DEGREE OF DOCTOR OF PHILOSOPHY IN
COMPUTER ENGINEERING AND SCIENCE

DOCTORATE SCHOOL IN
INFORMATION AND COMMUNICATION TECHNOLOGIES

XXVI Cycle

UNIVERSITY OF MODENA AND REGGIO EMILIA

Department of Engineering "Enzo Ferrari"

Ph.D. DISSERTATION

Learning Visual Models from Incomplete Data

Candidate:
DALIA COPPI
Advisor:
PROF. RITA CUCCHIARA
The Director of the School:
PROF. GIORGIO MATTEO VITETTA

DOTTORATO DI RICERCA IN
COMPUTER ENGINEERING AND SCIENCE

SCUOLA DI DOTTORATO IN
INFORMATION AND COMMUNICATION TECHNOLOGIES

XXVI Ciclo

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Ingegneria "Enzo Ferrari"

TESI PER IL CONSEGUIMENTO DEL TITOLO DI DOTTORE DI RICERCA

Apprendimento Automatico di Modelli Visuali con Dati Incompleti

Tesi di:
DALIA COPPI
Relatore:
PROF. RITA CUCCHIARA
Il Direttore della scuola:
PROF. GIORGIO MATTEO VITETTA

Learn from yesterday,
live for today,
hope for tomorrow.
The important thing is
not to stop questioning.
Curiosity has its own reason
for existing.

Albert Einstein

ACKNOWLEDGMENTS

This thesis is the results of three years of PhD, three years of work, of collaboration and discussion but also support of many great people.

Firstly I would like to express my gratitude to the *Imagelab group*, a group made of brilliant people and also good friends. These are the people that most influenced my PhD, thanks to them I learnt the secrets of Pattern Recognition and Machine Learning, I started to appreciate programming and I discovered that curiosity is one of the most important skills in research. Thanks, especially, to Marco Manfredi, Michele Fornaciari, Davide Baltieri and Daniele Borghesani that made this experience enjoyable and unforgettable. Daniele, in particular, deserves an individual thanks for the countless discussions on whichever topic, for the criticisms and the suggestions. Thanks also to Francesco Solera, the machine learning lover, for his not-so-funny jokes, to Martino Lombardi for the entertainment with his immense and diverse knowledge about life, to Fabio Battilani. Thanks to Costantino Grana, the C++ expert, Giuseppe Serra, and his Florentine accent, Roberto Vezzani, Simone Calderara, Andrea Prati and Paolo Santinelli. Finally thanks to Prof. Rita Cucchiara, that makes all this possible and especially that offered me the opportunity to undertake the experience of a PhD.

Scientifically speaking my gratitude goes also to the *CVSSP* of the University of Surrey, where I have spent eight months of my PhD. My deepest gratitude goes to Prof. Josef Kittler, Teo de Campos and Fei Yan that guided my

research and provided me with very valuable opinions. I should also express my appreciation to my officemates that welcomed me and my approximate English language skills: Charles Gray, Stuart James and the intruder Cemre Zor. Many other people made the experience in UK unforgettable, Daniele Raví, Luca Comi, Nazli Faraji Davar, Rangika Abeygunasekera and finally the space-guys Elisabetta Iorfida and Andrea Turconi.

Outside of the academic world the most intense gratitude goes to Fabio, my love. Without him I would not have gone so far, he always supported and encouraged me, he followed me in any of my ideas and choices, even the weirdest, such as take a one-way flight to London, he has been simply essential for me. Thanks also to my family, my parents Ermanna and Angelo and my sister Eleonora, they guided me, went along with my choices and concretely supported my studies. I would like to express my gratitude also to my grandparents, with their being strong even when they were supposed to be the weakest, they demonstrate me the importance of keep trying in front of each difficulty. Finally, outside University, I should also thank my oldest friends, Sara, Silvia, Simona, Daniele and Federico.

Grazie.

Dalia Coppi,
Modena, February 2014

ABSTRACT

The goal of a learning system is to capture patterns and regularities in training data which allow for future classification. Machine learning methods are able to generalize a classification model from labelled training data but difficulties arise when the distribution of the training data is not explicitly modelled. Real world applications offer a massive amount of visual data, but unfortunately labelled data are not always easy to find and the labelling process is costly and time consuming or may not be possible for a lack of knowledge. This work is focused on the learning of discriminative visual models in scenarios with partially annotated or incomplete data. With incomplete data we refer either to the case where only a subset of the training data is labelled or where only a fraction of the training classes is known. We evaluate the problem of learning from incomplete data in three separate computer vision applications, namely people tracking, novel image classification and document image analysis.

In video surveillance the input of a tracking system might be interpreted as a set of partially labelled data where there are only few annotated instances of the target and several not annotated samples. Not annotated test data might also deviate from training data because of occlusions, changes in pose or appearance making the target association problem challenging. We exploit a semi supervised learning method to solve the problem of people tracking and we demonstrate with an experimental analysis the effectiveness of the proposed approach.

Regarding image categorization, an interesting challenge is represented by the detection of novel categories and subcategories of objects. Assuming that objects can be organized in taxonomies, the instances to be classified may differ from the hierarchy learned from training data and they might share only parent nodes. Our work is devoted to derive a learning model from labelled data able to generalize over data coming from classes not seen during training.

Finally, the last part addresses the picture segmentation in document images of old books. Dealing with the layout segmentation of old documents results in a variety of pictorial elements, thus in the difficulty of being able to collect samples representative of this heterogeneity. We propose an effective feature representation and a Support Vector Machines classification along with an experimental evaluation that demonstrate an improvement over baseline methods of document layout analysis even if a detailed model of the input space is not available.

SOMMARIO

L'obiettivo di un sistema di apprendimento automatico è catturare la struttura e le regolarità presenti nei dati in ingresso in modo da permettere la classificazione di dati futuri. I metodi di apprendimento artificiale sono in grado di astrarre modelli di classificazione da dati di training precedentemente annotati, ma riscontrano difficoltà quando la distribuzione di tali dati non è esplicitamente modellata. Una considerevole quantit di dati visuali è oggi disponibile in varie applicazioni, le difficoltà, sfortunatamente, risiedono nell'avere a disposizione dati annotati e nella possibilit di etichettare i dati sulla base delle risorse di tempo disponibili o della conoscenza accessibile. Questa tesi è focalizzata sull'apprendimento automatico di modelli discriminativi in scenari con una scarsa disponibilit di dati annotati o con dati incompleti. Con dati incompleti ci riferiamo sia al caso in cui solamente un sottoinsieme dei dati di ingresso sia annotato, sia al caso in cui solo una frazione delle classi di addestramento sia annotata. Il problema dell'apprendimento automatico con dati parzialmente etichettati stato qui valutato in tre diverse applicazioni nel campo della visione artificiale, ovvero localizzazione e inseguimento di persone, classificazione di nuove categorie di immagini e analisi di immagini di documenti.

Nella video sorveglianza l'input di un sistema di tracking può essere visto come un insieme di dati solo parzialmente annotati, dove sono presenti alcuni esempi del target da seguire e diversi esempi non etichettati. Tali dati non etichettati possono discostarsi anche notevolmente dal modello dei dati anno-

tati a causa di occlusioni, cambiamenti di posa o di illuminazione, rendendo il problema di associazione tra dati etichettati e non ancora pi complicato. In questa tesi viene proposto un metodo di apprendimento automatico semi supervisionato per risolvere il problema di inseguimento di persone e viene dimostrato mediante unanalisi sperimentale lefficacia della soluzione proposta.

Riguardo alla classificazione di immagini, un'interessante sfida rappresentata dall'individuazione di nuove categorie e sottocategorie di oggetti. Assumendo che gli oggetti siano organizzati in tassonomie, può verificarsi il caso in cui gli elementi da classificare differiscano dalla gerarchia appresa o condividano solo parte dei nodi parentali. Il lavoro è qui dedicato allapprendimento di un modello dai dati di training che sia in grado di generalizzare anche su classi non viste durante la fase di apprendimento.

Infine, l'ultima parte affronta la segmentazione di figure in scansioni di testi antichi e il recupero di immagini simili da altre sorgenti. Lavorare sulla segmentazione di documenti datati risulta in una considerevole quantit di elementi illustrativi e quindi nella difficoltà di avere a disposizione esempi rappresentativi di questa eterogeneit. Viene proposta una rappresentazione efficace delle caratteristiche delle immagini e lutilizzo di Support Vector Machines come metodo di classificazione. L'uso di queste due tecniche ha condotto ad un miglioramento nei confronti di altri metodi esistenti anche nel caso in cui un modello dettagliato dei dati di training non disponibile.

CONTENTS

CHAPTER

ONE

INTRODUCTION

Many practical application in computer vision and pattern recognition deal with a massive amount of data coming from a large number of input sources. While the data acquisition is usually simply and reliable, annotation and classification of those data is slow and expensive. Equivalently, we can state that unlabelled data are readily available while labelled data are difficult to obtain, due to these reasons the problem of dealing with incomplete data is of increasing importance.

The term incomplete data may refer both to incompleteness in the observation X, *missing features*, or in the labels Y, *incomplete labels*. The problem of missing features can result from many scenarios, for example applications in activity recognition, affect recognition and all that applications where information from multiple sensors is fused. Some of the hardware and algorithms associated with some of the nodes might fail occasionally, leading to missing features and making the standard pattern recognition machinery unusable on that chunk of data.

When the incompleteness of the data is reflected on the labels a further categorization can be used and incomplete label problems can be decomposed in:

- **Partially labelled data** This is probably the most general definition and refers to the case where labels are difficult and expensive to obtain but observations are available at training time. This problem is usually addressed in semi-supervised classification, where only a part of the available data is annotated and the learning systems exploits both labelled and unlabelled data to solve the classification problem.

- **Coarse labels** There are some scenarios where labels are available with a level of abstraction higher than the real definition, this occurs especially when dealing with hierarchies and taxonomies. For example in activity recognition, each activity can be broken down into sub-activities (*e.g* a service in a tennis game is composed by multiple movement of the player), in object recognition images can be divided in sub-parts (*e.g* a lion is made of head, body, legs and tail), or again in object categorization a category is composed of subcategories (*e.g* a car can be a sport car, a compact car, a family car). Sometimes it might be unclear, not possible or not necessary to select the subclasses, thus, the training data might be just annotated for the higher level classes and the challenge of the learning system is to exploit only the coarse labels even if it is often easier to learn the single components.

- **Missing labels** This last problem occurs when the availability of class labels is limited to a subset of the classes we need to detect. Even if training objects are provided with the corresponding annotation, the issue is that the training set does not cover all the possible classification categories. There are several circumstances where this problem arises, for instance when the aim is to detect unexpected behaviours of people or crowds only knowing the definition of normal behaviour, or in applications of intrusion detection, where the intrusion is, of course, not modelled by training data. In machine learning the problem of learning with missing labels is addressed by anomaly and novelty detection, and aims to identify events that do not conform to the expected pattern learned from the training set.

- **Noisy labels** Finally, there are scenarios where the labels provided might be incorrect or there is an uncertainty about training labels. For example in large scale object categorization a recent trend is to exploit web based tagging, however tags might only vaguely correspond to the visual object or might be wrong and correspond to concept only related to the one contained in the image but not to the concept itself. Or, some data might have incorrect labels due to annotation mistakes.

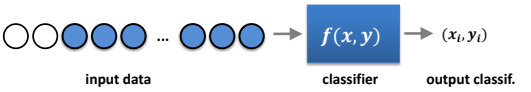
In the rest of the thesis we will consider problem arisen when learning with the first three categories of incomplete data.

1.1 Contribution of this work

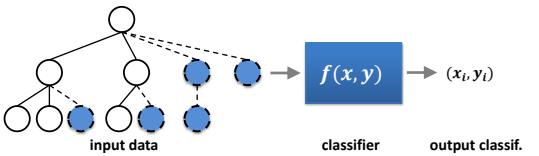
The thesis provides an analysis of the learning of visual models when dealing with incomplete data. Figure 1.1 gives a graphical overview of the three cases of learning with input data only partially annotated. All the three models assume a set of input data $X = \{x_1, \ldots, x_k, \ldots, x_n\}$ with corresponding labels $Y \in \{y_1, \ldots, y_k\}$ with $k < n$, the differences rely in the structure of the input data.

In the figure shaded elements represent unlabelled elements, and dotted lines distinguish elements not available during training. In the first case (a) only few examples are labelled while a large number of input instances are not labelled, the learner is given both labelled and unlabelled elements, collectively referred as *partially labelled*. In (b) all the elements given to the classifier are labelled but there are *missing labels*, or in other words the training set is not representative of all the possible labels. Finally in (c) we can broadly define the input data as *coarse labelled* because the labels are given at a high level of abstraction without modelling all the existing subcategories. The coarse training might also be not comprehensive of all the existing subcategories.

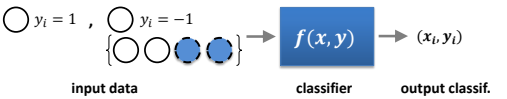
The problem addressed in this work is motivated by real world problems in Computer Vision, we thus analyse and propose solution to application scenarios instead of focusing on the problem from a theoretically perspective. The



(a)



(b)



(c)

Figure 1.1: Three different examples of learning with incomplete data that are addressed in this thesis. Shaded elements are classes without input labels while dotted lines distinguish elements not available during training.

main contribution of this work are represented by

a graph based Transductive Learning approach for people tracking in unconstrained scenarios Coppi et al. [2011a,b]. With the term unconstrained scenario we consider any case study without posing restrictions on the motion pattern, the type of camera, the light condition, *etc.*. Assuming that the target is known and is visible in the first frames of the considered videos we only have few labelled elements and a large number of unlabelled elements, thus we formulate the problem exploiting graph transduction and we aim at iteratively propagate the information from labelled to unlabelled nodes. We also introduced **an evolutionary**

update strategy to enforce the knowledge encoded in the labelled elements Coppi et al. [2011a] and an **extension to multi target tracking**. Using an iterative approach and updating the labelled elements allows to enforce time continuity in the tracking process and guarantees robustness to appearance changes and occlusions of the target. We also extend the solution to a multi target scenario, by using a formulation of the graph transduction based on game theoretic notions previously introduced in literature.

a hierarchical discriminative framework to detect novel categories and subcategories of visual objects Coppi et al. [2014b]. Assuming that visual object form a taxonomy where similar categories share the same parent node we propose this approach based on Support Vector Machines that aims at detecting novel categories as incongruence at different level of classification. We show how this method is effective while there is a tight relation between the visual appearance of the objects and the conceptual taxonomy, but we also demonstrate a weakness when this assumption is relaxed.

a method for illustration segmentation in digitized documents based on a discriminative classification of local correlation features Coppi et al. [2014a]. Illustration segmentation is considered as a binary classification problem where illustrations are only coarse labelled despite their heterogeneity (*e.g* drawings, charts, photos, sketches, etc.). Textual and pictorial regions are characterized by different local patterns, specifically textual regions have a regular and horizontal structure. We exploit the autocorrelation function to compute local descriptors able to emphasize the regular patterns of textual regions and the multi-modal distributions of pictorial elements and we build a discriminative model using Support Vector Machines.

1.2 Outline of the thesis

The rest of the thesis is organized in three main parts. *Chapter 2* starts with the motivations of the thesis, then overviews the main challenges of learning with incomplete annotation and the state-of-the-art of the proposed solutions. The second part, instead, gives a brief review of classification techniques highlighting the differences between supervised and unsupervised learning.

Then, we examine the problem of learning with incomplete data in three different computer vision scenarios stating the problems and the background and reporting experimental results for each proposed solution.

Part I examines the problem of tracking seen as a data association problem. *Chapter 3* introduces the people tracking in video surveillance and explains our solution based on graph transduction. More precisely it describes the theoretical derivation and properties of the method pointing out the motivations because the problem can be considered as one of partially labelled data and underling the tight relation with semi-supervised learning. Starting from single target tracking the consideration is extended to multi target in *Chapter 4*. Experiments on several dataset are reported showing the effectiveness of the method.

In *Part II*(*Chapter 5*) we consider the classification of visual object in semantic categories, specifically we focus on the problem of learning with missing labels. We introduce the topic of novelty detection describing some related works. We explore the possibility of having a taxonomy of visual categories where only a sub-part of the tree is known during training, and we illustrate our proposal for the detection of novel categories and subcategories based on a hierarchical structure of Support Vector Machines. Thorough experiments show the performance of the proposed approach with different classifiers settings.

The third part, regards document analysis, and specifically *Chapter 6* focuses on the extraction of images from digitized historical books. The page segmentation and image extraction steps are described and an effective feature representation bases on local correlation is introduced. The challenge of this problem is represented by the diversification of illustrations that share the same

coarse label (*i.e* photos, drawings, charts, etc.). The effectiveness of this descriptor in detecting the repeating patterns of text regions and differentiating them from pictorial elements is proven by experiments on two different dataset and with a comparison against the state of the art.

Finally, in *Chapter 7* the thesis summarizes the problem of learning discriminative models with partially labelled data, pointing out the main aspects arisen in the analysis of the three presented problem and concludes possible extensions and future works.

CHAPTER

TWO

LEARNING WITH INCOMPLETE DATA

The following introductory chapter aims at motivating the thesis topic and explaining the basic principles used. The first part analyses the challenge encountered by learning system when dealing with a limited amount of training data, or when the training data are incomplete. A large part also concentrates on the description of the Machine Learning techniques we used to tackle the problem.

2.1 Motivation

A typical learning framework is usually based on a set of input data and their corresponding annotation. A vectorial feature representation of the input elements is extracted with pattern recognition techniques and the task of the learning system is to identify the repeating patterns and structure in the input data and to build a model according with the input labels. Future data are then classified using the model learned from training data (as shown in Fig. 2.1).

Unfortunately there are many scenarios that far exceed from this simplistic paradigm, often training data are not enough, or are plagued by incompleteness, as introduced in Chapter 1, leading to a challenging classification. For example when working with large datasets, is infeasible to obtain labels for all

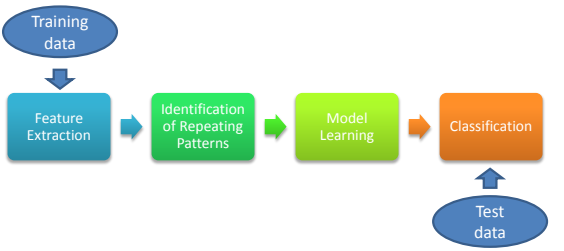


Figure 2.1: The basics steps of a learning system.

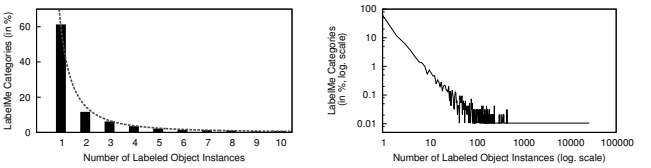


Figure 2.2: LabelMe statistics.

examples, and is also unlikely to have a large number of labelled examples for all the classes represented in the dataset. Analysing current large-scale dataset for object recognition, such as *LabelMe* [Russell et al. [2008]], is evident how the problem of the lack of training data is more common than one might expect, Fig. 2.2 shows the number of labelled instances for each training category, and demonstrate how over the 60% of all categories only have one single labelled instance. Similar statistics can be computed also for other datasets. For this reason pattern recognition system should deviate from the traditional assumption that all relevant classes are known or adequately represented.

Problems due to lack of adequate training exist also in industrial and multimedia application. The data might contain errors or might only be partially complete. In automatic quality control, for instance, it is possible to encounter a new blemish or fault that was not experienced previously and thus that was not available during the training step. The system should then be able to recognize all the elements that differ from the expected behaviour rather than being

able to classify only known failure. In multimedia retrieval, where the images might have been annotated using automatic tools (for example exploiting the content of the web page) is common that some objects are only vaguely specified or are wrongly specified, and are therefore considered uncertain in their representation.

Conversely to machine learning tools that barely generalize imprecise data, humans are able to correctly classify data even in severe situations. Our learning capacity allows us to reliably and rapidly assimilate different kind of regularities and this enable us to make inductive inference even from very small amount of data. We are able to generalize and learn even from small set of data and we can identify new objects/patterns and elements even without any prior knowledge. Biederman [1987b] evidenced how humans know approximately 30000 visual categories, which correspond to learn 5 new categories a day in our childhood. Moreover we can recognizes a large number of diverse objects despite large variations in the object's position, pose, lighting and background clutter. We can easily segment an object, analyse its shape and track it. Building computational systems that are able to achieve the same performance is extremely hard.

In this thesis we aim at reducing the gap between humans and machines when working with missing or coarse labels or few training example. We work on real world pattern recognition problems and we try to bridge the gap by learning powerful discriminative models on the training data.

2.2 Challenges and literature review

The motivations of the previous section pointed out how learning with incomplete and partially labelled data is of increasing interest in pattern recognition. In this section we give a structured overview of the main challenges and of the proposed solution in literature.

2.2.1 Learning from labelled and unlabelled data

Conventional learning system are usually based on provided training samples and training labels, both given as input by an external entity and, for this reason, are called supervised methods. If enough and proper training samples exist, these approaches can obtain very high recognition and classification performances. However when this assumption is not satisfied, the problem reduces to an ill-posed problem, where the input data do not carry enough information to learn the classification model. The lack of sufficient labelled training data and the problems involved in the labelling process have recently focused the attention on *semi-supervised methods*, hence on avoiding the problem by exploiting both labelled and unlabelled data in the learning process. Semi-supervised learning uses the discriminative power of the labelled data and also benefits from the potential of a massive amount of unlabelled data. It is also supported by the hypothesis that its learning process is similar to the human reasoning. There exists, indeed, a large agreement among experts that the power of human perception is also a result of a long-winded learning process where we continuously observe a considerable number of data, yet, most of it unlabelled. Zhu et al. [2007] empirically demonstrate how the model we learn changes when we are showed only labelled data or both labelled and unlabelled data. In particular they showed that the classification decision changes accordingly with unlabelled data, which is a typical assumption of SSL. A similar conclusion was also drawn by Zaki and Nosofsky [2007]. For more complicated tasks, however, it was also shown that unlabelled data does not always help humans in order to improve their performance on a given task [Vandist et al. [2009]]. Which is, once again, what happens in machine learning with Semi Supervised Learning where unlabelled data can also lead to worse results. Many works address the problem of analysing when unlabelled data help the classification task of not [Balcan and Blum [2005], Lafferty and Wasserman [2008], Singh et al. [2009]].

In literature exists a wide variety of semi supervised methods that allows to work with partially labelled data. For example Seeger [2001] and Lawrence

and Jordan [2004] used Gaussian processes introducing a null category to model unlabelled data, Kapoor [2006] used a mixture of Gaussian processes and extended the standard Gaussian process prior by introducing regularization based on the unlabelled and the labelled data points. Nigam et al. [2000] used a generative model and EM to classify text documents. Szummer [2002] introduced two different approaches: a kernel expansion method to account for unlabelled data in the input feature vectors and used a classification method trained with EM [Szummer and Jaakkola [2000]], and also a Markov Random Walk method that exploits clusters and low dimensional structure in the data in a probabilistic manner [Szummer and Jaakkola [2001]].

SSL has been used also to solve several computer vision problems: Leistner et al. [2008] presented an algorithm that combines semi-supervised boosting and visual similarity learning and demonstrate its effectiveness in object detection, Leistner et al. [2009] extended the Random Forest approach to unlabelled data and used it for object categorization, Liu and Chang [2009] proposed an interactive image segmentation based on SSL. Transductive Learning has been applied to actions recogniton [FarajiDavar et al. [2011]], faces detection [Li and Wechsler [2005]], image retrieval and user relevance feedback [Huang et al. [2010], Borghesani et al. [2011]], and also to tracking [Zha et al. [2010], Wu and Huang [2000]].

2.2.2 Learning with few examples

Another possibility to introduce further knowledge into a system is represented by *transfer learning*. Transfer learning refers to the ability of a learning system to transfer knowledge learned in one or more source (or support) tasks and use it to improve learning in a related target task. Transfer learning helps in that situations where a given task is only equipped with few training samples, but there exist a certain number of related task with more training samples.

The advantage, compared to traditional machine learning methods is that the classification rules are not built from scratch. Similarly to semi supervised learning, also the concept of transfer learning is supported by human reason-

ing and learning abilities. We tend to organize all our knowledge into useful taxonomies and to group concepts on the basis of common properties. This intrinsically means that new concepts are not learnt in isolation, but considering the connections with the prior knowledge [Hofstadter and Group. [1998], Brown and Kane [1988]]. For example, it might be easier to learn Italian if a person knows Spanish or French, or it might be easier to learn snowboarding if one already knows skiing. Other examples of this cognitive ability can also be found in visual task, it is spontaneous to recognize a new category if we already known related categories: a tiger might be recognized as similar to a leopard but with stripes and a guava is a fruit that can be easily recognized if apples and limes are known.

The research interest on transfer learning has increased starting from the late '90s in the machine learning community. The topic has been addressed with different terms: learning to learn, life-long learning, knowledge transfer and its theoretical foundations has been firstly discussed in the NIPS-95 workshop on "Learning to Learn". A review of the state-of-the-art and of the recently proposed methods can be found in the comprehensive survey of Pan and Yang [2010].

Developing transfer learning techniques able to get advantages from support task, requires proper answer to answer three questions: what to transfer, how to transfer and when to transfer. To get all these advantages, is however necessary to properly answer to three questions: what to transfer, when to transfer and how to transfer. *What to transfer* addresses the type of knowledge that can be transferred from support tasks to a new target task, the three main forms of transfer are instances, feature representation and model parameters. In instance transfer approaches a certain part of the source data are sampled and considered together with the few available labelled data in the target problem, this strategy is used by Dai et al. [2007] and Wu and Dietterich [2004]. In the second case, feature transfer, the representation of the target domain is learnt encoding in it some useful knowledge extracted from the source. Argyriou et al. [2008] showed how it is possible to exploit few labelled target samples together with source data and apply a feature learning mechanism to build a

shared across multiple tasks. Finally, parameter or model transfer approaches assume that individual models for related tasks should share some parameters or prior distributions of hyper-parameters.

By evaluating *when to transfer* the focus is on the relatedness of the source and target task. Transfer learning, indeed, does not always improves the recognition performance of a new target task, or, in other words, there is no guarantee on the generalization ability of the learner from the training set to the unseen examples of the target task. Has been empirically showed that if two task are dissimilar, knowledge transfer fails and leads to worse performance compared to independent learning producing the so called negative transfer.

After solving the problems of what and when to transfer, the question is *how to transfer*. Learning algorithms need to be developed to properly pass information, a big variety of methods exist in this sense: boosting approaches [Dai et al. [2007], Yao and Doretto [2010]], KNN [Zhang and Yeung [2010]], Markov logic [Davis and Domingos [2009]], graphical models [Dai et al. [2009]].

In computer vision and pattern recognition transfer learning has been exploited especially for the tasks of object detection and classification. Yao and Doretto [2010] introduced two different algorithms based on boosting techniques to transfer from multiple sources and applied them to object category recognition and specific object detection. F.F. et al. [2006] represented object categories with probabilistic models expressing posteriors models for categories with few training samples with knowledge transfer from other categories. Tommasi et al. [2010] tackled the problem of object category recognition with small sets of training samples introducing a Least Square SVM (LSSVM) based model adaptation algorithm able to select and weight appropriately prior knowledge coming from different categories. Kuzborskij et al. [2013] extended this approach to a multiclass formulation of LSSVM. Lim et al. [2011] proposed an object detector formulated by borrowing and transforming examples from other classes learning a mixed norm regularized SVM model.

When dealing with the lack of training data of the target task (neither la-

belled nor unlabelled) transfer learning is known as *zero-shot* learning. In this scenarios, other data sources have to be used to perform knowledge transfer, a new trend in the recent years is the concept of learning with attributes. Where the term attribute refers to category-specific features. Lampert et al. [2009] use a large database of human-labelled abstract attributes of animal classes (*e.g* brown, stripes, water, eats fish). Rohrbach et al. [2011] evaluate different approaches of zero-shot learning in large scale settings on the ImageNet (ILSVRC10) dataset [Berg et al. [2010]].

2.2.3 Learning anomalies and outliers

Another need of modern learning systems is to detect samples that differ from the model learnt during training, or in other words to detect outliers and discern new classes. Traditional models of learning, such as the standard PAC (Probably Approximately Correct [Valiant [1984]]) model, assume that training instances are drawn according to the same probability distribution as the unseen test examples. However in many real world applications, this assumption does not hold, and the hypothesis of a closed world can not be considered. It often happens that the training data is different from that available for testing, and that the two sets are actually drawn from different distributions. Identifying new categories or samples is referred as *novelty detection* or *anomaly detection*, and, given the fact that we can never train a machine learning system on all possible object classes, it is important an important and challenging task. Considering, for instance, zero-shot learning and one-shot learning is important for an autonomous system to firstly identifying new classes.

