shown in Equation (3.17). A loss function that is suitable to the new nature of the atomic actions discards the spatial ordering of actionlets which is unknown (hence the latent actionlet formulation). The temporal composition is known, so we can compute a list A_t of possible actionlets for frame t, and include that information on the loss function as

$$\Delta((y_i, \vec{v}_i), (y, \vec{v})) = \lambda_1(y_i \neq y) + \lambda_2 \frac{1}{RT_i} \sum_{r=1}^{R} \sum_{t=1}^{T_i} \delta(v_t \notin A_t) \tag{5.3}$$

Also, the step of inferring latent labels changes. In the *core model*, only labels Z must be inferred. Using V as latent in the spatial arrangement, we still use the temporal information encoded in the list A_t to restrict the possible atomic actions that can be selected in every frame of the video, but we must let the model to choose the best spatial ordering of the inferred atomic actions during training. The new inference of latent variables is stated as

$$\begin{aligned}
(V^*, Z^*) &= \underset{\{v_{t,r} | v_{t,r} \in A_t\}, Z}{\operatorname{argmax}} \quad E(X_i, Z, V, y_i) \\
&= \underset{\{v_{t,r} | v_{t,r} \in A_t\}, Z}{\operatorname{argmax}} \quad \sum_{r,a,t} \alpha^r_{y_i,a} \delta^a_{v_{t,r}} + \sum_{r,a,t,k} \beta^r_{a,k} \delta^k_{z_{t,r}} \delta^a_{v_{t,r}} + \sum_{r,t,k} {w^r_k}^\top x_{t,r} \delta^k_{z_{t,r}} \\
&\quad + \sum_{r,t,a,a'} \gamma^r_{a',a} \delta^{a'}_{v_{t-1,r}} \delta^a_{v_{t,r}} + \sum_{r,t,k,k'} \eta^r_{k',k} \delta^{k'}_{z_{t-1,r}} \delta^k_{z_{t,r}} \\
&= \underset{\{v_{t,r} | v_{t,r} \in A_t\}, Z}{\operatorname{argmax}} \quad \sum_{r,t} \Big(\alpha^r_{y,v_{t,r}} + \beta^r_{v_{t,r},z_{t,r}} + {w^r_{z_{t,r}}}^\top x_{t,r} \\
&\quad + \gamma^r_{v_{t-1,r},v_{t,r}} + \eta^r_{z_{t-1,r},z_{t,r}} \Big) .
\end{aligned} \tag{5.4}$$

To solve Equation (5.4) we use dynamic programming, in the same way as in the core model, with the difference that the possible states (Z, V) are limited by the temporal action annotations.

The second step, related to finding the optimal parameters, is the same as the core model, replacing the loss for the new loss function stated in Equation (5.3).

5.2. Multi-modal representation of atomic actions

A single linear classifier does not offer enough flexibility to identify atomic actions that exhibit high visual variability. As an example, the atomic action "open" can be associated with "opening a can" or "opening a book", displaying high variability in action execution. Consequently, we augment our hierarchical model including multiple classifiers to identify different modes of atomic action execution.

Inspired by (Raptis, Kokkinos, & Soatto, 2012), we use the *Cattell's Scree test* to find a suitable number of actionlets to model each atomic action. Specifically, using the atomic action labels, we compute a descriptor for every video interval using normalized histograms of initial pose labels Z obtained with k-means. Then, for a particular atomic action s, we compute the eigenvalues $e(s)$ of the affinity matrix of the atomic action descriptors, which is built using χ^2 distance. For each atomic action $s \in \{1, \ldots, S\}$, we find the number of actionlets $H_s = \text{argmin}_i \, e(s)^2_{i+1} / (\sum_{j=1}^{i} e(s)_j) + c \cdot i$, with $c = 2 \cdot 10^{-3}$. Finally, we cluster the descriptors from each atomic action s running k-means with $k = H_s$. This scheme generates a set of non-overlapping actionlets to model each single atomic action. In our experiments, we notice that the number of actionlets used to model each atomic action varies typically from 1 to 8.

To transfer the new labels to the model, we define $u(v)$ as a function that maps from actionlet label v to the corresponding atomic action label u. A dictionary of actionlets provides a richer representation for actions, where several actionlets will map to a single atomic action. This behavior resembles a max-pooling operation, where at inference time we will choose the set of actionlets that best describes the performed actions in the video, keeping the semantics of the original atomic action labels. The mapping from actionlets to atomic actions, $u(v)$, is known a priori. We now describe in detail the model using the proposed multimodal actionlets and motion poselets.

Figure 5.3 shows a schematic of our model. At the top level, our model assumes that each input video has a single complex action label y. Each complex action is composed

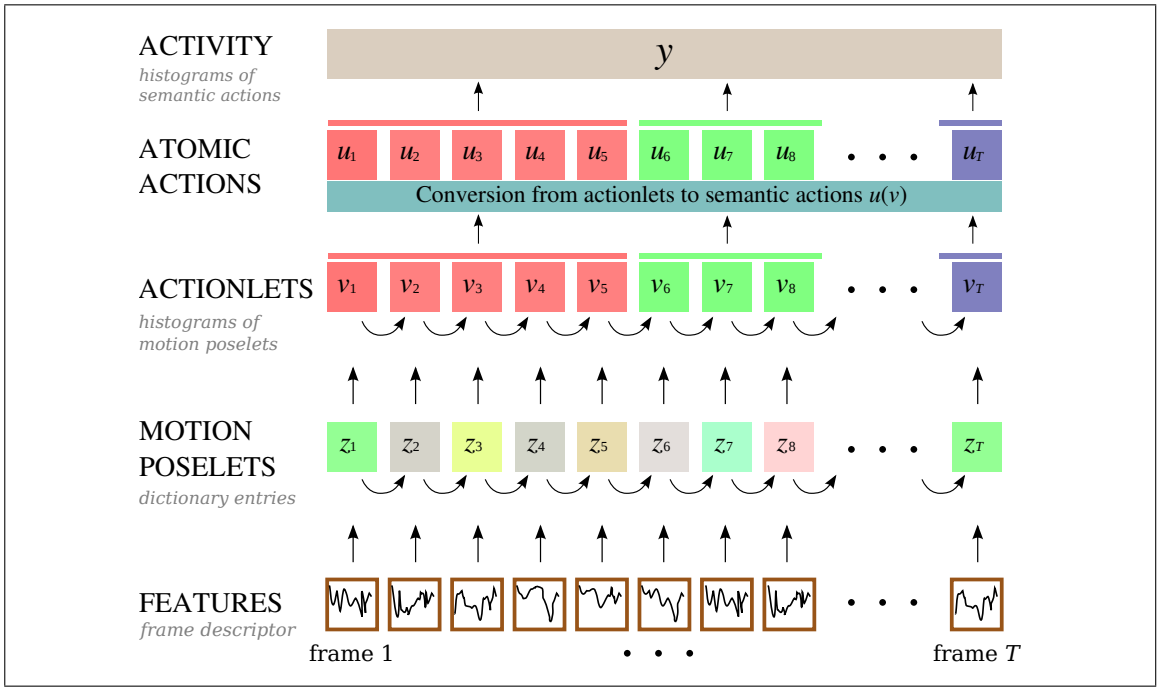

Figure 5.3. Graphical representation of the discriminative hierarchical model for recognition of complex human actions including multi-modal atomic actions (actionlets) and motion poselets, which we call *multimodal model*. At the top level, activities are represented as compositions of atomic actions that are inferred at the intermediate level. These actions are, in turn, compositions of poses at the lower level, where pose dictionaries are learned from data. Our model also learns temporal transitions between consecutive poses and actions.

of a temporal and spatial arrangement of semantic actions (former *atomic actions* in the *core model*) with labels $\vec{u} = [u_1, \ldots, u_T]$, $u_i \in \{1, \ldots, S\}$. In turn, each semantic action consists of several non-shared *actionlets*, which correspond to representative sets of pose configurations for action identification, modeling the multimodality of each atomic action. We capture actionlet assignments in $\vec{v} = [v_1, \ldots, v_T]$, $v_i \in \{1, \ldots, A\}$. Each actionlet index v_i corresponds to a unique and known actomic action label u_i, so they are related by a mapping $\vec{u} = \vec{u}(\vec{v})$. At the intermediate level, our model assumes that each actionlet is composed of a temporal arrangement of a subset from K body poses, encoded in $\vec{z} = [z_1, \ldots, z_T]$, $z_i \in \{1, \ldots, K\}$, where K is a hyperparameter of the model. These subsets capture pose geometry and local motion, so we call them *motion poselets*. Finally, at the bottom level, our model identifies motion poselets using a bank of linear classifiers that are applied to the incoming frame descriptors.

As in the *core model*, we build each layer of our hierarchical model on top of BoW representation of labels. To this end, at the bottom level of our hierarchy, and for each body region, we learn a dictionary of motion poselets. Similarly, at the mid-level of our hierarchy, we learn a dictionary of actionlets, using the BoW representation of motion poselets as inputs. At each of these levels, spatio-temporal activations of the respective dictionary words are used to obtain the corresponding histogram encoding the BoW representation.

We formulate our hierarchical model using an energy function similar to the *core model*, but using an additional mapping to translate actionlets into semantic actions. Given a video of T frames corresponding to complex action y encoded by descriptors $\vec{x}$, with the label vectors $\vec{z}$ for motion poselets, $\vec{v}$ for actionlets and $\vec{u}$ for semantic actions, we define an energy function for a video as:

$$\begin{aligned} E(X,V,Z,y) &= E_{\text{motion poselets}}(Z,X) \\ &+ E_{\text{motion poselets BoW}}(V,Z) + E_{\text{atomic actions BoW}}(u(V),y) \\ &+ E_{\text{motion poselets transition}}(Z) + E_{\text{actionlets transition}}(V). \end{aligned} \tag{5.5}$$

Besides the BoW representations and motion poselet classifiers described above, Equation (5.5) includes two energy potentials that encode information related to temporal transitions between pairs of motion poselets ($E_{\text{motion poselets transition}}$) and actionlets ($E_{\text{actionlets transition}}$). The energy potentials, not including the non-informative pose handling for sake of simplicity, are given by:

$$E_{\text{mot. poselet}}(Z,X) = \sum_{r,t}\sum_{k} {w_k^r}^\top x_{t,r}\delta^k_{z_{(t,r)}} \tag{5.6}$$

$$E_{\text{mot. poselet BoW}}(V,Z) = \sum_{r,a,k} \beta^r_{a,k}\delta^a_{v_{(t,r)}}\delta^k_{z_{(t,r)}} \tag{5.7}$$

$$E_{\text{atomic act. BoW}}(u(V),y) = \sum_{r,s} \alpha^r_{y,s}\delta^s_{u(v_{(t,r)})} \tag{5.8}$$

$$E_{\text{mot. pos. trans.}}(Z) = \sum_{r,k,k'} \eta^r_{k,k'} \sum_t \delta^k_{z_{(t-1,r)}}\delta^{k'}_{z_{(t,r)}} \tag{5.9}$$

$$E_{\text{acttionlet trans.}}(V) = \sum_{r,a,a'} \gamma^r_{a,a'} \sum_t \delta^a_{v_{(t-1,r)}}\delta^{a'}_{v_{(t,r)}} \tag{5.10}$$

Our goal is to maximize $E(X,V,Z,y)$, and obtain the spatial and temporal arrangement of motion poselets Z and actionlets V, as well as, the underlying complex action y.

In the previous equations, we use δ_a^b to indicate the Kronecker delta function $\delta(a = b)$, and use indexes $k \in \{1, \ldots, K\}$ for motion poselets, $a \in \{1, \ldots, A\}$ for actionlets, and $s \in \{1, \ldots, S\}$ for atomic actions. In the energy term for motion poselets, w_k^r are a set of K linear pose classifiers applied to frame descriptors $x_{t,r}$, according to the label of the latent variable $z_{t,r}$. In the energy potential associated to the BoW representation for motion poselets, $\vec{\beta}^r$ denotes a set of A mid-level classifiers, whose inputs are histograms of motion poselet labels at those frame annotated as actionlet a. At the highest level, α_y^r is a linear classifier associated with complex action y, whose input is the histogram of semantic action labels, which are related to actionlet assignments by the mapping function $\vec{u}(\vec{v})$. Note that all classifiers and labels here correspond to a single region r. We add the contributions of all regions to compute the global energy of the video. The transition terms act as linear classifiers η^r and γ^r over histograms of temporal transitions of motion poselets and temporal transitions of actionlets, respectively.

