Toward Joint Acquisition-Annotation of Images with Egocentric Devices for a Lower-Cost Machine Learning Application to Apple Detection

Salma Samiei, Pejman Rasti, Paul Richard, Gilles Galopin, David Rousseau

sensors

MDPI

Article

Toward Joint Acquisition-Annotation of Images with Egocentric Devices for a Lower-Cost Machine Learning Application to Apple Detection

Salma Samiei [1,2], Pejman Rasti [1,3], Paul Richard [1], Gilles Galopin [2] and David Rousseau [1,2,*]

1 Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), Université d'Angers, 62 Avenue Notre Dame du Lac, 49035 Angers, France; salma.samiei@univ-angers.fr (S.S.); pejman.rasti@univ-angers.fr (P.R.); paul.richard@univ-angers.fr (P.R.)

2 UMR 1345 Institut de Recherche en Horticulture et Semences (IRHS), INRAe, 42 Rue Georges Morel, 49071 Beaucouzé, France; gilles.galopin@agrocampus-ouest.fr

3 Department of Data Science, école D'ingénieur Informatique et Environnement (ESAIP), 49124 Angers, France

* Correspondence: david.rousseau@univ-angers.fr

Received: 9 May 2020; Accepted: 24 July 2020; Published: 27 July 2020

Abstract: Since most computer vision approaches are now driven by machine learning, the current bottleneck is the annotation of images. This time-consuming task is usually performed manually after the acquisition of images. In this article, we assess the value of various egocentric vision approaches in regard to performing joint acquisition and automatic image annotation rather than the conventional two-step process of acquisition followed by manual annotation. This approach is illustrated with apple detection in challenging field conditions. We demonstrate the possibility of high performance in automatic apple segmentation (Dice 0.85), apple counting (88 percent of probability of good detection, and 0.09 true-negative rate), and apple localization (a shift error of fewer than 3 pixels) with eye-tracking systems. This is obtained by simply applying the areas of interest captured by the egocentric devices to standard, non-supervised image segmentation. We especially stress the importance in terms of time of using such eye-tracking devices on head-mounted systems to jointly perform image acquisition and automatic annotation. A gain of time of over 10-fold by comparison with classical image acquisition followed by manual image annotation is demonstrated.

Keywords: egocentric vision; image annotation; apple detection; eye-tracking

1. Introduction

In the era of machine learning-driven image processing, unequaled performances are accessible with advanced algorithms, such as deep learning, which are highly used in computer vision for agriculture and plant phenotyping [1]. The bottleneck is no more the design of algorithms than the annotation of the images to be processed. When performed manually, this annotation can be very time consuming, and therefore very costly. Consequently, it is useful to investigate all possibilities to accelerate this process. Annotation time can be reduced via multiple approaches, which have all started to be investigated in the domain of bioimaging and especially plant imaging [2–9]. First, (i) annotation time can be reduced by parallelizing the task via online platforms [5]. Additionally, (ii) it can be reduced by using shallow machine learning algorithms that automatically select the most critical images or parts of the images to be annotated via active learning [4]. Transferring segmentation models (iii) learned over available datasets can significantly reduce the need for annotated data [10]. Another approach to reducing annotation time (iv) is to do the training on synthetic datasets that are automatically annotated [2,3,6,7,9,11]. At last, (v) annotation time can be reduced via the

use of ergonomic tools, which enable human annotators to accelerate the process without loss of annotation quality [8]. In this article, we contribute to the latter approach (v) to reduce annotation time. We introduce a novel use of egocentric devices in computer vision for plant phenotyping and assess their value to speed up image annotation.