In literature there exists a large number of algorithms of novelty detection: Markou and Singh [2003a,b] give a detailed analysis of such methods categorizing them in statistical and neural network based approaches. *Statistical approaches* are mostly based on modelling data according with their statistical properties and using this information to estimate whether a test sample comes from the same distribution or not. Different techniques vary in terms of complexity. Probability distributions can be estimated both using parametric

or non-parametric methods. The former category assumes that data are drawn from a family of known distributions, such as the normal distribution and certain parameters are calculated to fit this distribution. Methods based on Gaussian Mixture Models (GMM), Hidden Markov Models (HMM) and Hypothesis testing belongs to parametric methods. A review of these methods is beyond the aim of this Section, we refer the reader to the survey of Markou and Singh [2003a,b]. An important study related to parametric models concerns the trade-off between the recognition rate and the proportion of data rejected, because of noise in the training data, errors are unavoidable and the rejection option is introduced to avoid misclassification [Hansen et al. [1995], Fumera et al. [2000]] In non-parametric methods the overall form of the density function is derived from the data as well as the parameters of the model. As a result non-parametric methods give greater flexibility in general systems. Nearest neighbour based density estimation, Parzen density estimation and string matching approaches belong to this second category. *Neural networks* have been widely used for novelty detection. The basic idea is to train the neural network on the normal training data and then detect novelties by analysing the response of the trained neural network to a test input. If the network accepts a test input, it is normal and if the network rejects a test input, it is novelty. Compared to statistical methods, some issues for novelty detection are more critical to neural networks such as computational expense, but the have the advantage that a smaller number of parameters is required. These approaches include multilayer perceptrons (MLP), self organising maps (SOM), radial basis function networks, etc.

Support Vector Machines are also used to solve the problem of novelty and anomaly detection. In particular One Class SVM have been used to separate class of objects represented by the training set and all other possibile objects in the feature space [Tax and Duin [2004, 1998], Schölkopf et al. [2000]]. Muandet and Schölkopf [2013] has recently introduced a formulation of One Class Support Measure Machines (OCSMMs) for group anomaly detection. Unlike traditional anomaly detection, group anomaly aims at recognizing anomalous aggregate behaviours of data points, and the OCSMMs is a generalization of

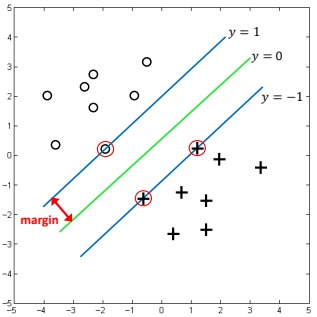
the OCSVMs to a space of probability measure.

Despite its importance also in visual systems, novelty detection has often been neglected. Weinshall et al. [2012] proposed to detect anomaly as incongruence of classifier at different level of a hierarchy. They demonstrate how the approach is general and applies to several problems, including image classification, audio classification, motion patterns detection. Bodesheim et al. [2013] instead introduced a kernel null space method for detection of novel categories of images. Specifically the approach makes use of a projection in a joint subspace where training samples of all known classes have zero intra-class variance.

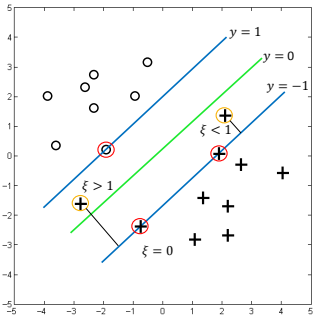
After the discussion on the main challenges encountered by learning systems when dealing with incompletely annotated data, in the rest of this Chapter we introduce the main machine learning methods we exploited in the rest of the thesis to solve the three different pattern recognition problems, namely: People Tracking, Novel image categories detection and Illustration Segmentation in document analysis. Since we used both supervised and semi supervised discriminative methods, we give the basic notation of both and we briefly introduce Support Vector Machines and Graph Based SSL.

2.3 Machine Learning: Supervised methods

The goal of *Supervised Classification* is to optimize a model, given a certain learning task and a limited amount of training data, with the best possible generalization performance. The training set is made of pairs (x_i, y_i), where $x_i \in X$ is a vector representation of the input data and $y_i \in Y$ are the labels corresponding to examples in X. When considering binary classification the labels $y_i \in \{1, -1\}$. The performance of the classifier is then tested with samples not used during training.



(a) SVM with linearly separable data



(b) SVM with non linearly separable data

Figure 2.3: Illustration of linearly (a) and non linearly (b) separable SVMs. The margin is defined as the perpendicular distance between the decision boundary and the closest of the data points. Red circled points represent the *support vectors*, the points with the orange circle around them (b) are on the wrong side of the marging and have corresponding *slack variables* $\xi_i \geq 0$.

2.3.1 Support Vector Machines

Support Vector Machines are probably one of the most known and widely used classification methods in pattern recognition, introduced in 1992 by Boser et al. [1992] and further formalised by Cortes and Vapnik [1995]. The aim of SVMs is to learn the hyperplane that better separates positive and negative training points with the largest *margin*. The concept of margin is defined as the smallest distance between the hyperplane and the nearest samples (*support vectors*), as illustrated in Fig. 2.3.

Maximal margin classifiers constitute the first and simplest SVMs model and only works for linearly separable data. It can be solved minimising the quadratic function under inequality constraints:

$$\begin{aligned} &\arg\min_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{s.t. } y_i(\mathbf{w}\cdot x_i - b) \geq 1 \end{aligned} \tag{2.1}$$

The term $y_i(\mathbf{w} \cdot x_i - b)$ of the constraint is the canonical representation of the decision hyperplane, $\mathbf{w}$ is the normal vector and b the bias term. From a geometrical point of view minimize 2.1 is equal to maximize the distance $\frac{2}{\|\mathbf{w}\|}$ between the hyperplanes $y = 1$ and $y = -1$.

Introducing the Lagrange multipliers α_i the problem is equivalently written as:

$$\arg\min_{(\mathbf{w},b)} \max_{\alpha} \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(\mathbf{w} \cdot x_i - b) - 1\right] \tag{2.2}$$

The dual formulation, finally, can be inferred by setting the derivatives of Eq. (2.2) with respect to b and $\mathbf{w}$ to zero obtaining:

$$w = \sum_{n=1}^{N} \alpha_i y_i x_i \tag{2.3}$$

and

$$\sum_{n=1}^{N} \alpha_i y_i = 0 \tag{2.4}$$

and then substituting Eq. (2.3) in Eq. (2.2):

$$\begin{aligned} \arg\min \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{s.t. } \alpha_i \geq 0 \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned} \tag{2.5}$$

So far we have assumed that the two data distributions are not overlapped, but in practice this assumption does not always hold and the input data points might not be linearly separable or an exact separation might lead to poor generalization. In order to handle this situation a possible solution is to consider a *soft margin* approach where the problem in Eq. (2.1) is modified adding a penalty for each data point wrongly classified. Introducing, for each training point on the wrong side of the margin, a *slack variables* $\xi_i > 0$ the problem

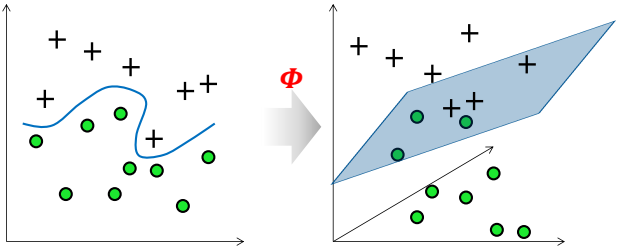


Figure 2.4: Kernel trick. Mapping Φ of non linearly separable data into a higher dimensional space.

becomes:

$$\arg\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \tag{2.6}$$

The left part of Eq. (2.6) is the regularization term, while the right part is the loss or error term and the soft parameter C controls the trade-off between the two terms.

Another possible approach to find a solution with non linearly separable data is to map the training vectors x_i into an higher dimensional feature space by a function Φ (See Fig. 2.4). Substitution the dot product in Eq. (2.6) with a kernel function $k(x_i, x_j)$ the equation can be rewritten as:

$$\begin{aligned} \min \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(x_i,x_j) \\ \text{s.t. } \alpha_i \geq 0 \\ \sum_{i=1}^{N}\alpha_i y_i = 0 \end{aligned} \tag{2.7}$$

The most commonly used kernel $k(x_i, x_j)$ are:

$$\begin{array}{ll} \textit{linear} & K(x_i, x_j) = x_i^T x_j \\ \textit{polynomial} & K(x_i.x_j) = \left(\gamma x_i^T x_j + r\right)^d, \gamma > 0 \\ \textit{RBF} & K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0 \\ \textit{sigmoid} & K(x_i, x_j) = \tanh\left(\gamma x_i^T x_j + r\right) \end{array} \tag{2.8}$$

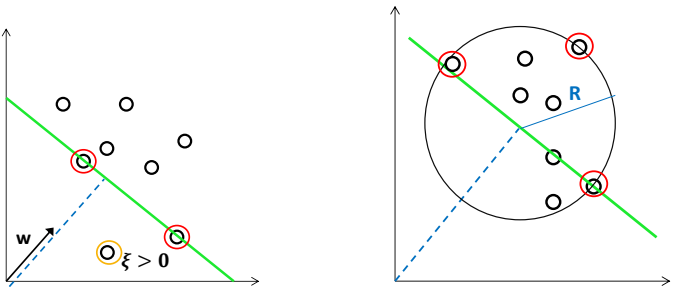
given γ, r and d are kernel parameters.

While being well suited to solve the problem of separability of non-linear separable training points, kernelized SVMs suffer from the drawback of being computationally expensive in practice. Typically the complexity of training a non linear SVM is between $\mathbf{O}(n^2)$ and $\mathbf{O}(n^3)$, where n is the dimensionality of the training vectors, depending on the chosen kernel. For this reason, using linear implementation is preferable when working with really high dimensional feature vectors and *additive* kernels [Vedaldi and Zisserman [2012]] have been recently introduced. Additive kernel are explained in Sec. 5.5 where we proposed to use Support Vector Machines in a hierarchical framework to solve the problem of novelty detection in image categories. In this thesis SVMs are also exploited in Chapter 6 to distinguish between textual and pictorial region in the context of document image classification.

2.3.2 OC-SVMs

Support Vector Machines, as presented in the previous section are a powerful tool to learn a binary classification, unfortunately working with incomplete data, and in particular with missing labels, training data for one of the two classes might not be available. In these situations classification corresponds to *One-Class Classification* (OCC), also referred to as outlier detection, novelty detection, anomaly detection.

One-Class Support Vector Machines (OC-SVMs), originally introduced by Schölkopf et al. [2000] and Schölkopf et al. [2001], aim to capture the region of the feature space where the distribution of the training points $(x_i, +1)$ with $i = \{1, \ldots, N\}$ lie. The equivalent geometric interpretation aims to find the



(a) OC-SVMs separating hyperplane (b) OC-SVMs separating hypersphere

Figure 2.5: Illustration of the two different interpretation of OC-SVMs by [Schölkopf et al., 2000] and Tax and Duin [2004] using respectively an hyperplane and a hypersphere.

hyperplane that maximizes the distance of the data points from the origin of the feature space (Fig. 2.5(a)).

The quadratic programming minimization function is slightly different from the original, given Φ a feature map of the input vectors, the problem is:

$$\begin{aligned} \arg\min_{w,\xi_i,\rho} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{N}\xi_i - \rho \\ \text{s.t. } (w\Phi(x_i)) \geq \rho - \xi_i \\ xi_i \geq 0 \end{aligned} \tag{2.9}$$

Where the regularization term $\|w\|$ and the error term given by the slack variables ξ_i have a meaning similar to the standard SVM formulation and the parameter ν controls the trade-off between these two terms. More formally ν represents here an upper bound on the fraction of outliers (*i.e* training points that lie outside the boundary) and a lower bound on the fraction of Support Vectors. Equivalently to the binary SVMs classification, introducing the Lagrange

multipliers α_i the dual minimization problem is:

$$\begin{aligned} \arg\min_{\alpha} \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_i, x_j) \\ \text{s.t. } 0 \geq \alpha_i \geq \frac{1}{\nu N} \\ \sum_{i=1}^{N} \alpha_i = 1 \end{aligned} \tag{2.10}$$

An alternative version of OC-SVMs has been introduced by Tax and Duin [2004] exploiting a spherical instead of planar approach where the aim is to minimize the volume of the hypersphere that better enclose the training points (Fig 2.5(b)).

In this thesis we use the implementation of Schölkopf et al. [2000] to address the problem of novelty detection in image classification considered in Chapter 5.

2.3.3 Mixed Norm SVMs

The linear formulation of Support Vector Machines in Eq. (2.2) is also referred to as *L2-regularized SVMs*, because the regularization term appears with a L2 norm. Variations of this formulation includes *L1-regularized SVMs* and *Mixed-norm regularized SVMs*. In the context of detection of novel categories of images analysed in Chapter 5, we exploited the latter formulation to enforce the discriminative power of classifiers at different level of a hierarchy.

Considering the linear SVMs of Eq. (2.2), we already mentioned how the weights vector w represents the normal of the separating hyperplane, an alternative interpretation of the vector w is related to the feature importance [Zhang et al. [2013]]. SVM weight, in fact, can be used to implement an implicit feature selection [Chang and Lin [2008]] with the advantage that classification and feature selection are both performed with the minimization of the same quadratic problem.

Mixed Norm SVMs promote the sparsity of the weights vector enforcing

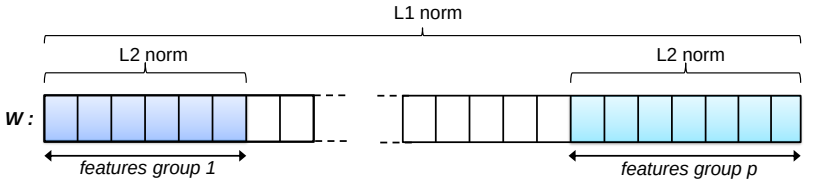


Figure 2.6: Mixed Norm SVMs weights and features grouping. Each group is L2 regularized, the whole feature vector is L1 normalized.

the feature selection, specifically they take into account the fact that features may be structured in some ways by using a mixed-norm that groups the features in different sets and regularizes them together.

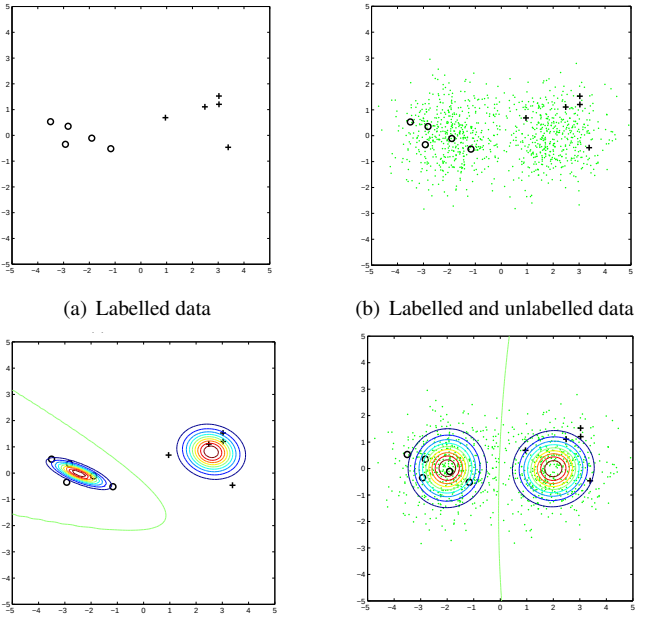
In the L1-L2 regularized SVMs, the regularization term is replaced by Ω_{1-2}:

$$\Omega_{1-2}(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|^2 \tag{2.11}$$

leading to the minimization problem:

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} \max(0, 1 - y_i(x_i^T w + b))^2 + \Omega_{1-2}(\mathbf{w}) \tag{2.12}$$

Intuitively the L1-L2 norm can be interpreted as an L1 norm applied to the vector containing the L2 norm of each group of features (see Fig. 2.6). In this way sparsity is promoted on each $\mathbf{w}_g$ norm and consequently on the components $w_{i,g}$ as well. Despite the advantages, it can be argued that L1-L2 norm SVMs strongly tie together the components of a group, for this reason a more flexible grouping can be found in L1-Lq norm, or in adaptive L1-Lq [Flamary et al. [2012]].



(a) Labelled data

(b) Labelled and unlabelled data

(c) Model learnt from labelled data

(d) Model learnt from labelled and unlabelled data

Figure 2.7: Semi Supervised Learning: Unlabelled data can help to learn the structure of the data distribution.

2.4 Machine Learning: Semi Supervised methods

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning. In SSL the dataset $X = (x_i)_{i\in 1,\dots,n}$ can be divided into two parts: the points $X_l = (x_1,\dots,x_l)$, for which labels $Y_l = (y_1,\dots,y_l)$ are provided, and the points $X_u = (x_{l+1},\dots,x_{l+u})$, the labels of which are not known. SSL is initially motivated by its practical value in learning faster and cheaper. In many real world applications indeed, it is relatively easy to acquire a large amount of unlabelled data, but their corresponding labels for the prediction task require slow human annotation. SSL methods assume that despite unlabelled data do not give any information about the learning task, which is determined by la-

belled data, they yield information about the data distribution in the input domain. Unlabelled data are thus used to either modify or re-prioritize hypotheses obtained from labelled data alone. In other words SSL learning techniques can help in that situations where there are too few labelled examples to reach a desired level of learning accuracy, but, by exploiting the structure of the domain learned from unlabelled examples the learning accuracy is improved. An example of this principle is shown in Fig. 2.7. The decision boundary learnt from labelled data only is different from the one learnt with semi supervised settings and would lead to a considerable number of classification errors.

Because of its potential of learning with less human effort and with higher accuracy SSL is of great interest both in theory and in practice. Semi Supervised techniques have been largely studied and adopted, the most commonly used classification methods are: *Generative Models*, *Transductive SVMs*, *Co-Training*, *Self-Training* and *Graph-based models*. Key assumptions of these methods is provided in Table 2.1, however a detailed description of SSL is beyond the goal of this Chapter and we refer the reader to the works of Chapelle et al. [2010], Zhu [2008] and Sammut and Webb [2010].

Table 2.1: Some representative Semi Supervised Learning methods

Method	Assumptions
mixture model, EM	generative mixture model
transductive SVM	low density region between classes
co-training	independent and redundant features splits
graph methods	labels smooth on graph

2.4.1 Transduction *vs* Induction

Considering the classification from an "output/result" perspective, SSL methods can be grouped in *inductive* and *transductive* semi-supervised methods (See Fig. 2.8). In the former category, the learner uses both labelled training data X_l and unlabelled data X_u to synthesize a prediction function $f : X \mapsto Y, f \in F$ where F is the hypothesis space. The goal is to find a predictor ca-

pable of classifying future test data more accurately than a predictor learned from only labelled data. Conversely, transductive learning is solely interested in the predictions of the unlabelled training data without any intention to derive a generalized function for future test data. According to Vapnik [2006] *"when trying to solve some problem, one should not solve a more difficult problem as an intermediate step"*, this equivalently means that a transductive learner only attempts to estimate the values of an unknown label function assignment at particular points of interest, while inductive inference attempts to estimate the unknown function over its entire domain of definition solving a more complicated and sometimes not necessary problem.

2.4.2 Graph-based methods

Recently, the most active area of research in semi-supervised learning is represented by Graph-Based methods [Culp and Michailidis [2008], Blum and Chawla [2001], Zhu et al. [2003], Belkin [2003]]. Graph methods are non-parametric, discriminative, and transductive in nature. They aim at naturally represent the geometry of labelled and unlabelled data using a representation based on a graph $G = (V,E)$ where the nodes $V = \{1,\ldots,n\}$ represent the training data and the edges E the similarity between them. The similarities are given by a weight matrix $W = \{w_{ij}\}$ where w_{ij} is non-zero only if x_i and x_j are neighbours:

$$w_{ij} = \begin{cases} 0 \text{ if } e = (i,j) \notin E \\ w(e) \text{ if } e = (i,j) \in E \end{cases} \tag{2.13}$$

The weight $w(e)$ of an edge e indicates the similarity of the incident nodes (and a missing edge corresponds to zero similarity). A frequently used way to represent the edges similarity are the Gaussian weight function $w_{ij} = \exp - \frac{\|x_i - x_j\|^2}{\sigma^2}$ or the *kNN* edge weight function with $w_{ij} = 1$ only if x_i is within the k nearest neighbours of x_j (or viceversa).

Underling that nodes of the graph represent both labelled and unlabelled data, the aim of graph based method is to estimate a label for unlabelled data

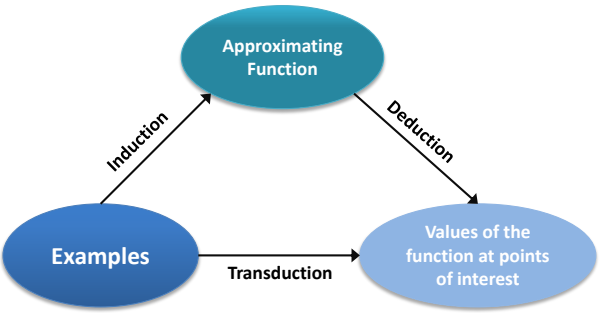
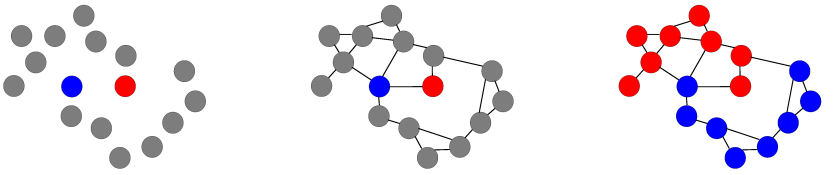


Figure 2.8: Induction *vs* Transduction Vapnik [1998].

points exploiting the knowledge contained in the labelled nodes. A first set of algorithm is based directly on the idea of *label propagation* from labelled to unlabelled samples. In their basic formulation these algorithms provides an iterative framework where each node start to propagate its label to its neighbours and the process is repeated until convergence [Zhu et al. [2003], Zhou et al. [2004], Szummer and Jaakkola [2002]]. An alternative set of approaches based on *graph regularization* that utilizes the graph Laplacian is proposed by Belkin [2003], Joachims [2003], Belkin and Niyogi [2004], Zhou et al. [2004]. The Graph Laplacian can be defined in different ways, with normalized and unnormalized formulations:

$$\begin{aligned} L &= I - D^{-\frac{1}{1}} W D^{-\frac{1}{1}} \\ L &= I - D^{-}1W \\ L &= D - W \end{aligned} \tag{2.14}$$

Regularize the graph structure consists in finding a labelling of the graph consistent with both the initial labelling and the geometry of the data induced by the graph structure, this can be done minimizing a cost function $C(y)$ on the graph. Given a labelling $\hat{Y} = (\hat{Y}_l, \hat{Y}_u)$ and considering L the Graph Laplacian, the cost function includes both a consistency term (to measure the consistency



(a) (b) (c)

Figure 2.9: Graph based SSL. Given labelled and unlabelled data (a) an undirected graph is created (b) and the final solution is found by label propagation on the graph (c).

with the initial labelling: $\|\hat{Y}_l - Y_l\|^2$) and a regularization that penalizes rapid changes in $\hat{Y}$ ($\hat{Y^T L Y}$).

$$C(\hat{Y}) = \sum_{i=1}^{l}(\hat{y}_i - \hat{y}_j)^2 + \sum_{i,j=1}^{n} w_{ij}(\hat{y}_i - \hat{y}_j)^2 \qquad (2.15)$$
$$C(\hat{Y}) = \|\hat{Y}_l - Y_l\|^2 + \mu \hat{Y}^T L \hat{Y}$$

Many different variants of Eq. (2.15) have been proposed with different regularization term to take also into account degenerate situations.

Despite the categorization in label propagation and graph regularization method Chapelle et al. [2010] demonstrate how the different algorithms can be cast into a common framework with a minimization of a quadratic function.

Part I

Semi supervised tracking

CHAPTER

THREE

SINGLE TARGET PEOPLE TRACKING

Object tracking is a challenging task and is addressed by several computer vision application: it is not only important by its own, but often represent the preliminary steps of any activity such as behaviour analysis or action recognition. In particular, the recent increase of surveillance videos from both indoor and outdoor cameras has focused the attention even more on the problem of object tracking. Among the others people and their mutual interactions are the elements that are considered the most. The complexity of tracking people however, is well known due to the articulated human shape and the variable motion patterns. Several advancements have been made by recent scientific works posing some constraints but the problem remains open: for example, a system with good performance on viewpoint variations might fail in clutter environments, a system based on motion estimation might have problems with a bouncing object, and so on (Fig. 3.1 shows some examples of challenging tracking situations). With the definition *tracking in unconstrained scenarios* we consider the situation where no constraints are given on the type of camera, the occlusions, the lighting condition, the clutter circumstances or the motion pattern. We would like to address the problem in a general environment even when time/spatial coherency is not completely guaranteed (*e.g* the target is not

Figure 3.1: Examples of challenging tracking situations: (a) zooming camera, (b) occlusion, (c) changing light, (d) cluttered situation, (d) moving camera. Images are taken from ALOV300++ dataset Smeulder et al. [2013]

.

visible in every frame), when the motion of the target is unpredictable or when its appearance changes rapidly (*e.g* the target take off a coat).

We consider the problem of single target people tracking and we assume to have the availability of the detection of the target at least in some frames and we re-interpret the tracking as a partially labelled data problem. The labelled data points are the initial detection and the unlabelled data points are the people patches extracted from the subsequent video frames exploiting a conventional people detector. The initial detection can be either given manually (as often happens in forensics applications) or extracted automatically (for example the detections of a person that enters a controlled area through a gate or a door). We explore the problem as a *tracking-by-learning* problem neglecting the motion prediction and motion estimation, and focusing instead on the data association problem among the video frames. We aim in this way at solving at least the problem of the unpredictable motion patterns that often arise in complex situation. Using a Semi Supervised Learning method and identifying frame-by-frame the target object we learn iteratively its appearance dealing also with the problem of unpredictable appearance. The appearance

model learnt is strengthen by updating the labelled instances with an evolutionary strategy. In this scenario the problem could not be considered just as a tracking problem, but more generally as a problem of pattern (of people) recognition by means of visual appearance. Thus during the work we will refer to the problem also with the term *people following*.

In the next sections of this chapter we firstly give an overview of the tracking problem in literature, and we successively describe the method we propose, detailing each step. We also test our proposal on several videos taken from publicly available datasets, comparing it with similar learning-based approaches and we demonstrate the effectiveness of the method in various situations.

3.1 Background and related work

Tracking overview Several tracking methods, since from the preliminary applications, rely on the background subtraction in order to identify moving objects [Calderara et al. [2008], Kuo et al. [2010], Zhao and Nevatia [2004], Hu et al. [2006], Kim [2008]]. Given the recent advances in people detection methods, now able to locate pedestrian even in complex scenes, tracking methods often rely on detection instead of background subtraction [Andriluka et al. [2008], Zhang et al. [2008], Leibe et al. [2008]]. Andriluka et al. [2008] proposed a combination of tracking and detection in a single framework, body parts are described by local features and a hierarchical Gaussian process latent variable model (hGPLVM) is exploited to model prior knowledge on articulations and temporal coherency in the walking cycle. The popularity of Particle Filtering [Doucet et al. [2001]](also known as Monte Carlo Estimation) that stems from its simplicity, generality and success over a wide range of challenging applications, has led to a large number of applications that exploit the combination of object detection and Particle Filtering [Khan et al. [2005]]. Vermaak et al. [2003] introduced a mixture of particle distribution to handle the multi-modality of target distributions, Okuma et al. [2004] combined the approach of Vermaak et al. [2003] with a boosted object detector for multitarget tracking in hockey games. Cai et al. [2006] later extended and improved the

same boosted Particle Filter introducing a mean-shift algorithm to stabilize the trajectories of the targets for tracking during mutual occlusion. To overcome detection inaccuracies Breitenstein et al. [2011], proposed to bias the detection using Particle Filtering directly on the people detector dense confidence response rather than on the discrete response improving the tracking performance especially in cluttered scenes.

The shortcoming of Particle Filter based methods is their effectiveness only on short sequences, at this aim data association tracking has been introduced to link together different tracklets or to recover from long term occlusions. In these approaches objects are tracked until they are visible and at a different level of processing tracklets are associated, the objective is thus to globally optimize a set of detected trajectories. This approach is typically interpreted as an a-posteriori optimization method where, given the data extracted from a video sequence, a specific optimization method concurs linking data along the time axis [Metternich et al. [2010]]. For instance Perera et al. [2006] improved the approach of Kaucic et al. [2005] introducing a technique able to deal with merge and splits of object trajectories; Li et al. [2009] instead used a hybrid boost approach to associate different tracklets. Classical approaches include the Joint Probabilistic Data Association Filter (JPDAF) [Svensson et al. [2010]] and Multi Hypotheses Tracking (MHT) [Boyd and Meloche [2003]]. MHT considers the possibility of multiple associations over several time steps but its complexity is such that it usually limits the analysis to only few steps. JPDAF otherwise try to make the best possible assignment in each time step by jointly considering all possible associations between targets and detections to the cost of an exponentially increasing complexity.

Occlusion handling and long term term tracking has also been addressed exploiting *3D* information when calibration is available. Leibe et al. [2008] fused information from stating and moving cameras with a Structure from Motion (SfM) approach and built 3D information merging 2D object detection and scene geometry. Gavrila and Munder [2007] and Ess et al. [2009] proposed approaches for tracking from moving cameras using stereo information. Differently, specific solutions like body part detectors Wu and Nevatia [2007]

or motion models based on the social behaviour analysis Pellegrini et al. [2009] can be adopted to deal with complex situation and occlusions.

