TRABAJO ESPECIAL DE GRADO

RECONOCIMENTO DE UBICACIÓN CON LAPSOS DE TIEMPO PARA LA ANOTACIÓN AUTOMÁTICA DE VIDEOS ANTIGUOS

LOCATION RECOGNITION OVER LARGE TIME LAGS FOR AUTOMATIC ANNOTATION OF ANCIENT MOVIES

Presentado ante la Ilustre

Universidad Central de Venezuela

Por la Br. Stephany B. Rosalía

Para optar al Título de

Ingeniera Mecánico

Caracas, Noviembre 2015

RECONOCIMENTO DE UBICACIÓN CON LAPSOS DE TIEMPO PARA LA ANOTACIÓN AUTOMÁTICA DE VIDEOS ANTIGUOS

Rosalía Stephany Brassesco[1]

[1] *Universidad Central de Venezuela. stephanybrassesco@gmail.com*

INTRODUCCIÓN

Los algoritmos basados en el aprendizaje han tenido mucho éxito recientemente en un amplio rango de problemas de visión artificial.

El uso extendido de dispositivos móviles dotados de cámaras, sensores GPS y conexión a internet, ha generado interés por el problema de reconocimiento de ubicación visual, ya que es visto como una herramienta valiosa en aplicaciones de turismo y de navegación.

Museos y otras instituciones han empezado a promover la recolección y digitalización de fotografías y vídeos antiguos con campañas de preservación del patrimonio cultural.

Las fotografías y vídeos antiguos frecuentemente incluyen monumentos, plazas, fuentes y otros lugares de interés y con valor histórico en pueblos y ciudades.

Se piensa que la visión artificial es una disciplina adecuada para explotar estos recursos, ya que los métodos manuales tienen un costo prohibitivo.

Sin embargo, la mayoría de los algoritmos clásicos de aprendizaje supervisado hacen la suposición fundamental de que las imágenes de prueba y las imagines que se usaron para entrenar al sistema provienen de una misma distribución, cuando en práctica la situación es menos ideal, y el rendimiento de los algoritmos se ve significativamente afectado con este cambio de dominio.

Tradicionalmente este problema ha sido tratado entrenando un nuevo modelo para cada dominio de interés, pero en el caso de fotografías antiguas esta no es una solución viable así que se propone alternativamente utilizar métodos de adaptación de dominio.

METODOLOGÍA

Reconocimiento de Ubicación

El reconocimiento de la ubicación puede ser definido de manera amplia como el problema de determinar dónde fue capturada una imagen. Sin embargo no existe un solo concepto de ubicación, y varias formas de representar un lugar han sido propuestas, como aquellas basadas en grafos, en imágenes icónicas, y representaciones que se apoyan en modelos 3D de la escena.

En este trabajo las ubicaciones se definen simplemente por sus coordenadas geográficas. En este escenario, el reconocimiento de ubicación puede ser tratado como un problema de clasificación de imágenes, donde los lugares, representados por sus coordenadas son un conjunto discreto de categorías. El modelo de un lugar es aprendido *offline* y al momento de una consulta el clasificador anota la imagen con la etiqueta de la ubicación que tenga mayor puntaje.

Adaptación de dominio

En la mayor parte de la investigación sobre aprendizaje visual supervisado, tanto teórico como empírico, se supone que las muestras utilizadas en la fases de entrenamiento y prueba son extraídas de la misma distribución. En contraste, en el escenario de adaptación de dominio no se hace esta suposición, y se consideran dos distribuciones desconocidas, la del conjunto fuente y la del conjunto objetivo. El problema es entonces aquel de aprender a completar una tarea cuando las dos distribuciones son distintas.

El desplazamiento entre dos distribuciones es un problema que se presenta con frecuencia en el ámbito de aprendizaje artificial, cuando las condiciones en las que un modelo fue desarrollado difieren de las condiciones en las que se hará uso práctico del mismo. Consideremos por ejemplo un filtro de correo basura como una aplicación donde está presente la adaptación de dominio. El desarrollo del sofisticado sistema pudo haber tomado un esfuerzo considerable, así que resultaría muy conveniente que el sistema funcione aún cuando hayan cambiado ligeramente las condiciones externas, ya sea porque se han unido nuevos usuarios que reciben una variedad distinta de correo, o porque el tipo de correo basura ha cambiado con el tiempo.

Un ejemplo en el contexto visual sería tener una gran base de datos etiquetada de imágenes provenientes de la red, y ser capaces de utilizarla para etiquetar imágenes de una base de datos de vídeo.

La meta de la adaptación de dominios es precisamente desarrollar algoritmos capaces de mejorar el rendimiento de una función predictiva sobre el conjunto objetivo utilizando el conocimiento disponible de la distribución fuente.

En este trabajo nos enfocamos en algoritmos de adaptación de dominio del tipo sin supervisión, en el que la totalidad de las imágenes pertenecientes al conjunto objetivo están sin etiquetar.

Se utilizaron los siguientes dos métodos de adaptación de dominio, basados en subespacios.

1. Geodesic Flow Kernel

La idea principal detrás de este enfoque es la de modelar los datos con subespacios lineales y empotrar el conjunto objetivo y el conjunto fuente en una variedad de Grassmann, al mismo tiempo construyendo un flujo geodésico entre los dos puntos. Los puntos en este flujo representan interpolaciones entre los subespacios. Las características de las imágenes son proyectadas en el flujo para obtener vectores de dimensión infinita. El producto interno de estos vectores definen una función que puede ser calculada de forma cerrada.

2. Subspace Alignment

Este método modela el conjunto de datos fuente y objetivo con subespacios lineales. La idea principal detrás del enfoque es la de optimizar una función de mapeo que alinea los dos subespacios, produciendo un espacio de vectores característicos invariantes al dominio.

Representación de la imagen

El primer paso para construir sistemas que tengan una comprensión semántica del ambiente visual es extraer características eficientes y efectivas. Una de las razones detrás de los impresionantes resultados que se han obtenido recientemente en varios problemas pertenecientes al campo de visión artificial ha sido el desarrollo de nuevas representaciones de las imágenes.

Vectores Característicos extraídos por una Red Neural Convolutiva

Los modelos basados en redes neurales profundas han tenido éxito en problemas de reconocimiento visual para los que existen datos de entrenamiento en cantidad abundante. Un ejemplo es aquel de reconocer caracteres escritos a mano. Recientemente este tipo de algoritmo ha superado el rendimiento de todos los otros métodos en bases de datos de referencia, como ImageNet.

En general, las arquitecturas profundas en esquema supervisado sufren de sobreajuste cuando hay una cantidad insuficiente de datos de entrenamiento.

Caffe + Imagenet

El programa Caffe permite definir, desarrollar y entrenar redes neurales convolutivas, y además ejecutar modelos pre-entrenados. Para extraer representaciones de imágenes eficientes los autores adoptan un modelo profundo propuesto por Krizhevsky como la arquitectura de base. Los vectores característicos son extraídos cuando se propagan los valores de intensidad RGB de lo píxeles a través de las capas convolutivas y las completamente conectadas de la red.

