I. Introduction

I.1 Motivation and Trajectory

The turn of the XXI century came with an explosion of multimedia content in the world. Digitalization of contents became a reality and it began with music digitalization when the standard technology changed from analog cassettes to digital Compact Discs. Still photography was also digitalized and personal cameras were popularized. Therefore, the growth of personal computer sales was a fact as users needed a PC to store and manage these new digital images. In parallel, advances were made in the video camera recorders. These evolved from low resolution cameras which recorded video on integrated DVDs to FullHD video files stored in high speed and capacity memory cards.

In addition to hardware and market evolution, internet and access to the internet also evolved accordingly. In 2000, 56kpbs modems were the trend in home communications, until it evolved to ADSL technology or more recent Fiber–To-The-Home.

Internet services matured on this accessibility evolution. In the music area Napster appeared in 1999 and people from the entire world were able to download a song in minutes and thereby transforming the way of to consume music. Services that allowed users to store their digital images on the cloud were popularized by Flickr in 2004. Video also had its own evolution that started with the uploading of videos to personal websites and evolved to major video sharing platforms as YouTube that was born in 2005.

As a user, I personally took active part in all these changes. My family bought me my first analog camera in 1996. Then my next camera was a digital camera with a resolution of 1.3Mp (4:3) with single-lens optics which was a gift to me in 2004. Less time passed until I had the next camera, an 8Mp camera in 2007 and in 2012 I bought my latest camera with 12Mp resolution and better optics. In so far as video is concerned, my first digital camera was not able to capture video but the following ones did , and they captured video at 2Mp resolution and a 720p, respectively. Now a new revolution is under the hood. The mobile devices are even more popular than the traditional PC, and they include still photography and full HD video capabilities and additionally they are always connected.

As digital cameras became popular and taking a picture was free of charge, every time I used a camera I took hundreds of images. At that moment I was annoyed by the amount of images and the time I spent retrieving one specific image from my hard disk drive. I realized that the problem could be solved if I used a software or internet service which

allowed me to tag every single photo with my own labels. At first this idea looked right to me but then I found out that it was useless: *if I wanted to search a label that I didn't use during the image tagging process, how could I search for it?*

In my personal life I started to study Telecommunications Engineering and I realized that I liked computers to perform tasks automatically. At that moment I thought about working on Artificial Intelligence, but later, during my Final-Year-Project, I realized that this was a huge field and I started to work in the sub-field of Computer Vision or, in other words: *teaching computers to see*. During that project my advisor, Dr. Koldo Espinosa, told me that if I wanted to develop novel ways for a computer to see I would need to become a researcher because there were so many things unsolved that I would need to learn the insights of the problem, study the state of the art and propose my own methods. This led me to start working in a Research Institution and I also started my personal project: obtaining a doctorate in research.

During my first years as researcher I found that there were many unsolved problems so companies couldn't rely on many computer vision algorithms. One company approached us asking for a system to search for content in videos automatically. That presented several problems. First, video analysis was very computationally intensive and, at that time, it was not possible to build a computing cluster using commodity hardware. Video processing algorithms were not also as mature as they are now. In addition, image processing capabilities were limited: face detection and person detection were the only two major accomplishments achieved by the research community. Lot of work was also taken in detecting a small set of single object in simple images but that wasn't enough for our customer.

In late 2010, I started to think about my thesis topic and everything came to my mind. In my adolescence I had problems searching for my own images in my computer. Internet services and search engines were also unable to search images by its content. Companies needed novel algorithms to retrieve visual information but the research community was mainly focused on solving small scale problems. All this generated the idea of developing improvements in the state of the art with the goal of solving these problems and all were related to one: *retrieve images based on its content*. As I said, the idea was not to create a completely novel approach, but to improve some parts that the research community missed.

In order to start with the thesis I understood that the key problem of image retrieval was automatic image annotation. If you have an algorithm that annotates the images with the most significant words, then a text based search engine will be smart enough to retrieve the most interesting tags (a lot of research is taking place, like *query expansion*, etc).