In all the revised method, an object or motion model is either learnt or defined to solve the tracking problem. If this is helpful to improve the tracker accuracy might sometimes limit its capability in terms of generalization. Specifically, learning appearance models of the targets results in an excellent tracking method when the temporal coherence among the target visual features is maintained but often fails when this constraint is not satisfied and the target appearance changes. Model-free tracking has thus been introduced and aims at learning specific target classifiers without relying on target models [Stalder et al. [2009]] and the tracking problem is interpreted as a classification problem where the knowledge about the target itself can be explicitly specified or learnt during the tracking process. Hence, the choice of the classifier is crucial to achieve good performance. Commonly used classifiers include adaptive classifiers [Stalder et al. [2009]] and on-line boosting [Kuo et al. [2010]]. The drawback is that model free trackers are subjected to drifting problems, Matthews et al. [2004] referred to this problem as *template update problem*. In order to mitigate this effect additional knowledge may be exploited, *e.g* geometric verification [Ess et al. [2009]], combination of generative and discriminative models, co-learning using different types of features [Tang et al. [2007]], or constrained updates [H. et al. [2010]].

Semi Supervised Learning and Transductive Learning A completely different perspective is to interpret the tracking as a Semi Supervised Learning (SSL) problem, where the actual knowledge about target object represents the only labelled information.

A comprehensive survey on SSL can be found in Zhu [2008] and, among the others, graph based algorithms have a relevant role. In such methods labelled and unlabelled input data are modelled as nodes of undirected graphs whose edges represent the similarity among data points. Labels for unlabelled instances are estimated by propagating the information available at labelled nodes to unlabelled nodes. As explained in Chapter 2 the main advantage of

graph methods is to be non-parametric and transductive in nature (the solution is a cut of the graph, thus a label for unlabelled input point and not a general classification function as in the case of inductive methods Zhu [2008]).

Transductive learning has been introduced by Vapnik in 90's Vapnik [1998] and has evolved over previous decades as an effective technique for solving several Computer Vision problems. Computer visions application of transduction spans a wide area of research. Transductive Learning (TL) has been applied as a method for interactive image segmentation [Liu and Chang [2009]], in conjunction with transfer learning for actions [FarajiDavar et al. [2011]] and faces recognition [Li and Wechsler [2005]], for image retrieval and user relevance feedback [Huang et al. [2010], Borghesani et al. [2011]]. The transductive learning paradigm has already been exploited also to solve tracking and re-identification problems. The seminal work of Wu and Huang [2000] introduced the TL as a solution to the severe variation of the models in color tracking. They fitted the TL problem into an EM frameworks to estimate the pixel labels in hand and face color tracking. Zha et al. [2010] proposed an on-line single target tracking using graph transduction applied to faces and cars. Coppi et al. [2011a] proposed an on-line single target tracking and re-identification method based on a graph based formulation of the TL problem. They enforced the knowledge encoded in the labelled instances introducing an update strategy to avoid drift in the tracking and to allow re-identification in case of occlusions. Independently from the chosen applications, many solutions for building transductive classifiers do exist Zhu [2008]. Among these, Spectral Graph transducer, Joachims [2003], exhibits consistent performances and a strong theoretical relation with incremental manifold learning. This solution aims to exploit the spectral properties of a similarity graph, built over a complete set of data, to learn the manifold structure from which the data have been sampled Lin and Zha [2008]. This intuition holds since graph Laplacian eigenvalues converge to the discretization of the Laplace Beltrami operator applied over the manifold and then are able to describe the functionals that regulate its structure.

3.2 Problem Statement

As introduced at the beginning of the chapter we consider the problem of single target people tracking in surveillance scenarios. We propose to tackle the problem with a Semi Supervised approach, that exploits some labelled instances of the target and aims at learning its visual appearance disregarding any temporal or spatial information. Semi-supervised learning is used to solve partially labelled problem since it aims to learn the classification from both labelled and unlabelled data assuming that also unlabelled points can help the classification task.

Specifically among semi supervised methods we adopt a *graph based transductive learning*: nodes of the graph represent the people patches and edges represent the similarity between them, only few nodes of the graph are labelled. The tracking is consequently formulated as the problem of label propagation from labelled to unlabelled nodes of the graph. We propose an iterative approach that, frame-by-frame, detects the person on the scene using a HOG based people detector [Dalal and Triggs [2005]] and estimates a label for unlabelled elements. To improve the discriminative power of the transductive learner, we exploit two different models of labelled instances, one representing the target and the other representing non-target elements. Finally, since the target's appearance can evolve in time, we designed an update strategy based on evolutionary spectral clustering to retain the best target images and capture its appearance variability. This mechanism explicitly avoids drifting error and is a novel aspect of our proposal.

The approach we propose is independent from the visual feature adopted for people representation. Among the possible plethora of descriptors we choose a covariance matrix descriptor because of its capacity of integrating information concerning colours, textures and shape (edges) in a single compact and highly discriminative formulation.

A schematic overview of the proposed method is depicted in Fig. 3.2. People are detected in the considered frames and represented with covariance matrices, the first association of target and non target elements is manually given

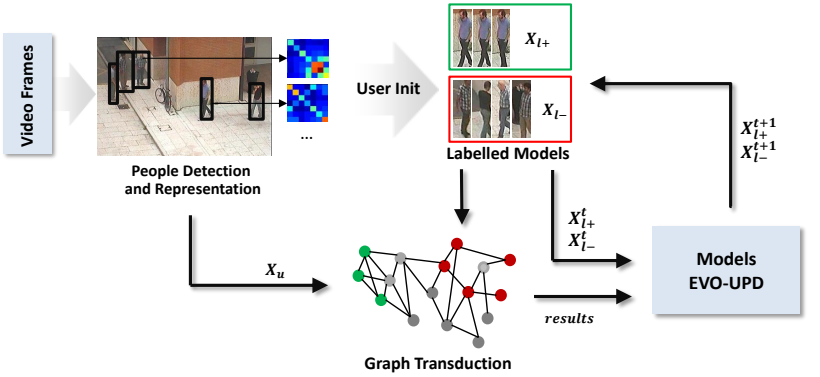


Figure 3.2: System general configuration. The scheme shows the main steps of our proposal, which consist in People detection, Transductive learning search and Model update.

by the user and the labelled models are initialized. Once examples of the target have been provided, the graph is built over labelled and unlabelled elements and the transductive learning searches for the target occurrences. The evolutionary update strategy is used to retain the patches that best represent the target (and non target) appearance.

3.3 Transductive Learning for People Following in Video

In this section we recall the general configuration of transductive learning using spectral graphs and devise how to exploit this semi supervised learning technique to follow people in video streams. Dealing with surveillance videos, many methods have been proposed to reliably detect and extract people on the scene and for these reasons we propose to work in a scenario where all the people patches are available, e.g. the people images are extracted by a detector, and the initial target is provided as well as examples of non-target objects. This

use-case directly leads to the use of a semi supervised learning algorithm where the labelled elements are the initial samples while the unlabelled elements are all the people patches extracted from the video frames. The purpose is thus to estimate the missing labels and to find the samples that correspond to the given target.

We propose an iterative approach where a transductive learning (TL) algorithm, is exploited to search the occurrences of the target when its appearance is learned and temporally updated, starting from the initial given samples.

The Transductive Learning (TL) configuration we propose here uses both positive and negative labelled elements and was previously introduced by Joachims Joachims [2003]. Suppose to have a set of labelled instances $X^L = \{(x_i, y_i)\}_{i=1}^{l}$ where x_i are the input elements described by their features, e.g. people patch, and $y_i = \pm 1$ the corresponding labels. The labels assume respectively the value $y_i = +1$ ($y_i = -1$) for target (not-target) elements. Suppose also to have a set of unlabelled instances $X^U = \{x_i\}_{i=l+1}^{l+u}$ which are the patches extracted from the video frames. The complete dataset comprises both the model X^L and the candidates samples X^U:

$$D(X,Y) = \{X^L \cup X^U, Y : y_i = \pm 1 \text{ iff } x_i \in X^L\} \tag{3.1}$$

By constructing the problem in this manner we aim at reproducing, frame-by-frame, the information encoded into the target model which can be equivalently interpreted as the problem of propagate the labels from labelled to unlabelled instances in order to classify the complete dataset.

We introduce the undirected graph $G = (V, E)$ with adjacency matrix A, V is the set of nodes, representing elements in X, and E are the edges representing the similarity among nodes. The elements a_{ij} are computed as the σ^2 fixed bandwidth exponential smoothing of nodes distances $\rho(x_i, x_j)$:

$$a_{ij} = \exp\left(-\frac{\rho(x_i, x_j)}{\sigma^2}\right) \tag{3.2}$$

The objective of the TL algorithm is to find a cut of the graph that separates positive and negative elements X^+ and X^-, estimating the labels for unlabelled

elements. Considering D as the diagonal degree matrix $D_{ii} = \sum_j A_{ji}$, and according to Luxburg [2007], the Laplacian graph can be computed as $L = D - A$, or in its normalized version $L = D^{-1}(D - A)$. The TL solves the problem by solving the following minimization problem:

$$\begin{aligned} &\min_{\overrightarrow{z}} \overrightarrow{z}^T L \overrightarrow{z} + c(\overrightarrow{z} - \gamma)^T I(\overrightarrow{z} - \overrightarrow{\gamma}) \\ &\text{s.t. } \overrightarrow{z}^T \overrightarrow{z} = n \text{ and } \overrightarrow{z}^T 1 = 0 \end{aligned} \tag{3.3}$$

where $\overrightarrow{z}$ is the generalized partition vector with elements z_i, $\overrightarrow{\gamma}$ equals $\gamma_+ = \sqrt{\frac{|\{i : z_i < 0\}|}{|\{i : z_i > 0\}|}}$ if $i \in X_+$ and $\gamma_- = -\sqrt{\frac{|\{i : z_i > 0\}|}{|\{i : z_i < 0\}|}}$ if $i \in X_-$ and c is a parameter that trades off training errors versus cut value.
Stepping back, the minimization problem in Eq. (3.3) can be obtained starting from the assumptions that the corresponding inductive learner should have low leave-one-out error (a) and constraining the problem to have averages over examples with similar expected value in the training and in the test set (b). The assumption (a) can easily be solved minimizing the LOO error on classification using a trivial kNN. The LOO error of the classifier can be bounded by:

$$Err_{loo}^{knn}(X,Y) \leq \sum_{i=1}^{N} (1 - \delta_i) \tag{3.4}$$

where δ_i is the kNN margin $\delta_i = y_i \frac{\sum_{j \in kNN(x_i)} y_j a_{ij}}{\sum_{m \in kNN(x_i)} a_{im}}$ with a_{ij} the similarity between x_i and x_j. The minimization of Eq. (3.4) can be obtained by maximizing the margin δ_i and imposing constrained values on the model labels. Margin maximization can be written in matrix form leading to the following constrained optimization problem:

$$\begin{aligned} &\max_y y^T A y \text{ s.t.} \\ &y_i = \pm 1 \text{ if } x_i \in X^L \\ &\forall y_{j \neq i} \in \{-1, 1\} \end{aligned} \tag{3.5}$$

Even if this problem can be efficiently solved using both the s-t mincut algorithm Blum and Chawla [2001] or transductive SVM it usually leads to unbalanced cuts. The assumption (b) is therefore necessary to avoid this issue and the cut size can be accounted using a ratio-cut algorithm Hagen and Kahng [2006]. The traditional ratio-cut is an unsupervised problem and find the optimal solution is known to be NP hard. The constraint on y makes the problem semi-supervised, however letting y to assume real values and exploiting spectral properties of the graph Laplacians yields to an efficient way to find a solution to the balanced ratio-cut problem in a semi-supervised way. The ratio-cut problem minimizes the average weight of the cut leading to balanced a cut of the graph.

$$\begin{aligned} &\max_{y} \frac{cut(G^+, G^-)}{\mid \{i : y_i = 1\} \mid\mid \{i : y_i = -1\} \mid} \text{ s.t.} \\ &y_i = 1 \text{ if } i \in Y^L \text{ and positive} \\ &y_i = -1 \text{ if } i \in Y^L \text{ and negative} \\ &\overrightarrow{y} \in \{+1, -1\}^n \end{aligned} \tag{3.6}$$

Given the Laplacian L of the graph and the partition vector $\overrightarrow{z}$ the ratio-cut becomes:

$$\min_{\overrightarrow{z}} \frac{\overrightarrow{z}^T L \overrightarrow{z}}{\overrightarrow{z}^T \overrightarrow{z}} \text{ with } z_i \in \{\gamma_+, \gamma_-\} \tag{3.7}$$

where γ_+ and γ_- are defined as previously. Even if this problem is NP complete its relaxed version is solved exploiting the *Courant-Fischer Minimax Principle* stating that the second eigenvalue of L is the non-degenerate solution and the corresponding eigenvector, i.e. the Fielder Vector, solves the *argmin* problem. The unconstrained ratio-cut problem is unsupervised, therefore in order to take into account labelled instances a quadratic penalty can be introduced to the objective function in Eq. (3.7) obtaining Eq. (3.3). The optimization problem in Eq. (3.3) can be recast as a *Quadratic Eigenvalue Problem*, (QEP) and solved analytically for positive semi definite matrices using eigen-

decomposition Tisseur and Meerbergen [2001]. Specifically, given the eigen-decomposition $L = U\Sigma U^T$ of the Laplacian, and introducing $\overrightarrow{w} = U^{-1}\overrightarrow{z}$, the constraint in Eq. (3.3) becomes equivalent to setting $w_1 = 0$ because the eigenvector of the smallest eigenvalue is always $\overrightarrow{1}$. Redefining V and Λ as the matrices containing, respectively, all eigenvectors U and eigenvalues Σ except the smallest one, Eq. (3.3) can be rewritten as

$$\begin{aligned} &\min_{\overrightarrow{w}} \overrightarrow{w}^T \Lambda \overrightarrow{w} + c(V\overrightarrow{w} - \gamma)^T I(V\overrightarrow{w} - \overrightarrow{\gamma}) \text{ s.t.} \\ &\overrightarrow{w}^T \overrightarrow{w} = n \end{aligned} \tag{3.8}$$

Finally introducing $G = (\Lambda + cV^T V)$ and $\overrightarrow{b} = cV^T C\overrightarrow{\gamma}$ the objective function can be one more time rewritten, disregarding continuous terms, as $\overrightarrow{w}^T G\overrightarrow{w} - 2\overrightarrow{b}^T\overrightarrow{w}$. Following again the *Courant-Fischer Minimax Principle* the minimization in Eq. (3.8) is then solved for $\overrightarrow{w}^* = (G - \lambda^* I)^{-1}\overrightarrow{b}$ where λ^* is the smallest eigenvalue of

$$\begin{bmatrix} G & -I \\ -\frac{1}{n}\overrightarrow{b}\overrightarrow{b}^T & G \end{bmatrix} \tag{3.9}$$

I is the identity matrix. The optimal value of Eq.(3.3) is computed as

$$\overrightarrow{z}^* = V\overrightarrow{w}^* \tag{3.10}$$

producing a predicted value for each example in the test set. The hard class assignment can be easily obtained thresholding the prediction vector $\overrightarrow{z}^*$.

The approach so far explained is the general configuration of our TL algorithm for people following in video, thereafter we propose to use this algorithm with two different iterative schemes: *single frame transduction* and *multiple frame transduction*. A graphical representation of the two proposed schemes is given in figures 3.3 and 3.4.

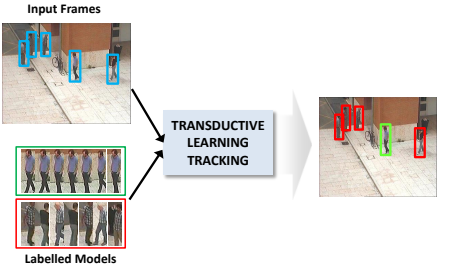


Figure 3.3: Single frame transduction. Only one frame F_i is used. Blue rectangles on the frames on the left side represent the unlabelled elements, while on the frames on the right side red rectangles represent the samples classified as *no target* and green rectangles represent the samples classified as *target* by the TL algorithm.

3.3.1 Single frame transduction

The first possible configuration we propose is based on a single frame transduction where the target is searched among the set of people detected only in the current frame. The input of the algorithm is thus constituted by the two models X_{l+} and X_{l-} of positive and negative labelled elements and the unlabelled elements X_u that are the patches extracted from the considered frame F_i (Fig. 3.3).

Labels for unlabelled elements are computed using Eq. (3.10) and the threshold for the class assignment is set to 0. Ideally only one label should be positive corresponding to the target whilst all the others should be negative. Due to noise of acquisition in real environments and to similarity between people appearances, multiple elements could have a positive predicted label value, we assume, with a good approximation, that the label with the highest value corresponds to the most similar element to the model of positive samples. Conversely, in case of absence of the target, *i.e* missed detection or real occlusion, all the returned labels should have a value less than 0, however the TL could return a positive value and a wrong example with similar appearance could be selected. In this case the update strategy helps avoiding the propagation of the error into the model of positive samples.

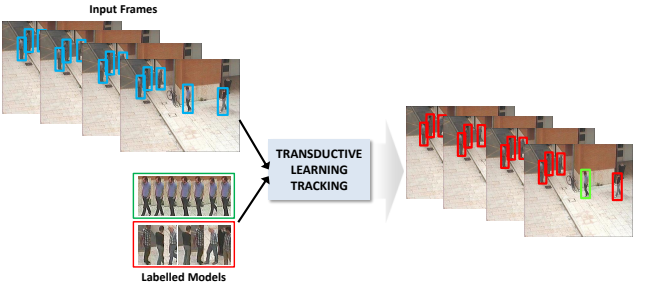


Figure 3.4: Multiple frame transduction. A set $F_i^1, \ldots, F_i^m$ of frames is employed. Blue rectangles on the frames on the left side represent the unlabelled elements, while on the frames on the right side red rectangles represent the samples classified as *no target* and green rectangles represent the samples classified as *target* by the TL algorithm.

3.3.2 Multiple frame transduction

The second setup, shown in Fig. 3.4, employs a multiple frame iterative scheme. The same transductive learning algorithm can be used over the samples of people extracted in multiple frames $\{F_i^1, \ldots, F_i^m\}$. Working with these settings the threshold for the predicted label values for unlabelled elements is again fixed to 0 but the expected number of elements above this threshold is upper-bounded by the number of processed frames, i.e. the target is detected in all the frames.

3.4 People detection and representation

So far we assumed that the input of our framework are the patches of the people on the scene, therefore the first step should consist in the extraction of people patches from each frame of the video sequence. Although the tool we propose is basically independent from the specific people detection method we decide to extract people on the scene using a state-of-the-art people detector based on Histogram of Oriented Gradient, HOG [Dalal and Triggs [2005]]. Dollar et al. [2011] stated how detectors based on sliding windows appears more promising for low to medium resolution settings, which are the typical

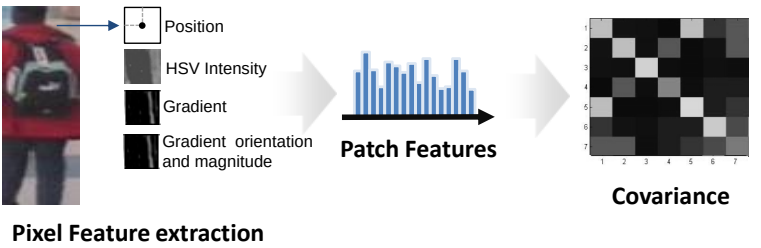


Figure 3.5: Covariance matrix computation.

settings in video surveillance application, with respect to segmentation or keypoint methods that tend to fail with these conditions. Despite HOG have been demonstrated by Dollar et al. [2009] to be one of the most reliable method to detect people in surveillance scenarios, in Dollar et al. [2011] the same authors show how the number of variants of HOG features has proliferated greatly in all modern detectors, and the best performing detectors tend to use a combination of cues overcoming the performances of the earlier approach proposed by Dalal and Triggs [2005]. For simplicity of implementation and since our focus is not placed on the detection stage, but, instead on the following tracking step, we prefer to use the basic formulation, anyway many other detectors can be applied without limitations and the same tracking results can be obtained using different detectors.

3.4.1 Covariance matrix

Once the people bounding boxes are extracted from each frame a predefined descriptor is computed to represent each snapshot and a specific metric is needed to match different occurrences of the same person providing a reliable comparison between multiple samples. The simplest descriptor to represent the patches of the extracted people is probably a color histogram. However this method is not able to distinguish between two persons wearing same colours

but in different position because it ignores all the shape and location information. Here, the growing literature in people re-identification can be evaluated since this step can be re-conducted to the problem of finding a feature to reliably re-identify people. A good survey on the possible feature for people search can be found in the work of Doretto et al. [2011]. To overcome the limits of the color histogram we decided to adopt a covariance matrix descriptor [Porikli et al. [2005]] since it combines color, shape and position cues. The same metric has been previously adopted by Metternich et al. [2010], Liu and Chang [2009], Bak et al. [2010] because of its robustness in matching the region in different views and poses. Covariance matrices exhibit scale and rotation invariance properties and are independent to changes in the average pixels intensity such as identical shifting of color values, i.e. changes in color due to shadows.

The covariance matrix is a square symmetric matrix $d \times d$, where d is the number of selected features, independently from the size of the image window, it has the advantage of being a low dimensional data representation. Given the covariance matrix C its diagonal entries represent the variance of each feature and the non-diagonal entries represent the correlations.

Considering I as a three-dimensional color image F is the $W \times H \times d$ dimensional feature descriptors extracted from I,

$$F(x,y) = \Phi(I,x,y) \tag{3.11}$$

where the function Φ can be any mapping such as intensity, color, gradients, filter responses, etc. Let $\{z_i\}_{i=1\ldots N}$ be the d-dimensional feature points inside F, with $N = W \times H$, the image I is represented with the $d \times d$ covariance matrix of the feature points:

$$C_R = \frac{1}{N-1} \sum_{i=1}^{n} (z_i - \mu)(z_i - \mu)^T \tag{3.12}$$

where μ is the vector of the means of the corresponding features for the points within the region R. In our case z_i is the 9-dimensional feature vector

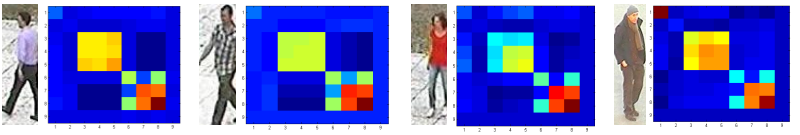


Figure 3.6: Covariance matrix computed from different people patches.

composed for each pixel by its spatial, color and edge information:

$$z_i = [\ x\ \ y\ \ H\ \ S\ \ V\ \ G_x\ \ G_y\ \ mag(x,y)\ \ o(x,y)\ ]^T \tag{3.13}$$

where x and y are the pixel location in the image grid, HSV the intensity values, G_x and G_y the first order derivatives of the intensities calculated through Sobel operator. Finally $mag(x,y) = \sqrt{G_x^2 + G_y^2}$ and $o(x,y) = arctan\left(\frac{G_y}{G_x}\right)$ are the magnitude and the angle of the first order derivatives. Based on this features vector the covariance of a region is a 9×9 matrix, as represented in Fig. 3.5. Figure 3.6 shows some examples of covariance matrices computed from different people patches. We would like to point out that we use HSV color space instead of the basic RGB color space because we experimented an higher invariance to scale and light changes of the HSV components when compared to RGB.

Finally, in order to compute the similarity between people patches, an distance between covariance matrices in necessary. Since covariance matrices do not lie in the Euclidean space any arithmetic subtractions or other simple operations between matrices is not correct. A robust distance metric between the covariance matrices is proposed in Forstner et al. [1999] as the sum of the squared logarithms of the generalized eigenvalues:

$$\rho(C_i, C_j) = \sqrt{\sum_{i=1}^{d} \ln^2 \lambda_k(C_i, C_j)} \tag{3.14}$$

where $\lambda_k(C_i, C_j)_{k=1\ldots d}$ are the generalized eigenvalues of C_i and C_j computed as:

$$\lambda_k C_i x_k - C_j x_k = 0 \;\; k = \{0 \dots d\} \tag{3.15}$$

where x_k are the generalized eigenvectors. Forstner et al. [1999] demonstrated that the distance measure ρ satisfies the metric axioms, positivity, symmetry, triangle inequality, for positive definite symmetric matrices.

After defining both the descriptor and the detector the complete tracking algorithm is depicted in Alg. 1.

3.5 Model Update

Using both positive and negative labelled samples as input of the transductive learning algorithm allows to have a more robust learning and a stable matching method. Of course, in order to have a powerful representation of both the target and the elements differing from the target, both labelled models should be iteratively updated by adding step by step new examples.

The basic idea is to exploit an update mechanism for each model in order to allow to iteratively add new samples to the earlier model, keeping a firm and accurate representation of the target object and avoid the injection of classification errors. Furthermore in this manner we can set the number of samples in each model limiting the number of input elements of the TL, and consequently the resource consumption. Updating the labelled samples is a straightforward solution to the drifting problem that would otherwise arise using only the initial model (which is likely to become obsolete limiting the search process after a certain number of frames). The target appearance, indeed, changes during the video, due for example to rotation, occlusions or different lighting. On the other hand, the main risk updating the models is the propagation of classification errors into the models making the successive labelling unreliable or leading to a deviation when small errors are introduced at each update step.

The easiest update strategy is, of course, a *first-in/first-out* scheme where the last results are iteratively added to the models while the oldest elements are removed keeping only the elements closest in time to the current appearance

Algorithm 1 Transductive Learning People Following

Initialization
$X_{l+} \longleftarrow \{x_i, y_i = +1\}$
$X_{l-} \longleftarrow \{x_i, y_i = -1\}$
while frames f **do**
    Set processing frames $F \longleftarrow \{f_j\}_{j=1}^{m}$
    HOG people detection over F:
        $X_u \longleftarrow \{x_i\}_{i=l+1}^{l+u}$
    Compute covariance matrix for X_l and X_u:
        $C_R = \frac{1}{N-1} \sum_{i=1}^{n} (z_i - \mu)(z_i - \mu)^T$
    Compute similarity matrix A_{ij}:
        $a_{ij} = \exp\left(-\frac{\rho(C_i, C_j)}{\sigma^2}\right)$
        where: $\rho(C_i, C_j) = \sqrt{\sum_{i=1}^{d} \ln^2 \lambda_k(C_i, C_j)}$
    Compute diagonal degree matrix: $Dii = \sum_j Aij$
    Compute Laplacian: $L = D - A$
    Perform Transductive Learning and compute predictions $\mathbf{z}^*$ (See Sec. 3.3)
    Threshold $\mathbf{z}*$ and get hard class
        $X_{r+} \longleftarrow X_{r+} \cup x_i | x_i \in X_u, z_i^* \geq 0$
        $X_{r-} \longleftarrow X_{r-} \cup x_i | x_i \in X_u, z_i^* < 0$
    if $|X_{r+}| >$ thresh **then**
        Update: $X_{l+} \longleftarrow X_{l+}^{new}$
    end if
    if $|X_{r-}| >$ thresh **then**
        Update: $X_{l-} \longleftarrow X_{l-}^{new}$
    end if
end while

of the target. A similar strategy is proposed by Zha et al. [2010], however, despite its simplicity, this update scheme has the main drawback of having no control over errors injection. In other words, by using the last results the models are always up-to-date but when an error in classification occurs there is no restriction on its insertion in the models leading to a drift in the search process. Our proposal is thus to exploit an update strategy based on a clustering where the errors are less likely to be kept in the models and where the elements

are not chosen on their proximity in time but on their representation of different appearances of the target.

3.5.1 Evolutionary Spectral Update

The update strategy we propose is based on *evolutionary spectral clustering*. Evolutionary clustering [Chakrabarti et al. [2006]] is a class of clustering techniques whose aim is to process continuously evolving and timestamped data.

The problem of update the labelled elements can be considered as a dynamic application: the labelled models should represent the evolving differences in appearance of the target and should preserve the time consistency with the previous historical appearances.

Regarding the positive labelled elements, we assume that the samples composing the model can be clustered in a certain number of sets, each one representing a different appearance of the target, i.e. a different pose. After some results are retrieved by the TL step we want to update the model adding the new elements and we want the new clustering depending on the current data features but also not deviating too dramatically from the most recent history. In this way the appearance changes are kept with temporal smoothness leading to a complete representation of the target. Referring to the model of negative elements the same construction can be applied, with the difference that each cluster should, in this case, represent a different person in the scene instead of a different appearance of the target.

Following the idea proposed by Chi et al. [2009] we decide to use an evolutionary version of the spectral clustering algorithm where the number of clusters in which data can be divided is easily derived exploiting the properties of the graph Laplacian. A general cost function is defined to measure the quality of the clustering result on evolving data points, this function embodies two costs. The first cost, called *snapshot cost (CS)*, measures the quality of the current clustering result with respect to the current data features, while the second cost, *temporal cost (CT)*, measures the temporal smoothness in terms of the goodness-of-fit of the current clustering result with respect to historic data

features. The overall cost function is thus defined as:

$$EC = \alpha CS + (1-\alpha)CT \tag{3.16}$$

Focusing on the spectral clustering algorithm in the special case where the number of nodes to be clustered does not change Eq. 3.16 becomes a linear combination of two Laplacians:

$$EC = \alpha \bar{D}_t^{-\frac{1}{2}} \bar{A}_t \bar{D}_t^{-\frac{1}{2}} + (1-\alpha)\bar{D}_{t-1}^{-\frac{1}{2}} \bar{A}_{t-1} \bar{D}_{t-1}^{-\frac{1}{2}} \tag{3.17}$$

In our case setting the number of nodes is equivalent to setting the number of the considered people patches classified by the TL algorithm as belonging to X_{r+} or X_{r-}. Referring to Eq. 3.17 $\bar{A}_t$ and $\bar{A}_{t-1}$ are the affinity matrices built over the model's elements and the last k retrieved results respectively at time intervals t and $t-1$ multiples of k, and $\bar{D}_t$ and $\bar{D}_{t-1}$ are the corresponding degree matrix. Specifically the two matrices $\bar{A}_t$ and $\bar{A}_{t-1}$ are computed exploiting a dissimilarity space where each element is represented in a vector space by the distances with the elements of the model at time t and $t-1$ in $\bar{A}_t$ and $\bar{A}_{t-1}$ respectively (the two representation sets at current and historical time). The affinity is then computed among the current model and last retrieved elements represented with such distances.

