TESIS DE LA UNIVERSIDAD
DE ZARAGOZA

2016

219

Mario Félix Rodríguez Martínez

The Understanding of Human Activities by Computer Vision Techniques

Departamento

Instituto de Investigación en Ingeniería [I3A]

Director/es

Orrite Uruñuela, carlos

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

ISSN 2254-7606

Tesis Doctoral

THE UNDERSTANDING OF HUMAN ACTIVITIES BY COMPUTER VISION TECHNIQUES

Autor

Mario Félix Rodríguez Martínez

Director/es

Orrite Uruñuela, carlos

UNIVERSIDAD DE ZARAGOZA

Instituto de Investigación en Ingeniería [I3A]

2016

Universidad Zaragoza

1542

Doctoral Thesis

The Understanding of Human Activities by Computer Vision Techniques

Author:

Mario F. Rodríguez Martínez

Supervisor:

Carlos Orrite Uruñuela

in the

Computer Vision Laboratory

Instituto de Investigación en Ingeniearía de Aragón, I3A

“Look Dave, I can see you’re really upset about this. I honestly think you ought to sit down calmly, take a stress pill, and think things over.”

HAL9000, *2001: A Space Odyssey*

Abstract

This thesis provides some novel frameworks for learning human activities and for further classifying them into categories. This field of research has been largely studied by the computer vision community however there are still many drawbacks to solve.

First, we have found few proposals in the literature for learning human activities from limited number of sequences. However, this learning is critical in several scenarios. For instance, in the initial stage after a system installation the capture of activity examples is time expensive and therefore, the learning with limited examples may accelerate the operational launch of the system. Moreover, examples for training abnormal behaviour are hardly obtainable and their learning may benefit from the same techniques. This problem is solved by some approaches, such as cross domain implementations or the use of invariant features, but they do not consider the specific scenario information which is useful for reducing the clutter and improving the results. Systems trained with scarce information face two main problems: on the one hand, the training process may suffer from numerical instabilities while estimating the model parameters; on the other hand, the model lacks of representative information coming from a diverse set of activity classes. We have dealt with these problems providing some novel approaches for learning human activities from one example, what is called a one-shot learning method. To do so, we have proposed generative approaches based on Hidden Markov Models as we need to learn each activity class from only one example. In addition, we have transferred information from external sources in order to introduce diverse information into the model. This thesis explains our proposals and shows how these methods achieve state-of-the-art results in three public datasets.

Second, we have studied the recognition of human activities in unconstrained scenarios. In this case, the scenario may or may not be repeated in training and evaluation and therefore the clutter reduction previously mentioned does not happen. On the other hand, we can use any labelled video for training the system independently of the target scenario. This freedom allows the extraction of videos from the Internet dismissing the implicit constrains when training with limited examples. Having plenty of training examples both, generative and discriminative, methods can be used and by the time this thesis has been made the state-of-the-art has been achieved by discriminative ones. However, most of the methods usually fail when taking into consideration long-term information of the activities. This information is critical when comparing activities where the order of sub-actions is important, and may be useful in other comparisons as well. Thus, we have designed a framework that incorporates this information in a discriminative classifier. In addition, this method introduces some flexibility for sequence alignment, useful feature when the activity segmentation is not exact. Using this framework we have obtained state-of-the-art results in four challenging public datasets with unconstrained scenarios.

Resumen

Esta tesis propone nuevas metodologías para el aprendizaje de actividades humanas y su clasificación en categorías. Aunque este tema ha sido ampliamente estudiado por la comunidad investigadora en visión por computador, aún encontramos importantes dificultades por resolver.

En primer lugar hemos encontrado que la literatura sobre técnicas de visión por computador para el aprendizaje de actividades humanas empleando pocas secuencias de entrenamiento es escasa y además presenta resultados pobres. Sin embargo, este aprendizaje es una herramienta crucial en varios escenarios. Por ejemplo, un sistema de reconocimiento recién desplegado necesita mucho tiempo para adquirir nuevas secuencias de entrenamiento así que el entrenamiento con pocos ejemplos puede acelerar la puesta en funcionamiento. También la detección de comportamientos anómalos, ejemplos de los cuales son difíciles de obtener, puede beneficiarse de estas técnicas. Existen soluciones mediante técnicas de cruce dominios o empleando características invariantes, sin embargo estas soluciones omiten información del escenario objetivo la cual reduce el ruido en el sistema mejorando los resultados cuando se tiene en cuenta y ejemplos de actividades anómalas siguen siendo difíciles de obtener. Estos sistemas entrenados con poca información se enfrentan a dos problemas principales: por una parte el sistema de entrenamiento puede sufrir de inestabilidades numéricas en la estimación de los parámetros del modelo, por otra, existe una falta de información representativa proveniente de actividades diversas. Nos hemos enfrentado a estos problemas proponiendo novedosos métodos para el aprendizaje de actividades humanas usando tan solo un ejemplo, lo que se denomina *one-shot learning*. Nuestras propuestas se basan en sistemas generativos, derivadas de los Modelos Ocultos de Markov, puesto que cada clase de actividad debe ser aprendida con tan solo un ejemplo. Además, hemos ampliado la diversidad de información en los modelos aplicado una transferencia de información desde fuentes externas al escenario. En esta tesis se explican varias propuestas y se muestra como con ellas hemos conseguidos resultados en el estado del arte en tres bases de datos públicas.

La segunda dificultad a la que nos hemos enfrentado es el reconocimiento de actividades sin restricciones en el escenario. En este caso no tiene por qué coincidir el escenario de entrenamiento y el de evaluación por lo que la reducción de ruido anteriormente expuesta no es aplicable. Esto supone que se pueda emplear cualquier ejemplo etiquetado para entrenamiento independientemente del escenario de origen. Esta libertad nos permite extraer vídeos desde cualquier fuente evitando la restricción en el número de ejemplos de entrenamiento. Teniendo suficientes ejemplos de entrenamiento tanto métodos generativos como discriminativos pueden ser empleados. En el momento de realización de esta tesis encontramos que el estado del arte obtiene los mejores resultados empleando métodos discriminativos, sin embargo, la mayoría de propuestas no suelen considerar la información temporal a largo plazo de las actividades. Esta información puede ser crucial para distinguir entre actividades donde el orden de sub-acciones es determinante, y puede ser una ayuda en otras situaciones. Para ello hemos diseñado un sistema que incluye dicha información en una Máquina de Vectores de Soporte. Además, el sistema permite cierta flexibilidad en la alineación de las secuencias a comparar, característica muy útil si la segmentación de las actividades no es perfecta. Utilizando este sistema hemos obtenido resultados en el estado del arte para cuatro bases de datos complejas sin restricciones en los escenarios.

Acknowledgements

I would like to take this opportunity to thank everyone who supported me through the followed path until reaching this thesis. All of you are part of it.

First of all, I want to mention my supervisor, Carlos Orrite, who has given me trustworthy guidance and thanks to him I have been able to complete the work. We have kept several discussions and rethinks of the work which have led to this thesis. Also, I would like to express my gratitude to Carlos M., who, from the distance, has been able to remember us the importance of a mathematical base. This thesis would not be completed without my stay in Kingston University, and there, my supervisor Dimitrios has contributed with a different point of view of the dealt issues. Finally, I want to remember the different colleagues in the CV_ Lab that have accompanied me during these years and who have been essential in this project, especially Miguel, without whom I would be still struggling with some code. Also people from DIRC in Kingston University made easier my stay there.

Of course, I would like to thank my family. They have supported me for reaching this goal in better and worse moments and they have been always there for what I have needed. To my friends from Logroño and Zaragoza, and any place they are, thank you all because your company have been very important for me.

This thesis was partially supported by the Spanish Government (TIN2013-45312-R MINECO and BES-2011-043752 FPI Grant), Gobierno de Aragon and the European Social Found.

Contents

List of Figures

List of Tables

1 Introduction

"Grandpa! What are you looking for? The keys? You are already holding them!"

In many situations it is not difficult for us, human beings, to guess what other people are doing and help them or anticipate to their moves. And not only that, but we are also able to guess with more or less success the mood of people around us. Thanks to these abilities many complex interrelation among people are achieved. On the other hand, remembering the initial quote of this thesis from the Sci-Fi movie *2001: A Space Odyssey*, where HAL9000, a computer, guesses the mood of Dave, a human astronaut, we can realize how this human-computer interrelation continues to be Sci-Fi. Nevertheless, the objective for a human-computer interrelation is different from the humans case and accomplishing the human behaviour understanding by computational techniques may lead to some useful interactions. However, despite the great advances accomplished in the last years by the machine learning community, this objective is still Sci-Fi in most environments.

In human behaviour understanding a key objective is the recognition of what the humans are doing, which is called the activity. In the literature there is not a consensus with the term activity and we can find the terms activity, action and event used interchangeably. Sometimes the use of these terms is defined by their level of granularity defining actions as the small meaningful movements that are indivisible whereas activities are defined by several concatenated actions. Other authors use the terms actions and atomic actions instead. An action, or atomic action, can be viewed as the basic activity composed by only one indivisible meaningful movement. This thesis proposes general purpose methods for the recognition of both, complex and basic activities (actions), therefore the term activity, which includes both concepts, is preferred. On the other hand, event is a more general concept that involves any change in the scenario related or not to humans.

Although an ambitious objective to accomplish would be the recognition of not only the activity but also the mood and intention of the subject, as described in the first paragraph, this thesis has been focused on some unresolved scenarios for activity recognition. Namely, the automatic recognition of activities based only in roughly acquired movement which still presents several difficulties and their solution would ease some human-computer interactions. Specifically, we have studied two fields: the learning of activities from limited training examples and the inclusion of the whole activity temporal information in a recognition system.

The ubiquity of sensors, mainly video cameras, and the increase of computational capability in the daily used devices are booming the research in applications involving the automatic understanding of human behaviour. In fact, really simple systems has been around for a while, for instance motion detectors present in automatic flushers, taps, hand-dryers, doors, lighting, in burglar alarms, etc. These systems guess the activity of the subject based only on the probability that someone present in the specific place is doing what it is supposed, and then the system responds accordingly. We have to go back to 1953 to find out what can be considered the first modern motion detector, when in a patent for a burglar alarm issued to Samuel Bagno in New York [Bagno, 1953]

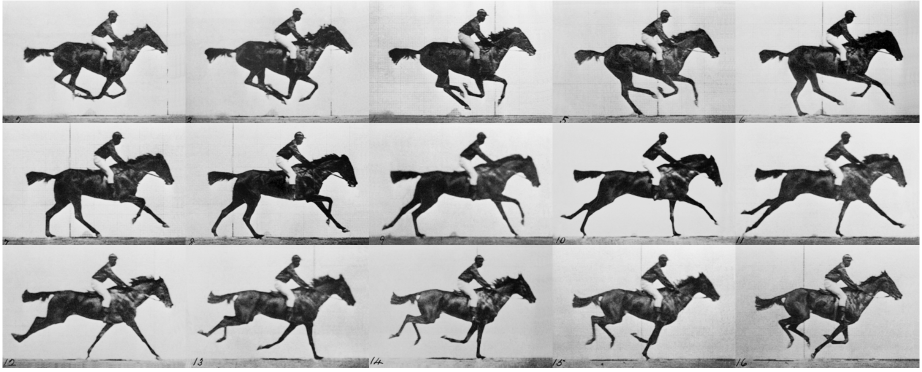


Figure 1.1: Pictures of a horse galloping captured in the study *"Sallie Gardner at a Gallop"* by Eadweard Muybridge.

was included the design of a presence detector operated by ultrasonic sound. But activity recognition has evolved since then, and the need for recognizing complex activities has led to the use of sensors that provide more information than the simple presence. Here comes handy the use of video cameras which nowadays capture a lot of information being relatively cheap. These characteristics have probably lead them to be the most used and studied sensors in activity recognition. Moreover, the video has been used to study the movement from its very beginning. On June 15, 1878, Eadweard Muybridge (1830 – 1904, Kingston Upon Thames), using 24 cameras in a row parallel to a horse's path, captured a sequence of pictures of a horse galloping at 58 km/h, published in his famous work *"Sallie Gardner at a Gallop"*[1] from which some frames are shown in Figure 1.1. The shutter of each camera was triggered by a thread as the horse passed, obtaining what is considered the first film. Former governor of California, Leland Stanford, hired Muybridge for finding a definitive answer

[1] I am happy to mention this work not only because of being the origin of the video, clearly related to this thesis, but also because I spent three months of my thesis research period in the University of Kingston Upon Thames, birth and death place of Muybridge.

to the question whether all four feet of a horse were off the ground at some point during the gallop, and thanks to the study they gave a scientific answer to that popularly debated question. In 1879, he also invented what is considered a precursor to cinematography, an early moving picture projector called the Zoopraxiscope.

Since the first video and the first activity recognition system were created, both fields have experienced a great evolution and merged in several applications. Nowadays video cameras are ubiquitous, from mobile devices and webcams to closed-circuit television (CCTV) for surveillance or television station cameras in crowded events. Thousands of videos are continuously recorded and many of them uploaded to the Internet. As an example, the most famous content platform, Youtube, states that 300 hours of video are uploaded to their servers every minute and hundreds of millions of hours are watched every day, and this is only one platform of many, and numbers are growing. This huge amount of data and people interacting makes it logical the investment in understanding the overwhelming information in an "intelligent" manner, helping people to find the meaningful content they are looking for. In most of the cases the search involves some kind of event, usually produced by humans, forming the research field of recognizing human activities in any kind of footage. Moreover, the ubiquity of cameras allows the use of activity recognition for on-line human-machine interaction in smart environments.

The search of meaningful clips in a footage where some activity is performed has led to many applications, facilitating the human-machine interaction and making some tedious tasks easier. Improvements in the training with limited data and the better inclusion of temporal information, pursued in this thesis, could be applied in some of these applications like the ones explained below.

A direct use of the video based activity recognition is indexing video clips which has a direct application in search engines. Current search engines are mainly based in tags written by hand, which implies a tedious work. Therefore, the tagging process usually relies on users comments from where the system extracts the meaningful information, as machine learning techniques for text documents are usually trustworthy. However, with this process, lots of videos are kept unlabelled or mislabelled. On the other hand, there are already systems that look for specific image classes, labelling them automatically with an interesting degree of accuracy. Looking for the video clips where certain activity is performed, however, involves one more dimension, time, complicating the system in two manners: first, the computational load is increased and then more time or more powerful computers are needed, and second, the time dimension has different characteristics to spatial dimensions thus, the systems should include complexer algorithms in order to deal with both space and time.

Surveillance is another field where the activity recognition is a key tool. Currently, most of the surveillance videos are never seen and they are only checked after the report of an important event. However, if a system is able to automatically discover when something important is happening, it would be possible to prevent several undesired behaviours as robberies, fights and so on, having a quick response. At the moment, most of the so called “smart” cameras incorporate simple systems based mainly in presence detectors but, slowly, more complex algorithms are being introduced which are trying to identify aspects of the subjects, for instance behaviour, object carrying, etc. Being located in fixed positions, the recording from a usual surveillance camera can be used to learn scenario based information reducing the clutter for the activity recognition. On the other hand, the so called important events rarely happen and therefore they are difficult to train in a specific scenario as a system needs to wait until they naturally happen or they have to be acted, loosing part of the natural behaviour.

Continuing with the review of application fields, in gaming industry activity recognition plays an important role as the player and the computer can interact without remote control, increasing the game possibilities. A well known technology is the one used by Microsoft with the Kinect, which is a range camera that increases the conventional rgb images information recorded by a conventional camera with a depth map, improving recognition in constrained environments.

This technology is also being used out of the gaming field as long as in a constrained environment.

Leaving the best for last, in my personal opinion, Ambient Assisted Living (AAL) is one of the most promising fields where human activity recognition can be a pillar. Thanks to medicine advances the life expectancy is increasing and so the number of dependant people, including not only elderly but also sick and disable ones. Furthermore, the total fertility rate (TFR) is low in many developed countries, as Spain for instance. This data can be validated in most of the statistics services, the National Statistic Institute of Spain (INE)INE [2014]. The increase of dependent people and the, at least, stagnation of the TFR makes it impossible to rely on younger generations for future assistance of the dependent population, and this quite certain future makes even more important the development of AAL. Even if it were possible to rely on human assistants, people would usually prefer to be independent as much as possible. Therefore, the help of automated assistance will improve the quality of life of many people. This automated assistance may benefit from the understanding of the subject behaviour and therefore the recognition of the activity being performed. As explained in surveillance systems, the constraints of the scenario present in an AAL system suppose an advantage for cluttering reduction but, on the other hand, the need of training examples in the new scenario restricts the suitability of the system in initial stages after its installation. Many times, the system should be able to detect unusual events, as for instance an erratic behaviour of the subject that might signal some kind of health alarm, and these unusual events are difficult to obtain due to their rare frequency and even more to obtain enough examples for a trustworthy training of the activity. It is worth noting that when developing AAL systems concerns about privacy arise, especially when working with video recording. The overcome of the different issues affecting an AAL system makes the field still immature but promising.

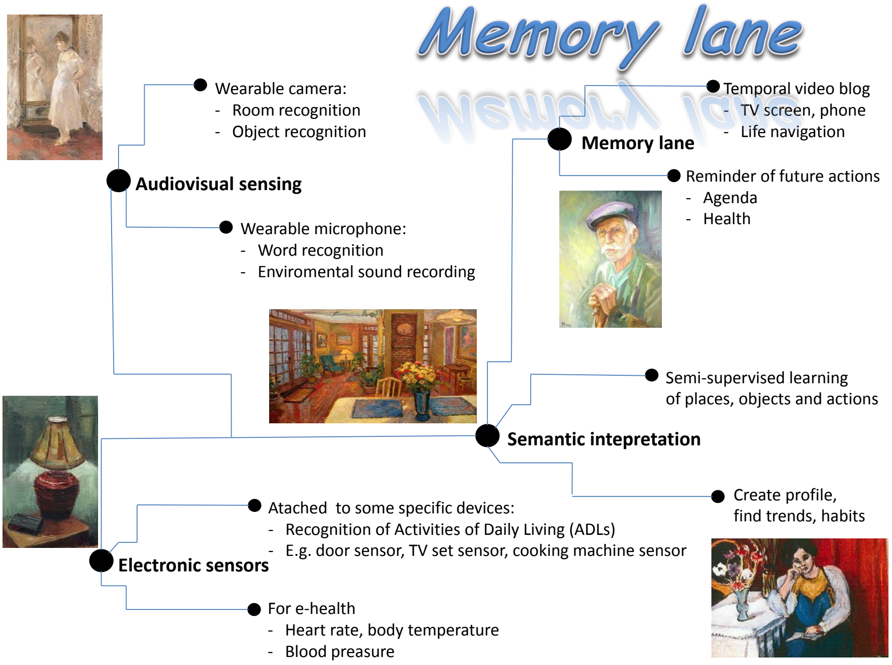


Figure 1.2: Diagram of Memory Lane project.

1.1 Context

This thesis has been made in the Computer Vision Laboratory (CVLab) of the Aragón Institute for Engineering Research (I3A) in the University of Zaragoza. The group has an extensive trajectory in computer vision techniques with one of its branches focused in human behaviour understanding.

The origin of the thesis dates back to a proposal made in the project (*Entorno para el seguimiento de personas y análisis de trayectorias encaminado a la comprensión del comportamiento social*) that could roughly be translated as (Human tracking and trajectory analysis framework focused on the social behaviour understanding). Moreover, previous know-how facilitated the evolution to an AAL project named Memory Lane which integrates different technologies in an intelligent environment to support the independent living of users. These two projects have made the environment where the thesis has been developed. Specifically, Memory Lane has supposed the base project for the thesis and, as we can see in its global diagram depicted in Figure 1.2, the semantic interpretation of input information is a key stage. A system that captures information along time, as the proposed in Memory Lane, has the interpretation of sequential data as an essential feature. In this regard, after identifying some drawbacks in the proposed methods for human activity recognition existing in the literature, we determined the following goals for the thesis.

1.2 Goals

This thesis covers two problems existing in the activity recognition field and looks for some solutions to them. First, the unstable training produced with a scarce number of available sequences, which happens for instance when a system is deployed in a new scenario. The second addressed problem is the poor long-term information representation of complex activities in state-of-the-art methods which may be essential in some cases, for instance, if the order of atomic actions determines the activity class.

Imagine a house where a *smart home* system for AAL is going to be installed, or a warehouse where the owners want to install a new surveillance system. This kind of systems usually includes fixed cameras which will always record from the same location avoiding many of the possible variabilities, so the systems can learn from the scenario reducing the clutter and improving the results in recognition tasks. However, just after the installation there is not a single recorded sequence of any activity in the scenario and it is expensive, at least in time, to capture several sequences for training the system. Moreover, many times, the desired recognizable activities are unusual ones with a low frequency

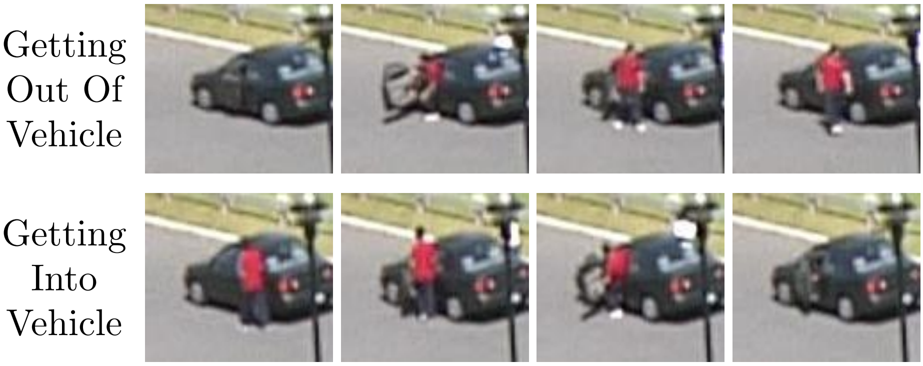


Figure 1.3: Two activities from Virat Release 2.0 dataset with contradictory order. First row: person opens car door, goes out of the vehicle, closes door and walks. Second row: person walks, opens car door, gets into the vehicle and closes door.

of happening and their recordings are hard to obtain. Most of the current reliable methods use lots of examples in the training process, being unusable with few training sequences, which delays the proper running of the system. The first goal of the thesis is to find a reliable recognition system using as few training sequences as possible. We consider two restricted scenarios: one where only one sequence per class is available and other even more restricted where just a single sequence from any class is used for training. These configurations are called one-shot learning methods.

On the other hand, despite having lots of training examples in unconstrained scenarios, most of the state-of-the-art methods in activity recognition make a poor representation of long-term information or just discard it, failing to recognize some complex activities. Long-term information is defined as the temporal structure of the whole activity as opposed to short-term information which only includes data from few frames being shorter than atomic actions and mid-term information which cover a length typical of atomic actions. The second goal of the thesis is the design of an activity recognition method that suitably incorporates the long-term information. This method should be able to keep at least the recognition rates of state-of-the-art methods in simple activity classes but at the same time improving the results in complex activities, which overall means an improvement of the system. As an example of complex activities, Figure 1.3 shows two activity classes from the Virat Release 2.0 dataset where the order of actions is the most determinant information to distinguish between them, deciding if the person is getting out of the vehicle or is getting into the vehicle.

1.3 Contributions

As stated before, two problems present in some activity recognition scenarios are covered: scarce data in training and complex activities where basic actions order determines the class. In both cases a pre-segmentation of the activity clips is assumed. In activity recognition, the classical application of a recognition system implies a previous segmentation of the footage where the activity happens, this stage is out of the scope of the thesis so the reader should check the literature if it is interested in it, for instance a good starting point is the review found in [Weinland et al., 2011]. As summary we divide the contributions of the thesis into two parts:

One-Shot Learning. One of the main goals of the thesis consists in training the recognition system with few examples, even as few as one. However, the lack of information may suppose the infeasibility of training the different parameters involved in the classifier. Therefore, we design a Transfer Learning process which enriches the available information using data from different sources and not only from the target scenario. Moreover, a system trained with only one sequence is not a classifier but an activity model obtained through a generative method. Thus, we propose the use of the generative method Hidden Markov

Model (HMM) [Rabiner, 1989]. HMM trains unstable models when operating with scarce data but our modified proposals overcome this problem. We have used two HMM modifications so that it works with global and local features. A global feature represents the information of the whole activity sequence in a single descriptor whereas several descriptors are obtained in the neighbourhood of some selected voxels of the activity video for local features. To work with global features we use a modification of the classical discrete HMM designed with a soft-assignment and avoiding the problems derived from the lack of training data. We use the approach Fuzzy Discrete Hidden Markov Model (FDHMM) proposed by [Uguz et al., 2008] applied to these observations. A more reliable method is developed from the HMM in order to use local spatio-temporal features such as Improved Dense Trajectories (IDT) introduced by [Wang and Schmid, 2013]. We propose to encode the video into a sequence of normalized Bag of Features (BoF) obtained from sliding frame-windows, with each sequence element belonging to a unit Simplex, $\Delta = \{\mathbf{v} \in \mathbb{R}^K : v_k \geq 0 : \sum_{k=1}^{K} v_k = 1\}$. The constraints of the Simplex allow us to design a method that produces stable models with scarce data. We show the improvement of our approaches applying them into three public and well known datasets: Weizmman, KTH and IXMAS. These datasets contain videos captured from fixed viewpoint cameras and therefore are suitable for the validation of the methods. The approaches description and the obtained results are detailed in chapters 4 and 5.

Long-term Information Inclusion. For the second main goal of the thesis, the recognition of complex activities where long-term information might be determinant, we have designed a Time Flexible Kernel. Thanks to this kernel formulation, the use of a Support Vector Machine (SVM) in the classification process without discarding the long-term temporal information is possible. This kernel improves the methods accuracy in two scenarios: when the descriptors are local spatio-temporal features with short temporal information and when the complexity of the activities produces errors if they depend on atomic actions order, even using IDT that includes some medium-term information. Moreover, its combination with the classical SVM approach produces a more reliable system. We apply this method using sequences of Fisher Vectors computed with short-term and mid-term features extracted from sliding frame-windows. The evaluation in the four large unconstrained datasets HMDB51, UCF50, Olympic-Sports and Virat Release 2.0, shows state-of-the-art-results. The method and the experimentation are explained in Chapter 6.

The thesis is completed with a comprehensive review of related works in Chapter 2, a description of the proposed feature extraction in Chapter 3 and a discussion of the contributions as well as a look into future work in Chapter 7.

2 Related Work

The automatic understanding of human activities through machine learning techniques has taken a great attention of the research community, producing an extensive literature. In this chapter we go across a comprehensive review of previous works in human activity recognition and specifically in vision-based techniques related to the main objectives of this thesis. First, we describe how a video-based recognition system works and the different issues that affect its performance and later we review how the community has dealt with the activity recognition contributing to different solutions.

2.1 Vision-Based Activity Recognition System

A video is a sequence of frames, when digitalized, can be viewed as a 3-dimensional matrix of pixels, two dimensions provide the space localization and the third one is related to time. A pixel can be represented by a scalar, if the frames are in gray scale, or an n-dimensional vector in other scales. Usually 3-dimensions, as in RGB scale, although more information can be added increasing the dimensionality, as for instance a depth map or an infra-red layer. When displayed on a

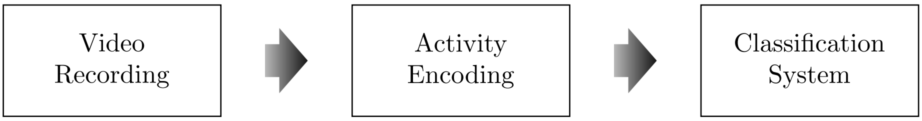


Figure 2.1: Video-based activity recognition system.

screen, as a sequence of images, humans are able to easily distinguish among activities, what denotes that the discriminant information is present in the video, however the same task is extremely challenging for a computational method in the current state of knowledge.

There is not just one way to capture the information of the performed activities using vision computer techniques since the more the technology has evolved the more recording techniques have appeared. The classical approach consists in an external camera, fixed or held by a cameraman, that records the activity. Nowadays, it is possible to add depth information using range imaging techniques, for instance using time-of-flight cameras, stereo-vision, interferometry, structure-light, and so on, however these technologies increase the complexity of the systems, sometimes needing a calibration stage and always being more expensive than the classical ones and therefore, they are less common. On the other hand, an approach that has gained importance recently is the use of wearable cameras where the classical external viewpoint is exchanged by an egocentric recording. A great advantage of this viewpoint is the reduction of concerns about privacy since the subject is not recorded, although it still keeps some issues about privacy, for instance the recording in private rooms or the appearance of interacting people. Additionally, despite the recent interest with wearable cameras there are still few researches and the technology is less developed than the external cameras, keeping some worries about if the images contains the discriminant information for the activities the subject is performing, as sometimes the activities are hardly recognizable even for humans.

In Figure 2.1 a simple diagram of a generic recognition system based on computer vision is depicted. The first stage is the video acquisition, where camera, viewpoint, scenario, etc. are defined. From the different techniques and technologies previously mentioned, this thesis is focused on the external cameras although most of the proposed methods can be exportable to any approach or even outside the video-based activity recognition field. The videos obtained from external cameras have several sources of variability that affect the recognition system performance and should be taken into consideration. In Figure 2.2 we observe a classification of the variability sources:

1. Actor. Activities are performed by subjects with various shapes, clothes, and no activity is exactly repeated in the same way even being performed by the same actor.

2. Scenario. There exist two kinds of approaches depending on the application. In the first case, as for instance in surveillance, the scenario is constrained by using a fixed camera, although a versatile recognition system should allow changes in the background, the illumination and the existence of occlusions. In the second case the recognition is processed in unconstrained scenarios, meaning that the videos have been recorded in multiple locations so there is not a common background or viewpoint.

3. Camera Settings. In the case of unconstrained scenarios, each recording can be obtained with a different camera, so resolution, frame rate, lens, etc. will vary from video to video.

The second stage of the recognition system is the activity encoding. Although the classical name of this stage is feature extraction we refer to it as activity encoding in order to distinguish between the actual extraction of local or global features encoded in specific descriptors and the final encoding of the data suitable for the classifier process, which normally is obtained after a processing of the descriptors. Later in this chapter the description of both stages

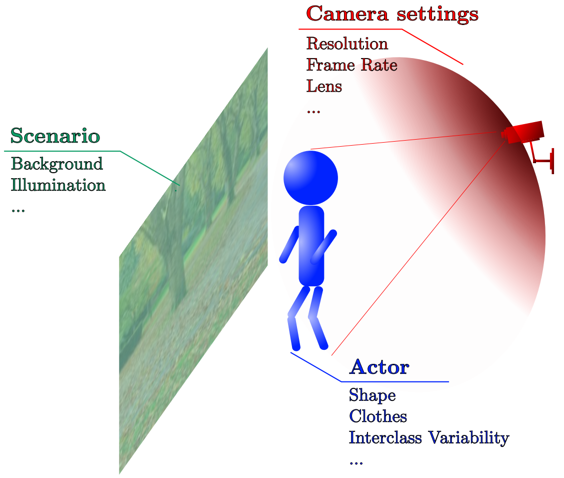


Figure 2.2: Sources of variability in a human action recognition system based on computer vision.

is extended. In the activity encoding design it is essential to keep in mind what kind of information has been captured, the first stage of the process, and how it is going to be classified, the last stage of the recognition system. As we have seen, there are several sources of video recordings and the encoding will vary depending if the videos come from a fixed camera or from a moving one, from a single scenario or from several ones, if the information is in gray scale, RGB or depth map, etc. Moreover, the kind of activities to recognize demarcates the discriminant information that should be kept or the clutter information that should be discarded.

Finally, the last stage is the classification process which depends among others on the goals of the system, the activity encoding, the availability of training data and the computational capacity. Although the three stages are defined separately, they are interdependent and any design should take into account all of them.

The rest of the Chapter is used to review several aspects involved in the activity recognition systems. First, in Section 2.2, we present the different sensors used for capturing information from human activity performances. In Section 2.3 we review how to extract discriminative information in different descriptors and the two main encoding applied to those descriptors. Later, we make a summary of classifiers in Section 2.4. In sections 2.5, 2.6 and 2.7 we review the literature in three aspects related directly with the two goals of the thesis: Transfer Learning, learning with limited examples and long-term information inclusion respectively. Finally, in Section 2.8 we describe the public datasets used for evaluate the different proposals made in the thesis.

2.2 Sensor-Based and Vision-Based Activity Capture

Before the use of any machine learning technique in a human activity recognition system it is necessary to acquire the data suitable for the task. In terms of the type of sensor used to capture the information, systems are generally divided into two groups: sensor-based and vision-based systems.

Sensor-based systems capture information from all kind of sensors but video cameras. An extensive review on sensor-based activity recognition systems is presented in [Chen et al., 2012]. The usual captured data is a time series of some parameter values such as orientation, location, acceleration, pressure, temperature, etc., which is advantageous over video recordings since the anonymity of the recorded subjects is almost completely preserved, or at least the user concerns are usually vanished. Depending on the location of the sensors it is possible to divide the approaches into two groups, namely using external or wearable sensors. In addition, the subgroup of wearable sensor approaches where recognition is performed with smartphone sensors is enough distinguishable to form its own group due to the relevance they have acquired in recent years. These three configurations are depicted in Figure 2.3.

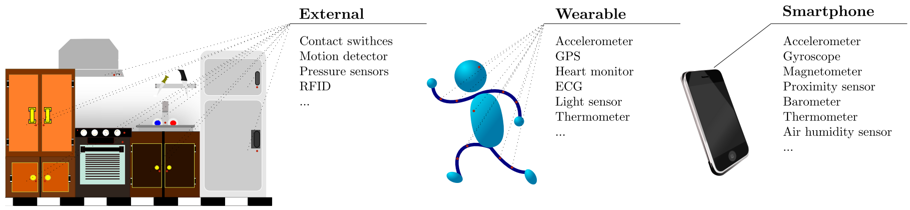


Figure 2.3: Sensor-based activity recognition environments using three configurations: external sensors (in a Smart-Home), wearable sensors and smartphone sensors

Since the introduction of the first external sensors, there has been a technological evolution reducing the size of the sensors, introducing wireless connectivity, increasing the computational capacity, reducing the battery consumption and including a diverse ecosystem of sensors that currently allows an ubiquitous presence of them, forming networks of sensors integrated in networks of processing devices. The easy availability of many classes of sensors makes it possible to attach them to most of the objects in a room creating what is called a *dense-sensing* environment, allowing a deep knowledge of subjects interaction. These sensor frameworks provide an information useful for many smart-home or smart-buildings utilities, [Ding et al., 2011], improving the AAL systems. The technological evolution has also favoured a blooming on wearable sensors. Looking for non uncomfortable and invasive systems the frameworks are evolving from bulky rucksacks with obtrusive sensors attached to small sensors integrated on the clothes and easily wearable devices such as watches or necklaces. These sensors are used in health monitoring, tracking and, of course, in human activity recognition. [Lara and Labrador, 2013] explain a comprehensive review of the human activity recognition systems using wearable sensors. Finally, the amount of sensors included in the current smartphones (gyroscope, accelerometer, barometer, GPS, and so on), and the fact that most of the people carry one constantly, makes it possible to recognize activities through the provided

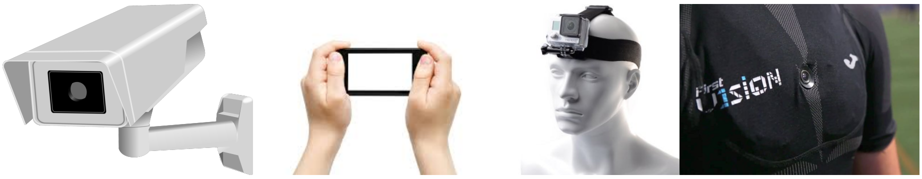


Figure 2.4: Vision-based activity recognition environments using two types of configurations: external cameras being fixed or held by a person, and wearable cameras using a framework or integrated in the clothes. Last two images belong to the commercial devices *GoPro Hero3* and *First V1sion* respectively.

information in a cheap and unobtrusive manner, as recent works gathered in surveys [Shoaib et al., 2015] [Su et al., 2014] confirm.

It is worth noting the audio-based human activity recognition, that although not being as fruitful as vision-based systems, it is easily separable from the other sensor-based methods due to the existence of a large community in audio and speech processing. Using a microphone, or several of them, the recorded audio is discriminative enough of several daily life activities we perform, although many others are indistinguishable. There are some researches pursuing this aim [Stork et al., 2012] but in most of the cases they are used as complementary information [Choudhury et al., 2008] [Kolovou and Maglogiannis, 2010].

From all the sensors available, video cameras are the ones that provide more information attracting the human activity recognition research community for this reason and making the vision-based methods the most studied among all [Turaga et al., 2008] [Poppe, 2010] [Weinland et al., 2011] [Ke et al., 2013]. In a similar process as the one experimented by other sensors, video cameras have evolved by reducing their size, consuming less energy, increasing their resolution and so on, what as well has driven to their ubiquitous presence. This ubiquity makes them present in several environments from where discriminative information of activity performances can be obtained. Attending to the perspective from where the information is captured there are two groups of systems: the ones based in external cameras that are used to record the subjects performing the different activities, and the systems that use egocentric cameras recording the subjects interaction with the environment but without the subject presence in the images. This last group has been less studied as until recently few devices had been developed. A recent comprehensive review has been written by [Nguyen et al., 2016]. In Figure 2.4 the first two images represent external cameras, one from a fixed viewpoint and the other with a moving viewpoint. On the other hand, the last two images represent wearable cameras that provide egocentric viewpoints, showing how these devices are becoming more integrated and less obtrusive. The video cameras have also evolved into more complex devices that provide a greater amount of information. Their use in activity recognition has covered infrared sensors [Han and Bhanu, 2005] and currently a great number of researches have focused in range cameras that give deep information, as the *Kinect* one [Xia et al., 2012] [Cottone et al., 2013]. An important issue with external cameras comes from the loss of privacy and, in order to preserve it, some researches have proposed the use of methods for hiding the subject without loss of discriminative information [Padilla-Lopez et al., 2015].

The great ecosystem of methods for capturing information makes it infeasible to deal with all of them and makes the specialization in a specific field necessary. Even after this specialization in a recording method, the extensive number of methods involving the different stages of the recognition process force to a subsequent specialization that reduce a work, like this thesis, to just a small portion of the whole cake. However, an overview of the whole ecosystem provides a global perspective that may facilitate the inclusion of advances in one field to others. As already mentioned, this thesis in particular is focused on external video-cameras, avoiding privacy concerns and restricting to the usual RGB channels captured by most devices, although the developed methods are general and might be included in diverse fields.

2.3 Activity Encoding

Once a specific event of an activity is recorded in a video-clip, a recognition system needs to extract the discriminative information and encode it in a manipulable format for the classifier algorithm. A video stores a lot of information that makes the use of raw video in the recognition systems unmanageable in most of the cases. However, the recent use of the raw information has gained importance in computer vision thanks to the success of Deep Learning methods based on Convolutional Neural Network (CNN), what is specially noteworthy is their use in images [Krizhevsky et al., 2012]. The increase in computational capacity has encouraged Deep Learning methods in human activity recognition but the cost is still high and the few developed researches still do not reach the most successful methods [Baccouche et al., 2011]. However, they can be used as feature extractors without ad-hoc definitions obtaining interesting results [Simonyan and Zisserman, 2014] [Xu et al., 2015]. Despite the high computational cost involved, Deep Neural Networks (DNN) is a research field with a promising future. Nevertheless, in this thesis the classical approach of feature extraction is used as the base for the proposed methods although a switch to DNN based feature extraction would be immediate.

