Introduction

As suggested by the title, this body of work is targeted on giving an automatic answer to a seemingly simple question: "What are you looking at?"

Humans often believe that they are incredibly good at answering this question. To convince the reader that this is a wrong belief, let us consider a simple example scenario in which we want to estimate the gaze point of a person which is sitting in front of a computer screen (Figure 1.1(a)). We are standing in a position in which the face of the person is clearly visible and, on our desk, we have an exact copy of the page that is displayed on the gazed screen (Figure 1.1(b)). When asked "what is this person looking at?", we will already struggle to roughly guess which side of the page is being gazed, while it will be basically impossible for us to name the exact gazed word, let alone the single character contained within it.

Although our gaze estimation capabilities are not as accurate as we would expect, we are still very confident about them. This is because we use them everyday, starting from the earliest cognitive developments when we were toddler.

Developmental psychology, in fact, argues that understanding the gaze of a person is fundamental in infants [14, 93, 15]: By learning how to make sense of the visual channel, infants start with recognizing the face of their caregiver, then learn to follow head movements, and finally learn to follow gaze directions. Later on, recognizing gaze enables an infant to engage in joint visual attention (i.e. sharing interest by looking at the same object), which in turn aids the infant in learning social, cognitive and communicative abilities. An example of this is given in [8, 11, 102], which argue that infants are facilitated to learn which utterance correlates to which object by understanding the gaze of the adult. Apart
from being useful in early learning, detecting the direction of another one’s gaze quickly becomes a crucial component of social interaction [3, 58, 65], as higher level cues about human behavior and intention (such as attention, eye contact, who is speaking to whom, cognitive state and non-verbal communication) can be extrapolated from it. For instance, gaze is a necessary ingredient for understanding the context and the full meaning of a displayed emotion [81, 89]. Examples occur when someone displays a scared face while gazing at a dangerous situation, or displays a happy face while gazing at his beloved one.

Since understanding gaze is fundamental for our development, especially for our human-human interaction, we seek to answer the question “what are you looking at?” in an automatic manner, so that it could be used to achieve more natural human-computer interaction.

### 1.1 Gaze Estimation and Inference

Two main cues are clearly involved in the estimation of the human’s gaze: the position of the head and the location of the eyes relative to it [66]. A number of studies have investigated which of the two cues is more important for gaze estimation. They found that the position of the head is often sufficient to determine the overall direction of attention [23, 34, 77], while the location of the eyes is used to fine tune attention and can hint to an additional layer of infor-
mation regarding thoughts, intentions, desires, and emotions. This is why, in many cultures, eyes are believed to be the "mirror of the soul".

Contrary to these findings, most automatic gaze estimation techniques in computer vision rely either on information about head pose [82] or eye location [44] in order to reduce the complexity of the problem. Furthermore, commercial eye-gaze tracking systems employ a calibration procedure and additional constraints (like restricting head movement or using head mounted cameras) to increase the accuracy and stability of the system. This often involves creating a direct mapping between the location of the eyes and a known position on the screen.

In the example scenario in Figure 1.1, using this technique would imply asking the subject to look at the corners of the text and to record the relative eye locations. Then use these locations as a reference for newly gazed locations. Furthermore, in order to avoid recalibrating at each head location, the subject would not be allowed to move from the calibrated position. Although the accuracy of our estimation would significantly improve with respect to the original rough gaze estimate, we can conclude that this procedure is very different than the way humans achieve the same task. But what is an alternative to this method?

According to [80] an ideal gaze estimator should:

- be accurate, i.e., precise to minutes of arc;
- be reliable, i.e., has constant, repetitive behavior;
- be robust, i.e., should work under different conditions, such as indoors and outdoors, for people with glasses and contact lenses;
- be non-intrusive, i.e., cause no harm or discomfort;
- allow for free head motion;
- not require calibration, i.e., instant set-up;
- have real-time response.

Existing gaze estimation systems forfeit some of these requirements for the sake of accuracy. On the other hand, as shown in the example scenario in Figure 1.1, the human gaze estimation system fulfills all of the requirements, except for high accuracy. Hence, here we want to investigate whether, by weakening the accuracy requirements (and therefore putting more focus on the usability requirements), it could be possible to develop a truly non-intrusive, more accessible and user-friendly gaze estimation system which is similar to the humans'.
To this end, in this work we will focus on appearance based methods only. Although they are considered as the less intrusive methods available, appearance based methods tend to be inaccurate. This is mainly due to the reduction of the amount of resolution available to capture both the head and the eye location information in a single frame.