The term "egocentric device" is used to designate all wearable imaging systems that record images from the first-person perspective. Images captured from egocentric devices are possibly of high value, since their field of view benefits from the attention of the person who wears the device and who is in charge of the targeted task to be done on the images. Reducing the field of view to a part of specific interest may reduce the complexity of the inspected scene and thus help the automatic processing of the acquired images. This is expected to be especially useful in complex scenes, such as those found outdoors in agriculture and phenotyping in the fields. Additionally, some egocentric devices, namely, head-mounted eye-trackers, can even include the capture of the ocular position of the annotator during the recording of the videos. This would, in theory, open up the possibility to annotate images directly, whereas acquisition and annotation are usually two separate steps. Such use of egocentric devices opens up the possibility to conduct these steps jointly and hence reduce annotation time. However, eye-trackers can never be perfectly calibrated, and their practical value in terms of both performance and time is still to be assessed in order to speed up annotation. That is what we propose here.

For the first application of egocentric devices to accelerate annotation, we considered as a proof of concept, a standard problem in computer vision for plant phenotyping. We chose the detection, i.e., segmentation, counting, and localization of apples in color images. This task has been addressed in many ways, including recently, with deep learning. This canonical problem is challenging for computer vision, since it includes self-occlusion of multiple instances, occlusion by the shoots of the apple trees, the variation of illumination, clutter from the self-similar background, variety in sizes and colors of fruits, etc. Additionally, this computer vision problem is significant for various agricultural applications, such as the design of automatic harvesters, automatic estimation of the fruit pack out, and variety testing. Most state-of-the-art methods developed for apple detection are currently working with supervised learning. Such methods require annotated images of apples to be efficient. In this article, we demonstrate how the use of egocentric devices can accelerate the annotation of apples in images. This acceleration in image annotation, illustrated here with apples, is of high value since it could benefit from reducing the annotation cost of any supervised learning segmentation method.

A visual abstract of the proposed original approach for a joint image acquisition-annotation process is illustrated with apple detection in Figure 1. For comparison, the conventional approach is also depicted in Figure 1 wherein a handy camera is used to acquire images, and after image transfer to a computer, images are manually annotated. We propose a single-step approach where hands-free, head-mounted cameras with embedded computational resources are jointly acquiring and annotating images. The article is organized as follows. After positioning our work with the most related work (Section 2), we present (Section 3) the egocentric devices used, the acquisition protocol, and the dataset created for this study. A classical algorithm adapted from the literature is described, as we use it to detect apples in color images (Section 4). The same algorithm is then applied to compare five different computational strategies, specially designed for this study, to reap benefits from egocentric vision (Section 5). We finally conclude on the best practice identified via this comparison.