As previously explained, a distinction between the feature extraction stage and the encoding is made below. Through this distinction two processes are separable: the extraction of information and its combination in a manageable format.

2.3.1 Feature Extraction

The idea behind features extraction is to obtain the discriminative information discarding the useless one and store it in what are called descriptors. Feature extraction is a key element in recognition systems thus a significant number of methods have been proposed, some have been collected in different surveys,

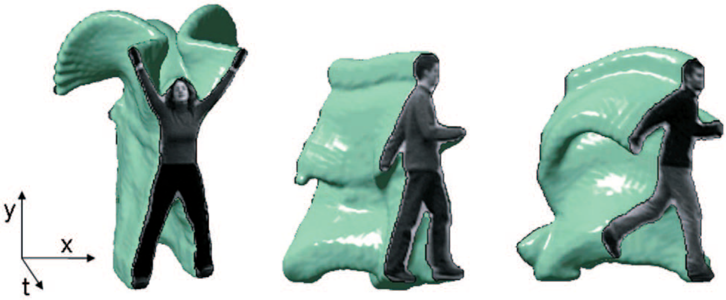


Figure 2.5: Space-time volumes of “jumping-jack”, “walk”, and “run” activities from [Gorelick et al., 2007].

[Weinland et al., 2011] [Ke et al., 2013]. The usual grouping of descriptors includes two methodologies: global descriptors and local or low-level descriptors.

Global Features

Global features are extracted from the whole body of the subject during the time the action is performed, and the descriptor encodes the activity in a holistic representation. The usual process of extraction starts detecting a region of interest (ROI) defined by a bounding box tightly enclosing the subject or a contour of the person body while performing the activity, and continues computing the descriptor in the ROI. Silhouette extraction is a method used to detect the contour and then the ROI, and it can be obtained by background subtraction. Early works used the silhouettes sequentially in Hidden Markov Models as [Yamato et al., 1992] and more recently [Wang and Suter, 2007] in a graphical model. Not only silhouettes have been used to obtain ROIs, but also other approaches as the extension to activity recognition of the adaboost-based face detector of Viola-Jones used in [Ke et al., 2005]. In addition to the sequential approach, ROIs and specifically silhouettes have been used to create volumetric descriptors as the proposed by [Bobick and Davis, 1996] [Bobick and

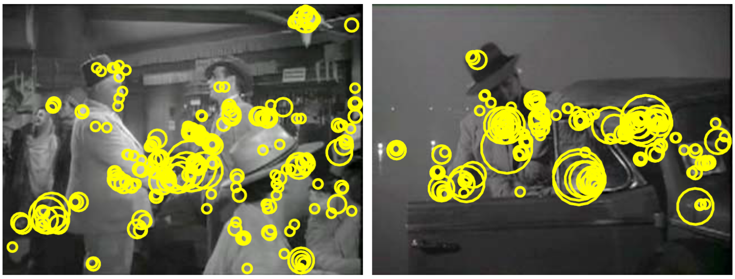


Figure 2.6: Space-time interest points detected in two frames of "hand-shake" and "get-out-car" activities from [Laptev et al., 2008].

Davis, 2001], where they accumulate the silhouette information throughout the clip creating Motion History Images (MHI) or Motion Energy Images(MEI). Other volumetric approaches track the body contours creating spatio-temporal shapes [Yilmaz and Shah, 2005] or obtain spatio-temporal volumes spanned by silhouette images [Gorelick et al., 2007]. A representation of this last method is depicted in Figure 2.5 as representative of global features.

Although global features encode powerfully the activity information, they rely on accurate localization, background subtraction or tracking, being more sensitive to variations in viewpoint, illumination changes, noise and partial occlusions which can be experienced in unconstrained scenarios. These problems has been fought by dividing the ROI into spatio-temporal grids obtaining the descriptors in the defined cells. There are several works in this trend, for example [Danafar and Gheissari, 2007] that compute the optical flow per cell, or [Dalal and Triggs, 2005] that obtain Histograms of Gradients (HOG). This solution has partially solved the problems and is a step towards the local spatio-temporal representation that currently has produced better performances.

Local Features

Without relying on the body detection or any part of it, the local features are extracted from the surrounding of any spatio-temporal point of the video that fulfils the interest conditions defined in the specific method. Local descriptors can be roughly categorised as: (i) image descriptors encoding information from patches surrounding the desired pixels or (ii) space-time descriptors encoding information from spatio-temporal volumes. As non assumption of the global structure of the activity is made, local descriptors are robust to unconstrained videos with challenging variations in background, viewpoint, illumination and so on. Therefore, recent researchers have mainly focused on local descriptors.

Some descriptors only encode the image appearance in a patch like SIFT descriptor [Lowe, 1999] [Lowe, 2004] or Histogram of Oriented Gradients (HOG) [Dalal and Triggs, 2005] which, captured sequentially, can be used in activity recognition. The Histogram of Optical Flow (HOF) descriptor proposed by [Chaudhry et al., 2009] encodes the optical flow value between two frames, considering pixels inside a patch around a key point, into a histogram, capturing mainly temporal information. Spatio-temporal descriptors, like the extensions of spacial descriptors as 3D-SIFT [Scovanner et al., 2007], HOG3D [Kläser et al., 2008] and spatio-temporal Harris interest points STIP [Laptev, 2005] or the previously mentioned HOF, have been proven more efficient than only spatial descriptors. On the other hand, some local spatio-temporal descriptors have been designed so to capture directly the video information as [Laptev and Lindeberg, 2004], Motion Interchange Patterns (MIP) [Kliper-Gross et al., 2012], SCISA [Le et al., 2011] or HOG-HOF [Laptev et al., 2008] do. A representation of this last method is depicted in Figure 2.6 as example of local features. The selection of key-points around where descriptors are computed has been also a field of study and [Wang et al., 2009] concluded that a dense sampling obtains better results than current key-points detectors at that time.

Recent approaches have obtained state-of-the-art results by expanding the local features to in-between models where the spatio-temporal interest points are tracked during some frames obtaining a trajectory around which the descriptor is computed. [Gaidon, 2012] call their method *trackelets* while [Wang et al., 2013] and [Jiang et al., 2012] define theirs as Dense Trajectories. The current state-of-the-art in several unconstrained and challenging datasets has been obtained with the Improved Dense Trajectories (IDT) of [Wang and Schmid, 2013].

2.3.2 Encoding

Global descriptors are usually well formatted so as to be used in a classification process whereas local descriptors suffer for excessive data that sometimes is irregular along the video and then need an encoding process that converts them into a manipulable format. Two encodings stand out in the literature: Bag of Words (BoW), also called Bag of Features (BoF), and Fisher Vectors (FV). Both of the encodings can be considered as global descriptors as they are a holistic representation of the videos, although they are composed from local features.

Bag of Features

This encoding was originally designed to efficiently represent textual documents as frequency of words [Salton and Buckley, 1988], thus the name Bag of Words. Later on, the computer vision community adopted it initially for modelling images as bags of patches[Sivic and Zisserman, 2009] [Cula and Dana, 2001], and finally in videos as bag of local spatio-temporal features, which has been the *de facto* standard until recently. There is an extensive literature using it in video-based recognition, being [Schuldt et al., 2004] [Dollar et al., 2005] [Laptev et al., 2008] [Niebles et al., 2008] [Kliper-Gross et al., 2012] [Wang et al., 2013] just a small portion. Due to this evolution of BoW, the method is also called Bag of Features (BoF) or other variants.

The method is simple, having a dictionary of "visual words" (the terminology is partially maintained from the original text documents but the "words" represent prototype features and sometimes the dictionary is referred as codebook or dictionary of features), the features present in a document (a video for instance) are counted in histogram bins, representing the “visual words” in the dictionary. Usually, such a dictionary is created using clustering of the extracted features in the training examples. In the case of BoF every feature is roughly assigned to a word of the dictionary. However, quantization errors are caused, being specially important when only a small amount of samples are used. Soft-assignment approaches have been proposed to deal with this problem [van Gemert et al., 2010] [Zhu et al., 2011].

Fisher Vector

The soft-assignment proposed for BoF partially solves the information loss produced by the BoF encoding however, Fisher Vector is a more effective method although using a smaller dictionary. This dictionary, also called codebook, is modelled as a Gaussian Mixture Model (GMM) that makes a better representation of the features space than a codebook composed by centroids obtained from a clustering process like K-means. FV stores second order information in large vectors of $K(2d+1)$ dimensions, much larger than BoF, where K is the number of Gaussians and d is the dimension of the local descriptors. This encodings have been recently used in state-of-the-art methods in both image classification [Sanchez et al., 2013] and human activity recognition in videos [Oneata et al., 2013] [Wang and Schmid, 2013].

An important drawback of FV, that also affects BoF, is the loss of global structural information, spatial and temporal, produced because the encodings are obtained from an unordered collection of local descriptors. Thus, they are not designed to classify classes characterized by their structure, and more complex encodings should be used in those cases. For instance, Spatial Pyramid Matching (SPM) [Lazebnik et al., 2006] and the extension to video in Spatio-Temporal Pyramid Matching (STPM) [Choi et al., 2013] divide the data into a grid and compute the encoding in each cell of the grid, keeping the structural information.

2.4 Classifiers

Once performed the activity representation the recognition process becomes a classification problem. Using a Bayes formulation, a classifier calculates the conditional probability $p(y|x)$ of having the label y given the input x. Classifiers are then divided into two groups in relation of the way they obtain this probability. Generative methods that learn a model of the joint probability $p(x, y)$ used later with the Bayes rules to obtain the conditional probability, and discriminative methods that directly learn the conditional probability [Ng and Jordan, 2002].

2.4.1 Discriminative Classifiers

Discriminative methods focus on separating the different classes, rather than learning their model. One of the simplest methods performs a direct classification by selecting the class of the Nearest Neighbour (NN), the class of the closest training sequence based in some distance measure. In order to avoid noise produced by spurious samples, k-NN methodology selects the most common label of the k closest sequences [Bobick and Davis, 2001] [Batra et al., 2008]. Some works compute the distance considering the temporal domain as Dynamic Time Warping (DTW) does, [Veeraraghavan et al., 2005] [Yao and Zhu, 2009]. Among discriminative models, Support Vector Machine (SVM) is widely used. It learns an hyperplane in feature space that discriminates between two classes [Schuldt et al., 2004] [Wang and Schmid, 2013]. SVM has been usually combined with BoF or FV encodings. Working with sequences, some discriminative methods have been designed taking into account the sequential information as Conditional Random Fields (CRF) [Sminchisescu et al., 2006] [Lafferty et al., 2001] and its evolution Hidden CRF (HCRF)Quattoni et al. [2007] do.

2.4.2 Generative Classifiers

On the other hand, generative classifiers learn the model of each class. In addition to their use in classification, they can be used to generate novel examples of a class. A simple generative model is the Naive Bayes classifier, that learns the joint probability of discrete inputs with their corresponding class, and uses this information to predict the conditional probability. In activity recognition they have been used in sensor-based environments with constrained inputs [Singla and Cook, 2009]. In more complex environments, the generative method Hidden Markov Model (HMM) has been widely used in classification for temporal series [Rabiner, 1989]. It models the class with several states, each one having two probability distributions: one modelling the transition among states and the other modelling the observation emissions. There is an extra probability distribution used to predict the initial state. The learning of these distributions is performed through and Expectation-Maximization (EM) method that facilitate the HMM application.

HMM have been used in activity recognition for several years as well [Yamato et al., 1992]. In addition to the basic applying of HMM for classification several variants have been proposed. The original [Rabiner, 1989] paper already proposed some, for instance some intelligence can be introduced in the transition distribution using an ergodic model if it is possible to transit from one state to any other, or a left-to-right model forcing the transition without return possibility. Moreover, two emission models where defined: one for discrete distributions in a Probability Mass Function (PMF) and other for continuous distributions in a Probability Density Function (PDF) with a Gaussian Mixture Model (GMM). Along years more variants have been introduced: [Feng and Perona, 2002] force a learning of the PMF outside the EM process, [Lu and Little, 2006] combine two Markov processes with a single observation in order to recognize the activity and track the subject, [Li et al., 2015] present an adaptive HMM where the emission function is adapted to the data. HMM has been used as well to combine classification and segmentation as in [Lv and Nevatia, 2006] where several weak HMM classifiers are learnt and combined using AdaBoost. Another well known variant is the Hierarchical HMM (HHMM), [Fine et al., 1998], where each emission probability of a main HMM is modelled by another HMM. [Karaman et al., 2014] used HHMM to combine segmentation and recognition in a single method.

2.5 Transfer Learning

As explained in the introduction, one of the goals of this thesis is the training of a human activity recognition system stable with scarce data. In this regard, Transfer Learning (TL) can be an adequate tool. Transfer learning is a straightforward concept, but an exact definition may clarify its use in machine learning. Having a source domain D_S with a learning task T_S and a target domain D_T with a learning task T_T the use of the information learnt in the source domain is transferred to the target domain in order to improve T_T. This idea has received a great attention from the machine learning and data mining community as showed in survey [Pan and Yang, 2010]. Depending on the way the transferred information is used the TL algorithms can be divided into three categories. *Inductive TL* operates with different source and target tasks, so the target task should be learnt in the target domain, but complemented with the source domain information, [Raina et al., 2007]. *Transductive TL* woks with equal source and target tasks, while source and target domains are different, [Daumé and Marcu, 2006]. In this case, there is not labelled data in the target domain, and the information of the source task is transferred. Finally, *Unsupervised TL* works with unlabelled data in both domains, implementing an unsupervised learning, such as clustering, in the target domain complemented with unlabelled data from the source domain, [Dai et al., 2008].

Transfer learning applied in human activity recognition has experimented a great increase of interest, reviewed in a recent survey [Cook et al., 2013]. A new attempt to organizing the different strategies is proposed in this survey. In [Liu et al., 2011] the authors use an inter-lingua in order to merge data from source domains and target domain, considering labelled data available only in the source domain, which classifies their method as *uninformed supervised* (US). In [Bian et al., 2012] a similar approach is implemented by creating a cross domain codebook where labelled actions from both domains are modelled with BoW, being an *informed supervised* (IS) transfer learning method. Authors in [Zhu et al., 2011] create a codebook with unlabelled data from the source domain and train the recognition with labelled data from the target domain, being an *informed unsupervised* (IU) transfer learning method. The literature is really extensive, but most of the methods use similar approaches [Hu et al., 2011] [Fei-Fei et al., 2006].

2.6 Learning with Limited Examples

Little research has been done in training an activity recognition system with limited number of labelled examples although being an essential feature in many practical situations. Some solutions propose the use of invariant features, for instance [Sun and Aizawa, 2013], or the application of Transfer Learning in a cross domain evaluation like [Cao et al., 2010] [Hu et al., 2011] [Bian et al., 2012]. These solutions reduce the number of labelled sequences in the target domain but still need some information. The limit of this information is the one-shot learning approach where only one example from the target scenario is needed. Some examples of it are found in [Seo and Milanfar, 2011] [Fanello et al., 2013] [Yang et al., 2013].

In the ideal case, only one sequence per class should be enough for activity representation, calling the learning from only one example as one-shot learning. Moreover, it is important to mention that the description of a one-shot learning approach differs among papers in the literature. In order to have a better understanding of the meaning we introduce two concepts that define two

Table 2.1: Strict One-shot Learning VS. Relaxed One-shot Learning

	min n^o examples learn class	min n^o examples learn recognition
Strict	**1**	n^o classes
Relaxed	**n^o classes**	n^o classes

different one-shot learning approaches. First, the **strict one-shot learning** assumes only one training example available which is used to model a single class. After training several models (one per available example) of different classes separately, it is possible to combine these models in order to train a recognition systems, which would be a generative one by nature. [Seo and Milanfar, 2011] proposed a nearest-neighbour classification using a strict one-shot learning approach. Second, the **relaxed one-shot learning** process uses simultaneously multiple training examples available, assuming one per class. This relaxation allows sharing some information among the examples in order to model the classes or directly training a recognition system, which would be a discriminative one by nature. The relaxed approach usually gives better results but at the expense of retraining the system with each new inclusion and with the inconvenience of requiring several examples from the beginning. [Yang et al., 2013] and [Fanello et al., 2013] approaches follow this description, the former method creates a vocabulary of features using sequences of the different classes while the latter trains a SVM with sub-sequences of the activity example as positive examples and sub-sequences of different classes examples as negative examples. We depict the properties of both models in Table 2.1, highlighting the difference. The strict method models a class with only one example while the relaxed approach needs information form one example per class. The strict approach is more versatile when including new examples as they are all independent and can be added to the system at any time. The relaxed approach usually gives better results but at the expense of retraining the whole system with each new inclusion and with the inconvenience of requiring several examples from the beginning.

2.7 Recognition using Temporal Information

Many approaches have been designed to account complex temporal structures recognizing complex human activities defined as composite multimedia semantics (e.g., birthday party, wedding ceremony) where orderless sub-actions appear in the video. These approaches consider the order of sub-actions as a distracter (not a discriminant) and hence they perform an alignment of similar sub-scenes disregarding their order. For instance, [Xu and Chang, 2008] divide the clips into sub-clips which later are matched with other sequence using the earth mover's distance (EMD). [Cao et al., 2012] have designed a kernel that makes a pooling of the frames into a fixed number of scenes called Scene Alignment Pooling (SAP). In [Vahdat et al., 2013] a detection of sub-scenes categories and a global scene category are combined in a Multiple Kernel Latent SVM where several features are used. In the work of [Li et al., 2013] the proposed method identifies the most representative segments of the actions using a dynamic pooling with a latent variable.

As opposed to the previously explained works, the temporal order of the sub-actions performed in an activity is considered essential to achieve the second goal of this thesis, as the objective is to distinguish between complex activities that can be composed of same sub-actions but in different order, even opposite. For instance the activities shown in the Introduction "Getting Out of Vehicle" and "Getting Into Vehicle" in Virat dataset depicted in Figure 1.3.

State-of-the-art results in activity recognition have been achieved using SVM classifiers, but some constraints inherent to SVM should be overtaken in order to include the temporal information for recognizing the class. Thanks to the kernel trick, used to compute an inner product in some arbitrary space, the SVM can classify in a dimensional space where samples may be linear-separable, different from the original one where the samples are not linear-separable. The standard kernel methods assume a fixed length D-dimensional vector per sample which is projected into a different space where the inner product is performed. However, this is not straightforward in activity recognition videos where the long-term activity dynamic information remains in the encoding because lengths of sequences may be arbitrary. Two solutions have been proposed in the literature: (i) obtaining some sort of inner product by aligning the sequences lengths of the patterns, as Dynamic Time-Alignment Kernel [Shimodaira et al., 2002] or Fast Global Alignment Kernel [Cuturi, 2011] do and (ii) training a HMM with a single sequence and posterior obtaining a Probability Product Kernel (PPK) [Jebara et al., 2004] like in [Jebara et al., 2007]. Both solutions have been used in sequence clustering tasks [Zhou et al., 2013] [Jebara et al., 2007] [Rodriguez-Serrano and Singh, 2012]. Sequence alignment enforces a common start and end of the segmented event which is not always the case. On the other hand, the PPK of HMMs implies that each HMM is trained with only one sequence which does not offer sufficient information to train properly the parameters of a complex model. Moreover, the optimization process in the HMM training is performed with the Baun-Welch algorithm which only assures a local optimum.

The long-term temporal information has been used in some other methods. A recent one is the extension of the work of Spatial Pyramid Matching (SPM) [Lazebnik et al., 2006] called Spatio-Temporal Pyramid Matching (STPM) [Choi et al., 2013]. The method suggests dividing the videos into equal number of spatio-temporal volumes at several scales, called pyramids, computing in each volume a BoF, and finally obtaining a similarity between two video clips by comparing the corresponding volumes. Fixing the number of divisions and comparing volumes one-to-one constrain the method to regular paced actions losing flexibility. In [Ryoo and Aggarwal, 2009] they encode each video into a 3D histogram with spatio-temporal information and design a specific kernel for the 3D histogram, but their pairwise feature comparison is not suitable for dense features extraction which recently have provide the state-of-the-art results. Using HMM, the authors of [Tang et al., 2012] propose to learn the temporal structure, including latent variables that determine the expected time to stay in a state. [Todorovic, 2012] keeps the long-term information using graph models of the video foreground and with a Kronecker product constructs a Kronecker graph model per action class used in the classification. In [Niebles et al., 2010] the sequences length is not constrained and the authors use a framework where appearance and temporal position of motion segments is included, however, due to high computational load, their learning process finds a local optimum.

2.8 Datasets

Once a method is proposed it should be carefully evaluated, verifying its benefits and finding its drawbacks. In this regard, from the large amount of available datasets, we have selected some public ones that fulfils the needed requirements. On the one hand they are adequate for the methods proposed, and on the other hand they are commonly used by the computer vision community and specifically for the objective tasks.

We have selected 9 different datasets because of their properties and their popularity. Weizmann, IXMAS and KTH offer several sequences recorded from the same point of view, what is ideal for validating our proposals in one-shot learning. Moreover, these datasets are among the most popular ones in activity recognition. On the other hand, recent researches have focused in challenging datasets where there is not control in scenarios, viewpoints and even movement of the cameras. HMDB51, UCF11, UCF50 (which is an extension of UCF11) and OlympicSports are among the most used, and Virat Release 2.0, although being recorded from fixed cameras as it is composed from surveillance videos, it offers a high variability in the classes and includes several viewpoints. Moreover, Virat offers labelled activities defined by their order of sub-actions. In addition, we use the ViHASi dataset for training in some experiments as being a synthetic one represents the possibility of creating training examples indefinitely. We can see some samples of these datasets in Figure 2.7.

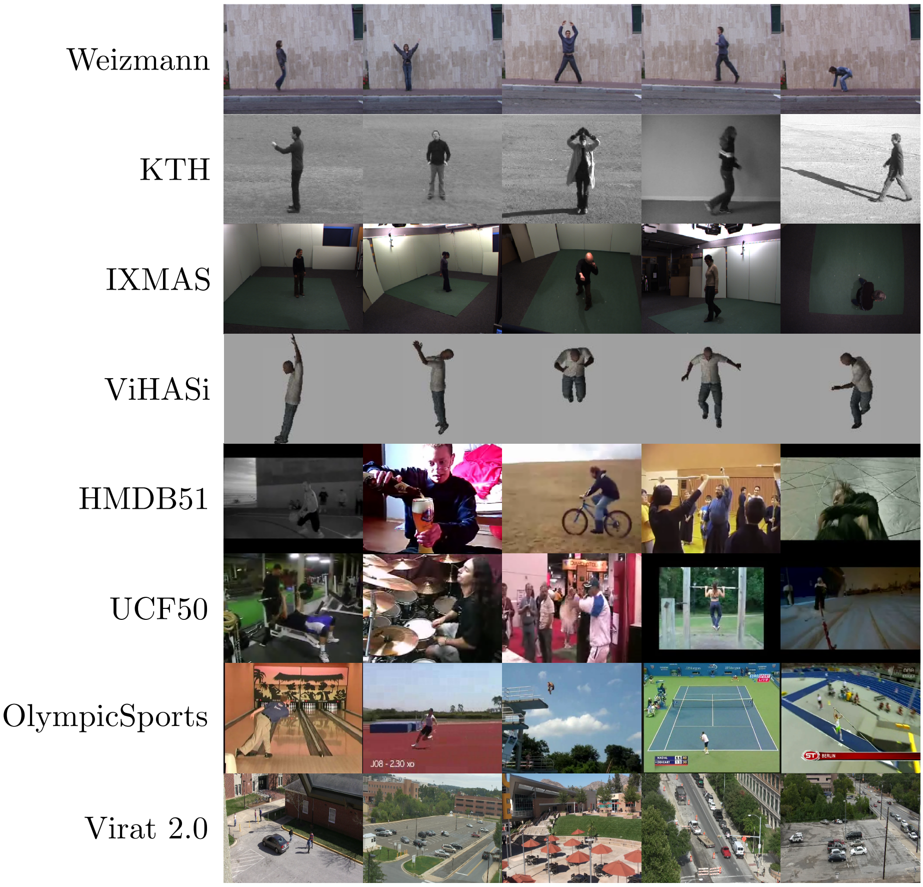


Figure 2.7: Samples of the datasets used in this thesis. UCF50 is an extension of UCF11 and the depicted images are representative enough of the kind of videos available in both datasets.

Weizmann [Gorelick et al., 2007] The Weizmann dataset is composed by 93 low-resolution (180 x 144, 50 fps) video sequences from the same viewpoint showing nine different people, each performing 10 natural activities: *bend, jumping-jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, run, gallop-side-ways, skip, walk, wave-one-hand* and *wave-two-hands.*

IXMAS [Weinland et al., 2006] The IXMAS dataset is composed by 5 camera viewpoints (390 x 291, 23 fps) of 11 actors performing 3 times each of the 13 activities included: *check-watch, cross-arms, scratch-head, sit-down, get-up, turn-around, walk, wave, punch, kick, point, pick-up* and *throw.*

KTH [Schuldt et al., 2004] The KTH dataset has been captured in 4 different scenarios where static cameras have recorded, at low-resolution (160 x 120, 25 fps), 25 subjects performing several times six types of activities: *walking, jogging, running, boxing, hand-waving* and *hand-clapping.*

ViHASi [Ragheb et al., 2008] The ViHASi dataset has been virtually created with 20 action classes, 9 different actors and 40 perspective camera views.

HMDB51 [Kuehne et al., 2011] The HMDB51 dataset is one of the most challenging datasets nowadays. It contains a collection of videos obtained from a variety of sources ranging from digitized movies to YouTube videos. The total of 6766 video clips contains 51 distinct activity categories each one represented by at least 101 examples. The dataset is divided by the authors into 3 splits, each one containing 70 training clips and 30 testing clips in order to display a representative variability of the recording sources. The dataset includes a stabilized version of the videos.

UCF11 [Liu et al., 2009] The UCF11 dataset, also called UCF YouTube, is obtained from YouTube videos. It contains 1160 video clips of 11 different activities, *basketball-shooting, biking/cycling, diving, golf-swinging, horse-back-riding, soccer-juggling, swinging, tennis-swinging, trampoline-jumping, volleyball-spiking* and *walking-with-a-dog.*

UCF50 [Reddy and Shah, 2013] The UCF50 dataset is obtained from YouTube videos and is an extension of UCF11. It contains 6681 video clips of

50 different activities. Some of these videos are segmentations of a longer one, so it is important to follow the authors' protocol. The authors suggest a division into 25 groups in order to apply a leave-one-group-out cross-validation strategy.

OlympicSprots [Niebles et al., 2010] The OlympicSprots dataset contains 783 videos of athletes practising 16 different sports. All video sequences were obtained from YouTube and have been annotated with the help of Amazon Mechanical Turk. The authors suggest a split for training and testing the recognition system.

Virat Release 2.0 [Oh et al., 2011] The Virat Release 2.0 dataset has been recorded in 11 different scenes of video surveillance, captured by stationary HD cameras (1080p or 720p). There are 11 different classes of activities annotated where persons and vehicles appear *Loading, Unloading, Opening Trunk, Closing Trunk, Getting Into Vehicle, Getting out of Vehicle, Entering Facility, Exiting Facility, Gesturing, Carrying* and *Running.* The authors suggest a scene-independent learning and recognition mode of evaluation.

3

Activity Encoding

As stated before, activity encoding is a middle process used to format the information available in a video clip in order to feed the classifier. Sometimes this stage is named feature extraction but we use this term for the initial process of extracting the information into local or global descriptors which latter are processed into a holistic encoding understandable by the classifier. The objective is to keep as much discriminative information as possible avoiding clutter and redundant information and being manageable by the classifier. A brief description of the encoding and slight variations of existing algorithms described in Annexe A is introduced below.

As shown in the related work the feature extractor can obtain global and local descriptors but, at the end, most classifiers work with holistic representations so, when working with local descriptors a second stage is applied in order to model the video into a holistic representation used in the classification process. Moreover, some post-processing may be applied to global descriptors as well in order to normalize or quantize the extracted data forming the final holistic representation.

Some holistic representations lose long-term temporal information, therefore

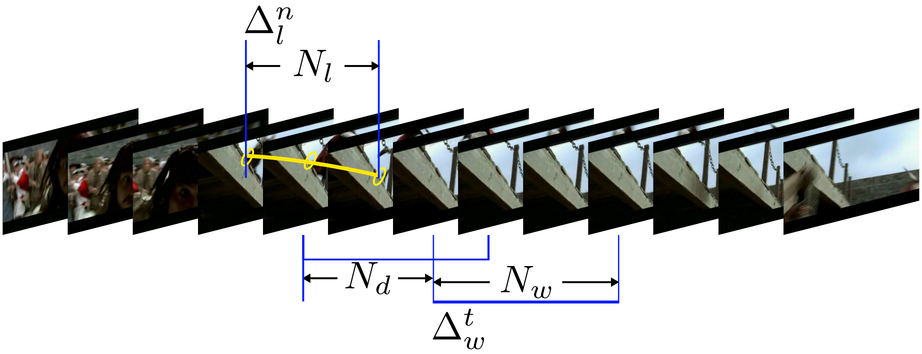


Figure 3.1: Frame-windowed video clip framework.

we have followed a common framework in every encoding which divides video clips into frame-windows, depicted in Figure 3.1, so to keep this important information used in all our methods. Each extracted descriptor $\mathbf{q}_n$, being global or local, is computed through a temporal window Δ_l^n of length N_l. The most temporally local descriptors use information of a single frame so $N_l = 1$, while the most temporally global descriptors cover all the frames of the activity, therefore $N_l = F$, being F the number of frames of the video-clip. On the other hand, we use a sliding frame-window process designed to capture the long-term temporal information by computing holistic descriptors in each frame-window. Δ_w^t is the t-th window with a duration of N_w frames. Finally, the windows stride is N_d frames, parameter used to define the overlapping of the windows. The values N_l, N_w and N_d are generally different and therefore the holistic descriptor should be accordingly computed as we explain later in this chapter.

At the end of the video encoding we obtain a time-series of observations $\mathcal{O} = \{O_1, \cdots, O_T\}$, where T is the number of windows, ready to be processed by the classifying algorithm.

3.1 Descriptors

From the different available methods in the literature we have selected two global descriptors ant two local spatio-temporal descriptors. Many more are described in the literature, [Weinland et al., 2011] [Ke et al., 2013], but the selected ones cover our needs as they include both worlds, local and global, and the last one has been reported as the state-of-the-art in many scenarios.

As global descriptors we have used human silhouettes and Motion History Images, described in A.1.1. Both, silhouette and MHI descriptors are matrices of a large size, equal to the raw frames. In order to feed the classification system every extracted descriptor should have the same size so, the bounding box cut image is resized to a common size. Moreover, the number of pixels can be excessive without contributing with discriminant information, so to reduce the computational cost and maximize the discriminant information we apply the trustworthy method Principal Component Analysis (PCA), [Pearson, 1901], trained with the descriptors extracted from the training video clips, codifying each descriptor in a single vector of a reduced dimensionality.

On the other hand, there are several low-level space-time features used in the literature, but we have selected two of them in order to evaluate short-term as well as medium-term local features. In the short-term range, we have selected the Motion Interchange Patterns (MIP) since they have been reported to provide some good results when dealing with HMDB51 and UCF50 datasets [Kliper-Gross et al., 2012], improving the method of [Laptev et al., 2008], STIP points with HOG-HOF descriptors, which has been a baseline in many experiments. For the sake of clarity we introduce a brief description of its extraction in A.1.2. At the moment of writing this thesis, Dense Trajectory (DT) [Wang et al., 2013], and the later version Improve Dense Trajectory (IDT) [Wang and Schmid, 2013] have achieved the state-of-the-art results in most of the current challenging datasets, for instance: HMDB51, UCF50 and OlympicSports. For this reason, we have selected this descriptor as the baseline of our systems in

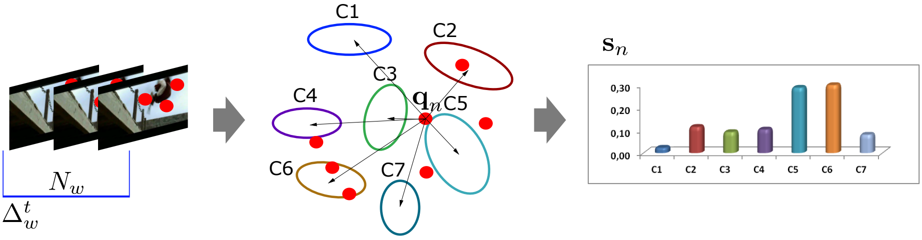


Figure 3.2: Soft-assignment of descriptor $\mathbf{q}_n$ in a clustered space, obtaining vector $\mathbf{s}_n$.

the final versions of our approaches. We introduce a brief description in A.1.3.

3.2 Holistic Encoding

After the feature extraction process, we propose to combine all the extracted features into a holistic descriptor per frame-window Δ_w^t so that to keep the long-term temporal information. Global descriptors already obey the holistic descriptor paradigm however we propose to apply the same proposed formulation for local descriptors just with the characteristic that there are much less global descriptors in a window than local descriptors, usually only one. This transformation of global descriptors quantize the information reducing the variability. We have selected two of the most common encodings to apply in a window, Bag of Words and Fisher Vectors. To perform them, both processes compute a clustering of the descriptors extracted from training examples in order to define the feature space. This clustering represents the "dictionary" of features used as scaffold of the descriptors.

3.2.1 Clustering

Despite the large number of methods available in the literature, the needed characterization of the features space is suitably obtained using K-means or Gaussian Mixture Models (GMM) optimization. Multiple works demonstrate their validity, including [Kliper-Gross et al., 2012] [Wang et al., 2013] and [Wang and Schmid, 2013]. K-means creates a Voronoi Tessellation of the features space through an iterative optimization method, explained in A.2.1, while the GMM is created usually with an Expectation Maximization process that maximizes the likelihood of the model, explained in A.2.2. The result of both methods are hyperellipsoidal clusters. On the other hand, both optimization algorithms only assure the finding of local optima [Jain, 2010].

Clustering is useful for a quantization of the feature space so to give a comparable structure of data obtained from the extracted descriptors. Using a function for obtaining a belonging measure of a descriptor to any cluster is possible to make a hard-assignment or a soft-assignment of descriptors. A hard-assignment uses the winner-takes-it-all rule so the cluster with highest belonging measure is assigned to the evaluated descriptors. On the other hand, the soft-assignment uses directly the value of the function, having every cluster a belonging value. In the proposed encodings, descriptors, $\mathbf{q}_n$, are extracted for every frame-window, Δ_w^t, which are soft-assigned to the clustering obtaining the vector $\mathbf{s}_n$.

3.2.2 Bag of Features

The fundamentals of BoF are described in A.3.1 and based on those fundamentals we adapt them to the windowed sequence. As we have seen at the beginning of this chapter, the descriptors are computed through a temporal frame-window of length N_l, generally different to the length of the windows where holistic descriptors are computed, N_w. Each descriptor, $\mathbf{q}_n$, has associated a temporal window Δ_l^n and therefore this descriptor influences proportionally to each Δ_w^t

window given the Equation 3.1

$$\rho_{nt} = \frac{|\Delta_w^t \cap \Delta_l^n|}{N_w} \tag{3.1}$$

Each observation O_t is a holistic descriptor, $BoF(t)$, calculated using Equation 3.2 in window Δ_w^t.

$$BoF(t) = \frac{1}{\sum_{n=1}^{N} \rho_{nt}} \sum_{n=1}^{N} \rho_{nt} \mathbf{s}_n \tag{3.2}$$

where $\mathbf{s}_n$ is the soft-assignment vector of each descriptor $\mathbf{q}_n$.

We finally convert the raw data into a new time-series of observations $\mathcal{O} = \{O_1, \cdots, O_T\}$ adequate for the following processes as we will see in next chapters. Depending on the working data, the possible $\sum_{n=1}^{N} \rho_{nt} = 0$ would cause numerical problems so, it is important to cope with those cases, for instance eliminating the corresponding O_t or giving a default value.

The vocabulary can be obtained with both, K-means algorithm or GMM optimization, and both, hard and soft assignment computed. However, the most common algorithms combine hard-assignment with K-means and soft-assignment with GMM.

3.2.3 Fisher Vectors

The fundamentals of FV are explained in A.3.2, but in order to cope with the windowed sequences we introduce some modifications.

Given the set of features $Q = \{\mathbf{q}_1 \ldots \mathbf{q}_N\}$ and a GMM $\lambda = \{\omega_k, \mu_k, \Sigma_k\}$ describing the features space, the associated FV is computed as follows.

First, the posterior probability of sample $\mathbf{q}_n$ to Gaussian λ_k is computed:

$$s_{nk} = \frac{e^{-\frac{1}{2}(\mathbf{q}_n - \mu_k)\Sigma_k^{-1}(\mathbf{q}_n - \mu_k)^T}}{\sum_{j=1}^{K} e^{-\frac{1}{2}(\mathbf{q}_n - \mu_j)\Sigma_j^{-1}(\mathbf{q}_n - \mu_j)^T}} \tag{3.3}$$

As FV encoding faces the issue of a sliding frame-windows, the length of the window N_w and the length of the descriptor N_l do not generally coincide, we introduce the factor of intersection ρ_{nt} obtained with Equation 3.1. Modifying the deviation vectors as follows:

$$u_{jk} = \frac{1}{\sum_{n=1}^{N} \rho_{nt}\sqrt{\omega_k}} \sum_{n=1}^{N} \rho_{nt} s_{nk} \frac{q_{jn} - \mu_{jk}}{\sigma_{jk}} \tag{3.4}$$

$$v_{jk} = \frac{1}{\sum_{n=1}^{N} \rho_{nt}\sqrt{\omega_k}} \sum_{n=1}^{N} \rho_{nt} s_{nk} \left[\left(\frac{q_{jn} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \tag{3.5}$$

3.3 Summary

Thanks to the proposed encodings we cover a wide spectrum of possibilities using both global and local descriptors. In Figure 3.3 we summarize the followed encodings through a diagram.

Global descriptors are transformed to a soft-assignment BoF. To do so, we first model the features space with a GMM. On the other hand, local descriptor encoding into a holistic one is more varied. Firstly, the features space is modelled with a GMM, but also it can be done through K-means. If K-means is used, then a hard-assignment BoF is performed. In the case of GMM, it is possible to perform hard and soft assignment BoF as well as FV.

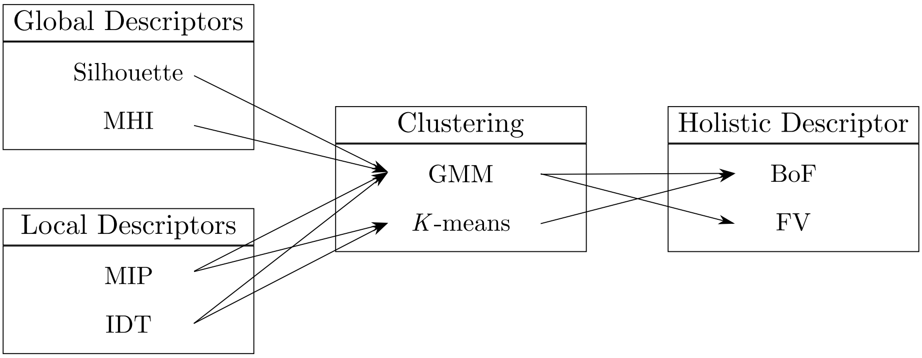


Figure 3.3: Summary of the encoding techniques and their combination followed in this thesis.