This solution turns out to be intuitive, since the historic similarity matrix, i.e. the similarity matrix at the previous iteration, is scaled with a coefficient α and combined with the current similarity matrix, i.e. the similarity matrix at the current iteration. The coefficient α is in $0,1$, with $\alpha = 1$ meaning that only the current data are clustered and $\alpha = 0$ meaning that the model is not updated maintaining the previous clustering.

We exploit the spectral properties of the Laplacian performing the unnormalized spectral clustering Luxburg [2007] in order to obtain k clusters $C_1, \ldots, C_k$, where k is the number of clusters chosen employing the eigengap analysis.

Supposing that $X_{l+} = \{x_i, 1\}_{i=1,\ldots,n_{mp}}$ and $X_{l-} = \{x_i, -1\}_{i=n_{mp}+1,\ldots,l}$ are the

Algorithm 2 Evolutionary Update

Input:
  $X \longleftarrow X_l \cup X_r$ where: X_l current model, X_r last results
  A_{t-1}, D_{t-1} similarity and degree matrix at previous iteration
Begin
Compute current similarity matrix over$X$, A_t:
  $a_{ij} = \exp\left(-\frac{\rho(C_i, C_j)}{\sigma^2}\right)$
Compute diagonal degree matrix: $Dii = \sum_j Aij$
Compute Evolutionary Cost:
  $EC = \alpha D_t^{-\frac{1}{2}} A_t D_t^{-\frac{1}{2}} + (1-\alpha) D_{t-1}^{-\frac{1}{2}} A_{t-1} D_{t-1}^{-\frac{1}{2}}$
Perform Spectral Clustering on EC and get clusters $C_1, \ldots, C_k$
Return X_l^{new}

positive and negative model constituted respectively by n_{mp} and $n_{mn} = l - n_{mp}$ elements and supposing that X_{r+} and X_{r-} are the last r positive and negative results, the updated models are X_{l+}^{new} and X_{l-}^{new} are built performing the evolutionary spectral clustering and contains the same number of elements of the previous models. The elements of the new models are chosen maintaining some elements for each cluster $C_1, \ldots, C_k$, representing with good approximation all the different appearances of the target in the positive model, or all the elements that differs from the target in the negative model. The update method is depicted by

$$\begin{aligned} & X_l^{new} = \left\{x_j, \pm 1\right\}_{j=1,\ldots,n_m} \text{ s.t.} \\ & |X_l^{new}| < n_m \\ & x_j \in C_i,\ i = 1, \ldots, k\ ,\ \bigcup_i C_i = X_l \bigcup X_r \end{aligned} \tag{3.18}$$

The full algorithm is depicted in Alg. 2 and Fig. 3.7 shows an example of the positive model X_{l+} in a sequence of updates, it can be seen how, at each update, samples representing the target in new and different positions are added to the previous model.

Figure 3.7: Example of a sequence of updates of the positive labelled model $X1+$.

3.6 Datasets

The evaluation is performed on three different datasets:

- **THIS** [1]. This dataset, shown in Fig. 3.8 has been introduced by Vezzani and Cucchiara [2010]. Specifically we used the video category *Train Station*, that includes video recorded along the platforms and underpasses of a train station, mostly representing people walking alone or in groups.

- **CAVIAR** [2] is a widely used dataset for people tracking. Sequences from this dataset, and in particular *Clips from shopping center in Portugal - Corridor view* are collected in the hallway of a shopping center and show people walking, meeting with other groups and entering or exiting shops, thus contain some occlusions.

- **3DPes** [3] introduced by Baltieri et al. [2011]. This datasets, Fig. 3.9, is specifically focused on re-identification, therefore we used them to evaluate critical aspects of the proposed method, *e.g* people occluding each other, changes in appearance and pose with respect to the camera and people leaving the scene multiple times.

[1] http://www.openvisor.org
[2] http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1
[3] http://imagelab.ing.unimore.it/visor/3dpes.asp

(a) CAVIAR

(b) THIS

Figure 3.8: Examples of frames taken from THIS and CAVIAR videos.

Referring to Fig. 3.9 sequence (a) has very long occlusions when people go behind the large pilaster. In this case classical trackers do not give the same label to the same people anymore because other similar people are visible in the area. Sequence (b) is an example of small size people which make people less distinguishable and finally sequence (c) depicts different people with variable aspects, in fact people change their dresses and are detected before in frontal and then in opposite way as well as appear and disappear from the scene often. Each sequence taken from CAVIAR and THIS datasets consists of approximatively $300-600$ frames, while sequences from 3DPes have approximatively $2000-2500$ frames, with a number of people variable from 2 to 5. We tested our system using all possible target in each sequence.

(a)

(b)

(c)

Figure 3.9: Examples of frames taken from videos of the 3DPeS dataset.

3.7 Experimental results

In this section we assess the performance of the proposed method for people following in video. Patches of the people on the scene are detecting using

a conventional HOG people detector [Dalal and Triggs [2005]], the proposed graph transduction method (Sec. 3.3) is then used to classify unlabelled patches using both positive and negative labelled elements. In our experiments the labelled models X_{l+} and X_{l-} are firstly manually initialized by the user and afterwards automatically updated with the strategy explained in Sec. 3.5. The systems works iteratively on each frame or set of frames depending on the choice of Single Frame Transduction (SFT) or Multiple Frame Transduction (MFT).

3.7.1 Evaluation Measures

In order to test the proposed solution we adopt measures similar to those proposed by Kuo et al. [2010] for tracking pedestrians in sparse scenes. The rationale behind this choice is that traditional per-pixel or coverage measures are not particularly suitable for tracking by detection methods. In particular we focus on evaluating how many times a target is correctly detected and tracked during the time it appears in the scene. Our proposal is specifically designed to be robust against drifting using the evolutionary spectral update algorithm for target model update, Sec. 3.5, thus the proposed evaluation measures aim at both accepting tracking errors and globally evaluating tracking results over the complete ground truth sequences, i.e. complete people trajectories.

We measure the algorithm performances in term of:

- Ground truth (GT): number of ground truth people sequences, i.e. trajectories.
- Mostly tracked (MT) : number of sequences that are successfully tracked for more than 80% of their duration (number of target images correctly tracked divided by the ground truth target images sequence).
- Partially tracked (PT): number of sequences that are successfully tracked in 20% of the ground truth frames.
- Mostly lost (ML): number of sequences that are successfully tracked for less than 20%

Table 3.1: Single target people tracking. Comparison of different SSL methods.

	THIS Dataset					
	GT	MT (%)	PT (%)	ML (%)	P (%)	R (%)
TFPN EvoU	48	94	3.8	2.2	96	97
TTP U	48	92	5.8	2.2	91	95
TLT	48	76	15.1	8.9	92	76
	CAVIAR Dataset					
TFPN EvoU	140	92	7.3	0.7	94	95
TTP U	140	82	14.6	3.4	87	89
TLT	140	72.5	18.4	9.1	91	73
	3DPeS Dataset					
TFPN EvoU	50	58.3	38.7	3	76	60
TTP U	50	51.2	32.4	16.7	39	54
TLT	50	44.1	34.5	21.4	35	49

Since object tracking can be viewed as a method able to recover missed detections and remove false alarms from the raw detection responses, we also provide the metrics for detection evaluation:

- Precision (P)the number of correctly matched detections divided by the number of output detections
- Recall (R): the number of correctly detected elements divided by the ground truth elements number

3.7.2 Evaluation of the impacts of negative labelled elements

We first assessed our complete method with single frame processing on different videos measuring precision and recall. The optimal parameters were heuristically selected in order to minimize errors in the classification and in the model update. As a result we set $\sigma = 0.4$ for the computation of the affinity matrix, $\alpha = 0.8$ for the evolutionary spectral update and the number of elements in the positive and negative labelled models respectively as $n_{l+}^{max} = 10$ and $n_{l-}^{max} = 15$.

Figure 3.10: 1st and 2nd rows: Some example frames of a sequence specifically focused on testing robustness in case of appearance changes. 3rd row: Obtained results.

To demonstrate the effective improvement given by the conjunctive use of the positive and negative labelled elements we compared our proposal (TFPN EvoU) with the work by Coppi et al. [2011a] (TTP U) where only positive labelled elements were exploited and with the baseline transductive tracking proposed by Zha et al. [2010] (TLT). In the work of Coppi et al. [2011a] the model is updated with a strategy based on the spectral properties of the graph laplacian while in [Zha et al. [2010]] the weights of the labelled elements are simply decreased in time. Obtained results on the different datasets are reported in Tab. 3.1. Results compares favourably with the previous methods. The gap in precision and recall demonstrates the effectiveness of our update strategy in modelling the changes in appearance and maintain an up-to-date representation of the target. Particularly the results display how the approach presented in this paper outperforms the other methods when working with the 3DPeS videos. These sequences are more challenging when compared to videos in THIS and CAVIAR datasets, in fact people exhibit less variability in dresses colours, move unpredictably changing the pose with respect to the camera, light conditions are different from point to point of the scene etc.; for all these reasons reaching an expressive gap in performances with these videos demonstrates the higher reliability and robustness of our method.

Table 3.2: Single target people tracking. Precision and Recall Values on test datasets using single or multiple frame processing.

	Average Precision (%)			**Average Recall (%)**		
	THIS	CAVIAR	3DPeS	THIS	CAVIAR	3DPeS
SFT	97	96	76	95	94	60
MFT 3	96	96	84	96	93	66
MFT 5	94	93	85	94	92	53
MFT 8	95	97	79	90	94	42

3.7.3 Single frame vs. Multiple frame processing

The second part of our experiments is focused on a comparison between the single and multiple frame processing explained in Sec. 3.3. In this experiment we evaluate the impact on precision and recall of the proposed solution when processing either a single frame (SFT) or multiple frames, respectively 3,5,8 (MFT3, MFT5, MFT8). Tab. 3.2 exhibits the value of precision and recall when the number of frames used as unlabelled input elements of the transductive learning increase.

We get comparable results on all the different configurations meaning that the system has a noticeable degree of robustness in terms of the number of unlabelled samples provided as input. A slight improvement in both precision and recall can be seen when using a number of 3 frames, whilst a decrease is obtained processing an increased number of frames. This can be reasonably explained because when the number of input frames increases beyond a certain value the appearance of the target changes with respect to the model leading to uncertain classification.

Some qualitative results obtained performing the proposed method on the 3DPeS videos in Fig. 3.9 are shown in Fig. 3.11. Fig. 3.12 lists few cases of failure, the first rows contains some wrong responses of the people detector which have been classified as the target by the transductive learning. Moreover the second and the third rows display two different cases where the transductive learning merge two people with very similar appearance. These few cases of failure are indicative of the weakness of our proposal: first of all using His-

Figure 3.11: Results on 3DPeS dataset. Black rectangles denote frames where the transductive learning did not give any positive result, while dots are placed instead of reporting all the target boxes. Numbers under the sequences are frame indexes.

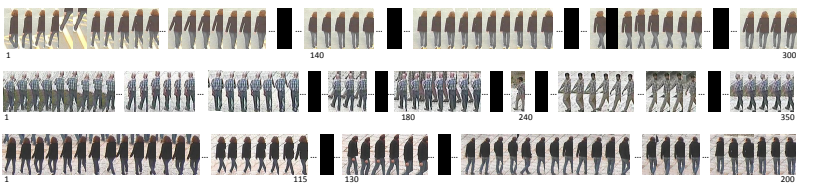


Figure 3.12: Some failure cases on 3DPeS dataset. Black rectangles denote frames where the transductive learning did not give any positive result, while dots are placed instead of reporting all the target boxes. Numbers under the sequences are frame indexes.

tograms of Oriented Gradients to detect the people on the scene sometimes returns imprecise detection that can increase the number of false positive elements when the detection contains significant parts of the background or the foreground of the target. Secondly using only appearance information represented as covariance matrix tend to fail when people do not show distinctive patterns and are very similar to each other, i.e. same all black dresses and hair on the same background.

Finally, as an evidence of the robustness of the proposed update method we recorded some sequences specifically focused on strong appearance changes, i.e. sequences where people rapidly turn on themselves or remove and wear their jacket. We can see in Fig. 3.10 some frames of these videos and the corresponding results of the search. Notice that, despite the complex appearance

changes and some occlusions with other people on the scene, the method is able to correctly detect and follow the target person.

3.7.4 Comparisons on CAVIAR dataset

Finally, on the CAVIAR dataset we compare our proposal with several learning based tracking by detection state-of-the-art approaches in order to show the performance improvement using transductive learning and spectral updating for people tracking in unconstrained scenarios. As in previous tests the detection of the people on the scene is obtained using the HOG based people detector [Dalal and Triggs [2005]] and transductive people tracking is run over the detected snapshots in every frame. Table 3.3 shows the result assessed by our proposal in comparison with different learning based tracking approaches [Kuo et al. [2010]].

Table 3.3: Single target tracking. Performance comparison of different methods on CAVIAR dataset

	MT (%)	PT (%)	ML (%)	P (%)	R (%)
Wu *et al*	75.7	**17.9**	6.4		75.2
Zha *et al*	85.7	10.7	3.6		76.4
Xing *et al*	84.3	12.1	3.6		81.8
Huang *et al*	78.3	14.7	7.0		86.3
Li *et al*	84.6	14.0	1.4		89.0
Kuo *et al*	84.6	14.7	**0.7**	**96.9**	89.4
TFPN EvoU	**92.7**	7.3	**0.7**	94.0	**95.0**

We compared our method against methods that perform data association among small fragments of reliable tracks, *tracklets*, and then perform a data association stage among them to recover full tracks proposed by Wu and Nevatia [2007], Zhang et al. [2008], Xing et al. [2009], Huang et al. [2008]. The performance improvement over these approaches is given by the fact that our proposal perform data association among every people snapshots independently being able to better recover in case of errors or model corruption. We additionally compared our solution with two model learning solutions that exploit

boosting to select and rank the best tracked candidates in subsequent frames introduced by Kuo et al. [2010] and Li et al. [2009]. In these cases, we are able to obtain a slight improvement due to the presence of transductive learning. It has been empirically demonstrated that when performing classification over a set of partially labelled data, *e.g* find a target among possible candidates target, transduction can bring some performance improvement making use of both the relation among labelled and unlabelled data and unlabelled data themselves. Additionally, comparing to Kuo et al. [2010] we do not make use of any assumption about objects motion obtaining a solution that is more generally applicable even in case of abrupt motion or strong occlusions. Moreover all these approaches do not explicitly provide any solution against drifting and model corruption. This problem is often delegated to the adopted classifier or to the optimization method used for data association. Our explicit *Evolutionary Spectral Update* algorithm helps avoiding drifting in many situations and facilitates to recover from sporadic errors allowing us to reliably follow the target in a long-term fashion.

CHAPTER

FOUR

MULTI TARGET PEOPLE TRACKING

In this Chapter we extend the data association tracking proposed in Chapter 3 to follow simultaneously all the person in the scene. Multi target tracking is an highly challenging task since people can have similar appearance and intersecting trajectories. Similarly to the previous Chapter we use an object detector to extract people patches from the video frames and we directly use them as data source for tracking.

We propose the graph transduction as a solution to track multiple people in videos formulating the problem as a multi-class partially labelled data problem. We exploit the formulation of Erdem and Pelillo [2012] based on a game theoretic framework and we prove its reliability in solving a real world problem. Graph transduction is here formulated in terms of a multi-player non-cooperative game where the players are the data points that take part in the game to decide their class memberships. We decide to exploit this novel formulation (instead of re-formulate the previously adopted approach) because it can be considered as intrinsically multi-class: labelled players directly represents the different classes. To our knowledge this is the first application of transductive learning to multiple target tracking.

4.1 Graph Transduction Game

The theoretical formulation of the *Graph Transduction Game* (GTG) has been recently introduced in Erdem and Pelillo [2012]. Starting from the basis of the transductive learning on undirected graph, they build a solution in which the label estimation is based on game-theoretic notions, in contrast to common solution based on the spectrum of the graph. Precisely the graph transduction is formulated as a non-cooperative multiplayer game and the labelling correspond to the Nash equilibria.

For notions about multi-player games we refer to Weibull [1995], here we only recall the main ideas and we give the basic definition of the game theory. In game theory a *game* is modelled as a strategic interaction among *players* where the goal of each player is to maximize its own *payoff* by choosing the best action (*pure strategy*) to play. The *graph transduction game* proposed by Erdem and Pelillo [2012] is formulated assuming that players $i \in I$ participating in the game corresponds to particular points in a data set $\mathcal{D} = \{d_1, \ldots, d_n\}$ and that the set of strategy among whom the players can choose is $S_i = \{1, \ldots, c\}$. Each strategy S_i expresses a certain hypothesis about its membership to a class and c is the total number of classes (*i.e* the the mixed strategy profile of each player $i \in I$ lies in the c-dimensional simplex Δ_i).

Since the problem is a problem of SSL, the players belongs to two disjoint groups: those which already have knowledge of their membership, referred to as *labelled players* and denoted with the symbol I_ℓ, and those which do not have any idea about their membership at the beginning of the game, which are hence called *unlabelled players* and correspondingly denoted with I_u. The labelled players $I_\ell = \{I_{\ell|1}, \ldots, I_{\ell|c}\}$ do not need to maximise their payoff since they always play their already chosen k^{th} pure strategy where $k = 1, \ldots, c$. The transduction game can be easily reduced to a game with only unlabelled players I_u that need to find their mixed strategy $e_i^k \in \Delta_i$ and the fixed strategies of labelled players I_ℓ act as bias over the choices of unlabelled players.

For the Nash equilibrium theorem Nash [1951] the GTG always has equilibrium in mixed strategies that corresponds to a steady state where each player

plays a strategy that could yield the highest payoff when the strategies of the remaining players are kept fixed, and it provides us a globally consistent labeling of the data set. Once an equilibrium is reached, the label of a data point (player) i is simply given by the strategy with the highest probability in the equilibrium mixed strategy of player i as:

$$\varphi_i = \arg\max_{h=1\ldots c} x_{ih}, \tag{4.1}$$

thereby yielding a crisp classification.

Similarly to other graph transduction methods the data are represented with an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of nodes representing both labelled and unlabelled points, and $\mathcal{E}$ are the edges weighted with an adjacency matrix $W = (w_{ij})$. Being the solution considered as the equilibrium in a non-cooperative game, the adjacency matrix W is used to compute the pay-off between players. And being the game considered as an instance belonging to the class of *polymatrix games* Quintas [1989], Howson [1972] where players are nodes of a graph and every edge denote a two-player game between corresponding pair of players, the partial pay-off matrix between two players i and j is computed as $A_{ij} = w_{ij} \times I_c$ where I_c is the identity matrix of size c. The pay-offs are then computed as:

$$u_i(e_i^h) = \sum_{j=1}^{n} (A_{ij} x_j)_h \tag{4.2}$$

and

$$u_i(x) = \sum_{j=1}^{n} x_i^T A_{ij} x_j \tag{4.3}$$

The Nash equilibria, thus the labelling for unlabelled points, is computed using the *evolutionary approach* Daskalakis et al. [2006], Daskalakis [2011]. The dynamic interpretation of Nash equilibria through the evolutionary approach imagines that the game is played repeatedly, generation after generation, during which a selection process acts on the multi-population of strategies, thereby resulting in the evolution of the fittest strategies. The particular class of dynam-

ics used in this are the so called *imitation dynamics* given by

$$\dot{x}_{ih} = x_{ih}\left[\sum_{l\in S_i} x_{il}\left(\phi_i\left[u_i\left(e_i^h - e_i^l, x_{-i}\right)\right] - \phi_i\left[u_i\left(e_i^l - e_i^h, x_{-i}\right)\right]\right)\right] \quad (4.4)$$

where the dot signifies derivative w.r.t. time and $\phi_i(u_i)$ is a strictly increasing function of u_i. The multi-population version of the replicator dynamics is obtained when ϕ_i is taken as the identity function, *i.e.* $\phi_i(u_i) = u_i$, as:

$$\dot{x}_{ih} = x_{ih}\left(u_i(e_i^h, x_{-i}) - u_i(x)\right) \quad (4.5)$$

Erdem and Pelillo [2012] demonstrate how in both the discrete and continuous time version of the imitation dynamics the fixed points of Eq. 4.5 are Nash equilibria. Here, the problem is empirically solved using the discrete counterpart of Eq. (4.5), where the mixed strategy of each player is initialized with uniform probabilities:

$$x_{ih}(t+1) = x_{ih}(t)\frac{u_i(e_i^h)}{u_i(x(t))} \quad (4.6)$$

4.2 Experiments

In this section we report the results obtained with the proposed configurations for multi-target people tracking. Similarly to Sec. 3.4 people are extracted from the video frames using a HOG based people detector and described with covariance matrices. The similarity is computed using the generalized eigenvalues between each couple of matrices as explained by Eq. (3.14).

We evaluate our proposal on a set of video sampled from the three datasets, **THIS**, **CAVIAR** and **3DPes**. We limit the evaluation to sequences where at least 2 targets were visible, and, regarding the 3DPes dataset we confine the investigation to the videos without strong appearance changes, since this extension to multi-target people tracking is only a seminal work and we aim at stressing the working condition as future work. We will thus refer to the dataset as 3DPes*, to differentiate from the experiments of the previous Chapter.

Figure 4.1: Examples of results obtained on the THIS (left) and the Caviar (right) datasets. Coloured bounding boxes show the obtained tracking results.

Fig. 4.1 and Fig. 4.2 show some frames taken respectively from **THIS** and **Caviar** and **3DPes** with the associated obtained results.

Working off-line on the people patches the problem can also be considered as one of multiple object data association, therefore we measured the performance in terms of mean object Precision and mean object Recall, where we considered:

- *True Positive* is a people patch classified correctly with its label;
- *False Positive* is a misclassified patch (*i.e* the i^{th} target label is assigned to a different person);
- *False Negative* is missing estimation (*i.e* the i^{th} target label is non assigned in the frame even if that target is present).

In order to abstract the overall performance we also evaluate the F-measure.

Since the approach we proposed is off-line and relies on pre-recorded sequence, we evaluate the results varying the number of initially labelled frames.

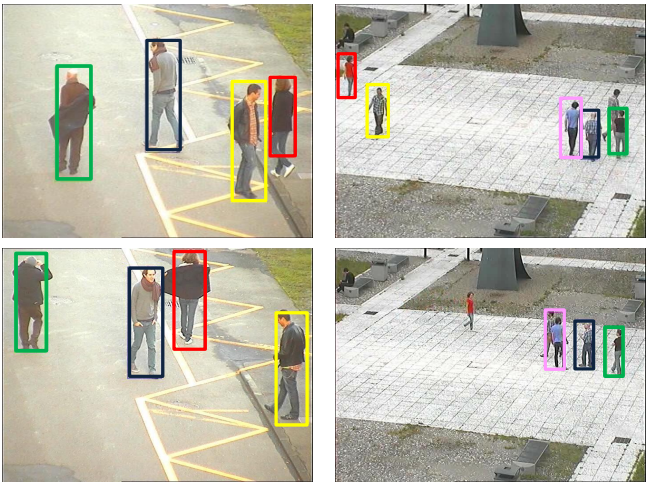


Figure 4.2: Examples of results obtained on the 3DPes datasets. Coloured bounding boxes show the obtained tracking results.

These frames were randomly chosen and the results have been averaged on 20 different executions. Table 4.1 reports the F-measure obtained using a initial labelling of five frames on the three datasets and summarizes the averaged length of the evaluated videos in number of frames and the number of targets to track for each dataset. Fig. 4.3 and 4.4 show the values of mean precision and recall on the three datasets.

Clearly the obtained results improve as we increase the labelled frames as illustrated in Fig. 4.3 and 4.4, but almost saturate to satisfactory values when the number of frames is fixed to 5 demonstrating good reliability without requiring a large number of labelled data. The results also show how increasing the number of labelled frames also increase the stability of the solution while labelling only one or three frames, despite the precision and recall reach adequate levels, the standard deviation is large. The best performing dataset is **THIS** reaching the 100 % of precision and accuracy, but results are good even on challenging dataset such as **3dPes**, where the number of target is higher and the targets are not always correctly identified by the detector due to occlusions,

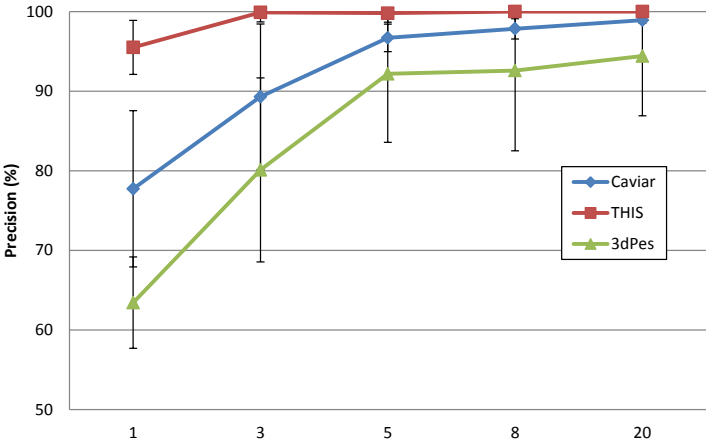


Figure 4.3: Results reported in terms of average precision varying the labelled input frames. The number of labelled frames is reported on the horizontal axis.

Table 4.1: Multitarget people tracking. GTG performances.

Dataset	# Frames	# Targets	F-measure
THIS	109	3	0.99
CAVIAR	140	4	0.96
3DPes*	280	5	0.92

though on the **3dPes** dataset the standard deviation of the results is higher. We would like to highlight that the video length reported in Table 4.1 is an average on different sequences, but, especially with the **3dPes** dataset we stressed the algorithm increasing the number of frames from 200 to 430 and even with the highest number of frames the results were satisfactory with the F-measure of roughly the 90%. Few obtained tracking results are proposed in Fig. 4.1 and Fig. 4.2.

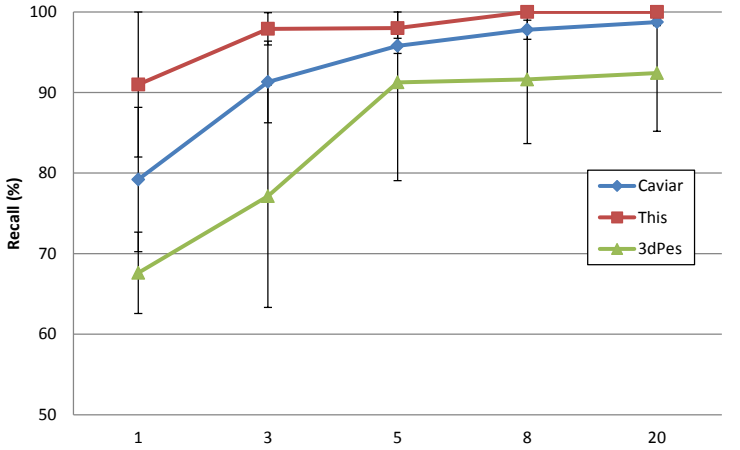


Figure 4.4: Results reported in terms of average recall varying the labelled input frames. The number of labelled frames is reported on the horizontal axis.

4.2.1 Further experiments

To further the evaluation we also compare our approach with the other tracking methods based on graph transduction. At this aim we recast our solution as one of single people tracking, labelling the input samples x_i with $y_i = +1$ when they correspond to the target and $y_i = -1$ otherwise. Results are reported in Tab. 4.2, in particular we evaluate the GTG framework against the work presented in the previous Chapter that exploits positive and negative labelled models with the evolutionary spectral update technique (TTPN SpUpd) and also against the preliminary work proposed by Coppi et al. [2011a] with only positive labelled elements and the update strategy (TTP EvoUpd). Recalling that the other methods work iteratively on-line on each frame of the video while this method performs batch all the video frames, we can see that the results are comparable with the results previously obtained with the single target method, even without any update strategy of the labelled models. The reason of this can be explained both by the robustness of the GTG framework as already demonstrated in Erdem and Pelillo [2012] for other classification tasks, and by the fact

Table 4.2: Graph transduction based people tracking. Comparison between the GTG multitarget tracking and the single target tracking method proposed in Chapter 3.