Las activaciones de las capas ocultas cercanas a la salida de la red se toman como las características *Caffe,* las cuales son adecuadas para tareas de reconocimiento visual. Son buenas representaciones del mundo, ya que capturan el conocimiento semántico, y son lo

suficientemente generales como para tener un rendimiento aceptable incluso en tareas para las cuales los modelos no fueron directamente entrenados.

Las características Caffe + Imagenet proviene de un modelo entrenado en la base de datos de gran escala ILSVRC, una base de datos de reconocimiento de objetos.

PlacesCNN

Para obtener los vectores característicos PlacesCNN se entrenó una red neural convolutiva con una base de datos de imágenes de lugares llamada Places que tiene más de 7 millones de imágenes etiquetadas, y 476 categorías diferentes.

Bases de datos

Los algoritmos estudiados en este trabajo fueron puestos a prueba en dos bases de datos distintas, LTLL y Roma, Ciudad Abierta. En las dos bases de datos consideradas, el desplazamiento de dominio entre las distribuciones del conjunto fuente y el conjunto objetivo es una consecuencia del pasar del tiempo.

Además de los desafíos usuales del reconocimiento de ubicación, incluyendo aspectos como cambios en las condiciones de iluminación, oclusión, y variaciones en el punto de vista de la cámara, las bases de datos reflejan problemas que surgen al considerar lapsos de tiempo extensos, como lo son la degradación del color, cambios en el proceso de adquisición de la imagen y cambios físicos en el lugar.

En lo que sigue se describe brevemente cada base de datos.

1. LTLL

Esta base de datos consiste de un conjunto de imágenes antiguas y un conjunto correspondiente de imágenes modernas. Contiene imágenes de 25 lugares de todo el mundo, incluyendo monumentos en ciudades Europeas y Asiáticas. Las imágenes modernas fueron descargadas a través€ś de internet de sitios como Flickr y Google Imágenes. EN total hay 225 imágenes antiguas y 275 imágenes modernas.

2. Roma, Ciudad Abierta

Esta base de datos consiste en un conjunto de fotogramas de la película italiana de 1945 *Roma, Ciudad Abierta.* Hay 16 imágenes en diferentes puntos de la ciudad de Roma. Para desarrollar los experimentos propuestos en este trabajo, se descargó un conjunto de imágenes modernas de las ubicaciones de Google Street View. Por cada ubicación se añadieron 72 imágenes modernas a la base de datos: dada una coordenada geográfica se descargaron imágenes correspondientes a 8 orientaciones, y moviendo las coordenadas unos metros se repitió el proceso 8 veces.

Experimentos

En esta sección se explica la metodología seguida para llevar a cabo los experimentos. La tarea de reconocimiento de la ubicación con lapsos de tiempo extensos fue examinada bajo el marco de clasificación supervisada de imágenes. Los experimentos se repitieron con dos bases de tatos distintas, LTLL y Roma, Ciudad Abierta.

En todos los experimentos se usó un clasificador basado en el algoritmo de 1-vecino más cercano. Las imágenes modernas son el conjunto *fuente* y están etiquetadas con una de las ubicaciones. Todas las imágenes modernas fueron usadas para entrenar el clasificador. Las imágenes antiguas son el conjunto objetivo. En este conjunto el clasificador es puesto a prueba anotando las imágenes con una de las etiquetas de ubicación.

El rendimiento final reportado es la precisión promedio de todas las categorías, en todo el conjunto de imágenes antiguas.

Como representación de la imagen se usaron vectores característicos extraídos por una red neural profunda en el programa *Caffe.* En particular se extrajo la séptima capa convolutiva. La dimensión original de los vectores es 4096. Esta dimensión se redujo a la dimensión intrínseca estimada a través de tres métodos: análisis de componentes principales conservando el 99% de la varianza (EIG), S*ubspace Dimensionality Measure* (SDM) y M*aximum Likelihood Estimation* (MLE).

Por último, para cada dimensión se llevaron a cabo tres experimentos de clasificación, uno sin adaptación alguna, uno con el método de adaptación de dominio *Geodesic Flow Kernel* (GFK) y otro con *Subspace Alignment (SA).*

RESULTADOS Y DISCUSIÓN

Los resultados de los Cuadros 1-4 muestran que el reconocimiento de ubicación automático con lapsos de tiempo se ve beneficiado por los métodos de adaptación de dominio, ya que en todos los casos el rendimiento de clasificación mejora al usar adaptación de dominio.

En los experimentos de la base de datos LTLL los mejores resultados se obtuvieron con el método GFK, con 57.8% de rendimiento, y con una dimensión intrínseca estimada de 16, una reducción considerable a partir de la dimensión original de 4096.

Se hace notar que los vectores característicos PlacesCNN, entrenados en una base de datos de escenas, tienen mejor rendimiento que los vectores Caffe+Imagenet, que fueron entrenados en una tarea de clasificación de objetos, y dista más de la tarea considerada en este trabajo.

Anotar automáticamente fotogramas de películas antiguas con la etiqueta de ubicación

geográfica correcta resulta ser un problema difícil, y en comparación con la base de datos de monumentos, el rendimiento de clasificación empeora. Esto era de esperarse ya que los fotogramas tienen frecuentemente puntos de vista menos aventajados con respecto a las imágenes históricas que buscaban capturar precisamente los monumentos.

En la base de datos de Roma, Ciudad Abierta, el método de adaptación de dominio SA tiene en promedio un mejor rendimiento que el método GFK. Y nuevamente la técnica de estimación de dimensión intrínseca MLE da los mejores resultados. En esta base de datos el rango de dimensiones estimadas es aún mayor.

Es importante hacer notar que para estos experimentos, el rendimiento de clasificación aleatoria estaba situado en 6,25%, y que sin adaptación de dominio, los vectores característicos *Caffe + Imagenet* no tienen ninguna capacidad de discriminación.

Cuadro 1. Resultados de clasificación (%) Caffe + Imagenet. LTLL

EIG (fuente)	28	41.18	52.45	53.43
EIG(meta)	36	41.18	52.20	52.45
SDM	19	41.18	50.00	50.98
MLE(fuente)	32	41.18	50.00	55.39
MLE(meta)	37	41.18	51.47	52.45

Cuadro 2. Resultados de clasificación (%) Caffe + Places. LTLL

	Dimensión	No – Adaptación	GFK	SA
EIG (fuente)	11	48.04	52.94	47.55
EIG(meta)	11	48.04	52.94	47.55
SDM	20	48.04	51.47	51.96
MLE(fuente)	16	48.04	57.84	51.47
MLE(meta)	19	48.04	53.92	51.96

Cuadro 3. Resultados de clasificación (%) Caffe + Imagenet. Roma

	Dimensión	No – Adaptación	GFK	SA
EIG (fuente)	7	6.25	6.25	18.75
EIG(meta)	13	6.25	18.75	18.75
SDM	178	6.25	6.25	12.5
MLE(fuente)	30	6.25	6.25	6.25
MLE(meta)	15	6.25	12.5	6.25

Cuadro 4. Resultados de clasificación (%) Caffe + Places. Roma

	Dimensión	No – Adaptación	GFK	SA
EIG (fuente)	4	18.75	18.75	6.25
EIG(meta)	10	18.75	18.75	25
SDM	77	18.75	25.00	31.25
MLE(fuente)	15	18.75	18.75	31.25
MLE(meta)	14	18.75	18.75	25

CONCLUSIONES

En este trabajo de grado se abordó el problema de reconocimiento visual de la ubicación de fotografías y vídeos antiguos bajo el marco de clasificación supervisada de imágenes. La tarea consistía en anotar las imágenes antiguas con una etiqueta de ubicación a partir de una base de datos de imágenes modernas del mismo conjunto de lugares.