4

Relaxed One-Shot Learning of Activities based on Soft-assignment and HMM

The easiness of acquiring and placing video cameras in almost any location has encouraged the advances in automatic human activity recognition and its application in many smart systems. Specifically, there are several applications based on fixed camera systems where each camera records the scenario always from the same viewpoint. We can think about all the surveillance cameras we have around us in our daily life, being in public transport, work places, bank offices, public buildings and even particular houses among others. All these cameras mainly record the scenario and are checked only after something happens, but rarely prevent any undesirable event. This performance could be improved if the surveillance systems were able to recognize activities and at least trigger an alarm after detecting a suspicious behaviour. Other applications are being created based on fixed cameras in devices like personal computers or consoles, mainly used for human-machine interaction, for instance by controlling the device or by improving interactivity in games. An additional scenario, still less developed because of privacy concerns, is the installation of fixed cameras at home which would improve the performance of Ambient Assisted Living (AAL) systems. Strategically placed cameras at home, for instance in the kitchen, could distinguish how daily activities are performed, helping the subject in some important tasks like cooking, cleaning or tidying, triggering alarms if risky behaviours are detected or telling the subject about some oversight or the next step to perform in an activity. Although for most users this helping may look unnecessary, many elderly or disable people might benefit from it by making their lives less handicapped providing an extension of our independent lives, as most people will became handicapped sooner or later.

A common pattern is present in all the previous applications, the need to deploy the system in each new scenario placing the cameras with fixed viewpoint, unique per scenario. And this pattern is the seed for our research in one-shot learning activity recognition systems. But, why the training with limited number of examples is important there?

In the activity recognition field, most of the proposed recognition approaches are trained with a large amount of labelled examples using large databases. In practice, this is reasonable for applications with unconstrained scenarios such as searching specific activities in movies, or indexing Internet videos, where the training examples can be obtained relatively easy from on-line videos. Results in large and unconstrained datasets such as HMDB51 [Kuehne et al., 2011] or OlympicSports [Niebles et al., 2010] are useful for general evaluation, because their examples have been collected from diverse sources, for instance Youtube or extracted from movies, but accuracy of algorithms is not yet at the level required in many commercial applications.

Higher accuracy can be achieved in constrained scenarios and with fixed cameras, such as the ones we have as goal, and they can be represented by the Weizmann [Gorelick et al., 2007], KTH [Schuldt et al., 2004] or IXMAS [Weinland et al., 2006] datasets. However, these good results are obtained assuming the availability of several labelled examples for training, which in many situations is impractical.

Recognizing in a fixed scenario has the advantage of suppressing in some degree the clutter introduced by the change of background and viewpoint and therefore higher reliability may be achieved. However, after the installation, the system should be re-trained again as any previously collected sequences may not be representative of the new scenario. Although the performance is constrained by the number of labelled sequences used for training, collecting and labelling large amount of data for the particular scenario is infeasible, as it is laborious and may require the involvement of the user and here comes the importance of training with limited number of examples. Reducing the number of training examples to a minimum for a suitable method reduces the set-up time of the system. Little research has been done in training an activity recognition system with limited number of labelled examples although being an essential feature in many practical situations, two examples are [Seo and Milanfar, 2011] and [Yang et al., 2013].

As we have already explained, we pursue a system able to operate from the first recorded sequence. A system like this implies the use of a **strict one-shot learning** scheme. It is not possible to use a discriminative recognition system at this stage, as we do not dispose of examples from different classes, instead, we have chosen the Hidden Markov Model generative method in order to build our system. However, as we will see, the traditional models suffer form some drawbacks when working with scarce data. Moreover, although the main goal is the use of a strict one-shot learning method, we have experimented with **relaxed one-shot learning** methods for two reasons: first, they have better performance as they dispose of more initial information and may be attainable in some cases and second, there are some previous works using the relaxed approach that allows us to check our proposals comparatively.

In this chapter we present a preliminary approach analysing some strategies for a suitable one-shot learning method. In this regard, we perform the work in a relaxed one-shot learning scenario but with the intention to migrate the learnt strategies to a strict one-shot learning method. We first start the chapter explaining briefly the HMM principles, and their use for a generative recogni-

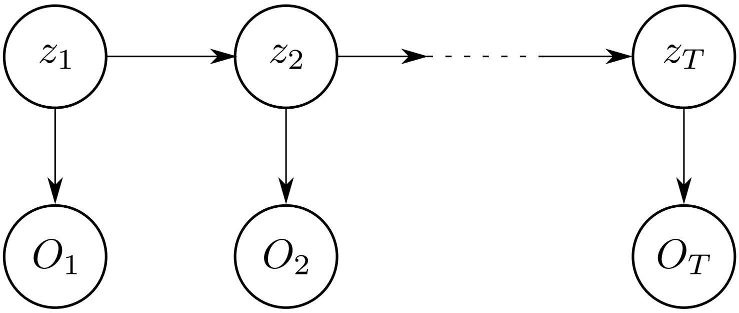


Figure 4.1: A first order Markov chain of latent variables Z that produces a series of observations $\mathcal{O}$.

tion system in Section 4.1. Later, in Section 4.2, we explain a soft-assignment encoding method of the activities and a specific HMM approach, called Fuzzy-discrete-HMM (FDHMM), which represents a stable solution for the defined encoding. We will see how the system needs the creation of a codebook and how we have designed two methods for its computation that represent the two approaches analysed in this chapter. Later, we present the experiments validating the approaches in Section 4.3. Finally, in Section 4.4, we conclude the results and expose the drawbacks of the systems that are addressed in next chapter, focusing there on a strict one-shot learning approach.

4.1 Hidden Markov Model

Hidden Markov Models are generative approaches designed to characterize sequential data. Specifically, the topic of this thesis is the understanding of human activities, which are encoded in time series as shown in Section 3. Being HMM a generative approach, and using the traditional training of the model, it is possible to compute the learning of the model using only one sequence example therefore, it fulfils the requirement for strict one-shot learning.

HMM supposes that any given time series of observations, $\mathcal{O} = \{O_1, \cdots, O_T\}$, is produced by a series of latent (or hidden) variables, $Z = \{z_1, \ldots, z_T\}$, that form a first order Markov chain, as shown in Figure 4.1. These latent variables are discrete multinomial variables describing which state is responsible for the generation of the observations. Formally, the parameters of the HMM are $\theta = \{N, A, B, \pi\}$. N is the number of states, i.e., $S = \{S_1, \ldots, S_N\}$. Each observation, O_t is the emission produced by the hidden state z_t. The set of hidden states forms a sequence, $Z = \{z_1, \ldots, z_T\}$ where $z_t \in S$. $A = \{a_{ij}\}$ is the state transition matrix where a_{ij} represents the transition probability from state i to state j, $a_{ij} = p(z_{t+1} = S_j | z_t = S_i)$. $\pi = \{\pi_i\}$ is the initial state probability distribution where $\pi_i = p(z_1 = S_i)$, $1 \leq i \leq N$ being S_i the state at the beginning of the time series. Finally, B represents the observation probability distribution in every state where $b_j(O_t) = p(O_t \mid z_t = S_j)$.

There is not restriction on the observations nature as long as their emission can be modelled. Traditionally the HMM have been divided into two groups attending to the way the B parameters are modelled: the discrete HMM for categorical observations where B represents a Probability Mass Function (PMF) per state, and the continuous HMM for multivariate continuous observations, where a Probability Density Function (PDF) is defined per state. The continuous HMM usually uses a GMM per state in order to define the PDF of the observations.

Given the previous definition of HMM, two of the three basic problems of interest defined in [Rabiner, 1989] need to be solved in order to use them for recognition.

Problem 1: Given an observation sequence $\mathcal{O} = \{O_1, \cdots, O_T\}$ and a model $\theta = \{N, A, B, \pi\}$, what is the probability of the observation sequence given the model $p(\mathcal{O}|\theta)$. The solution to this problem allows for applying a classification criterion. Suppose W different classes and then W different models, $\theta_w, 1 \leq w \leq W$, one per class. The most likely class is then selected through $w^* = \mathrm{argmax}_w(p(\mathcal{O}|\theta_w))$.

Problem 3: Before the $p(\mathcal{O}|\theta)$ computation is possible, the adjustment of the model parameters $\theta = \{N, A, B, \pi\}$ to some observation sequences available for training is necessary.

The Problem 2 that identifies the most probable state sequence is irrelevant for our objective.

Problem 1 can be efficiently solved using the *forward-backward procedure*. In particular, it is solved using the forward variant of the procedure. In this regard the method defines the forward variable $\alpha_t(i)$ as

$$\alpha_t(i) = p(O_1, \cdots, O_t, z_t = S_i|\theta) \tag{4.1}$$

Using the forward procedure, $p(\mathcal{O}|\theta)$ is solved inductively as follows:

1. Initialization:
$$\alpha_1(i) = \pi_i b_i(O_1) \tag{4.2}$$

2. Induction:
$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \tag{4.3}$$

3. Termination:
$$p(\mathcal{O}|\theta) = \sum_{i=1}^{N} \alpha_T(i) \tag{4.4}$$

Additionally, a backward procedure is defined in a similar manner which is used in the resolution of Problem 3 as we will see later. In this regard the backward variable $\beta_t(i)$ is defined as

$$\beta_t(i) = p(O_{t+1}, \cdots, O_T|z_t = S_i, \theta) \tag{4.5}$$

And the computation of $\beta_t(i)$ is inductively solved as follows

1. Initialization:
$$\beta_T(i) = 1 \tag{4.6}$$

2. Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{4.7}$$

A solution to the Problem 3 is the optimal adjustment of the HMM parameters to the training observations through the maximum likelihood, $\theta^* = \mathrm{argmax}_\theta\, p(\mathcal{O}|\theta)$. In this case $\mathcal{O}$ may represent one or several training sequences although, for the sake of clarity, we only refer to one in the formulation. Being the adaptation to several sequences direct through a summation term. The computation of the global optimum is analytically impractical and therefore an acceptable estimation has been proposed becoming standard process. Given one or several training observations, the HMM parameters can be estimated finding a maximum in the Likelihood function through a Baum-Welch algorithm. This process maximizes the Baum's auxiliary function $Q(\hat{\theta}, \theta)$, Equation 4.8 through an iterative estimation [Rabiner, 1989] [Bishop, 2006].

$$Q(\hat{\theta}, \theta) = \sum_{Z} p(Z|\mathcal{O}, \theta) \ln p(\mathcal{O}, Z|\hat{\theta}) \tag{4.8}$$

Defining $\gamma_t(i) = p(z_t = S_i|\mathcal{O}, \theta)$ and $\xi_t(i,j) = p(z_t = S_i, z_{t+1} = S_j|\mathcal{O}, \theta)$, the function Q can be expressed as:

$$Q(\hat{\theta}, \theta) = \sum_{j=1}^{N} \gamma_1(j) \ln \pi_j + \sum_{t=1}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \xi_t(i,j) \ln a_{ij} + \sum_{t=1}^{T} \sum_{j=1}^{N} \gamma_t(j) \ln(b_j(O_t)) \tag{4.9}$$

The mentioned Baum-Welch algorithm can readily be interpreted as an Expectation Maximization (EM) algorithm as we following present it.

E-step: The expectation step implies the calculation of functions $\xi_t(i,j)$ and $\gamma_t(i)$, with which the auxiliary function $Q(\hat{\theta}, \theta)$ is calculated. Their computation is performed making use of the *forward-backward procedure*

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\Sigma_{i=1}^{N}\Sigma_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \tag{4.10}$$

And $\gamma_t(i)$ is obtained with Equation 4.11.

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{4.11}$$

M-step: The maximization step is implied to maximize $Q(\hat{\theta}, \theta)$ over $\theta = \{A, B, \pi\}$, performed using Lagrange multipliers. After this process $\hat{\pi}_i$, $\hat{a}_{ij}$ and $\hat{b}_j$ are obtained. The optimizations of $\hat{\pi}_i$, $\hat{a}_{ij}$ are independent of the emission model and are obtained with Equations 4.12 and 4.13.

$$\hat{\pi}_i = \gamma_1(i) \tag{4.12}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{4.13}$$

On the other hand, the maximization of $\hat{b}_j$ depends on the emission model selected. In this regard, for a discrete HMM where the observations belongs to a codebook $C = \{c_1 \ldots c_K\}$, $O_t \in V$, it is easily obtained with Equation 4.14.

$$\hat{b_j}(v_k) = \frac{\sum_{t=1}^{T} \gamma_t(j) \cdot r_{tk}}{\sum_{t=1}^{T} \gamma_t(j)} \tag{4.14}$$

where we use the parameter $r_{tk} \in \{0, 1\}$ where $r_{tk} = 1$ if, when being in state j, observation O_t is exactly v_k and 0 otherwise.

For continuous HMM more complex optimizations are performed, but optimization algorithms are available in the literature and several code libraries.

The main general drawback of the EM algorithm is as follows: there is in general no guarantee of reaching a global optimum, and local optima can at times be problematic, particularly with small training sets. Additionally, we have focused in small training sets which produce overfitting as we increase the number of parameters to train. In discrete models the problem is less serious as the number of parameters is smaller however, the lack of information may produce incorrect estimations. On the other hand, the usual continuous HMM approach that incorporates GMM distributions may lead to an unstable EM algorithm as explained in [Shinozaki and Ostendorf, 2008].

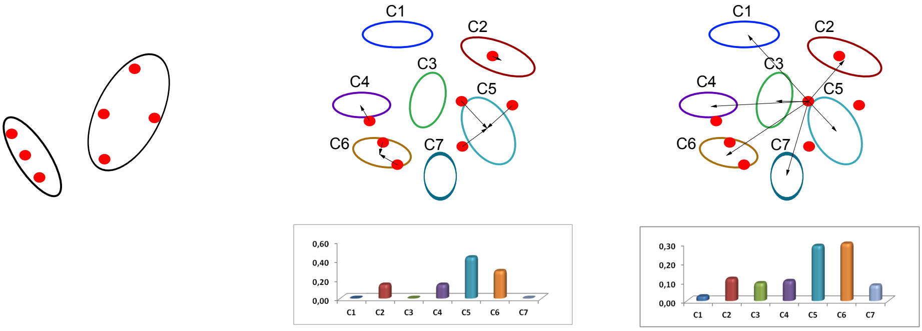
4.2 Soft-assignment and FDHMM

In a pattern recognition system, when only limited information is available for training, two unavoidable main problems should be addressed: first, the loss of discriminative information in the different stages should be reduced to a minimum as it cannot be compensated with extra data; second, the risk of numerical instabilities in the training process if it is not properly designed.

Initially, we opted for representing each activity example as a sequence of subject poses. In this regard, we selected the silhouette in each frame or the MHI of a window of frames, as explained in Section A.1.1. The MHI is a richer feature as it includes short term temporal information. However, the number of frames of the window is a critical parameter as too many might suppose the occlusion of some important information. On the other hand, as the temporal information is coped in the HMM, the simple silhouette may simplify the method. We experimented both approaches in this preliminary exploration. It is worth noting that the implementation of the proposed approach is independent of the descriptor, as long as it is a global descriptor. Moreover, as explained in the activity encoding, we use the training sequences for modelling a PCA dimensionality reduction.

Using a pose sequence as descriptor of the activity, we opted for an implementation of the Fuzzy Discrete HMM (FDHMM) [Uguz et al., 2008] which, as shown later, allows a stable training with scarce information preserving as discriminative information as possible by exploiting a soft-assignment of the observations in a clustered pose space.

4.2.1 Soft-assignment



(a) (b) (c)

Figure 4.2: Different observation parametrizations. Continuous using GMM (a). Discrete (b). Soft-assignment(c).

After applying the PCA dimensionality reduction, the pose descriptors are codified as D-dimensional real vectors. Afterwards, we model an activity class with a HMM where each state needs a model of the observation probability distribution in the space $\mathbb{R}^D$. We are dealing with a situation where a limited number of labelled sequences are available for training and therefore, scarce number of pose examples are available. In this case both classical approaches, continuous and discrete HMM, suffer from limitations in the training.

Working with real values, it would seem reasonable to use the continuous-HMM designed to deal with this kind of data but the limited training data produces some unwanted results. Looking at the continuous example in Figure 4.2a we see red dots representing the training observations and two ellipses surrounding them representing a 2 Gaussians GMM that models the observation probability distribution. With an adequate number of training observations it is possible to train the GMM and to obtain a reliable distribution however, as we dispose of sparse information, the obtained model is over fitted and therefore unreliable. Human motion has a high variability and the probability that at least one pose in validation examples reach values far from the model is too high for a robust system as it would produce an almost zero likelihood in the Problem 1 of HMM explained in Section 4.1.

The discrete-HMM, on the other hand, constrains the freedom of a real vector to a specified dictionary (also named codebook or vocabulary). In Figure 4.2b we see the same red dots in the same data space represented in the continuous example, but in this case we force the observation to belong to a specific codeword of a designed codebook. A codebook is a quantization of the data space so a hard-assignment may suppose the loss of some information. On the other hand, this constraint may force the limited information to be more meaningful as similar data would belong to same codeword. In the figure, every codeword is represented with an ellipse, and the training observations in a specific state are used to obtain the observation probability distribution using the winner-takes-it-all rule, which means that one observation belong to the closest codeword. The advantage of grouping similar data is lost when training a PMF with limited data as there is a high probability that none of the training observations land on some codewords whereas some of the validation ones do. Again a high number of training examples would minimize this performance but, as this possibility is discarded, a different approach should be used. The effect of a zero likelihood is usually minimized using a regularization that assigns a minimum probability to every codeword, therefore there is not zero likelihood, but it is clear that with few examples and taking into account the high variability of human motions this improvement is limited and the information loss using a hard-assignment could be reduced using a soft-assignment approach.

Finally, in Figure 4.2c we can see a modification in the use of the observations for training and testing the HMMs based on a Soft-assignment. In this approach, we use a similar type of model as the discrete one, using the same codebook but applying it in a different way. Working with D-dimensional real vectors the winner-takes-it-all rule removes much of the position information, so we replace this rule with a soft-assignment. Every observation has a probability of belonging to every codeword depending on the distance to the specific codeword. With this modification we solve the problem of zero likelihood but also we assign a more reliable probability in the codewords because some examples may land in an unclear place that the winner-takes-it-all rule forces to a specific codeword and the fuzzy observation distributes among every codeword. It is worth noting that this is just an improvement and it is not possible to make a robust system with limited information.

Assuming an existing dictionary composed by a GMM, it is possible to assign a belonging probability from a sample to every Gaussian as shown in the posterior probability in Equation 3.3. The set of features $Q = \{\mathbf{q}_1 \dots \mathbf{q}_T\}$ is now the sequence of dimensionally reduced poses and the GMM $\lambda = \{\omega_k, \mu_k, \Sigma_k\}$ is the clustering describing these features space, where the weights are discarded.

The number of clusters K defines the size of the codebook (i.e., the number of observation symbols). If K is too small it might not explain some important differences that should be detected in the temporal features. A large K avoids this performance but in the limit, when the number of training samples is too small, the training of the GMM is not possible and we force that every sample constitutes a cluster c_k, being its model defined by its own value $\mathbf{m}_k$. The soft-assignment of a new observation $\mathbf{q}_t$ to every sample (cluster) is computed as:

$$s_{tk} = \frac{e^{-\|\mathbf{m}_k - \mathbf{q}_t\|^2}}{\sum_{i=1}^{K} e^{-\|\mathbf{m}_i - \mathbf{q}_t\|^2}} \tag{4.15}$$

After the assignment of a pose to the clusters we obtain a vector $\mathbf{s}_t = \{s_{t1} \dots s_{tK}\}$ where $\sum_{k=1}^{K} s_{tk} = 1$. Although it represents a real vector, a modification on the discrete-HMM assures a stable solution.

4.2.2 Fuzzy Discrete HMM (FDHMM)

Given a set of observation symbols $C = \{c_1 \dots c_K\}$, the sequence of observations, $\mathcal{O} = \{O_1, \cdots, O_T\}$, in a discrete-HMM corresponds to a hard-assignment, $\mathcal{O} = \{\mathbf{r}_1, \cdots, \mathbf{r}_T\}$, and the computation of Problems 1 and 3 of the HMM has been already explained in Section 4.1. In an FDHMM the sequence of observations corresponds to a soft-assignment, $\mathcal{O} = \{\mathbf{s}_1, \cdots, \mathbf{s}_T\}$ to a set of Gaussians $\lambda = \{\lambda_1 \dots \lambda_K\}$, and the observation model is defined as follows:

$$b_j(O_t) = \sum_{k=1}^{K} s_{tk} m_{jk} \tag{4.16}$$

where $O_t = \{s_1 \dots s_K\}$ and $\sum_{k=1}^{K} m_{jk} = 1$.

The codebook creation should be addressed in advance in order to define the model, but for the time being we assume its existence, dealing with this issue in next section. On the other hand, the change of the observation model implies some modifications in the HMM algorithms.

We first define the vector $\mathbf{H}_k = \{h_1 \dots h_K\}$ where $h_i \in \{0, 1\}$ and $h_i = 1$ if $i = k$ and 0 otherwise. Therefore, $\mathbf{H}_k$ is a vector of zeros with a one in a specific element defined each time and $m_{jk} = b_j(\mathbf{H}_k)$.

The forward procedure is modified as follows:

1. Initialization:

$$\alpha_1(i) = \pi_i \sum_{k=1}^{K} s_{1k} b_i(\mathbf{H}_k) \tag{4.17}$$

2. Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] \left[\sum_{k=1}^{K} s_{(t+1),k} b_j(\mathbf{H}_k) \right] \tag{4.18}$$

3. Termination:

$$p(\mathcal{O}|\theta) = \sum_{i=1}^{N} \alpha_T(i) \tag{4.19}$$

The same is applied to the backward algorithm.

Additionally, the Baum-Welch algorithm is modified accordingly.

E-step: The computation of the $\xi_t(i,j)$ and $\gamma_t(i)$ values is performed with the new forward and backward procedures.

M-step: The computation of $\hat{\pi}_i$, $\hat{a}_{ij}$ values is unmodified while the computation of $\hat{m}_{jk} = \hat{b}_j(\mathbf{H}_k)$ suffers the following modification.

$$\hat{m}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j) \cdot s_{tk}}{\sum_{t=1}^{T} \gamma_t(j)} \tag{4.20}$$

In this sense, we change the winner-takes-it-all rule, given by r_{tk} in Equation 4.14, for a distributed probability of membership, given by s_{tk}. As $\sum_{k=1}^{K} s_{tk} = 1$, all the probability conditions are fulfilled and therefore, the convergence of the Baum-Welch algorithm to a local minimum is guaranteed.

4.2.3 Codebook creation

So far, an existing GMM has been taken for granted but, as its obtaining is a critical step when limited training examples are available, we define two codebook creation strategies. These two strategies define the difference between the two followed methodologies using FDHMM in the experiments.

Shared poses approach (FDHMM)

Inspired by codebook approaches in speech recognition, where one phoneme can be used more than once in a word and additionally different words can share several phonemes in common, we propose to use this idea in activity recognition changing phonemes by poses and words by activity examples. In a relaxed one-shot learning approach one activity example is available per class and we can transfer poses information among all the classes in order to construct the codebook by applying a clustering algorithm to all poses extracted form the examples. All these examples are initially used for training the PCA reduction that is subsequently applied to them. So, the input to the clustering algorithm is a set of vectors $Q = \{\mathbf{q}_i\} \ni \mathbf{q}_i \in \mathbb{R}^D$, where we assume that some activities are composed by poses repeated in different times, and there are some activities of different classes that share same poses. Although there is an information transferring among activity examples, all the examples belong to the same domain and therefore we call the method directly FDHMM.

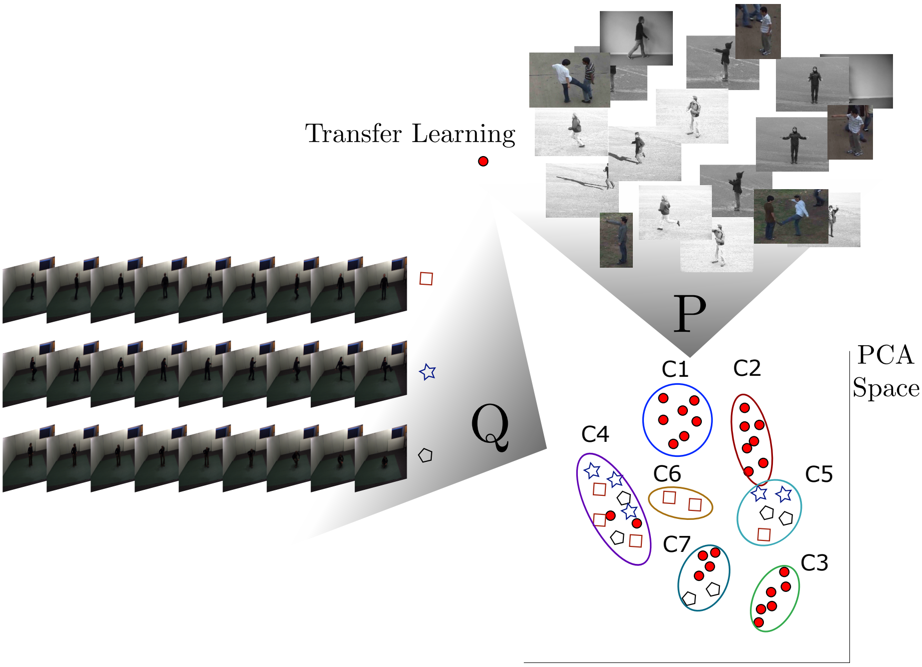


Figure 4.3: Codebook creation with Transfer Learning

Transfer learning approach (FDHMM+TL)

The limited number of labelled data for training gives rise to a sparse projected data space, which produces at least two problems: first, the data space is not conveniently modelled with the obtained clusters and second, the system needs several sequences in the initial training stage, although could be only one per class, in order to include the variability of the activities. The second problem implies a rigid system where the introduction of a new activity class may suppose the repetition of the complete training process including this new class.

Attending to the large amount of human activity videos already available in the Internet and the easiness to record the new ones, and the premise of existing many shared poses among activities we propose the extraction of poses descriptors in the source domain and the transferring to the target space using the PCA trained with Q. After this process we obtain a new set of vectors $P = \{\mathbf{p}_i\} \ni \mathbf{p}_i \in \mathbb{R}^D$, which we merge with the target domain ones, Q, creating an extended set $E = \{Q, P\}$ which is now clustered in order to obtain the codebook, as shown in Figure 4.3. The idea is to introduce poses information from source datasets, being any external video, so to make a general pose model which provides a freedom of introducing new classes without a clustering retraining. As the original approach is complemented with a transfer learning process we call the method FDHMM+TL.

Attending to the HMM recognition task and following [Pan and Yang, 2010] definitions of transfer learning, we propose an *inductive TL* stage where the task in the target domain is the classification while the data from the source domain is using only for clustering. However, the real use of the source domain information is the clustering in the target domain mixing unlabelled data from both domains, so, form this point of view we are proposing an *unsupervised TL* . Using [Cook et al., 2013] definition, our method is an IU transfer learning, although in the clustering poses form source and target domains are unlabelled.

4.3 Experiments and Results

FDHMM

We start the experiments with the Weizmann dataset, which is a small dataset, given the average length of activity datasets currently available. The authors of the dataset provide the mask of the silhouettes for all the activity videos obtained by background subtraction. We use these silhouettes for creating MHIs, as introduced in A.1.1, with temporal frame-windows of 5 frames, $N_l = N_w = 5$. The window stride is the minimum of one frame, $N_d = 1$. After applying the PCA, where we include a 85% of the data variance, to each MHI, and once a codebook has been created with the shared poses approach, each activity is codified by the soft-assignment of poses, $\mathcal{O} = \{\mathbf{s}_1, \cdots, \mathbf{s}_T\}$.

The experiments follow the one-subject-out model: sequences from one subject are selected for testing and *num* training sequences per class are randomly selected from the remaining subjects. The value of *num* goes from 1 to the maximum available sequences that in the Weizmann dataset case is 8. We compare the classification performance with several methods proposed in the literature in addition to our proposed framework based in FDHMM.

Dynamic Time Warping (DTW): DTW was firstly introduced by [Sakoe and Chiba, 1978]. Given two time series, $\mathcal{O}^1 = \{O^1_1, \cdots, O^1_{T_1}\}$ and $\mathcal{O}^2 = \{O^2_1, \cdots, O^2_{T_2}\}$, DTW aligns the two series so that their difference is minimized. DTW provides a way to align both temporal series by obtaining a warping path that has the minimum distance between both series. At the same time DTW provides a way to quantify the goodness of the matching by means of an accumulative cost along the warping path. The label for an unknown sequence $\mathcal{O}$ is estimated by assigning the label of the training sequence with the lowest accumulative cost.

Conditional Random Fields (CRF): Introduced in [Lafferty et al., 2001]. They are generative undirected graphical models which provide an estimated class per observations. In comparison to HMM, the estimation is not based only in the current observation but can also consider any combination of past and future observations. In this case, the CRF predicts labels for each observation in a sequence, not one for the entire sequence. Therefore, during evaluation using a Viterbi algorithm, the optimum path is found, and the label of the sequence is assigned to the most frequently occurring label. We have tested different long range dependencies, denoted by w. This parameter w accounts for the amount of past and future history to be used when predicting the state at time t.

Hidden Conditional Random Fields (HCRF): HCRF was first introduced by [Gunawardana et al., 2005] for phone-conversation/speech classification and has later been applied to gesture and object recognition in [Quattoni et al., 2007]. The authors argued that CRFs are limited in the sense that they cannot capture intermediate structures using hidden-sate variables. HCFRs use intermediate hidden variables to model the latent structure of the input domain. In a more recent work, [Zhang and Gong, 2010] have presented a method for action categorization with a modified HCRF. With a single HCFR it is possible to train a classifier where the class selection is performed by $\omega^* = \text{argmax}_{\omega}(p(\omega|\mathcal{O}, \theta))$. Where θ are the model parameters. We also conducted experiments that incorporated different long range dependencies in the same way done with CRF.

continuous-HMM: We train a continuous-HMM per class, where the observation PDF per state is computed with only one Gaussian so to limit the number of parameters to train. Therefore, the training with limited number of training examples is eased.

discrete-HMM: We train a discrete-HMM per class. We solve the effect of a zero likelihood produced in some clusters when training the PMF with limited data by regularizing the minimum value of the PMF. The minimum value in any cluster is at least a threshold which is obtained in relation to the number of words K in the codebook, $th = 1/(100K)$.

FDHMM: Finally we tested our framework with a FDHMM per class.

The HMM approaches are 2 states models with an ergodic transition matrix

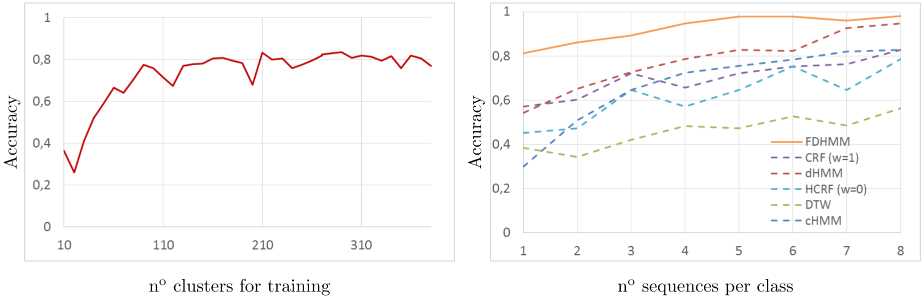


Figure 4.4: (Left) Average accuracy (over 93 test sequences of the Weizmann dataset) provided by FDHMM approach with a single training sequence per class for different number K of clusters. (Right) Average accuracy for all methods obtained by increasing the number of training sequences.

randomly initialized and the initial probabilities $\pi_1 = 1$ and $\pi_2 = 0$.

As shown in Figure 4.4 (left), we ran several experiments with FDHMM using different number of clusters, starting from a single training sequence per class. The number of training poses in different trials were around 400, so the maximum number of clusters is limited to this amount. The graph shows how the increase of the number of clusters has the tendency to improve the accuracy, although after reaching around 100 clusters the improvement is slight. After realizing this performance, and in order to introduce a FDHMM proposal free of any parameter and independent of any clustering process, we carry out the next experiments with the number of clusters K exactly equal to the number of samples obtaining a 81.1% of accuracy which is among the best results in the previous experiments, being the maximum 84.3% for $K = 290$. For a K equal to the number of samples, the soft-assignment is performed with Equation 4.15.

Figure 4.4 (right) shows the average accuracy over 93 testing sequences in Weizmann dataset using the literature methods DTW, CRF, HCRF, continuous-HMM and discrete-HMM in comparison to our proposal using FDHMM. The *num* training sequences per class goes from 1 to 8. Our method converges to

Table 4.1: Comparisons of recognition performance (average percentage accuracy) for action recognition using only 1 sequence for training

Models	DTW	cHMM	dHMM	CRF (w=0)	CRF (w=1)	CRF (w=2)	HCRF w=0)	HCRF (w=1)	HCRF (w=2)	FDHMM
Avg.(%)	38.4	29.9	54.2	54.8	57.0	55.9	45.1	40.8	34.4	81.1

an average accuracy maximum value of 98% with only 5 training sequences. As it can be noticed in Figure 4.4 (right), no other method outperforms FDHMM using up to 8 sequences. It is worth noting how HMM methods exhibit the best improvement when the number of sequence increases, dHMM reducing the gap with FDHMM till a close result while increasing the number of training sequences and cHMM reducing the gap from being the worst method with one example to tie in the third position using 8 sequences.

Focusing our experiments on using one training sequence per class, Table 4.1 shows the average accuracy for all methods, where the CRF and HCRF have been trained with different values for the long rage dependencies parameter w. As mentioned before, this parameter takes into account the amount of past and future samples used when predicting the state at a particular time. Clearly our approach provides the best result with just one training sequence per class, i.e., 81.11%, followed by CRF (w=1) with 56.99%.

FDHMM+TL

So far, we have seen how a relaxed one-shot learning activity recognition method can be successfully implemented using FDHMM. However, the need of one activity example per class constrains the method. We deal with this issue by using a transfer learning stage in the clustering process named FDHMM+TL.

For the new experiments we have selected the IXMAS dataset as target domain, specifically only the camera 1. By choosing only one camera we simulate

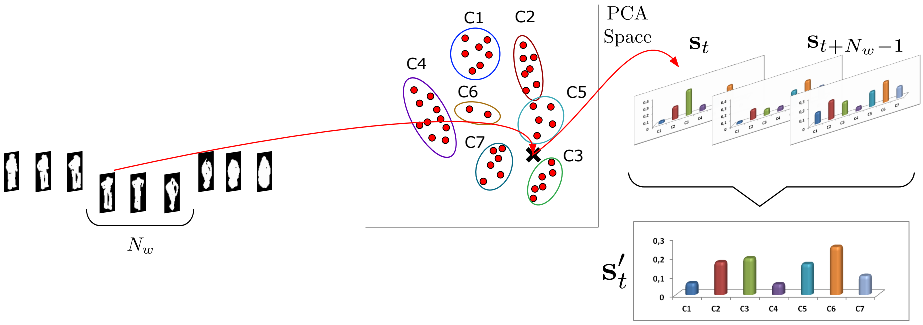


Figure 4.5: Mean filter in a window of N_w soft-assigned silhouette frames.

a fixed camera scenario. Following the authors of the IXMAS dataset suggestion we discard two of the actors due to their irregular performance, so we dispose of 27 repetitions per action which is a much larger dataset than Weizmann. Additionally, we discard 2 classes suggested as well by the authors so we dispose of 11 actions.

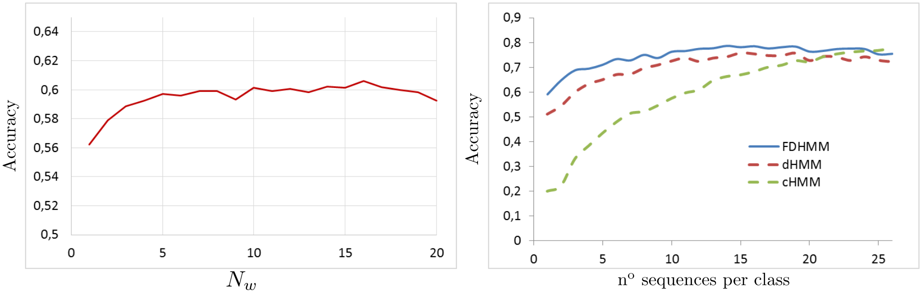
In these experiments we simplify the descriptor extraction by using the silhouettes directly, however we have visually realized how some of the extractions fail significantly, thus the $p(\mathcal{O}|\theta)$ computation in a HMM might be highly unstable due to spurious. Therefore, we apply a mean filter to the soft-assignations $\mathbf{s}_t$ as shown in Figure 4.5 and detailed in Equation 4.21.

$$s'_{tk} = \frac{\sum_{j=t}^{t+N_w-1} s_{jk}}{N_w} \tag{4.21}$$

As MHI case, $N_l = N_w$ and the window stride $N_d = 1$.

The initial experiments are carried out over 5 diverse classes of IXMAS (sit down, walk, wave, punch and kick) so as to recreate a scenario with few initial training classes and then we can simulate later the inclusion of 6 new classes not used in the codebook and PCA training.

As we have changed the dataset, we first apply directly the FDHMM frame-

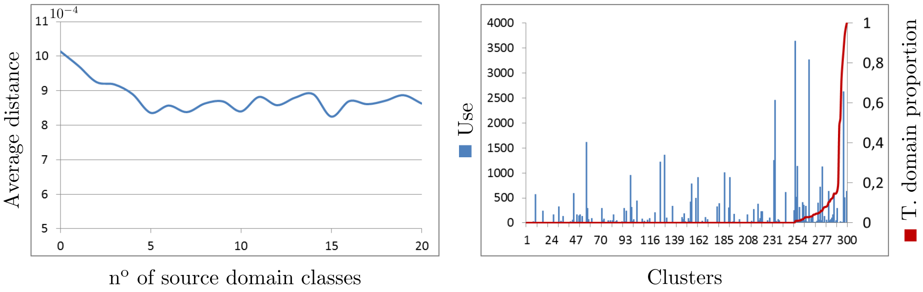


(a) (b)

Figure 4.6: (a) Accuracy while increasing the window size. (b) Comparison of cHMM, dHMM and FDHMM increasing the training sequences.

work without transfer learning. Results shown in Figure 4.6a are used for selecting the window size. We can observe how the use of a window improves the results over using directly one frame, and how after reaching $N_w = 5$ there is not clear improvement. Although FDHMM has been already proved as a good method for training limited labelled datasets, we carried out a new experiment, shown in Figure 4.6b, comparing its results with cHMM and dHMM. In this comparison FDHMM and dHMM use the same codebook. The interest of this experiment lies in the large number of training sequences available in IXMAS comparing with Weizmann in addition to the confirmation of the method performance. We observe that FDHMM improves both methods when training with few sequences as expected from the previous experiments and as previously suspected cHMM overtakes it when using a lot of them (over 24 for this experiment). On the other hand, dHMM appears to maintain an offset with respect to FDHMM after initially reduce the gap.

We continue the experiments evaluating the transfer learning improvement using IXMAS dataset as target domain and ViHASi dataset as source domain. The ViHASi dataset has been virtually created, see Section 2.8, so if necessary it would be easy to implement new examples.



(a)

(b)

Figure 4.7: (a)Average distance between estimated and actual pose vectors. (b) Use of clusters ordered by theirs domain proportion.