Chapter 6. EXPERIMENTS

In this chapter we present experimental results using action recognition benchmarks to illustrate the benefits of our approach in complex activity recognition, which also outputs spatio-temporal annotations as a side product. We split the evaluation of the proposed models into three sections: first, the *core model* presented in Chapter 3 is evaluated, showing the benefits of a hierarchical and compositional model applied to complex activities datasets; then, we test the *sparse model* of Chapter 4, including the addition of a *motion descriptor* and pose *garbage collector*, which shows advantages in terms of recognition accuracy and quality of the poses that the model found during training; and finally, starting from the *core model*, we show how the improvements presented in the *multimodal model* of Chapter 5 leverages the generalization to action recognition benchmarks which are not suitable to use with the *core model*, facilitating a more general usage and expanding the possibilities for the model. We start describing the implementation details that are shared between the models and explain the benchmark datasets used for evaluation. Later, each model is evaluated on the same datasets to provide a complete overview of the strengths of each proposed variation of the core model.

6.1. Common implementation details

Our models need to assign a single atomic action $a_r \in \{1, \ldots, A\}$ to every single region $r \in \{1, \ldots, R\}$ of the human body. For this reason, we split the human body into a fixed set of $R = 4$ overlapping regions in all experiments: right arm, right leg, left arm, and left leg, as shown in Figure 3.2.

At the lowest level of the hierarchical model, each frame t and body region r is represented by two feature vectors, encoding geometry and local motion. Geometry is estimated using 3D joint poses, and local motion is computed using RGB data and extracting features from trajectories in the video. For the *core model*, only geometry is used, whereas for the *sparse model* and *multimodal model* we use geometry and motion. The geometric feature vector for each frame and region is explained in Section 3.1, while the motion feature vector is described in Section 4.2. When no RGB information is present in the dataset, we use velocity features from the raw joint poses of wrists and ankles.

To initialize latent variables Z which correspond to latent poses, in the *core model* and in the *sparse model* we obtain an initial dictionary of body poses by clustering the low level descriptors $x_{t,r}$ using the standard k-means algorithm. Using this initial dictionary, the value of each latent variable is obtained by associating the corresponding descriptor to its closest centroid. For the *multimodal model*, we use a different mechanism to assign the pose latent assignments Z, as explained in Section 5.2.

It is important to note that during training, our algorithm takes a single global annotation for each video at the activity level, while the per-frame annotations for actions associated to each body region r depend on the type of model. Full spatio-temporal annotations are needed for the *core model* and *sparse model*, but only temporal annotations for the *multimodal model*, which can infer spatial annotations in training and testing stages. At test time, for all the models these labels are not available and it is the task of our algorithm to infer them using the learned model. Furthermore, for training and evaluation purposes, we augment the annotations in each dataset with an additional *idle* or *background* action. We add this annotation at each frame where the subject is not executing any action, separated for each region.

We also reduce the temporal resolution for faster processing, so the effective frame rate for all videos in training and recognition steps is close to 5 fps.

6.2. Common benchmark datasets

6.2.1. MSR-Action3D

The MSR-Action3D dataset (Li, Zhang, & Liu, 2010) consists of 10 actors performing 20 simple atomic actions related to gaming in front of a TV. This dataset provides pose estimation data (joint locations) and low resolution depth maps. We use this dataset to compare the model strengths using simple action videos when only one level of action annotation is provided. To allow our model to learn from this kind of data, we augment the dataset annotations creating an intermediate level of $A + 1$ atomic actions, with A the number of activities, adding an *idle* action to this level. The atomic actions are related to the time span when the activity is being performed, detecting the *idle* poses in training in all frames via a simple heuristic: we assume that in the first frame the person is in *idle*

state, and assign all remaining frames where the geometric descriptor is similar to the one of the first frame via thresholding and filtering out isolated frames. Then, we assign the remaining frames with the corresponding activity label. In this way, the complete video is assigned a single activity action (the original annotation), while only a subset of frames and regions are assigned with the atomic action in the intermediate level.

In our experiments, we use a total of 557 sequences as in (J. Wang et al., 2012). We test our approach using the Setup 2 as in (J. Wang et al., 2012), testing our model using all 20 action categories at once, with subjects 1-3-5-7-9 assigned to the training set and the rest to the testing set. In our evaluation, we use $K = 100$ poses for each body region, for a total of 400 pose dictionary entries.

6.2.2. Composable Activities Dataset

In order to test the suitability of our model for recognizing complex and composable activities, we use the Composable Activities benchmark dataset from (Lillo et al., 2014). This dataset consists of 694 RGB-D videos that contain activities in 16 classes performed by 14 actors. Each RGB-D sequence was captured using a Microsoft Kinect sensor, and the dataset is provided with the 3D position of relevant body joints estimated according to the SDK described in (Microsoft, 2012). Each activity in this dataset is spatio-temporally composed by a variable number of mid-level (atomic) actions. Every actor performs each activity 3 times in average. The total number of actions in the videos is 25 (plus an *idle* action), while the number of actions that compose each particular activity fluctuates between 2 to 10 actions. Figure 6.1 summarizes the composition of atomic actions for each activity, note that activities such as *Composed activity 4* can be composed of up to 10 atomic actions. The RGB-D data and annotations for this dataset are publicly available[1].

For evaluation, we use leave-one-subject-out cross-validation for all models.

[1] `http://web.ing.puc.cl/~ialillo/ActionsCVPR2014/`

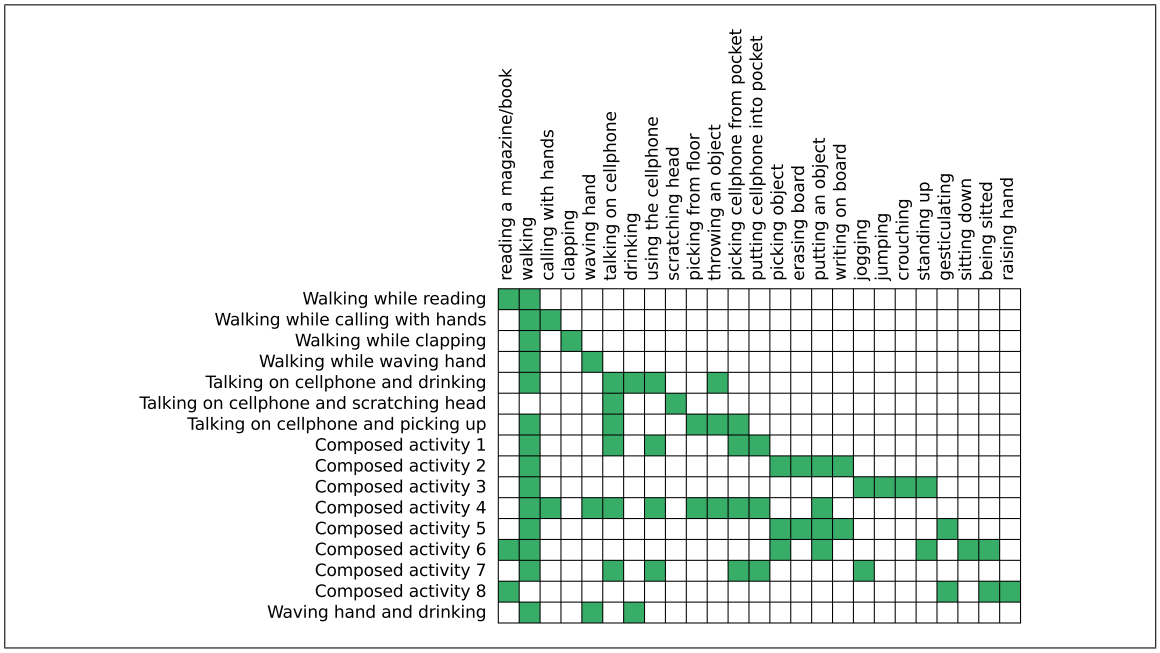

Figure 6.1. Composition of actions (columns) into activities (rows). Note that some activities are simpler, composed only by two actions, while others are very complex, including up to ten different atomic actions per video. Activities also share an *idle* action, not shown in the table.

6.3. Evaluation of the *core model*

To validate the *core model*, we perform a series of evaluations using the benchmark datasets of Section 6.2. In particular, the Composable Activities Dataset is a more suitable benchmark given the hierarchical structure of activities and atomic actions, providing spatio-temporal labels as required by the *core model*. Nevertheless, we show competitive results in a simpler action recognition dataset as MSR-Action3D with the advantage of retrieving the temporal action labels for every region of the human body in a single framework, which is not possible when using a classifier which output a single action label, illustrating the benefits of using a hierarchical and compositional model.

6.3.1. Performance of the *core model* with single action videos

To evaluate the model's action recognition performance using single action videos, we used the MSR-Action 3D dataset, using the annotated frames as described in Section 6.1. In order to illustrate the hierarchical capabilities of our model, we keep the global activity

Algorithm	Accuracy
Core model	89.5%
Core model, GEO+MOV features	90.6%
Core model, GEO+MOV features and NI handling	93.0%
Multimodal model	93.0%
J. Luo et al. (Luo, Wang, & Qi, 2013)	**96.7**%
Tao and Vidal (Tao & Vidal, 2015)	93.6%
Lu & Tang (Lu & Tang, 2014)*	95.6%
Yang & Tian (X. Yang & Tian, 2014)*	93.1%
C. Wang et al.(C. Wang et al., 2013)	90.2%
Vemulapalli et al. (Vemulapalli et al., 2014b)	89.5%
Lillo et al. (Lillo et al., 2014)	89.5%
J. Wang et al. (J. Wang et al., 2012)	88.2%

Table 6.1. Recognition accuracy rates in the MSR-Action3D dataset for our approach and alternative state-of-the-art methods. i) Using only the core model; ii) using the core model but with improved pose feature descriptor GEO+MOV; iii) using the *garbage collector* mechanism for non-informative (NI) poses; and iv) Using the proposed multimodal model. * indicates the use of depth features instead of 3D pose estimation data.

label but also annotate all frames in the video with the given action class, except those frames where the subject is standing still, which we label as *idle* using a simple heuristic.

Our *core model* achieves an action classification accuracy of 89.5%, a recognition performance that is on par with state-of-the-art. Although this dataset does not provide a rich hierarchy of complex activities composed by atomic actions, the results allow us to validate that our model performs well on the task of single action recognition.

As with competing methods, most of the actions can be recognized with almost perfect accuracy, but some actions are still dificult to discriminate like *hand catch* and *high throw*. Table 6.1 shows the accuracy of our method compared to state-of-the-art in this dataset.

6.3.2. Performance of the *core model* with complex activity videos

For complex activity videos we perform exhaustive tests in Composable Activity dataset (Lillo et al., 2014) as benchmark, which has two levels of annotations: the top level is a single activity, and the second level is a spatio-temporal composition of atomic actions. For instance, the activity *walk while hand waving* has a spatio-temporal composition of 3 single actions: *walk*, *hand wave*, and *idle*; while the activity *composed-activity-4* is composed of 11 single actions: *idle*, *walk*, *call a friend with hands*, *hand wave*, *talking on cellphone*, *pick*

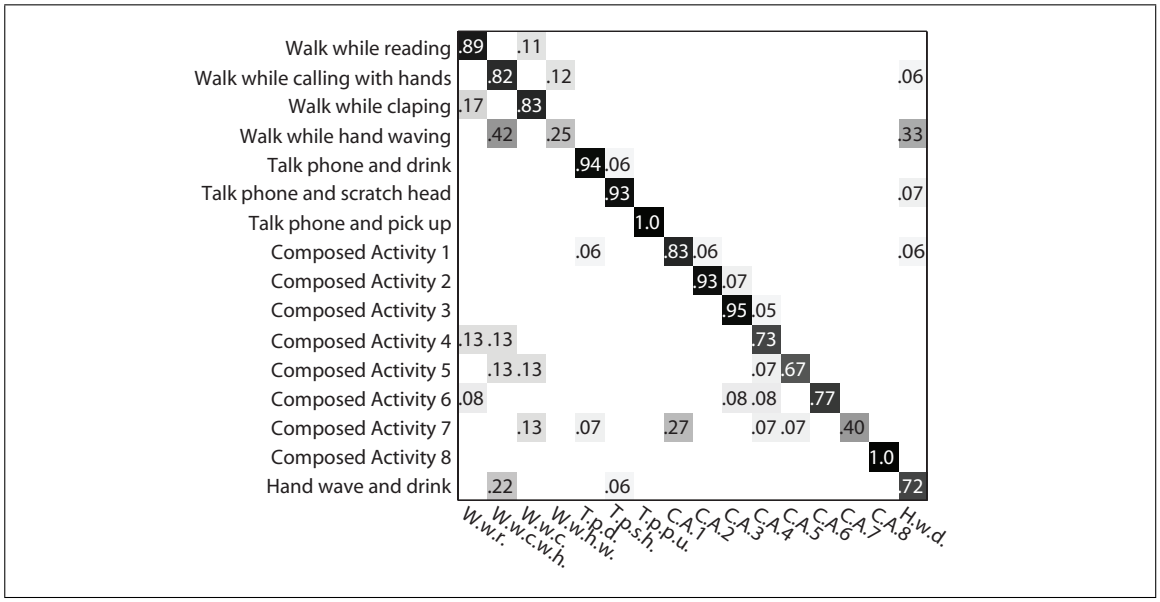

Figure 6.2. Confusion matrix for the activity classification task in the new Composable Activities dataset.

from the floor, *dial cellphone*, *put an object*, *pick cellphone from pocket*, and *put cellphone in pocket* (see Figure 1.1).