El lapso temporal entre las dos colecciones de imágenes se modeló como un desplazamiento del dominio visual, y se usaron métodos de adaptación de dominio. En particular fueron evaluados los métodos *Geodesic Flow Kernel* y *Subspace Alignment,* y se demostró que en efecto son capaces de atenuar el efecto negativo de la presencia de un lapso temporal en el desempeño del clasificador.

Por otra parte se pudo observar que los vectores característicos extraídos por una red neural profunda que fue entrenada en una tarea más relacionada al problema a tratar muestran mejor desempeño en el nuevo dominio.

En este trabajo se consideraron solamente métodos de adaptación de dominio sin supervisión, una posible dirección futura de investigación sería considerar una pequeña cantidad de imágenes etiquetadas pertenecientes al conjunto objetivo, y determinar qué tanto se beneficiaría la tarea con este esfuerzo adicional.

Por otro lado, la estimación de la dimensión intrínseca utilizada para modelar el subespacio del dominio visual muestra grandes variaciones según la técnica empleada, un esquema de validación cruzada podría aplicarse para escoger este parámetro libre en futuros experimentos de este tipo.

Finalmente, cuando se quiere clasificar imágenes tomadas de un vídeo, la elección del fotograma es crucial, ya que las escenas urbanas al aire libre se ven afectadas por cambios de apariencia y de punto de vista entre el objetivo y las imágenes utilizadas para entrenar el modelo. Otro problema es que este tipo de escenas presentan estructuras no discriminatorias a causa de la vegetación, el cielo, paredes de edificios y calles. Bases de datos más extensas atenuarían el problema que ocasionan estos fotogramas no informativos.

SAPIENZA
UNIVERSITÀ DI ROMA

Department of Computer, Control, and
Management Engineering Antonio Ruberti
Sapienza University of Rome, Italy

Location Recognition Over Large Time Lags for Automatic Annotation of Ancient Movies

by

Rosalía Stephany Brassesco

A thesis presented for the degree of
Master in Artificial Intelligence and Robotics

October 2015

Supervisor: Prof. Barbara Caputo

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Barbara Caputo, for her continuous encouragement, expertise and motivation. I could not have hoped for better guidance.

To my parents and my sisters: thank you for everything.

Abstract

In visual recognition systems it is often the case that the conditions in which a model is developed differ from those in which the model is deployed. The performance of a system that has been trained on label data from a *source* domain is severely affected when tested on samples drawn from a different *target* domain distribution. This issue imposes a limit on the real-world applications of a system, as expensive manually annotated data is needed for every new domain. Domain adaptation aims to handle this mismatch between distributions, and to develop models that will perform well on the test data in the new domain.

In this work the explored domain shift is caused by a temporal gap between ancient photographs and frames of movies and modern images, and the task considered is that of location recognition.

The *Geodesic Flow Kernel* and the *Subspace Alignment* domain adaptation methods are evaluated on two datasets. The first one consists of historical and modern images depicting 25 locations of cities across the world, and the second one consists of frames of 1945 Italian movie *Rome, Open City* and modern geo-tagged images that cover the same set of locations downloaded from Google Street View.

We find that the classification performance improves when domain adaptation, together with state-of-the-art features, are used on the task of location recognition with a temporal gap between training and test sets.

Contents

Chapter 1

Introduction

1.1 Overview and Motivation

Learning based methods have proved successful in a range of computer vision tasks. Machine learning systems are developed and used in practical real world scenarios. However, there is still much work to be done.

Nowadays, with the use of mobiles devices supplied with cameras, GPS and Internet connectivity, location recognition has received much attention, as it is a potentially useful tool for tourism, smart navigation aid and many creative applications like re-photography [1].

Museums and other institutions have started to promote the collection and digitalization of ancient photographs and footage as part of cultural heritage campaigns. Together with ancient commercial films, old photographs form an important visual documentation of the culture and history of the last century.

Machine learning is thought to be a well suited framework for exploiting these resources, as manual annotation is labor intensive and expensive.

Ancient movies and photographs often depict monuments, important buildings, fountains, statues and places of interest in towns and cities. These landmarks often have historical value and are also popular touristic attractions. How-

ever, these samples have usually not been included in location recognition attempts, and therefore the difficulties raised by the visual variations in color, texture, etc., that they present for learning tasks with respect to modern images have not been widely considered.

Indeed, most standard supervised machine learning algorithms make the fundamental assumption that the test data has been drawn from the same distribution as the training data. In practice, the situation is often much less ideal and the performance of these algorithms is significantly affected when the system is deployed on test data from a *target* domain that differs from the *source* domain where the model was trained.

This problem has traditionally been dealt with by training a new model for every target domain. This involves a high training cost, and relies on the availability of abundant labeled data (expensive) for every new domain.

Adapting systems that have been trained on a source domain to perform well the same task on instances from a new target domain has been recognized as an important problem that is recently receiving ample attention.

The task of location recognition consists of determining where a photo was taken by comparing it to a database of images of previously seen locations [2]. The problem has seen recent interest due to the widespread use of devices that have a camera and Internet connectivity, and due to the availability of vast amounts of geo-tagged images from Internet sites such as Flickr and Google Street View that depict urban outdoor scenes.

1.2 Related Work

In this section an overview of related work is presented to help the reader put the present work into context.

Location recognition has been treated under an image classification framework, with the models being learned offline from a training set that covers a given set of locations [3, 4]. Location recognition has also been treated under the

alternative, though related setup of image retrieval [5, 6].

Although urban scenes and monuments appear in ancient photographs and paintings, these samples are usually not included in location recognition attempts, and difficulties raised for learning tasks by the visual variations in color, texture and brush strokes they present with respect to modern images have not been considered.

The relevant task of aligning 2D depictions of an architectural site including drawing paintings and historical photographs to a 3D model of the scene is addressed in [7] The starting point for this thesis is the work done by Fernando et al. in [8] where the task of location recognition over large time lags was first defined and explored. To our knowledge no other work to date has focused on this task. They also contributed a dataset of ancient and current photographs of matching locations of cities and towns around the world that was used in this thesis.

One of the main challenges in visual recognition is that of choosing image descriptors that are robust to variations in illumination, viewpoint, and occlusion, and that have a rich enough representational capacity to capture the semantic information required by the task.

The performance of gradient-based image representations has likely plateaued in recent years and has been exceeded in many tasks by CNN features. Deep models trained on large image databases are publicly available [9, 10]. In particular, in [11, 12] the authors develop a framework for training, testing, finetuning, and deploying models. They also provide a CNN whose architecture follows that of Krizhevsky et al..

The variability of visual data from two time periods is a case of domain shift, as the situation can be modeled as the training and test data being drawn from two different distributions. Domain adaptation aims to solve this problem. A survey of the literature and recent advances is offered in [13, 14]. The two methods used in the experiments in this thesis are based on subspace learning ad are introduced in [15] and [16].