We first design an experiment that provides an idea of how well the clustering after the transfer learning represents the poses space. Considering the original pose vector $\mathbf{e}_t$ (column vector), we can compare it with a reconstruction based on its soft-assignment representation, $\mathbf{s}_t$. If $\mathbf{s}_t$ is considered a sparse coding of $\mathbf{e}_t$, having $\mathbf{m}$ a matrix which columns represent the cluster centroids, we can estimate the original pose vector with Equation 4.22, which is a linear combination of the cluster means

$$\widehat{\mathbf{e}}_t = \mathbf{m}\mathbf{s}_t \tag{4.22}$$

We use the euclidean distance between the estimated pose vector $\widehat{\mathbf{e}}_t$ and the actual pose vector $\mathbf{e}_t$ to estimate how good the codebook is. We have carried out an experiment evaluating the average distance in all the IXMAS pose vectors but those from the 5 initial IXMAS sequences. We have trained the PCA and the codebook with the initial sequences. The codebook has been trained with one random sequence per IXMAS class and adding each time from the ViHASi dataset all the sequences from a new class up to 20.

The average distance results using up to 300 clusters are depicted in Figure

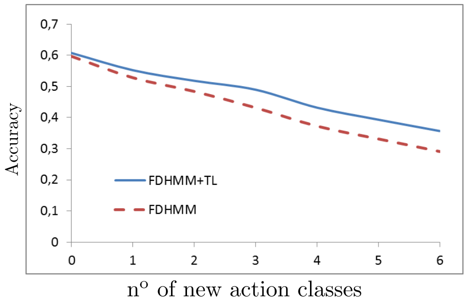
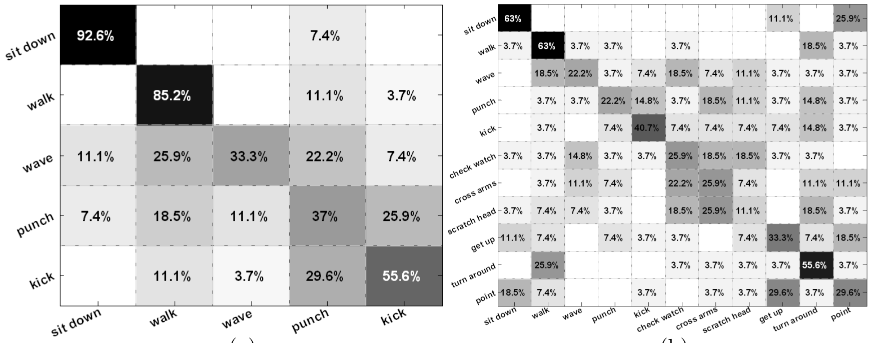


Figure 4.8: Accuracy increasing the number of classes not used in the codebook.

4.7a where we observe how the inclusion of source actions decreases the distance (improving the representation), but after 5 classes it does not decrease or even increases. There are several possible reasons, like the inclusion of not representative poses, but also the use of the same number of clusters when increasing the number of source data can imply a worse clustering and then a worse representation. In Figure 4.7b we show how the clustering is conformed and used. From the previous experiment we select the case with 7 actions from ViHASi (collapse, grenade, hero door slam, jump kick, punch, walk and walk turn 180). In the graph we observe the result of obtaining the Fuzzy Observation from the 6 actions remaining unused in IXMAS, which are composed by 43728 poses. The bars graph shows how each cluster is activated by the soft-assignments of those poses, and we have ordered them attending their proportion of target and source domain composition. We can observe how the most used clusters are found along all proportions and we can not conclude any correlation between use and proportion.

Finally we want to compare the recognition rate between the use of Transfer Learning and the original FDHMM, where the only difference is the codebook. Comparing both cases when recognizing the 5 target sequences of IXMAS we obtain a slight difference between the use of Transfer Learning (FDHMM+TL) and the direct use of FDHMM. In the first case we obtain a 61% of recognition,



	sit down	walk	wave	punch	kick
sit down	92.6%			7.4%	
walk		85.2%		11.1%	3.7%
wave	11.1%	25.9%	33.3%	22.2%	7.4%
punch	7.4%	18.5%	11.1%	37%	25.9%
kick		11.1%	3.7%	29.6%	55.6%

(a)

	sit down	walk	wave	punch	kick	check watch	cross arms	scratch head	get up	turn around	point
sit down	63%								11.1%		25.9%
walk	3.7%	63%	3.7%	3.7%		3.7%				18.5%	3.7%
wave		18.5%	22.2%	3.7%	7.4%	18.5%	7.4%	11.1%	3.7%	3.7%	3.7%
punch		3.7%	3.7%	22.2%	14.8%	3.7%	18.5%	11.1%	3.7%	14.8%	3.7%
kick		3.7%		7.4%	40.7%	7.4%	7.4%	7.4%	7.4%	14.8%	3.7%
check watch	3.7%	3.7%	14.8%	3.7%	3.7%	25.9%	18.5%	18.5%	3.7%	3.7%	
cross arms		3.7%	11.1%	7.4%		22.2%	25.9%	7.4%		11.1%	11.1%
scratch head	3.7%	7.4%	7.4%	3.7%		18.5%	25.9%	11.1%		18.5%	3.7%
get up	11.1%	7.4%		7.4%	3.7%	3.7%		7.4%	33.3%	7.4%	18.5%
turn around		25.9%				3.7%	3.7%	3.7%	3.7%	55.6%	3.7%
point	18.5%	7.4%			3.7%		3.7%	3.7%	29.6%	3.7%	29.6%

(b)

Figure 4.9: (a) Confusion Matrix with 5 classes and (b) with 11 classes

and with FDHMM a 59%. From this results we might conclude that the transfer learning only complicates the process without any benefit. But the strongest point of the method is the availability to train new sequences using the original codebook. Using the remaining 6 actions in IXMAS, we include one by one and we test the results in both methods, shown in Figure 4.8. Although the starting point of both methods is close, when the number of actions increase we detect how the lines separate, obtaining FDHMM+TL the best results. The method is a relaxed one-shot learning approach but, on the other hand, it is more constrained and more similar to a strict approach.

In Figure 4.9 we show the confusion matrices using Transfer Learning in two cases, classifying 5 and 11 IXMAS classes respectively. We can see how the initial classes are worse recognized when including new classes as expected, but in Figure 4.9b we can observe as well how the initial five classes keep a better recognition rate in comparison to the new ones. So, the average recognition rate of the initial 5 classes shown in the (b) confusion matrix is a 42% while the overall recognition rate is close to 35%. This is explained considering that the codebook includes points from those initial classes.

4.4 Conclusions

This preliminary approach validates the use of FDHMM applied to a relaxed one-shot learning and the Transfer Learning stage shows how the enrichment of the information improves the performance. FDHMM is a stable solution for training a HMM with limited training examples. On the other hand, the use of source domain information improves the pose space modelling and consequently the accuracy of the classification. However, some issues arise and should be addressed.

The first issue in the previous approach is the feature extraction. We have dealt with two different configurations, MHI and mean filter of silhouettes, both using extracted silhouettes from rgb images. The extraction of robust silhouettes is difficult in many scenarios and it is highly dependent to background changes, illumination changes and occlusions. Additionally, they have been clearly overtaken by local spatio-temporal features, much more stable in the mentioned situations. As the experimentation presented in this section is a preliminary approach, we have not completed a comparison between MHI and the mean filter of silhouettes as both are currently overtaken by other approaches and the needed time to implement the experiments is not worthy. For instance, Improve Dense Trajectories are currently the state-of-the-art [Wang and Schmid, 2013] obtaining far better results. However, the change to IDTs implies a different type of encoding which is not directly compatible with FDHMM as the soft-assignment of a global feature is. We explain a different approach in Chapter 5 which deals with the issue.

On the other hand, the transfer learning proposed lies in the manually selected source domain. This approach is useful to corroborate the enhancing produced with a transfer learning but it lacks of generality. Moreover, it would be better to make a smart domain transformation of the transferred data instead of the direct projection.

Furthermore, several activity examples are used for the codebook creation and even with the transfer learning five initial examples are proposed in order to obtain a meaningful codebook. It would be interesting to reduce this to only one, so from the first recorded sample it is possible to train the method, and therefore to apply a strict one-shot learning method.

As with the comparison between descriptors, the experiments are not completed using both datasets in all the approaches. However, the results are convincing enough to assure the previous conclusions and they lead us to look for a more robust approach. Thus, the solution proposed in Chapter 5 looks forward to the solving of the arisen issues.

This preliminary work were presented in two communications [Orrite et al., 2011] and [Rodriguez et al., 2013a].

5

Strict One-Shot Learning of Activities with an MAP adapted GMM and Simplex-HMM

In the previous chapter we have presented some advances using a solution of HMM for limited training sequences where, thanks to a transfer learning stage, we have reduced the number of activity examples needed in a relaxed one-shot learning framework. We can extract two important ideas from these methods: first, a properly designed HMM can be stably trained with limited training examples and second, the existing activity videos available on-line can enrich the system minimizing the target scenario examples needed. In this chapter we constrain the method to a strict one-shot learning of human activities framework progressing in both ideas. In this regard, we have designed a novel framework which only uses one sequence for activity representation. To carry out with this restriction we propose the transfer learning of a Universal Background Model (UBM), as done in [Reynolds et al., 2000]. The UBM is trained with features extracted from extensive datasets available on-line and later transferred to the target scenario where only one, or just a few, labelled sequence is available. The transferred vocabulary is used for a sequence of BoF encoding which feeds the

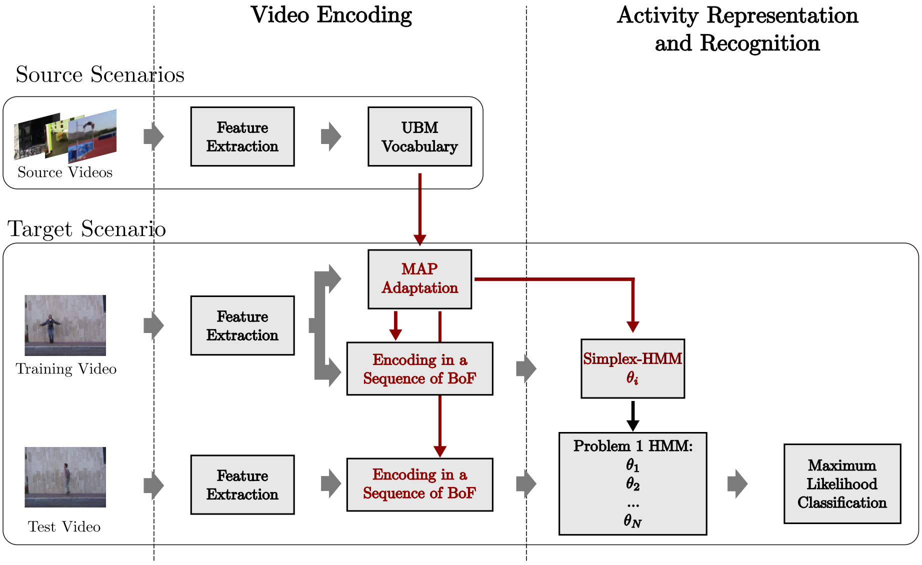


Figure 5.1: Flow diagram of the proposed approach, highlighting in red the stages where the main novelties are introduced.

training and testing of a new specially designed case of a HMM which provides a stable solution with limited training data.

A flow diagram of the proposed approach is depicted in Figure 5.1. From the wide range of features extractors available in the literature, IDT [Wang and Schmid, 2013] have shown state-of-the-art performance in several challenging datasets and so we use them in the Feature Extraction stage. The proposed method extracts the IDTs from videos in public datasets of human activities, considered the source domain, and creates a UBM vocabulary modelled with a GMM as done by [Reynolds et al., 2000], representing general, person and scenario independent features. Unlike [Reynolds et al., 2000], the trained UBM represents a universe of features, and not a universe of activities (or speakers in their case). Once selected the target scenario, an initial labelled training video is recorded. The corresponding IDTs are extracted from this video and used in a twofold task. First, with the unordered IDTs, the UBM vocabulary is transferred to the target scenario using a Maximum a Posteriori (MAP) Adaptation, and obtaining a sequence specific vocabulary. Second, the IDTs are grouped into temporal windows where they are soft-assigned to the adapted vocabulary, obtaining a BoF per window. The BoF histogram is normalized so that it sums one, equivalent to say it belongs to a unit simplex. This way, the video is encoded as a sequence of BoF, and the activity is then modelled with an HMM which is a well known generative approach applied on time series.

The EM algorithm is used to solve the difficult task of estimating a HMM. However, as mentioned before, EM has some complications. EM does not guarantee reaching a global optimum, and local optima can at times be problematic, particularly with small training sets. Moreover, small training sets produce overfitting as we increase the number of parameters to train and the usual continuous HMM approach that incorporates GMM distributions may lead to an unstable EM algorithm [Shinozaki and Ostendorf, 2008]. In order to obtain a reliable system, the proposed Simplex-HMM is numerically stable, even with a single training sequence. Besides, the soft assignment that leads to BoF seems more suitable than a hard assignment for the case of scarce training data.

Testing follows a process flow similar to training. First the IDTs are extracted and a sequence of BoFs is obtained using a temporal sliding window. The encoded video is then evaluated, given the Simplex-HMM, as described in Section 4.1 for the Problem 1. Maximum Likelihood Classification is used to identify the model that fits better the observation and therefore to assign an activity label. Using the defined one-shot learning system we have obtained state-of-the-art results in the public datasets Weizmann, KTH and IXMAS.

The two main contributions of this framework are summarised below:

1. A video encoding based on Transfer Learning where a UBM vocabulary is obtained training a GMM with features from sequences in the source

domain, and adapted to the feature space extracted from the target scenario by a MAP adaptation of the GMM, conforming a target domain vocabulary.

2. The definition of an HMM constrained to a sequence of vectors in a Simplex (Simplex-HMM), avoiding the numerical problems produced in the HMM training with scarce data.

5.1 Video Encoding

Following the diagram depicted in the *Video Encoding* section of Figure 5.1 we can observe that video encoding comprises two different tasks. Firstly, a UBM is modelled by a GMM using source videos, widely available, and afterwards, the adaptation of this UBM-GMM takes place on the target scenario. Secondly, as explained in Section A.3.1 a BoF video encoding looses the temporal information of the activity and therefore, a temporal sliding window is used to recover this kind of information.

5.1.1 Transfer Learning with MAP adaptation

The need of a codebook trained with scarce data has been shown a difficult task and the use of only one activity example is clearly insufficient. The naive Transfer Learning previously proposed is now replaced with a MAP adaptation of the source domain information.

As mentioned before, although few samples are available in the target domain, plenty of videos can be obtained on-line as source domain from where the learning is transferred to the target domain. Some machine learning methods have used this approach being the speaker verification systems based on UBM-GMM especially successful in this regard. The improvement in activity recognition systems using target domain information that complements the

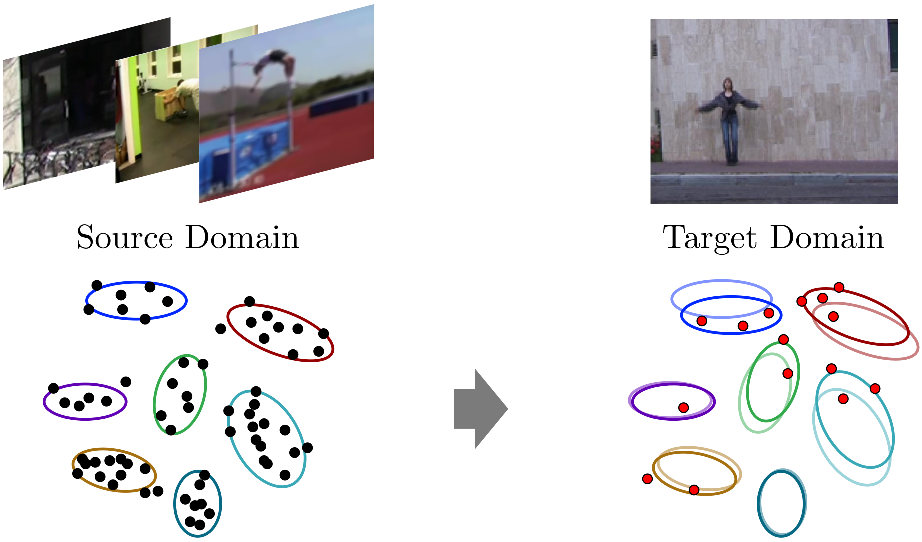


Figure 5.2: Source GMM and MAP adapted GMM

source domain information has been proven in previous researches being [Cao et al., 2010] and our proposal [Rodriguez et al., 2013a] two of them.

Figure 5.2 represents the proposed Transfer Learning process showing a simplified 2D GMM trained in the source domain and adapted to the target domain. From the source domain a large number of IDTs is randomly selected from the extracted ones, $\mathbf{P} = \{\mathbf{p}_j\}$, $\mathbf{p}_j \in \mathbb{R}^D$, being unlabelled data, and then used in a EM process to train the GMM which represents the UBM. The general model of GMM supports full covariance matrices, but diagonal covariance matrices can satisfactorily approximate the original density modelling with a higher order GMM and they are computationally more efficient. Therefore, the framework uses diagonal covariance matrices and in addition it disregards GMM weights obtaining a simplified model $\lambda = \{\mu_i, \Sigma_i\}$. This UBM is later MAP adapted to the target domain using only the available samples in this domain, they can be as few as the extracted from a single sequence, $\mathbf{Q} = \{\mathbf{q}_j\}$, $\mathbf{q}_j \in \mathbb{R}^D$. For Gaussian λ_i in the UBM, the probabilistic alignment of the feature vectors is computed with Equation 5.1.

$$p(\lambda_i|\mathbf{q}_j) = \frac{\mathcal{N}(\mathbf{q}_j|\mu_i, \Sigma_i)}{\sum_{k=1}^{K} \mathcal{N}(\mathbf{q}_j|\mu_k, \Sigma_k)} \tag{5.1}$$

These probabilistic alignments and the features vectors are used to compute the sufficient statistics of the mean with Equations 5.2 and 5.3. Weights have been disregarded so every Gaussian has the same weight, and covariance adaptation has been proven to be dispensable in most of the systems, so the system keeps the original covariance matrices in every new adapted GMM.

$$n_i = \sum_{j=1}^{M} p(\lambda_i|\mathbf{q}_j) \tag{5.2}$$

$$E_i(\mathbf{Q}) = \frac{1}{n_i} \sum_{j=1}^{M} p(\lambda_i|\mathbf{q}_j)\mathbf{q}_j \tag{5.3}$$

The sufficient statistics, computed with the target domain training data, are used to update the UBM, estimating the new means with Equation 5.4.

$$\widehat{\mu}_i = \alpha_i E_i(\mathbf{Q}) + (1 - \alpha_i)\mu_i \tag{5.4}$$

The parameter α_i $(0 \leq \alpha_i \leq 1)$ is an adaptation coefficient controlling the balance between old and new estimates and can be obtained through Equation 5.5

$$\alpha_i = \frac{n_i}{n_i + rM} \tag{5.5}$$

where r is the controlling variable for adaptation and the rM term assures an equal adaptation independent on the number of IDT samples per example. After the MAP adaptation, a new GMM is obtained per video activity, $\widehat{\lambda} = \{\widehat{\mu}_i, \Sigma_i\}$,

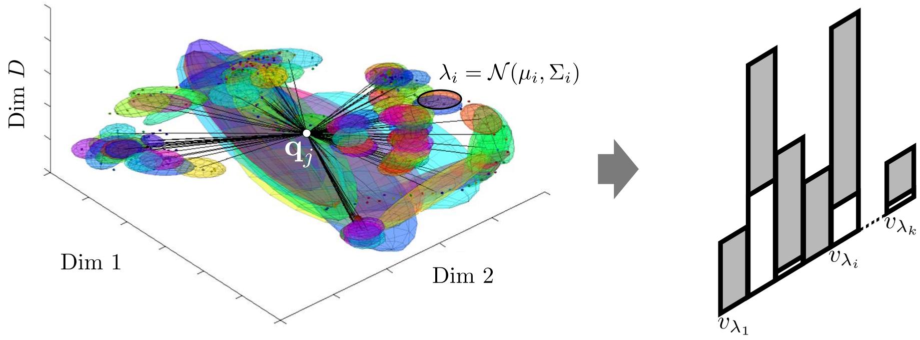


Figure 5.3: GMM representation and Soft-assignment-BoF. Grey bars represents the BoF while white highlighted bars represents the $\mathbf{q}_j$ sample contribution.

representing the new codebook used in the encoding.

5.1.2 Temporal Windowed Soft-assignment-BoF

Over the past several years, many methods have modelled activities by encoding the extracted features in a single BoF, obtaining the codebook bins through a clustering algorithm of the training samples. Two of the most common clustering algorithms are the ones explained in Section A.2, K-means and GMM. The former is defined only by the mean while the latter encodes second order information as includes both, the mean and the covariance and even a weight of the cluster. The proposed encoding uses the IDT features extracted from the activity videos and models the features space through the simplified GMM, $\widehat{\lambda} = \{\widehat{\mu}_i, \Sigma_i\}$. The number of clusters K can vary a lot in different approaches and empirically has been proven that a large number, in the order of thousands, is appropriate for BoF encoding of local features, see [Wang et al., 2013].

From each video activity, a set of IDT feature vectors $\mathbf{Q} = \{\mathbf{q}_j\}$, $\mathbf{q}_j \in \mathbb{R}^D$ is extracted. After using them for MAP adapt the UBM, each sample is soft assigned to the Gaussians using the Equation 3.3, which coincides with Equa-

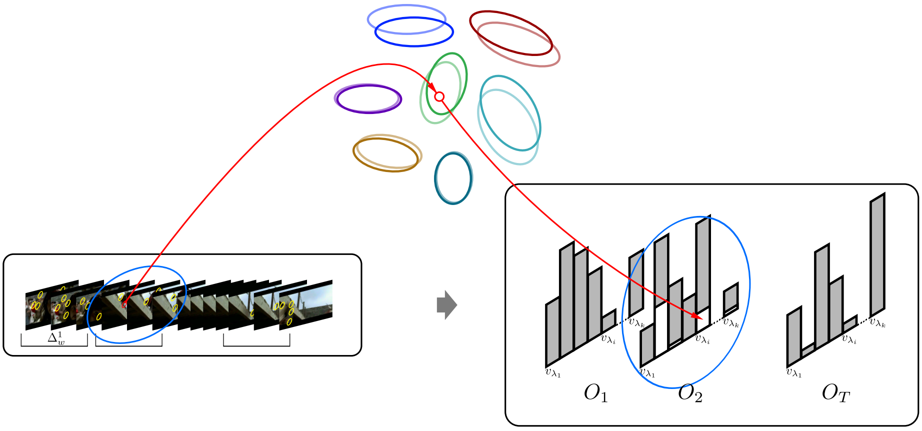


Figure 5.4: BoF sequence of Windowed video

tion 5.1. Figure 5.3 shows a sample evaluated in every Gaussian of the GMM used to encode the data in a Soft-assignment-BoF. With many feature samples, the proposed soft-assignment is unnecessary and the winner-takes-all rule usually applied in BoF approaches is sufficient. However, the one-shot learning objective is to obtain a representative model with only one activity example which contains few feature samples, therefore keeping as much information as possible, as a proper soft-assignment does, is essential. Moreover, in order to keep the long-term temporal information of the activity, a temporally windowed Soft-assignment-BoF encoding is proposed, as shown in Figure 5.4.

Every N_d frames a new BoF is obtained in a window of N_w frames, keeping the long-term temporal information in a sequence of BoF, $\mathcal{O} = \{O_1, \cdots, O_T\}$. We use the encoding explained in Section A.3.1, using IDT as descriptors which are computed through a temporal window of length N_l, generally different to N_w. Each IDT, $\mathbf{q}_j$, has associated a temporal window Δ_l^j and influences proportionally to each Δ_w^t window given the Equation 3.1. Each bin value, $v_{\lambda_i}^t$, associates to a specific BoF, O_t, is then calculated using Equation 5.6

$$v^t_{\lambda_i} = \frac{1}{\sum_{j=1}^{M} \rho_{jt}} \sum_{j=1}^{M} \frac{\rho_{jt} \mathcal{N}(\mathbf{q}_j | \widehat{\mu}^t_i, \Sigma_i)}{\sum_{k=1}^{K} \mathcal{N}(\mathbf{q}_j | \widehat{\mu}^t_k, \Sigma_k)} \tag{5.6}$$

Thanks to the applied normalizations every Soft-assignment-BoF, we denote as $BoF(t)$ in Equation 3.2 and in general denoted as O_t, belongs to the unit simplex, $\Delta = \{\mathbf{v}^t \in \mathbb{R}^K : v^t_{\lambda_i} \geq 0 : \sum_{k=1}^{K} v^t_{\lambda_k} = 1\}$.

5.2 Activity Representation and Recognition using Simplex-HMM

Given an activity video, the encoding represents the activity as a sequence of normalized BoF, $\mathcal{O} = \{O_1, \cdots, O_T\}$, each one belonging to the unit simplex Δ. These observations are $\mathbb{R}^K$ vectors although the real dimensionality of the space is $(K-1)$. In Figure 5.5a, a simplex of 3 dimensions is represented, and it can be observed that it is a triangle in a plane, so in reality it has only 2 dimensions. The *Activity Representation and Recognition* step shown in Figure 5.1 depicts the flow chart of a classifier where the sequence of normalized BoF is used to feed a Simplex-HMM, which is explained later in this section.

An HMM is formally represented by the parameters $\theta = \{N, A, B, \pi\}$, as explained in Section 4.1. In the simplex space, the emission probability can be modelled for instance with a Dirichlet distribution. However, the high dimensionality of the space causes numerical problems [Minka, 2009]. To exemplify these problems, we consider the simple case of a uniform distribution. In Equation 5.7 we observe how a uniform distribution would require the PDF $f(\mathbf{x}) = (K-1)!$, $\mathbf{x} \in \Delta$, so that the integral in the simplex is 1. So, with a high K any PDF is numerically infeasible.

$$\int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-\sum_{i=1}^{K-2} x_i} (K-1)! d_{x_{K-1}} \ldots d_{x_2} d_{x_1} = 1 \tag{5.7}$$

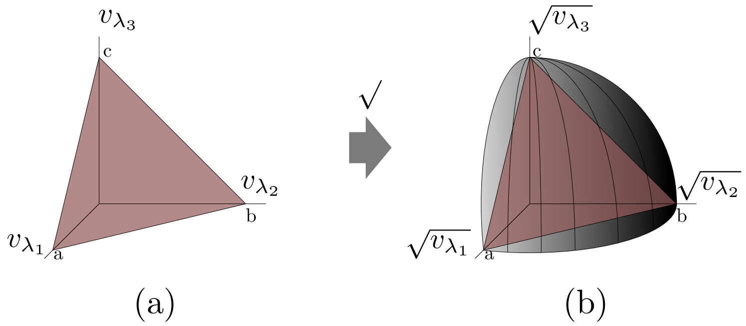


Figure 5.5: (a) 3-dimensional simplex representation and (b) unit sphere portion encompassing the square root transformation of the simplex.

The emission probability can be modelled in the $\mathbb{R}^K$ space, where there is no variance in the perpendicular dimension to the plane, or in a $\mathbb{R}^{K-1}$ space obtained for instance by performing the Aitchison's solution to the compositional data. However, both cases suffer for the same problem produced when samples have a high dimensionality and their number is limited for training. The lack of available data produces overfitting of the parameters, and the high dimensionality of the data intensifies the problem, so the training of any sort of parameter related to covariance is hopeless. We have corroborated this performance with some preliminary experiments with the continuous HMM and one Gaussian per state, obtaining numerical problems as the log-likelihood has always gone to $-\infty$.

Making the assumption of having an emission function that does not represent a PDF but is numerically manageable, we have found a stable solution in spite of the high dimensionality. We define the function having a maximum in a specific point of the simplex equal to 1, and decreasing its value while distancing from this point, but always with a positive or zero value in the simplex. Doing this, the training parameters of the B function are reduced to only a $\mathbb{R}^K$ vector. This assumption defines a generic Simplex-HMM approach but, in order to apply it, we need to define this function.

There are infinity functions that fulfil the requirements from where we have selected one with which we study the validity of the framework. Specifically, we simplify the observation model by defining the exponential of the Euclidean distance between a mean vector and the observations. Thus, we apply the observation model defined in Equation 5.8.

$$b_j(O_t) = e^{-\varphi\sqrt{\sum_{k=1}^{K}(v^t_{\lambda_k}-m_{jk})^2}} \tag{5.8}$$

Equation 5.8 shows φ and m_{jk} as free parameters but, as there are few samples in training and it is important to reduce as much as possible the parameters to learn, we experimentally fix the value of φ.

Considering that each $v^t_{\lambda_k}$ is an element in a histogram and represents the frequency of a specific feature model, it is possible to appreciate a drawback in the direct use of the Euclidean distance. Considering two normalized histograms of the same dimensionality, $A = \{a_i\}$ and $B = \{b_i\}$, $1 \leq i \leq N$ and $\sum_{i=1}^{N} a_i = \sum_{i=1}^{N} b_i = 1$, if $a_j = 1$, so $a_i = 0, \forall i \neq j$, and $b_j = 0$, then the dissimilarity between A and B should be maximum independent on the values of $b_i, \forall i \neq j$, and then the distance from A to B should remain constant for all values of $b_i, \forall i \neq j$. In Figure 5.5a this distance would be represented by the distance from point a to any point in segment bc, which is not constant and the middle point of the segment is closer to a than the edges, being more significant with high dimensionality. To tackle this drawback, we propose to replace the Euclidean distance with the Hellinger distance, which is equivalent to transforming the points in the simplex to a portion of a hypersphere of unit radius by applying the square root of the vector element ($v^t_{\lambda_k} \to \sqrt{v^t_{\lambda_k}}$), as shown in Figure 5.5b. The Hellinger distance implies the same transformation in the free parameters ($m_{jk} \to \sqrt{m_{jk}}$). However, for the sake of simplicity we do not impose the condition that $\sum_{k=1}^{K} m_{jk} = 1$ in the optimization process so we avoid the square root notation without loss of generalization. Moreover, we can change the notation in variables $v^t_{\lambda_k}$ avoiding again the square root by imposing the condition of $\sum_{k=1}^{K} v_{\lambda_k}^2 = 1$ for Hellinger distance instead of $\sum_{k=1}^{K} v_{\lambda_k} = 1$ for the Euclidean distance. These changes in notation make the formulation equivalent in both Euclidean and Hellinger distances. In order to differentiate both observation models we name them accordingly: EOM, for the Euclidean observation model, and HOM, for the Hellinger observation model.

Although the function $b_j(O_t)$ prevents the numerical problems, it is not a PDF and some characteristics should be analysed for validate its utility. First, if $b_j(O_t)$ is not a PDF then equation 4.4 is not a probability and then we can not assure that the comparison made in a maximum likelihood classifier is valid. In spite of this issue, we proved its validity experimentally but as it is not a likelihood we call it pseudo-likelihood. On the other hand, we need to assure the adequate training of the model, so we prove the EM convergence to a local maximum in the Annex B.1.

The EM algorithm is adapted to this new emission model. As explained in Section 4.1, the maximization of the likelihood function is performed by maximizing the Baum's auxiliary function $Q(\hat{\theta}, \theta)$, equation 4.9. This maximization is performed in an iterative process composed by two steps.

E-step: This step implies the calculation of functions $\xi_t(i,j)$ and $\gamma_t(i)$. Using the proposed observation model $b_j(O_t)$, the equations are the same explained in the original case.

M-step: This process calculates the $\hat{\pi}_i$, $\hat{a}_{ij}$ and $\hat{m}_{jk}$ that maximize $Q(\hat{\theta}, \theta)$. The optimizations of $\hat{\pi}_i$, $\hat{a}_{ij}$ and $\hat{m}_{jk}$ are obtained by maximizing Equation 4.9. In our method the optimizations of $\hat{\pi}_i$, $\hat{a}_{ij}$ are the same as in the traditional method but $\hat{m}_{jk}$ has a specific computation. Equation 5.9 has to be maximized with respect to m_{jk}:

$$\sum_t \sum_j \gamma_t(j) \ln(b_j(O_t)) = \sum_t \sum_j \gamma_t(j) \left(-\varphi \sqrt{\sum_{k=1}^{K} (v_{\lambda_k}^t - m_{jk})^2} \right) \tag{5.9}$$

By setting $\frac{\partial}{\partial m_{jk}} = 0$, the following equation is obtained:

$$\varphi \sum_{t=1}^{T} \gamma_t(j) \frac{(v^t_{\lambda_k} - m_{jk})}{\sqrt{\sum_{k'=1}^{K} (v^t_{\lambda_{k'}} - m_{jk'})^2}} = 0 \tag{5.10}$$

Since m_{jk} does not depend on t and $\gamma_t(j)$ are treated as constants in the M-step once computed in the E-step, equation 5.11 can be easily derived:

$$m_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j) \frac{v^t_{\lambda_k}}{\sqrt{\sum_{k'=1}^{K} (v^t_{\lambda_{k'}} - m_{jk'})^2}}}{\sum_{t=1}^{T} \gamma_t(j) \frac{1}{\sqrt{\sum_{k'=1}^{K} (v^t_{\lambda_{k'}} - m_{jk'})^2}}} \tag{5.11}$$

Then, since m_{jk} is on the left and on the right side, the equation is solved by a fixed point iteration, obtaining $\hat{m}_{jk}$ when convergence is achieved.

Each Simplex-HMM trained with a single video has a specific GMM computed with the MAP adaptation, so the Simplex-HMM model is defined by $\Gamma(A, B, \pi, \widehat{\mu}_i, \Sigma_i)$. As each training sequence is modelled with a Simplex-HMM the inclusion of new training sequences implies a linear increment on the storage space and the required computational power for testing.

5.2.1 Experiments and Results using an HOM in a Simplex-HMM

This work focuses on human activity recognition applied on constrained scenarios, where videos are obtained by fixed viewpoint cameras. The model of the feature space is trained using human motion information from external video and MAP adapted to the target domain, as described in Section 5.1.1. Later, the target domain is used for both, training and testing the algorithm. Our method is evaluated using several datasets that accomplish the source and target domain constraints which description is found in Section 2.8.

Source Domain Datasets In a previous approach we used a virtually created dataset easily scalable if needed however, the realism of the movements is dubious. On the other hand, the quantity of available human motion examples in the Internet is measureless and therefore we can extract the needed information from on-line videos, being real human movements. We have selected three source domain datasets including a high variability in unconstrained video clips that simulate the easily obtainable ones from the Internet: HMDB51, Olympic-Sprots and Virat Release 2.0. They include a high variability of movements in several locations. The three datasets combined have 79 different activity classes extracted from Youtube, movies and surveillance cameras in 7878 video clips.

Target Domain Datasets We have selected three popular datasets in the human activity recognition field as target domain where the videos are recoded from fixed cameras, namely: the Weizmann dataset, the IXMAS dataset and the KTH dataset. Additionally, a dataset with unconstrained video recording has been selected as target dataset in order to evaluate the performance of the algorithm in general purpose applications. In this regard the UCF11 dataset has been chosen.

All videos are processed by means of the state-of-the-art IDT extractor. From the extracted IDTs in the Source Domain datasets, 100000 are randomly selected and used for the GMM training, obtaining 5000 Gaussians, which represent the UBM vocabulary, later MAP adapted to the target domain. On the other hand, sequences from the target domain are used for learning the activity models using a Simplex-HMM (SHMM) and for testing the classifier performance. We have defined an SHMM in a generic manner without specifying the observation model and we have described two possible observation models. Based on the analysis performed in the Section 5.2, we are going to select the HOM for the following experiments. Moreover, we define the SHMM as an ergodic two-state model and set $\varphi = 1.5$, as experiments have shown that the performance is rather insensitive when φ is in the range $(1, 2)$. We perform the following initialization for the EM algorithm used in the SHMM with 2 states:

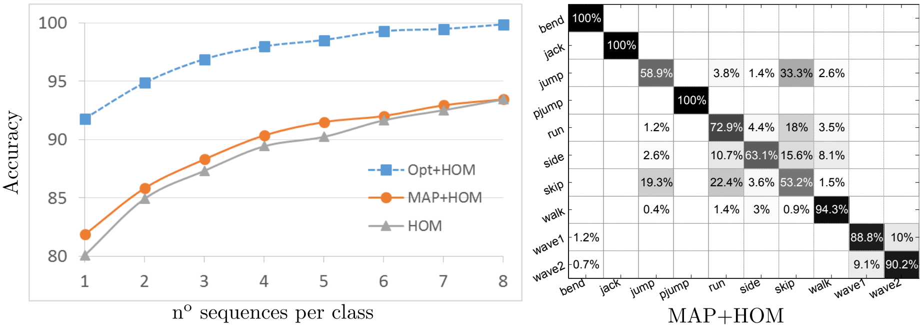
$\pi_1 = 1$ and $\pi_2 = 0$, the transition matrix is randomly initialized, and finally the initial mean vectors, $\mathbf{m}_1$ and $\mathbf{m}_2$, are the observations closest to times $\frac{T-1}{4} + 1$ and $\frac{3(T-1)}{4} + 1$ of the training sequence.

Below we detail the performed experiments using both approaches: strict one-shot learning and relaxed one-shot learning.

Strict One-shot Learning

Considering the proposed activity representation, there are two possible ways of modelling an activity. First, the simplest and fastest method uses the original UBM trained with the source datasets, which is represented by a GMM $\lambda = \{\mu_i, \Sigma_i\}$. An SHMM is then trained per example obtaining the model $\Gamma(A, B, \pi, \{\mu_i, \Sigma_i\})$ where μ_i and Σ_i is shared by all models. The second approach, on the other hand, adapts the UBM vocabulary to a specific GMM per example using the explained MAP adaptation. In [Reynolds et al., 2000] they suggest the insensitivity of the method to the parameter r that we experimentally corroborate, selecting finally $r = 0.014$, where only the mean is modified. Again, the previous configuration of SHMM is trained per activity example, but each activity model is represented in the adapted UBM vocabulary, which implies different GMM means per example as specific information and only Σ_i is shared among models, $\widehat{\Gamma}(A, B, \pi, \{\widehat{\mu}_i, \Sigma_i\})$. In the experiments we call both approaches HOM and MAP+HOM respectively.

In addition to the previous representations, a special instance of the SHMM approach is performed in Weizmann dataset in order to validate the proposed algorithms. If the source dataset used to train the UBM is the whole target domain, including both, train and test examples, then the obtained UBM is the optimal that can be reached with the method configuration, so this special case is considered the UBM ground truth and is labelled with the name Opt+HOM in the experiments. It is worth noting that in a real world application this is infeasible and the Opt+HOM configuration is only used to represent the



	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100%									
jack		100%								
jump			58.9%		3.8%	1.4%	33.3%	2.6%		
pjump				100%						
run			1.2%		72.9%	4.4%	18%	3.5%		
side			2.6%		10.7%	63.1%	15.6%	8.1%		
skip			19.3%		22.4%	3.6%	53.2%	1.5%		
walk			0.4%		1.4%	3%	0.9%	94.3%		
wave1	1.2%								88.8%	10%
wave2	0.7%								9.1%	90.2%

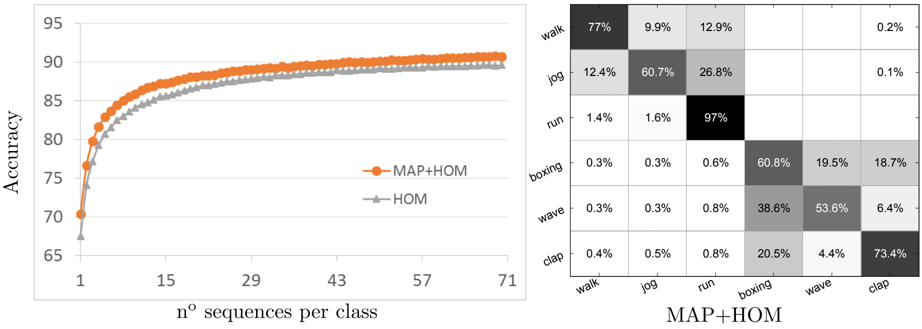
Figure 5.6: Strict one-shot learning results in Weizmann dataset. (Left) Accuracy obtained with different number of training examples per class. (Right) Confusion Matrix of the MAP+HOM method using only one example per class in training

experimental ceiling of our methodology.