In our cross-validation setup, the accuracy of our model is 84.2%, when using $K = 50$ poses for each body part (a total of 200 poses), which provides a good compromise between model complexity and accuracy. We also set the model parameters λ_1 to 100, and λ_2 to 20. In general, we use cross-validation to adjust the value of all our main parameters.

We compare the performance of our method with respect to three baselines techniques: a BoW representation plus a lineal SVM classifier (BoW-approach), a version of our model without learning the pose dictionary (H-BoW-approach), and a Hidden Markov Model approach (HMM-approach). In the case of BoW-approach we use k-means algorithm to obtain a pose dictionary that is then used to quantize observed poses. We build a histogram of poses using all frames of the sequence and considering also a Spatial Pyramid Matching (SPM) scheme. The accuracy of this baseline is 66.7%. Our model demonstrates a substantial accuracy improvement, exploiting the ability to model activities and actions, and jointly learning a pose dictionary. A second baseline consists of a simplified version of our hierarchical model that does not learn the pose dictionary, but uses a pose quantization

Algorithm	Codebook size	Accuracy
Core model	200 (learned)	**84.2%**
Core model	600 (learned)	**82.7%**
BoW	200 (fixed)	66.7%
BoW	600 (fixed)	61.2%
H-BoW	200 (fixed)	73.0%
H-BoW	600 (fixed)	70.5%
HMM	200 (fixed)	73.6%
HMM	600 (fixed)	72.6%

Table 6.2. Recognition accuracy of our method compared to three baselines: Bag-of-Visual-Features (BoW), our method but without learning pose dictionary (H-BoW), and a Hidden Markov Model approach (HMM).

given by the k-means algorithm. In this case, the accuracy drops by 11%, validating our discriminative learning scheme to learn the pose dictionary. The third baseline is an HMM model that we learn using atomic actions as states and poses as observed variables. In this case, we independently learn a model for each class. At test time, we score a new sequence using all models, and select the activity label that corresponds to the model with highest log-likelihood. In this case, we obtain an accuracy of 73.6%. Recognition rates are summarized in Table 6.3.

Effects of size of pose dictionary The proposed method is relatively robust to the size of the pose dictionary. A low number of poses per body part (5 to 20) lacks reprensentativity, and a high number increases the computational load. In our experiments, we observe similar performance for the case of 50, 100 and 150 poses per body part. When testing with 25 poses the accuracy drops by 6%. We did not test larger dictionaries due to the processing time, which is quadratic with respect to dictionary size. We chose 50 poses per body part in our experiments as a compromise between good accuracy and processing speed.

Importance of transition terms in the model When we simplify our model by fixing the pose dictionary and dropping the energy terms related to action and pose transitions, we observe a drop in accuracy of 11.2%. If we learn the pose dictionary, but ignore the temporal transition components γ and η, accuracy drops by 4.8%. As expected, learning temporal cooccurrence improves the accuracy of our method, as it links poses and actions over time. In our current model, we use a single frame correlation; this short-term relation

could be expanded to middle or long term correlations, with the cost of an increased running time.

Action annotation Beyond activity categorization, the hierarchical structure and compositional properties of our model enable it to perform per-frame annotation of the atomic actions that compose each activity. Furthermore, it can also indicate which body parts are associated to the atomic actions present in a frame, as well as, the temporal span of each action. We illustrate this capability in Figure 6.3. We also evaluate the effectiveness of our algorithm in correctly annotating mid-level actions at each frame. This is again a multiclass classification task that we also summarize in a confusion matrix (Figure 6.4) whose diagonal average is 46.5%.

Robustness to occlusion Our method is also capable of inferring action and activity labels even if some joints are not observed. To illustrate this, we simulate an occluded part by fixing it to the position observed in the first frame. We select a part to be occluded in every sequence using a uniform sampling. In this scenario, the accuracy of our model drops by 7.2%, while the drops in performance of BoW is (12.5%) and HMM (10.3%). Also, Figure 6.12 shows some qualitative results.

6.4. Evaluation of the *sparse model*

Now we validate the effectiveness of the *sparse model* variarion of our proposed model at the activity recognition task by measuring activity classification performance on RGB-D videos. Like the *core model*, we consider two experimental scenarios. First, we test the ability of our approach to discriminate simple actions on MSR-Action3D. Second, we test the performance of our model in recognizing complex and composable human activities in Composable Activities dataset. We also study the contribution of key components of our model and their impact in recognition performance. In particular, we highlight the contribution of the extensions of the *sparse model* framework with respect to the *core model*.

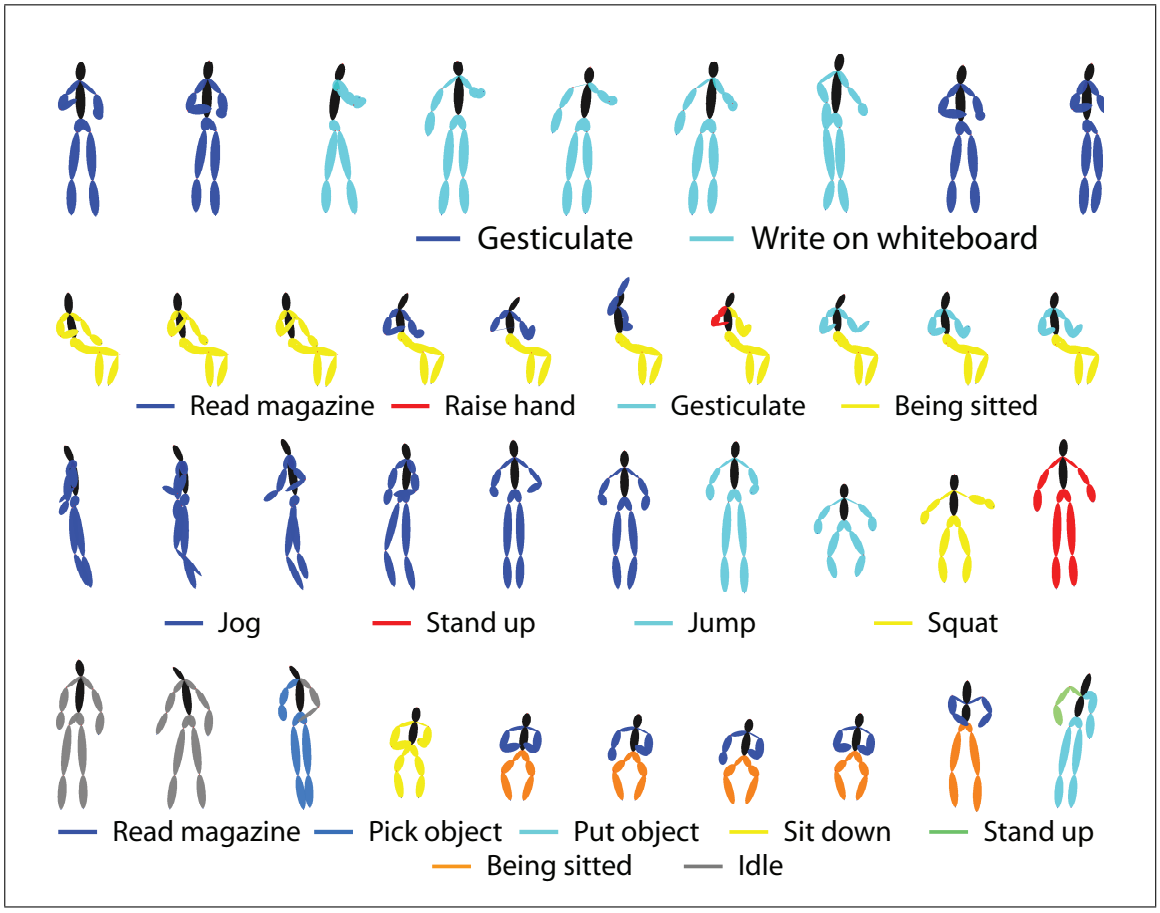

Figure 6.3. Per-frame simple action annotation results. This figure shows several example sequences which our algorithm classifies to the correct activity category. Furthermore, we show how our algorithm is able to correctly predict the atomic actions that compose each activity and which body parts contribute to those actions. Here, we color each body part according to the predicted action label. Best viewed in color.

6.4.1. Performance of the *sparse model* with single action videos

While our main goal is the recognition of human activities that can be composed of simpler atomic actions, we experimentally verify that the model can also handle the recognition of atomic actions. Towards this goal, we evaluate our algorithm on the MSR-Action3D dataset.

Tables 6.1 report recognition accuracy of our model for the MSR-Action3D dataset using Setup 2. Accuracy is measured by the average of the diagonal of the normalized confusion matrix. We can observe that, although designed for complex activity recognition,

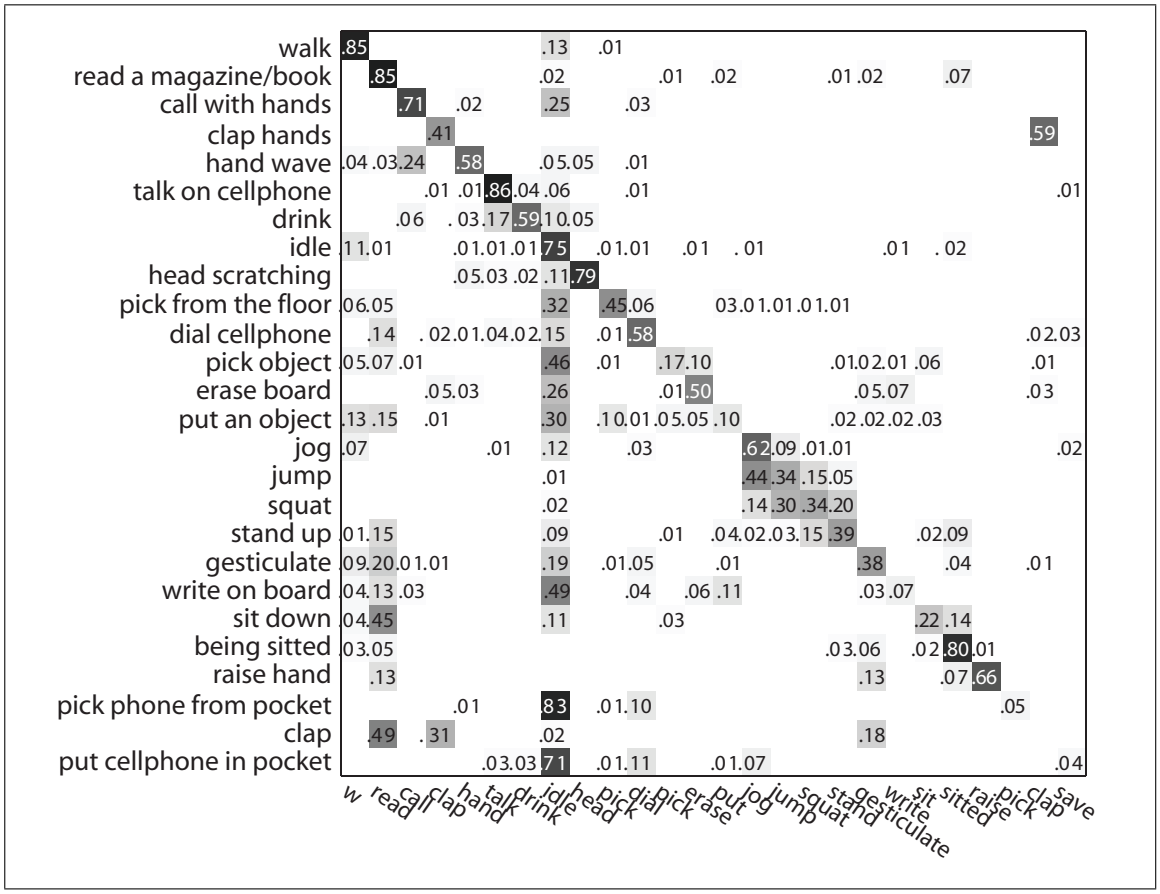

Figure 6.4. Confusion matrix for the action classification task using the *core model*. Rows are the ground truth actions at each frame, while columns are the predicted mid-level action label inferred by our model.

our model can also achieve competitive results in the task of single atomic action recognition. An important aspect of these results is that our model achieves this performance using a total of 400 pose dictionary entries (100 entries per body region). This compares favorably with the results reported in (J. Wang et al., 2012) and (C. Wang et al., 2013), which use dictionaries with thousands of entries. Our reduced dictionary translates to compact representations that provide more meaningful interpretarions and require less computation at the inference stage.