1.3 Thesis Outline

The contribution of this thesis is investigating the task of automatically annotating ancient movies with the location labels of the scenes. Domain adaptation methods combined with state-of-the-art features are evaluated on this locations recognition task.

The chapter structure and organization of the rest of this thesis is described below.

Chapter 2 presents the background to the location recognition problem, and describes the type of image representation used in this thesis.

Chapter 3 presents the issue of domain shift, and describes the domain adaptation problem, as well as the particular methods used in this thesis.

Chapter 4 presents the datasets used in this work, as well as the experimental setup and analysis of the results.

Chapter 5 draws the overall conclusions of this work and provides suggestions for some future directions of research.

Chapter 2

Location Recognition

In this chapter the problem of location recognition is examined under an image classification framework. Image representation is also addressed.

2.1 Location Recognition

Location recognition has received a lot of recent attention and demand due to the ubiquitousness of mobile phones and other devices endowed with cameras, navigation systems and Internet connectivity.

Location recognition can be very broadly defined as the problem of determining where an image was taken. There is, however, not a single concept of location and various ways of representing a place have been used in research: graph-based representations [4], iconic-image representation [17], and representations that leverage on 3D models [18]. The possibility we consider is simply to define locations by their geographical coordinates.

In this setting location recognition becomes an image classification problem, where the locations, represented by their coordinates, are a discrete set of categories. The model for a place is learned offline, and the classifier annotates a query image with the label of the best scoring location.

Location recognition is a hard problem, affected by large appearance and viewpoint variation between the training images and test images as well as by the presence of a large number of non-discriminative structures due to vegetation, sky walls, and roads.

2.1.1 Location Recognition Over Large Time Lags

The task of location recognition over large time lags was defined in [8] as: annotate an ancient photograph with the correct location label, given a set of labeled modern images that cover the same pre-defined places.

2.2 Image Representation

The first step for building systems that have a semantic understanding of the visual world is extracting efficient, effective features.

One of the reasons behind the impressive recent advances in computer vision tasks such as location recognition is the development of effective image representations.

2.2.1 CNN-features

Deep models have had success in visual recognition tasks for which there is abundant training data, an early example is that of digit classification [19]; they have in recent years outperformed all other methods on important large scale benchmark datasets for image recognition such as ImageNet [20].

Supervised deep architectures will in general suffer from overfitting when there is an insufficient amount of training data.

Caffe + Imagenet features

In [11] and [12] the authors developed a framework that allows to train convolutional neural network models and to execute pre-trained models.

For developing a new state-of-the-art image representation they adopt the deep CNN model proposed by Krizhevsky et al. in [9] as the underlying architecture of the feature. The features are extracted by propagating the mean-centered raw RGB pixel intensity values through the convolutional and fully connected layer of the deep model. The activations of the hidden layers close to the output of the model are taken as the Caffe features.

Caffe features are well suited for visual recognition tasks. They are good representations of the visual world, as they capture semantic knowledge and are general enough that they perform reasonably well even in tasks the models weren't directly trained for. Such is the case of the task of scene recognition, as Caffe is learned on the ILSVRC, an object recognition dataset.

PlacesCNN features

In [10] the authors introduce a scene-centric database called Places that has more than 7 million labeled images spanning 476 place categories.

Using the Caffe framework described above, a CNN model is learned from this new corpus of training data. The features that are extracted in this way have state-of-the-art performance on scene-related tasks.

Chapter 3

Domain Shift

This chapter introduces dataset shift in the context of visual recognition, and describes domain adaptation as one of the approaches for dealing with the issue. Dataset shift, the problem that occurs when the conditions change between the training and testing situations, has wide implications in machine learning, as it weakens the generalization potential of models to new tasks.

3.1 Domain Adaptation

In most standard supervised learning research, both theoretical and empirical, it is assumed that the samples used in the training and test stages are drawn independently from the *same* data distribution. In contrast, in the domain adaptation setting this assumption is dropped, and two *different*, unknown data distributions are considered: the distribution $\mathcal{D}_S$ of the source data, and the distribution $\mathcal{D}_T$ of the target data. The problem becomes that of learning to perform a task when these two distributions are different.

The problem of a shift in the distribution of the training and test datasets arises often in machine learning applications. Indeed, it is frequently the case that the conditions under which a model was developed differ from the conditions in which practical use of the model will be made.

The performance of state-of-the-art machine learning based systems degrades significantly when there exists a large domain shift between the test data and the training data used to learn the model.

This effect is present even in artificial laboratory conditions such as that of object classifiers optimized on benchmark vision datasets, as was documented by Torralba et al. in [21]. Image datasets play an important role in computer vision research, they provide training data, and allow to measure and compare the performance of competing algorithms. Nevertheless it is the case that classification algorithms consistently and significantly suffer a drop in performance when they are trained and tested on two different image datasets, even when those datasets assume the same general task. This issue is known in the literature as dataset bias [21].

Consider an email spam filtering system as an example of a real world application where the domain adaptation problem is present. The sophisticated filtering system could have taken a significant effort to build, so it is very desirable that the system is usable at test time, even if the situation has changed slightly, perhaps because of the addition of new users that receive a different type of emails, or because the type of spam has evolved with time.

An application in the computer vision context would be having a large annotated corpus of web images as a source dataset and an unlabeled dataset of interest from a video corpus. It would be useful to be able to take advantage of the knowledge available from the first domain and apply it when learning a model that performs well a task such as object detection in the second, different, but somehow related domain.

The goal of domain adaptation is precisely that of developing algorithms that can improve the performance of a target predictive function over the target domain using the available knowledge from the source domain.

Domain adaptation has received attention in a variety of applications that involve different discriminative learning tasks and types of data where the current systems adapt poorly to new domains. Some examples are semantic role labeling, part-of-speech tagging and sentiment analysis in the context of natural language

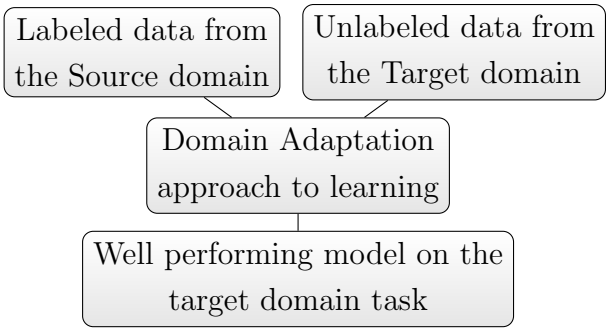

Figure 3.1: Diagrammatic Domain Adaptation problem overview

processing, and image classification, location recognition and object detection in the context of computer vision.

Types of DA and related problems

Domain adaptation is a particular case of the more general problem of transfer learning, which deals with how to transfer information from previous environments to help with learning, inference and prediction in a new environment. In [22] there is a comprehensive overview of transfer learning for classification, regression and clustering tasks applies to different areas. It also situates domain adaptation among related problems like covariance shift and multi-task learning.

Domain adaptation can be categorized into *supervised domain adaptation*, in which there is plenty of labeled data for both the source and the target domain, *semi-supervised domain adaptation*, in which the source data is labeled, a small portion of the target data is labeled and the rest is unlabeled, and *unsupervised domain adaptation* problem, which is the one focused on this work, and in which all the target domain data is unlabeled.