Dataset	Method	Precision	Recall	F-measure
THIS	***GTG***	1.00	0.85	0.92
	TLPN EvoUpd	0.96	0.97	0.96
	TTP EvoUpd	0.91	0.95	0.93
CAVIAR	***GTG***	0.90	0.92	0.91
	TLPN EvoUpd	0.94	0.95	0.94
	TTP EvoUpd	0.87	0.89	0.88
3DPes	***GTG***	0.87	0.99	0.93
	TLPN EvoUpd	0.86	0.94	0.89
	TTP EvoUpd	0.67	0.71	0.68

that working with graphs with more nodes increase the possible paths among nodes representing the target patches in different frames. In other words this overcomes the inevitable errors of the on-line methods when the people detection are imprecise in the evaluated frames and also offers a straightforward solution to the problem of appearance changes. We can see how, this method achieves better performance with the multi-target configuration, instead of the single target. This is explained because in the multi-class configuration, the different classes (people) are better represented and separated from each other, while in the single target the *non-target* class models the appearance of different target that might also be very dissimilar or even close to the target.

Despite this satisfactory results we would like to test the effectiveness and the robustness of the method also under difficult circumstances expanding the evaluation to the other videos in the 3DPes dataset and eventually other datasets.

Part II

Novelty detection

CHAPTER

FIVE

DETECTION OF NOVEL CATEGORIES AND SUBCATEGORIES OF IMAGES

In object categorization classifiers are trained with positive and negative examples of each class (Fig. 5.1) in order to learn the decision boundary that allows to identify and distinguish one class from the others in the feature space. The countless amount of visual objects unfortunately makes infeasible the task of collecting labelled training data for each existing category. The object classification problem can be seen as a incompletely labelled data problem since it is not possible to have labelled data for each existing category, regardless to the number of category actually considered.

One interesting trend in machine learning is, thus, the detection of new categories of data. This problem can be posed as one of novelty or anomaly detection, where novelty or anomaly is defined in relation to the current knowledge of the system. In other words novel objects are objects that were not seen during training. Despite the considerable advances in object classification, with the state-of-the-art classification methods now reporting impressive results even on difficult and large datasets [Vedaldi and Zisserman [2012], Duchenne et al. [2011], Chatfield et al. [2011], Everingham et al. [2010]], the detection of unknown categories is still an open problem. Yet humans are able

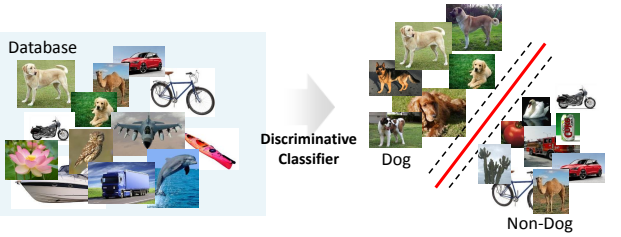


Figure 5.1: Object classification scheme.

to detect novel patterns and to learn new categories of data without having any a priori knowledge about new stimuli. This competence stems from their ability to relate a new structure to meaningful concepts. In other words humans can generalise concepts from their understanding of known objects, and new experiences are then compared with and differentiated from the classes in an existing categorisation framework [Eysenck and Keane [2005]]. While it is likely that a single or few examples are enough for people to understand a new category and infer it as belonging to a novel instance [Biederman [1987a]], machine learning algorithms need a large set of training samples to learn a classification model.

Inspired by human perceptual and reasoning abilities, and considering that training instances could not be given for new classes of objects, we would like to solve at least the first aspect of the problem, namely detecting novel categories whereby novelty is defined in relation to the knowledge that has been compiled from seen categories. Assuming that object categories form a taxonomy where similar categories share the same parent node, we aim at developing techniques that distinguish instances of categories placed outside the considered taxonomy, e.g. that create a new internal node in the tree representing the taxonomy, or correspond to unseen objects. The particular focus is on detecting sub-categories that belong to a known super-category but were not specialised during training, in other words classes that originate a new leaf in the tree. The objective of our work is to investigate different classifier architectures and their

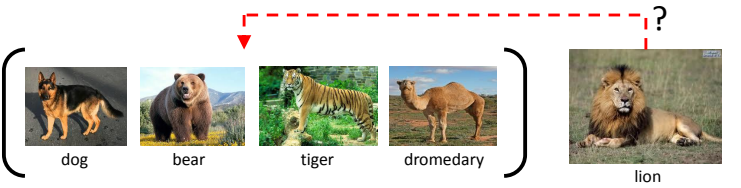


Figure 5.2: New category related to siblings classes.

associated anomaly flagging mechanisms for novel class detection.

5.1 Background and Related work

Recently, there has been an increased interest in novelty detection, *i.e* , the ability to detect if new data is of a type (class, model or domain) that has not been seen during training. In Markou and Singh [2003a,b], comprehensive surveys are offered on novelty detection with the main distinction being made between statistically based approaches and neural network approaches. The reviews also identify various application domains where novelty detection is important. In Computer Vision, novelty detection has recently been apporached by Lampert et al. [2009] and Weinshall et al. [2012], among others. Lampert et al. [2009] focuses on detecting unseen categories of objects by using attributes. In order to make predictions about classes with no training data, they learn a representation that goes beyond the class boundaries merging images of the object classes that are characterized by the same attribute. Our approach shares the idea that the knowledge about unseen classes should come to related known classes but we assume that this expertise comes from the global representation of the images, and not from a disjoint set of attributes.

On the other hand, Weinshall et al. [2012] demonstrate how in a hierarchicaly organised object class taxonomy, novelty can be identified in terms of disagreement between two classifiers making decisions at different levels of

the hierarchy. In particular, a novel object is defined as an input whose probability of belonging to a parent class (general concept) is high but at the same time the probability of membership in any known specific (child) class is low. Despite demonstrating how this framework can be applied to several domains, ranging from detecting novel classes of visual and audio objects, through out-of-vocabulary word detection, to detecting novel patterns of motion, the experiments presented are at proof of concept level.

A broader notion of novelty is anomaly [Chandola et al. [2009]], which refers to the problem of finding patterns in data that do not conform to expected behaviour. Kittler et al. [2014] proposed a taxonomy of anomalies, that include outlier detection, novel class detection and domain change detection. Based on this definition, a direct solution to the problem of detecting anomalies is to determine a region in the observation space representing the normal behaviour and classify any object that lies outside this area as an outlier or anomaly. Many approaches propose to identify anomalies using generative methods in a statistical framework [Almajai et al. [2012], Deng et al. [2012], Rodner et al. [2011], Pauwels and Ambekar [2011]]. Although such solutions might be appealing in low dimensional spaces, the problem is very challenging in other (more common) situations.

The problem of novelty detection and modeling new classes also relates to that of zero shot learning, which refers to the ability to recognise classes that were not seen during training, see Rohrbach et al. [2011] and references therein. Our work can be viewed as an extension of the concept of zero shot learning to a taxonomy of different categories. While zero shot learning aims at modelling a new class, our approach has the ability to (i) first of all, identify if the novel sample belongs to a novel (sub-)class and (ii) define the location of this novel (sub-)class with respect to a known taxonomy, *i.e* , it indicates how to modify a class hierarchy to accommodate for this new (sub-)class.

5.2 Problem Statement

We follow an approach similar to the one proposed in Weinshall et al. [2012] and try to identify novel classes of data using a hierarchy of classifiers by employing an incongruence detector which measures disagreement between classifier outputs at different levels of the hierarchy. This approach is compared with a novel class detection approach using a flat classifier structure where every concept or its subgroup is considered as a separate class, and novelty is flagged by none of the classifiers detecting a positive stimulus. We initially confine the investigation to a group of concepts studied in Weinshall et al. [2012], which contains images of different types of motorbikes, and demonstrate the feasibility of the methods and the relative advantages of the two classifier architectures. While the hierarchical structure using an incongruence measure offers better computational efficiency and a better understanding of nuances of novelty, the flat structure exhibits a slightly better detection performance. We extend the experiments to a larger taxonomy of objects and we show that the simpler hierarchical approach of Weinshall et al. [2012] breaks down when the semantic object categorisation does not map onto a corresponding visual similarity hierarchy, and we also demonstrate how a blend of the hierarchical and flat approaches yield better performance.

Starting from the assumption that an object from an unknown class belongs to a sibling class of known categories, Weinshall et al. [2012] define the incongruent or novel event in relation to partial order on a set of classes. The partial order can be represented by a directed graph and subset-superset relations in the graph can be modelled as conjunctive and disjunctive hierarchies. More precisely a conjunctive hierarchy models part-of membership, *e.g* head, legs and tail combine to form a dog and can be considered as more general concept than the dog. A disjunctive hierarchy models class membership, where an object can be defined at different levels of generality, *e.g* Beagle and Collie are specific concepts of dog. In this paper, we explore disjunctive hierarchies for object classification, where semantically related subcategories of images share the same parent node.

5.2.1 Notation

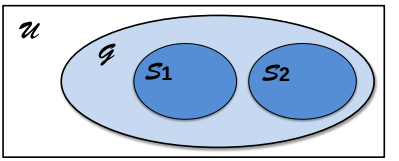
Following section 3.1 of Weinshall et al. [2012], we define a general concept $\mathcal{G}$ as a superset of more specific concepts $\{\mathcal{S}_i\}$. However, the union of all the known specific concepts does not form the complete set that comprises the general concepts, i.e., $\cup_i\{\mathcal{S}_i\} \subset \mathcal{G}$ (rather than $\cup_i\{\mathcal{S}_i\} = \mathcal{G}$). We also assume that during training, samples are given from the set of known subcategories $\cup_i\{\mathcal{S}_i\}$ and also from a small set of a background class that does not belong to $\mathcal{G}$.

As illustrated in Fig. 5.3, the input space of the algorithm is therefore defined by the union of the disjoint sets $\mathcal{S}_i$, while the output space is represented by the three possible classification results *Known, Unknown, Background*. In details:

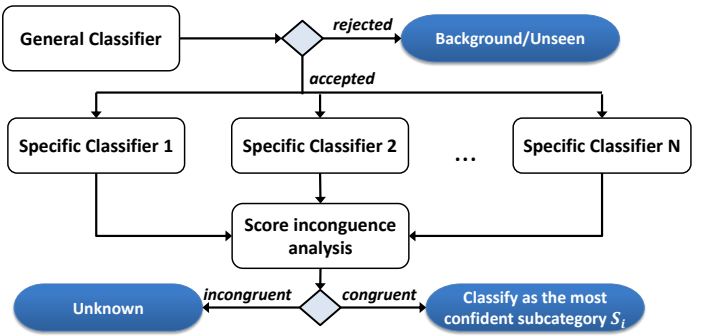
- *Known*: samples that belong to the set of subcategories $\cup_i\{\mathcal{S}_i\}$ that are known from the training set.
- *Unknown*: samples that belong to $\mathcal{G} \setminus \cup_i\{\mathcal{S}_i\}$, i.e., they are from a known general category but do not belong to any of the subcategories that were known during training.
- *Background*: samples that are rejected by the general level classifier, i.e., they do not belong to the general concept $\mathcal{G}$ and are detected because the general classifier used background samples at training (they belong to $\mathcal{U} \setminus \mathcal{G}$).

Background samples are collected using images that clearly do not belong to the known general class, such as background regions of images or textures. They certainly do not cover the infinity set of possible object classes that do not belong to the known general category $\mathcal{G}$. Therefore, further to testing with a test split of the *Background* set we also test with another set of samples collected from other foreground object classes that do not belong to $\mathcal{G}$. This set is labelled as the *Unseen* set, as it is not similar to any object category seen at training.

Therefore, even though the method produces 3 types of labels (*Known*, *Unknown* subcategory and *Background*), the test set samples have four types



(a) Sets diagram.



(b) Classification flowchart.

Figure 5.3: Class types shown in a sets diagram (a) and a flowchart that summarises the incongruence detection method for disjunctive hierarchies, proposed in Weinshall et al. [2012].

of labels (the above plus *Unseen*). Following Weinshall et al. [2012], we consider that if *Unseen* samples are classified as *Background*, the method has succeeded.

5.3 Classification schemes

The following sections detail the four different classification schemes we evaluate to detect novel categories of images and describe the object description method we adopted.

5.3.1 Disjunctive hierarchies with Binary SVMs (B-SVMs)

Weinshall *et al* proposed to identify novel classes or subclasses of images using the incongruence between classifiers at different levels of a hierarchy. Let $M^{\mathcal{G}}$ be the model learnt on a general concept using samples from $\cup_i\{\mathcal{S}_i\}$ and $M_i^{\mathcal{S}}$ the models learnt on specific concepts or subcategories. The detection of a novel category is based on the disagreement between the predictions of the different models. In other words, a sample is identified as novel when is accepted by $M^{\mathcal{G}}$, but rejected by all $M_i^{\mathcal{S}}$. Conversely, a sample belonging to one of the known categories is accepted both by $M^{\mathcal{G}}$ and one of $M_i^{\mathcal{S}}$, as illustrated in Figure 5.3.

At the specific level, a decision score $V_i(x)$ is obtained for each sample x and for each learnt model $M_i^{\mathcal{S}}$. The binary-SVMs method (B-SVMs) uses SVMs for classification at all levels in a one-against-all scheme. Since SVMs are discriminative, Weinshall et al. [2012] propose to whiten the classification scores as follows:

$$S_i(x) = \frac{V_i(x) - V_i^w}{V_i^c - V_i^w} \ , \tag{5.1}$$

where V_i^c is the average confidence of train or validation examples classified **correctly** using $M_i^{\mathcal{S}}$ and V_i^w is the same for examples classified **wrongly** using $M_i^{\mathcal{S}}$.

Weinshall *et al* rely on the assumption that sibling classes semantically grouped in the same super class also have similar feature vectors. This theory is generally accepted and exploited in hierarchical image classification methods, and can also be exploited in the context of novelty detection for classes that were not seen during training. Later in this paper, we will demonstrate that this assumption is not sufficient when a wider taxonomy of images is considered and the visual hierarchy is not trivial.

5.3.2 One-class SVMs (OC-SVMs)

We propose to use the same architecture as Sec. 5.3.1 (Fig. 5.3), but replacing binary SVMs by OC-SVMs. The configuration is explained in Fig.5.5(a). OC-

SVMs [Schölkopf et al. [2001]] are usually exploited in the context of outlier detection when only positive training samples are given. They aim to find the hypersphere that best encloses the training data, differently from common binary SVMs that try to find the hyperplane that best separates two training classes, *i.e* , they are designed for outlier detection. By setting the parameter ν, OC-SVMs can be properly tuned to recognise a fraction of the training samples as outlier and allow for errors and uncertainty in the training set, so there is no need to use (5.1) to normalise the scores. Similar to common binary SVMs, OC-SVMs can be used in their dual formulation.

5.3.3 B/OC-SVMs

We also evaluated a hybrid combination of binary and one-class SVMs in which a binary SVM was used as the general classifier for $\mathcal{G}$ and OC-SVMs are used as specific subcategory classifiers for $\mathcal{S}_i$. The motivation for this combination is that both positive and negative training samples are given at the general level (*i.e* $\mathcal{G}$ and *Background* samples), but for each of the specific level classifiers, only positive samples are given for training ($\mathcal{S}_i$). *Unknown* samples are not given.

5.3.4 Flat model

In contrast to the previous approaches, the class hierarchy is not explored by this method. Instead, it treats the novel subclass as a category of objects that differs from all the known subclasses *and the background class*. In a problem with N subcategories (regardless of the number of super-categories), a set of $N+1$ one-vs-all binary SVM classifiers is trained: one for each known subcategory and one for the background class (see Fig.5.5(b)). A new object is classified as novel if it is rejected by all the $N+1$ classifiers. Having N subcategories plus the background category makes it possible to disregard the normalisation in (5.1) since each classifier has a larger number of negative training examples that results in a decision boundary that better encloses the positive training data points.

5.3.5 B-SVMs/Flat model

This configuration of classifiers can be considered halfway between the B-SVMs and the Flat model. The model $M^{\mathcal{G}}$ is learnt on a general concept in the same way as Sec. 5.3.1, using samples from $\cup_i\{\mathcal{S}_i\}$. The specific model $M^{\mathcal{S}_i}$ is learnt using instances from $\mathcal{S}_{\hat{i}}$ as positive samples and, for the negative class, all the subcategories $\mathcal{S}_{i,i\neq\hat{i}} \cup \mathcal{S}_j$ and the *background* category, *i.e* samples from all the N subcategories that differ from the current one, as explained in Fig. 5.4(b). This is in contrast to the configuration of 5.3.1, where only samples from $\mathcal{S}_{i,i\neq\hat{i}}$ were used as negative training instances, Fig. 5.4(a).

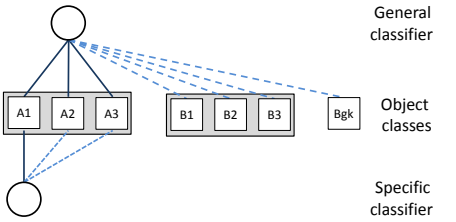
Proposing this approach we would like to benefit from the advantages of the hierarchical configuration which reduces the number of candidate subclasses to evaluate for each sample, and to benefit from the classification performance of the Flat structure which is able to learn a better decision boundary.

5.4 Image Representation

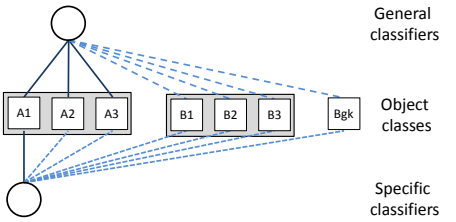
In order to build a vectorial representation x of each image, we used a method that has proven to be state-of-the-art in the benchmark presented in Chatfield et al. [2011]. Images are represented using Pyramid Histograms Of visual Words (PHOW – based on Lazebnik et al. [2006]), encoded with Fisher Vectors [Perronnin and Dance [2007], Perronnin et al. [2010]]. More specifically, SIFT descriptors are computed on a dense grid at four different scales defined by setting the width of the spatial bins of SIFT to 4, 6, 8 and 10 pixels. PCA is performed on the obtained local features and the dimensionality is reduced to 80 components.

Fisher Vectors (FV) are built by concatenating Gaussian gradient vectors $x = [\cdots, \mathcal{F}^F_{\mu,i}, \mathcal{F}^F_{\sigma,i}, \cdots]$ w.r.t. mean μ_i and standard deviation σ_i (the variables are assumed independent), for each Gaussian i in a GMM that models all training features f, where

$$\mathcal{F}^F_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{f_t - \mu_i}{\sigma_i} \right) \tag{5.2}$$



(a) B-SVMs classification scheme



(b) B-SVMs/Flat classification scheme

Figure 5.4: Representation of the B-SVMs and B-SVMs/FLAT schemes. Object classes belonging to the same general category are grouped in gray boxes. Connections from classifier to categories with straight lines means the object category is used as positive training samples, connections with dotted lines means negative training samples.

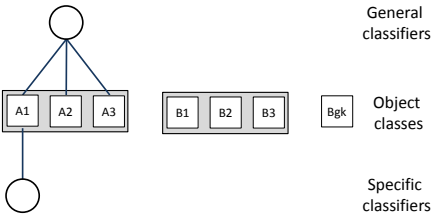
and

$$\mathcal{F}_{\sigma,i}^{F} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(f_t - \mu_i)^2}{\sigma_i^2} - 1 \right] \tag{5.3}$$

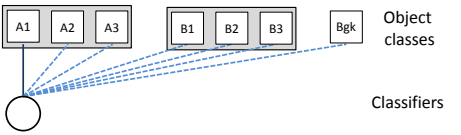
where $\gamma_t(i)$ represents the soft assignment of the descriptor of patch f_t to the Gaussian i and F is the set of T descriptors f_i of an image region.

This is done in each region of the spatial pyramid, which was set up combining regions in this configuration: 1×1, 2×2 and 3×1 and the FVs of each of these regions are concatenated for each image (See Fig. 5.6). This results in a vectorial representation x of $D = M \times 2G \times R$ dimensions per image, where $M = 80$ is the local feature dimensionality (after PCA), $G = 256$ is number of Gaussians in the mixture and $R = 8$ is the number of pyramid regions.

For the above, we used the implementation publicly available in the VLFeat



(a) OC-SVMs classification scheme



(b) Flat classification scheme

Figure 5.5: Representation of the OC-SVMs and FLAT schemes. Object classes belonging to the same general category are grouped in gray boxes. Connections from classifier to categories with straight lines means the object category is used as positive training samples, connections with dotted lines means negative training samples.

toolbox Vedaldi and Fulkerson [2008].

5.5 Kernels

In all the experiments we used the Hellinger (or Bhattacharyya) kernel, which is an additive kernel. Vedaldi and Zisserman [2012] state how additive kernels usually yield classification results similar to non-linear kernels while being at the same time efficient to compute. Additive kernels are in the form $K(x,y) = \sum_{i=1}^{D} k(x_i, y_i)$ where k is itself a kernel (and $x, y \in \mathbb{R}^D$). In particular the Hellinger kernel can be computed for non-negative vectors as $k(x,y) = \sqrt{xy}$. This can be easily extended to arbitrary vectors: $k'(x,y) = sign(xy)k(|x|,|y|)$. The interesting advantage of these kernels is to allow to perform an explicit embedding of the data and then learn a linear classifier in the new space. For

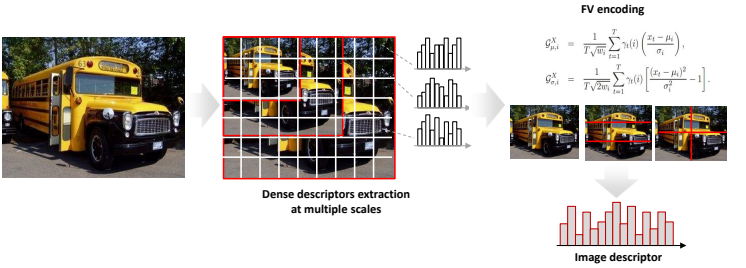


Figure 5.6: Fisher vectors computation.

example for the Hellinger kernel we can define a feature map as $\varphi(x_i) = \sqrt{x_i}$ and then

$$K(x,y) = \langle \varphi(x), \varphi(y) \rangle = \sum_{i=1}^{D} \sqrt{x_i}\sqrt{y_i} = \sum_{i=1}^{D} \sqrt{x_i y_i}. \tag{5.4}$$

5.6 Datasets

5.6.1 Caltech256 - Motorbikes

The first evaluation setting is based on a small sub-hierarchy of the Caltech256 dataset introduced by Griffin et al. [2007], which was used in the novelty detection experiments of Weinshall et al. [2012]. The category *Motorbikes* is chosen as the general concept and the hierarchy is represented by the three more specific subclasses: *Cross*, *Road* and *Sport*. Finally the *Clutter* class images are used as negative examples for the general level classifier and twenty two object classes (different from *Motorbikes*) are sampled to serve as *Unseen* objects. Fig. 5.7 shows the structure of the taxonomy.

5.6.2 Caltech256 - Transportation

In addition, we evaluate a more extensive hierarchy of images using the transportation hierarchy in Caltech256, and specifically *Air* and *Ground transportation*, Fig. 5.8(a). These two super-categories are respectively divided

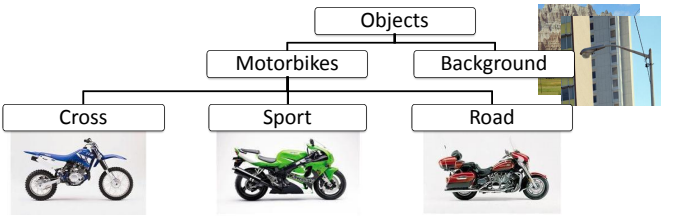


Figure 5.7: Samples of the *Caltech256 - Motorbikes* dataset in the taxonomy of Weinshall et al. [2012].

in *Blimps*, *Fighter Jets*, *Helicopters*, *Airplanes* and *Fire Trucks*, *Motorbikes*, *Car Sides*, *School Bus*. As in the *Motorbikes* dataset, the *Clutter* category is used as negative class at the general level and a set of samples of twenty-two different classes are used as *Unseen* samples.

5.6.3 SUN397

In order to further evaluate the proposed framework we used a subset of the SUN397 Scene Categorization dataset introduced by Xiao et al. [2010]. This dataset has a hierarchical division of the scenes in *Indoor*, *Outdoor natural*, *Outdoor man-made*, Fig. 5.9(b). We sampled four more specific classes for each one of the three super-categories.Specifically Indoor scenes are divided in *Cathedral*, *Classroom*, *Library* and *Stage*; Outdoor natural scenes are divided in *Hill*, *Islet*, *Skislope* and *Snowfield* and finally Outdoor man-made are divided in *Chalet*, *Train railway*, *Runway* and *Windfarm*.

Differently from Caltech256, this dataset does not contain a category that can be used as negative example for the general level classifier, *e.g* the *Clutter* category. For this reason, and because the chosen taxonomy covers all the three super categories of which the dataset is composed, we decided to focus only on the detection on *Novel* subcategories disregarding the *Unseen* classes.

Figure 5.8: Samples of the taxonomies chosen from *Caltech256 - Transportation* dataset.

5.6.4 Oxford Flowers 17

The last hierarchy we used is taken from the Oxford Flowers 17 dataset [Nilsback and Zisserman [2006]], this dataset, usually adopted for segmentation, is composed of 17 categories of flowers. We built our hierarchy grouping in the same parent node flowers of the same color. Specifically we used as general level categories *White flowers* and *Yellow flowers*, respectively partitioned in *Snowdrop*, *Windflower*, *Lily valley*, *Daisy* and *Dandelion*, *Colts foot*, *Cowslip*, *Buttercup*. The background category used as negative examples, instead, is composed by *Fritillary*, *Iris* and *Tigerlily* (Fig. 5.10).

We exploit this taxonomy, for some aspects easier than Caltech256 - Transportation and SUN, to perform some further experiments described in Sec. 5.8.

5.7 Experimental Results

Using the datasets explained in Sec. 5.6 the experiments were repeated with a leave-one-class-out approach on the subcategories to simulate the novel class.

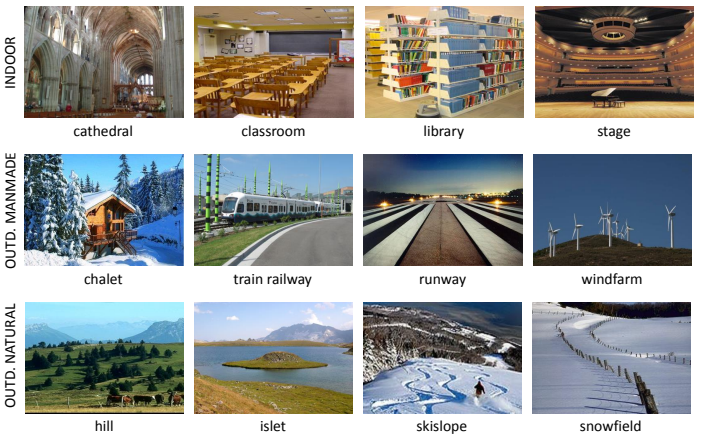


Figure 5.9: Samples of the taxonomies chosen from *SUN* dataset.

The training data of the general level classifier consists of the combination of the known subcategories as positive samples and the clutter class as negative samples. The specific level classifiers were trained using, as positive samples, objects from one of the subclasses used to train the general level SVM and, as negative samples, objects from the other subclasses depending on the used approach. For each subcategory, 39 images were chosen randomly for training and 20 for testing, as done in Weinshall et al. [2012]. The experiments were repeated 25 times, sampling different train and test samples. Experiments are carried out exploiting the open-source *LibSvm* library. The parameters of the SVM classifiers were optimised using cross validation. Table 5.1 shows the average detection scores obtained on the two datasets for each classification scheme, our results are compared with the results of Weinshall et al. [2012]. For the evaluation of the **B-SVMs/Flat** model, since the positive and negative classes were unbalanced, we used a SVM implementation with weighted cost functions, *i.e* the cost parameter C were set to $w_+ \times C$ and $w_- \times C$, with $w_+ \neq w_-$ for positive and negative training samples.

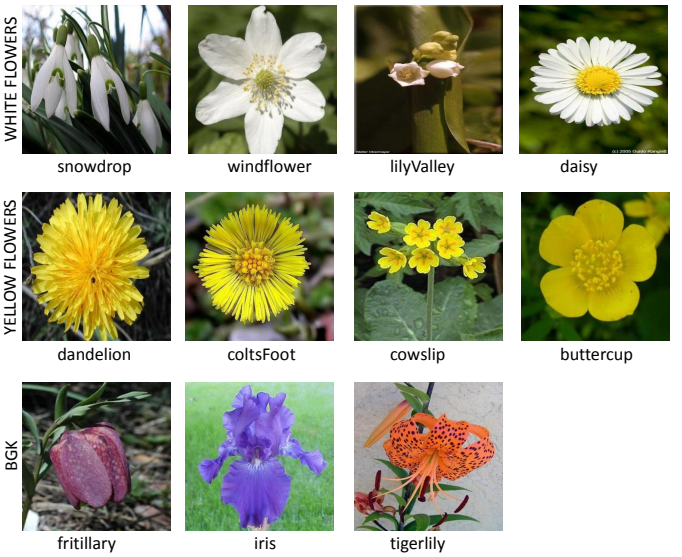


Figure 5.10: Samples of the taxonomies chosen from *Oxford Flowers 17* dataset.

5.7.1 Evaluation on Caltech256 - Motorbikes

We exploited this hierarchy to evaluate all the classification schemes detailed in Sec. 5.3. As expected, our implementation of **B-SVMs** gave significantly better results than Weinshall et al. [2012] thanks to the better image representation we adopted[1]. This scheme yields good novelty detection rates, as shown in Fig. 5.11(a), but has the main drawback of strongly relying on a threshold on the score values normalized with Eq. (5.1). In Weinshall et al. [2012], the authors fixed this threshold to 0.5. Here we used 0. This threshold directly controls the number of elements classified as *Known* or *Unknown*: with a value of the threshold near 0 almost all unknown objects are classified correctly but also a relatively high percentage of known objects is classified as unknown, while moving the threshold to 0.5 the effect is the opposite.