Table 6.1 also reports the performance of our model under several settings. First, we report performance when the model uses the geometric descriptor (GEO) described in Section 3.1. This corresponds to the feature used in the *core model*. We also consider

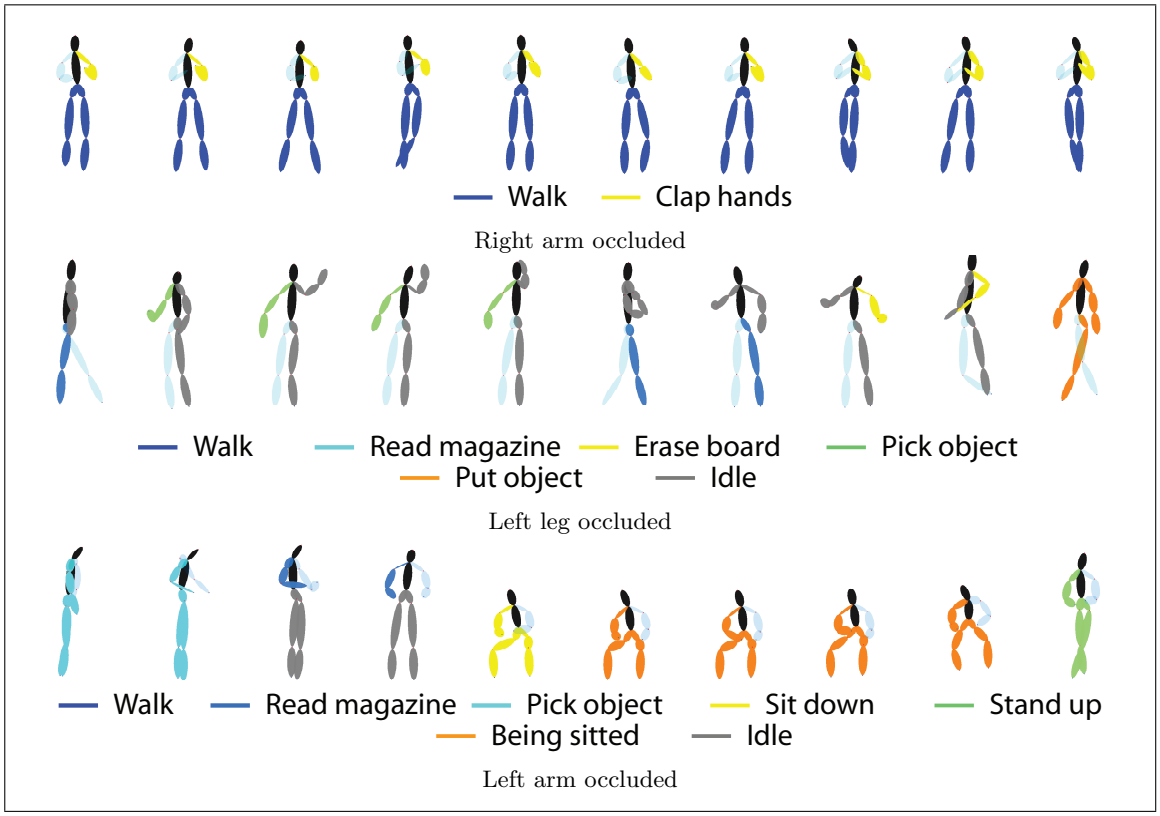

Figure 6.5. The occluded body parts are depicted in light blue. When an arm or leg is occluded, our method still provides a good estimation of actions in each frame.

the case when we combine the geometric descriptor with the motion descriptor described in Section 4.2 (GEO/MOV). Additionally, we consider the case when the model includes the method to identify and discard non-informative frames (NI). In the specific case of MSR-Action3D, we use the quadratic regularizer (same as *core model*) since every video only contains a single action, so the poses that compose every action are naturally sparse. Also, MOV descriptor includes actions involving fine motions of hands and foots, we also incorporate to our descriptor the spatial gradient of motions of body joints associated to wrists for upper body regions and ankles for lower body regions, adding three features (differences in x, y and z) for every region.

6.4.2. Performance of the *sparse model* with complex activity videos

In order to test the suitability of our model for recognizing complex and composable activities, we use the Composable Activities benchmark dataset as in the *core model*.

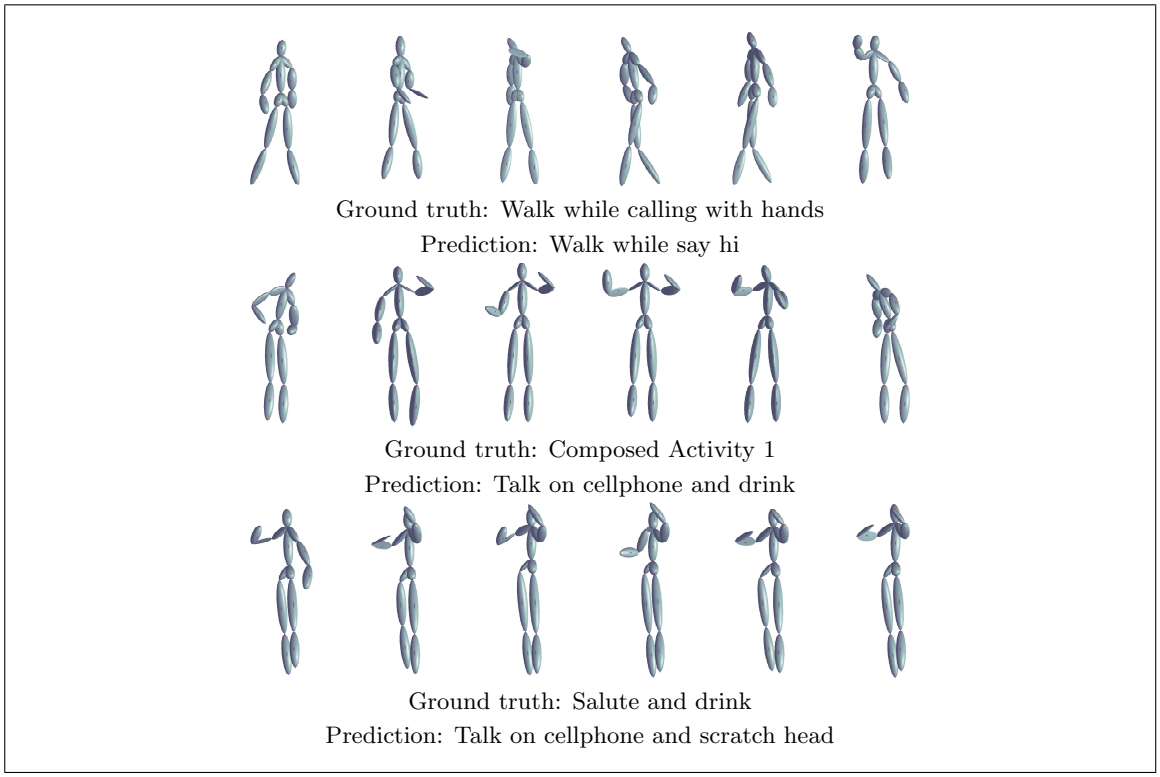

Figure 6.6. Failure cases. Our algorithm tends to confuse activities that share very similar body postures.

Recognition rates for the Composable Activities Dataset are summarized in Table 6.3. We report performance averaged over multiple runs using a leave-one-subject-out cross-validation strategy. We use a validation set to experimentally adjust the value of all the main parameters. In practice, we set $\lambda_1 = 100$ and $\lambda_2 = 20$. We set $K = 50$ pose dictionary entries per body region when using the quadratic regularizer (QR), and $K = 150$ per body region when using the proposed sparse regularizer (SR). Also, we use fixed parameter values $\mu = 0.1$ and $\rho = 5$ for SR, and $\mu = 1$ and $\rho = 0$ for QR.

Table 6.3 reports our recognition results in context by comparing them to the performance of two simplified versions of our model, and a state-of-the-art algorithm. The first baseline is a bag of words model (BoW), which only captures very coarse per-region pose orderings and uses an independently pre-trained pose dictionary. Specifically, this baseline uses k-means to quantize pose descriptors for each body region independently, which are

Algorithm	GEO	GEO/MOV
Core model	84.2%	90.9%
Core model+NI	88.5%	91.8%
Sparse model	84.9%	90.6%
Sparse model+NI	–	**92.2%**
BoW	67.2%	74.1%
HMM	76.5%	78.9%
H-BoW	74.2%	82.4%
2-lev-HIER	79.6%	83.8%
LG (Vemulapalli et al., 2014b)	74.7%	–
Cao et al. (Cao, Zhang, & Lu, 2015)	79.0%	

Table 6.3. Recognition accuracy of our method compared to several baselines (see Section 6.4.1). It is noteworthy that our 3-level model outperforms all 2-levels models. Also, including motion cues in the descriptor (GEO/MOV) and using non-informative poses handling (NI) improve the accuracy over our previous model. The best performance is obtained when using all the contributions described in this work.

aggregated into a temporal pyramid histogram representation. This is then fed into a multi-class linear SVM for directly mapping from video descriptors to activities. The accuracy of this baseline is 74.1% when using the combined descriptor based on geometric and motion information.

As a second baseline, we implement a Hidden Markov Model (HMM). The HMM model can directly encode pose and action transitions built upon an independently pre-trained pose dictionary. In our implementation, states are trained with supervision by assigning one state to each atomic action. Quantized poses are the observed variables. We train models independently for each class, and at testing time, we classify new sequences by assigning the label that corresponds to the highest scoring model. The accuracy of this baseline is 78.9% when using the combined descriptor based on geometric and motion information. While the ordering encoded by the HMM model helps to improve accuracy over BoW, it still lacks the discriminative power provided by the joint learning of mid and top level representations, as performed by our proposed model.

We also compare performance against two simplified versions of our hierarchical model. The first simplified version (H-BoW) does not jointly learn the pose dictionary, but uses a fixed pose quantization obtained with k-means, and omits the transition terms. Unlike our full model, this simplified version does not take advantage of jointly learning the pose

dictionary, which leads to sub-optimal pose encoding at the lower level of the hierarchy. Also, by omitting the transition terms, the model cannot capture patterns in the evolution of actions and poses. These simplifications lead to a 10% drop in performance in comparison to our full model.

As a second simplification of our full model, we construct a hierarchical model with only two coupled layers (2-lev-HIER) that are jointly trained to encode poses and atomic actions. In this simplified model, activity recognition is performed by an independently trained linear classifier that operates on top of the inferred atomic actions. In this case, the performance of this model simplification is 8.4% lower than our full model. This indicates the clear benefit of jointly learning the mid-level representations and the top level activity classifier.

We also compare against an existing state-of-the-art algorithm from the literature. In this case, we compare to the method recently described in (Vemulapalli et al., 2014b) (LG). We select this algorithm because it achieves state-of-art performance on several pose-based action datasets. We train this model to directly predict the activities from poses, omitting the mid-level annotations, as in the BoW baseline. While the accuracy of LG is above BoW, it is still 11% lower than our model that only uses geometric information (GEO).

The confusion matrix obtained with our full model is reported in Figure 6.7. Note that for some activities the prediction is perfect, while for others there is high confusion between some activities. Large confusion may be caused by highly similar poses, for instance between *calling with hands* and *waving hand*, where many actors perform the *calling with hands* action with only one arm.

As a main advantage, in addition to the high recognition performance, our model also generates a rich video interpretation in the form of detailed per-frame and per-body-region action annotations. Figure 6.8 shows the action labels associated to each body part in a test video from the Composable Activities dataset. This example illustrates the capability of our model to correctly identify, spatially and temporally, the main body parts that are involved in the execution of a given action.