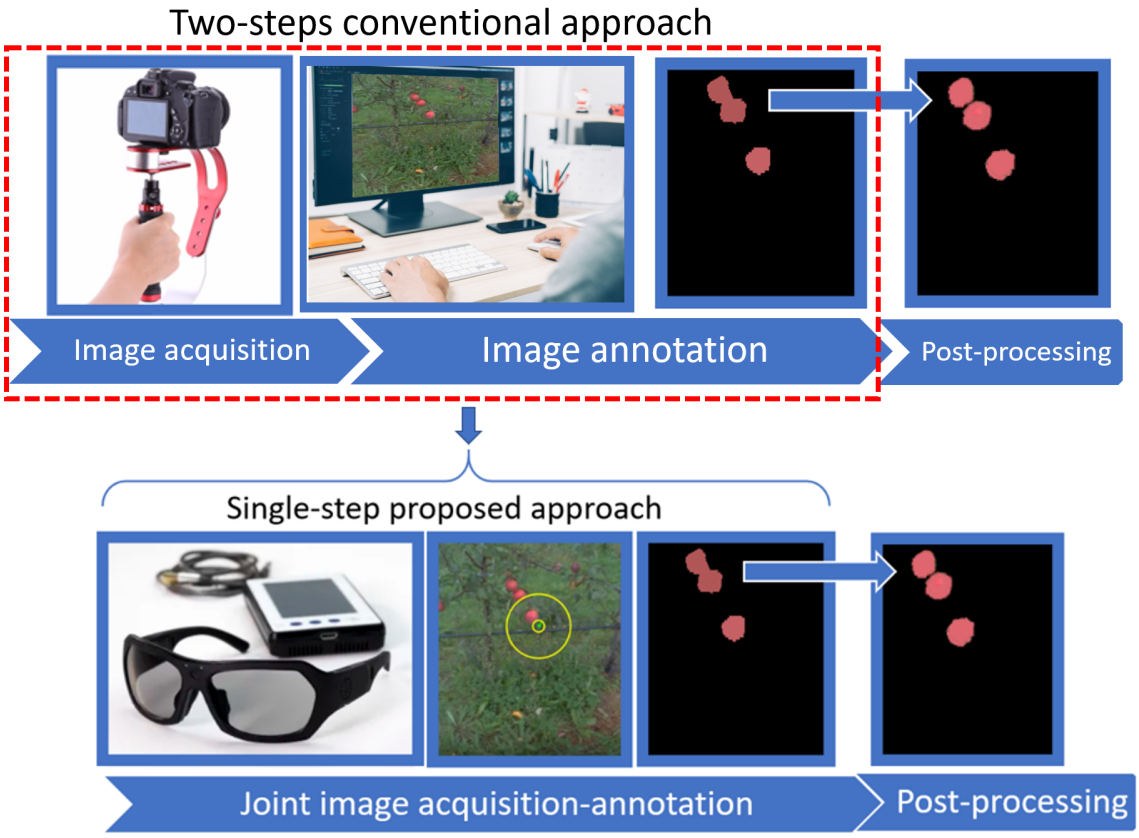

Figure 1. Visual abstract of the article. The red dotted-line encapsulates the conventional two steps of the acquisition and annotation process. We jointly perform image acquisition and image annotation by the use of a head-mounted egocentric device, which simultaneously captures images and the gaze of the person who wears the device and reaps benefits from both factors to annotate images automatically. It is to be noted that the post-processing step to separate touching annotated objects is not included here. It remains a step necessary in the conventional two-step approach and our proposed single-step approach.

2. Related Work

Egocentric (first-person) vision is a relatively new research topic in the field of computer vision which is increasingly attracting interest for understanding human activities [12–15], object detection [16,17], creation of models of the environment with different levels of precision [18,19], perception of social activity[20], user–machine interactions [21], driving assistance [22], and medical applications [23–25]. There are different types of egocentric systems, such as smart glasses, action cameras, and eye-trackers. Based on the processing capabilities, embedded sensors, such as the one used in this article, are now more and more utilized in conjunction with egocentric video analysis [21]. Features such as hand appearance and head motion give essential cues about the attention, behavior, and goals of the viewer [26–29]. In our case, we also used the fact that, usually, in egocentric vision, salient objects of interest tend to occur at the center of the image, since they attract the attention of the viewer [16,30]. In this article, we primarily used an eye-tracking system for egocentric vision to speed up image annotation. The use of eye-tracker to speed up image annotation has been proven useful for annotation with a screen-based system in [8,31,32]. Those studies demonstrated a possible gain of time for annotation of 30-fold (approximately) by comparison with manual annotation. Here, we use, for the first time to the best of our knowledge, an embedded eye-tracking system in the form of glasses (see Figure 1) to jointly conduct image acquisition and annotation and thus extend the results of [8,31,32]. Embedded eye-tracking systems are known to be less accurate than screen-based eye-tracking systems because they can move slightly on the head of the observer during acquisition. However, embedded

eye-tracking systems open the door for an accelerated procedure with joint acquisition and annotation, as illustrated in Figure 1. In this article we will compare the performances in terms of accuracy of apple detection and annotation time of both screen-based eye-tracking systems and embedded eye-tracking systems for image annotation.