In order to annotate images we need a couple of things: a mathematical descriptor of the images and an algorithm that performs the annotation. So I started to analyze the state of the art in these two areas and I found the basis of my thesis: the most promising approach in 2010 was based on image annotation by searching similar images and then propagating the tags from the most similar images to the query image.

This sounds simple but a lot of problems arise, as stated in a seminal work by Makadia et al [MAKA10]: you need to describe the images with a compact descriptor, you need to perform near real time similarity searches on a big database of images and you need to transfer the tags. Several methods were proposed on each part and I analyzed most of them, which led me to have a deep knowledge about the state of the art and allowed me to write several review papers and invitations to symposia. This generated one of the main pillars of this thesis.

Regarding the step of image description, we found that if you describe your images using MPEG7 based descriptors, you can yield the same performance as state of the art but at lower computational and storage cost. But the MPEG7 descriptors were peculiar. The Scalable Color Descriptor was based on human perception of color, but the Edge Histogram Descriptor was based on man-made texture filters. At that time I asked myself: *Why didn't they use a Human Based texture descriptor?* I found that it was not an easy solution as a valid cortex model for this purpose did not exist. Then I asked myself again: *Could we achieve better results using a bio-inspired texture descriptor?* At first, we didn't know so we started to work in this research line taking the foundations of the retinal and color processing from my mentor Dr. Estibaliz Garrote, which become the second main pillar of this thesis.

During my analysis of the state of the art, I also found that state of the art approaches to transfer tags were complex but they didn't include any visual information at all. In this case, I also had a question: *If the research community is working with images, why did they not use image features to propagate the tags?* That question led to me to contact

Prof. Lorenzo Torresani, leader of the Visual Learning Group at Dartmouth College (USA), and to explore during 4 months a novel technology to combine textual and visual data in order to generate rankings. With such knowledge as the inception, we started to work in a novel approach to validate if using the textual and the visual features during the tag transference stage achieves better results than using only textual features. This defined the third pillar of this thesis.

I.2 Introduction to the field

Nowadays we are living in a digital world, where the number of sensors, electronic devices and users increase every second. In 2014 more internet connected mobile phones are expected than computers are expected [MART13]. Every single digital process or every information exchange produces huge amounts of data which is added to existing data collections. This production generates several problems related to the management of such vast amounts of data. The magnitude of the problem is huge and it can be seen by the fact that the data generated during two days in 2010 is greater than the accumulated data since the origin of the civilization to 2013 [TECH10].These magnitudes scare so there are lots of efforts from the scientific community to work on solving problems generated by this *data wave*, from the bottom layer (e.g. storage of the data) to the top layer (e.g. speech recognition).

Depending on its nature, data can be classified into two main groups: structured data, which follows a model that gives meta-information about data and helps during the processing; and unstructured or raw data, which does not have any pre-defined structure or any meta-information, so it is harder to analyze. In this second group we can find multimedia data and specifically images.

“A picture is worth a thousand words". This old saying is a trend in the modern digital era. .Actually, images are one of the most important pieces of exchanged data. This can be seen in the huge amount of web services and mobile applications related to the topic, from the most traditional services, like Flickr or image search engines, to the most novel applications that let the user to modify the pictures or automatically process them to achieve a better aesthetics. One specific example of the increase of images on the internet in the last few years is that the number of new images uploaded to *Instagram* each day is more than 60 Million [INST14], which is 2.5 million images uploaded per hour.

Despite being one of the data type most relevant to users, search engines are unable to handle this data correctly [RODR14]. This means that even when users are uploading their photos to the cloud, there is no way that a user can retrieve images in a user friendly system. Most services allow users to incorporate textual labels or tags, so in order to retrieve one image users can use a search engine and query it with the tag they desire. The problem occurs when a user doesn't add any label, which is very common as users can upload several hundred photos and tag only some of them with sparse tags.

The solution to this problem was proposed in early 1990. It consisted on an automatic algorithm that analyzed the content of the images and assigned labels related to the content. The concept is clear but the implementation is not so obvious. Computer vision algorithms are needed to analyze the content of the images, but these algorithms are not yet mature enough to extract information from all the objects around the world.