To illustrate the semantic interpretation of the poses learned by our model, Figure 6.9 shows top-scoring frames for three activities executed by different subjects. In general,

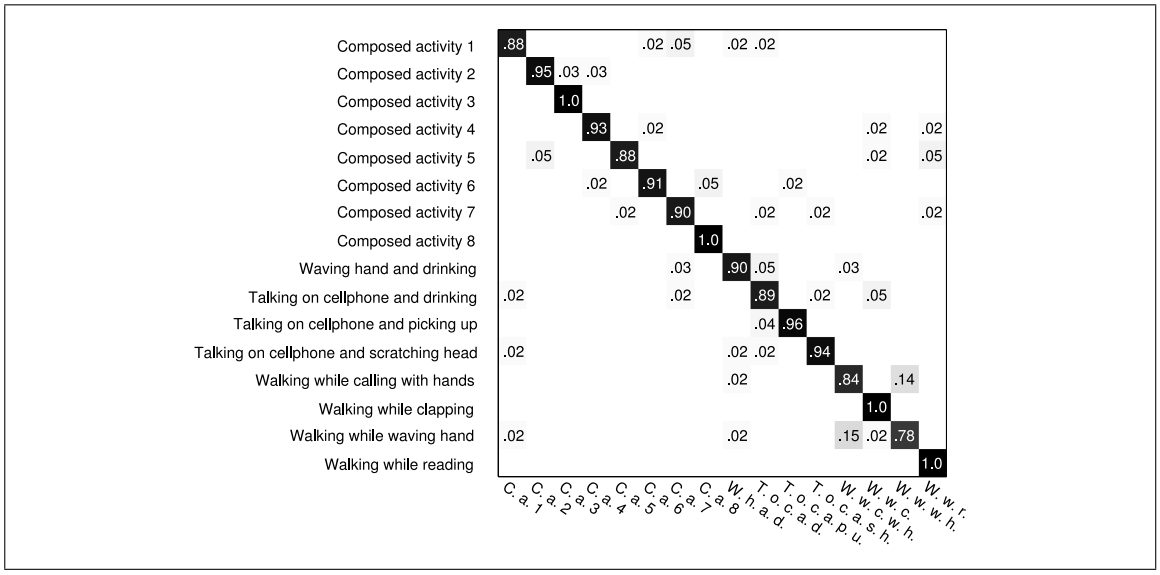

Figure 6.7. Confusion matrix for the activity classification task in the Composable Activities dataset, using sparse regularization, a GEO/MOV descriptor, and NI handling.

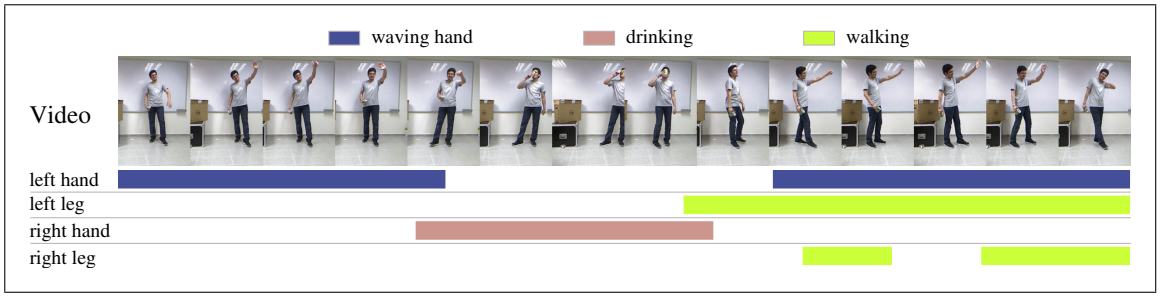

Figure 6.8. Automatic spatio-temporal annotation of atomic actions. Our method automatically detects the temporal span and spatial body regions that are involved in the performance of atomic actions in videos.

our model produces highly interpretable poses that are associated to characteristic body configurations of the underlying atomic actions. To further illustrate this observation, Figure 6.10 shows the highest activations for eight pose dictionary entries associated to the body region corresponding to the left arm. In each case, Figure 6.10 also indicates the atomic action that assigns a greatest relevance (weight) to the corresponding pose.

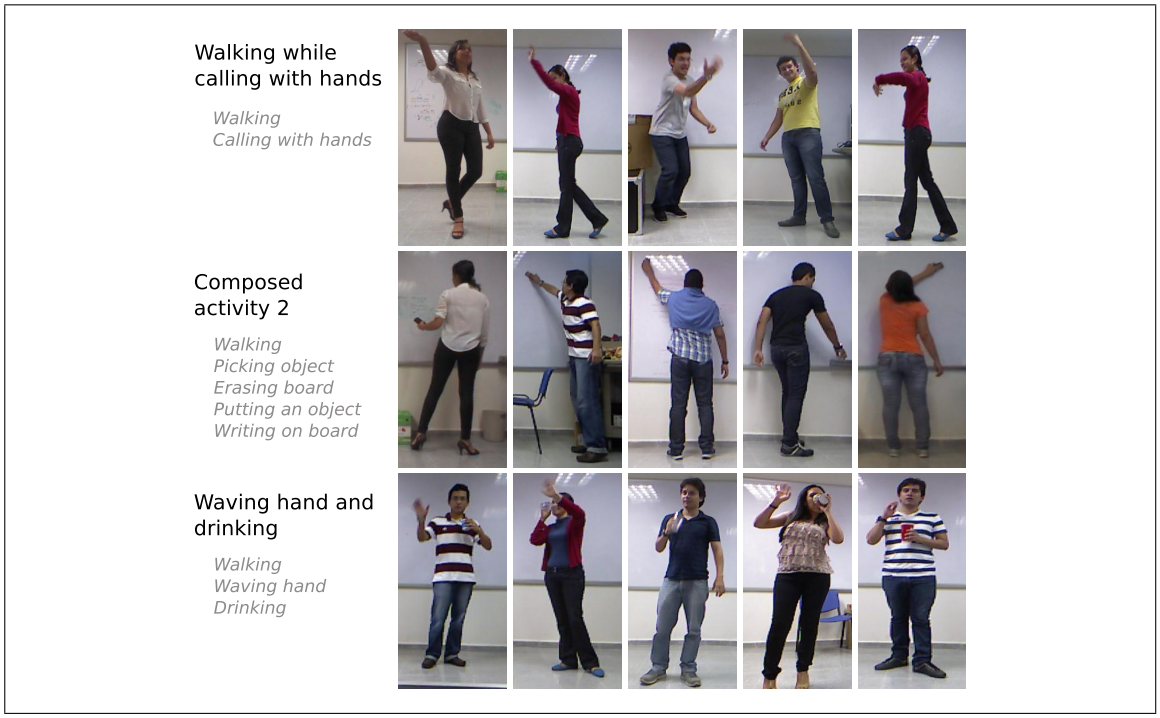

Figure 6.9. Examples of top scoring frames for three activities. Note the high correlation between the actions that compose each activity and the pose of the actors.

6.4.3. Impact of motion descriptor

Tables 6.1 compare the performance of the geometric (GEO) and combined (GEO/MOV) descriptors under different model configurations using the MSR-Action3D and Composable Activities datasets.

Since the MSR-Action3D dataset does not include RGB information, we modify the motion descriptor presented in Section 4.2. Specifically, we encode differences of displacements for every joint, in a similar setup to (H. Wang et al., 2011). By incorporating this motion descriptor, we achieve a recognition accuracy of 90.6% in the MSR-Action3D dataset, reducing error rate by 1.1% with respect to the model presented in (Lillo et al., 2014). In case of the Composable Activities dataset, we use the motion descriptor presented in Section 4.2. By incorporating this descriptor, we achieve a recognition accuracy of 90.9% in the Composable Activities dataset when using a QR, reducing error rate by 5.2% with respect to the model presented in (Lillo et al., 2014).

Figure 6.10. Examples of top scoring poses for the body region corresponding to the left arm. Also shown, it is the label of the action with the highest classifier weight associated to the pose. In this case the model is trained using SR and $K = 150$ for each body region.

6.4.4. Impact of handling non-informative poses

During learning, we need to initialize the set of candidate frame poses that can be considered as NI. In practice, we initialize the NI poses by using the initial pose dictionary obtained with k-means, and selecting as NI the poses that are most distant to their assigned cluster centers. For each video, we initially assign a total of 20% of the frames to the NI bucket. As learning progresses, on each iteration poses can be reassigned to a pose dictionary entry or to NI. In general, we observe that after convergence approximately 17% of all training frames are assigned as non-informative. When we initialize the NI assignment with 40% of the frames, our final model reduces this to 19%. This high degree of robustness with respect to the initialization, indicates that the model is effectively learning to detect non-informative frames. Moreover, near 40% of the initial pose assignments are updated throughout the learning iterations. We can compare this with the version of the model in (Lillo et al., 2014), that does not include non-informative pose handling, and where only

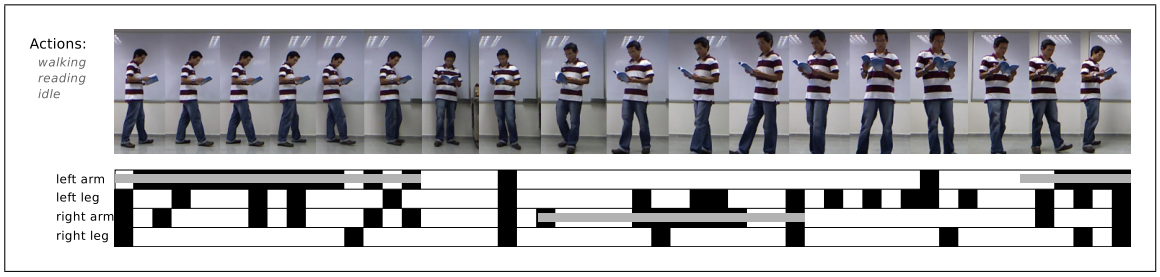

Figure 6.11. Non-informative pose sequence for the four regions of the body, in a video from the activity *Walking while reading*. The black squares represent frames labeled as a non-informative pose. A thick gray line shows when the corresponding region is occluded. We can observe a relation between body region occlusions and identification of non-informative poses. Specifically, when there is no occlusion, the identification of non-informative poses tends to be temporally sparse, but for occluded intervals, many consecutive frames are selected as non-informative.

25% of the initial pose assignments are updated throughout the learning iterations. This indicates that our full model is capable of updating the pose representations more effectively.

In terms of accuracy, the use of NI reduces error rate by 1.8% and raises recognition accuracy from 90.6% to 92.4% in the MSR-Action3D dataset. In the Composable Activities dataset, the introduction of NI reduces error rate by 1.4% and raises recognition accuracy to 92.2%. An important ability of the NI mechanism is that occluded regions are often assigned as NI poses. Figure 6.11 shows a sequence of the activity *Walking while reading*. In this figure, the bottom graph shows with black boxes frames where a body region is identified by our method as corresponding to a non-informative pose. Observing the body region corresponding to the arms, the long sequences of non-informative poses nearly coincides with the occlusion periods of the arms (thick gray lines). Other frames considered as non-informative tend to be sparser in time, and they can be explained by rare poses or noisy body joints estimation. This behavior is advantageous in two ways: during learning, it allows the model to automatically disregard many occluded regions when learning the pose classifiers; and during testing, it allows the model to identify possible occluded regions.

6.4.5. Inference of per-frame annotations.

The hierarchical structure and compositional properties of our model enable it to output a predicted global activity, as well as per-frame annotations of predicted atomic actions and

poses for each body region. We highlight that, in the generation of the per-frame annotations, no prior temporal segmentation of atomic actions is needed. Also, no post-processing of the output is performed. The proficiency of our model to produce per-frame annotated data, enabling atomic action detection temporally and spatially, is a key advantage of our model.

Figure 6.8 illustrates the capability of our model to provide per-frame annotation of the atomic actions that compose each activity. The accuracy of the mid-level action prediction can be evaluated as in (Wei et al., 2013). Specifically, we first obtain segments of the same predicted action in each sequence, and then compare these segments with ground truth action labels. The estimated label of the segment is assumed correct if the detected segment is completely contained in a ground truth segment with the same label, or if the Jaccard Index considering the segment and the ground truth label is greater than 0.6. Using these criteria, the accuracy of the mid-level actions is 79.4%. In many cases, the wrong action prediction is only highly local in time or space, and the model is still able to correctly predict the activity label of the sequence. Considering only the correctly predicted videos in terms of global activity prediction, the accuracy of action labeling reaches 83.3%. When consider this number, it is important to note that not every ground truth action label is accurate: the videos were hand-labeled by volunteers, so there is a chance for mistakes in terms of the exact temporal boundaries of the action. In this sense, in our experiments we observe cases where the predicted labels showed more accuracte temporal boundaries than the ground truth.