Object detection in agricultural conditions has been investigated with a large panel of computer vision approaches [33–45]. In the early works, such as [33], methods were handcrafted both from the hardware side and the software side. Nowadays, it is more common practice to use standard RGB cameras, and base the detection of apples on supervised machine learning methods learned end-to-end via deep learning, as in [44,45]. Such modern methods, neural network-based, show high performances but require large amounts of annotated images. Manual pixel-wise annotation is, in general, a time-consuming operation, taking approximately 1.5 h per 100 images (308 × 202 pixels). In practice, apple detection is also challenging because of illumination conditions [46–48]. In this article, we will not provide a novel method to detect apples automatically. Instead, we will investigate the possibility of performing acquisition and annotation of apples in an orchard environment simultaneously by using head-mounted egocentric devices. Indeed, while there has been significant recent interest in fruit detection, segmentation, and counting in orchard environments, the cost of providing a unified annotated dataset of the fruit on trees makes it the bottleneck in the state-of-the-art literature [49].

The head-mounted egocentric camera provides areas of interest located in the vicinity of the targeted objects in the scene. Therefore, these areas of interest are less accurate than if a manual annotator was pointing at the object with a mouse. We propose in this article to test a standard image segmentation approach to detect the targeted object in the areas of interest provided by the head-mounted egocentric camera. As a consequence, the work relates to the literature on weakly or semi-supervised learning [50] with inexact supervision; that is, the training data are given with labels that are not as exact as desired. Different semi-supervised learning models have been introduced, such as iterative learning (self-training), generative models, graph-based methods, and vector-based techniques [51,52]. The color-based clustering technique for apple detection by using Gaussian Mixture Models was explained in [53]. In this approach, the SLIC superpixel was applied to the input image using the LAB color space. The superpixel's results were clustered into approximately 25 color classes. Finally, based on the KL-divergence between Gaussian Mixtures, each superpixel was classified into an apple or background [54], from hand-labeled classes. Our objective was not to design a novel semi-supervised algorithm. Instead, we revisited existing standard methods based on superpixels and assessed the value of the areas of interest extracted by the head-mounted egocentric camera for a given task of object detection.

3. Material and Method

3.1. Egocentric Vision Device

The egocentric imaging system used was VPS-16 head-mounted eye-tracking glasses equipped with stereoscopic cameras in the nose bridge, a front camera with a diagonal coverage of 88 degrees, and an audio microphone sampling at 10 kHz. The front camera was calibrated with the eye-tracker before acquisition. The visual task defined to the wearer was to find apples on the targeted trees. The acquisition time was nearly 90 s for the whole dataset (calibration time included). This acquisition time is quite similar to the time required with a digital camera fixed on a tripod or hand-held, the former of which would need to be located in different positions to cover all apples located on a tree. The distance of the viewer and the tree was set approximately to one and a half meters. The viewer was counting the number of apples as evidence of the ground-truth, which was recorded via the audio microphone. Fixation points were recorded by the eye-tracker to investigate how they could serve to automatically annotate apples on the trees.

3.2. Dataset

With the sensor described in the previous subsection, we generated a new dataset of 10 videos (25 fps) from 10 various apple trees in the orchard environment captured by the egocentric head-mounted glasses' eye-tracker. The total number of extracted images from the entire dataset was 24,618 (frames). A fundamental parameter of eye-tracking analysis depends on the definition of the fixation and the algorithm used to separate fixation from saccades [55]. Fixation refers to a person's point-of-focus as they look at a stationary target in a visual field. Although the mean duration of a single fixation may depend on the nature of the task [56], numerous studies have been done to measure the average duration for a single fixation [56–65]. The mean fixation duration for visual search is 275 ms, and for tasks that require hand-eye coordination, such as typing, the mean fixation can be 400 ms [56]. Among our dataset, the number of frames which received gazing of at least 275 ms was 419. The acquisitions were made on two days at midday with different weather conditions at the orchard of INRAE Angers, France. No difference was found in the results of the data coming from the two days. This dataset includes a variety of apple colors together with apple and foliage density, which are representative of the dataset found in the literature for apple detection [66–68]. Due to the complexity of each orchard tree, the illumination, and the environment itself, different natural colors were found in the images, including various shades of green, red, yellow, brown, or gray for the appearance of foliage, grass, apples, and tree trunks. Ground-truth was created by manual annotation of the raw color images at approximately 54 s per image by using the Image Segmenter application in Matlab 2017a. A sample of raw color images from different apple trees and their corresponding manual ground-truth are illustrated in Figure 2. For the whole dataset, which consists in 419 images, it roughly took 6 h to manually annotate all images. These manual annotations were generated for evaluation of the accuracy of the egocentric vision methods presented in the next section.