The experiments in this section are conducted again with the one-subject-out model. The examples form one subject are selected for testing and from the remaining subjects, *num* sequences per class are randomly selected for training. The value of *num* goes from 1 to the maximum available sequences. The result per subject are the average of 100 runs, and the final result is the average of all subjects.

Figure 5.6 shows on the left a graph with the performance of the method proposed in this paper using the three activity representations described previously. It is worth noting the Opt+HOM results because they justify the suitability of the selected activity representation, i.e. the IDT features encoded in a sequence of BoFs which trains a Simplex-HMM based on a HOM. The results are impressive with only one sequence of each class, as it reaches a 91.8%, and when using the 8 sequences available it almost reaches the 100%, which is comparable to the state-of-the-art results. However, as explained before, this is only possible if the feature space is properly modelled with a GMM, and in this case we have used information from the whole dataset. We can conclude from this result that a proper feature space representation becomes an important objective to be achieved. Thus, we have trained a UBM vocabulary, represented with a GMM of IDTs, obtained from a set of videos as diverse as possible (using the three source domain datasets). Applying the HOM configuration with the GMM trained with three source datasets, the method performance falls significantly against the Opt+HOM configuration, although it still obtains a satisfactory 80.11% using only one sequence per class. Finally, the results are improved by adapting the GMM to the scenario, but as each model uses a single example available, the MAP adaptation has to be performed to this limited available data using the MAP+HOM configuration. Figure 5.6 demonstrates how this adaptation improves the results during all the series, which implies that the adaptation to the target scenario improves the features representation. On the right of Figure 5.6, the Confusion Matrix of the MAP+HOM method using only one sequence for training is depicted. From all the classes the greatest confusion is produced among activities that involve subject displacement (*jump, run, side, skip*), caused by the model giving more importance to the displacement information than the limbs movements. Some works as the one proposed by [Gorelick et al., 2007] incorporates a preprocessing that compensates the displacement, but because of the use of background subtraction and its complications in some scenarios we have opted to avoid it.

Figure 5.7 shows the performance of our proposed method on KTH dataset, only including results for HOM and MAP+HOM representations. In this case we use up to 70 examples for training, and as the graph shows there is a constant gap between HOM and MAP+HOM almost independent on the training examples, which clearly demonstrates the improvement obtained using the proposed adaptation to the scenario. Attending to the Confusion Matrix we can observe the same phenomenon produced in Weizmann dataset, the classes with displacement (*walk, jog, run*) are mainly confused among them. Additionally, we observe in this dataset how the “static” classes (*boxing, wave, clap*) are con-

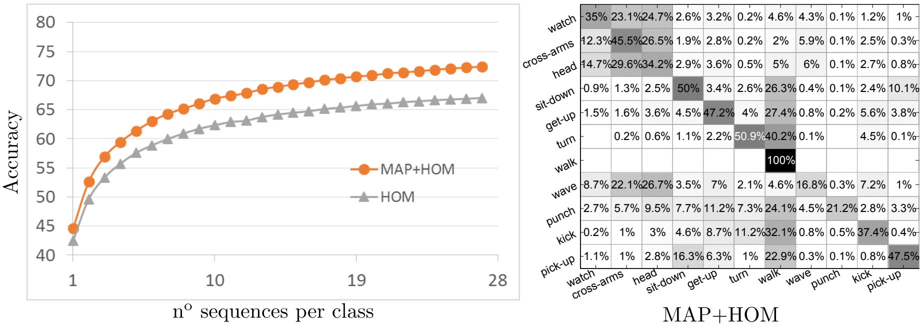


	walk	jog	run	boxing	wave	clap
walk	77%	9.9%	12.9%			0.2%
jog	12.4%	60.7%	26.8%			0.1%
run	1.4%	1.6%	97%			
boxing	0.3%	0.3%	0.6%	60.8%	19.5%	18.7%
wave	0.3%	0.3%	0.8%	38.6%	53.6%	6.4%
clap	0.4%	0.5%	0.8%	20.5%	4.4%	73.4%

Figure 5.7: Strict one-shot learning results in KTH dataset. (Left) Accuracy obtained with different number of training examples per class. (Right) Confusion Matrix of the MAP+HOM method using only one example per class in training

fused as well among them. Unlike Weizmann, the “static” activities in KTH are all described by the movement of the subject arms, which results to the difficulty to distinguish them.

Finally, we repeat the experiments using the IXMAS dataset but avoiding the *point* and *throw* activities as suggested in [Weinland et al., 2006] and Figure 5.8 shows these results. It is worth noting that IXMAS dataset is recorded simultaneously with 5 fixed cameras each one with different viewpoint. We conduct the experiment separately per camera and the shown results is the average of these experiments. Due to the free subject position in the scenario some activities, like *check-watch, cross-arms* but also others, are occluded by the subject body in some cameras generating a worse performance in comparison to the other datasets. However, we ratify the suitability of the MAP Adaptation in the graph. The Confusion Matrix highlights two phenomena: First, the *walk* activity is always right but several other activities are confused with it, which indicates a bias in that model that could be compensated but potential solution is beyond the scope of this work. Second, there are four activities (*check-watch,*



	watch	cross-arms	head	sit-down	get-up	turn	walk	wave	punch	kick	pick-up
watch	35%	23.1%	24.7%	2.6%	3.2%	0.2%	4.6%	4.3%	0.1%	1.2%	1%
cross-arms	12.3%	45.5%	26.5%	1.9%	2.8%	0.2%	2%	5.9%	0.1%	2.5%	0.3%
head	14.7%	29.6%	34.2%	2.9%	3.6%	0.5%	5%	6%	0.1%	2.7%	0.8%
sit-down	0.9%	1.3%	2.5%	50%	3.4%	2.6%	26.3%	0.4%	0.1%	2.4%	10.1%
get-up	1.5%	1.6%	3.6%	4.5%	47.2%	4%	27.4%	0.8%	0.2%	5.6%	3.8%
turn		0.2%	0.6%	1.1%	2.2%	50.9%	40.2%	0.1%		4.5%	0.1%
walk							100%				
wave	8.7%	22.1%	26.7%	3.5%	7%	2.1%	4.6%	16.8%	0.3%	7.2%	1%
punch	2.7%	5.7%	9.5%	7.7%	11.2%	7.3%	24.1%	4.5%	21.2%	2.8%	3.3%
kick	0.2%	1%	3%	4.6%	8.7%	11.2%	32.1%	0.8%	0.5%	37.4%	0.4%
pick-up	1.1%	1%	2.8%	16.3%	6.3%	1%	22.9%	0.3%	0.1%	0.8%	47.5%

Figure 5.8: Strict one-shot learning results in IXMAS dataset. (Left) Accuracy obtained with different number of training examples per class. (Right) Confusion Matrix of the MAP+HOM method using only one example per class in training

cross-arms, scratch-head and *wave*) that are mainly confused among them. The lower amount of movement in these activities contributes to a higher dependence in the camera viewpoint and therefore their confusion.

As the experiments in IXMAS have been performed per camera we can evaluate them separately. The results of each camera accuracy is shown in Table 5.1 using only one example per training class or using the maximum of 27. In addition to the MAP+HOM improvement, we can highlight that this improvement is more significant in camera 5, which is a zenith camera. The viewpoint of this camera is rarely used and then the movements are worse represented in the UBM. Thus, the MAP Adaptation produces a greater impact.

We have found only one paper covering the strict one-shot learning paradigm [Seo and Milanfar, 2011], although indirectly. They propose a sequence representation based on their defined local space-time descriptors. Afterwards, computing a distance among sequences they select the class of the closest one, being a strict one-shot learning as the representation do not need information of other sequences. The experimental results are presented by two graphs, one for Weiz-

Table 5.1: Accuracy using strict one-shot learning in each of the IXMAS cameras using 1 and 27 examples for training.

	camera1		camera2		camera3		camera4		camera5	
	1	27	1	27	1	27	1	27	1	27
SHMM	45.7	73.6	44.5	71.5	43.1	69.7	48.5	69.7	30.6	50.6
MAP+SHMM	48.3	78.5	45.8	76.1	44.7	71.5	50	74.8	34	61.2

mann and other for KTH, that show the performance obtained using only one sequence per class in the training. In Table 5.2 we compare our results with this work, extracting their results approximately from the graphs, showing clearly how our approach improves the previous results significantly in the two comparable datasets. In IXMAS we compare with our previous work explained in the section 4.2 of relaxed one-shot learning and published in [Rodriguez et al., 2013a], where silhouettes ware used as descriptors. In that work, only camera1 was used, and the method is not a strict one-shot learning but a relaxed one-shot learning constrained to 5 sequences initialization where the inclusion of new activity classes is studied approximating the method to a strict configuration. The new proposed method clearly overcame the results using the five cameras (camera5 gives the worst results), and learning from the first sequence. Moreover, we include in the experiments the stable HMM solution, FDHMM, of section 4.2 using the sequence of vectors in a unit simplex encoding, based on BoF of IDTs. However, FDHMM does not consider the histogram distribution and therefore it fails in the experiments as shown in Table 5.2 where we conduct the FDHMM experiments with and without the MAP adaptation. In the three target datasets we obtain the best results using the MAP+HOM configuration, what encourages the use of this configuration.

After proving the suitability of the method for constrained scenarios, we evaluate the strict one-shot learning in the unconstrained dataset UCF11. We randomly chose 1, 10, 20, 30, 40 and 50 video-clips of each class for training

Table 5.2: Strict one-shot learning with one example per class.

	Weizmann	KTH	IXMAS
FDHMM	68.17%	67.16%	25.36%
MAP+FDHMM	69.61%	67.6%	33.7%
HOM	80.11%	67.53%	42.47%
MAP+HOM	**81.88%**	**70.39%**	**44.58%**
[Seo and Milanfar, 2011]	75%	65%	-
[Rodriguez et al., 2013a]	-	-	35%*

*Not directly comparable as they use only one camera and 5 initial sequences, which is less restrictive.
(-) Lack of results in the referenced papers.

Table 5.3: Strict one-shot learning for UCF11 dataset. Results for our approach compared with [Yang et al., 2013].

	1	10	20	30	40	50
HOM	30.7%	59.2%	69%	75%	79.2%	82%
[Yang et al., 2013]	19.3%	31.3%	39.2%	46.3%	50%	51%

and we test the rest. The average of 50 runs is the result shown in Table 5.3. This experiment follows the same configuration found in [Yang et al., 2013]. Initially, when choosing 1 sequence for training, we compare HOM with MAP-HOM, obtaining similar results (30.67% and 30.83% respectively). As sequences belong to unconstrained scenarios the UBM adaptation does not improve the accuracy and we discard it. On the other hand, we verify how our approach has a good performance compared with [Yang et al., 2013].

Relaxed one-shot learning

As the literature is scarce in one-shot learning methods for human activity recognition, we add a new experiment using the relaxed one-shot learning methodol-

Table 5.4: Relaxed one-shot learning with one example per class. Results of our approach and two state-of-the-art methods.

	Weizmann	KTH	IXMAS
HOM	84.18%	76.65%	52.84%
MAP+HOM	**87.12%**	**80.21%**	**56.43%**
[Yang et al., 2013]	80%	-	-
[Orrite et al., 2011]	81.1%	-	-

(-) Lack of results in the referenced papers.

ogy. In this experiment we select the video examples of one subject, and from them only one example per class. These examples are used for training one Simplex-HMM per class, and all the other subjects are used in testing. The relaxed one-shot learning allows us to apply the MAP adaptation to the features extracted from all the training examples, which implies a better adaptation in comparison to the previous experiments. Moreover, the GMM is now shared among the Simplex-HMMs as in the HOM method. However, this process has some constraints in comparison to the Strict methodology as a sequence per class is necessary from the beginning, which initially can be expensive to obtain, and implies a less flexible addition of new examples and classes.

Table 5.4 shows the results in relaxed one-shot learning using our method in comparison with some results found in the literature. In [Yang et al., 2013], spatio-temporal sub-actions based on optical flow are defined and modelled from all the sequences available for training. In our previous approach of FDHMM without Transfer Learning of section 4.2, published in [Orrite et al., 2011], the descriptors are MHIs based on silhouettes computed in a fixed temporal window. MHIs from all the sequences in training are used to model the feature space. Both methods are outperformed, demonstrating the suitability of the proposed method. Again, it is shown the improvement achieved by the use of the MAP Adaptation to the scenario. The improvement of these results compared to the

Strict methodology happens not only because of the adaptation, as the HOM method does not benefit from it, but also because every Simplex-HMM uses the same actor, and therefore the difference is not on the actor features but only on the activity.

5.3 Extended Study for One-shot Learning of Human Activity by a Simplex-HMM

In previous section we have defined the Simplex-HMM and we have analysed its working performance using the Hellinger observation model. The possibility of using different emission functions opens the framework to different performances. Therefore, this section extends the analysis by studying four different emission functions based on an exponential of the Euclidean distance, the mentioned exponential of the Hellinger distance, the Bhattacharyya coefficient and the Manhattan distance. None of them is a PDF but as proved in Annexe B.1 the EM algorithm converges.

Moreover, the original MAP adapted Simplex-HMM approach is computationally expensive and the cost increases linearly with the number of training sequences. Therefore, two modifications are proposed here: first, reducing the dimensionality of the simplex and second, accelerating the optimization process without a reduction in accuracy.

5.3.1 Observation Models

In order to minimize the training instabilities of the HMM training process we have proposed the use of a numerically manageable observation model. Therefore, we have designed $b_j(O_t)$ as a decreasing function with a maximum equal to 1 in a point in the unit simplex, and its value decreases while it separates from that point, being always greater or equal to 0. The training of an HMM is computed by maximizing the likelihood, traditionally obtained with the EM iterative algorithm. This maximization can be processed by maximizing the Baum's auxiliary function, $Q(\hat{\theta}, \theta)$, where different observation models only modify the last term of Equation 4.9 expressed in Equation 5.12:

$$\sum_t \sum_j \gamma_t(j) \ln(b_j(O_t)) \tag{5.12}$$

Below, we define the four studied observation models. The maximization processes of those with analytic solution are described in Annexe B.2.

Euclidean Observation Model (EOM)

A function that decreases exponentially based on the euclidean distance between the observation, O_t, and the trained point in the simplex for each state, $\mathbf{m}_j$, as expressed in Equation 5.13. A drawback is present in this function as the observations are histograms while the euclidean distance makes an unfair comparison between histograms.

$$b_j(O_t) = e^{-\varphi \sqrt{\sum_{k=1}^{K} (v^t_{\lambda_k} - m_{jk})^2}} \tag{5.13}$$

where φ is a constant and $v^t_{\lambda_k}$ the bin values of the histograms.

The optimization of the model can be obtained by an iterative process of fixed point optimization.

Hellinger Observation Model (HOM)

The drawback present in EOM due to the comparison of histograms can be solved by computing the square roots of each element in the vectors. With this transformation the Hellinger distance is calculated which is a fair comparison between histograms.

$$b_j(O_t) = e^{-\varphi \sqrt{\sum_{k=1}^{K} (\sqrt{v^t_{\lambda_k}} - \sqrt{m_{jk}})^2}} \tag{5.14}$$

We already have analysed the HOM model and how it can be optimized with a fixed point iteration.

Bhattacharyya Observation Model (BOM)

Another fair comparison between histograms is the Bhattacharyya coefficient which is used in this new emission function.

$$b_j(O_t) = \sum_{k=1}^{K} \sqrt{v^t_{\lambda_k}} \sqrt{m_{jk}} \tag{5.15}$$

This model can be optimized as well with a fixed point iteration.

Manhattan Observation Model (MOM)

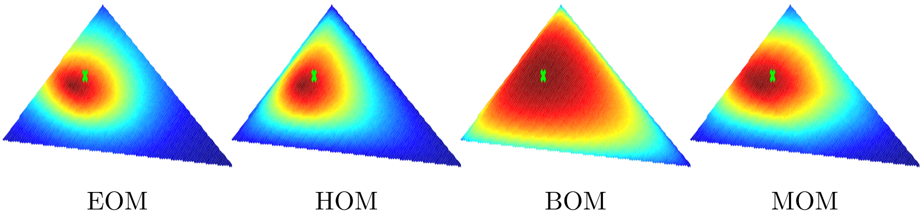
Finally, a decreasing function based on the Manhattan distance is considered. As the maximum Manhattan distance inside a Simplex is 2, we use this factor in the observation model described in Equation 5.16.

$$b_j(O_t) = \frac{2 - \sum_{k=1}^{K} |v^t_{\lambda_k} - m_{jk}|}{2} \tag{5.16}$$

No analytical solution for the maximization of MOM is available, instead general purpose maximization methods implemented in the *fmincon* function of Matlab is used.

5.3.2 Computational cost reduction

In the MAP-Adapted Simplex-HMM framework we have identified two processes that may be computationally expensive. First, the previously explained optimizations are based on iterative processes and, if no analytical solution is available, general purpose optimizations usually lead to slower processes. Second, the soft-assignment of the local features in a high dimensional space is performed for every Simplex-HMM trained with each training sequence. Such



EOM HOM BOM MOM

Figure 5.9: Difference between the maximum of Equation 5.12 and the fast estimation provided by Equation 5.17, for the four methods studied (EOM, HOM, BOM and MOM). The green cross is the estimation and the function is evaluated all over the simplex, with red colours representing the highest values.

a high dimensionality not only slows down the soft-assignment but also any computation in the method. We propose an estimation of the optimum and a reduction of the dimensionality based on the MAP-adaptation in order to reduce the computational cost.

Fast Estimation of the Optimal Parameters

It is worth noting that in one-shot learning paradigm the activity model is optimised using a single sequence which might lead to an overfitting. Then, we propose to avoid the optimization methods and substitute them with a direct estimation of the maximum of Equation 5.12.

The maximum value of Equation 5.12 might be estimated using a trade-off among all the training observations, when higher values of $\ln b_j(O_t)$ are weighted by higher values of $\gamma_t(j)$. Although the trade-off can be different depending on the emission function, we propose the simple direct weight of the observations with the $\gamma_t(j)$ values through (5.17).

$$\hat{m}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j) v_{\lambda_k}^t}{\sum_{t=1}^{T} \gamma_t(j)} \tag{5.17}$$

Experimentally we have observed that this method not only is faster, but also maintains the performance of Simplex-HMM. Figure 5.9 shows the representation of this approach, in a 3D simplex, for the four methods under study. We have designed an experiment where $\gamma_t(j)$ values and O_t observations are randomly obtained, and we have sample the $\ln b_j(O_t)$ along the whole simplex, represented by the figure colours being red the maximum and blue the minimum. Additionally, we have added a green cross in the estimated value. The figure shows that the direct estimation provides a satisfactory approximation. Since only one sequence is available for training, the optimum is found for this specific sequence, which may not be the optimum for the activity class and the use of an approximation may not suppose a performance deterioration. It is worth noting that this estimation is not optimal and it may lead to a decrease in the log-likelihood of the EM process. Therefore, we have included a condition in the algorithm so that the iterative process will be terminated if the log-likelihood decreases.

Reduced MAP Adaptation

One of the pillars that supports Simplex-HMM framework is the MAP adaptation of the features space. In the strict one-shot learning process the MAP adaptation is processed for every training sequence in the target scenario, obtaining a GMM per sequence $\widehat{\lambda}_n(\widehat{\mu}_n, \Sigma), 1 \leq n \leq N$. The proposed framework assumes shared covariance matrices among models and no weights for the Gaussians, restricting the modification per model to the centroids only. For a BoF encoding the number of Gaussians, K, is several thousands and if we have N training examples we need to store KN centroids. But the computational cost is even more restricting. These KN centroids suppose that the evaluation of every new descriptor should be made against all of them to obtain the soft-assignation. Given the set of descriptors $\mathbf{Q} = \{\mathbf{q}_m\}$, $\mathbf{q}_m \in \mathbb{R}^D, 1 \leq m \leq M$, for Gaussian $\widehat{\lambda}_{ni}$ in the MAP-adapted UBMs, the probabilistic alignment of the feature vectors is computed with Equation 5.18. If $\mathbf{Q}$ represents the test set of descriptors, this equation should be processed KNM times.

$$p(\widehat{\lambda}_{ni}|\mathbf{q}_m) = \frac{\mathcal{N}(\mathbf{q}_m|\widehat{\mu}_{ni}, \Sigma_i)}{\sum_{k=1}^{K} \mathcal{N}(\mathbf{q}_m|\widehat{\mu}_{nk}, \Sigma_k)} \tag{5.18}$$

On the other hand, if $\mathbf{Q}$ represents the set of descriptors of a training example then, the MAP Adaptation uses this same probabilistic alignment for the calculus of the GMM transformation. In this case the UBM is represented only by one GMM, $\lambda(\mu, \Sigma)$. These probabilistic alignments and the feature vectors are used to compute the sufficient statistics and an initial step is the calculus of the global probabilistic alignment with Equation 5.19.

$$n_i = \sum_{m=1}^{M} p(\lambda_i|\mathbf{q}_m) \tag{5.19}$$

For each Gaussian λ_i in the UBM the global probabilistic alignment gives the information on how related are the training data to the specific Gaussian. A large n_i means that there are some $\mathbf{q}_m$ samples close to the Gaussian and then this Gaussian is representative for the data, a small n_i means that the data is separated from the Gaussian and this Gaussian has low relevance to the samples in $\mathbf{Q}$. Therefore, we can use n_i not only for the MAP-adaptation but also as an activation parameter per Gaussian. Thus, we use a threshold over n_i. If $n_i \geq \xi$ the Gaussian is active and MAP adapted and discard otherwise. Finally, the MAP-Adapted GMM is composed by the adapted Gaussians that fulfil the threshold restriction as described in Equation 5.20.

$$\widehat{\lambda} = \{\widehat{\lambda}_i\}, \; \forall i \ni n_i \geq \xi \tag{5.20}$$

Depending on the threshold ξ the computational cost and storage are more or less reduced and we need to find a trade-off between cost and accuracy. After the reduction each model has a different number K of Gaussians, and if the average number is $\overline{K} < K$ then, the final number of times the probabilistic alignment equation should be processed is $\overline{K}NM < KNM$.

5.3.3 Experiments

So far, we have presented four observation models and we have defined two modifications that reduce the computational cost. To validate them we divide the experiments into three sets. First, modifying the approach based on HOM we introduce the fast estimation and the reduced MAP adaptation, comparing the accuracy performance of different combinations. Second, after the validation of the two improvements, we extend the experiments to the other three models: EOM, BOM and MOM. Finally, we check the time reduction achieved by the proposed improvements.

Fast estimation and reduced MAP adaptation

HOM is used in all experiments of this section to evaluate the performance of the techniques that reduced the computational cost, proposed in section 5.3.2. Two subset of experiments are performed. First, no MAP adaptation is applied, to evaluate the fast estimation of the optimal parameters (Fast) against the original approach (Orig). Second, MAP adaptation is applied to compare the reduced MAP adaptation ($\overline{K}$), also in combination with the fast estimation of parameters ($\overline{K}$+Fast), against the original MAP adaptation. The average number of Gaussians used after reduction is different en each dataset. Considering an original UBM of $K = 5000$ Gaussians and using a threshold of $\xi = 10^{-10}$, $\overline{K} = 373$ in Weizmann are obtained, $\overline{K} = 712$ in KTH and $\overline{K} = 1155$ in IXMAS. Two justification may be given for the variety of the $\overline{K}$ values: first, activities with varied movements have more Gaussians activated; second, longer sequences have more extracted features increasing the activated Gaussians as the activation equation is a direct addition without normalization.

Table 5.5 shows the results of this set of experiments in the strict one-shot learning framework. In Weizmann dataset the original approach obtains

Table 5.5: Accuracy in strict one-shot learning using the proposed improvements.

	Weizmann				KTH				IXMAS			
	Orig	Fast	$\overline{K}$	$\overline{K}$+Fast	Orig	Fast	$\overline{K}$	$\overline{K}$+Fast	Orig	Fast	$\overline{K}$	$\overline{K}$+Fast
UBM	**80.1**	79.9	-	-	67.5	**72.4**	-	-	42.5	**44.3**	-	-
MAP	**81.9**	-	81.1	81.5	70.4	-	71.8	**74**	44.6	-	46.6	**47.4**

Table 5.6: Accuracy in strict one-shot learning using the four observation models studied for direct UBM (Fast) and MAP-adapted UBM ($\overline{K}$+Fast).

	Weizmann		KTH		IXMAS	
	UBM	MAP	UBM	MAP	UBM	MAP
HOM	**79.9**	81.5	72.4	74	*44.3*	47.4
EOM	*78.1*	*80.7*	72	*73.5*	45.8	*46.5*
BOM	79.2	81.8	*71.8*	*73.5*	46.3	47.4
MOM	79.5	**82.1**	**72.8**	**74.1**	**50.3**	**51.2**

Bold for maximum and Italic for minimum

better results, although close to the others, while in KTH and IXMAS both, fast estimation and reduced MAP adaptation outperforms the results obtained without their use. From these results we can conclude that both modifications not only maintain a similar level of performance but in most cases even improve it. The reduction of Gaussians and the Fast approximation are likely to decrease the clutter originated from an over-fitting of the one-shot learning.

Observation Models

Table 5.6 compares the four observation models applied for strict one-shot learning. The experiments include the direct use of the UBM, assuming fast estimation of parameters (Fast), and the MAP adaptation, assuming fast estimation of parameters and reduction of Gaussians ($\overline{K}$+Fast). Attending these two approaches, in every case the reduced MAP adaptation improves the accuracy of the method. On the other hand, comparing among observation models differences are insignificant but the trend is that MOM achieves the best performance and EOM the worst. The drawback previously explained for EOM about the comparison between histograms might explain its worse results however, the difference with the other methods is tight.

Computational cost

In order to encourage the use of the proposed improvements we evaluate the computational times in three stages of the algorithms. Every value is obtained using the Weizmann dataset and the experiments are run in Matlab in an Intel i7-4790 CPU at 3600 GHz. In the case of $K = 5000$ we have obtained the computational times when applying the direct use of the UBM without MAP adaptation but as the process is the same results should be also the same.

First, the time spent in the training of the Simplex-HMM for different observation models (EOM, HOM, BOM and MOM) is evaluated. Table 5.7 shows how both, the reduction in the number of Gaussians, and the use of the fast approximation reduce the time consumed. The fast approximation is clearly useful if a good optimization method is lacked, like in MOM, where the training is terminated after 100s. In the rest of methods the improvement is still clear as the time required is in average a 57% of the original one although with a high variability, being as low as a 23% for HOM and as high as a 85% for BOM. Since training is not highly restrictive in time we can use the real optimization but results show how the fast approach obtains even better accuracy. On the other hand, the reduction of Gaussians increases the speed of training significantly reducing the time in average a 75%.

In both stages, training and likelihood evaluation, the cluster assignment is equivalent and the computational cost is highly dependent on the number of

Table 5.7: Average time (s) consumed for training the Simplex-HMM in the different observation models using Weizmann dataset. Comparison of methods that use all the Gaussians or the reduced number, and optimizing Equation 5.12 (Orig) or using the fast approximation (Fast).

	$K = 5000$		$\overline{K} = 373$	
	Orig	Fast	Orig	Fast
EOM	0.4636	0.2644	0.1073	0.0766
HOM	0.7035	0.1631	0.1324	0.0313
BOM	0.1224	0.1145	0.0376	0.0289
MOM	>100	0.1076	>100	0.0383

Table 5.8: Average time (s) spent in the activity encoding (after the IDT have been extracted).

$K = 5000$	$\overline{K} = 373$
0.8653	0.042

clusters. Table 5.8 shows that using a reduced set of the dictionary ($\overline{K} = 373$), leads to more than 95% efficiency, compared to using the whole dictionary ($K = 5000$). It is worth noting that this time is computed per model, so the time is increased linearly with the number of training examples in the target domain.

Finally, Table 5.9 shows the time spent for obtaining the log-likelihood in a single Simplex-HMM once the encoding has been computed, using $K = 5000$ and $\overline{K} = 373$ in the four observation models. Results confirm that the reduction of the number of Gaussians reduces the computational time in the Simplex-HMM evaluation, with a 87% of efficiency. It is worth noting again that the times are reported for a single training sequence, therefore they should be multiplied appropriately for multiple training sequences. On the other hand, the whole algorithm is highly parallelizable, therefore a proper implementation could further

Table 5.9: Average time (s) spent in log-likelihood computation using different number of Gaussians.

	$K = 5000$	$\overline{K} = 373$
EOM	0.0141	0.0014
HOM	0.0054	0.0012
BOM	0.0105	0.0016
MOM	0.0151	0.0009

minimise the computational time.

5.4 Conclusion

It is worth noting the shortage of relevant work about one-shot learning for human activities available in the literature. Especially, in the strict one-shot learning paradigm very little work has been done, and our point of view is that it is the most appropriate in many real world applications thanks to its flexibility including new examples and classes. Although the human activity recognition community is tending to focus in large unconstrained datasets, more research in this field can accelerate the installation of recognition systems in new scenarios incorporating scenario based information and therefore reducing the clutter.

In this topic, we have been able to design a framework that gives good results in the three tested datasets, and even in the unconstrained dataset UCF11. The introduced Simplex-HMM facilitates the modelling of an activity using limited amount of data, as few as one example per class, thanks to the reduction of parameters to train. Moreover, in the Weizmann dataset, we have seen how these representation obtains great results if the feature space is properly modelled with a GMM. The difficult to properly model the features space with only one sequence is reduced by using a MAP adaptation of a UBM trained with source domain datasets. The experiments performed in all datasets have confirmed that the adaptation of the model to the specific scenario improves the method accuracy.

It is worth noting that the more labelled sequences, the more storage space and computational power is required for inferring sequences. This problem is reduced by applying two modifications in the algorithm. An estimated optimization and a reduction of the transferred Gaussians form the UBM. Thanks to these changes the algorithm modifications have reduced significantly the computational cost facilitating the introduction of new sequences. Computational times can be further improved, if more efficient languages as C or C++ are used for the implementation, instead of Matlab.

The proposed algorithm assumes a limited number of available labelled sequences from the target dataset. So, in spite of the significant efficiency achieved with the modifications, computational cost will still increase linearly with the increase of the training sequences in the target scenario.

The learning of human activities with MAP adaptation and Simplex-HMM was communicated in the journal paper, [Rodriguez et al., 2016a], and another communication has been submitted to the international journal IEEE Transactions on Cybernetics.

6 Time Flexible Kernel

Significant research effort has been invested in video-based activity recognition during the last few years, supported by the widespread availability of video cameras and by the benefit it may suppose to many applications. In chapters 4 and 5 we have focused in scenarios where the scene is recorded form a fixed viewpoint however, if we spread the recognition to unconstrained scenario new possibilities appear. For instance, classification methods can be used for video indexing where source videos are unconstrained, such as film scenes or clips in video-sharing websites. Furthermore, the use of moving cameras, such as egocentric or attached to drones, can be used in surveillance or entertainment tasks. As well some AAL approach are benefited from robust recognition systems in unconstrained scenarios. As mentioned before, one field of study in AAL gaining momentum is the one based on egocentric cameras which are less intrusive than classical approaches from fixed cameras although, they imply several issues such as the camera movement or the absence of the subject in the image. A comprehensive review by [Nguyen et al., 2016] has been recently published.

The design of sophisticated low-level descriptors, mentioned for instance in [Laptev, 2005] [Scovanner et al., 2007] or [Kläser et al., 2008], has been central in recent advances for this research challenge. Specifically, space-time feature codification has been present in the state-of-the-art approaches where video sequences are represented by a Bag of Features (BoF) or a Fisher Vector (FV), encoding the extracted features. The recognition process is carried out afterwards by applying a multi-class Support Vector Machine (SVM) [Oneata et al., 2013], which takes advantage of the kernel trick. Despite the promising performance of these approaches there are two drawbacks due to the characteristics of the descriptors: (i) using image or short-term descriptors, the lack of explicit temporal information withhold them from reliable recognition of activities [Kliper-Gross et al., 2012], and (ii) mid-term descriptors may describe better the activities [Wang and Schmid, 2013] but still lack of information of the whole temporal structure making them unreliable for complex activities where the order of sub-actions describes the activity.

On the other hand, some state space models such as HMM [Yamato et al., 1992] or more recent Conditional Random Fields [Wang et al., 2011], codify the long term temporal information of the sequences. Although they have provided satisfactory results in human activities they usually work in constrained scenarios where there is no camera motion and the point of view is fixed, as the work we have presented in chapters 4 and 5 for one-shot learning. Due to these restrictions it is possible to train states encompassing common characteristics among the videos and thus to achieve high accuracy. However, when working with unconstrained scenarios their results decrease. Moreover, in the case of our proposal, large training datasets are computational expensive and therefore we advise against its use. In Table 5.3 we shown results in the unconstrained dataset UCF11, which where good for one-shot learning but are still far from recent methods using the discriminative SVM classifier.

New databases, such as HMDB51, UCF50, OlympicSports and Virat Release 2.0 have been produced aiming at challenging tasks such as the previously mentioned. These datasets were recorded in unconstrained environments with random viewpoints, camera movements and/or dynamic changes in the back-

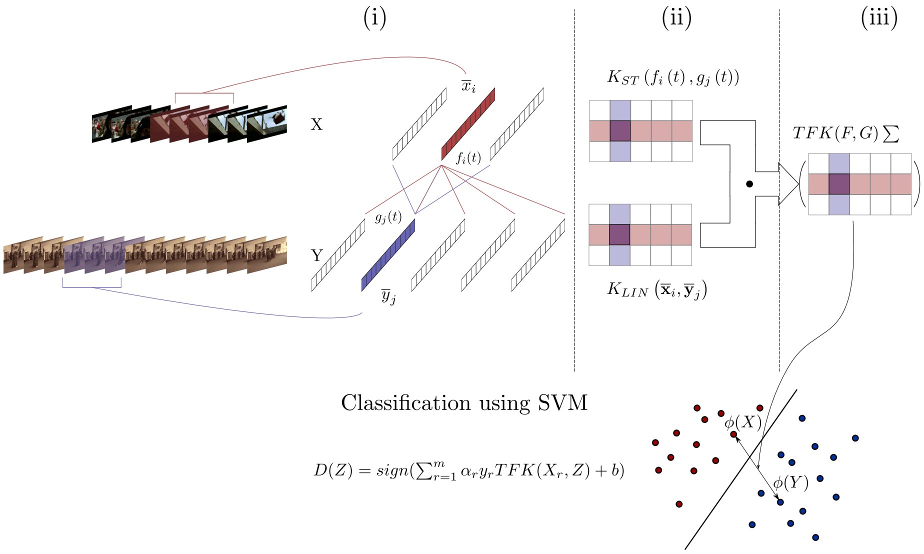

(i)
(ii)
(iii)
$\overline{x}_i$
X
$f_i(t)$
$g_j(t)$
Y
$\overline{y}_j$
$K_{ST}\left(f_i\left(t\right), g_j\left(t\right)\right)$
$K_{LIN}\left(\overline{\mathbf{x}}_i, \overline{\mathbf{y}}_j\right)$
$TFK(F, G)\sum$
Classification using SVM
$D(Z) = sign(\sum_{r=1}^{m} \alpha_r y_r TFK(X_r, Z) + b)$
$\phi(X)$
$\phi(Y)$

Figure 6.1: Graphical Abstract: (i) sliding frame-windows are used to encode the low-level descriptors information obtaining a sequence of vectors, (ii) two different kernels between sequence elements are computed using video and structure information and (iii) a multiple kernel learning is calculated in order to recognize using a SVM.

ground and the recognition methods should be designed accordingly. ahSome of the best results in these challenging benchmark datasets have been obtained with variations of the mentioned SVM approach [Gaidon et al., 2012] [Oneata et al., 2013] [Wang and Schmid, 2013], which has been proven to be a convenient method in spite of the lack of long-term dynamic information. Nevertheless, the long-term temporal information is important in the description of complex activities and thus, we propose the recognition framework depicted in Figure 6.1 where such information is maintained. Using any of the encodings based in BoF or FV, we create sequences of BoFs or FVs as explained in Section 3.2. These sequences preserve the long-term dynamic information needed for the recognition of complex activities. However, as the sequences length and the pace of actions are variable, standard kernels obtained between vectors of same length are not applicable and novel approaches, like Spatio-Temporal Pyramid Matching (STPM) [Choi et al., 2013], keep the long-term information, but they rely on perfect alignment of the sequences with regular pace of actions. Nevertheless, it is worth noting the improvement achieved using several encoding scales, proposed in STPM, that we also apply into our work. So, our contribution includes the design of a novel kernel formulation between arbitrary length sequences that allows the use of the long-term dynamic information in a SVM with matching flexibility, named Time Flexible Kernel (TFK). In order to validate our contribution we have carried out several experiments in four challenging datasets: HMDB51, UCF50, OlympicSports and Virat Release 2.0.

The rest of the chapter is divided as follows. First, we make an introduction of the SVM in Section 6.1. Section 6.2 explains the proposed framework: the Time Flexible Kernel, as well as its application for activity recognition. Section 6.3 presents our experimental validation and section 6.4 concludes the work.

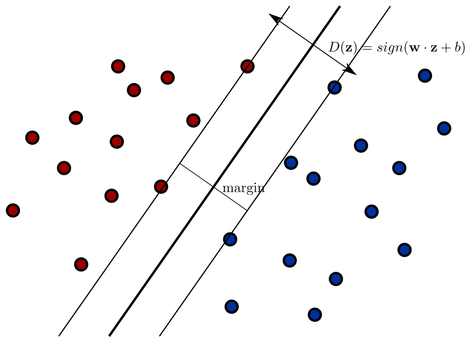


Figure 6.2: SVM binary decision hyperplane that maximizes the margin.

6.1 Support Vector Machines

Support Vector Machines are supervised learning algorithms used for solving classification and regression problems. In this thesis the different goals share the objective of obtaining better recognition rates of activity classes and therefore this short introduction to SVMs is focused in the classification task. The original SVMs only provide a binary classifier however different techniques have been implemented in order to extend their use for multi-class classification. In this section I am going to give some general ideas of how SVM techniques work in order to ease the understanding of the TFK contribution. Thus, this introduction lacks of an in deep analysis and therefore I recommend to go to the literature to find out the whole working of the method. A good start is the work of [Campbell and Cristianini, 1998].

Given a set of labelled input vectors $\{X, Y\} = \{(\mathbf{x}_i, y_i)\}, 1 \leq i \leq M$, where $\mathbf{x}_i$ is the vector and $y_i \in \{+1, -1\}$ the label, the binary classifier is represented by an hyperplane in the vectors' space that separates vectors from one class to vectors from the other, as shown in Figure 6.2. This classifier can be represented through the decision function in Equation 6.1.

$$D(\mathbf{z}) = sign\left(\mathbf{w} \cdot \mathbf{z} + b\right) \tag{6.1}$$

where $\mathbf{w}$ is a weight vector, b a bias and $\mathbf{z}$ the vector to classify.