Most of the algorithms proposed by the computer vision community are related to the fields of statistics, mathematics and machine learning as a whole. This means that they are based on empirical measurements of the current world, so they need huge amounts of information to model all the possible objects in the world. But human beings' brain contains complex architectures that are able to learn this kind of knowledge with less information, so using bio-inspired models to process real world data is a clear advantage [GARR11].

Until now, lots of works have been carried out in the field of automatic image annotation and most of them have improved the state of the art. Now, some of these algorithms are spreading across the industry and future opportunities are appearing .Searching of multiple objects in a single image is still a challenge, and video analysis in real time, detecting actions, emotions, concepts,… will be one of the key future research lines.

I.3 Aim of the Thesis

This thesis is related to the automatic image annotation task. The main idea of this job is how to generate the labels that most likely describe the content of a particular photo.

To this end, multiple approaches have been proposed and are analyzed in chapters II and III. This analysis has led to the conclusion that a Nearest Neighbor based model is the main baseline to be considered.

Considering that baseline, the aim of this thesis is to propose:

- A **novel mathematical description** of the content of a natural image based on an animal's visual system.
- A **novel tag transference algorithm** that relies on visual and textual information to transfer the most relevant tags from a set of possible tags.

Therefore, in order to fulfil the aim, the following elements must be looked at:

- Study of the state of the art: the automatic image annotation field involves lots of technologies from the computer vision community, multimedia community and also from other related fields like machine learning. An in-depth study of all the technologies must be performed, but also a detailed study of the image annotation field, as lots of research groups are participating as it is one of the most extensively researched fields.
- Study of color and texture descriptors: Edges and color are two of the elements that compose an image, so it is clear that low level visual descriptors that analyze such information need to be carefully looked at.
- Study of the physiological component involved in the process of perception and processing of color and texture information in an animal's visual system.
- Modelling the relevant components of the processing chain of an animal's visual system.
- Experiments running and testing of the output obtained by the models. These models must be compared with biological responses of an animal's visual system but also with other proposed bio-inspired approaches.
- Modelling and evaluation of machine learning techniques to combine textual and visual information.

I.4 Content of the Thesis

This thesis is divided into 7 chapters that provide a detailed description of the different areas included in this work.

Chapter I is devoted to the personal motivation of this work and the introduction to the topic. It also presents the objectives to be achieved and the steps to be performed.

Chapter II provides a general framework of the knowledge in which to develop this research. General information on base technologies like visual descriptors and similarity

search techniques is provided. A detailed perspective of the field of image annotation is also presented and divided into the main three research lines, which are generative models, discriminative models and nearest neighbor based models.

Chapter III is dedicated to an in-depth analysis of the main trends in the state of the art presented in the previous chapter. To this end, a common methodology composed by common image databases and common validation metrics are proposed and followed thorough the thesis. As the result of such analysis in Chapter III a baseline that represents the current state of the art is proposed, so in the rest of the thesis all the experiments are done comparing the proposals against the baseline.

Chapter IV is focused on how texture information is extracted and represented in a primate's brain. This chapter studies the different parts of the visual system and specially looks at the study of the primary visual cortex, its neural structure and functionality, as it is in charge of basic texture information extraction and representation. It also presents the existing functional and computational models of the cortex, and it shows the proposal of several models which mimic the real effects generated in a macaque's visual cortex.

Chapter V presents how the proposed cortex models can be used to make a new implementation of the MPEG7 standard, and more particularly the Edge Histogram Descriptor. A detailed proposal of integration is exposed and then a fine tunning of the biological parameters is performed leading to a visual descriptor that achieves better results than the standard MPEG7 and the state of the art.

Chapter VI looks at a different point in the image annotation chain, and more particularly it is focused in the label transference models. In this chapter a novel tag transference algorithm is proposed which uses visual information and textual information to propagate the tags. In addition a fast and adapted training algorithm is proposed, obtaining an accurate result with less overhead in real time queries.

Finally, Chapter VII sets out the final conclusions and main contributions of this thesis as well as proposals for future works.