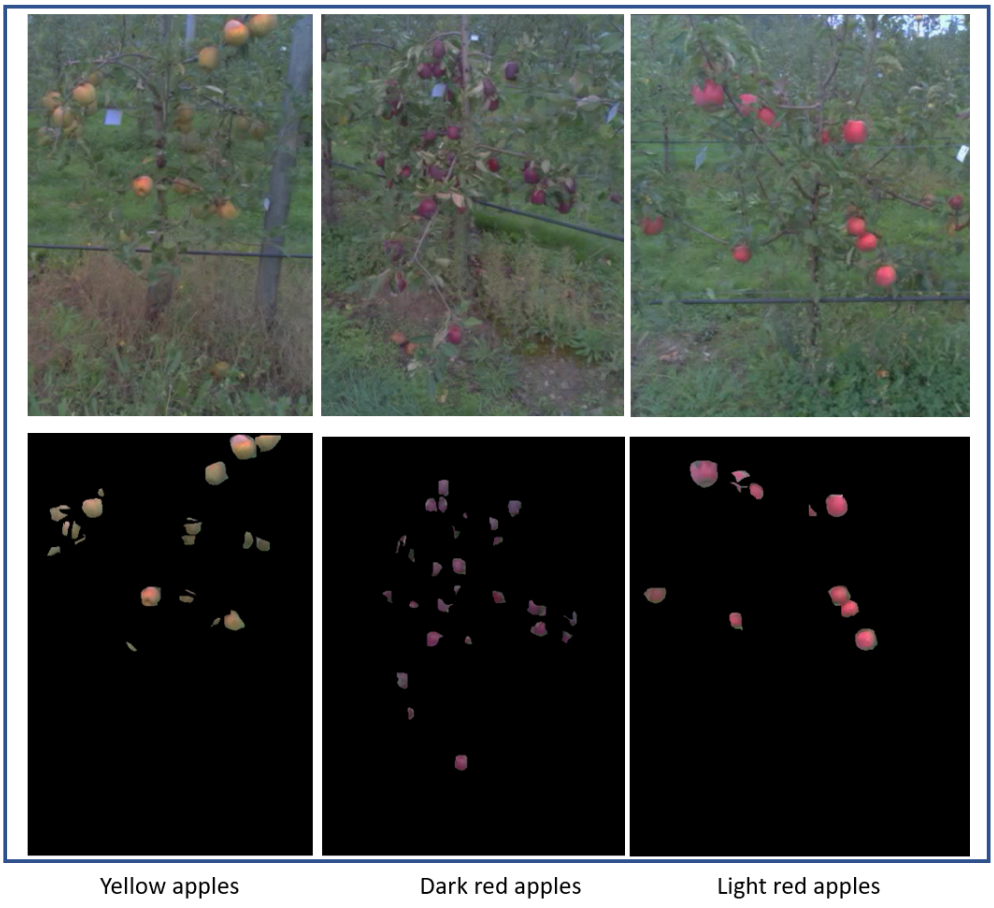

Figure 2. Example of RGB images of apple trees from our dataset and the corresponding ground-truth (manually annotated).

4. Image Processing Pipeline

In this section, we present the image processing pipeline developed to automatically annotate apples from the attention areas captured with egocentric vision. A global view of this pipeline is depicted in Figure 3 and includes three main steps: image pre-processing, segmentation, and performance evaluation.