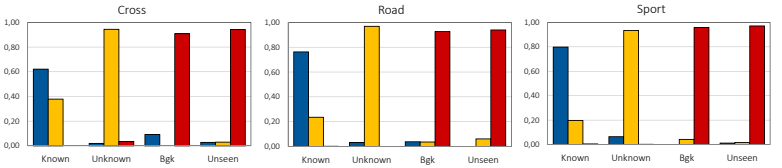
[1]This is evident from Tab. 5.1 and by comparing Fig. 5.11(a) with Fig. 3 of Weinshall et al. [2012].

Table 5.1: Correct detection rates for Known subcategories, Novel subcategories and Unseen classes. The first row (**B-SVMs***) shows results from Weinshall et al. [2012] and the remaining rows show our results on these datasets: Caltech256 - **Motorbikes**, Caltech256 - **Trans**portation, (iii) **SUN**397.

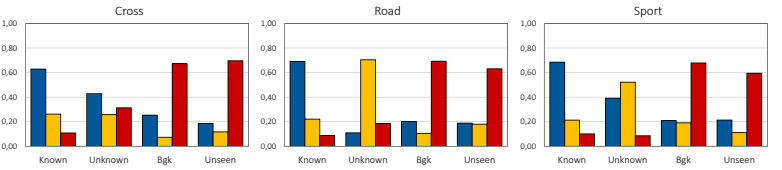
Data	Method	Known	Subcat.	Unseen
Motorbikes	***B-SVMs****	0.57	0.71	0.74
	B-SVMs	0.73	0.95	0.95
	OC-SVMs	0.67	0.49	0.64
	B/OC-SVMs	0.71	0.68	0.97
	FLAT	0.84	0.86	0.95
Trns.	***B-SVMs***	0.66	0.20	0.40
	B-SVMs/FLAT	0.67	0.39	0.62
SUN	***B-SVMs***	0.65	0.46	-
	B-SVMs/FLAT	0.68	0.57	-

Despite **OC-SVMs** being theoretically well suited for outlier detection, our experiments demonstrate their limitations in this context, where the data points lie in a high dimensional space. Each set still had the highest rate in its own category, but the overall performance is significantly lower than other approaches, as shown in Fig. 5.11(b). Fig. 5.11(c) shows the average classification rates for the hybrid scheme **B/OC-SVMs**. The performance in this case was worse than the **B-SVMs** configuration, but this method has the advantage of not requiring score normalisation.

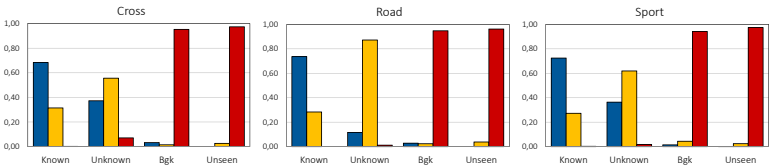
Finally Fig. 5.11(d) shows the results obtained with the **Flat** model. It can be observed that these results are similar to the ones obtained with the **B-SVMs** scheme. There is an improvement in the detection rate of the *known* subcategories, which is relevant when it is preferable to have a lower number of misclassified known objects. One disadvantage is that this scheme can not be exploited in larger hierarchies because novel objects can only be identified if they are rejected by all classifiers. It is therefore unable to detect subclasses belonging to different super-categories. The **B-SVMs/Flat** model has not been evaluated in this set of experiments because dealing with only one general category reduces this scheme to the initial **B-SVMs** approach.



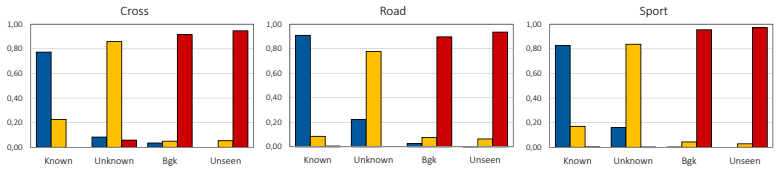
(a) B-SVMs configuration



(b) OC-SVMs configuration



(c) B/OC-SVMs configuration



(d) Flat configuration

Figure 5.11: Results leaving out one subcategory of motorbikes from the training set (Cross, Road and Sport, from left to right) for each classification scheme. The x-axis represents the ground truth subcategory type and the y-axis is detection rate. Blue, yellow and red bars correspond respectively to Known, Unknown, Background category type detection (see Sec. 5.2).

5.7.2 Evaluation on Caltech256 - Transportation

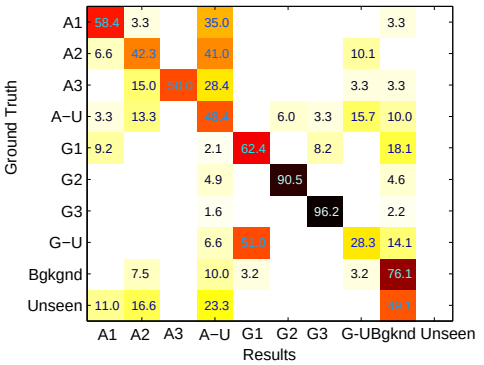
We explored the possibility of extending this framework to more complex hierarchies of images using the taxonomy **Caltech 256 - Transportation**. Based on the results discussed in the previous section, we decided to restrict these

experimets to using the **B-SVMs** scheme, which was the best performing hierarchical method in the previous experiment. We extended the evaluation to the **B-SVMs/Flat** model, to benefit from the Flat model at the specific level of classification and because the Flat method alone is unable to deal with multiple super-categories. Noting that the score normalisation of (5.1) makes the framework sensitive to the threshold, we decided not to use that normalisation in these experiments. In the previous settings with only two known subcategories, it was necessary to normalise the classifiers score, otherwise the two specific level SVMs would become the same classifier with swapped output signals, i.e., trained on opposite labels. When more than two subcategories are known, the normalisation of (5.1) becomes unnecessary (using the one-against-all setting) and does not produce any performance improvement.

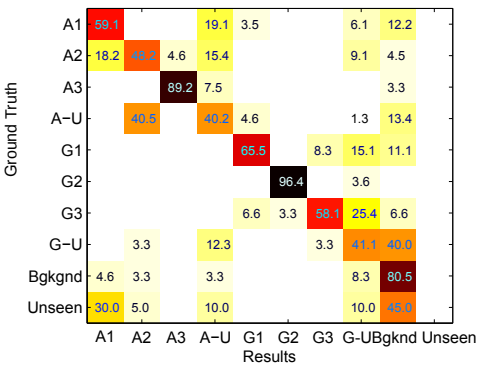
The best results are obtained with the **B-SVMs/Flat** model. The confusion matrices in Fig. 5.12 show that most of the known sub-categories were correctly classified. The roughly block-diagonal structure of the matrices show that the majority of the samples were classified to the correct super-category. Most of the unseen samples were correctly detected as background. On the other hand, most of the mistakes were either false *background* detection or false *unknown* subcategory detection. The true positive rates for *unknown* (novel) subcategory are substantially improved with respect to the **B-SVMs** scheme, where the obtained rates were disappointing, as most of those samples were either misclassified as other sub-categories or as background, the gain is quantified in Table 5.1 in nearly the 20% of corrected classified samples. This shows that the incongruence-based method of Weinshall et al. [2012] breaks down when the taxonomy of the concepts is not strictly related with the visual hierarchy (*i.e* the structure in the feature space), while the approach we proposed is stronger and yields better results.

5.7.3 Evaluation on SUN 397

We finally evaluated our framework on a taxonomy built over the **SUN397** dataset for scene recognition. Similarly to the previous experiments we re-



(a) Novel Subcategories: Airplanes - School Bus



(b) Novel Subcategories: Helicopters - Car Sides

Figure 5.12: Confusion matrices (in %) obtained on Caltech 256 with the **B-SVMs/Flat** scheme by removing these subcategories from the training set (a) *Airplanes* and *School Bus*, (b) *Helicopters* and *Car Sides*. 'A' and 'G' indicates *Air* and *Ground* transportation super-category, respectively. '-U' indicates the unknown subcategory. Note that 'Unseen' is not a label in the training set and unseen samples are expected to be classified as background.

stricted the evaluation to the **B-SVMs** and **B-SVMs/Flat** schemes.

Using the taxonomy described in Sec. 5.6.3 we iteratively sampled one of the four subcategories as *Unknown* and we used the other three to train the

classifiers. The average results are shown in Table 5.1. As already mentioned in Sec. 5.6.3 in this case we limited our aim to the detection of novel subcategories without focusing on *unseen* classes. The detection rate on *known* categories is similar to the one obtained for Caltech256 - Motorbikes taxonomy, while the *unknown* detection rate is significantly better than the previous one. Also in this case the average values demonstrate that the proposed framework, despite the satisfactory results, needs improvements when the visual hierarchy is not trivial to avoid the misclassification of the novel samples.

5.8 Further Experiments

We used the Oxford Flowers 17 dataset in order to deepen the evaluation of the hierarchical framework we proposed. In order to evaluate the performance on this dataset we described each image exploiting the descriptors that showed the best results in Nilsback and Zisserman [2008]. The feature vector is thus composed by a concatenation of the quantization histograms obtained with a Bag Of Words (BOW) on HSV colour values, SIFT (Scale Invariant Feature Transform) and HOG (Histogram of Oriented Gradients) descriptors.

We chose to extend the experiments on the flower taxonomy because of the good results obtained also with the basic B-SVMs configuration.

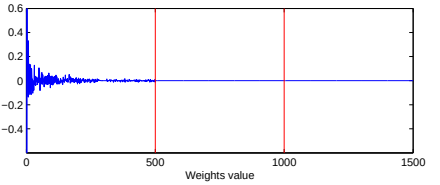
5.8.1 Mixed Norm SVMs

Inspired by human classification and deduction ability we assume that categories at different levels of the taxonomy might be recognized using different features, for instance a dog can be recognized for its four legs, but if the task is to distinguish among different breeds of dogs the colour, the size, the type of fur will be important. Recent features encoding methods tend to be always more descriptive and to embed both local and global information. Spatial Pyramid Representation has been widely adopted with different quantization techniques and has demonstrate impressive results in generic object recognition [Lazebnik et al. [2006], Perronnin and Dance [2007], Perronnin et al. [2010],

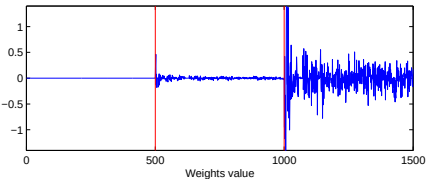
Sanchez et al. [2012]]. However the drawback of all these approaches is that, in order to capture fine differences between features, large size codebooks are always used, leading to high dimensional feature vectors and thus suffering from insufficient training examples. Recently Harada et al. [2011] proposed a solution to the large memory need introducing an appropriate partitioning and a weighted sum of semi-local features over the pyramid levels, where the weights are automatically selected to maximize the discriminative power. Following a similar principle and remembering that features might have different discriminative power at different level of the hierarchy, we would like to be able to give different weights to different features, without introducing any pre-processing step. At this aim, we used mixed norm SVMs with a L1-L2 regularization and we separate into three different groups the three histograms computed on HSV, SIFT and HOG features. As introduced in Sec. 2.3.3, the L1-L2 norm formulation of SVMs, regularize with an L2 norm each group of features and apply an L1 norm to the entire feature vector. Adding this regularization allows to achieve a sparse weight vector w. Recalling that each value w_i of w is proportional to the importance of the *i-th* feature in the feature vector for the classification, the regularization consequently allows to enforce the feature selection.

Similarly to the previous evaluations, we used a leave-one-class-out approach to simulate the novel class repeating the experiments several times and averaging the results. The experiments are performed using the *G-SVM* toolbox provided by Flamary et al. [2012]. We used the hierarchical configuration introduced in Sec. 5.3, and we replaced each classifier with a L1-L2 regularized SVM. The obtained results are reported in Figure 5.15, the comparison with the standard SVM formulation shows an improvement, especially in the detection of novel subclasses. Figures 5.15(a) and 5.15(b) illustrate respectively the confusion matrix obtained with SVMs and mixed norm SVMs.

In Figure 5.13 an exemplification of the weights vector obtained for the general (a) and specific (b) level classifier shows the difference with a standard L2 normalized SVM and highlights the sparsity of the vectors, pointing out how the general classifier hyperplane is based on color features, while the



(a) General classifier weights.



(b) Specific classifier weights.

Figure 5.13: L1-L2 Weights.

specific classifier uses HOG and SIFT values.

5.8.2 Reject Option

Kittler et al. [2014] defined a *reject option* as the lack of convincing support for any of the hypotheses associated with an application domain. In other words a reject option is flagged when the confidence of the classifier is below a certain threshold due for example to ambiguity in the input data. Such confidence can be measured in terms of a posteriori probability on the output. The aim of the reject option is to establish whether the response of the discriminative classifier should be considered reliable and accepted or not. This procedure is motivated by the fact that sometimes it may be preferable to avoid the classification and label the data as ambiguous rather than generate a wrong classification result. Geometrically this can be interpreted as a rejection of the data points that lie close to the decision boundary.

Standard Support Vector Machines are discriminative classifiers and does not provide probabilistic output, thus in order to obtain a posterior class prob-

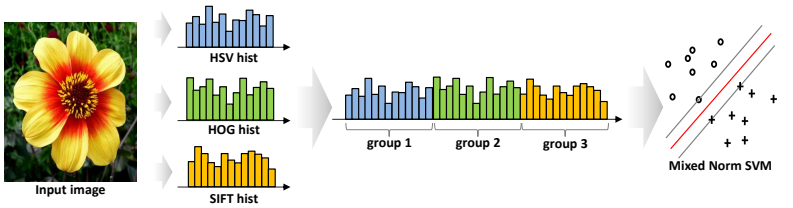


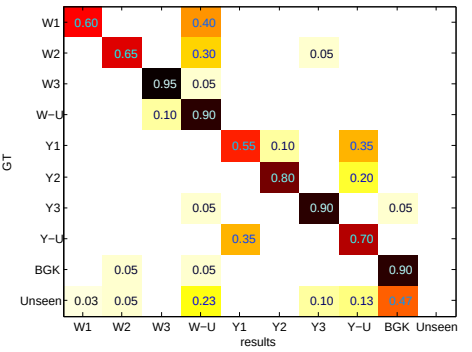
Figure 5.14: Mixed Norm SVMs applied on Oxford Flowers 17 dataset. The feature vectors are the concatenation of HSV, HOG and SIFT quantization histograms, each group of feature is L2 regularized, the complete feature vector is L1 normalized.

ability $P(y = 1 \mid x)$ is necessary to exploit the Platt's scaling Platt [1999]. Platt proposed to approximate the posterior by a sigmoid function:

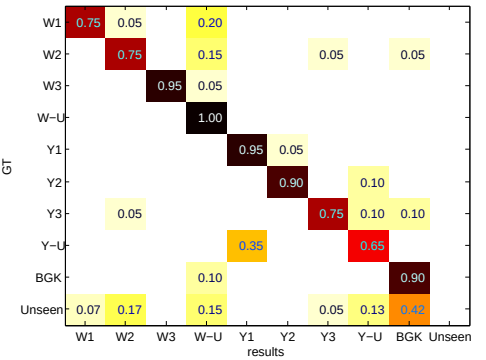
$$P(y = 1 \mid x) \approx P_{A,B}(f) \equiv \frac{1}{1 + \exp(Af + b)} \tag{5.5}$$

Where the set of parameters A and B are estimated using an EM framework.

Table 5.2 shows the different results obtained setting a varying reject threshold on the posterior class probabilities. Known categories are indicated with K, while unknown subcategories are indicated with U; *TPR* indicate the True Positive Rate and *FNR* the False Negative Rate (*i.e* the instances wrongly classified as belonging to a different subcategory). The aim of this evaluation was to find a threshold able to reject the wrongly classified objects, since we prefer not to classify ambiguous objects rather than classify them in the wrong category. However it can be observed that, despite the meaningful results obtained in Kittler et al. [2014], the improvements in our experiments are not considerable. The true positive rate of each category decreases as the threshold on the posterior class probability increase, but the false negative rate does not change. This probably means that the errors made by our classification scheme are not due to points that lie close to the classification boundary, but instead to true ambiguity in the input images, or, in the feature representation.



(a) SVMs



(b) Mixed Norm SVMs

Figure 5.15: Confusion matrices (in %) obtained on Oxford Flowers 17 by removing the subcategories *Windflower* and *Buttercup* from the training set. 'W' and 'Y' indicates *White* and *Yellow* flowers super-category, respectively. '-U' indicates the unknown subcategory. Note that 'Unseen' is not a label in the training set and unseen samples are expected to be classified as background.

Table 5.2: Detection of novel categories and subcategories. Reject option results on Oxford Flowers 17

Thresh	K-TPR	K-FNR	K-REJ	U-TPR	U-FNR	U-REJ
	79	21	0	85	15	0
0.20	77	21	1	75	15	10
0.25	76	21	2	75	12,5	10
0.30	75	20	6	75	12,5	10
0.35	74	20	6	72,5	12,5	12,5
0.40	73	20	8	62,5	12,5	25

Part III

Document Layout Analysis

CHAPTER

SIX

ILLUSTRATIONS SEGMENTATION IN DIGITIZED DOCUMENTS

Digitized documents play an important role in the preservation of historical contents and in their diffusion to the general public. Without digital editions the huge amount of old archives and documents would not be easily accessible. Digitization allows a pervasive diffusion and also the augmentation of masterpieces with multimedia details and appealing contents. Though, the huge amount of historical archives and books make desirable that their annotation and analysis were automatic and require the minimum user's intervention.

Pattern recognition and machine learning offer significant tools for automatically analysing the content of digitized documents and improving their presentation. If Optical Character Recognition (OCR) methods almost yield completely reliable results, the task of identifying textual regions and separate them from other components of the page is more challenging especially in old documents. The variety of exiting layouts and the heterogeneity of illustration and page elements is considerable as shown in Fig. 6.1. State of the art methods for Document Layout Analysis and segmentation have been proposed, and, among them, one of the most important is represented by Tesseract Smith [2007]. This OCR engine, sponsored by Google, not only offers a powerful

Figure 6.1: Examples of the variety of different layout existing.

tool for text recognition, but also provides layout analysis and image recognition modules. Despite the satisfactory results achieved on contemporary documents, the analysis of historical archives is still challenging due to the high variability of possible contents.

6.1 Background and Related work

Document layout analysis is an active area of research and a vast number of works have been presented in the literature. The focus of the problem is often the segmentation of text regions and the subsequent Optical Character Recognition (OCR) step of both printed and handwritten text, but approaches dealing also with pictures segmentation have been studied.

Chen and Blostein [2007] give a comprehensive survey on document image classification dividing it in three main components: the problem statement, the classifier architecture and the performance evaluation, separately analysing each component:

- The *problem statement* is related to the set and type of documents to be analysed.

- The *classifier architecture* is the core of the problem and includes the aspects related to page segmentation, feature representation and classification.

- *Performance evaluation* is also a critical and important component of a document classifier. The challenge is often the variety of documents considered, either fixed layout documents (books or forms) and flexible layout (newspapers) that inevitably produce different sets of classes as possible outputs.

The page segmentation problem can be further decomposed in *Geometrical Layout Analysis* and *Logical Layout Analysis*. The former step is solely based on the geometric characteristics of the document image and aims at finding homogeneous content portions, while the purpose of the latter is to segment the page image into a hierarchy of regions based on the human- perceptible meaning of the content. Regions are therefore assigned a logical label (*e.g* title, caption, paragraph) and a logical relation among the regions is determined (*e.g* reading order, inclusion in the same article).

The geometrical analysis approaches can be categorized into top-down, bottom- up or mixed segmentation approaches. Top-down methods, such as XY cuts [Ha et al. [1995], Cesarini et al. [2001]] or methods that exploits white streams [Appiani et al. [2001], Pavlidis and Zhou [1991]] or projection profiles [Esposito et al. [2000]] are usually fast but tend to fail when dealing with complex layouts. Bottom-up methods are instead more flexible and process the image page from the pixel level and subsequently aggregate into higher level regions but with an higher computational complexity. These approaches are usually based on mathematical morphology, Connected Components (CCs), Voronoi diagrams [Kise et al. [1998], Agrawal and Doermann [2009], Winder et al. [2011]] or run-length smearing [Sebastiani and Ricerche [2002]].

Many other methods exist which do not fit exactly into either of these categories: the so called mixed or hybrid approaches try to combine the high speed of the top-down approaches with the robustness of the bottom-up ones. Chen et al. [2013] proposes a method based on whitespace rectangles extraction and grouping: initially the foreground CCs are extracted and linked into chains according to their horizontal adjacency relationship; whitespace rectangles are then extracted from the gap between horizontally adjacent CCs; progressively

CCs and whitespaces are grouped and filtered to form text lines and afterward text blocks. Lazzara et al. [2011] provides a chain of steps to first recognize text regions and successively non-text elements. Foreground CCs are extracted, then delimiters (such as lines, whitespaces and tab-stop) are detected with object alignment and morphological algorithms. Since text components are usually well aligned, have a uniform size and are close to each other, the authors propose to regroup CCs by looking for their neighbors. Filters can also be applied on a group of CCs to validate the link between two CCs.

Once homogeneous region are extracted, the other important subtask is the classification of the regions into a set of logical predefined classes (*e.g* text blocks, tables, drawings, photos, etc.). The XY tree representation is a commonly used approach for describing the physical layout of the documents: it is used by Diligenti et al. [2003] with Hidden Tree Markov Models to perform classification, and by Appiani et al. [2001] and Baldi et al. [2003] with decision trees and a KNN based classifier, respectively. Feature vectors composed by a combination of different features are also common. Wang et al. [2006] propose fixed length feature vectors composed by a combinations of run length encoding, correlation of text lines, spatial and area features. Meng et al. [2007] suggest a combination of projection histograms and crossings number histograms. Hu et al. [1999] use interval encoding features to capture elements of spatial layout, modelled with HMMs. A more complex approach based on effective thresholding, morphology and connected component analysis has been used by Kitamoto et al. [2006], while Fataicha et al. [2002] proposes a multiscale approach.

Without expressing any hypothesis about the physical or logical structure of the analysed documents, a different approach is proposed by Journet et al. [2008] using texture features based on frequencies and orientations and a block based page analysis. With a similar assumption Nicolas et al. [2007] propose a 2D conditional random field model.

Finally the problem of the evaluation of the performance has been addressed by Clausner et al. [2011] introducing a system for accurate ground truthing and evaluation of large amounts of documents based on a flexible

Figure 6.2: Examples of digitized pages of the *Encyclopaedia Treccani*.

XML scheme and a set of automated and semi-automated tools to assist the user in the annotation process.

6.2 Problem Statement

The aim of our work is to correctly identify and extract the illustrations contained in digitized books. Documents in general, but in particular encyclopaedias and books, present a wide variety of illustrations, ranging from photos to drawings, sketches, charts and schemes. In such situations, the state-of-the-art methods, and especially the approaches based on morphological and pixel level operation encounter difficulties and tend to fail. Given the heterogeneity of the pictorial elements, the problem can be considered as one of incomplete data classification, where the classification has to deal with coarse labels: the sub-categorization of the different types of illustration is not given, and also, some categories might not appear in the training set. For instance, considering the case of the encyclopaedia, it may happen that a specific type of illustration can be found only in a section of the book. We propose to address the problem of pictures segmentation using a supervised classification approach and introducing a feature descriptor based on local correlation of regular blocks. The feature is improved with respect to previous approaches thanks to an effective encoding aimed at detecting the repeating patterns of text regions and differentiate them from pictorial elements. The purpose is to extract all the illustrations in order to allow subsequent steps for the construction of a digital library. The

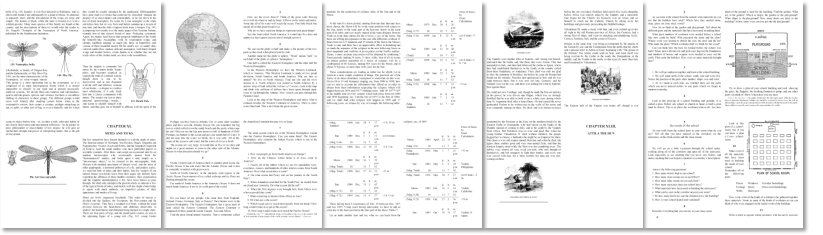


Figure 6.3: Examples of digitized pages of the *Gutenberg13* dataset.

images extracted can serve as a starting point for the enrichment of the digital editions with, for example, the retrieval of similar images.

The main dataset we used for our evaluation is shown in Fig.6.2 and is composed by digitized pages of the famous *Italian Encyclopaedia Treccani*. Additionally we tested our proposal on the *Gutenberg13 dataset*, a collection of digitized books extracted from the Gutenberg project database (see Fig. 6.3). In the rest of the chapter we detail our method and we showed how our proposal is able to deal with the problem of incomplete training data and to learn an effective model of textual and pictorial elements.

6.3 Page Layout Segmentation

We approach the page segmentation problem starting from the idea that textual and pictorial regions in a document are characterized by different local patterns: lines of text exhibit a regular structure which can be exploited to successfully differentiate them from illustrations in a classification framework. The method we propose can be decomposed in the steps depicted in Fig. 6.4. The geometric layout analysis is performed by extracting the main regions from the page using the XY cut segmentation, then each region is divided in small squared blocks of size n, and local correlation features are computed on each block and classified using a Support Vector Machine.

The XY cut is a well known recursive algorithm for top-down page seg-

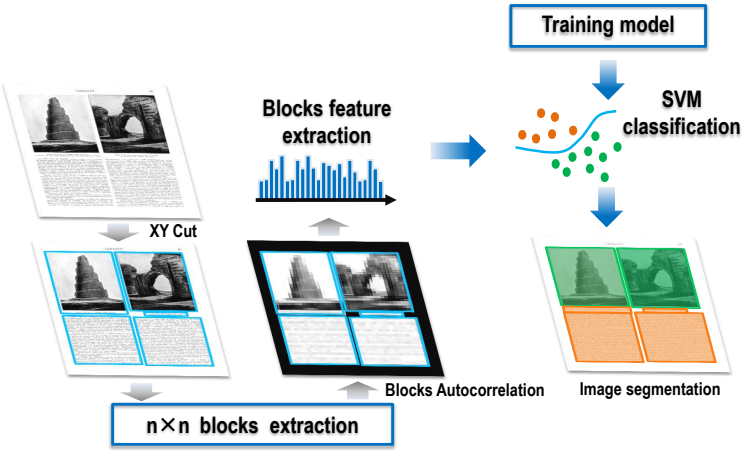


Figure 6.4: System overview

mentation. The method works by firstly projecting the pixels' values on the vertical and horizontal axis of the image and subsequently by finding a low density region of the projection histograms, *i.e* by finding the white spaces of the page. In this way the page is recursively segmented in rectangular regions.

We used the recursive XY cut with a preprocessing phase of binarization and morphological closure on the page in order to filter out the white interline spaces. The closure is performed with a squared structuring element of size 41×41 pixels. At each iteration the white borders surrounding the regions are removed before calling the next recursive step to find the internal cuts. Algorithm 3 provides the pseudo code of our approach to recursive XY cuts. By exploiting this algorithm, we obtain a segmentation of the page, usually corresponding to the two columns of text or parts of them, and, if existing, full page images.

Algorithm 3 Recursive XY Cut (RXYC)

Input: *image*
Output: *list* ▷ list of regions on the page

```
Binarize(image)
MorphologicalClosure(image)
RXYC_Step(image)

procedure RXYC_STEP(ROI)
    Remove white borders
    vProj ← vertical projection (sum of rows values)
    hProj ← horizontal projection (sum of columns values)
    hCut ← first y such as vProj(y) < Thresh
    if hCut then                                   ▷ Horizontal cut found
        ROI ← ROI
        ROI.h ← hCut
        RXYC_Step(ROI)                  ▷ XYCut on the first sub-image
        ROI.y ← ROI.y + hCut
        ROI.h ← ROI.h − hCut
        RXYC_Step(ROI)                  ▷ XYCut on the second sub-image
    else
        vCut ← first x such as hProj(x) < Thresh
        if vCut then                               ▷ Vertical cut found
            ROI ← ROI
            ROI.w ← vCut
            RXYC_Step(ROI)              ▷ XYCut on the first sub-image
            ROI.x ← ROI.x + vCut
            ROI.w ← ROI.h − vCut
            RXYC_Step(ROI)              ▷ XYCut on the second sub-image
        else
            list ← list ∪ ROI
        end if
    end if
end procedure
```

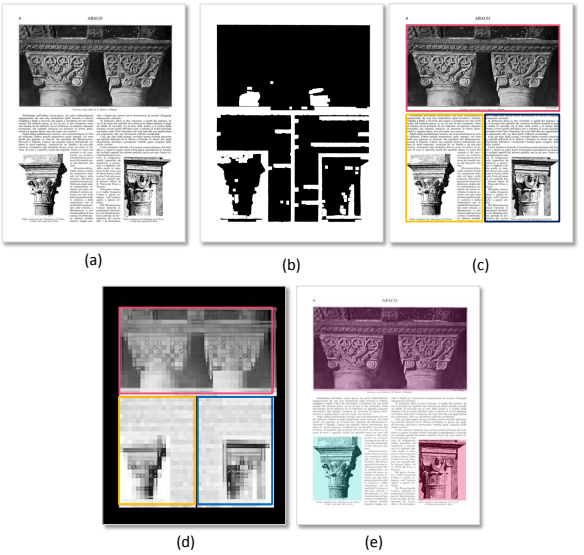


Figure 6.5: Sequence of step: (a) Input image, (b) Closure, (c) XY-Cut, (d) Local correlation features, (e) Final illustration segmentation.