If the classes are linearly separable, every sample from one class would be in one side of the hyperplane and every sample from the other class would be in the other side, as shown in the Figure 6.2. Being linearly separable, the SVM learning maximizes the margin of the classifier which is defined as

$$\gamma = \min_{\mathbf{x} \in X} |\mathbf{w} \cdot \mathbf{x} + b| \tag{6.2}$$

The margin maximization can be proven to be obtained by maximizing the following function which includes the Lagrange multipliers α_i in order to introduce some constraints:

$$L = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) - \sum_{i=1}^{M} \alpha_i \left(y_i \left[\mathbf{w} \cdot \mathbf{x}_i + b\right] - 1\right) \tag{6.3}$$

Solving the saddle point equations $\frac{\partial L}{\partial b}$ and $\frac{\partial L}{\partial w}$ gives respectively:

$$\sum_{i=1}^{M} \alpha_i y_i = 0 \qquad \mathbf{w} = \sum_{i=1}^{M} \alpha_i y_i \mathbf{x}_i \tag{6.4}$$

If we substitute these solutions in Equation 6.3 then we obtain:

$$L = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{6.5}$$

subject to constraints:

$$\alpha_i \geq 0 \qquad \sum_{i=1}^{M} \alpha_i y_i = 0 \tag{6.6}$$

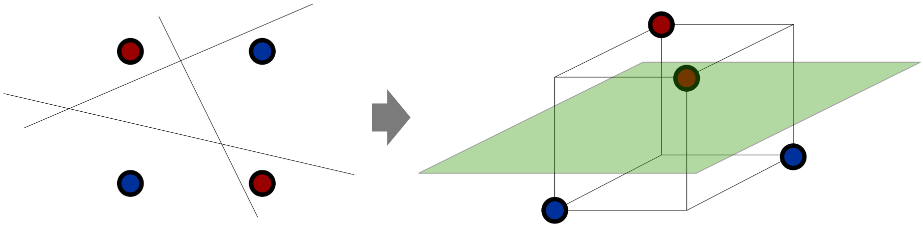


Figure 6.3: Transformation of non linearly separable 2D vectors into linearly separable 3D vectors.

The literature provides information of how this optimization is solved. I would only like to note that the function is convex and therefore there is only one solution to the optimization problem therefore, independently of the optimization process or the starting point we will always reach the same solution (or a close to the optimum one if an iterative process has a stop condition). Remembering the HMM classifiers, with which we have previously work, the EM only assured local optimums therefore, in this regard, SVM supposes an advantage. After maximizing Equation 6.5, the decision function with maximal margins is:

$$D(\mathbf{z}) = sign\left(\sum_{j=1}^{M} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{z}) + b\right) \tag{6.7}$$

This solution is only useful when data is linearly separable, but real world data do not use to be so nice, and some datasets give data samples that are not linear separable. In order to solve this problem an elegant solution is the transformation of the data vectors into a higher dimensional space (even infinity dimensional space) where the classes are linearly separable and then the apply of the SVM is performed in the new space. In Figure 6.3 we observe a representation of what is pursuit. However, this transformation is not always possible and can be computationally very expensive. It is in this part where the kernel trick is used. If we check all the previous equations used for the binary classifier we observe how the data is always used in a inner product. The kernel trick assumes that, without making any data transformation, a inner product in a different dimensional space (unknown for most of the cases) can be easily computed. This kernel is then computed and substituted in the previous equations.

If the kernel function $K(\mathbf{x}, \mathbf{y})$ is semi-definite positive then, there exist a function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$, with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\phi(\mathbf{x}), \phi(\mathbf{y}) \in \mathbb{R}^D$, defining a inner product of the vectors:

$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y}) \tag{6.8}$$

There are several kernels defined in the literature and following we define two popular ones:

Radial Basis Function (RBF) kernel

$$K(\mathbf{x}, \mathbf{y}) = \mathrm{e}^{-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}} \tag{6.9}$$

and polynomial kernel

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p \tag{6.10}$$

Using the kernel trick in the SVM equations we obtain that the maximization function is:

$$L = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{6.11}$$

subject to constraints:

$$\alpha_i \geq 0 \qquad \sum_{i=1}^{M} \alpha_i y_i = 0 \tag{6.12}$$

After maximizing we obtain the decision function:

$$D(\mathbf{z}) = sign\left(\sum_{j=1}^{M} \alpha_j y_j K(\mathbf{x}_j, \mathbf{z}) + b\right) \tag{6.13}$$

With the kernel trick the method solves the non-linear separable problem by introducing non-linear transformations. However, real world classes have additional issues produced by noise in the data and outliers. An SVM can separate the data but this can lead to overfitting of the method and therefore a poor generalization. In order to solve this problem there exist the soft margin solution that consist in allowing some training samples to violate the margin and even to be misclassified. We are not going to explain the process in this introduction to SVMs as there are better tutorials in the literature. What it is interesting for this thesis is to know that they allow the kernel trick and therefore the proposed method that we explain in following sections.

We have seen how the SVMs provide of a binary decision function for classification but, as we pursuit a multi-class classification we need a different solution. Nowadays, there exist different schemes that provide appropriate multi-class classification, as for instance the methods explained in [bo Duan and Keerthi, 2005]. In this regard two schemes are prominent:

- *One-against-all* This method trains every class independently in its own SVM where the samples from that class are labelled as positive examples and the samples form the rest of the classes are labelled as negative examples. The decision function of each class $\omega \in \Omega$ is transformed to a weight function by eliminating the sign:

$$D^{\omega}(\mathbf{z}) = \sum_{j=1}^{M_{\omega}} \alpha_j^{\omega} y_j^{\omega} K(\mathbf{x}_j^{\omega}, \mathbf{z}) + b^{\omega} \tag{6.14}$$

And the class is decided by the maximum of all decision functions:

$$\omega^* = \underset{\omega \in \Omega}{\operatorname{argmax}}(D^\omega(\mathbf{z})) \tag{6.15}$$

This is the scheme we use in our experiments as in the literature has provide the best results for the datasets and descriptors evaluated.

- *One-against-one* For this method, an SVM is trained between each possible pair of classes, having $W(W-1)/2$ SVM classifiers if W is the number of classes. The decision can be processed by organizing the pairwise SVMs in a directed acyclic graph.

6.2 Activity Recognition Framework using a Time Flexible Kernel (TFK)

In the previous section we have seen how the use of the kernel trick allows to obtain a classifier in a dimensional space where the data samples are separable. If we check the formulation of the two popular examples defined there we observe how they use samples of the same dimensionality. However, the activity encodings we are working with, explained in Chapter 3, are sets of vectors of different lengths which difficult the use of kernels. It is worth noting that the sets of vectors are sequences and therefore their order is important. Some solutions proposed in the literature include dynamic alignment of the sequences or forcing the encodings to sequences of the same length and afterwards the matchings of the corresponding elements. These solutions force a perfect matching of the sequence elements in contrast to our proposal where we provide a flexibility in the matching.

Figure 6.4 summarises our proposed framework for activity recognition and compares it to the standard approach. Specifically, it assumes a pipeline of

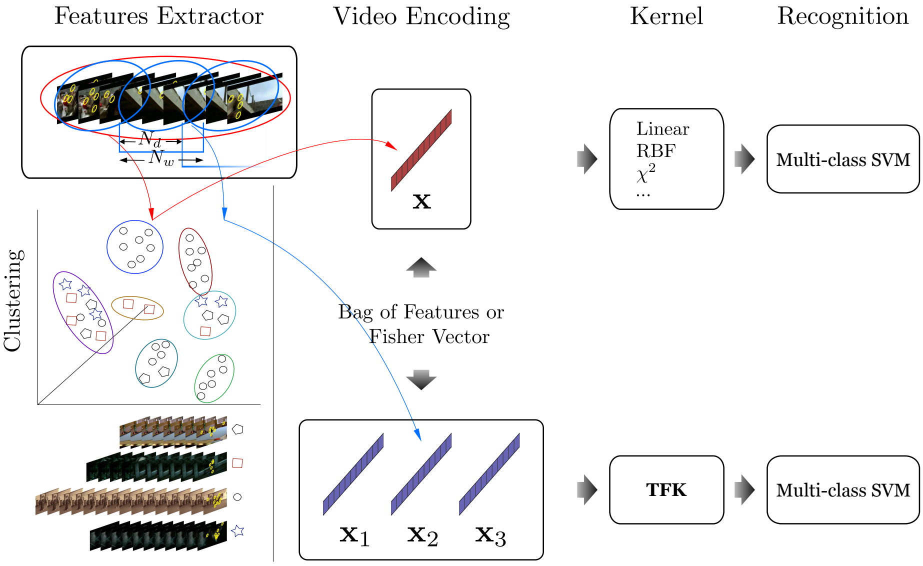


Figure 6.4: Standard (up) and proposed (down) approaches: The features extraction and the clustering stages are common. The standard approach encodes a video into a single BoF using hard-assignment to clusters or a single FV. Our novel approach encodes the video splitting it into sliding frame-windows (window duration, N_w frames, and window stride, N_d frames) obtaining a BoF using soft-assignment or a FV in each window. The new encoding needs a specific kernel (TFK) instead of standard kernels, such as linear, RBF, etc. Finally a multi-class SVM performs the recognition.

feature extraction and clustering, video encoding using BoF or FV, applying kernel and finally using multi-class SVM.

6.2.1 Video Encoding

In Chapter 3 we have explained two activity encodings using BoFs or FVs. We use both approaches for encoding the activities in this work so we are able to validate the suitability of the TFK in any of them. We use an encoding that maintains the temporal information of sequences as shown in Figure 6.4 where we can compare the standard approach (up) and our encoding method (down). There is a common stage of features extraction and codebook generation by clustering in the standard approach and in the proposed one, but the video is represented differently. Our proposal keeps temporal information by computing the FV or BoF on sliding frame-windows on the video. The width of the window is N_w frames and it is displaced N_d frames each time.

A limitation may be introduced because of the width of the window, as the narrower the window the sparser the data used for encoding. As shown in the case of BoF a descriptor is commonly assigned to the closest cluster which is a rough assignation because much of the spatial information in the descriptors space is lost. FV, on the other side, keeps information related to the mean and variance of each cluster which addresses this limitation. Soft-assignment has been proven a good improvement representing continuous data with a codebook model [van Gemert et al., 2010] and then in order to cope with the BoF limitation a soft-assignment is proposed. Specifically, first the relative distance between a descriptor $\mathbf{q}_n \in \mathbb{R}^d$ and a cluster centroid $\mu_k, 1 \leq k \leq K$, in relation to the nearest cluster centroid is obtained based on the euclidean distance d^E.

$$d\left(\mathbf{q}_n, \mu_k\right) = \frac{d^E\left(\mathbf{q}_n, \mu_k\right)}{\min_j\left(d^E\left(\mathbf{q}_n, \mu_j\right)\right)} \tag{6.16}$$

But, instead of performing a hard-assignment (one for the closest cluster and zero for the rest), a soft-assignment is applied as follows:

$$s_{nk} = \left(\frac{1}{d\left(\mathbf{q}_n, \mu_k\right)} \right)^{\beta} \tag{6.17}$$

We assure that the maximum value, $\max_{k=1}^{K} s_{nk} = 1$, is obtained for the closest cluster while smaller values are assigned for more distant centroids. Also, high values of β approximate to the hard-assignment, which is achieved when $\beta \rightarrow \infty$.

Finally, we obtain the observation sequence per activity example, $\mathcal{O} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, being $\mathbf{x}_t$ a D-dimensional vector. In the case of BoF $\mathbf{x}_t$ elements are obtained applying Equation 3.2 and D is the number of clusters in the codebook, while in the case of FV, where soft-assignment is not used, the observation sequence $\mathcal{O}$ is obtained using Equations 3.4 3.5 and A.18 and therefore $D = K2d$.

6.2.2 Time-Flexible Kernel (TFK)

In the standard approach we have a fixed size D-dimensional vector per video so only standard kernels (linear, polynomial, RBF, χ^2, ...) may be applied before using a multi-class SVM. In contrast, our encoding produces arbitrary length sequences of vectors and therefore our novel formulation of a kernel between sequences of different length is applied.

Having two sequences $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ we define the function space $\Gamma : \mathbb{R} \longrightarrow \mathbb{R}^D$ where $F, G \in \Gamma$:

$$F(t) = \sum_{i=1}^{N} f_i(t)\mathbf{x}_i \tag{6.18}$$

$$G(t) = \sum_{j=1}^{M} g_j(t)\mathbf{y}_j \tag{6.19}$$

with $\mathbf{x}_i, \mathbf{y}_j \in \mathbb{R}^D$. We link each vector element with a specific function $f_i, g_j : \mathbb{R} \longrightarrow \mathbb{R}$ used to introduce the temporal position of each element. These functions weigh each sequence element according to variable t.

The TFK is then defined as:

$$TFK(F,G) = \int_t F\left(t\right)^T G\left(t\right) dt \tag{6.20}$$

With the aim of demonstrating that TFK is indeed a kernel we reorder the equation:

$$\begin{aligned} TFK(F,G) = \int_t \sum_{i=1}^{N} \left(f_i(t)\mathbf{x}_i^T\right) \sum_{j=1}^{M} \left(g_j(t)\mathbf{y}_j\right) dt = \\ \sum_{i=1}^{N}\sum_{j=1}^{M} \left(\int_t \left(f_i(t)\right)\left(g_j(t)\right) dt\right) \left(\left(\mathbf{x}_i^T\right)\left(\mathbf{y}_j\right)\right) = \\ \sum_{i=1}^{N}\sum_{j=1}^{M} K_{ST}\left(f_i\left(t\right), g_j\left(t\right)\right) K_{LIN}\left(\mathbf{x}_i, \mathbf{y}_j\right) \end{aligned} \tag{6.21}$$

To prove that Equation 6.21 represents a kernel, we follow several steps. First we check whether K_{ST} and K_{LIN} inside the summation are indeed kernels. The linear kernel, K_{LIN}, is well known. On the other hand, to assure that the structural kernel, K_{ST}, is a kernel we impose the following initial conditions on f_i and g_j: First, they should be square integrable, so $\int_t \left(f_i(t)\right)^2 dt$ and $\int_t \left(g_j(t)\right)^2 dt$ are well defined (not infinity). Second, $f_i(t), g_j(t) \geq 0, \forall t$. Thus, $f_i(t)$ and $g_j(t)$ belong to the Hilbert-space L_2, hence the kernel is semi-positive definite [Jebara et al., 2004].

We still need to prove that the summation of these kernels is a kernel. In this regard, we proceed with the following steps: First, we extend the vector representing the shortest sequence with zeros. Without loss of generality, let's assume that $N > M$, so we extend $\mathbf{y}_j$ in such a way that $\mathbf{y}_j = \overline{\mathbf{0}}$ for $j > M$ and we can use any function $g_j(t)$ for $j > M$ that fulfil the initial conditions. We also denote x_i^p the $p-th$ component of the $\mathbf{x}_i$ vector, and the N-dimensional vector $\hat{\mathbf{x}}^p = (x_i^p, i = 1 \cdots N)$. Then, we develop the scalar product in Equation 6.21 as:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{p} K\left(f_i\left(t\right), g_j\left(t\right)\right)(x_i^p)(y_j^p) = \sum_{p} ((\hat{\mathbf{x}}^p))^T \mathbf{K} (\hat{\mathbf{y}}^p) \tag{6.22}$$

where $\mathbf{K}$ is a $N \times N$ matrix $K_{i,j} = K\left(f_i\left(t\right), g_j\left(t\right)\right)$. The matrix $\mathbf{K}$ is a positive semidefinite matrix, since it corresponds to a kernel in the space of functions. Thus, each of the addends in Equation 6.22 is a kernel in a subspace, and the sum of kernels in all the subspaces is also a kernel in the global space [Bishop, 2006].

As $\mathbf{x}_i, \mathbf{y}_j$ are vectors in an arbitrary $\mathbb{R}^D$ space, we can consider any projection of them in a different $\mathbb{R}^S$ space obtaining $\phi\left(\mathbf{x}_i\right), \phi\left(\mathbf{y}_j\right)$. Then, we can consider any kernel $K\left(\mathbf{x}_i, \mathbf{y}_j\right)$ as a linear kernel in the projected space $K_{LIN}\left(\phi\left(\mathbf{x}_i\right), \phi\left(\mathbf{y}_j\right)\right)$ so, in the previous proof, the linear kernel can be substituted by any arbitrary kernel.

6.2.3 Application of TFK in Activity Recognition

In a real world application there are video sequences with variable lengths, and the recording or segmentation of same event classes are not perfect and then they might start and end in different positions. This implies that when comparing two repetitions of the same activity class it is possible that only a portion of the sequence coincides. We can see this fact in Figure 6.5 where two sequences of the same activity class (somersault), extracted from the HMDB51 benchmark, coincide only in the final portions of the sequences.

Thanks to TFK we are able to compare sequences of different lengths and by selecting the appropriate associated function we can deal with non perfect alignment. In this regard we design the following framework.

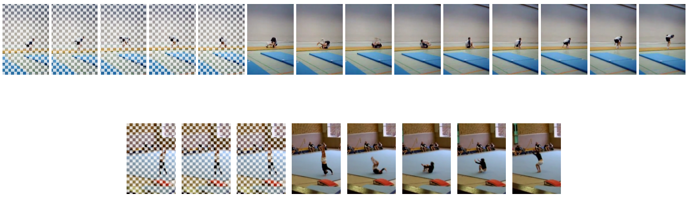


Figure 6.5: Activity correspondence: Two videos of somersault from HMDB51 where only the final portions of the sequences coincide.

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ be two sequences of vectors representing two different activity executions (same or different class). These sequences are obtained with the process explained in Section 6.2.1 so each vector of the sequence is a BoF or a FV. On the other hand, the proper alignment between two sequences is unknown and the computation of an algorithm seeking for this alignment can increase notably the computational cost. Moreover, a proper segmentation is assumed in advance so the core of the activity is most probably located in the middle of the sequences. Therefore, without an alignment process and simply ensuring that centres of both sequences coincide, the proposed method uses the structural kernel of TFK to provide the desired degree of flexibility in compression and stretching of the activity representation. As depicted in Figure 6.6, we center both sequences and assign a Gaussian distribution to each element of the sequences constrained to fixed temporal positions, being $f_i(t)$ and $g_j(t)$ the probability density functions of $\mathcal{N}_i$ and $\mathcal{N}_j$ respectively, $f_i(t) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(t-\mu_i)^2}{2\sigma_x^2}}$ and $g_j(t) = \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(t-\mu_j)^2}{2\sigma_y^2}}$ being $\mu_i = (i - \frac{N+1}{2})\Delta_t^x$ and $\mu_j = (j - \frac{M+1}{2})\Delta_t^y$.

The associated Gaussians weigh the inner product between sequence elements in relation to their temporal position, obtaining a maximum when their time coincides. These functions provide flexibility in the temporal position of the elements, allowing an irregular expansion or narrowing of the sequences as

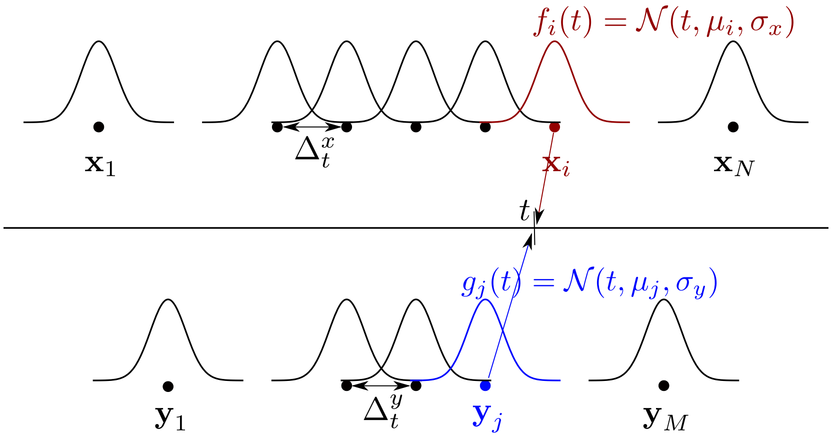


Figure 6.6: Kernel structure: Two centred sequences. Every vector of both sequences have a normal distribution associated. All the vector elements are compared in t weighted with their respective Gaussian function.

well as a displacement. In order to define the functions $f_i(t)$ and $g_j(t)$ it is possible to fix the vector spacing Δ_t and then only the standard deviation of the Gaussian σ modifies the precision of the sequence position. The smaller is σ, the narrower are the Gaussians and then the lesser is the degree of temporal flexibility. Moreover, as the number of elements in a sequence is variable and each element has a Gaussian associated, it is possible to normalize the functions so that the length of the sequence does not influence the kernel value using the normalized Gaussians: $f_i'(t) = \frac{1}{N} f_i(t)$ and $g_j'(t) = \frac{1}{M} g_j(t)$.

Taking into account all previous concerns, we use the following kernel:

$$TFK(F, G) = \sum_{i=1}^{N} \sum_{j=1}^{M} K_{GAUSS_\rho} \left(f_i'(t), g_j'(t) \right) K_{LIN} \left(\overline{\mathbf{x}}_i, \overline{\mathbf{y}}_j \right) \tag{6.23}$$

We use the kernel between Gaussians that was proposed in [Jebara et al., 2004] as K_{ST} in Equation 6.21, which in our one-dimensional case is simplified as:

$$K_{GAUSS_\rho}\left(f_i'(t), g_j'(t)\right) = \frac{\left(2\pi\sigma_x\sigma_y\right)^{(1-2\rho)/2}}{NM\sqrt{2\rho}} e^{\frac{-\|\mu_i-\mu_j\|^2}{4\sigma_x\sigma_y/\rho}} \tag{6.24}$$

Selecting $\rho = 1/2$ we obtain the Bhattacharyya kernel.

Inspired by the idea exposed in STPM [Choi et al., 2013], we explore the addition of two levels of granularity in the sequence division. Therefore, using a simple linear combination of kernels, that keeps the kernel property [Gönen and Alpaydın, 2011], we combine the TFK previously explained with a linear kernel between the vectors obtained from the feature extraction of the whole video. In Figure 6.4 this would mean to combine the two pipelines of the diagram in the following kernel.

$$CombK(v_1, v_2) = TFK(F, G) + K_{LIN}\left(\mathbf{x}, \mathbf{y}\right) \tag{6.25}$$

If the means and variances of the functions $f_i(t)$ and $g_j(t)$ are only dependant on the length of the sequences it is possible to precompute in advance the $K_{GAUSS_\rho}\left(f_i'(t), g_j'(t)\right)$ values for most of the possibles combinations of N and M, so the computational cost is only influenced by the kernel between the vectors. Considering the computational cost of $K\left(\mathbf{x}_i, \mathbf{y}_j\right)$ be $O(D)$, the increase is linear with the increase of one of the sequences length $O(NM(D+1))$.

6.3 Validation

We test the performance of our framework in four challenging Activity-Recognition benchmarks described in Section 2.8 (HMDB51, UCF50, OlympicSports, Virat Release 2.0), and we compare the results against other published state-of-the-art methods.

6.3.1 Parameter Identification

The framework performance is tested using two different descriptors. First, as representative low-level descriptor with short-term information we have selected the MIP descriptor, shortly explained in Section A.1.2, which performs better than other common short-term descriptors like SIFT or HOG-HOF. Second, as state-of-the-art descriptor, and with mid-term temporal information captured, we have selected the IDT descriptor, shortly explained in Section A.1.3. Once we prove the suitability of TFK for short-term descriptors we carry out a more exhaustive experimentation in the mid-term descriptors as they currently represent the state-of-the-art for activity recognition. Anyway, both cases need a parameter tuning and, therefore, we proceed with a parameter analysis below.

The MIP descriptors are extracted with the original specifications proposed by the authors. The video is encoded with a BoF approach creating a codebook of 5000 codewords per channel obtaining a (8×5000)-dimensional vector. Our proposed video encoding depends on three parameters: β (in Equation 6.17) controlling the softness of the assignment and N_d and N_w (in Figure 6.4) modelling the sliding frame-windows. Fixing the temporal spacing $\Delta_t = 1$ we let σ as the free parameter of the Gaussian functions.

We firstly perform experiments for multiple combinations of the parameter β, the window width N_d, the window displacement N_w and the the standard deviation σ of the Gaussian function of the kernel, using the first split of the HMDB51 dataset.

From initial experiments we have found that a sliding frame-window without overlapping provides best results, therefore we fix $N_d = N_w$ and then only 2 parameters of the video encoding are analysed: N_w and β. We show in Figure 6.7 the results obtained by fixing two of the three analysed parameters to the values finally selected, so the graphs represent the performance of the remaining one.

We can observe the importance of using the soft-assignment approximation

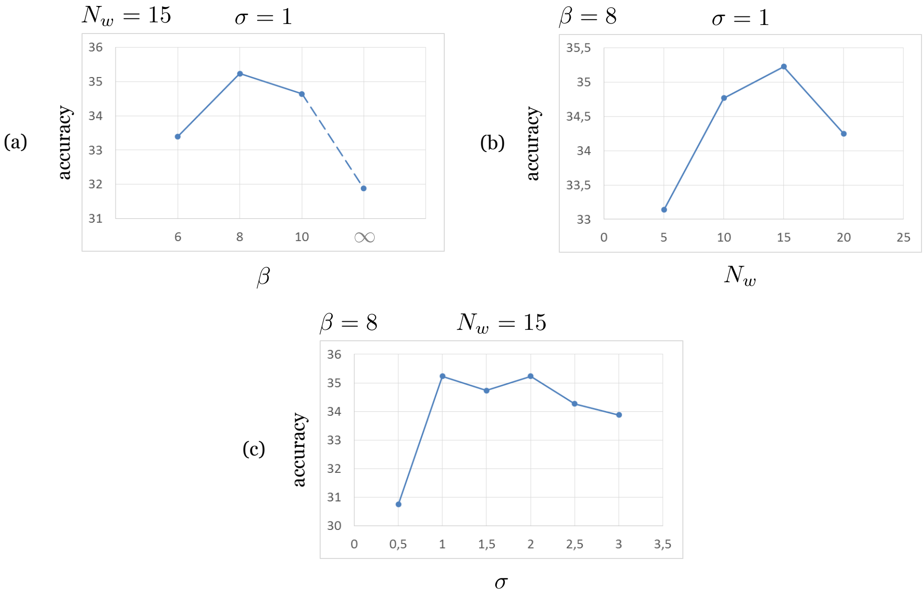


Figure 6.7: Parameters performance using MIP descriptors: Performance of the system evaluated in the first split of HMDB51 in relation to three parameters: (a) β, (b) N_w, and (c) σ.

in Figure 6.7(a) where different values of β are evaluated. If we use a hard-assignment with a window of width $N_w = 15$ the system performance declines in relation to a proper soft-assignment, which can be explained by the lack of sufficient data in a window. On the other hand, any of the three analysed values of β (6, 8 and 10) gives better results than the hard-assignment, what implies that the use of a soft-assignment is an adequate optimization in a wide range of values.

The width of the window N_w, Figure 6.7(b), does not impose significant variations in the system performance either, although the value $N_w = 15$ seems to be optimal.

Figure 6.7(c), depicts the performance of the system while varying the standard deviation σ of the kernel Gaussian functions. The bigger it is the wider are the Gaussians which means that the sequences are more flexible to asymmetrically expand or shrink but also that the temporal position is less influential. Very small values of σ lead to low accuracy, but then the variation in performance is minor and we find the optimum between $\sigma = 1$ and $\sigma = 2$.

After analysing the parameters influence in MIP descriptors, we extend the parameters influence experiments to the IDT descriptors. In this case the encoding is performed with FV using a mixture model of 256 Gaussians as, in addition to the use of IDT, FVs provide better results than BoF in the literature. Hence, the study is performed over N_w and N_d for the sliding frame-window and σ for the temporal structure. We perform the analysis using the training examples of the OlympicSports dataset, dividing it into two groups randomly selected: 70% for training and 30% for validation. The use of this subgroup have some advantages in comparison to the analysis made with HMDB51 first split: first, the group is smaller and then the computational cost is reduced, and second, there is not a validation using test samples and therefore it is valid as a real world implementation and the direct comparison with literature results is legit. In Figure 6.8 we show three graphs fixing two of the parameters to the final value selected and varying the remaining one.

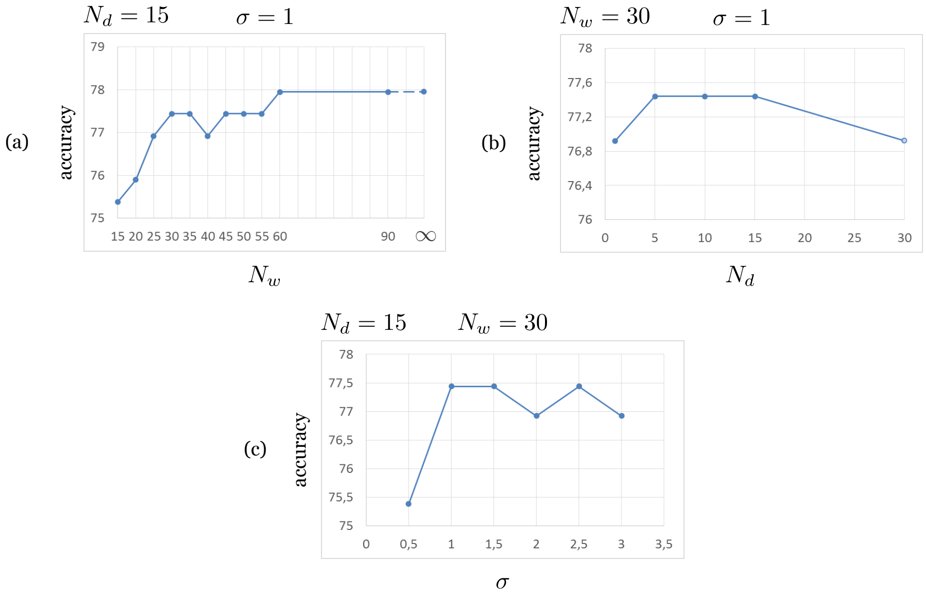


Figure 6.8: Parameters performance using IDT descriptors: Performance of the system evaluated in the OlympicSports dataset in relation to three parameters: (a) w_K, (b) w_D and (c) σ.

Figure 6.8 (a) shows low variation in the performance of the system for varying N_w except with short windows close to the IDT length, once reached $N_w = 25$ the accuracy variation is produced only by two examples recognition, slightly tending to the maximum when increasing the size until reaching the whole video represented in the axes as $N_w \to \infty$. The case where $N_w \to \infty$ corresponds to the use of a single FV per video, which is the standard approach of IDT presented in [Wang and Schmid, 2013]. Therefore, in order to compare our framework with the original approach we carry out the experiments using $N_w = 30$. The value of N_d has even a lower influence in the performance and although the optimal value is between $N_d = 5$ and $N_d = 15$, it makes little difference in the final result. The parameter σ has a similar behaviour to the one in the MIP analysis, although we can consider that the optimum value extends from $\sigma = 1$ to at least the maximum analysed $\sigma = 3$ because the accuracy decreases only in one incorrectly recognized example.

6.3.2 Framework Validation

Our proposed framework is advantageous in two scenarios: (i) when using short-term temporal descriptors, as it allows including longer temporal information in the classifier and (ii) recognizing complex activities where the order of actions may be crucial for correct recognition and therefore the temporal information given by mid-term descriptors is still insufficient. In the latter scenario, our framework can benefit even cases where longer time descriptors are used, as it encodes all the temporal information of the activity. However, when dividing the videos in small sub-clips the information can be scarce and produces unstable encoding that leads to bad classification. This problem is overcome by combining the two granularities of the video division proposed in Equation 6.25, sequences with scarce data would be better encoded with a single descriptor while sequences with enough data would be better encode in a sequence.

Our first experiment analyses the proposed framework performance using short-term descriptors. It is carried out using the MIP descriptor and the BoF encoding over the datasets HMDB51 and UCF50. In the previous section, we have seen that the performance of the framework is not significantly influenced when parameters are chosen within acceptable ranges. For the following experiments we select $\beta = 8$ for the soft-assignment, the width $N_w = 15$ and the displacement of sliding frame-windows $N_d = 15$, so they are not overlapped, and the standard deviation of the Gaussian functions of the structural kernel $\sigma = 1$. We divided both datasets according to the authors' recommendations: 3 splits in HMDB51 and 25 groups in UCF50. We have used publicly available code for MIP [1] and SVM [2], using the default parameters. The randomness of initialisation of the K-means algorithm justifies why our results do not exactly coincide with those given in the MIP original paper. To ensure fair comparison between the standard method and our proposed framework, clustering and features extraction, as well as the one-against-all SVM classification, coincide in both pipelines. The difference lays in the middle stages. The standard method encodes the video with a single BoF obtained with a hard-assignment and applies a linear kernel between BoFs (BoF + LinK) as suggested by the authors for best results in [Kliper-Gross et al., 2012]. The proposed framework encodes each video in a sequence of BoFs (SeqBoF) using soft-assignment and applies the proposed TFK (SeqBoF + TFK). In Table 6.1 we can see how the inclusion of long-term temporal information using TFK clearly improves the results in both datasets which validates our first assumption regarding short-term descriptors. However, clearly better results have been obtained in the state-of-the-art using mid-term descriptors and specifically with IDT therefore, we continue the experiments with these descriptors.

The second battery of experiments has been performed using the state-of-the-art descriptor, IDT[3], and comparing the novel framework against the origi-

[1] MIP descriptor code can be downloaded in http://www.openu.ac.il/home/hassner/projects/MIP/MIPcode.zip

[2] SVM code can be downloaded in http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[3] IDT descriptor code can be downloaded in

Table 6.1: Average accuracy (in %) in the HMDB51 and UCF50 datasets using the splits suggested by the authors. First row: results provided in [Kliper-Gross et al., 2012]. Second row: own implementation of [Kliper-Gross et al., 2012]. Third row: novel framework using TFK.

	HMDB51	UCF50
[Kliper-Gross et al., 2012]	29.17	68.51
MIP + BoF + LinK	30.9	66.0
MIP + SeqBoF + TFK [OURS]	**34.4**	**72.4**

nal work in [Wang and Schmid, 2013] and other related works in Activity Recognition. The sliding frame-window varies form the MIP experiments as $N_w = 30$ and $N_d = 15$. We compute the experiments in all the evaluation datasets: OlympicSports, HMDB51, UCF50 and Virat Release 2. In all the experiments we follow the authors' recommendations: one division for training and testing in OlympicSports, 3 splits in HMDB51, leave-one-group-out from 25 groups in UCF50 and leave-one-scene-out with 11 scenes in Virat Release 2.0. As in the literature we find mainly results of Accuracy (acc) or Mean Average Precision (mAP), we compute both in the different approaches evaluated. Table 6.2 shows the results obtained with the original IDT as well as our proposed frameworks. To assure fair comparison we perform all the experiments using the same IDT extraction and GMM estimation. First, we obtain an unique FV per Video and apply a linear kernel for a SVM classification (IDT+FV+LinK), obtaining our own implementation of the approach in [Wang and Schmid, 2013]. Following, we use the extracted IDT features to obtain a sequence of FV (SeqFV) that are used in our TFK approach (IDT+SeqFV+TFK). Finally, we combine both approaches in the CombK kernel (IDT+CombK) following Equation 6.25.

The TFK approach is suitable in complex activities where the order of sub-actions determines the class. This can be confirmed with the results in Virat

http://lear.inrialpes.fr/people/wang/download/improved_trajectory_release.tar.gz

Table 6.2: Mean Average Precision (mAP) and average accuracy (acc) (in %) results in the OlympicSports, HMDB51, UCF50 and Virat Release 2 datasets using the Improved Dense Trajectories. First row: own implementation of [Wang and Schmid, 2013]. Second row: novel framework using TFK. Third row: novel framework using the combination of TFK with [Wang and Schmid, 2013].

	OlympicSports		HMDB51		UCF50		Virat2	
	mAP	acc	mAP	acc	mAP	acc	mAP	acc
IDT+FV+LinK	89.8	83.6	57.8	57.4	93.8	90.0	43.5	55.7
IDT+SeqFV+TFK	86.5	82.8	58.3	57.7	92.7	89.5	**52.1**	**63.6**
IDT+CombK	**89.9**	**84.3**	**59.1**	**58.6**	**94.1**	**90.3**	47.9	58.1

dataset where there are 4 activities with their respective “opposites”, depicted in the second row, two last columns of Table 6.2. However, TFK also has some drawbacks, as can be observed in the other three datasets where it performs similarly to the original IDT method. TFK relies in the extracted features in each window, and if they are scarce, the computed FV can be less robust to clutter. These datasets have complex activities, but not all of them depend on the order of sub-actions, therefore, although some activities are better classified with TFK, others are worse. The solution for this lack of robustness against clutter is the linear combination of both kernels. As we can see in the last row, this approach improves the results in all datasets but Virat where, even improving the original approach, the result is worse than the direct use of TFK because all the activities but 3 have “opposite” counterparts and the combination of kernels lowers the importance of order.

In Table 6.3 we observe a comparison of the proposed approach against some of the best results in the literature. It is worth noting that there are two rows representing the original approach with IDT [Wang and Schmid, 2013], the third-to-last and the second-to-last. In the third-to-last row we show the results provided in the original paper, but a direct comparison between this results and

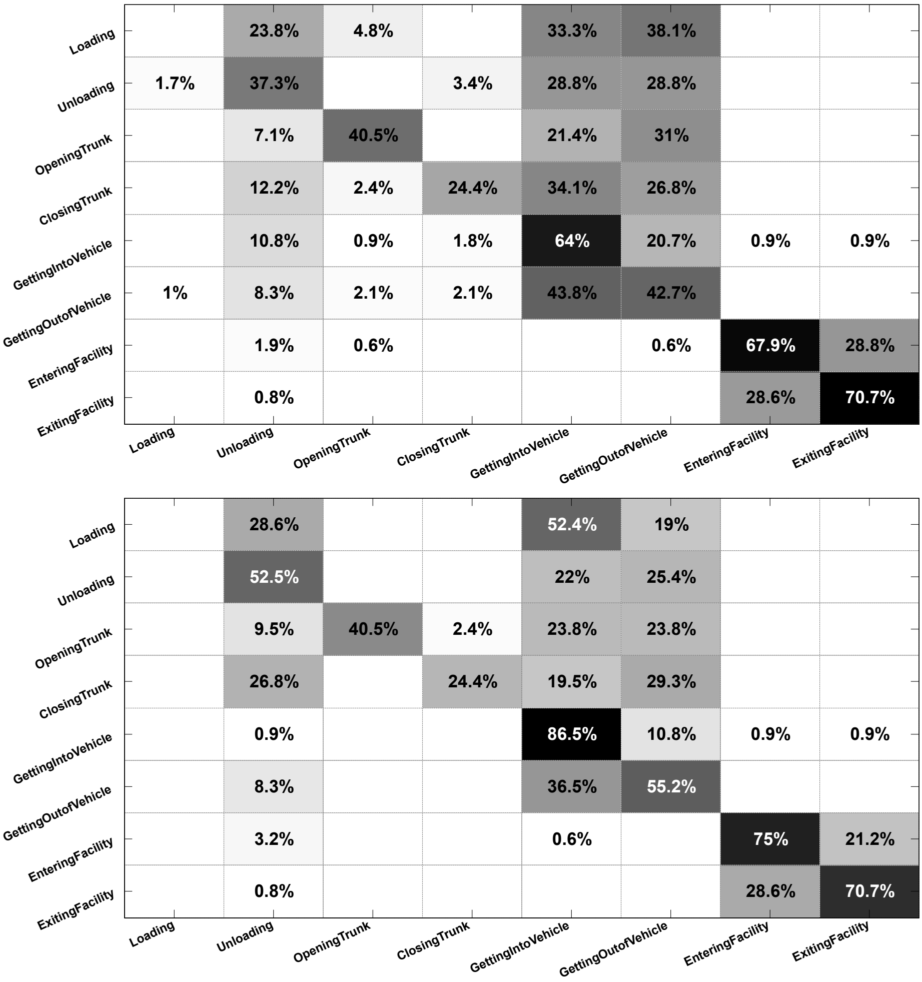
Table 6.3: Mean Average Precision (mAP) and average accuracy (acc) (in %) results in the OlympicSports, HMDB51, UCF50 and Virat Release 2 datasets. Comparison of the proposed framework (last row) with several state-of-the-art approaches.