6.4.6. Robustness to occlusion and noisy joints.

Our method is also capable of inferring action and activity labels even if some joints are not observed. This is a common situation in practice, as body motions induce temporal self-occlusions of body regions. Nevertheless, due to the joint estimation of poses, actions, and activities, our model is able to reduce the effect of this problem. To illustrate this, we simulate a totally occluded region by fixing its geometry to the position observed in the first frame. We select which region to be completely occluded in every sequence using uniform sampling. In this scenario, the accuracy of our preliminary model in (Lillo et al., 2014) drops by 7.2%. Using our new SR setup including NI handling, the accuracy only

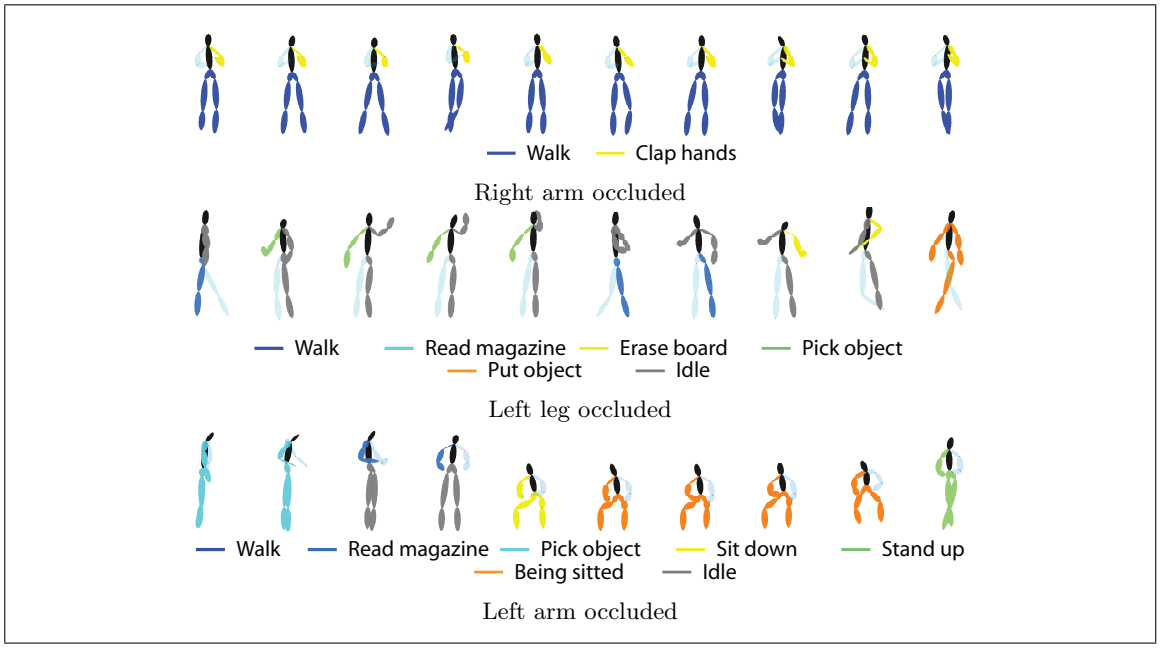

Figure 6.12. The occluded body regions are depicted in light blue. When an arm or leg is occluded, our method still provides a good estimation of the underlying actions in each frame. Best viewed in color.

drops by 4.3%, showing that the detection of non-informative poses helps to reduce the effect of occluded regions. In fact, as we show in Section 6.4.4, many of truly occluded regions in the videos are identified using NI handling. In contrast, the drop in performance of BoW is 12.5% and HMM 10.3%. This is expected, as simpler models are less capable of robustly dealing with occluded regions, since their pose assignments rely only on the descriptor itself, while in our model the assigned pose depends on the descriptor, sequences of poses and actions, and the activity evaluated, making inference more robust. Figure 6.12 shows some qualitative results for cases displaying occluded regions.

In terms of noisy joints, we manually add random Gaussian noise to change the joints 3D location of testing videos, using the SR setup and the GEO descriptor to isolate the effect of the joints and not mixing the motion descriptor. Figure 6.13 shows the accuracy of testing videos in terms of noise dispersion σ_{noise} measured in inches. For little noise, there is no much effect in our model accuracy, as expected due to the robustness of the geometric descriptor. However, for more drastic noise levels, the accuracy drops dramatically. This

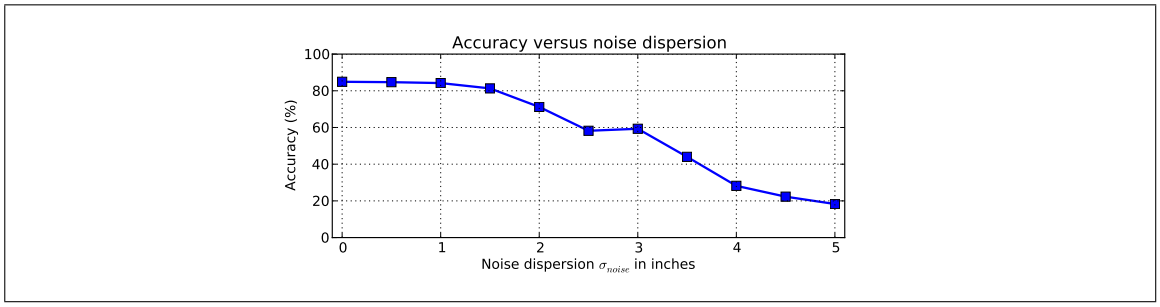

Figure 6.13. Performance of our model in presence of simulated Gaussian noise in every joint, as a function of σ_{noise} measured in inches. When the noise is less than 3 inches in average, the model performance is only slightly affected, while under higher noise dispersion the model accuracy is drastically affected. It is important to note that in real data high levels of noisy joint estimation tend to occur rarely.

behavior is expected, since for highly noisy joints the model can no longer predict well the sequence of actions and poses.

6.5. Evaluation of the *multimodal model*

Our experimental validation focuses on evaluating the new properties of the model compared to the *core model*.

We evaluate our method on four action recognition benchmarks: the MSR-Action3D dataset (Li et al., 2010), Concurrent Actions dataset (Wei et al., 2013), Composable Activities Dataset (Lillo et al., 2014), and sub-JHMDB (Jhuang et al., 2013). Using cross-validation, we set $K = 100$ in Composable Activities and Concurrent Actions datasets, $K = 150$ in sub-JHMDB, and $K = 200$ in MSR-Action3D. In all datasets, we fix $\lambda_y = 100$ and $\lambda_u = 25$. The number of *actionlets* to model each atomic action is estimated using the method described in Section 5.2.

The garbage collector (GC) label $(K + 1)$ is automatically assigned during inference according to the learned model parameters θ^r. We initialize the 20% most dissimilar frames to the $K+1$ label. In practice, at test time, the number of frames labeled as $(K+1)$ ranges from 14% in MSR-Action3D to 29% in sub-JHMDB.

Computation is fast during testing. In the Composable Activities dataset, our CPU implementation runs at 300 fps on a 32-core computer, while training time is 3 days, mostly due to the massive execution of the cutting plane algorithm. Using Dynamic Programming,

complexity to estimate labels is linear with the number of frames T and quadratic with the number of actionlets A and motion poselets K. In practice, we filter out the majority of combinations of motion poses and actionlets in each frame, using the $P = 400$ best combinations of (k, a) according to the value of non-sequential terms in the dynamic program, as described in Section 3.2.1

6.5.1. Classification of Simple and Isolated Actions

As a first experiment, we evaluate the performance of our model on the task of simple and isolated human action recognition in the MSR-Action3D dataset (Li et al., 2010). Although our model is tailored at recognizing complex actions, this experiment verifies the performance of our model in the simpler scenario of isolated atomic action classification. Table 6.1 shows that in this dataset our model achieves classification accuracies comparable to state-of-the-art methods. Note that in the *multimodal model* for MSR-Action3D there is no need for multiple *actionlets* per atomic action, since every actor executes the same movement in a single action. Nevertheless, it is noteworthy that in this case we did not have to use the heuristic to create the *idle* frames.

Although our model did not achieve state-of-the-art accuracy, it provides rich semantic action annotations and semantic pose templates that other models do not provide.

6.5.2. Detection of Concurrent Actions

Our second experiment evaluates the performance of our model in a concurrent action recognition setting. In this scenario, the goal is to predict the temporal localization of actions that may occur concurrently in a long video. We evaluate this task on the Concurrent Actions dataset (Wei et al., 2013), which provides 61 RGBD videos and pose estimation data annotated with 12 action categories. We use a similar evaluation setup as proposed by the authors. We split the dataset into training and testing sets with a 50%-50% ratio. We evaluate performance by measuring precision-recall: a detected action is declared as a true positive if its temporal overlap with the ground truth action interval is larger than 60% of their union, or if the detected interval is completely covered by the ground truth annotation.

Algorithm	Precision	Recall
Our full model	0.92	0.81
Wei et al. (Wei et al., 2013)	0.85	0.81

Table 6.4. Recognition accuracy in the Concurrent Actions dataset.

Our model is tailored at recognizing complex actions that are composed of atomic components. However, in this scenario, only atomic actions are provided and no compositions are explicitly defined. Therefore, we apply a simple preprocessing step: we cluster training videos into groups by comparing the occurrence of atomic actions within each video. The resulting groups are used as complex actions labels in the training videos of this dataset. At inference time, our model outputs a single labeling per video, which corresponds to the atomic action labeling that maximizes the energy of our model. Since there are no thresholds to adjust, our model produces the single precision-recall measurement reported in Table 6.4. Our model outperforms the state-of-the-art method in this dataset at that recall level.

6.5.3. Recognition of Composable Activities

In this experiment, we evaluate the performance of our model to recognize complex and composable human actions. In the evaluation, we use the Composable Activities dataset (Lillo et al., 2014). We train our model using two levels of supervision during training: i) spatial annotations that map body regions to the execution of each action are made available ii) spatial supervision is not available, and therefore the labels $\vec{v}$ to assign spatial regions to actionlets are treated as latent variables.

Table 6.5 summarizes our results. We observe that under both training conditions, our model achieves comparable performance. This indicates that our weakly supervised model can recover some of the information that is missing while performing well at the activity categorization task. In spite of using less supervision at training time, our method outperforms state-of-the-art methodologies that are trained with full spatial supervision.

6.5.4. Action Recognition in RGB Videos

Our experiments so far have evaluated the performance of our model in the task of human action recognition in RGBD videos. In this experiment, we explore the use of our

Algorithm	Accuracy
Core model + GC, GEO desc. only, spatial supervision	88.5%
Core model + GC, with spatial supervision	91.8%
Our full model, no spatial supervision (latent $\vec{v}$)	91.1%
Lillo *et al.*(Lillo et al., 2014) (without GC)	84.2%
Cao et al. (Cao et al., 2015)	79.0%

Table 6.5. Recognition accuracy in the Composable Activities dataset.

Figure 6.14. Examples of actionlets using high-scored frames, for testing videos in sub-JHMDB dataset. Actionlets 2 ans 3 belong to the *catch* action, Actionet 10 and 11 to *golf* action, Actionlet 25 to *pick* and Actionlet 32 to *push*. Note that actionlets are highly related to poses and movements of the subjects in the videos.

model in the problem of human action recognition in RGB videos. For this purpose, we use the sub-JHMDB dataset (Jhuang et al., 2013), which focuses on videos depicting 12 actions and where most of the actor body is visible in the image frames. In our validation, we use the 2D body pose configurations provided by the authors and compare against previous methods that also use them. Given that this dataset only includes 2D image coordinates for each body joint, we obtain the geometric descriptor by adding a depth coordinate with a value $z = d$ to joints corresponding to wrist and knees, $z = -d$ to elbows, and $z = 0$ to other joints, so we can compute angles between segments, using $d = 30$ fixed with cross-validation. We summarize the results in Table 6.6, which shows that our method outperforms alternative state-of-the-art techniques. In Figure 6.14, we show high scored frames which were labeled with the corresponding actionlet during testing, using sub-JHMDB dataset. In Figure 6.15, we show an extended set of motion poselets learned from Composable Activities dataset.

Algorithm	Accuracy
Our model	77.5%
Huang et al. (Jhuang et al., 2013)	75.6%
Chéron et al. (Chéron et al., 2015)	72.5%

Table 6.6. Recognition accuracy in the sub-JHMDB dataset.

Videos	Annotation inferred	Precision	Recall
Testing set	Spatio-temporal, no GC	0.59	0.77
Testing set	Spatio-temporal	0.62	0.78
Training set	Spatial only	0.86	0.90
Training set	Spatio-temporal	0.67	0.85

Table 6.7. Atomic action annotation performances in the Composable Activities dataset. The results show that our model is able to recover spatio-temporal annotations both at training and testing time.

