What are you looking at? Automatic estimation and inference of gaze
Valenti, R.

Citation for published version (APA):
Summary and Conclusions

6.1 Summary

In this thesis, we have analyzed important techniques for gaze estimation and explored ways to improve each of them. We have proposed a gaze estimation pipeline consisting of the following components:

- **Detection**: Accurately detect and extract eye centers and head pose information.
- **Estimation**: Combine head and eye location information to obtain a multi-cue visual gaze estimation system.
- **Inference**: Use information about the gazed environment to infer the most probable gazed object in the scene.

The results obtained in this thesis are discussed in the following paragraphs:

**Chapter 2: Accurate Eye Center Location through Invariant Isocentric Patterns.**

In Chapter 2 we argue that, although accurate eye center location can already be determined using commercial eye-gaze trackers, additional constraints and expensive hardware make these existing solutions unattractive and impossible to be used on standard (i.e. visible wavelength), low resolution images of eyes. Systems based solely on appearance are proposed in literature, but their accuracy does not allow to accurately locate and distinguish eye centers movements in these low resolution settings. Therefore, we investigate a fast method which uses circular symmetry based on isophotes, where centers of the oscu-
lating circles of the isophotes are computed from smoothed derivatives of the image brightness, so that each pixel can provide a vote for its own center. The use of isophotes in the proposed method yields low computational cost (which allows for real-time processing), robustness to rotation and linear illumination changes (contrast and brightness), and achieve (in-plane) rotational invariance. Furthermore, a scale space framework is used to improve the accuracy of the proposed method and to gain robustness to scale changes.

The proposed approach is extensively tested for accurate eye center location and robustness to changes in illumination, occlusion, eye rotation, pose and scale by using the BioID database, the color FERET database, the Yale Face B database and the CMU Multi-PIE database.

Chapter 3: Combining Head Pose and Eye Location Information for Gaze Estimation.

In Chapter 3, we discussed that head pose and eye location for gaze estimation are closely related topics, but their combination has not been studied in the literature. We argue that, in presence of non frontal faces, eye locators are not adequate to accurately locate the center of the eyes. On the other hand, head pose estimation techniques are able to deal with these conditions, hence they may be suited to enhance the accuracy of eye localization. Therefore, in Chapter 3, we propose a proper integration of a head pose tracker and the isophote based eye locator discussed in Chapter 2 in a complementary manner, so that both system could benefit from each other’s evidence.

In the proposed method, the transformation matrix obtained from the head pose is used to normalize the eye regions, while the transformation matrix generated by the extracted eye locations is used to correct the pose estimation procedure. The scheme is designed to (1) enhance the accuracy of eye location estimations, especially in low resolution videos, (2) to extend the operative range of the eye locators and (3) to improve the accuracy of the head pose tracker. The experimental results show that the proposed unified scheme improves the accuracy of eye estimations by 16% to 23%. Further, it considerably extends its operating range by more than 15°, by overcoming the problems introduced by extreme head poses. Also the accuracy of the head pose tracker is improved by 12% to 24%.

These enhanced estimations are then combined to obtain a gaze estimation system which uses both eye location and head information in order to project the visual gaze of a person on the gazed scene, by retargeting a set of known points on the gazed scene using the head pose information and pose-normalized eye locations. This allows the proposed method to estimate locations gazed by the
eyes at different head locations. The experimentation on the proposed combined gaze estimation system shows that it is accurate (with a mean error between $2^\circ$ and $5^\circ$) and that it can be used in cases where classic approaches would fail without imposing restraints on the position of the head.

**Chapter 4: Image Saliency by Isocentric Curvedness and Color.**
To locate the most probable gazed object, we need a way to find the most salient objects in the scene. Therefore, in Chapter 4, we propose a computational bottom-up model to detect saliency in images. The method is based on the idea that salient objects should have local characteristics that are different than the rest of the scene, being edges, color or shape. By using a novel operator, these characteristics are combined to infer global information. The obtained information is used as a weighting function for the output of a segmentation algorithm so that the salient object in the scene can easily be distinguished from the background.

The proposed approach is fast and it does not require any learning. The experimentation shows that the system can enhance interesting objects in images and it is able to correctly locate the same object annotated by humans with an F-measure of 85.61% when the object size is known, and 79.19% when the object size is unknown, improving the state of the art performance on a public dataset.