	OlympicSports		HMDB51		UCF50		Virat2	
	mAP	acc	mAP	acc	mAP	acc	mAP	acc
[Tang et al., 2012]	66.8	-	-	-	-	-	-	-
[Niebles et al., 2010]	72.1	-	-	-	-	-	-	-
[Li et al., 2013]	76.2	-	-	-	-	-	-	-
[Gaidon et al., 2012]	-	82.7	-	-	-	-	-	-
[Cao et al., 2012]	-	-	-	27.8	-	-	-	-
[Kliper-Gross et al., 2012]	-	-	-	29.2	-	68.5	-	-
[Reddy and Shah, 2013]	-	-	-	27.0	-	76.9	-	-
[Wang and Schmid, 2013]	**90.2**	-	-	55.9	-	**90.5**	-	-
IDT+FV+LinK	89.8	83.6	57.8	57.4	93.8	90.0	43.5	55.7
IDT+CombK [OURS]	89.9	**84.3**	**59.1**	**58.6**	**94.1**	90.3	**47.9**	**58.1**

(-) Lack of results in the referenced papers.

our approach is not fair as several stages have some randomness and slightly alter the final results. On the other hand, the second-to-last row shows our own implementation of [Wang and Schmid, 2013] which share the features extraction and the clustering with the TFK so the comparison is fair. We can see how our novel approach overtakes all the compared methods in the "fair" comparison. To our knowledge there is no method with better results for all the datasets.

The results on Virat Release 2.0 are further analysed using only the activities with "opposites", (Loading, Unloading, Opening Trunk, Closing Trunk, Getting Into Vehicle, Getting out of Vehicle, Entering Facility and Exiting Facility). The Confusion Matrices of the (IDT+FV+LinK) and (IDT+SeqFV+TFK) methods are depicted in Figure 6.9. In addition to the improvement obtained with the



	Loading	Unloading	OpeningTrunk	ClosingTrunk	GettingIntoVehicle	GettingOutofVehicle	EnteringFacility	ExitingFacility
Loading		23.8%	4.8%		33.3%	38.1%		
Unloading	1.7%	37.3%		3.4%	28.8%	28.8%		
OpeningTrunk		7.1%	40.5%		21.4%	31%		
ClosingTrunk		12.2%	2.4%	24.4%	34.1%	26.8%		
GettingIntoVehicle		10.8%	0.9%	1.8%	64%	20.7%	0.9%	0.9%
GettingOutofVehicle	1%	8.3%	2.1%	2.1%	43.8%	42.7%		
EnteringFacility		1.9%	0.6%			0.6%	67.9%	28.8%
ExitingFacility		0.8%					28.6%	70.7%

	Loading	Unloading	OpeningTrunk	ClosingTrunk	GettingIntoVehicle	GettingOutofVehicle	EnteringFacility	ExitingFacility
Loading		28.6%			52.4%	19%		
Unloading		52.5%			22%	25.4%		
OpeningTrunk		9.5%	40.5%	2.4%	23.8%	23.8%		
ClosingTrunk		26.8%		24.4%	19.5%	29.3%		
GettingIntoVehicle		0.9%			86.5%	10.8%	0.9%	0.9%
GettingOutofVehicle		8.3%			36.5%	55.2%		
EnteringFacility		3.2%			0.6%		75%	21.2%
ExitingFacility		0.8%					28.6%	70.7%

Figure 6.9: Confusion Matrices in Virat Release 2.0 using activities with "opposites": First the Confusion Matrix using IDT + FV + LinK approach, second using IDT + SeqFV + TFK.

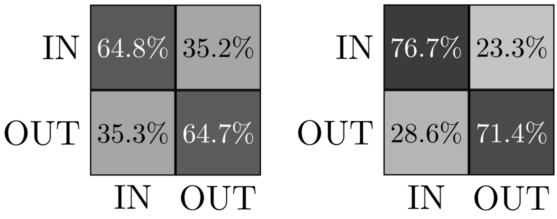


Figure 6.10: Confusion Matrices in Virat Release 2.0 using two classes (IN and OUT): First the Confusion Matrix using IDT + FV + LinK approach, second using IDT + SeqFV + TFK.

proposed framework, these matrices confirm our premise that the TFK is suitable for better learning of complex activities defined with the sub-actions order. The improvement achieved by the proposed framework is clear as every element of the diagonal is greater or equal. But the improvement does not restrict to this as in addition of a general improvement, the wrong classified activities are now confused with activities with similar temporal structure. For instance, the two first activities ('Loading' and 'Unloading') are mainly confused with ('Getting Into Vehicle' and 'Getting Out of Vehicle'). If we observe the Confusion Matrix of the (IDT + FV + LinK) method, the confusion is more or less random, but using TFK we can see how the structure is learnt and then "loading" is mainly confused with "getting into vehicle" and "unloading" is mainly confused with "getting out of vehicle". We achieve a clearer representation of this idea by gathering all the activities with similar temporal structure into one class so, activities (Loading, Opening Trunk, Getting Into Vehicle and Entering Facility) with structure (approaching and opening-closing) are grouped in class IN and activities (Unloading, Closing Trunk, Getting out of Vehicle and Exiting Facility) with structure (opening-closing and moving away) are grouped in class OUT. Figure 6.10 depicts the Confusion Matrices of these two classes. Here it is clear how the TFK approach keeps better the temporal structure of the activities.

Finally, we introduce one more experiment in order to compare our framework with the STPM approach which also preserves the temporal structure of the activities. In [Choi et al., 2013], they designed an experiment called *Quality of binary decision* where one example is compared to other two, one with the same class and other with a different class. Whenever the example of the same class is more similar to the initial example than the other one, the binary decision is correct. With this experimentation the authors obtains a maximum of 95.3% of Precision. In order to get a similar process we select single-class SVM to provide binary decisions between two randomly selected examples (one form the same class and one from a different class). Whenever the example of the same class has a greater note than the other one, the binary decision is correct. Using this experimentation we obtain a 99.3% of Precision, which is clearer better than the provided by [Choi et al., 2013].

6.4 Conclusion

We have introduced a new framework that improves accuracy in human activity classification taking into account the long-term information. The framework can be used with a wide variety of low-level feature descriptors, such as MIP and IDT, and video encoding methods, such as BoF and FV. The specific technical novelties of our work is a video encoding method that preserves the temporal information and the Time Flexible Kernel that is able to compare sequences of different lengths and random alignment.

Our experiments demonstrated the value of the novel framework in two cases: First, low-level descriptors with short-term information lose the long-term temporal information of the sequences. Our framework is able to consider such temporal information and therefore can improve the performance in activity recognition. Second, although modern state-of-the-art descriptors like IDT include some temporal information for recognizing several activities in spite of the unordered encoding of BoF or FV, they fail in case of complex activities that are defined by the order of the same short events. Again, our framework is able to preserve such complex temporal structure and distinguish between activities that consist of similar events but in different order.

The novel formulation of TFK is not restricted to activity sequences but it can be applied in any comparison between two sets that their structure of information can be defined using the functions f_i and g_j. For instance, an interesting extension for future research will be its application in image-based recognition, where the spacial structure is an important source of information.

Finally, the TFK approach can introduce some noisy results if the number of low-level extracted features is small in some windows. Using several levels of granularity in window width reduces this effect.

The proposed approach and the obtained results have been published in the journal paper [Rodriguez et al., 2016b].

7 Discussion

7.1 Conclusions

Along the previous chapters we have dealt with the initial objectives of this thesis proposing and analysing some novel algorithms which improve the state-of-the-art, as results have demonstrated.

The raw video obtained from the activity recordings possesses lots of information unmanageable for most systems and some preprocessing may be useful in order to obtain a correct encoding of the activities, reducing the clutter and facilitating the differentiation among classes. Using some state-of-the-art encodings from the literature and adapting them to our objectives we have been able to develop different strategies to provide the desired information in the adequate format for each recognition method.

The learning of human activities with limited training examples has been covered in chapters 4 and 5. Based on HMM we have applied two different frameworks that assure a stable learning process of the activities. First, in a relaxed one-shot learning structure we have applied the FDHMM and we have been able to obtain encouraging results with FDHMM and with the use of a transfer learning stage. However, the used global features are difficult to extract and their results are far from the state-of-the-art approaches. Thus, a different framework is proposed to improve the results and reproducibility of the approach. This framework is designed for the use of local spatio-temporal features and applied to a strict one-shot learning, which is more useful in the real world. Called Simplex-HMM, the proposed framework obtains great results in the evaluated datasets and is an adequate method when working with limited data. The Simplex-HMM provides an stable training of the model with limited data and it benefits from a transfer learning of the features space trained with source domain sequences. However, the increase in training examples supposes a linear increase in computational cost and storage needs. We have reduced this computational cost with some improvements without losing accuracy in the classification but the linear increase of the cost is still present.

It is worth emphasizing the importance of an adequate modelling of the features space, verified in the experiments. This modelling might be difficult to obtain from limited information. Therefore, we propose the use of the extensive information available on the Internet and transfer the learning of the features space to the specific target scenario, which we have confirmed as a good solution.

The migration from fixed cameras to unconstrained video clips removes the need of training with limited number of sequences as the Internet provides a lot of video clips easily obtainable. In this case, the use of generative classifiers is not recommended while discriminative ones like SVM are. The introduction of long-term temporal information in a recognition system with unconstrained datasets, absent in most methods, has been developed in Chapter 6. The Time Flexible Kernel we have proposed allows a flexible alignment between sequences and the posterior application of a multi-class SVM. Using this framework and combining two levels of time granularity we have been able to outperform state-of-the-art results in four challenging datasets, proving the suitability of the TFK for differentiating among complex activities. The long-term temporal information inclusion in the recognition system is especially useful when some activities are differentiated by the order of the atomic actions.

7.2 Future work

While researching, the given solutions, sometimes, are as important as the new open questions that arise. In this regard, we continue with the reflection on some future work that the current research has left open.

In this thesis, the pre-segmentation of the activities has been assumed, but real world applications need to make this process. Thus, it would be interesting to evaluate the performance of the Simplex-HMM in a continuous video, segmenting the activities automatically. Moreover, the proposed learning of new activities is supervised what should be avoided for instance with a novelty threshold in the trained Simplex-HMM. So, future researches should focus on sequences segmentation and unsupervised learning.

The linear increase of computational cost on Simplex-HMM when increasing the training sequences should be avoided. Therefore, future work should investigate a smooth transition from the Simplex-HMM algorithm to other designed for several training sequences so that the method can start working initially but also adapt for new sequences while they are being recorded.

In the proposed TFK we have made a naive alignment between sequences coinciding the center of the sequences in the same time and aligning the rest accordingly. The flexibility provided by the K_{ST} kernel, in our proposal modelled with Gaussians, reduces the dependence of the alignment but a smarter alignment would improve the performance. In future work, it would be interesting to analyse the sequences and incorporate some kind of intelligence to this process.

An interesting property of the TFK framework is that, although being designed for time series, it can be adapted to any data structure just designing the K_{ST} kernel suitable for the data. For instance, it would be interesting in activity recognition to augment the granularity not only in time but also in space so time and space information is incorporated. Applications in image recognition would be a direct use of the adapted TFK as well.

Finally, it is worth noting that computationally expensive methods are getting easier to implement thanks to the hardware evolution. For instance, Deep Learning has gained interest since its great results in image recognition and currently is being applied in many fields of research. In activity recognition, Recurrent Neural Networks can be used for sequential data and Convolutional Neural Networks are being used for feature extraction. Still the computational cost is restrictive when working with videos although more methods are being implemented, and in the near future the use of these technologies looks promising for activity recognition. Attending to the proposals of this thesis, Deep Learning could be used for feature extraction and applied in both *Simplex-HMM* and TFK. It would be interesting to evaluate the performance of the systems using the novel feature extraction. This implementation of Deep Learning would not void the results of the thesis as it would be implemented in a stage where we have used literature methods which are not our contribution.

7.3 Publications

This thesis has led to several publications, some directly associated to the main objectives that has been already mentioned and some associated to side works. These publications are summarized below.

Journal papers

[**Rodriguez et al., 2016a**] Rodriguez, M., Orrite, C., Medrano, C., and Makris, D. *One-shot learning of human activity with an map adapted gmm and simplex-hmm.* IEEE Transactions on Cybernetics, PP(99):1–12. (2016)

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

[**Submitted**] Rodriguez, M., Orrite, C., Medrano, C., and Makris, D. *Extended Study for One-shot Learning of Human Activity by a Simplex-HMM.* IEEE Transactions on Cybernetics

Conference publications

[**Orrite et al., 2016**] Orrite, C., Rodriguez, M., Medrano, C. *One-shot learning of temporal sequences using a distance dependent Chinese Restaurant Process.* In Proceedings of the 23nd International Conference Pattern Recognition ICPR (Accepted December 2016)

[**Rodriguez et al., 2015**] Rodriguez, M., Medrano, C., Herrero, E., and Orrite, C. *Spectral Clustering Using Friendship Path Similarity* Proceedings of the 7th Iberian Conference, IbPRIA (June 2015)

[**Orrite et al., 2015**] Orrite, C., Soler, J., Rodriguez, M., Herrero, E., and Casas, R. *Image-based location recognition and scenario modelling.* In Proceedings of the 10th International Conference on Computer Vision Theory and Applications, VISAPP (March 2015)

[**Castán et al., 2014**] Castán, D., Rodríguez, M., Ortega, A., Orrite, C., and Lleida, E. *Vivolab and cvlab - mediaeval 2014: Violent scenes detection affect task.* In Working Notes Proceedings of the MediaEval (October 2014)

[**Orrite et al., 2014**] Orrite, C., Rodriguez, M., Herrero, E., Rogez, G., and Velastin, S. A. *Automatic segmentation and recognition of human actions in monocular sequences* In Proceedings of the 22nd International Conference Pattern Recognition ICPR (August 2014)

[**Rodriguez et al., 2013a**] Rodriguez, M., Medrano, C., Herrero, E., and Orrite, C. *Transfer learning of human poses for action recognition.* In 4th International Workshop of Human Behavior Unterstanding (HBU). (October 2013)

[**Rodriguez et al., 2013b**] Rodriguez, M., Orrite, C., and Medrano, C. *Human action recognition with limited labelled data.* In Actas del III Workshop de Reconocimiento de Formas y Analisis de Imagenes, WSRFAI. (September 2013)

[**Orrite et al., 2013**] Orrite, C., Monforte, P., Rodriguez, M., and Herrero, E. *Human Action Recognition under Partial Occlusions* . Proceedings of the 6th Iberian Conference, IbPRIA (June 2013)

[**Orrite et al., 2011**] Orrite, C., Rodriguez, M., and Montañes, M. *One sequence learning of human actions.* In 2nd International Workshop of Human Behavior Unterstanding (HBU). (November 2011)

Discusión

7.4 Conclusiones

A lo largo de los capítulos previos hemos tratado los objetivos iniciales de la tesis proponiendo y analizando nuevos algoritmos que mejoran el estado del arte, como han demostrado los resultados.

Las grabaciones en crudo de las diferentes actividades poseen mucha información que es inmanejable para la mayoría de los sistemas, por lo que puede ser útil realizar un pre-procesamiento de los datos a fin de obtener una codificación de las actividades adecuada, reduciendo el ruido y facilitando la distinción entre clases de actividades. hemos sido capaces de desarrollar diferentes estrategias para codificar los datos deseados en un formato adecuado par cada método de reconocimiento, adaptando para ello codificaciones descritas en la literatura.

En los capítulos 4 y 5 encontramos descritas las propuestas para el aprendizaje de actividades humanas a partir de escasos ejemplos de entrenamiento. Hemos aplicado dos estructuras diferentes basadas en HMM que nos aseguran la estabilidad en el proceso de entrenamiento. Primero, hemos aplicado FDHMM en un escenario relajado de aprendizaje con una secuencia (*relaxed one-shot learning*) y hemos observado las mejoras tanto al usar FDHMM directamente como al añadir una etapa de transferencia de aprendizaje. Sin embargo, las características globales empleadas son difícilmente extraibles y producen resultados alejados del estado del arte. Por tanto hemos propuesto un modelo diferente para mejorar los resultados y la reproducibilidad del método. Esta nueva estructura se ha diseñado para emplear características espacio-temporales locales y aplicar a un escenario estricto de aprendizaje con una secuencia (*strict one-shot learning*), lo cual es más útil en el mundo real. El método propuesto, llamado *Simplex-HMM*, ha obtenido muy buenos resultados en las bases de datos evaluadas y por tanto es un método adecuado para trabajar con pocas secuencias de entrenamiento. *Simplex-HMM* ofrece un entrenamiento del modelo estable con pocos datos y se beneficia de la adaptación de un modelo del espacio de características obtenido con datos externos. Sin embargo, al aumentar el número de secuencias de entrenamiento en coste computacional y de almacenaje aumenta linealmente. Hemos conseguido reducir este coste con varias mejoras y sin perder eficacia aunque el aumento lineal del coste sigue existiendo.

Merece la pena destacar la importancia de modelar adecuadamente el espacio de características, algo que se puede verificar en los resultados experimentales. Como la obtención de este modelo puede ser complicada si se dispone de información limitada, proponemos el uso de la vasta información disponible en Internet y transferir el modelo aprendido al escenario objetivo, lo cual se ha comprobado que es una buena solución.

El cambio entre el uso de cámaras fijas y la eliminación de restricciones en las grabaciones elimina la limitación de secuencias de entrenamiento, puesto que Internet provee de gran cantidad de vídeos fácilmente accesibles. En este caso se desaconseja el uso de clasificadores generativos en oposición de los discriminativos como SVM. En el capítulo 6 se desarrolla un método de inclusión de información temporal de larga duración en el sistema de reconocimiento que habitualmente es descartada. El método *Time Flexible Kernel* que proponemos permite un alineamiento flexible entre secuencias y el uso de un SVM multiclase. Usando este sistema y combinando dos niveles de granularidad en el tiempo hemos mejorado el estado del arte en cuatro bases de datos exigentes, probando que TFK es adecuado para distinguir entre actividades complejas. El uso de información a largo plazo es especialmente útil cuando la clase de las actividades se diferencia debido al orden de las sub-acciones que la conforman.

7.5 Trabajo Futuro

En investigación, tan importante como las soluciones dadas son las nuevas cuestiones que surgen. A este respecto, presentamos a continuación algunas reflexiones sobre posible trabajo futuro.

En esta tesis se ha asumido la pre-segmentación de las actividades, algo que en una sistema real debe ser hecho. Por lo tanto sería interesante evaluar *Simplex-HMM* en un vídeo continuo, donde la segmentación se realizase de forma autónoma. Además, el proceso de aprendizaje de nuevas actividades propuesto es supervisado lo que debería evitarse, por ejemplo introduciendo un umbral en el *Simplex-HMM* para detección de novedad. Así, un trabajo futuro debería enfocarse en segmentación y aprendizaje no supervisado.

En *Simplex-HMM* existe un incremento lineal del coste computacional al aumentar el número de secuencias de entrenamiento, y esto debería ser evitado. Para ello en el futuro se debería investigar una transición suave desde *Simplex-HMM* hasta otro algoritmo diseñado para muchas secuencias de entrenamiento, de modo que el sistema pueda empezar a trabajar desde el principio pero también adaptarse a las nuevas secuencias que se van grabando.

En el método propuesto de TFK hemos realizado un alineamiento naíf entre secuencias, haciendo coincidir el centro de ambas y alineando el resto correspondientemente. Gracias a la flexibilidad aportada por el kernel K_{ST}, utilizando nuestra propuesta de modelarlo como Gausianas, se ha conseguido reducir la dependencia de un alineamiento correcto, pero un método más inteligente mejoraría el funcionamiento. En un trabajo futuro sería interesante realizar un análisis de las secuencias previo para introducir algo de inteligencia a este proceso.

También sería interesante aprovechar que TFK, aunque inicialmente diseñado para series temporales, puede ser adaptado para cualquier estructura de datos tan solo diseñando el kernel K_{ST} adecuado. Por ejemplo, sería interesante añadir información tanto temporal como espacial para el reconocimiento de actividades. Su utilización en reconocimiento de imágenes sería otro uso directo adaptando los kernels de TFK.

Para finalizar es interesante destacar que la evolución del hardware está facilitando el uso de métodos cada vez más costosos computacionalmente. Por ejemplo, el empleo de *Deep Learning* ha ganado interés gracias a los magníficos resultados que ha conseguido en el reconocimiento de image y actualmente está siendo utilizado en muchos campos. En reconocimiento de actividades se pueden emplear redes neuronales recurrentes para aprender series temporales o en la etapa de extracción de características se pueden emplear redes neuronales convolucionales (CNN). El problema es que cuando se trabaja con vídeos el coste computacional sigue siendo excesivo en la mayoría de casos, aunque poco a poco van saliendo nuevos métodos que hacen que esta tecnología tenga un futuro prometedor. Atendiendo a lo expuesto en esta tesis, *Deep Learning* podría utilizarse para la extracción de características cuya aplicación sería válida tanto para *Simplex-HMM* como para TFK. Sería interesante evaluar el funcionamiento de ambos métodos con esta nueva extracción de características. Además, esta implementación de *Deep Learning* mantendría la validez de los resultados expuestos en la tesis puesto que afectaría a etapas donde hemos utilizado métodos de la literatura que no son contribución nuestra.

7.6 Publicaciones

Esta tesis ha propiciado varias publicaciones que se enumeran a continuación.

Artículos en Revistas

[Rodriguez et al., 2016a] Rodriguez, M., Orrite, C., Medrano, C., and Makris, D. *One-shot learning of human activity with an map adapted gmm and simplex-hmm.* IEEE Transactions on Cybernetics, PP(99):1–12. (2016)

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

[Submitted] Rodriguez, M., Orrite, C., Medrano, C., and Makris, D. *Extended Study for One-shot Learning of Human Activity by a Simplex-HMM.* IEEE Transactions on Cybernetics

Publicaciones en Conferencias

[Orrite et al., 2016] Orrite, C., Rodriguez, M., Medrano, C. *One-shot learning of temporal sequences using a distance dependent Chinese Restaurant Process.* In Proceedings of the 23nd International Conference Pattern Recognition ICPR (Accepted December 2016)

[Rodriguez et al., 2015] Rodriguez, M., Medrano, C., Herrero, E., and Orrite, C. *Spectral Clustering Using Friendship Path Similarity* Proceedings of the 7th Iberian Conference, IbPRIA (June 2015)

[Orrite et al., 2015] Orrite, C., Soler, J., Rodriguez, M., Herrero, E., and Casas, R. *Image-based location recognition and scenario modelling.* In Proceedings of the 10th International Conference on Computer Vision Theory and Applications, VISAPP (March 2015)

[Castán et al., 2014] Castán, D., Rodríguez, M., Ortega, A., Orrite, C., and Lleida, E. *Vivolab and cvlab - mediaeval 2014: Violent scenes detection affect task.* In Working Notes Proceedings of the MediaEval (October 2014)

[**Orrite et al., 2014**] Orrite, C., Rodriguez, M., Herrero, E., Rogez, G., and Velastin, S. A. *Automatic segmentation and recognition of human actions in monocular sequences* In Proceedings of the 22nd International Conference Pattern Recognition ICPR (August 2014)

[**Rodriguez et al., 2013a**] Rodriguez, M., Medrano, C., Herrero, E., and Orrite, C. *Transfer learning of human poses for action recognition.* In 4th International Workshop of Human Behavior Unterstanding (HBU). (October 2013)

[**Rodriguez et al., 2013b**] Rodriguez, M., Orrite, C., and Medrano, C. *Human action recognition with limited labelled data.* In Actas del III Workshop de Reconocimiento de Formas y Analisis de Imagenes, WSRFAI. (September 2013)

[**Orrite et al., 2013**] Orrite, C., Monforte, P., Rodriguez, M., and Herrero, E. *Human Action Recognition under Partial Occlusions* . Proceedings of the 6th Iberian Conference, IbPRIA (June 2013)

[**Orrite et al., 2011**] Orrite, C., Rodriguez, M., and Montañes, M. *One sequence learning of human actions.* In 2nd International Workshop of Human Behavior Unterstanding (HBU). (November 2011)

A
Video Encoding

This annexe describes the different base techniques used in the proposed algorithms of the thesis for encoding the video clips.

A.1 Descriptors

Following we describe two global descriptors (human silhouettes and Motion History Images) and two local spatio-temporal descriptors (Motion Interchange Patterns and Improved Dense Trajectories).

A.1.1 Silhouettes and Motion History Images

We define the Human Silhouette descriptor as a binary image, I, with zeros in the background pixels, $I(x, y) = 0$, and ones in the human body pixels, $I(x, y) = 1$. The human silhouette can be extracted using complex and robust methods, like the Laplacian fitting proposed by [Al-Maadeed et al., 2014], or using specific cameras like Kinect, which provides a depth map, making easier the segmentation of the human body, [Megavannan et al., 2012]. However, simple scenarios are needed for a proper running of the extraction and in many of

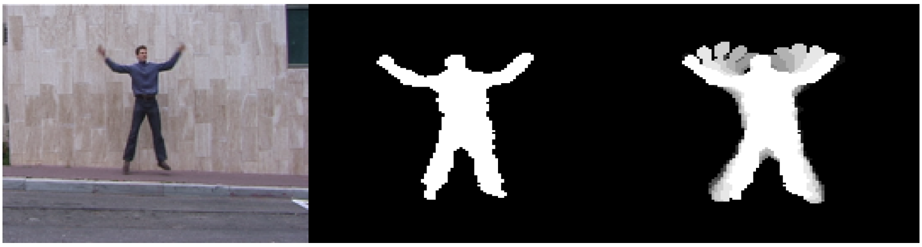


Figure A.1: Visualization of a frame (left), the human silhouette extracted from this frame (centre), and the MHI obtained from ten frames previous to this one (right).

these scenarios the simple background subtraction is reasonably robust. Two of the possible ways of modelling the background are using a mean filter or modelling each pixel as a RGB Gaussian. Afterwards, when the distance between the same pixel in a new frame and in the background exceeds a threshold we consider the pixel as foreground and background otherwise. The directly extracted foreground with background subtraction produces noisy frames which are cleaned with specific technique as for instance erosions and dilatations of the foreground blobs. It is possible to compensate the global motion produced in some activities (e.g., walking) so to emphasize the motion of the limbs and not the global motion. One technique is done by fitting a 2nd order polynomial to the center of mass trajectory, and aligning this trajectory to a reference point. Sometimes a bounding box of centred silhouettes is selected, as can be observed in Figure A.1.

The Motion History Image (MHI) [Bobick and Davis, 2001] technique captures motion information, specifically the history of motion in a sequence of frames. The input of the method is a binary image I representing the movement regions. However, as suggested by [Bobick and Davis, 2001], the computation with human body templates gives better performance and thus, the use of I representing the human silhouette is more appropriate. Therefore, we compute the MHI using the I binary frames obtained in the human silhouette extraction. Although MHI can be obtained from the whole video clip, we apply the encoding in temporal frame-windows Δ_l^n of N_l frames, as shown in Equation A.1. Large temporal frames-windows can produce excessive occlusions of the older frames information and therefore N_l should be short enough to minimize this problem.

$$h_{N_l}(x,y,t) = \begin{cases} 1 & if \quad I(x,y,t) = 1 \\ max\left(0, h_{N_l}(x,y,t-1) - \dfrac{1}{N_l}\right) & otherwise \end{cases} \tag{A.1}$$

A.1.2 Motion Interchange Patterns

For each pixel $p = (x, y, t)$, in the current frame t, an encoding using a triplet of frames (i.e., previous, current, and next) is obtained. As shown in Fig. A.2 these three frames are used to obtain the displacement of small patches of 3×3 pixels. The location of the center of the patch in the current frame is considered (0,0), and the eight possible locations in each of the previous and the next frames are denoted i and j (respectively) and numbered from 0 to 7. The eight index values correspond to center pixel locations of (-4,0), (-3,3), (0,4), (3,3), (4,0), (3,-3), (0,-4), and (-3,-3).

The possible locations of the patch in the three frames is encoded by combining i and j in a 64-trit (trinary digit) code denoted by $S(p)$. The images are considered as gray value images with intensities scaled between 0 and 255. Each trit $S_{ij}(p)$ is calculated by obtaining the sum of squared differences between the patch in previous frame located in i and the current patch (SSD1), and the patch in the next frame located in j and the current patch (SSD2) with the

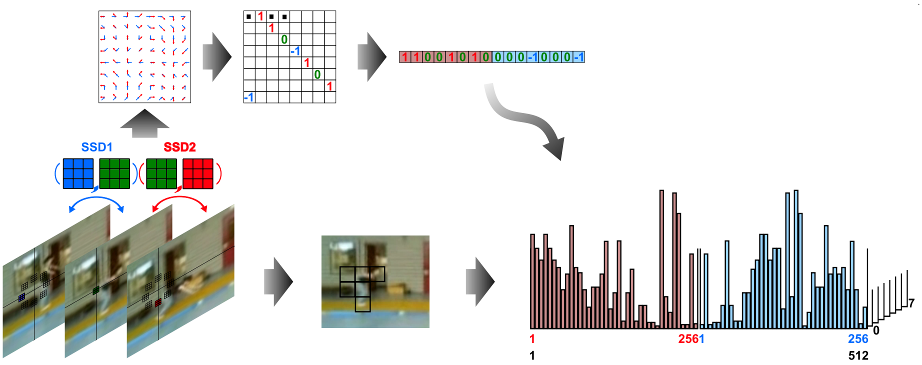


Figure A.2: MIP Encoding. Using a triplet of frames each pixel is encoded as a 8 × 8 matrix of 3 × 3 pixel patches displacements. Eight α channels are coded with 2 UINT8, which can be represented as 512-dimensional code words. Finally, using small 16 × 16 pixel patches encodings, 8 bag of words of 512 bins are created

following assignation:

$$S_{ij}(p) = \begin{cases} 1 & if \quad SSD1 - \theta > SSD2 \\ 0 & if \quad |SSD2 - SSD1| \leq \theta \\ -1 & if \quad SSD1 < SSD2 - \theta \end{cases} \tag{A.2}$$

The threshold is selected as $\theta = 1296$. The 64-trit matrix $S(p)$ obtained is divided into 8 channels each one representing a fix angle between i and j denominated $\alpha = 0, 45 \ldots 315$. Each channel is processed separately. The 8-trit per channel is represented separating the positive and negative parts obtaining 2 UINT8. Each UINT8 can be read as a number between 0 and 255 and concatenating them we obtain a 512 vector of zeros and 2 ones. Selecting patches of 16 × 16 pixels we obtain a histogram of 512 bins per channel which represents the MIP descriptor in that region. We have described the MIP descriptors slightly, but for a deeper understanding we refer to [Kliper-Gross et al., 2012].

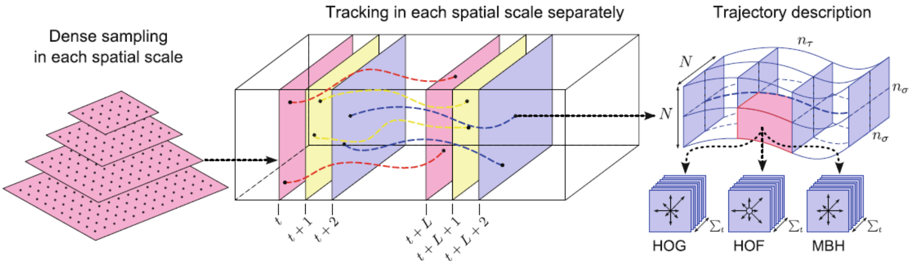


Figure A.3: Illustration of the Dense Trajectory extraction, [Wang et al., 2013]. *Left* Feature points are densely sampled on a grid for each spatial scale. *Middle* Tracking is carried out in the corresponding spatial scale for $N_l = L$ frames by median filtering in a dense optical flow field. *Right* The trajectory shape is represented by relative point coordinates, and the descriptors (HOG, HOF, MBH) are computed along the trajectory in a $N \times N$ pixels neighbourhood, which is divided into $n_\tau \times n_\sigma \times n_\sigma$ cells.

MIP descriptor code can be downloaded in `http://www.openu.ac.il/home/hassner/projects/MIP/MIPcode.zip`

A.1.3 Improved Dense Trajectories

IDT is a local spatio-temporal descriptor considered a mid-term temporal descriptor as it is computed along a trajectory that last several frames, specifically in the original papers $N_l = 15$. Following, we briefly specify the extractor method of this feature which is represented in Figure A.3.

The first process of the method is the tracking of densely sampled points along $N_l = 15$ frames. To do this, the dense optical flow field of frame I_t with respect to next frame I_{t+1} is computed using [Farnebäck, 2003] algorithm and obtaining $w_t = (u_t, v_t)$ where u_t is the horizontal component and v_t the vertical component. Once the optical flow is computed, points can be tracked without additional cost. The pixels are selected with a dense sampling of 5 pixels step at several scales incremented by a factor of $1/\sqrt{2}$. Trajectories from points in homogeneous image areas, without displacement, with an excessive displacement or truncated before the N_l frame are discarded. Once a trajectory is discarded or reaches the end, a new point is sampled to start a new trajectory.

In addition, the improved version of DT compensates the camera motion by estimating the background motion, assuming that two consecutive frames are related by a homography. Therefore, the matching between two consecutive frames is performed by matching SURF points [Bay et al., 2006] with nearest neighbour rule and matching points tracked with optical flow previously selected with the good-features-to-track criterion [Shi and Tomasi, 1994]. Then, RANSAC [Fischler and Bolles, 1981] is used to select the best matching, discarding spurious, with which the homography is estimated. Using the homographies, consecutive frames are wrapped and the optical flow is computed in those wrapped frames so to discard trajectories obtained by the camera motion. In the original paper of IDT, [Wang and Schmid, 2013] also use a human detector to avoid interference between human motion and camera motion but, as it implies a previous stage increasing the computational cost and complexity, we avoid it in our implementation. Each extracted trajectory is encoded concatenating the sequence of points obtained $(P_t, P_{t+1} \dots P_{t_{N_l}})$.

The trajectory information is complemented with three descriptors obtained in the surrounding cuboid. First, a cuboid of $N \times N \times N_l$, with $N = 32$ pixels and $N_l = 15$ frames is selected in the neighbourhood of the trajectory, and divided into a spatio-temporal grid of size $n_\tau \times n_\sigma \times n_\sigma$, with $n_\tau = 3$ subdivisions in time and $n_\sigma = 2$ subdivisions in space. In each sub-cuboid of the grid HOGs, HOFs [Laptev et al., 2008] and the Motion Boundary Histograms (MBH) descriptors [Dalal et al., 2006] are computed. HOG orientations are quantized in 8 bins, while HOF orientations are quantizing in 9 bins including a zero bin when the displacement is small. Full orientation and magnitude of each vector is used for weighting. The HOG descriptor adds 96 elements while HOF adds 108. MBH is the histogram of the gradients of optical flow $w = (u, v)$ and it is computed twice, one per component. Therefore, quantizing the gradient into 8 bins, MHBx is obtained from horizontal component and MHBy from vertical component. Each of the descriptors add 96 elements, obtaining at the end a descriptor of size $(30 + 96 + 108 + 96 + 96) = 426$. The HOG descriptors are obtained in the cuboid surrounding the original trajectory, as they only depend on the image information, while HOF and MBH are obtained wrapping every two consecutive frames. Finally, each of the descriptors components are individually normalized using RootSIFT method [Arandjelović and Zisserman, 2012].

For an in deep knowledge of the method it is better to read the original papers of DT [Wang et al., 2013] and IDT [Wang and Schmid, 2013]. IDT descriptor code can be downloaded in `http://lear.inrialpes.fr/people/wang/download/improved_trajectory_release.tar.gz`.

A.2 Clustering

The information created and digitalized every day coming from multiple sources like video recording, users meta-data, texting, etc, is laboriously manageable and therefore, the automatic labelling of the information into meaningful groups facilitates its usability. Clustering can roughly be defined as the discovering of natural groups of samples coming from lager set of samples. However, there is not a consensus of a specific definition and depending on the objective task different clustering techniques are appropriate. Comprehensive surveys related to this field are found in [Xu and Wunsch, 2005] [Schaeffer, 2007] and [Jain, 2010].

Despite the large number of methods available in the literature, we have use the the traditional methods of K-means and Gaussian Mixture Models (GMM) optimization. K-means creates a Voronoi Tessellation of the features space through an iterative optimization method while the GMM is created usually with an Expectation Maximization process that maximizes the likelihood of the model. Both methods find hyperellipsoidal clusters and their optimization processes only find local optima [Jain, 2010]. Figure A.4 depicts the Voronoi

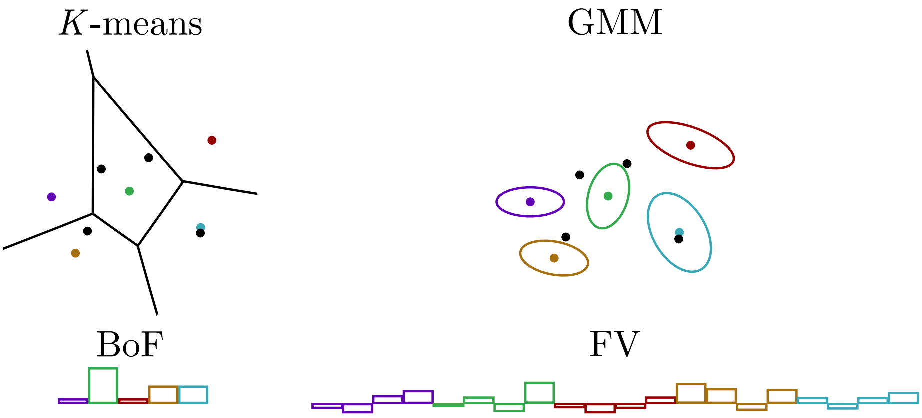


Figure A.4: *First row:* Voronoi tessellation of a K-means and K ellipsoids representing the Gaussians of the GMM, in a 2-dimensional space with $K = 5$. Coloured points represent the centroids means and black points represent the samples to encode. *Second row:* 5-bins histogram representation of a BoF, and $2(mean) \times 2(variance) \times 5(cluster)$ vector of FV values.

tessellation created by a K-means clustering and the ellipsoid representation of the Gaussians composing a GMM. We find an exhaustive study of both methods in [Bishop, 2006], although we briefly describe them below.