3.1.1 Unsupervised Domain Adaptation for Visual Classification

Motivation

Supervised machine learning models are trained on large amounts of labeled data. Annotating large images is a task that requires a significant human effort, making it time consuming and expensive.

It would be advantageous to be able to re-use already available labeled data from one domain to perform a machine learning task such as classification in a new domain.

Notation and problem statement

Let $\mathcal{X}_S$ denote the source feature subspace and let $X_S = \{\mathbf{x}_{Si}\}_{i=1}^{N_1}$ denote the source domain data pertaining to M categories, where $\mathbf{x}_{Si} \in \mathcal{X}_S$ is a source data instance. Also, let $\mathcal{Y}$ denote the label space, and $y_{Si} \in \mathcal{Y}$ the label that corresponds to $\mathbf{x}_S i$. Assume all source data to be labeled.

Similarly, let $\mathcal{X}_T$ denote the target feature subspace and let $X_T = \{\mathbf{x}_{Ti}\}_{i=1}^{N_1}$ denote the target domain data, where $\mathbf{x}_{Ti} \in \mathcal{X}_T$ is a target data instance. Assume all target domain data to be unlabeled, but pertaining to the same categories as the source domain data.

In an image classification problem the goal of the learning algorithm is to use a set of training data to select a function $f : \mathcal{X} \mapsto \mathcal{Y}$ that has a low expected error in predicting a class label for an unseen input $\mathbf{x}$. In the unsupervised domain adaptation setting, the model is learned from the knowledge in the labeled source dataset, $\{\mathbf{X}_S, y_S\}$, and the knowledge in the unlabeled target dataset, $\mathbf{X}_T$. The aim is to obtain from the available information a classifier function $f : \mathcal{X}_T \mapsto \mathcal{Y}$ that will perform well on the novel, unlabeled data from the target domain.

Directly training the classifier on the source data will minimize some loss function that measures the cost of the predictions with respect to the unknown distribution from which the source samples were drawn. Solving this problem

will not coincide in the case of different domains with obtaining the minimal loss with respect to the unknown distribution from which the target samples are drawn. Thus the need of the domain shift with novel algorithms.

3.2 Domain Adaptation Methods

Domain adaptation is, as discussed in [23], a very challenging problem. It is an open problem, and there is not yet an efficient general framework for tackling the problem. The type of method to use when deploying a system, must be chosen in function of what works best with the type of data and task being considered. That said, there are three main classes of domain adaptation algorithms:

Re-weighting or instance based methods which attempt to correct the domain shift by considering that the data instances from the source domain that are close to data instances from the target domain are more important for the task at hand, and should consequently be given a higher weight, reducing the discrepancy between the domains. In [24], [25], and [26] the domain adaptation problem is approached from this point of view.

Iterative methods which attempt to correct the domain shift by iteratively adjusting the model by incorporating new pseudo-labeled information, or by selectively adding or removing instances at each step. An example of this kind of method is presented in [27].

Projection based methods which attempt to correct the domain shift by finding a common space where the source and target data are close. Some of the methods that enter into these category are [28] which uses a metric learning approach and [29] which is based on feature augmentation. Also in this category are the methods based on subspace learning, the Geodesic Flow Kernel [16] and the Subspace Alignment methods [15], [30] are described in further detail later.

3.2.1 Geodesic Flow Kernel (GFK) method

Gong et al. proposed the Geodesic Flow Kernel (GFK), a kernel-based domain

adaptation method. The main idea behind the approach is to model the data with linear subspaces and to embed the source and target datasets in a Grassmann manifold, while also constructing a parametrized geodesic flow between the source and target points. The points along this flow represent the interpolating subspaces between the two domains. The raw features are projected into these intermediate subspaces to form infinite dimensional feature vectors $\mathbf{z}^\infty \in \mathcal{H}^\infty$, where $\mathcal{H}^\infty$ is an infinite dimensional feature space. Inner products in $\mathcal{H}^\infty$ define a kernel function that can be computed efficiently over the original feature space in closed form.

Constructing the geodesic flow

The source and target domain are modeled by linear subspaces of dimension d, and embedded onto a Grassman manifold $G(d, D)$, which is the collection of all the d-dimensional subspaces of the feature vector space $\mathbb{R}^D$.

Let $\mathbf{P}_S, \mathbf{P}_T \in \mathbb{R}^{D\times d}$ denote the basis of the principal component analysis subspaces for the source and target domains, respectively. The geodesic flow $\{\Phi(t) : t \in [0, 1]\}$ parametrized by t, between the two subspaces $\mathbf{P}_S$ and $\mathbf{P}_T$ on the manifold $G(d, D)$ is a path connecting the source and target subspaces. Each point on the path is an intermediate subspace, which in the beginning of the flow is more similar to the source subspace, and towards the end of the flow is more similar to the target subspace.

Then, for the extreme values of t, $\Phi(0) = \mathbf{P}_S$ and $\Phi(1) = \mathbf{P}_T$. For all other values of t,

$$\Phi(t) = \mathbf{P}_S\mathbf{U}_1\mathbf{\Gamma}(t) - \mathbf{R}_S\mathbf{U}_2\mathbf{\Sigma}(t) \tag{3.1}$$

where $\mathbf{R}_S \in \mathbb{R}^{D\times(D-d)}$ denotes the orthogonal complement to $\mathbf{P}_S$, $\mathbf{U}_1 \in \mathbb{R}^{d\times d}$ and $\mathbf{U}_2 \in \mathbb{R}^{(D-d)\times d}$ are orthonormal matrices obtained from the following singular vale decompositions,

$$\mathbf{P}_S^T\mathbf{P}_T = \mathbf{U}_1\mathbf{\Gamma}\mathbf{V}^T, \quad \mathbf{R}_S^T\mathbf{P}_T = -\mathbf{U}_2\mathbf{\Sigma}\mathbf{V}^T \tag{3.2}$$

The diagonal matrices $\mathbf{\Gamma}, \mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ have elements $\cos\theta_i$ and $\sin\theta_i$ for $i = 1, 2, ..., d$. The θ_i are called the principal angles between subspaces P_S and P_T and measure how much the subspaces overlap.

Computing the kernel

The geodesic flow can be seen as a smooth interpolation between the source and target subspaces. A point on the flow for a particular $t \in [0, 1]$ is an intermediate subspace $\Phi(t)$. Then, $\Phi(t)^T\mathbf{x}$ is the projection of a feature vector $\mathbf{x}$ into this subspace. These projections can be used to build a classifier.

When the original feature vector $\mathbf{x}$ is projected into each of the subspaces that form the geodesic flow, an infinite dimensional feature vector is obtained,

$$\mathbf{z}^\infty = \{\Phi(t)^T\mathbf{x} : t \in [0, 1]\} \tag{3.3}$$

This new feature representation is insensitive to the particularities of either domain, that is it forces the classifier to use domain-invariant features.