6.4 Local Correlation Features

In order to distinguish between textual and pictorial regions we used a texture analysis method similar to the one proposed in Journet et al. [2008] and Grana et al. [2010]. The approach exploits the autocorrelation matrix, an effective feature for finding repeating patterns which is particularly suited in this case since textual textures have a pronounced orientation that heavily differs from that of illustrations. The autocorrelation function is defined as the cross correlation of a signal with itself, and represents a measure of similarity between two signals. Once applied to a grayscale image, it produces a central symmetric matrix, that gives an idea of the degree of regularity of the texture. The methods consists in the subdivision of the original image into square blocks of

size n. The formal definition of the autocorrelation of a block is:

$$C(k,l) = \sum_{y=\max(0,l)}^{n-1+\min(0,l)} \sum_{x=\max(0,k)}^{n-1+\min(0,k)} I(x,y) \cdot I(x+k,y+l) \tag{6.1}$$

where l and k are defined in $[-n/2, n/2]$. It is important that the size n of the blocks is chosen such that the repeating pattern of the textual blocks is highlighted.

Implementing the autocorrelation following this definition gives an algorithm with a high computational complexity, roughly proportional to n^4. According to the cross-correlation theorem the cross-correlation of two signals is equal to the product of the Fourier Transform of each one, where one of them has been complex conjugated.

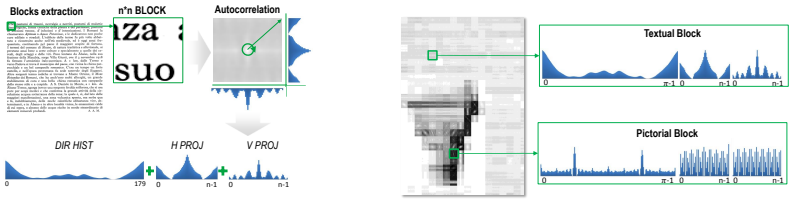
$$f \star g \Leftrightarrow F' \cdot G \tag{6.2}$$

where F and G denote the transformed signals and F' is the complex conjugate of F. Since the autocorrelation is a special case of the cross- correlation, Eq. 6.2 becomes:

$$f \star f \Leftrightarrow F' \cdot F = |F|^2 \tag{6.3}$$

and for the Wiener-Khinchin theorem, the Fourier Transform of an autocorrelation function is the power spectrum, or equivalently, the autocorrelation is the inverse Fourier transform of the power spectrum. Following these properties, we efficiently compute the autocorrelation of the blocks only using two steps of the Fast Fourier Transform (FFT) with to a reduction of the complexity from $O(N^4)$ of the naive approach to $O(N \log N)$.

The result of the autocorrelation can be employed to extract an estimate of the relevant directions within the texture. Usually, the autocorrelation matrix is encoded with a *directional histogram*, a polar representation in which each direction is determined by an angle $[0°, 360°)$ and the bin value is given by the



(a) Descriptors computation. (b) Feature vectors.

Figure 6.6: (a) feature vectors computation from an image block. (b) image showing the autocorrelation on every block of a page and example feature vectors obtained from a text area and from an illustration.

sum of the pixels along that direction.

$$w(\theta) = \sum_{r \in (0, bs/2]} C(r\cos\theta, r\sin\theta) \tag{6.4}$$

Since the autocorrelation matrix has a central symmetry by definition, we consider only the first half of the direction histogram in the range $[0°, 180°)$. ω and r are quantized: the step of ω is set to $1°$ and the step of r is set to 1 pixel. Using this encoding a text block is characterized by peaks around $0°$ and $180°$ because of the horizontal dominant direction, conversely an image block is described by a generic multi-modal distribution.

We finally enrich the descriptor concatenating the directional histogram with the projections of the autocorrelation matrix along the vertical and horizontal directions in order to enhance the repeating pattern of the text lines. Fig. 6.6 provides a visual summary of the feature extraction process and few example results.

6.5 Datasets

In this section we report the results obtained on the digitized pages of the "Enciclopedia Treccani"[1]. The encyclopedia consists of a set of volumes firstly

[1] http://www.treccani.it

published between 1925 and 1936. In our evaluation we only considered the first volume which is composed of 1183 pages. The original size of each digitized page is 22110×28819 pixels at 1 bpp, and we used a downscaled version with a factor of 0.125 along each dimension and converted to gray -scale so that the new images are 2763×3602 pixels at 8 bpp. The pages are two-column text with a number of illustration per page that may vary from 0 to 10. The main difficulty of this dataset, compared to other document analysis datasets, is the variety of graphical elements that is not limited to photos, but also includes drawings and charts. The overall number of manually annotated illustrations is 1157.

For a more comprehensive analysis we also built the Gutenberg13 dataset. This dataset, available online at `http://www.imagelab.unimore.it`, has been created using a set of publicly available e-books from Project Gutenberg[2]. The e-books have been converted to grayscale with a resolution of 300 pixel/inch printing 4 pages for sheet, resulting in a two column text layout with a font size similar to the Treccani dataset. The total number of pages is 268 and each page is 2481×3509 pixels. Some example images of the two datasets are provided in Fig. 6.7 and in Fig. 6.8.

6.6 Experimental results

In our experiments we set the block size to 64×64 pixels, enough to consider a line of text. Recalling that our descriptor is the concatenation of the vertical and horizontal projection of the autocorrelation matrix and its directional histogram, we obtained a 308 dimensional feature vector for each block. We used an SVM with RBF kernel whose C and γ parameters have been estimated using cross validation (the values $C = 4096$ and $\gamma = 0.5$ have been used). The training set consists of 4000 blocks randomly sampled half from text regions and half from image regions.

We compared our proposal with the results given by the layout analysis

[2] `http://www.gutenberg.org`

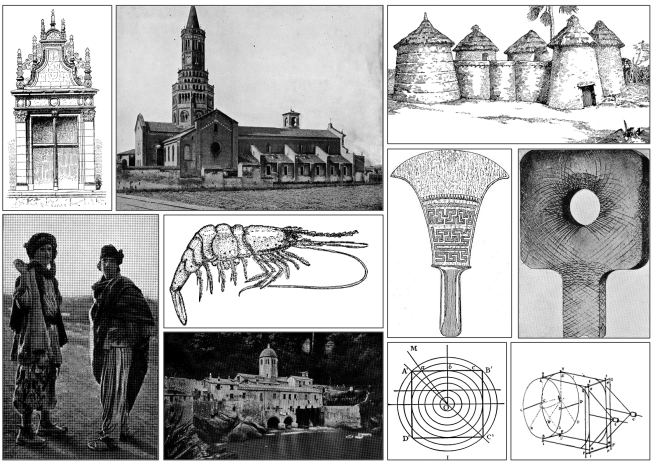


Figure 6.7: Sample images from the Treccani dataset

module of Tesseract [3], the Optical Character Recognition engine sponsored by Google. Since the image-find function of Tesseract is essentially based on morphological operations, in order to maximize Tesseract performance we threshold the pages to the maximum level of 255, meaning that everything that is not white is considered as black.

In order to evaluate the performance we solved the matching problem between the ground truth and the results obtained by the two methods exploiting the *Hungarian Method*. The nodes of the bipartite graph correspond to the GT bounding boxes and to the automatically extracted bounding boxes, while the edges are weighted using a measure of the overlap between bounding boxes. Given two bounding boxes l and r, the weight is computed as:

$$w_{lr} = \frac{l \cap r}{l \cup r} \tag{6.5}$$

Results are reported in terms of the number of True Positives (TP), False Negatives (FN) and False Positives (FP). Table 6.1 shows the performance compar-

[3] https://code.google.com/p/tesseract-ocr

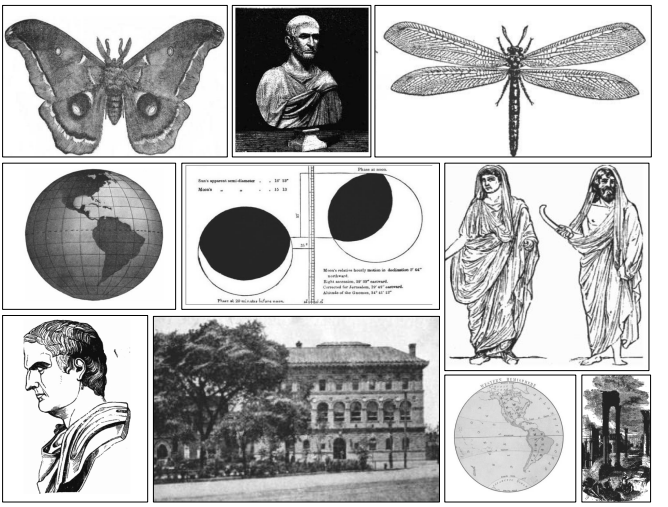


Figure 6.8: Sample images from the Gutenberg13 dataset

ison of our method and Tesseract on the Treccani and Gutenberg13 datasets, reporting the number of images of the three classes (TP, FN, FP). Table 6.2, instead, focuses on the performance in terms of percentage of pixels with respect to the total number of pixels annotated as illustration.

As Table 6.1 shows, our method outperforms Tesseract on the Treccani dataset. Tesseract has indeed the main drawback of not being able to recognize most of the drawings although it makes a good job in finding photographs. Our proposal, instead, exploiting a supervised classification approach, has the capability of learning the correct classification also of drawings and charts. Table 6.1 also demonstrates how our method produces a higher number of false positives, but we would like to highlight that despite the high quantity, false positive results correspond to small areas as shown in Table 6.2. These errors usually correspond to portions of music sheets, tables or drawings as displayed in Fig. 6.9. Some examples of illustration segmentations obtained with Tesseract, that highlight the difficulty in extracting drawings and charts, are reported in Fig. 6.10. In order to further evaluate the reliability of our proposal and the

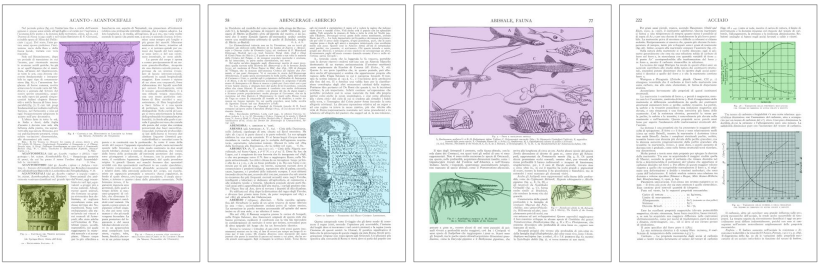
Table 6.1: Comparison between our method based on local correlation, the same method with cross testing and Tesseract on the Treccani and Gutenberg13 datasets. Results are reported in terms of number of TP, FN, FP. The values are the number of images of the three classes.

Dataset	Method	TP	FN	FP
Treccani	***Our Method***	361	5	132
	Tesseract	200	166	16
Gutenberg13	***Our Method***	524	11	62
	Our Method XT	461	20	64
	Tesseract	486	40	9

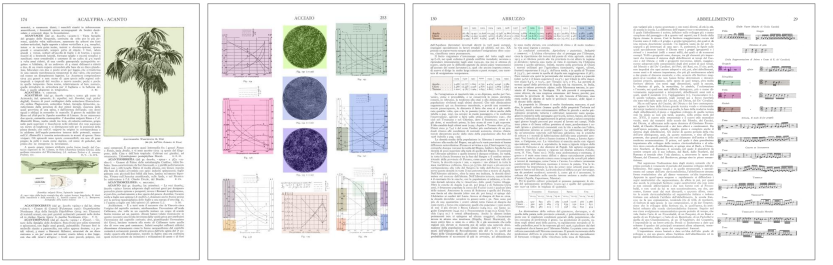
Table 6.2: Comparison between our method based on local correlation, the same method with cross testing and Tesseract on the Treccani and Gutenberg13 datasets. Results are reported in terms of % of TP, FN, FP of illustration pixels.

Dataset	Method	TP pxls (%)	FN pxls (%)	FP pxls (%)
Treccani	***CorrSegm***	99.57	0.43	4.53
	Tesseract	52.28	47.72	0.39
Gutenberg13	***Our Method***	99.23	2.39	0.77
	Our Method XT	91.30	2.66	8.70
	Tesseract	83.13	16.87	1.02

robustness of the model learnt by the Support Vector Machine, we performed a cross-testing. Specifically we used the model learnt on the Treccani dataset to test the segmentation on the Gutenberg dataset. The results, reported in Tables 6.1 and 6.2, show how our proposal is generic and applies to different types of documents provided the similar structure, and that does not over fit on the specific training data. Despite the better performance achieved when training and testing on the same dataset (more than 99% of true positive images segmented), Table 6.2 reports how, even without a re-training phase, more than the 90% of the pictorial elements are correctly identified.

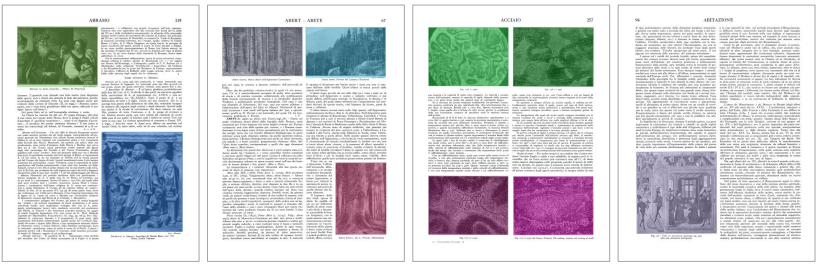


(a) Correct segmentation results.

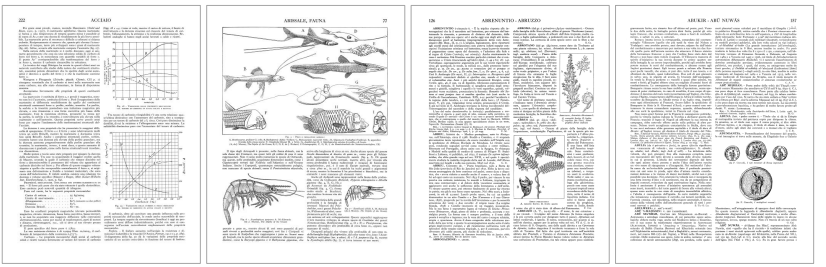


(b) Wrong segmentation results.

Figure 6.9: Illustration segmentation obtained with our method Treccani Dataset. (a) Photographs, draws and charts correctly segmented. (b) Examples of oversegmented regions and wrong detections.



(a) Correct segmentation results.



(b) Wrong segmentation results.

Figure 6.10: Illustration segmentation obtained with Tesseract Layout Analysis module on the Treccani Dataset. (a) Examples of photographs correctly segmented. (b) Charts and drawings not detected as illustrations.

CHAPTER

SEVEN

CONCLUSION

7.1 Summary

In this thesis we addressed the problem of learning with incomplete or partial knowledge. As humans, we are able to learn from small sets of data and to identify new object categories even without prior knowledge. Our learning abilities allows us to rapidly generalize different kind of regularities from both labelled and unlabelled data with a continuous learning process. Conversely automated learning systems usually need large sets of annotated data. We tried to fill the gap between human and machine learning using ad-hoc ML and pr techniques.

Our motivations originates from the assumption that machine learning methods should follow and imitate human learning abilities when dealing with small set or incomplete training data.

We studied the topic in three different computer vision applications instead of reasoning at a higher level of abstraction, assuming that a specific solution should be found for every problem.

In Part I we considered the problem of people tracking in surveillance scenarios and we re-interpreted the problem of identifying a given target among

subsequent video frames as a semi supervised learning problem. Two models of labelled samples have been considered for single target tracking, they were respectively constituted by patches of the target person and by patches of people that differ from the target. Unlabelled patches have been exploited together with the labelled models in a graph transduction framework to iteratively and on-line propagate labels to the unlabelled patches extracted from the video sequence. Moreover a model update strategy based on evolutionary spectral clustering has been used to update labelled models and avoid drifting in the target process. In Chapter 4 a different formulation of graph transduction based on game theory has been used to extend the tracking to a multi-target scenario. Different targets to track have been considered as players of a game: labelled samples have been interpreted as players that already knew their pure strategies, while unlabelled samples have been interpreted as players that needed to find their strategies by maximising pay-offs according to the similarities with labelled samples. Missing labels have been estimated by finding the Nash equilibria over the set of possible strategies. Both methods have been evaluated on different video surveillance sequences and obtained satisfactory results compared to conventional state-of-the-art trackers. Specifically, single target tracking, thanks to thanks to the positive and negative labelled samples and the update strategy, outperforms similar methods based on semi supervised learning. We also demonstrated how SSL can effectively exploit the structure of the manifold modelled by unlabelled data to help the association of people patches even with challenging situations.

In Part II we explored methods to detect novel categories and subcategories in hierarchies of images. Dealing with image categorization, is infeasible to collect annotated data for all existing categories of objects, for this reason being able to detect novel classes is an important task for visual classification systems. Inspired by human reasoning and learning, we considered object categories as organized in a taxonomy, and we analysed a hierarchical novelty detection framework which leverages the incongruence between classifiers at different levels of the hierarchy. We evaluated this framework on three datasets for object and scene classification, and using different classification schemes

based on binary and One-Class SVMs in different parts of the hierarchy. Despite the theoretical foundation of OC-SVMs for novelty detection, configurations based on standard binary SVMs achieved better results, and especially we demonstrated how the best performance are achieved with a configuration that exploits all known subcategories in a *1-vs-all* configuration at the lower level of the hierarchy. The pure hierarchical method, instead, achieves satisfactory novelty detection rates for small taxonomies and when the semantic hierarchy matches the appearance hierarchy of image classes, but breaks down when these assumptions are not satisfied. We also evaluated and analysed an approach based on Mixed Norm SVMs, preliminary obtained results obtained have been promising and demonstrate the ability to enforce the selection of differently discriminative features at different level of the hierarchy.

Finally in Part III (Chapter 6) we proposed a method to automatic extract illustrations from digitized historical documents. We considered the case where only two labels were given: *text-illustration*, but the heterogeneity of illustrations contained in the documents and the lack of all the different types of illustration in the training set yield a challenging classification. Starting from the assumption that text regions are characterized by a strong horizontal repeating pattern we introduced a descriptor based on the autocorrelation of squared blocks that has demonstrate to be effective even with drawings and charts. The feature representation method proposed allows to learn a robust classification model able to distinguish textual regions from the variety of pictorial regions. We compared our method with Tesseract and we showed how it is able to detect challenging illustration also when the state-of-the-art fails.

7.2 Open Issues

Solving a problem always generates a large number of unsolved and interesting new research topics. A further property of research, which occurs especially in computer vision and machine learning, is that a problem is never completely solved but only up to a certain accuracy. Due to this reason, there is always space for improvement concerning the methods presented in this thesis. In the

following, we give some ideas for future research directions.

Considering the problem of single target tracking in Chapter 3 we assumed the people patches given and extracted with a HOG based people detector. A possible extension is the evaluation of a detection-free system, where unlabelled patches are extracted from video frames with a random or Gaussian sampling. This extension would also allow to test the proposal on different categories of targets.

Regarding instead the multitarget configuration and given the promising results obtained, future work includes the extension to longer video sequence in order to test the robustness of the system in long term tracking and in occlusion handling. Additionally, we would like to evaluate an on-line (iterative) approach similar to the one applied to single target tracking and to introduce a mechanism to handle the insertion of new targets (equivalently to handle a variable number of target classes).

Transfer learning techniques (as introduced in Section 2.2) can be exploited to improve the classification of novel categories in Chapter 5. A common representation of categories that share the same parent node might be useful to generalize unseen concepts sibling of known classes. Once detected, new classes, can be added to the set of known categories in a life-long learning framework. Moreover, also attributes and zero-shot learning might represent a feasible alternative. Once the learning and detection steps are consolidate we would like to extend the evaluation to large scale problems, such as the ImageNet dataset.

Lastly, regarding the problem addressed in the segmentation of illustration in digitized historical books, a possible improvement is to analyse the use of SSL instead of conventional SVMs and also to extend the categorization to specialized classes of illustrations. A multi-class segmentation might be considered to differentiate among different type of illustrations instead of limiting the analysis to coarse labels. The system finally represents an useful tool for a future enrichment of historical manuscripts with renovated contents. Starting from the appearance of the extracted images, and eventually exploiting the keywords contained in their captions, is in fact possible to automatically retrieve similar images, for example from the web.

APPENDIX

A

RELEVANCE FEEDBACK: THE USER IN THE LOOP

The use of relevance feedback strategies in information retrieval and in particular in content-based image retrieval systems is widely considered a very precious (sometimes necessary) addition to the system itself. Automatically retrieved instances might be really heterogeneous, a typical example is given by any Google image search of a term with several meaning, (*e.g* Fig. A.1).

Relevance feedback allows the system to somehow bypass the semantic gap that solely may limit the effectiveness of a query by similarity, given the heterogeneity of the visual appearance of some concept prototypes. Actually it is the most effective way to capture user's need and, more generally, user's search intention. The reason is pretty straightforward: the automatic association of low-level features to high-level semantics is still a very open problem, and the only practical way to identify what the user is looking for is by including him in the retrieval loop, with the input of feedbacks (positive, negative or both). The common scenario in which relevance feedback is used within content-based image retrieval systems is the following:

1. An initial query-by-keyword or query-by-example is performed, in the

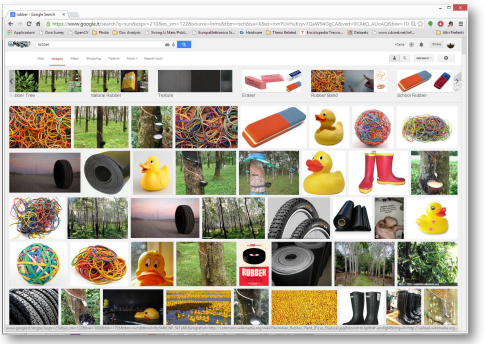


Figure A.1: Google image search of a term with plural meanings.

form of a list of results ranked with increasing distances from the query in the feature space;

2. The user provide some good (and bad, implicitly or explicitly) feedbacks given the displayed images, choosing in other words what is relevant and what is irrelevant;

3. An algorithm uses these information to change the displayed results in a ”refinement” step to accommodate user’s judgements;

4. Back to step 2 and loop until a certain condition (or satisfaction) is reached.

Here we focus on the third step, and, in particular we demonstrate how the Transductive Learning (TL) algorithm introduced in Chapter 3 is an effective tool to account for users’ feedbacks. Particularly we show how this approach can improve the retrieved results in an application based on surveillance data Coppi et al. [2013].

A.1 Relevance feedback: an overview

The idea behind Relavance Feedback (RF) is to take into consideration the users' browsing behaviour in image retrieval and overcome the difficulties of Content Based Image Retrieval (CBIR). Despite the advances in image representation method, there is still a semantic gap between the low level features and the high level concepts, that RF techniques try to bridge. Starting from the seminal RF approach for CBIR introduced in the early and mid '90s by Kurita and Kato [1993], Picard et al. [1996], Rui et al. [1998] the literature on this topic is countless. Representative surveys are given by Zhou and Huang [2003], Crucianu et al. [2004], Sivakamasundari and Seenivasagam [2012]. Moreover, aside the research on the algorithm for relevance feedback, there is a wide literature about the way in which the performance of a system with relevance feedback can be safely evaluated in order to provide fair comparison with different techniques.

A solution for enhancing the accuracy of image retrieval is moving the query point toward the contour of the user's preference in the feature space, this approach is called Query Point Movement (QPM). The query point is moved in order to create a more complete query. Works based on QPM are proposed by Liu and Chang [2009], Ishikawa et al. [1998], Porkaew et al. [1999].

Tackling the problem from a different perspective, a further category of approaches applies some machine learning procedures (like SVM, Boosting or Gaussian Mixture Models) to learn how to separate relevant samples from irrelevant ones Tao et al. [2006], Tieu and Viola [2000], Cox et al. [1998]. A final category of approaches, finally, followed an idea similar to the one we proposed based on Semi Supervised learning, (see Wu et al. [2000], Sahbi et al. [2008, 2007], Radosavljevic et al. [2008]). The idea is to take advantage both of the unlabeled and labeled samples in a transductive inference manner, learning from an incremental amount of training samples (feedbacks, in this case). In particular Borghesani et al. [2011] successfully exploited graph transduction together with covariance representation for historical images retrieval.

Concerning instead the evaluation metric, Luo and Nascimento [2004] in-

troduced a complete set of measures and distinguished among *actual*, *new*, *cumulative* precision and recall allowing the evaluation of both the learning aptitude of the system (*i.e* how fast relevant images are retrieved) and the history-based completeness of the system (*i.e* the cumulative performance).

A.2 Transductive Relevance Feedback

The relevance feedback problem can be analysed as a semi-supervised learning problem, in which the positive and the negative feedbacks given by the users constitute iteratively (and incrementally) the labelled training instances of the algorithm.

As proposed in Sec. 3.3 we tackle the problem with a graph-based transductive learning method. Figure A.2 gives a glimpse of the approach: the user initially queries an image, a set of results is retrieved and presented to the user that can give his feedbacks selecting relevant and ton relevant (*i.e* positive and negative) samples. Using the feedbacks as labelled elements the graph transduction re-ranks the retrieved images potentially moving away from the query center dissimilar images and attracting toward the query center similar images. This process might be repeated iteratively to further improve the results according to the user's suggestion.

The undirected graph $G = (V, E)$ is defined representing as vertices V the labelled and unlabelled images of the dataset, and, as edges the distances between them.

A.3 An application on surveillance data

Video surveillance data are often unreliable due to many factors, such as poor color resolutions, low frame-rate, occluded viewpoints, bad luminance. Most of the applications devoted to digital surveillance forensics allow to retrieve data similar to a requested instances based on some similarity measure, but, so far, the user involvement in the search process has been limited to execute

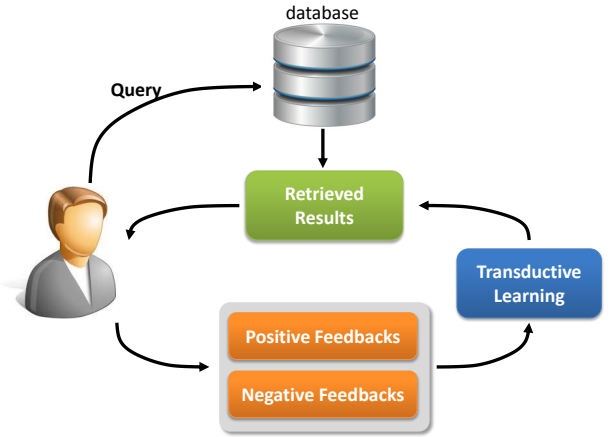


Figure A.2: Overview of the iteractive query system with relevance feedback and transductive learning.

the initial query. However, working with these uncertain data the experience of investigators becomes essential and the relevance feedback constitute an useful tool to transfer the human knowledge to automatic systems.

When processing video surveillance data several elements are interesting and can be automatically acquired by modern video surveillance systems. Among these people trajectories and people appearances constitute a proper choice that carry important information about people behaviour in the scene. Investigators may want to find the occurrences of a suspect person in a video stream or set of video or eventually searching for all the people that follow a certain path. We tested the effectiveness of the graph transduction as a relevance feedback tool using snapshots of people collected from the *Caviar* dataset [1] already used in Sec. 3.7 and we also decided to add to the evaluation people trajectories acquired from Edinburgh Informatics Forum Pedestrian Database [2]. Snapshots of the application are given in Fig. A.3, where green bounded elements represent positive feedbacks and red bounded elements rep-

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1
[2] http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING

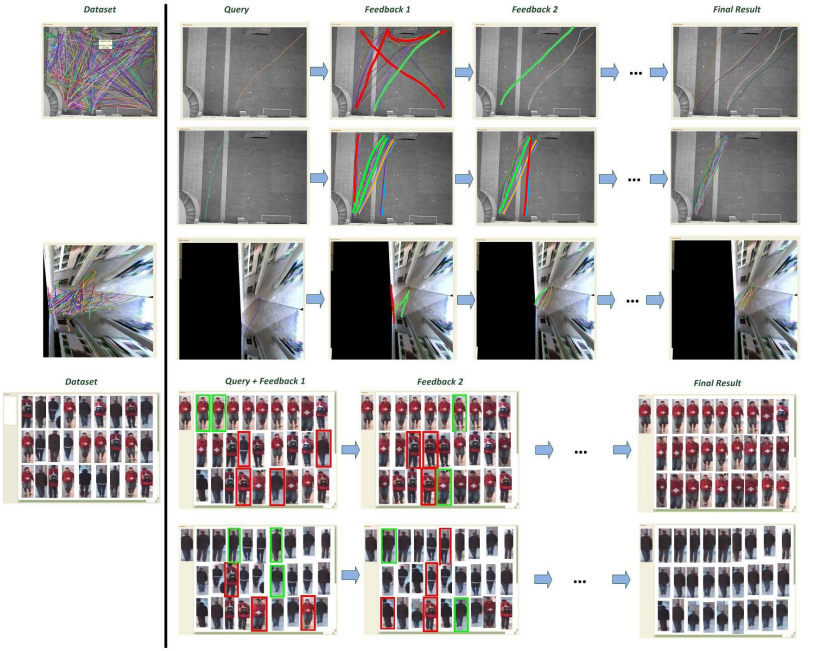


Figure A.3: Queries performed on our active query system for video surveillance forensic analysis.

resent the negative ones. Final results are shown on the rightmost sides of each rows.