A.2.1 K-means

Given a data set $Q = \{\mathbf{q}_1 \dots \mathbf{q}_N\}$ of n-dimensional points, the goal of K-means is to find a partition of K clusters that minimize the sum of squared errors obtained between each data point and the centre of the assigned cluster. We can use any error measure, but the most common, and the used by us, is the Euclidean distance. We formally define $\boldsymbol{\mu} = \{\mu_k\}, k = 1 \dots K$ as the centres of the clusters. We introduce the parameter $r_{nk} \in \{0, 1\}$ where $n = 1 \dots N$ and $k = 1 \dots K$ which represents if the point n belongs to cluster k, $r_{nk} = 1$, or not,

$r_{nk} = 0$. The objective function is therefore defined as:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{q}_n - \mu_k\|^2 \tag{A.3}$$

The minimizing of the objective function is a known NP-hard problem and then a numerical iterative method is performed, which only assures the convergence to a local minima:

1. Select initial values for the cluster centres $\boldsymbol{\mu}^1$ and compute the objective function J.

2. Assign each sample of the dataset to the cluster with the closest centre:

$$r_{nk} = \begin{cases} 1 & if \qquad k = \operatorname{argmin}_j \|\mathbf{q}_n - \mu_j^t\|^2 \\ 0 & otherwise \end{cases} \tag{A.4}$$

3. Compute new cluster centres:

$$\mu_k^{t+1} = \frac{\sum_{n=1}^{N} r_{nk} \mathbf{q}_n}{\sum_{n=1}^{N} r_{nk}} \tag{A.5}$$

4. Evaluate the objective function J with the new centres and repeat from point 2 until convergence.

The number of clusters K and the initial cluster centres $\boldsymbol{\mu}^1$ are input parameters that should be provided to the method. The initialization of the centres is a key factor for reaching the optimum and a usual method for reducing its impact is to repeat the process with several randomly initializations, selecting the trial that reaches the minimum in the objective function.

A.2.2 Gaussian Mixture Model

A GMM is composed by K Gaussians $\{\lambda_k(\mu_k, \Sigma_k)\}$, where μ_k represents the mean of the k-th Gaussian and Σ_k the covariance, and the mixing coefficients $\boldsymbol{\omega}$ obeying $\sum_{k=1}^{K} \omega_k = 1$, where ω_k is the k-th mixing coefficient. A GMM is formally defined as $\lambda = \{\omega_k, \mu_k, \Sigma_k\}$. and is a model more complex than the clustering obtained with K-means, as in addition to the cluster centres $\boldsymbol{\mu}$ it includes covariances and mixing components. These additional parameters complicate the training but provide a better modelling of the data space through the marginal distribution of the random variable $\mathbf{q}$.

$$p(\mathbf{q}) = \sum_{k=1}^{K} p(\lambda_k)p(\mathbf{q}|\lambda_k) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{q}|\mu_k, \Sigma_k) \tag{A.6}$$

Suppose that the dataset $Q = \{\mathbf{q}_1 \dots \mathbf{q}_N\}$ is now used to train the GMM $\lambda = \{\omega_k, \mu_k, \Sigma_k\}$. This objective is fulfil by finding the GMM parameters that maximize the log-likelihood of the dataset in the model, defined as:

$$\ln(p(Q|\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma})) = \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{q}_n|\mu_k, \Sigma_k)\right) \tag{A.7}$$

Unfortunately, the solution of the likelihood maximization cannot be obtained analytically and therefore, a suboptimal iterative process is used. The elegant and powerful method Expectation Maximization (EM) is the most commonly used approach, although it only finds a local optimum following the subsequent steps:

1. Initialize the GMM parameters. The mixing coefficients $\boldsymbol{\omega}^1$, the means $\boldsymbol{\mu}^1$ and the covariances $\boldsymbol{\Sigma}^1$. Using these initial parameter, evaluate the log-likelihood.

2. **E-step**. Evaluate the responsibilities $\gamma_n(k)$, name given to the posterior probability that component λ_k was responsible for generating $\mathbf{q}_n$, using

the current parameters.

$$\gamma_n(k) = \frac{\omega_k \mathcal{N}(\mathbf{q}_n|\mu_k^t, \Sigma_k^t)}{\sum_{j=1}^{K} \omega_j^t \mathcal{N}(\mathbf{q}_n|\mu_j^t, \Sigma_j^t)} \tag{A.8}$$

3. **M-step**. Re-estimate parameters using the obtained responsibilities.

$$N_k = \sum_{n=1}^{N} \gamma_n(k) \tag{A.9}$$

$$\mu_k^{t+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_n(k)\mathbf{q}_n \tag{A.10}$$

$$\Sigma_k^{t+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_n(k)(\mathbf{q}_n - \mu_k^{t+1})(\mathbf{q}_n - \mu_k^{t+1})^T \tag{A.11}$$

$$\omega_k = \frac{N_k}{N} \tag{A.12}$$

4. Evaluate the log-likelihood and repeat from step 2 until convergence.

Like in K-means, the EM algorithm needs the number of Gaussians K fixed and the initialization parameters, conditioning the found optimum.

Both methods have been deeply study and several approaches developed. As we are going to work with large datasets we have used developed code for this problematic. In particular we have used the VLFeat library, available in `http://www.vlfeat.org/`.

A.3 Holistic Descriptors

Two of the most common encodings are Bag of Words (BoW), also called Bag of Features (BoF), and Fisher Vectors (FV). To perform them, both processes compute a clustering of the descriptors extracted from training examples in order to define the feature space. This clustering represents the "dictionary" of features used as scaffold of the descriptors. BoF can use any clusterings while FV uses a GMM. Following we describe how BoF and FV work.

A.3.1 Bag of Features

BoF is a simple method coming from text documents classification, where originally was called Bag of Words, consisting in counting the frequency of word appearance provided a vocabulary. In the video encoding process the words are features and the vocabulary is created using a clustering of training features. A graphical representation is shown at the beginning of the second row in Figure A.4. Formally, let $Q = \{\mathbf{q}_1 \dots \mathbf{q}_N\}$ be a set of features extracted from a video, and $C = \{c_1 \dots c_K\}$ the vocabulary created through a clustering algorithm. We recover the parameter $r_{nk} \in \{0, 1\}$ where $r_{nk} = 1$ if sample $\mathbf{q}_n$ belongs to cluster c_k and $r_{nk} = 0$ otherwise. Using **K**-means we have described how to compute r_{nk} in Equation A.4. We define the vector $\mathbf{r}_n = \{r_{n1} \dots r_{nK}\}$ which obeys $\sum_{k=1}^{K} r_{nk} = 1$. Therefore, the BoF value is calculated with:

$$BoF = \frac{1}{N} \sum_{n=1}^{N} \mathbf{r}_n \tag{A.13}$$

When the number of samples is reasonably large, the original BoF model obtained with a hard-assignment is acceptable. However, in our approaches we work with limited data as we compute the BoF in windows of size N_w where the number of features N can be small and therefore, the loss of the spatial information in the descriptors space produced by this rough assignation is unacceptable. Soft-assignment has been proven a good improvement representing continuous data with a codebook model [van Gemert et al., 2010] and then in order to cope with the BoF limitation we apply a soft-assignment. We define s_{nk} as the value assigned from sample $\mathbf{q}_n$ to cluster c_k. Similar to hard-assignment we define the vector $\mathbf{s}_n = \{s_{n1} \dots s_{nK}\}$ with $s_{nk} \geq 0$ and $\sum_{k=1}^{K} s_{nk} = 1$. Therefore, the

BoF value is computed accordingly:

$$BoF = \frac{1}{N}\sum_{n=1}^{N}\mathbf{s}_n \tag{A.14}$$

The vocabulary can be obtained with both, K-means algorithm or GMM optimization, and both, hard and soft assignment computed. However, the most common algorithms combine hard-assignment with K-means and soft-assignment with GMM.

A.3.2 Fisher Vectors

Fisher Vector, [Perronnin and Dance, 2007] [Perronnin et al., 2010], is a feature pooling technique used in document encoding that recently has gained attention in image and video encoding, providing the best performance in recent works, as for instance in [Sanchez et al., 2013], [Oneata et al., 2013] or [Wang and Schmid, 2013]. Unlike BoF, FV encodes both first and second order statistics between the features and the GMM, recording for each Gaussian the mean and variance statistics per dimension. A graphical representation is shown at the end of the second row in Figure A.4.

Given the set of features $Q = \{\mathbf{q}_1 \dots \mathbf{q}_N\}$ and a GMM $\lambda = \{\omega_k, \mu_k, \Sigma_k\}$ describing the features space, the associated FV is computed as follows.

First, the posterior probability of sample $\mathbf{q}_n$ to Gaussian λ_k is computed:

$$s_{nk} = \frac{e^{-\frac{1}{2}(\mathbf{q}_n-\mu_k)\Sigma_k^{-1}(\mathbf{q}_n-\mu_k)^T}}{\sum_{j=1}^{K} e^{-\frac{1}{2}(\mathbf{q}_n-\mu_j)\Sigma_j^{-1}(\mathbf{q}_n-\mu_j)^T}} \tag{A.15}$$

We can make a connection with the soft-assignment explained in BoF, as the posterior probability is one of the possibilities there.

For each Gaussian k, the mean and covariance deviation vectors are computed:

$$u_{jk} = \frac{1}{N\sqrt{\omega_k}} \sum_{n=1}^{N} s_{nk} \frac{q_{jn} - \mu_{jk}}{\sigma_{jk}} \tag{A.16}$$

$$v_{jk} = \frac{1}{N\sqrt{\omega_k}} \sum_{n=1}^{N} s_{nk} \left[\left(\frac{q_{jn} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \tag{A.17}$$

Finally, the FV encoding is formed by the concatenation of the deviation vectors, composing a vector of size $2KD$.

$$FV = \begin{bmatrix} \vdots \\ \mathbf{u}_k \\ \vdots \\ \mathbf{v}_k \\ \vdots \end{bmatrix} \tag{A.18}$$

In some cases the encoding is complemented with the information provided by the soft-assignment BoF computed with the s_{nk} values in a vector of size $K(2D+1)$. The FV can be improved trough normalizations as proposed by [Perronnin et al., 2010] so, we apply a power and L2 normalizations.

$\mathcal{B}$

Simplex-HMM Analytical Demonstrations

B.1 Proof of EM convergence with a pseudo probability observation emission function in a HMM

Given an observation emission function $b_j(O_t) \geq 0$ that does not represent a PDF the convergence of the EM algorithm remains as proved below following the proof given in [Bishop, 2006]. As $b_j(O_t)$ does not define a probability distribution we call it pseudo-probability distribution and change the notation to $b_j(O_t) = \Upsilon(O_t|z_t = S_j, B)$. Using this notation, we define the joint pseudo-probability distribution $\Upsilon(\mathcal{O}, Z|\theta)$ over the observed variable given the latent variable Z and the HMM parameters.

$$\Upsilon(\mathcal{O}, Z|\theta) = p(z_1|\pi) \left[\prod_{t=2}^{T} p(z_t|z_{t-1}, A)\right] \prod_{t=1}^{T} \Upsilon(O_t|z_t, B) \tag{B.1}$$

and the pseudo-likelihood is defined as:

$$\Upsilon(\mathcal{O}|\theta) = \sum_{Z} \Upsilon(\mathcal{O}, Z|\theta) = \sum_{Z} \pi_{z_1} \left[\prod_{t=2}^{T} a_{z_{t-1} z_t} \right] \prod_{t=1}^{T} b_{z_t}(O_t) \tag{B.2}$$

Using the previous definitions, we define the $p(Z|\mathcal{O}, \theta)$ probability function by normalizing EquationB.1.

$$p(Z|\mathcal{O}, \theta) = \frac{\Upsilon(\mathcal{O}, Z|\theta)}{\Upsilon(\mathcal{O}|\theta)} = \frac{\Upsilon(\mathcal{O}, Z|\theta)}{\sum_{Z} \Upsilon(\mathcal{O}, Z|\theta)} \tag{B.3}$$

We introduce the distribution $q(Z)$ defined over the latent variable. For any choice of $q(Z)$ we can decompose the pseudo - log likelihood as:

$$\ln(\Upsilon(\mathcal{O}|\theta)) = \mathcal{L}(q, \theta) + KL(q||p) \tag{B.4}$$

where

$$\mathcal{L}(q, \theta) = \sum_{Z} q(Z) \ln \left\{ \frac{\Upsilon(\mathcal{O}, Z|\theta)}{q(Z)} \right\} \tag{B.5}$$

$$KL(q||p) = -\sum_{Z} q(Z) \ln \left\{ \frac{p(Z|\mathcal{O}, \theta)}{q(Z)} \right\} \tag{B.6}$$

In order to prove this decomposition we apply the following process.

First, we derive from Equation B.3

$$\Upsilon(\mathcal{O}, Z|\theta) = p(Z|\mathcal{O}, \theta)\Upsilon(\mathcal{O}|\theta) \tag{B.7}$$

and then the decomposition in Equation B.4 is verified by applying the substitution of $\Upsilon(\mathcal{O}, Z|\theta)$.

$$\ln(\Upsilon(\mathcal{O}|\theta)) = \sum_Z q(Z) \ln\left\{\frac{p(Z|\mathcal{O},\theta)\Upsilon(\mathcal{O}|\theta)}{q(Z)}\right\} - \\ \sum_Z q(Z) \ln\left\{\frac{p(Z|\mathcal{O},\theta)}{q(Z)}\right\} \quad \text{(B.8)}$$

$KL(q||p)$ is cancelled leaving the right-hand term as $\sum_Z q(Z) \ln \Upsilon(\mathcal{O}|\theta)$. After noting that $q(Z)$ sums to 1, the decomposition is proved.

$KL(q||p)$ is the Kullback-Leibler divergence between $q(Z)$ and the posterior distribution $p(Z|\mathcal{O},\theta)$, being $KL(q||p) \geq 0$ with equal to 0 if, and only if, $q(Z) = p(Z|\mathcal{O},\theta)$. Therefore, $\mathcal{L}$ is a lower bound as $\mathcal{L}(q,\theta) \leq \ln(\Upsilon(\mathcal{O}|\theta))$.

The EM algorithm uses two steps to perform the maximization. Using the decomposition in Equation B.4 and following both steps we can demonstrate the maximization of the pseudo-log likelihood. Supposing that θ is the current value of parameters, the E step maximizes the lower bound $\mathcal{L}(q,\theta)$ with respect to $q(Z)$. As $\ln(\Upsilon(\mathcal{O}|\theta))$ does not depend on $q(Z)$ that maximization is achieved by vanishing $KL(q||p)$, which happens when $q(Z) = p(Z|\mathcal{O},\theta)$. At this point, $\ln(\Upsilon(\mathcal{O}|\theta)) = \mathcal{L}(q,\theta)$.

If we substitute $q(Z) = p(Z|\mathcal{O},\theta)$ in Equation B.4, considering Equation B.5 and Equation B.6, $KL(q||p)$ disappears, leaving only the lower bound, $\mathcal{L}(q,\theta)$. As $q(Z)$ is fixed to the old parameter θ we use the notation θ where $q(Z)$ is substituted by $p(Z|\mathcal{O},\theta)$ and $\hat{\theta}$ otherwise.

$$\mathcal{L}(q,\hat{\theta}) = \sum_Z p(Z|\mathcal{O},\theta) \ln \Upsilon(\mathcal{O},Z|\hat{\theta}) - \\ \sum_Z p(Z|\mathcal{O},\theta) \ln p(Z|\mathcal{O},\theta) = Q(\hat{\theta},\theta) + const \quad \text{(B.9)}$$

In this equation the first term is the Baum's auxiliary function and the second term is the negative entropy of the q distribution which is constant in $\hat{\theta}$.

The M step keeps $q(Z)$ fixed and maximizes $\mathcal{L}(q,\hat{\theta})$ with respect to $\hat{\theta}$ obtaining θ^{new}. This will increase $\mathcal{L}$ unless it is a maximum. Moreover, $KL(q||p)$ was vanished using the old parameters and the new $p(Z|\mathcal{O},\theta^{new})$ is different to $q(Z)$, giving $KL(q||p)$ a positive value and increasing even more the value of $\ln(\Upsilon(\mathcal{O}|\theta))$.

As we have seen in Equation B.9 the maximization of $\mathcal{L}(q,\theta)$ is equivalent to maximize the Baum's auxiliary function, that for HMM is:

$$Q(\hat{\theta},\theta) = \sum_{j=1}^{N}\gamma_1(j)\ln\pi_j + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\xi_t(i,j)\ln a_{ij} + \sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(j)\ln(b_j(O_t)) \quad \text{(B.10)}$$

B.2 Optimization of different observation models

In an HMM, different observation models suppose changes in the term of the Baum's auxiliary function defined in the following equation:

$$\sum_t\sum_j\gamma_t(j)\ln(b_j(O_t)) \quad \text{(B.11)}$$

The maximization of the model likelihood given an observation implies the maximization of Equation B.11 which varies with the observation model. Following we detail the optimization of this term of the Baum's auxiliary function for the observation models: EOM, HOM and BOM.

B.2.1 Euclidean Observation Model (EOM)

$$b_j(O_t) = e^{-\varphi\sqrt{\sum_{k=1}^{K}(v^t_{\lambda_k}-m_{jk})^2}} \tag{B.12}$$

where φ is a constant and $v^t_{\lambda_k}$ the bin values of the histograms.

Substituting Equation B.12 into Equation B.11 we obtain:

$$\sum_t \sum_j \gamma_t(j)\left(-\varphi\sqrt{\sum_{k=1}^{K}(v^t_{\lambda_k}-m_{jk})^2}\right) \tag{B.13}$$

Then, Equation B.13 has to be maximized with respect to m_{jk}. By setting $\frac{\partial}{\partial m_{jk}} = 0$, the following equation is obtained:

$$\varphi\sum_{t=1}^{T}\gamma_t(j)\frac{(v^t_{\lambda_k}-m_{jk})}{\sqrt{\sum_{k'=1}^{K}(v^t_{\lambda_{k'}}-m_{jk'})^2}} = 0 \tag{B.14}$$

Since m_{jk} is independent on t and the $\gamma_t(j)$ values are treated as constants in the M-step once computed in the E-step, Equation B.15 can be easily derived:

$$m_{jk} = \frac{\sum_{t=1}^{T}\gamma_t(j)\frac{v^t_{\lambda_k}}{\sqrt{\sum_{k'=1}^{K}(v^t_{\lambda_{k'}}-m_{jk'})^2}}}{\sum_{t=1}^{T}\gamma_t(j)\frac{1}{\sqrt{\sum_{k'=1}^{K}(v^t_{\lambda_{k'}}-m_{jk'})^2}}} \tag{B.15}$$

Since m_{jk} is on the left and on the right side, the equation is solved by a fixed point iteration, obtaining the optimal values $\hat{m}_{jk}$ when convergence is achieved. The process is explained in Algorithm 1.

It should be noted that the optimization process can be performed separately for each j.

Algorithm 1

```
Randomly initialize $m_{jk} \ni \sum_{k=1}^{K} m_{jk} = 1$
$\epsilon_m = \infty$
while $\epsilon_m > \epsilon$ do
    $m'_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j) \frac{v_{\lambda_k}^t}{\sqrt{\sum_{k'=1}^{K} (v_{\lambda_{k'}}^t - m_{jk'})^2}}}{\sum_{t=1}^{T} \gamma_t(j) \frac{1}{\sqrt{\sum_{k'=1}^{K} (v_{\lambda_{k'}}^t - m_{jk'})^2}}}$
    $\epsilon_m = \max_k |m_{jk} - m'_{jk}|$
    $m_{jk} = m'_{jk}$, $k = 1...K$
end while
```

B.2.2 Hellinger Observation Model (HOM)

$$b_j(O_t) = e^{-\varphi \sqrt{\sum_{k=1}^{K} (\sqrt{v_{\lambda_k}^t} - \sqrt{m_{jk}})^2}} \tag{B.16}$$

Doing a change in notation Equation B.16 is equal to B.12 and the optimization process is equivalent to the Algorithm 1 with the only change of the vector constraints. The difference in the model is kept by these constraints. The terms $\sqrt{v_{\lambda_k}^t}$ can be named without the square root so $\sqrt{v_{\lambda_k}^t} \rightarrow \tilde{v}_{\lambda_k}^t$. This change implies a change in the constraints so $\sum_{k=1}^{K} v_{\lambda_k} = 1 \rightarrow \sum_{k=1}^{K} \tilde{v}_{\lambda_k}^2 = 1$. On the other hand, the same change can be done for the terms $\sqrt{m_{jk}}$, naming them $\tilde{m}_{jk}$ directly.

B.2.3 Bhattacharyya Observation Model (BOM)

$$b_j(O_t) = \sum_{k=1}^{K} \sqrt{v_{\lambda_k}^t} \sqrt{m_{jk}} \tag{B.17}$$

It is possible to eliminate the notation of the square root in the free parameters transforming $\sqrt{m_{jk}} \rightarrow \tilde{m}_{jk}$ as long as $\sum_{k=1}^{K} \tilde{m}_{jk}^2 = 1$ and $\tilde{m}_{jk} \geq 0$ without loss of generalization. Then, the substitution of Equation B.17 after the transformation into Equation B.11 leads to the objective function to maximize Equation B.18:

$$\sum_t \sum_j \gamma_t(j) \ln \left(\sum_{k=1}^{K} \sqrt{v_{\lambda_k}^t} \tilde{m}_{jk} \right) \tag{B.18}$$

The incorporation of the restriction $\left(1 - \sum_{k=1}^{K} \tilde{m}_{jk}^2 = 0 \; \forall j\right)$ into Equation B.18 using Lagrange multipliers modifies the objective function to maximize as follows:

$$\sum_t \sum_j \gamma_t(j) \ln \left(\sum_{k=1}^{K} \sqrt{v_{\lambda_k}^t} \tilde{m}_{jk} \right) + \sum_j \lambda_j \left(1 - \sum_{k=1}^{K} \tilde{m}_{jk}^2 \right) \tag{B.19}$$

This new function is analytically derived by setting $\frac{\partial}{\partial \tilde{m}_{jk}} = 0$, obtaining the following result:

$$\sum_t \gamma_t(j) \frac{\sqrt{v_{\lambda_k}^t}}{\sum_{k'=1}^{K} \sqrt{v_{\lambda_{k'}}^t} \tilde{m}_{jk'}} - 2\lambda_j \tilde{m}_{jk} = 0 \tag{B.20}$$

later it is partially derived again by setting $\frac{\partial}{\partial \lambda_j} = 0$, obtaining the following equation:

$$1 - \sum_{k=1}^{K} \tilde{m}_{jk}^2 = 0 \tag{B.21}$$

As λ_j is independent on k it can be considered some sort of normalization coefficient. Then, the maximization of Equation B.11 with the incorporated restriction leads to:

$$\tilde{m}_{jk} \propto \sum_{t=1}^{T} \gamma_t(j) \frac{\sqrt{v_{\lambda_k}^t}}{\sum_{k'=1}^{K} \sqrt{v_{\lambda_{k'}}^t} \tilde{m}_{jk'}} \tag{B.22}$$

and the maximization process can be solved with the fixed point iteration of Algorithm 2 until convergence (that is, when the difference between two consecutive solutions is less than a predefined value ϵ):

Algorithm 2

```
Randomly initialize $\tilde{m}_{jk} \ni \sum_{k=1}^{K} \tilde{m}_{jk}^2 = 1$
$\epsilon_m = \infty$
while $\epsilon_m > \epsilon$ do
  $\gamma_t'(j) = \frac{\gamma_t(j)}{\sum_{k'=1}^{K} \sqrt{v_{\lambda_{k'}}^t} \tilde{m}_{jk'}}$
  $\tilde{m}_{jk}' = \sum_{t=1}^{T} \gamma_t'(j) \sqrt{v_{\lambda_k}^t}$, $k = 1...K$
  $\tilde{m}_{jk}'' = \frac{\tilde{m}_{jk}'}{\sqrt{\sum_{k'=1}^{K} \tilde{m}_{jk'}'^2}}$, $k = 1...K$
  $\epsilon_m = \max_k |\tilde{m}_{jk} - \tilde{m}_{jk}''|$
  $\tilde{m}_{jk} = \tilde{m}_{jk}''$, $k = 1...K$
end while
```

It should be noted that the optimization process can be done separately for each j.

Bibliography

Al-Maadeed, S., Almotaeryi, R., Jiang, R., and Bouridane, A. (2014). Robust human silhouette extraction with laplacian fitting. *Pattern Recognition Letters*, 49:69 – 76.

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

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., and Baskurt, A. (2011). Sequential deep learning for human action recognition. In *Human Behavior Understanding*, pages 29–39.

Bagno, S. (1953). Method and apparatus for detecting motion in a confined space. US Patent 2655645.

Batra, D., Chen, T., and Sukthankar, R. (2008). Space-time shapelets for action recognition. In *IEEE Workshop on Motion and video Computing (WMVC)*, pages 1–6.

Bay, H., Tuytelaars, T., and Gool, L. V. (2006). Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417.

Bian, W., Tao, D., and Rui, Y. (2012). Cross-domain human action recognition. *IEEE Transactions on Systems, Man, and Cybernetics. B Cybernetics*, 42(2):298–307.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA.

bo Duan, K. and Keerthi, S. S. (2005). Which is the best multiclass svm method? an empirical study. In *International Workshop on Multiple Classifier Systems*, pages 278–285.

Bobick, A. and Davis, J. (1996). Real-time recognition of activity using temporal templates. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 39–42.

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

Campbell, C. and Cristianini, N. (1998). Simple learning algorithms for training support vector machines. Technical report.

Cao, L., Liu, Z., and Huang, T. S. (2010). Cross-dataset action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1998–2005.

Cao, L., Mu, Y., Natsev, A., Chang, S.-F., Hua, G., and Smith, J. (2012). Scene aligned pooling for complex video recognition. In *European Conference on Computer Vision (ECCV)*.

Castán, D., Rodríguez, M., Ortega, A., Orrite, C., and Lleida, E. (2014). Vivolab and cvlab - mediaeval 2014: Violent scenes detection affect task. In *Working Notes Proceedings of the MediaEval 2014 Workshop*.

Chaudhry, R., Ravichandran, A., Hager, G. D., and Vidal, R. (2009). Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1932–1939. IEEE.

Chen, L., Hoey, J., Nugent, C., Cook, D., and Yu, Z. (2012). Sensor-based activity recognition. *IEEE Transactions onSystems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(6):790–808.

Choi, J., Wang, Z., Lee, S.-C., and Jeon, W. J. (2013). A spatio-temporal pyramid matching for video retrieval. *Computer Vision and Image Understanding*, 117(6):660 – 669.

Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., Lamarca, A., Legrand, L., Rahimi, A., Rea, A., Bordello, G., Hemingway, B., Klasnja, P., Koscher, K., Landay, J., Lester, J., Wyatt, D., and Haehnel, D. (2008). The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing*, 7(2):32–41.

Cook, D., Feuz, K., and Krishnan, N. (2013). Transfer learning for activity recognition: a survey. *Knowledge and Information Systems*, pages 1–20.

Cottone, P., Maida, G., and Morana, M. (2013). User activity recognition via kinect in an ambient intelligence scenario. In *International Conference on Applied Computing, Computer Science, and Computer Engineering (ICACC)*, volume 7, pages 49 – 54.

Cula, O. and Dana, K. (2001). Compact representation of bidirectional texture functions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–1041–I–1047 vol.1.

Cuturi, M. (2011). Fast global alignment kernels. In *International Conference on Machine Learning (ICML)*.

Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. (2008). Self-taught clustering. In *International Conference on Machine Learning (ICML)*, pages 200–207.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893 vol. 1.

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

Danafar, S. and Gheissari, N. (2007). Action recognition for surveillance applications using optic flow and svm. In *Asian Conference on Computer Vision (ACCV)*, pages 457–466.

Daumé, III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1):101–126.

Ding, D., Cooper, R. A., Pasquina, P. F., and Fici-Pasquina, L. (2011). Sensor technology for smart homes. *Maturitas*, 69(2):131 – 136.

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *International Conference on Computer Communications and Networks (ICCCN)*, pages 65–72.

Fanello, S. R., Gori, I., Metta, G., and Odone, F. (2013). Keep it simple and sparse: real-time action recognition. *Journal of Machine Learning Research*, 14(1):2617–2640.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis (SCIA)*, pages 363–370.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:594–611.

Feng, X. and Perona, P. (2002). Human action recognition by sequence of movelet codewords. In *First International Symposium on 3D Data Processing Visualization and Transmission*, pages 717–721.

Fine, S., Singer, Y., and Tishby, N. (1998). The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32(1):41–62.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

Gaidon, A. (2012). *Structured Models for Action Recognition in Real-word Videos*. Theses, Université de Grenoble.

Gaidon, A., Harchaoui, Z., and Schmid, C. (2012). Recognizing activities with cluster-trees of tracklets. In *British Machine Vision Conference (BMVC)*, pages 30.1–30.13.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

Gunawardana, A., Mahajan, M., Acero, A., and Platt, J. C. (2005). Hidden conditional random fields for phone classification. In *International Conference on Speech Communication and Technology*. International Speech Communication Association.

Han, J. and Bhanu, B. (2005). Human activity recognition in thermal infrared imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17–17.

Hu, D. H., Zheng, V. W., and Yang, Q. (2011). Cross-domain activity recognition via transfer learning. *Pervasive and Mobile Computing*, 7:344–358.

INE (2014). Mujeres y hombres en España. Technical report, Catálogo de publicaciones de la Administración General del Estado.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Jebara, T., Kondor, R., and Howard, A. (2004). Probability product kernels. *Journal of Machine Learning Research*, 5:819–844.

Jebara, T., Song, Y., and Thadani, K. (2007). Spectral clustering and embedding with hidden markov models. In *European Conference on Machine Learning (ECML)*.

Jiang, Y.-G., Dai, Q., Xue, X., Liu, W., and Ngo, C.-W. (2012). Trajectory-based modeling of human actions with motion reference points. In *European Conference on Computer Vision (ECCV)*.

Karaman, S., Benois-Pineau, J., Dovgalecs, V., Mégret, R., Pinquier, J., André-Obrecht, R., Gaëstel, Y., and Dartigues, J.-F. (2014). Hierarchical hidden markov model in detecting activities of daily living in wearable videos for studies of dementia. *Multimedia Tools and Applications*, 69(3):743–771.

Ke, S.-R., Thuc, H. L. U., Lee, Y.-J., Hwang, J.-N., Yoo, J.-H., and Choi, K.-H. (2013). A review on video-based human activity recognition. *Computers*, 2(2):88.

Ke, Y., Sukthankar, R., and Hebert, M. (2005). Efficient visual event detection using volumetric features. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 166–173 Vol. 1.

Kläser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference (BMVC)*, pages 995–1004.

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

Kolovou, X. and Maglogiannis, I. (2010). Video-surveillance and context aware system for activity recognition. In *International Conference on Pervasive Technologies Related to Assistive Environments (PETRA)*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105.

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

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, pages 282–289.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, 64:107–123.

Laptev, I. and Lindeberg, T. (2004). Local descriptors for spatio-temporal recognition. In *In First International Workshop on Spatial Coherence for Visual Motion Analysis*.

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

Lara, O. and Labrador, M. (2013). A survey on human activity recognition using wearable sensors. *Communications Surveys Tutorials, IEEE*, 15(3):1192–1209.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178.

Le, Q. V., Zou, W. Y., Yeung, S. Y., and Ng, A. Y. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3361–3368.

Li, W., Yu, Q., Divakaran, A., and Vasconcelos, N. (2013). Dynamic pooling for complex event recognition. In *IEEE International Conference on Computer Vision (ICCV)*.

Li, Z., Wei, Z., Yue, Y., Wang, H., Jia, W., Burke, L. E., Baranowski, T., and Sun, M. (2015). An adaptive hidden markov model for activity recognition based on a wearable multi-sensor device. *Journal of Medical Systems*, 39(5):57.

Liu, J., Luo, J., and Shah, M. (2009). Recognizing realistic actions from videos "in the wild". *IEEE Conference on Computer Vision and Image Understanding (CVPR)*.

Liu, J., Shah, M., Kuipers, B., and Savarese, S. (2011). Cross-view action recognition via view knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3209–3216.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157 vol.2.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Lu, W.-L. and Little, J. (2006). Simultaneous tracking and action recognition using the pca-hog descriptor. In *Canadian Conference on Computer and Robot Vision*, pages 6–6.

Lv, F. and Nevatia, R. (2006). Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *European Conference on Computer Vision (ECCV)*, pages 359–372.

Megavannan, V., Agarwal, B., and Venkatesh Babu, R. (2012). Human action recognition using depth maps. In *Signal Processing and Communications (SPCOM), 2012 International Conference on*, pages 1–5.

Minka, T. P. (2009). Estimating a Dirichlet distribution. Technical report.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press.

Nguyen, T.-H.-C., Nebel, J.-C., and Florez-Revuelta, F. (2016). Recognition of activities of daily living with egocentric vision: A review. *Sensors*, 16(1):72.

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

Niebles, J. C., Chen, C.-W., and Fei-Fei, L. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision (ECCV)*, pages 392–405.

Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., Aggarwal, J. K., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., and Desai, M. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3153–3160.

Oneata, D., Verbeek, J., and Schmid, C. (2013). Action and event recognition with fisher vectors on a compact feature set. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1817–1824.

Orrite, C., Monforte, P., Rodriguez, M., and Herrero, E. (2013). Human action recognition under partial occlusions. In Sanches, J. M., Micó, L., and Cardoso, J. S., editors, *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, pages 398–405, Berlin, Heidelberg. Springer Berlin Heidelberg.

Orrite, C., Rodriguez, M., Herrero, E., Rogez, G., and Velastin, S. A. (2014). Automatic segmentation and recognition of human actions in monocular sequences. In *International Conference on Pattern Recognition (ICPR)*, pages 4218–4223.

Orrite, C., Rodriguez, M., and Medrano, C. (2016). One-shot learning of temporal sequences using a distance dependent chinese restaurant process. In *International Conference on Pattern Recognition (ICPR)*.

Orrite, C., Rodriguez, M., and Montañes, M. (2011). One-sequence learning of human actions. In *Human Behavior Unterstanding*, pages 40–51.

Orrite, C., Soler, J., Rodríguez, M., Herrero, E., and Casas, R. (2015). Image-based location recognition and scenario modelling. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 216–221.

Padilla-Lopez, J. R., Chaaraoui, A. A., and Florez-Revuelta, F. (2015). Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9):4177 – 4195.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572.

Perronnin, F. and Dance, C. R. (2007). Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society.

Perronnin, F., Sánchez, J., and Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, pages 143–156.

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

Quattoni, A., Wang, S., Morency, L.-P., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Ragheb, H., Velastin, S., Remagnino, P., and Ellis, T. (2008). ViHASi: virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods. In *ACM Workshop on Vision Networks for Behavior Analysis (VNBA)*, pages 77–84.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007). Self-taught learning: Transfer learning from unlabeled data. In *International Conference on Machine Learning (ICML)*, pages 759–766.

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

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41.

Rodriguez, M., Medrano, C., Herrero, E., and Orrite, C. (2013a). Transfer learning of human poses for action recognition. In *Human Behavior Understanding.*

Rodriguez, M., Medrano, C., Herrero, E., and Orrite, C. (2015). *Spectral Clustering Using Friendship Path Similarity*, pages 319–326. Cham.

Rodriguez, M., Orrite, C., and Medrano, C. (2013b). Human action recognition with limited labelled data. In *Actas del III Workshop de Reconocimiento de Formas y Analisis de Imagenes, WSRFAI.*

Rodriguez, M., Orrite, C., Medrano, C., and Makris, D. (2016a). One-shot learning of human activity with an map adapted gmm and simplex-hmm. *IEEE Transactions on Cybernetics*, PP(99):1–12.

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

Rodriguez-Serrano, J. A. and Singh, S. (2012). Trajectory clustering in CCTV traffic videos using probability product kernels with hidden markov models. *Pattern Analysis & Applications*, 15:415–426.

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

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523.

Sanchez, J., Perronnin, F., Mensink, T., and Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245.

Schaeffer, S. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition (ICPR)*.

Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *International Conference on Multimedia*, pages 357–360.

Seo, H. J. and Milanfar, P. (2011). Action recognition from one example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):867–882.

Shi, J. and Tomasi, C. (1994). Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593 – 600.

Shimodaira, H., ichi Noma, K., Nakai, M., and Sagayama, S. (2002). Dynamic time-alignment kernel in support vector machine. In *Advances in Neural Information Processing Systems (NIPS).*

Shinozaki, T. and Ostendorf, M. (2008). Cross-validation and aggregated em training for robust parameter estimation. *Computer Speech & Language*, 22(2):185 – 195.

Shoaib, M., Bosch, S., Incel, O. D., Scholten, H., and Havinga, P. J. (2015). A survey of online activity recognition using mobile phones. *Sensors*, 15(1):2059.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, pages 568–576.

Singla, G. and Cook, D. J. (2009). Interleaved activity recognition for smart home residents. In *International Conference on Intelligent Environments*, pages 145–152.

Sivic, J. and Zisserman, A. (2009). Efficient visual search of videos cast as text retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):591–606.

Sminchisescu, C., Kanaujia, A., and Metaxas, D. (2006). Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2):210–220.

Stork, J., Spinello, L., Silva, J., and Arras, K. (2012). Audio-based human activity recognition using non-markovian ensemble voting. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 509–514.

Su, X., Tong, H., and Ji, P. (2014). Activity recognition with smartphone sensors. *Tsinghua Science and Technology*, 19(3):235–249.

Sun, L. and Aizawa, K. (2013). Action recognition using invariant features under unexampled viewing conditions. In *ACM International Conference on Multimedia*, pages 389–392.

Tang, K., Fei-Fei, L., and Koller, D. (2012). Learning latent temporal structure for complex event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1250–1257.

Todorovic, S. (2012). Human activities as stochastic kronecker graphs. In *European Conference on Computer Vision (ECCV)*.

Turaga, P., Chellappa, R., Subrahmanian, V., and Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488.

Uguz, H., Ozturk, A., Saracoglu, R., and Arslan, A. (2008). A biomedical system based on fuzzy discrete hidden markov model for the diagnosis of the brain diseases. *Expert Systems with Applications*, 35(3):1104–1114.

Vahdat, A., Cannons, K., Mori, G., Oh, S., and Kim, I. (2013). Compositional models for video event detection: A multiple kernel learning latent variable approach. In *IEEE International Conference on Computer Vision (ICCV)*.

van Gemert, J. C., Veenman, C. J., Smeulders, A. W. M., and Geusebroek, J. M. (2010). Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283.

Veeraraghavan, A., Roy-Chowdhury, A., and Chellappa, R. (2005). Matching shape sequences in video with applications in human movement analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1896–1909.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103:60–79.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision (ICCV)*.

Wang, H., Ullah, M. M., Kläser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference (BMVC)*, page 127.

Wang, J., Liu, P., She, M. F., and Liu, H. (2011). Human action categorization using conditional random field. In *Robotic Intelligence In Informationally Structured Space (RiiSS)*.

Wang, L. and Suter, D. (2007). Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257.

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

Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 20–27. IEEE.

Xu, D. and Chang, S.-F. (2008). Video event recognition using kernel methods with multilevel temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1985–1997.

Xu, R. and Wunsch, D., I. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645 –678.

Xu, Z., Yang, Y., and Hauptmann, A. G. (2015). A discriminative CNN video representation for event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, Y., Saleemi, I., and Shah, M. (2013). Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1635–1648.

Yao, B. and Zhu, S.-C. (2009). Learning deformable action templates from cluttered videos. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1507–1514.

Yilmaz, A. and Shah, M. (2005). Actions sketch: a novel action representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 984–989.

Zhang, J. and Gong, S. (2010). Action categorization with modified hidden conditional random field. *Pattern Recognition*, 43(1):197–203.

Zhou, F., De la Torre Frade, F., and Hodgins, J. K. (2013). Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:582–596.

Zhu, Y., Zhao, X., Fu, Y., and Liu, Y. (2011). Sparse coding on local spatial-temporal volumes for human action recognition. In *Asian Conference on Computer Vision (ACCV)*.