To make the use of this new feature representation computationally feasible the geodesic flow kernel is introduced. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be two original, raw feature vectors whose projections into $\Phi(t)$ for a continuous $t \in [0, 1]$ are computed and concatenated into infinite dimensional feature vectors $\mathbf{z}_i^\infty$ and $\mathbf{x}_j^\infty$. The geodesic flow kernel is defined by the inner product between these two vectors,

$$\mathbf{G}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{z}_i^\infty, \mathbf{z}_j^\infty \rangle = \mathbf{x}_i^T \int_0^1 \Phi(t)\Phi(t)^T dt \mathbf{x}_j = \mathbf{x}_i^T \mathbf{G} \mathbf{x}_j \tag{3.4}$$

where $\mathbf{G} \in \mathbb{R}^{D \times D}$ is a positive semidefinite matrix. The matrix G can be computed in closed form,

$$\mathbf{G} = \begin{bmatrix} \mathbf{P}_S\mathbf{U}_1 & \mathbf{R}_S\mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}_1 & \mathbf{\Lambda}_2 \\ \mathbf{\Lambda}_2 & \mathbf{\Lambda}_3 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T\mathbf{P}_S^T \\ \mathbf{U}_2^T\mathbf{R}_S^T \end{bmatrix} \tag{3.5}$$

where $\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \mathbf{\Lambda}_3$ are diagonal matrices whose elements are expressions involving the principal angles.

The only free parameter present in the geodesic flow kernel method for domain adaptation is the dimensionality of the subspaces. In this method $d = d_S = d_T$, and the problem of estimating this dimensionality is addressed in a subsequent section.

3.2.2 Subspace Alignment (SA) method

In [30] and [15] Fernando et al. proposed the subspace alignment method for domain adaptation. In this method the data from the source and target datasets is modeled with linear subspaces. The main idea behind the approach is to optimize a mapping function that aligns the source subspace to the target subspace, producing a domain-invariant feature space.

Subspace generation

Let d be the intrinsic dimensionality of the source and target subspaces. The basis of the subspaces $\mathcal{X}_S$ and $\mathcal{X}_T \in \mathbb{R}^{D \times d}$ are selected by performing PCA on the source and target data, and finding the eigenvectors that correspond to the d largest eigenvalues.

Learning the mapping function

In this method a the basis vectors of subspace $\mathcal{X}_S$ are aligned to the basis vectors of the target subspace $\mathcal{X}_T$ by a using a linear transformation matrix $M \in \mathbb{R}^{d \times d}$. The matrix M is learned by minimizing the Bregman matrix divergence,

$$M^* = argmin_M(\|X_S M - X_T\|_F^2) \tag{3.6}$$

where the norm used in the equation is the Frobenius norm. As the Frobenius norm is invariant to orthonormal operations, the equation above can be rewritten

more conveniently as

$$M^* = argmin_M(\|X_S^T X_S M - X_S^T X_T\|_F^2) \quad (3.7)$$
$$= argmin_M(\|M - X_S^T X_T\|_F^2) \quad (3.8)$$

This means that the optimal linear transformation M* is obtained when $M^* = X_S^T X_T$.

Matrix M can now be used to transform the source subspace and align it with the target subspace, so that the target aligned source coordinate system $X_a = MX_T = X_S^T X_T X_T$ can be obtained. This transformation permits to compare the source and target data instances in their respective subspaces. Next, when using a nearest neighbor classifier, the source data is projected into the target aligned subspace by using X_a and classification is performed in this d-dimensional domain invariant feature space.

3.3 Subspace Dimensionality Estimation Techniques

The only free hyper-parameters in the DA methods discussed above is the dimensionality d of the considered domains subspaces. In what follows, some techniques for estimating the intrinsic dimensionality of the feature space are outlined.

3.3.1 Eigenvalue-based estimation (EIG)

This is the standard statistical technique used for dimensionality reduction. In order to estimate the dimension d, principal component analysis (PCA) is performed on the data. Then, d is selected by finding the smallest number of most significant eigenvectors for which the projected data retains 99% of the variance.

With this technique the source and target datasets are considered separately, which implies that in general $d_S \neq d_T$.

3.3.2 Maximum likelihood estimation (MLE)

Fernando et al. propose to use in conjunction with their Subspace Alignment domain adaptation method an efficient dimensionality estimation technique presented in [31] that aims to retain the sample local neighborhood after dimensionality reduction. The intrinsic dimensionality of the domain is estimated to be the average of the maximum likelihood estimation (MLE) of the dimensionality for each data point.

In order to derive the maximum likelihood estimator of the intrinsic dimension from i.i.d. observations $\mathbf{x}_i$ for $i = 1, ..., n$ in $\mathbb{R}^D$, it is first assumed that the observations represent an embedding of a lower dimensional sample. So, an observation $\mathbf{x}_i$ can be written as $\mathbf{x}_i = \phi(\mathbf{z}_i)$ where $\mathbf{z}_i$ are sampled from an unknown density function f on $\mathbb{R}^d$, with unknown $d \leq D$ and ϕ is a continuous and sufficiently smooth mapping. With this assumption, it is ensured that lower dimensional instances that are close in $\mathbb{R}^d$ are mapped to close neighbors in the embedding.

The idea is then to fix a data point $\mathbf{x}$ and assume that $f(\mathbf{x})$ is constant in a small sphere of radius R around $\mathbf{x}$. Then, the inhomogeneous binomial process which counts observations within a distance $t \leq R$ from $\mathbf{x}$ as a Poisson process.

Under these assumptions it can then be shown that the maximum likelihood estimate of the dimensionality for instance $\mathbf{x}$ is:

$$\hat{d}(\mathbf{y}) = \left[\frac{1}{N(R, \mathbf{x})} \sum_{j=1}^{N(R,\mathbf{x})} \log \frac{R}{\Phi_j(\mathbf{x})} \right]^{-1} \tag{3.9}$$

where $\Phi_j(\mathbf{x})$ is the distance from observation $\mathbf{x}$ to its j-th nearest neighbor, and R is set to the mean pair-wise distance between the observations.

Finally, the intrinsic dimensionality of the considered domain is obtained as the average MLE dimension over all the instances

$$d = \frac{1}{n} \sum_{i=1}^{n} \hat{d}(\mathbf{x}_i) \tag{3.10}$$

Again, the two datasets are considered separately and a different dimension must be computed for each domain.

3.3.3 Subspace disagreement measure (SDM)

In the context of unsupervised domain adaptation, it is desirable to have a way of selecting the optimal dimensionality d of the data in an automatic way. To address this problem, in [16] Gong et al. proposed a subspace disagreement measure (SDM) which is optimized in order to choose d. The SDM evaluates the similarity between the subspaces of the source, target and combined source+target datasets.

In order to compute the SDM, the PCA subspaces of the source, target and combined datasets are computed. If the source and target datasets are similar, the three computed subspaces should be close to each other on the Grassmannian. This intuition is formalized in terms of the principal angles as

$$\mathcal{D}(d) = 0.5[\sin \alpha_d + \sin \beta_d] \tag{3.11}$$

where α_d is the d-th principal angle between source and combined subspaces, and β_d is the d-th principal angle between the target and combined subspaces.

The optimal d* is selected as

$$d^* = \min\{d|\mathcal{D}(d) = 1\} \tag{3.12}$$

That is, the optimal d* should be as high as possible without letting the two subspaces have orthogonal directions (i.e., $\alpha_d = \beta_d = \pi/2$).