**Chapter 5: Improving Gaze Estimation by Saliency.**
Using the salient objects detected by the method discussed in Chapter 4, in Chapter 5, we argue that raw gaze estimates should be combined with the saliency information about the scene, to infer the most likely gazed object. Hence, we propose to add a third step in the visual gaze estimation pipeline, which considers salient parts of the gazed scene in order to compensate for the errors which might have occurred during the gaze estimation procedure. Depending on the amount of confidence of the gaze estimate, the saliency framework discussed in Chapter 4 is used to locally define a probability density function, which can be used to find the best adjustment of the raw gaze estimate. The best solution which minimizes the error between the candidate adjustments is then taken to correct the current and future gaze estimates.

The system is tested on three scenarios: using eye tracking data, enhancing a low-accuracy webcam-based eye tracker, and using a head pose tracker. The experimentation shows that the correlation between the subjects in the commercial eye tracking data is improved by an average of 13.91%. The correlation on the low accuracy eye gaze tracker is improved by 59.85%, and for the head pose tracker we obtain an improvement of 10.23%. These results show the potential of the system as a way to enhance and self-calibrate different visual gaze
estimation systems.

6.2 Conclusions

The goal of this thesis was to build an appearance based gaze estimation system which would combine eyes, head pose and information indicating important objects in the gazed scene. Therefore, we have investigated, improved and extended the steps necessary to achieve gaze estimation using a normal webcam. To this end, as the first objective (detection) we targeted the accurate detection of eye location and head pose information in normal, low resolution images. We proposed a fast eye location method which can be used in a variety of situations, and we showed that it is robust to changes in illumination, pose, scale, eye rotations and occlusions. Due to its high accuracy, the method was combined with an appearance based head pose tracker and used to adjust erroneous pose estimations, while the head pose estimations were used to improve the eye location estimation in non-frontal head poses. We showed that this synergetic combination provides improved results in both head pose estimation and eye location.

The second objective of the thesis (estimation) was to combine the obtained head pose and eye information into a framework which would allow restraint-free gaze estimation. To this end, instead of considering the two cues independently, we proposed a method in which the head pose information is used to recalibrate the mapping between the eye positions and known points on the gazed area. This allowed the system to reduce the 3D gaze estimation problem to a set of 2D problems, allowing the eye gaze mapping to be computed at different head pose settings. Although the resulting estimate gives a good approximation of the gaze of the user, this estimate might still be noisy or erroneous due to the resolution of the off-the-shelves hardware and to image noise.

Therefore, the third objective of the thesis (inference) was to adjust these possibly erroneous estimates by using information about the gazed scene. To this end, we firstly proposed a computational method which extracts early saliency in images. The algorithm analyzes local image structures to determine global structures, and proved to achieve higher accuracy than other state of the art methods. Then, the errors which influenced the accuracy of gaze estimation system were taken into account to define a maximum area in which errors were plausible. Using the described saliency framework on the candidate gaze estimates, a possible overall adjustment was computed. This adjustment was then
used to correct the previous and future gaze estimates, obtaining a self calibrating and self adjusting system.

Overall, the methods and techniques proposed in this thesis can be used by devices with a low-resolution camera to address the question "What are you looking at?", in a user-friendly way, without requiring specialized hardware or enforcing specific restraints on the user.

I believe that the ideas discussed in this thesis and the results obtained by their combination are able to provide enabling technology with which computers, starting from a sequence of frames, could understand the interest of the user and react consequently. I foresee the proposed methods and their combination to be actively used in fields such as human behavior analysis and human-computer interfaces, with direct application to assistive interfaces, gaming, interactive advertisement, market research, and many more. Nowadays, for instance, more and more portable devices with integrated cameras are becoming available. Smartphones with front and back facing cameras and with sufficient computational power are already pervasive and, consequently, interesting new computer vision applications and techniques are emerging to take advantage of these new platforms. Using the discussed techniques, it is possible to envision applications in which one camera studies the user, while the other studies the environment. This information could then be combined to understand the interests of the user, and augmented information could flow through the screen of the smartphones, like in an information-magnifier lens. In this way, the proposed technology could enable these devices to effectively become our third eye, pointed at the world of information.