People Snapshots People patches have been extracted from the video frames using a conventional HOG people detector Dalal and Triggs [2005], the dataset contains nearly 2000 snapshots manually annotated. Similarly to Sec.3.4 images are represented using covariance matrix, but here we used a slightly simplified version with a 7-dimensional feature vector based on normalized pixels locations $(x/W, y/H)$, RGB values in the range $[0,1]$ and the norm of the first derivatives of the intensity with respect to x and y, calculated through the Sobel filter. The covariance matrix of a region thus results a 7×7 matrix, and the distance between different matrices is computed using the metric proposed in

Forstner et al. [1999] as the sum of the squared logarithms of the generalized eigenvalues.

Performance have been evaluated with a user centric perspective, thus considering a number of iterations considered as not redundant and boring for the user. In these experiments we fixed the number to $T = 5$ iterations and we evaluated the performances in terms of cumulative recall at each step, where *cumulative* recall means the recall computed at each iteration and relative to the whole set iterations so far. Finally we tried to concentrate the analysis on *feasible search task*, *i.e* visual topics with a good number of representatives with a low degree of uncertainty in the evaluation, in order to assure a valuable reference ground truth.

We report a comparison between two alternative of our method using only positive feedbacks and positive and negative feedbacks, and some other algorithms for relevance feedback originally introduced for CBIR. Results are reported in Table A.1 The algorithms are the following:

- Baseline kNN classifier: no multiple iterations of relevance feedback, the system proposes the firs k results according retrieved by the classifier;
- Naive relevance feedback (actually no relevance feedback at all): the system discards the current set of k results and proposes to the user the next k, following the original rank given by the visual similarity;
- Mean Shift Feature Space Warping, MSFSW, proposed by Chang et al. [2009], where the feature space or the metric space are manipulated, in order to shape it in the direction of the users' feedbacks;
- Transductive Learning with Positive feedbacks, TLP: our transductive learning approach which uses positive feedbacks as labelled samples. The affinity matrix is filled only for the $k = 30$ nearest neighbors;
- Transductive Learning with Positive and Negative feedbacks, TLPN: our transductive learning approach with both positive and negative feedbacks. The affinity matrix is filled only for the $k = 30$ nearest neighbors.

Table A.1: Relevance feedback. Recall values at different iteration steps.

method	1	2	3	4	5
Naive	68,4	72,1	75,2	78,1	81,7
MSFSW	69,2	75,3	76,9	79,7	82,4
TLP	70,7	79,0	83,7	88,3	91,3
TLPN	72,7	80,2	86,2	91,1	95,0

The comparison shows how the approach we proposed reaches an higher value of recall within the same number of iterations.

People Trajectories To further the evaluation we propose to use also people trajectories, that can be compared either on their position in the scene, **spatial analysis**, or their shape, **shape analysis**, at this aim Coppi et al. [2013] exploited two different similarity measures based on trajectory points coordinates and trajectory shape.

The former analysis is based on a spatial model proposed by Calderara et al. [2009], that combines a statistical representation of the data with a point-to-point approach, where each point in the trajectory is modelled as a bi-variate Gaussian. Briefly, given the k^{th} rectified trajectory (i.e. with perspective distorsion corrected by a projection on the ground plane) projected on the ground plane $T_k = \{t_{1,k}, \ldots, t_{n_k,k}\}$, where $t_{i,k} = (x_{i,k}, y_{i,k})$ with n_k is the number of points of trajectory T_k, a bivariate Gaussian $S_i^k = \mathcal{N}(x, y \mid \mu_{i,k}, \Sigma)$ is centred on each data point $t_{i,k}$ (*i.e* having the mean equal to the point coordinates $\mu_{i,k} = (x_{i,k}, y_{i,k})$) and with fixed covariance matrix Σ). After assigning a Gaussian to every trajectory point, the trajectory can be modelled as a sequence of symbols corresponding to Gaussian distributions. The latter method, focused on trajectories shape, discard the paths position and exploits instead their shape modelled as a sequence of angles. This statistical representation has been introduced by Calderara et al. [2011], and exploits circular statistics, precisely a mixture of Von Mises distribution is used as pdf to describe a trajectory T_j by

means of its angles:

$$p(\theta) = \sum_{k=1}^{K} \pi_k \mathcal{V}(\theta|\theta_{0,k}, m_k) \tag{A.1}$$

where $\mathcal{V}(\theta|\theta_{0,k}, m_k) = \frac{1}{2\pi I_0(m)} e^{m\cos(\theta-\theta_0)}$ and with I_0 the modified Bessel function of order 0.

Finally, in order to compute the distances among trajectories an alignment algorithm is used and then distances are computed using the Bhattacharyya coefficient between respectively normal distributions and angular distributions.

The dataset for quantitative accuracy evaluation consists of 3000 trajectories.

Charts in Figures A.4 and A.5 report the cumulative precision and recall obtained respectively with shape and point coordinate analysis averaged on several executions. Light grey bars refer to transduction with only positive feedbacks while dark grey bars refer to transduction with both positive and negative feedbacks. The boost on performances obtained through relevance feedback (bars portions over the dashed lines) is evident and demonstrates the capability of the system to obtain satisfying results even when simple and fast similarity measures are employed to compare the elements. It is remarkable to note that the final average precision and recall are closer, in most of the considered cases, to 90% even with a low number of iterations of the transductive classifier.

The three examples reported in Fig. A.3 correspond to a query based on trajectories shape (the first row) and to queries based on trajectories points (second and third rows). It is clear how bad results are progressively moved away from the presented results even when shape based analysis returns elements sharing shape but with dissimilar orientations.

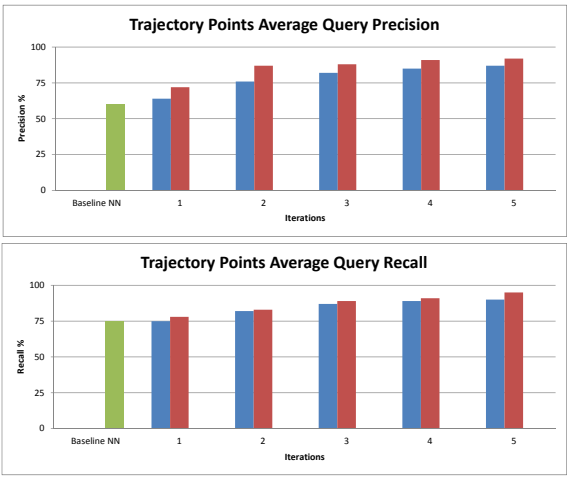


Figure A.4: Queries average precision and recall on people trajectories. Points analysis.

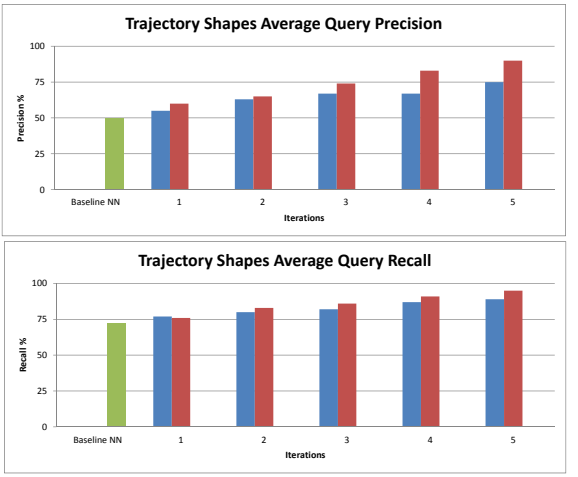


Figure A.5: Queries average precision and recall on people trajectories. Shape analysis.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

M. Agrawal and D. Doermann. Voronoi++: A dynamic page segmentation approach based on voronoi and docstrum features. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1011–1015, 2009.

I. Almajai, F. Yan, T. deCampos, A. Khan, W. Christmas, D. Windridge, and J. Kittler. Anomaly detection and knowledge transfer in automatic sports video annotation. In *Studies in Computational Intelligence*. Springer, 2012.

M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1 –8, june 2008.

E. Appiani, F. Cesarini, A.M. Colla, M. Diligenti, M. Gori, S. Marinai, and G. Soda. Automatic document classification and indexing in high-volume applications. *International Journal of Document Analysis and Recognition*, 4(2):69–83, 2001.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, pages 243–272, 2008.

S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Proceedings of the International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 435 –440, 29 2010-sept. 1 2010.

M.F. Balcan and A. Blum. A pac-style model for learning from labeled and unlabeled data. In *In Proceedings of the 18th Annual Conference on Computational Learning Theory*, pages 111–126. COLT, 2005.

S. Baldi, S. Marinai, and Soda. G. Using treegrammars for training set expansion in page classification. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 829–833, 2003.

D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 1st International ACM Workshop on Multimedia access to 3D Human Objects*, Scottsdale, Arizona, USA, November 2011.

M. Belkin. *Problems of Learning on Manifolds*. PhD thesis, 2003.

M. Belkin and P. Niyogi. Semi-supervised learning on riemannian manifolds. *Machine Learning*, 56(1-3), June 2004.

A. Berg, J. Deng, and F.F. Li. Large scale visual recognition challenge, 2010. URL `http://www.image-net.org/challenges/LSVRC/2010/`.

I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Rev.*, 97:115–147, 1987a.

I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987b.

A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 19–26, 2001.

P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler. Kernel null space methods for novelty detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3374–3381, 2013.

D. Borghesani, D. Coppi, C. Grana, S. Calderara, and Cucchiara R. Feature space warping relevance feedback with transductive learning. In *ACIVS*, pages 70–81, 2011.

B.E. Boser, I. Guyon, and V. Vapnik. Proceedings of the fifth annual acm conference on computational learning theory. In *COLT*, pages 144–152. ACM, 1992.

J.E. Boyd and J. Meloche. Evaluation of statistical and multiple-hypothesis tracking for video traffic surveillance. *Machine Vision and Applications*, 13(5-6), 2003.

M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1820 –1833, 2011.

A. L. Brown and M. J. Kane. Preschool children can learn to transfer: Learning to learn and learning from example. In *Cognitive Psychology*, pages 493—523, 1988.

Y. Cai, N. de Freitas, and J.J. Little. Robust visual tracking for multiple targets. In *Proceedings of the European Conference on Computer Vision(ECCV)*, pages 107–118, 2006.

S. Calderara, R. Cucchiara, and A. Prati. Bayesian-competitive consistent labeling for people surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):354–360, February 2008.

S. Calderara, A. Prati, and R. Cucchiara. Video surveillance and multimedia forensics: an application to trajectory analysis. In *Proc. of the ACM workshop on Multimedia in forensics, mifor*, pages 13–18, 2009.

S. Calderara, A. Prati, and R. Cucchiara. Mixtures of von mises distributions for people trajectory shape analysis. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(4):457 –471, 2011.

F. Cesarini, M. Lastri, S. Marinai, and G. Soda. Encoding of modified x-y trees for document classification. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1131–1136, 2001.

D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 554–560, New York, NY, USA, 2006. ACM.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009.

YJ. Chang, K. Kamataki, and T. Chen. Mean shift feature space warping for relevance feedback. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 1849–1852, Nov 2009.

YW. Chang and CJ. Lin. Feature ranking using linear svm. In *Journal of Machine Learning Research*, pages 53–64, 2008.

O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

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

K. Chen, F. Yin, and C.L. Liu. Hybrid page segmentation with efficient whitespace rectangles extraction and grouping. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 958–962, 2013.

N. Chen and D. Blostein. A survey of document image classification: Problem statement, classifier architecture and performance evaluation. *Interna-*

tional Journal of Document Analysis and Recognition, 10(1):1–16, May 2007.

Y. Chi, X. Song, D. Zhou, K. Hino, and B.L. Tseng. On evolutionary spectral clustering. *ACM Trans. Knowl. Discov. Data*, 3:17:1–17:30, 2009.

C. Clausner, S. Pletschacher, and A. Antonacopoulos. Aletheia - an advanced document layout and text ground-truthing system for production environments. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 48–52, 2011.

D. Coppi, S. Calderara, and R. Cucchiara. People appearance tracing in video by spectral graph transduction. In *ICCV Workshops*. IEEE Computer Society, 2011a.

D. Coppi, S. Calderara, and R. Cucchiara. Appearance tracking by transduction in surveillance scenarios. In *Proceedings of the International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 142 – 147, September 2011b.

D. Coppi, S. Calderara, and R. Cucchiara. Active query process for digital video surveillance forensic applications. *Signal Image and Video Processing*, June 2013.

D. Coppi, C. Grana, and R Cucchiara. Illustrations segmentation in digitized documents using local correlation features. In *Prooceeding of the 10th Italian Conference on Digital Libraries*, 2014a.

D. Coppi, J. Kittler, T. deCampos, F. Yan, and R. Cucchiara. On detection of novel categories and subcategories of images using incongruence. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*. ACM, 2014b.

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

I. J. Cox, M.L. Miller, T.P. Minka, and P.N. Yianilos. An optimized interaction strategy for bayesian relevance feedback. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 553–558, 1998.

M.l Crucianu, M. Ferecatu, and N. Boujemaa. Relevance feedback for image retrieval: a short survey. In *In State of the Art in Audiovisual Content-Based Retrieval, Information Universal Access and Interaction including Datamodels and Languages (DELOS2 Report*, 2004.

M. Culp and G. Michailidis. Graph-based semisupervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:174–179, January 2008.

W. Dai, Q. Yang, GR. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 193–200. ACM, 2007.

W. Dai, O. Jin, GR. Xue, Q. Yang, and Y. Yu. Eigentransfer: a unified framework for transfer learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. ACM, 2009.

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

C. Daskalakis, P. W. Goldberg, and C.H. Papadimitriou. The complexity of computing a nash equilibrium. In *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing*, STOC '06, pages 71–78. ACM, 2006.

C.s Daskalakis. On the complexity of approximating a nash equilibrium, 2011.

J. Davis and P. Domingos. Deep transfer via second-order markov logic. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 217–224. ACM, 2009.

J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

M. Diligenti, P. Frasconi, and M. Gori. Hidden tree markov models for document image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:2003, 2003.

P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 304 –311, june 2009.

P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99, 2011.

G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, 2: 127–151, 2011.

A. Doucet, N. De Freitas, and N. Gordon, editors. *Sequential Monte Carlo methods in practice*. 2001.

O. Duchenne, A. Joulin, and J. Ponce. A graph–matching kernel for object categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

A. Erdem and M. Pelillo. Graph transduction as a noncooperative game. *Neural Computation*, 24(3):700–723, 2012.

F. Esposito, D. Malerba, Lisi F.A., and W. Ras. Machine learning for intelligent processing of printed documents. *Journal of Intelligent Information Systems*, 14:175–198, 2000.

A. Ess, B. Leibe, K. Schindler, and L. van Gool. Robust multiperson tracking from a mobile platform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1831–1846, oct. 2009.

M. Everingham, L. J. Van Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2), 2010.

M. W. Eysenck and M. T. Keane. *Cognitive psychology: a student's handbook*. Psychology Press, 2005.

N. FarajiDavar, T. de Campos, J. Kittler, and Fei Yan. Transductive transfer learning for action recognition in tennis games. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1548 –1553, 2011.

Y. Fataicha, M. Cheriet, J.Y. Nie, and C.Y. Suen. Content Analysis in Document Images: A Scale Space Approach. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 3, pages 335–338, 2002.

Li F.F., R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, April 2006.

R. Flamary, N. Jrad, R. Phlypo, M. Congedo, and A. Rakotomamonjy. Mixed-norm regularization for brain decoding. Technical report, Laboratoire LITIS, Université de Rouen, 2012.

W. Forstner, B. Boudewijn Moonen, and C.F. Gauss. A metric for covariance matrices. Technical report, Dept.Geodesy Geoinform., Stuttgart Univ., Stuttgart, Germany, 1999.

G. Fumera, F. Roli, and G. Giacinto. Reject option with multiple thresholds. *Pattern Recognition*, 33:2099–2101, 2000.

D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1): 41–59, jun 2007.

C. Grana, D. Borghesani, and R. Cucchiara. Automatic segmentation of digitalized historical manuscripts. *Multimedia Tools and Applications*, pages 1–24, July 2010.

G. Griffin, A. D. Holub, and P. Perona. The Caltech 256 object category dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007.

Kjellstro H., D. Kragic, and M.J. Black. Tracking people interacting with objects. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 747–754, june 2010.

J. Ha, R.M. Haralick, and I.T. Phillips. Recursive x-y cut using bounding boxes of connected components. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 952–955 vol.2, 1995.

L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *Trans. Comp.-Aided Des. Integ. Cir. Sys.*, 11(9):1074–1085, November 2006.

L.K. Hansen, C. Liisberg, and P. Salamon. The error-reject tradeoff. *Open Systems and Information Dynamics*, 4:159–184, 1995.

T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi. Discriminative spatial pyramid. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1617–1624, Washington, DC, USA, 2011. IEEE Computer Society.

D. R. Hofstadter and Fluid Analogies Research Group. *Fluid concepts and creative analogies : computer models of the fundamental mechanisms of thought*. Penguin, 1998.

Joseph T. Howson. Equilibria of polymatrix games. *Management Science*, 18: 312–318, 1972.

J. Hu, R. Kashi, and R. Wilfong. Document Classification Using Layout Analysis. In *Proceedings of the International Conference on Database and Expert Systems Applications (DEXA)*, pages 556–560, 1999.

W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):663–671, april 2006.

C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *Proceedings of the European Conference on Computer Vision(ECCV)*, ECCV '08, pages 788–801, 2008.

Y. Huang, Q. Liu, S. Zhang, and D.N. Metaxas. Image retrieval via probabilistic hypergraph ranking. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, june 2010.

Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples, 1998.

T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 290–297, 2003.

N. Journet, J.Y. Ramel, R. Mullot, and V. Eglin. Document image characterization using a multiresolution analysis of the texture: application to old documents. *International Journal of Document Analysis and Recognition*, 11(1):9–18, 2008.

A. Kapoor. *Learning discriminative models with incomplete data*. PhD thesis, Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Arts and Sciences, 2006.

R. Kaucic, A.G.A. Perera, G. Brooksby, J. Kaufhold, and A. Hoogs. A unified framework for tracking through occlusions and across sensor gaps. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 990–997 vol. 1, 2005.

Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1805–1918, 2005.

ZW. Kim. Real time object tracking based on dynamic feature grouping with background subtraction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, june 2008.

K. Kise, A. Sato, and M. Iwata. Segmentation of page images using the area voronoi diagram. *Computer Vision and Image Understanding*, 70(3):370–382, 1998.

A. Kitamoto, M. Onishi, T. Ikezaki, D. Deuff, E. Meyer, S. Sato, T. Muramatsu, R. Kamida, T. Yamamoto, and K. Ono. Digital Bleaching and Content Extraction for the Digital Archive of Rare Books. In *Proceedings of the International Conference on Document Image Analysis for Libraries (DIAL)*, pages 133–144, 2006.

J. Kittler, W. Christmas, T. deCampos, D. Windridge, F. Yan, J. Illingworth, and M. Osman. Domain anomaly detection in machine perception: A system architecture and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

CH. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 685–692, june 2010.

T. Kurita and T. Kato. Learning of personal visual impression for image database systems. In *Document Analysis and Recognition, 1993., Pro-*

ceedings of the Second International Conference on, pages 547–552, 1993.

I. Kuzborskij, F. Orabona, and B. Caputo. From n to n+1: Multiclass transfer incremental learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3358–3365, 2013.

J. Lafferty and L. Wasserman. Statistical analysis of semi-supervised regression. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, pages 801–808. MIT Press, Cambridge, MA, 2008.

C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by betweenclass attribute transfer. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

N.D. Lawrence and M.I. Jordan. Semi-supervised learning via gaussian processes. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, 2004.

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

G. Lazzara, R. Levillain, T. Geraud, Y. Jacquelet, J. Marquegnies, and A. Crepin-Leblond. The scribo module of the olena platform: A free software framework for document image analysis. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 252–258, 2011.

B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 30(10):1683–1698, oct. 2008.

C. Leistner, H. Grabner, and H. Bischof. Semi-supervised boosting using visual similarity learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

C. Leistner, A. Saffari, J. Santner, and H. Bischof. Semi-supervised random forests. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 506–513, 2009.

F. Li and H. Wechsler. Open set face recognition using transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11), nov. 2005.

Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2953–2960, june 2009.

J.J. Lim, R. Salakhutdinov, and A. Torralba. Transfer learning by borrowing examples for multiclass object detection. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, 2011.

T. Lin and H. Zha. Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):796–809, 2008.

W. Liu and SF. Chang. Robust multi-class transductive learning with graphs. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 381 –388, june 2009.

J. Luo and M. A. Nascimento. Content-based sub-image retrieval using relevance feedback. In *ACM International Workoshop on Multimedia Databases*, pages 2–9, 2004.

U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17, 2007.

M. Markou and S. Singh. Novelty detection: A review - part 1: Statistical approaches. *Journal of Signal Processing*, 83, 2003a.

M. Markou and S. Singh. Novelty detection: A review - part 2: Neural network based approaches. *Journal of Signal Processing*, 83, 2003b.

L. Matthews, T. Ishikawa, and S. Baker. The template update problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):810 – 815, june 2004.

G. Meng, N. Zheng, Y. Song, and Y. Zhang. Document images retrieval based on multiple features combination. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 143–147, 2007.

M.J. Metternich, M. Worring, and A.W.M. Smeulders. Color based tracing in real-life surveillance data. In Yun Q. Shi, editor, *Transactions on Data Hiding and Multimedia Security V*, Lecture Notes in Computer Science, pages 18–33. Springer Berlin / Heidelberg, 2010.

K. Muandet and B. Schölkopf. One-class support measure machines for group anomaly detection. *CoRR*, 2013.

John Nash. Non-Cooperative Games. *The Annals of Mathematics*, 54(2):286–295, September 1951.

S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte. Document Image Segmentation Using a 2D Conditional Random Field Model. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 407–411, 2007.

K. Nigam, A.K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, pages 103–134, May 2000.

M-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1447–1454, 2006.

M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

K. Okuma, A. Taleghani, N. De Freitas, O. De Freitas, J.J. Little, and D.G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Proceedings of the European Conference on Computer Vision(ECCV)*, pages 28–39, 2004.

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

E. J. Pauwels and O. Ambekar. One class classification for anomaly detection: Support vector data description revisited. In *Advances in Data Mining. Applications and Theoretical Aspects*, volume 6870 of *Lecture Notes in Computer Science*. Springer, 2011.

T. Pavlidis and J. Zhou. Page segmentation by white streams. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 945–953, 1991.

S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009.

A.G.A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and Wensheng Hu. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 666–673, june 2006.

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

F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *IEEE European Conference on Computer Vision*, 2010.

R. W. Picard, T. P. Minka, and M. Szummer. Modeling user subjectivity in image libraries. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 777–780, 1996.

J.C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on lie algebra. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 728–735. IEEE Computer Society, 2005.

K. Porkaew, K. Chakrabarti, and S. Mehrotra. Query refinement for content based multimedia retrieval in mars. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 747–751, 1999.

L.G. Quintas. A note on polymatrix games. *International Journal of Game Theory*, 18(3):261–272, 1989.

V. Radosavljevic, N. Kojic, G. Zajic, and B. Reljin. The use of unlabeled data in image retrieval with relevance feedback. In *Symposium on Neural Network Applications in Electrical Engineering*, pages 21–26, 2008.

E. Rodner, E. S. Wacker, M. Kemmler, and J. Denzler. One-class classification for anomaly detection in wire ropes with gaussian processes in a few lines of code. In *IAPR Conference on Machine Vision Applications*, 2011.

M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):644–655, 1998.

B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, May 2008.

H. Sahbi, J.-Y. Audibert, and R. Keriven. Graph-cut transducers for relevance feedback in content based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

H. Sahbi, P. Etyngier, J.-Y. Audibert, and R. Keriven. Manifold learning using robust graph laplacian for interactive image search. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

C. Sammut and G. I. Webb, editors. *Encyclopedia of Machine Learning*. Springer, 2010.

J. Sanchez, F. Perronnin, and T. E. deCampos. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*, 33(16), 2012.

B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection, 2000.

B. Schölkopf, J.C. Platt, J.C. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

F. Sebastiani and Consiglio Nazionale Delle Ricerche. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

M. Seeger. Input-dependent regularization of conditional density models. Technical report, Institute for Adaptive and Neural Computation, 2001.

A. Singh, R.D. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn't. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, pages 1513–1520, 2009.

G. Sivakamasundari and V. Seenivasagam. Different relevance feedback techniques in cbir: A survey and comparative study. In *Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on*, pages 1115–1121, 2012.

A. Smeulder, D. Chu, R. Cucchiara, S. Calderara, A. Deghan, and M. Shah. Visual tracking: an experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 2013.

R. Smith. An Overview of the Tesseract OCR Engine. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633, Washington, DC, USA, 2007.

S. Stalder, H. Grabner, and L. van Gool. Beyond semi-supervised tracking: Tracking should be as simple as detection, but not simpler than recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1409–1416, 2009.

D. Svensson, M. Ulmke, and L. Danielsson. Joint probabilistic data association filter for partially unresolved target groups. In *Information Fusion (FUSION), 2010 13th Conference on*, pages 1–8, july 2010.

M. Szummer. *Learning from partially labelled data.* PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2002.

M. Szummer and T. Jaakkola. Kernel expansions with unlabeled examples. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, pages 626–632, 2000.

M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, pages 945–952, 2001.

M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, pages 945–952. MIT Press, 2002.

F. Tang, S. Brennan, Qi Zhao, and Hai Tao. Co-tracking using semi-supervised support vector machines. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, oct. 2007.

D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7): 1088–1099, 2006.

D. M. J. Tax and R. P. W. Duin. Outlier detection using classifier instability. In *SSPR/SPR*, Lecture Notes in Computer Science, pages 593–601. Springer, 1998.

D. M. J. Tax and R. P. W. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, January 2004.

K. Tieu and P. Viola. Boosting image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 228–235 vol.1, 2000.

F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Review*, 43(2), 2001.

T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, pages 1134–1142, November 1984.

K. Vandist, M. De Schryver, and Y. Rosseel. Semisupervised category learning: The impact of feedback in learning the information-integration task. *Attention, Perception & Psychophysics*, 71(2):328–341, February 2009.

V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics) 2nd Edition*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008. URL `http://www.vlfeat.org/`.

A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:480–492, 2012.

J. Vermaak, A. Doucet, and P. Pérez. Maintaining multi-modality through mixture tracking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1110–1116, 2003.

R. Vezzani and R. Cucchiara. Video surveillance online repository (visor): an integrated framework. *Multimedia Tools and Applications*, 50(2):359–380, November 2010.

Y. Wang, I.T. Phillips, and R.M. Haralick. Document zone content classification and its performance evaluation. *Pattern Recognition*, 39(1):57–73, 2006.

J. W. Weibull. Evolutionary game theory, 1995.

D. Weinshall, A. Zweig, H. Hermansky, S. Kombrink, F. W. Ohl, J. Anemuller, J. H. Bach, L. Van Gool, F. Nater, T. Pajdla, M. Havlena, and M. Pavel. Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10), 2012.

A. Winder, T. Andersen, and E.H.B. Smith. Extending page segmentation algorithms for mixed-layout document processing. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1245–1249, 2011.

B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, November 2007.

P. Wu and T. G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 110–117. ACM, 2004.

Y. Wu, Q. Tian, and T.S. Huang. Integrating unlabeled images for image retrieval based on relevance feedback. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 1, pages 21–24 vol.1, 2000.

Ying Wu and T.S. Huang. Color tracking by transductive learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 133–138. IEEE Computer Society, 2000.

J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1200–1207, june 2009.

Y. Yao and G. Doretto. Boosting for transfer learning with multiple sources. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1855–1862, June 2010.

Safa R. Zaki and Robeir M. Nosofsky. A high-distortion enhancement effect in the prototype-learning paradigm: dramatic effects of category learning during test. *Memory & cognition*, 35(8):2088–2096, December 2007.

Y. Zha, Y. Yang, and D. Bi. Graph-based transductive learning for robust visual tracking. *Pattern Recognition*, 43(1), 2010.

C. Zhang, Y. Shao, J. Tan, and N. Deng. Mixed-norm linear support vector machine. *Neural Computing and Applications*, 23:2159–2166, 2013.

L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, june 2008.

Y. Zhang and DY. Yeung. Transfer metric learning by learning task relationships. In *KDD*, pages 1199–1208. ACM, 2010.

T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208–1221, sept. 2004.

D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, pages 321–328. MIT Press, 2004.

X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, April 2003.

X. Zhu. Semi-supervised learning literature survey, 2008.

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 912–919, 2003.

X. Zhu, T. Rogers, R. Qian, and C. Kalish. Humans perform semi-supervised classification too. In *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 1*, AAAI'07, pages 864–869, 2007.