Unlike in the previous two dimensionality estimation techniques, it allows to compute a shared dimension for the source and target datasets.

3.4 Temporal Gap as a Visual Domain Shift

Two sets of images that are taken with a large time lag in between can be considered as belonging to two domains that have different underlying data distri-

butions. Thus, classifying historical images proves to be a candidate application for the domain adaptation framework, as was first proposed by Fernando et al. in [8].

Many institutions such as universities and museums, as well as city archives have started cultural heritage campaigns that promote the collection and digitalization of historical images and footage. Ancient photographs and images often depict buildings, monuments, statues, fountains and other popular landmarks of towns and cities. This historical visual data can be exploited in many ways, and machine learning is thought to be a well suited approach for doing so automatically.

Cinema is an important part of the culture of the 20th century. Many iconic ancient movies were shot in outside locations of cities around the world. Annotation of many hours worth of film heritage is a labor intensive activity that would greatly benefit from machine learning methods, allowing the development of many creative and educational tools for present and future generations.

Visual data can vary considerably from domain to domain. Typical causes of visual domain shift include changes in the camera, image resolution, lighting, background, viewpoint, and post-processing. All of these aspects are markedly present when considering images taken at different time periods, indeed ancient images have different colors, texture and contrast characteristics when compared to modern digital images [32]. Additionally, the alteration in the locations themselves, due to the evolution of urban planning, and other changes in the surroundings contributes to the domain shift, it can even be the case that some viewpoints are no longer accessible.

The drop in the classification performance mentioned in section 3.1 can be expected to be particularly significant when the domain shift between training and test sets is as acute as in the case of time lags.

In this thesis existing approaches to domain adaptation are evaluated on the setting of classifying historical visual data, including frames from ancient movies. The long term goal is that of automatically annotating historical images, footing and movies which can then be used to create rich cultural heritage maps with

didactic or tourism purposes.

Chapter 4

Experimental Setup and Results

In this chapter a description of the datasets that were used in this thesis and the experimental setup on the task of location recognition over large time lags are provided.

4.1 Datasets

The algorithms and experimental setups studied in this work were tested on two different datasets. In both of the considered datasets, the domain shift between the source and the target distributions is the consequence of the passage of time.

Apart from the usual and well known challenges of location recognition, including factors such as illumination conditions, occlusion, and viewpoint variation, the datasets reflect issues that arise when considering color degradation, changes in the image acquisition process across time, and changes in the physical places themselves.

In what follows there is a brief description of each dataset.

4.1.1 Large Time Lags Location recognition dataset (LTLL)

The Large Time Lags Locations (LTLL) recognition dataset was introduced in [8]. The dataset consists of a set of ancient images and a matching set of modern ones.

It contains pictures from 25 locations around the world. These include well known landmarks in various European and Asian towns and cities such as Paris, London, Leuven, Agra, Colombo and Kandy, for which both old and new pictures were easily available from the web. Twelve of the locations are in the municipality of Merelbeke in Belgium, as historical pictures of these locations were provided by the city archive of Merelbeke. The corresponding modern images for these locations were obtained by the collectors of the dataset from Flickr, Google Street-View and the Google Images search engine.

In total the dataset is made up of 225 ancient images and 275 modern images.

Figure 4.1: Sample of the modern images in the LTLL dataset.

Figure 4.2: Sample of the ancient images in the LTLL dataset.

4.1.2 Rome Memory Project dataset

This dataset consists of a set of old images taken from the 1945 italian movie Rome, Open city *(Roma città aperta)*. There is a total of 16 images spread

across different locations in Rome. The images are annotated with the their corresponding geographical coordinates.

The annotated images were graciously provided by the E-learning Laboratory of La Sapienza University.

For the development of this thesis the dataset from the ancient movie was enriched by adding a set of modern time images matching the locations from the movie, that were downloaded from Google Street View.

For each Rome location from the movie, 72 modern images were added to the dataset. This was accomplished in the following way: given some geographical coordinates, 8 images were downloaded for this location, corresponding to 8 different heading values. Then moving a few meters in each of these directions the process was repeated.

This collection method seeks to ensure that scene of interest (that of the old movie) will be present in at least some portion of the modern images.

Figure 4.3: Dataset of images from the 1945 Italian film Rome, Open City

4.2 Experimental Setup

This section presents a detailed description of the experimental setup used to test the proposed approaches to automatically annotate historical images.

The task of location recognition over large time lags is examined under an image classification framework. The experiments are carried out for two different datasets, the LTLL and the Rome, Open City datasets, both described in section 4.1.

For all the experiments, a 1-Nearest Neighbour classifier is employed. The modern images comprise the *source* set, and are labeled with one of the considered locations. All the images in the source set are used for training the classifier. The ancient images comprise the *target* set, on this set the classifier is tested by annotating the images with one of the location labels.

The reported final performance is the multi-class classification accuracy, that is, the average of the accuracies of all the categories obtained over the whole test set of ancient images.

The image features used in the experiments are CNN features. Two different types are evaluated, the *Caffe + Imagenet* and the *Caffe + CNNplaces* features. The 7th layer of the convolutional neural network is chosen in both cases. The features are originally 4096-dimensional vectors. This dimension is reduced to an estimated intrinsic dimensionality which is computed by three different subspace dimensionality estimation techniques: PCA retaining 99% of the variance (EIG), subspace dimensionality measure (SDM) and maximum likelihood estimation (MLE).

For a given dimensionality, three classification experiments are performed. A simple 1-Nearest Neighbor classification without any adaptation scheme, and classification using two domain adaptation methods, the geodesic flow kernel (GFK) and subspace alignment (SA).

4.3 Classification Results

4.3.1 Results on the LLTL dataset

Classification results **Caffe + Imagenet** FC 7 layer features:

Dimensionality		No-Adapt	GFK	SA
EIG (source)	28	41.18	52.45	53.43
EIG(target)	36	41.18	52.20	52.45
SDM	19	41.18	50.00	50.98
MLE(source)	32	41.18	50.00	55.39
MLE(target)	37	41.18	51.47	52.45

Table 4.1: Performance (Acc.all %) of a 1-Nearest Neighbour classifier for different domain adaptation methods and different subspace dimensionality estimations. The full source dataset was used in training. *Caffe + Imagenet* FC7 layer features.

Classification results **Caffe + CNNplaces** FC 7 layer features:

Dimensionality		No-Adapt	GFK	SA
EIG (source)	11	48.04	52.94	47.55
EIG(target)	11	48.04	52.94	47.55
SDM	20	48.04	51.47	51.96
MLE(source)	16	48.04	57.84	51.47
MLE(target)	19	48.04	53.92	51.96

Table 4.2: Performance (Acc.all %) of a 1-Nearest Neighbour classifier for different domain adaptation methods and different subspace dimensionality estimations. The full source set was used in training. *Caffe + CNNplaces* features.

4.3.2 Results on the Annotation of Ancient Movies Location

Classification results **Caffe + Imagenet** FC 7 layer features:

Dimensionality		No-Adapt	GFK	SA
EIG (source)	7	6.25	6.25	18.75
EIG(target)	13	6.25	18.75	18.75
SDM	178	6.25	6.25	6.25
MLE(source)	30	6.25	6.25	6.25
MLE(target)	15	6.25	12.5	6.25

Table 4.3: Performance (Acc.all %) of a 1-Nearest Neighbour classifier for different domain adaptation methods and different subspace dimensionality estimations. The full source dataset was used in training. *Caffe + Imagenet* FC7 layer features.