6.5.5. Spatio-temporal Annotation of Atomic Actions

In this experiment, we study the ability of our model to provide spatial and temporal annotations of relevant atomic actions. Table 6.7 summarizes our results. We report precision-recall rates for the spatio-temporal annotations predicted by our model in the testing videos (first and second rows). Notice that this is a very challenging task. The testing videos do no provide any label, and the model needs to predict both, the temporal extent of each action and the body regions associated with the execution of each action. Although the difficulty of the task, our model shows satisfactory results being able to infer suitable spatio-temporal annotations.

We also study the capability of the model to provide spatial and temporal annotations during training. In our first experiment, each video is provided with the temporal extent of each action, so the model only needs to infer the spatial annotations (third row in Table 6.7). In a second experiment, we do not provide any temporal or spatial annotation, but only the global action label of each video (fourth row in Table 6.7). In both experiments, we observe that the model is still able to infer suitable spatio-temporal annotations.

6.5.6. Effect of Model Components

In this experiment, we study the contribution of key components of the proposed model. First, using the sub-JHMDB dataset, we measure the impact of three components of our model: garbage collector for motion poselets (GC), multimodal modeling of actionlets, and

Algorithm	Accuracy
Core model, GEO descriptor only	66.9%
Core Model	70.6%
Core Model + GC	72.7%
Core Model + Actionlets	75.3%
Our full model (Actionlets + GC + latent $\vec{v}$)	77.5%

Table 6.8. Analysis of contribution to recognition performance from each model component in the sub-JHMDB dataset.

Initialization Algorithm	Accuracy
Random	46.3%
Clustering	54.8%
Ours	91.1%
Ours, fully supervised	91.8%

Table 6.9. Results in Composable Activities dataset, with latent $\vec{v}$ and different initializations.

use of latent variables to infer spatial annotation about body regions (latent $\vec{v}$). Table 6.8 summarizes our experimental results. Table 6.8 shows that the full version of our model achieves the best performance, with each of the components mentioned above contributing to the overall success of the method.

Second, using the Composable Activities dataset, we also analyze the contribution of the proposed self-paced learning scheme for initializing and training our model. We summarize our results in Table 6.9 by reporting action recognition accuracy under different initialization schemes: i) Random: random initialization of latent variables $\vec{v}$, ii) Clustering: initialize $\vec{v}$ by first computing a BoW descriptor for the atomic action intervals and then perform k-means clustering, assigning the action intervals to the closer cluster center, and iii) Ours: initialize $\vec{v}$ using the proposed self-paced learning scheme. Our proposed initialization scheme helps the model to achieve its best performance.

6.5.7. Qualitative Results

Finally, we provide a qualitative analysis of relevant properties of our model. Figure 6.15 shows examples of moving poselets learned in the Composable Activities dataset. We observe that each moving poselet captures a salient body configuration that helps to discriminate among atomic actions. To further illustrate this, Figure 6.15 indicates the most

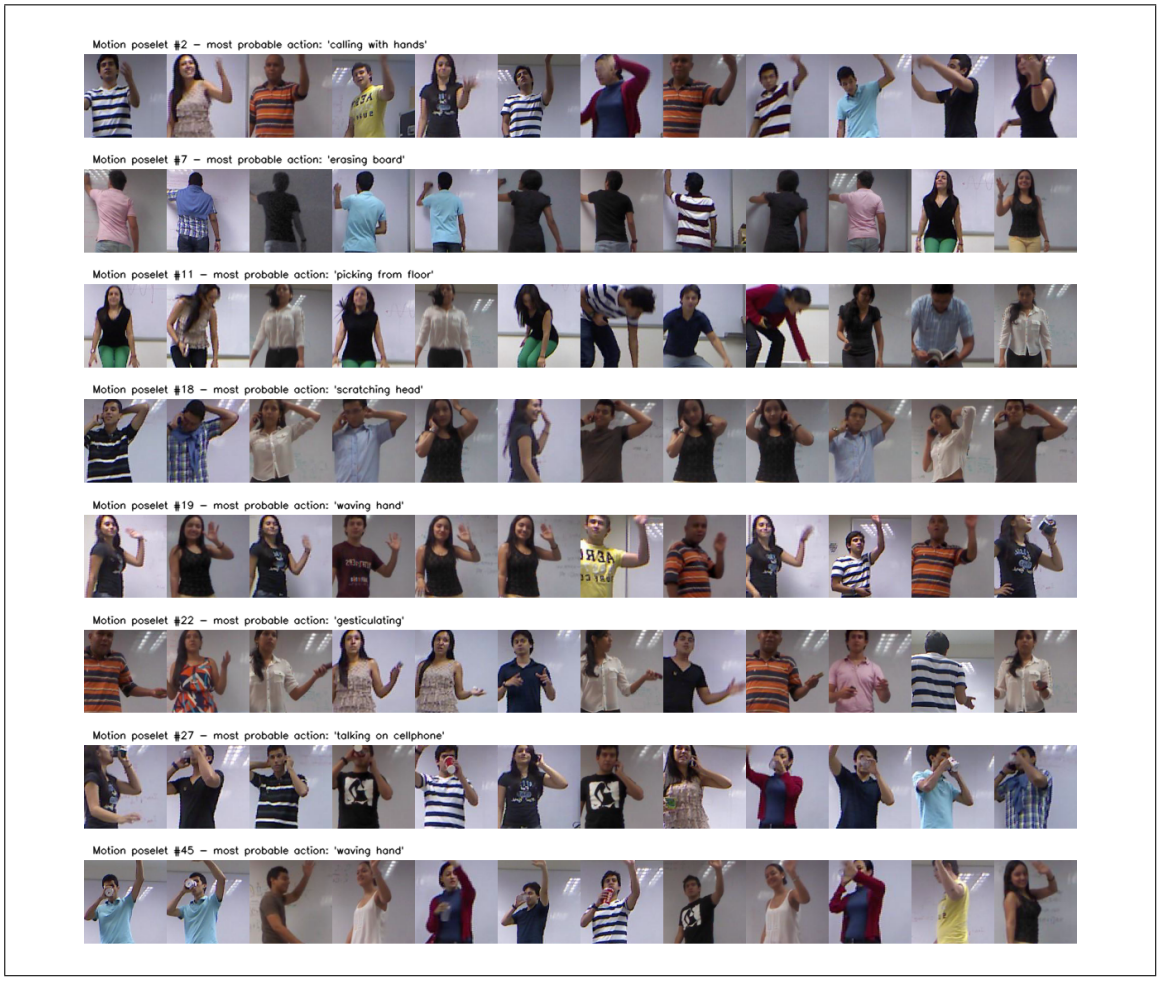

Figure 6.15. Motion poselets learned from the Composable Activities dataset.

likely underlying atomic action for each moving poselet. Figure 6.16 presents a similar analysis for moving poselets learned in the MSR-Action3D dataset.

We also visualize the action annotations produced by our model. Figure 6.17 (top) shows the action labels associated with each body part in a video from the Composable Activities dataset. Figure 6.17 (bottom) illustrates per-body part action annotations for a video in the Concurrent Actions dataset. These examples illustrate the capabilities of our model to correctly annotate the body parts that are involved in the execution of each action, in spite of not having that information during training.

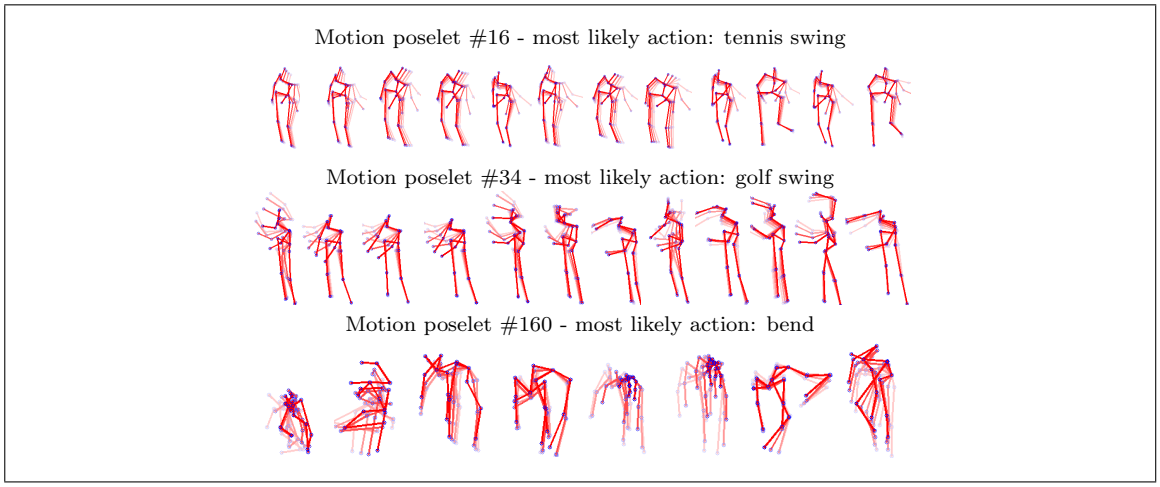

Figure 6.16. Motion poselets learned from the MSR-Action3D dataset.

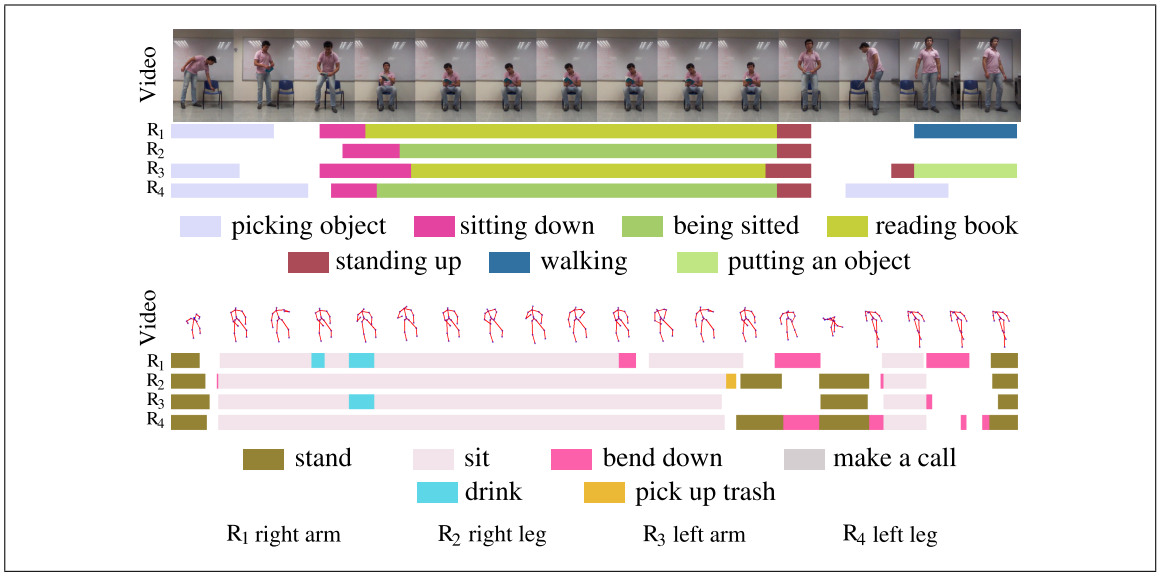

Figure 6.17. Automatic spatio-temporal annotation of atomic actions. Our method detects the temporal span and spatial body regions that are involved in the performance of atomic actions in videos.

Chapter 7. CONCLUSIONS AND FUTURE WORK

In this thesis, we incrementally build a new method for visual recognition of actions and activities using RGB-D data, proposing a novel hierarchical compositional model. As its main strengths, the proposed method is able to jointly learn suitable representations at different abstraction levels leading to compact and robust models. In particular, our model achieves powerful multi-class discrimination while providing useful annotations at the intermediate semantic level. The compositional capabilities of our model also provides robustness to partial body occlusions.

We presented three versions of our model. We started presenting the *core model*, where the main hierarchical and compositional elements were put together to provide a simple but powerful method to recognize complex activities, increasing the recognition performance when compared to similar methods. The experiments showed that using a hierarchical compositional model using poses as latent variables provide pose and action representations that are not achievable with similar methods that focus only on classification with pre-built pose representations. The proposed model outputs the complex activity label as well as full annotations for atomic actions in every frame and region of the body, therefore spatio-temporal action localization is a byproduct of the proposed model.