Therefore, we argue that the gaze estimates obtained by the head pose and eye location should only be considered as a rough indication of the direction of interest, and that additional information about the gazed scene needs to be considered to improve accuracy. For instance, in the famous scene from the movie "Taxi Driver", the main character asks himself: "Are you talking to me?" (Figure 1.2), while definitely looking at himself in the mirror. He then turns around to check the rest of the scene. Although it is an acted scene, this behavior clearly hints that taking the context of the gazed scene into consideration is important to completely understand an uncertain gaze estimate. But how can the information about the gaze scene help in gaze estimation? Recalling the scenario in Figure 1.1, if the plain text in Figure 1.1(b) would be replaced by a small red ball on the right side of the image, it would be natural to assume that the subject is gazing at it, and the assumption would probably be correct. Therefore, we argue that the gazed scene needs to be inspected to infer the most likely gazed object in the scene to adjust uncertain gaze locations, obtaining an improvement in the accuracy of the gaze estimation system.
1.2 Objectives and Approach

To automatically answer the question "What are you looking at?", in this work we arrive at the following three main research objectives:

• **Detection:**
  In order to fulfill the requirements for an ideal gaze estimator, we need to restrict it to only use appearance information, as it is the sole non-intrusive system which can be used in different conditions. The difficulty here is to find a way to extract, starting from low resolution images of faces, information that is accurate enough to be used for gaze estimation. Furthermore, we need to investigate whether the inaccurate estimations of the head pose and eye location could be used together to reinforce each other in order to yield better overall results.

The first question we need to answer is how to perform accurate eye center location (i.e. within the area of the pupil) on low resolution images (i.e. captured by a simple webcam). Accurate eye center location can already be achieved using commercial eye-gaze trackers, but additional constraints (e.g. head mounted devices or chin rests) and expensive hardware make these solutions unattractive and impossible to be used on standard (i.e. visible wavelength), low-resolution images of eyes. Systems based solely on appearance (i.e. not involving active infrared illumination) are present in the literature, but their accuracy does not allow locating and being able to distinguish eye centers movements in these low-resolution settings. The question is discussed in Chapter 2.

A second question relates to head pose estimation. Head pose and eye location for gaze estimation have been studied separately in numerous works [82, 44]. They show that satisfactory accuracy in head pose and eye location estimation can be achieved in constrained settings. However, due to distorted eye patterns, appearance based eye locators fail to accurately locate the center of the eyes on extreme head poses. In contrast, head pose estimation techniques are able to deal with these extreme conditions, so they may be suited to enhance the accuracy of eye localization. Therefore, in Chapter 3 we consider the question whether a hybrid scheme to combine head pose and eye location information could achieve enhanced gaze estimation.

• **Estimation:**
  Once relevant cues are extracted, they need to be combined in a sensible way to generate a single estimate. However, the low resolution of the
images limits us in the construction of a geometrically accurate eye and head model. Therefore, the aim is to find a way to interpret and combine the obtained information: Instead of trying to geometrically combine the gaze vectors suggested independently by the head pose and by the eye location, in Chapter 3 we want to investigate whether these gaze vector could be combined in a cascade in order to reduce the problem of 3D gaze estimation into subsets of 2D problems.

- **Inference:**
  Finally, our last objective is to investigate an algorithm that, without using prior knowledge about the scene, will efficiently extract likely gaze candidates in the scene. As common salient features as edges, contrast and color only defines local saliency, in Chapter 4 we study whether it is possible to estimate global salient features starting from the local ones, and whether these features help into estimating interesting objects in the scene. To this end, in Chapter 4 we investigate a computational method to infer visual saliency in images.

  In Chapter 5 we will subsequently investigate whether it is possible to adjust the fixations which were roughly estimated by the gaze estimation device. We will test whether the approach can be efficiently applied to different scenarios: using eye tracking data, enhancing a low accuracy webcam based eye tracker, and using a head pose tracker.

Therefore, starting from finding the location of the eyes in a face and ending on finding the most probable object of interest in the gazed scene, the overall focus of this work is to investigate a unified way to improve each single step of the gaze estimation pipeline, in order to achieve a more accurate and natural gaze estimation system.