Classification results **Caffe + CNNplaces** FC 7 layer features:

Dimensionality		No-Adapt	GFK	SA
EIG (source)	4	18.75	18.75	6.25
EIG(target)	10	18.75	18.75	25.00
SDM	77	18.75	25.00	31.25
MLE(source)	15	18.75	18.75	31.25
MLE(target)	14	18.75	18.75	25.00

Table 4.4: Performance (Acc.all %) of a 1-Nearest Neighbour classifier for different domain adaptation methods and different subspace dimensionality estimations. The full source set was used in training. *Caffe + CNNplaces* features.

4.4 Observations and Analysis

The classification results of all the experiments are shown in Table 4.1, Table 4.2, Table 4.3 and Table 4.4. The results show that the location recognition over large time lags task benefits from domain adaptation approaches, as in all cases the performance of either of the considered DA methods is better than having no adaptation.

For the experiments carried out on the LTLL dataset, the best performance is 57.8%, achieved by the Geodesic Flow Kernel (GFK) domain adaptation method with an intrinsic dimensionality of 16 estimated by MLE on the source set, a considerable reduction from the original space dimensionality of 4096.

Method	Performance %
HA-rSIFT-FV + SA	56.1
Caffe + Imagenet + SA	55.4
Caffe + CNNPlaces + GFK	57.8

Table 4.5: Comparison of best classification rates. The first line refers to the work of Fernando et al. in [8]

.

Table 4.5 compares the classification performance of different settings. The first line reports the best result obtained in [8] by using the Hessian Affine (HA) as a feature detector [33], root-SIFT (rSIFT) [34] as the descriptor and Fisher Vectors (FV) [35] as the representation. The result is outperformed by the GFK method in combination with *CNNPlaces* features. This reinforces the idea that the features extracted from a convolutional neural network should be a primary candidate for visual recognition tasks.

The annotation of ancient movies proved to be a more challenging problem. This is to be expected, as the frames often have less advantageous viewpoints than the old postcards and historical images, and due to the prominent presence of people.

In the Rome, Open City dataset, the SA method seems to perform better

on average than GFK. Again the MLE dimensionality estimation performed on the source set gives the best result. On this dataset there is a wider range of dimensionality estimations, in one extreme case d is reduced from 4096 to 4.

It is important to note that for these experiments chance performance was equal to 6.25%, and without domain adaptation, the *Caffe + Imagenet* features had no discriminative power at all, while the *Caffe + CNNPlaces* still performed better than chance.

We conclude the section by pointing out that the *PlacesCNN* features are better suited than the *Caffe + Imagenet* for the task of location recognition. This is not surprising, as the deep model was trained on a scene-centric database, which is more closely related to the task at hand.

Chapter 5

Conclusions

5.1 Conclusions

In this thesis the task of visual location recognition of ancient photographs and movie frames was approached by placing it under a supervised image classification framework. The task was to annotate the ancient images with a location label using a dataset of modern images covering the same set of pre-defined locations.

The temporal gap between the images was modeled as a visual domain shift, and domain adaptation methods were used. In particular, the Geodesic Flow Kernel (GFK) and Subspace Alignment (SA) methods were evaluated, and were shown to attenuate the drop in classification performance that this distribution shift produced.

Features extracted from the activation of a deep model that was trained on a task that is more closely related to the one being considered translate better to the new domain.

5.2 Future Work

Only unsupervised domain adaptation methods were considered in this thesis. A future direction of research would be to consider a small number of labeled images from the target set and investigate how much the task would benefit from this additional effort.

The estimation intrinsic dimensionality of the subspaces used to model the visual domains varies a lot from one technique to another, perhaps a cross-validation scheme for choosing this parameter could be explored in future research.

In the task of annotating ancient movies, the choice of frame is crucial, as urban outdoor scenes are affected by large appearance and viewpoint variation between the target view and training dataset and presence of large number of non-discriminative structures due to vegetation, sky walls and roads. Having larger datasets would attenuate the influence of uninformative frames.

List of Figures

List of Tables

Bibliography

[1] Soonmin Bae, Aseem Agarwala, and Frédo Durand. Computational rephotography. *ACM Trans. Graph*, 29(3):1–15, 2010.

[2] Yunpeng Li, Noah Snavely, and Daniel P Huttenlocher. Location recognition using prioritized feature matching. In *Computer Vision–ECCV 2010*, pages 791–804. Springer, 2010.

[3] Petr Gronat, Guillaume Obozinski, Josef Sivic, and Tomas Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 907–914. IEEE, 2013.

[4] Song Cao and Noah Snavely. Graph-based discriminative learning for location recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 700–707. IEEE, 2013.

[5] Akihiro Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 883–890. IEEE, 2013.

[6] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–7. IEEE, 2007.

[7] Mathieu Aubry, Bryan C Russell, and Josef Sivic. Painting-to-3d model alignment via discriminative visual elements. *ACM Transactions on Graphics (TOG)*, 33(2):14, 2014.

[8] Basura Fernando, Tatiana Tommasi, and Tinne Tuytelaars. Location recognition over large time lags. *Computer Vision and Image Understanding*, 139: 21–28, 2015.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[11] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[13] Anna Margolis. A literature review of domain adaptation with unlabeled data. *Rapport Technique, University of Washington*, page 35, 2011.

[14] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: An overview of recent advances. *Submitted*, 1: 1–8, 2014.

[15] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Subspace alignment for domain adaptation. *arXiv preprint arXiv:1409.5241*, 2014.

[16] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012.

[17] Edward Johns and Guang-Zhong Yang. From images to scenes: Compressing an image cluster into a single scene model for place recognition. In *Computer*

Vision (ICCV), 2011 IEEE International Conference on, pages 874–881. IEEE, 2011.

[18] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 667–674. IEEE, 2011.

[19] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014.

[21] Antonio Torralba, Alexei Efros, et al. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.

[22] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[23] Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Algorithmic Learning Theory*, pages 139–153. Springer, 2012.

[24] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[25] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88. ACM, 2007.

[26] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[27] Amaury Habrard, Jean-Philippe Peyrache, and Marc Sebban. A new boosting algorithm for provably accurate unsupervised domain adaptation. *Knowledge and Information Systems*, pages 1–29, 2015.

[28] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010*, pages 213–226. Springer, 2010.

[29] Abhishek Kumar, Avishek Saha, and Hal Daume. Co-regularization based semi-supervised domain adaptation. In *Advances in neural information processing systems*, pages 478–486, 2010.

[30] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2960–2967. IEEE, 2013.

[31] Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2004.

[32] Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *Computer Vision–ECCV 2012*, pages 499–512. Springer, 2012.

[33] Michal Perd'och, Ondrej Chum, and Jose Matas. Efficient representation of local geometry for large scale object retrieval. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 9–16. IEEE, 2009.

[34] Zhenhua Wang, Bin Fan, and Fuchao Wu. Local intensity order pattern for feature description. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 603–610. IEEE, 2011.

[35] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.