The second version presented, the *sparse model*, shows how a fused descriptor, composed of geometric features and motion descriptors, improves the accuracy of the activity prediction by over 5% compared to the geometric descriptor alone in the Composable Activities dataset. Moreover, the model makes an efficient use of learned body poses by imposing sparsity over coefficients of the mid-level (atomic action) representation. We observe that this capability produces specialization of poses at the higher levels of the hierarchy. Recognition accuracy using sparsity over mid-level classifiers is similar to the case of using a quadratic regularizer, however, the semantic interpretation of the output of the model is improved, since interactions of pose classifiers with atomic action and activity classifiers are more efficient. The *garbage collector* mechanism introduced in this model also helps in the interpretation of poses, since only frames corresponding to strong poses in terms of activity and action classification remain to be used for the upper levels of the model. For learning the model, we develop an optimization scheme based on a regular quadratic solver applied

to an Elastic-Net objective function using a dual formulation. Under this optimization, our core model can be considered as a Latent Structural SVM, but adding sparsity in the poses with respect to atomic actions.

For the third version of the model, which we call *multimodal model*, we augment the representativity of the data in terms of learning dictionaries of *motions poselets* and *action-lets*. This model demonstrates to be very flexible and informative, capable of handling visual variations and providing spatio-temporal annotations of relevant atomic actions and active body part configurations even when no spatial action information is present. In particular, the model demonstrates to be competitive with respect to state-of-the -art approaches for complex action recognition, while also proving highly valuable additional information. We show how the model keeps perfoming well even when the spatial information of actions were discarded, allowing the model to be applied to a broader set of action recognition benchmarks. We also show how an initialization scheme based on k-means, taht takes also ideas from *self-paced learning*, impacts the final recognition performance and allows the model to be competitive with respect to the version using full spatio-temporal annotations.

There are several research avenues for future work. In particular, during training our model requires annotated data at the level of action, which can be problematic for a large scale application. An improvement could be treating all spatial and temporal action labels as latent variables, and using only a list of possible action labels for every activity. Also, for real-time video recognition, we also need to include inference with respect to the temporal position and span of each activity, which can be also considered as latent variables. Finally, as we mentioned before, our model can be extended to the case of identifying the composition of novel activities that are not present in the training set.

References

Aggarwal, J. K., & Ryoo, M. S. (2011, April). Human activity analysis. *ACM Computing Surveys*, *43*(3), 16:1–16:43.

Aggarwal, J. K., & Xia, L. (2014). Human activity recognition from 3D data: A review. *Pattern Recognition Letters*, *48*, 70 - 80.

Amer, M. R., & Todorovic, S. (2012). Sum-product networks for modeling activities with stochastic structure. In *CVPR* (pp. 1314–1321).

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127.

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3), 257–267.

Bourdev, L., & Malik, J. (2009). Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV* (pp. 1365–1372).

Boureau, Y., Bach, F., LeCun, Y., & Ponce, J. (2010). Learning mid-level features for recognition. In *CVPR* (pp. 2559–2566).

Brendel, W., & Todorovic, S. (2010). Activities as time series of human postures. In *ECCV* (pp. 721–734).

Brendel, W., & Todorovic, S. (2011). Learning spatiotemporal graphs of human activities. In *ICCV* (pp. 778–785).

Cao, C., Zhang, Y., & Lu, H. (2015). Spatio-temporal triangular-chain crf for activity recognition. In *Proceedings of the 23rd annual acm conference on multimedia conference* (pp. 1151–1154).

Castrodad, A., & Sapiro, G. (2012). Sparse modeling of human actions from motion imagery. *International Journal of Computer Vision*, *100*(1), 1–15.

Chaaraoui, A. A., Padilla-López, J. R., & Flórez-Revuelta, F. (2013). Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices. In *IEEE international conference on computer vision workshops* (pp. 91–97).

Chen, C., Zhuang, Y., Nie, F., Yang, Y., Wu, F., & Xiao, J. (2010, December). Learning a 3D Human Pose Distance Metric from Geometric Pose Descriptor. *IEEE Transactions on Visualization and Computer Graphics*, *17*(11), 1676–1689.

Chéron, G., Laptev, I., & Schmid, C. (2015). P-CNN: Pose-based CNN Features for Action Recognition. *ICCV*, 3218–3226.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, eccv* (pp. 1–2).

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior Recognition via Sparse Spatio-Temporal Features. In *VS-PETS workshop at the international conference on computer vision (iccv)* (pp. 65–72).

Du, Y., Wang, W., & Wang, L. (2015). Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. In *Cvpr* (pp. 1110–1118).

Escorcia, V., Davila, M. A., Golparvar-Fard, M., & Niebles, J. C. (2012). Automated vision-based recognition of construction worker actions for building interior construction operations using RGBD cameras. In *Construction research congress* (pp. 879–888).

Felzenszwalb, P., Mcallester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *CVPR* (pp. 1–8).

Feng, X., & Perona, P. (2002). Human action recognition by sequence of movelet codewords. In *3dpvt* (Vol. 16, pp. 717–721). IEEE.

Gaidon, A., Harchaoui, Z., & Schmid, C. (2013, March). Temporal Localization of Actions with Actoms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(11), 2782–2795.

Hu, J.-F., Zheng, W.-S., Lai, J., & Zhang, J. (2015). Jointly learning heterogeneous features for RGB-D activity recognition. In *Cvpr* (pp. 5344–5352). doi: 10.1109/CVPR.2015.7299172

Hu, N., Englebienne, G., Lou, Z., & Krose, B. (2014). Learning latent structure for activity recognition. In *Icra.* doi: 10.1109/ICRA.2014.6906983

Ikizler, N., & Forsyth, D. A. (2008). Searching for Complex Human Activities with No Visual Examples. *International Journal of Computer Vision*, *80*(3), 337–357.

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., & Black, M. J. (2013). Towards understanding action recognition. In *Iccv* (pp. 3192–3199). doi: 10.1109/ICCV.2013.396

Joachims, T., Finley, T., & Yu, C. (2009). Cutting-plane training of structural SVMs. *Machine Learning*, *77*(1), 27–59.

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

Jurie, F., & Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *ICCV* (pp. 604–610).

Kong, Y., & Fu, Y. (2015). Bilinear heterogeneous information machine for RGB-D action recognition. In *Cvpr* (pp. 1054–1062). doi: 10.1109/CVPR.2015.7298708

Koppula, H., Gupta, R., & Saxena, A. (2013). Learning human activities and object affordances from RGB-D videos. *International Journal of Robotics Research*, *32*(8), 951-970.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS* (pp. 1097–1105).

Kumar, M. P., Packer, B., & Koller, D. (2010). Self-paced learning for latent variable models. In *Nips* (p. 1189-1197).

Laptev, I. (2005). On Space-Time Interest Points. *International Journal of Computer Vision*, *64*(2-3), 107–123.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Cvpr* (pp. 1–8).

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR* (pp. 2168–2178).

Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3D points. In *CVPR* (pp. 9–14).

Lillo, I., Niebles, J. C., & Soto, A. (2016). A hierarchical pose-based approach to complex action understanding using dictionaries of actionlets and motion poselets. In *CVPR.*

Lillo, I., Niebles, J. C., & Soto, A. (2017). Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos. *Image and Vision Computing*, *59*(March), 63-75.

Lillo, I., Soto, A., & Niebles, J. C. (2014). Discriminative hierarchical modeling of spatio-temporally composable human activities. In *CVPR* (pp. 812–819).

Lu, C., & Tang, C.-k. (2014). Range-Sample Depth Feature for Action Recognition. In *CVPR* (pp. 772–779).

Luo, J., Wang, W., & Qi, H. (2013). Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *ICCV* (pp. 1809–1816).

Luo, J., Wang, W., & Qi, H. (2014). Spatio-temporal feature extraction and representation for RGB-D human action recognition. *Pattern Recognition Letters*, *50*(1), 139-148.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., & Zisserman, A. (2008). Discriminative learned dictionaries for local image analysis. In *CVPR* (pp. 1–8).

Maji, S., Bourdev, L., & Malik, J. (2011). Action recognition from a distributed representation of pose and appearance. In *Cvpr* (p. 3177-3184).

Microsoft. (2012). *Kinect for Windows SDK.*

Nie, B. X., Xiong, C., & Zhu, S.-c. (2015). Joint action recognition and pose estimation from video. In *Cvpr* (pp. 1293–1301). doi: 10.1109/CVPR.2015.7298734

Niebles, J. C., Chen, C.-W., & Fei-Fei, L. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV* (pp. 392–405).

Oreifej, O., & Liu, Z. (2013). HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR* (p. 716-723).

Ramanan, D., & Forsyth, D. A. (2003). Automatic annotation of everyday movements. In *NIPS.*

Raptis, M., Kokkinos, I., & Soatto, S. (2012). Discovering discriminative action parts from mid-level video representations. *CVPR*, 1242–1249. doi: 10.1109/CVPR.2012.6247807

Raptis, M., & Sigal, L. (2013). Poselet key-framing: A model for human activity recognition. In *CVPR* (pp. 2650–2657).

Savarese, S., Winn, J., & Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons. In *CVPR* (pp. 2033–2040).

Shahroudy, A., Wang, G., & Ng, T.-T. (2014). Multi-modal feature fusion for action recognition in rgb-d sequences. In *Isccsp* (pp. 1–4).

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., ... Blake, A. (2011a). Real-time human pose recognition in parts from a single depth image. In *Communications of the acm* (pp. 116–124).

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., ... Blake, A. (2011b). Real-time human pose recognition in parts from a single depth image. In *Communications of the ACM* (pp. 116–124).

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2012). Unstructured human activity detection from RGBD images. In *ICRA* (pp. 842–849).

Tao, L., & Vidal, R. (2015). Moving Poselets : A Discriminative and Interpretable Skeletal Motion Representation for Action Recognition. In *IEEE international conference on computer vision workshops* (pp. 61–69).

Thurau, C., & Hlavac, V. (2008). Pose primitive based human action recognition in videos or still images. In *CVPR* (pp. 1–8).

Tsochantaridis, I., Hofmann, T., Joachims, T., & Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. In *ICML* (pp. 104–111).

Vemulapalli, R., Arrate, F., & Chellappa, R. (2014a). Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In *Cvpr* (pp. 588–595). doi: 10.1109/CVPR.2014.82

Vemulapalli, R., Arrate, F., & Chellappa, R. (2014b). Human action recognition by representing 3D skeletons as points in a Lsmie group. In *CVPR* (pp. 588–595).

Vishwakarma, S., & Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, *29*(10), 983–1009.

Wan, J., Ruan, Q., Li, W., An, G., & Zhao, R. (2014). 3D SMoSIFT: three-dimensional sparse motion scale invariant feature transform for activity recognition from RGB-D videos. *Journal of Electronic Imaging*, *23*(2), 023017-023017.

Wang, C., Wang, Y., & Yuille, A. L. (2013). An approach to pose-based action recognition. In *CVPR* (pp. 915–922).

Wang, H., Klaser, A., Schmid, C., & Liu, C.-L. (2011). Action recognition by dense trajectories. In *CVPR* (pp. 3169–3176).

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *CVPR* (pp. 1290–1297).

Wang, L., Qiao, Y., & Tang, X. (2014). Video Action Detection with Relational Dynamic-Poselets. In *Eccv* (pp. 565–580). doi: 10.1007/978-3-319-10602-1_37

Wang, Y., & Mori, G. (2008). Learning a discriminative hidden part model for human action recognition. In *NIPS* (pp. 1721–1728).

Wang, Y., & Mori, G. (2011). Hidden part models for human action recognition: Probabilistic versus max margin. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(7), 1310-1323.

Wei, P., Zheng, N., Zhao, Y., & Zhu, S.-C. (2013). Concurrent action detection with structural prediction. In *ICCV* (pp. 3136–3143).

Weinland, D., Ronfard, R., & Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Journal of Computer Vision and Image Understanding*, *115*(2), 224–241.

Xia, L., Chen, C.-C., & Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3D joints. In *IEEE conference on computer vision and pattern recognition workshops* (pp. 20–27).

Yang, J., Yu, K., Gong, Y., & Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *CVPR* (pp. 1794–1801).

Yang, X., & Tian, Y. (2014). Super Normal Vector for Activity Recognition Using Depth Sequences. In *Cvpr* (pp. 804–811).

Yao, B., & Fei-Fei, L. (2010). Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. In *CVPR* (pp. 17–24). IEEE.

Yu, C., & Joachims, T. (2009a). Learning structural SVMs with latent variables. In *ICML* (pp. 1169–1176).

Yu, C., & Joachims, T. (2009b). Learning structural svms with latent variables. In *Icml* (p. 1169-1176).

Yuille, A., & Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation*, *15*(4), 915-936.

Zanfir, M., Leordeanu, M., & Sminchisescu, C. (2013). The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection. In *Iccv* (pp. 2752–2759). doi: 10.1109/ICCV.2013.342

Zhu, Y., Chen, W., & Guo, G. (2013). Fusing spatiotemporal features and joints for 3D action recognition. In *IEEE conference on computer vision and pattern recognition workshops* (pp. 486–491).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.