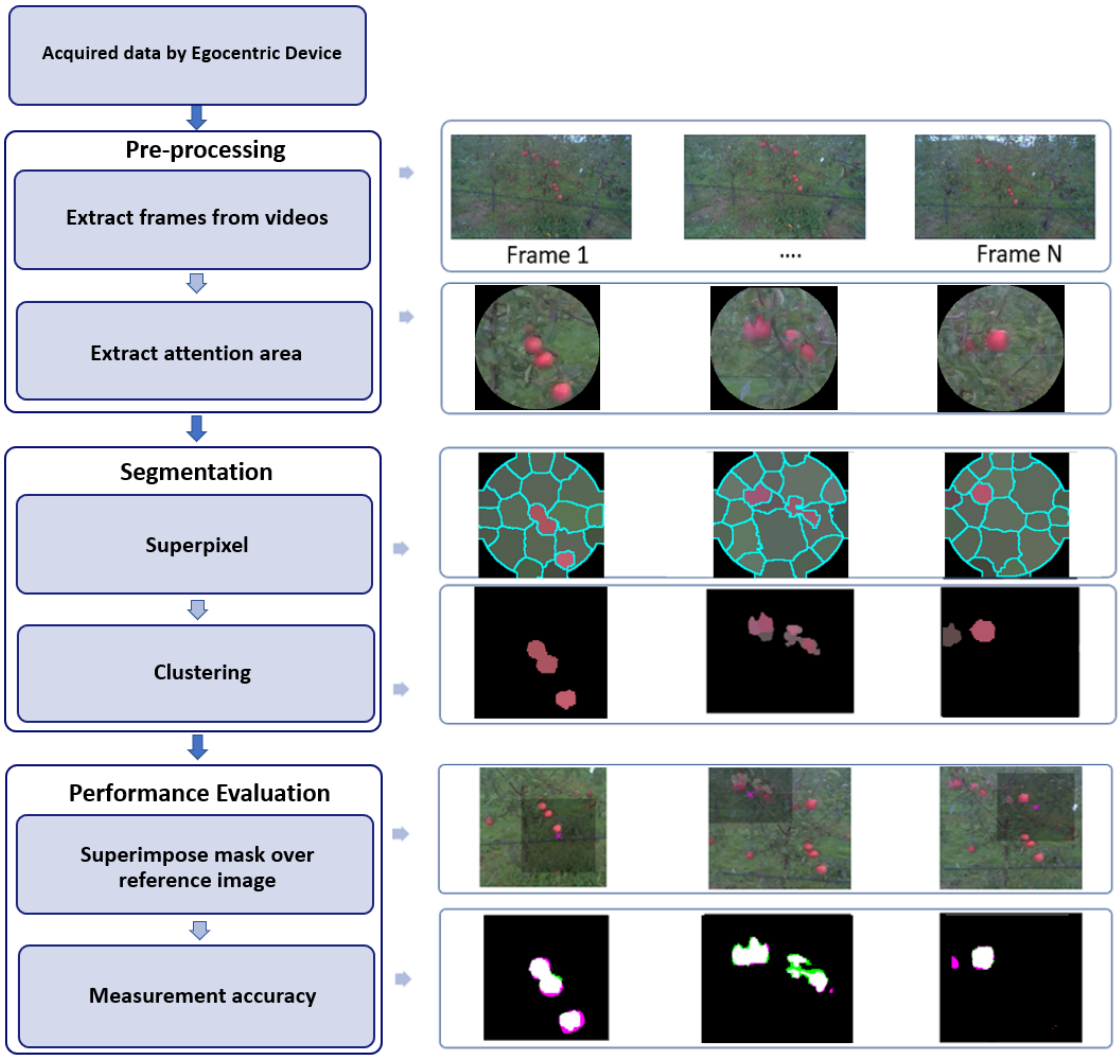

Figure 3. The three-step image processing pipeline proposed to automatically segment apples from the attention areas captured with egocentric devices.

The pre-processing started with the extraction of the frames with a resolution of 960 × 544 pixels from recorded videos. Next, an attention area was extracted from each frame based on egocentric priors. The extraction of this attention area constitutes the main contribution of the article. Several strategies have been tested and are presented in the next section. The pre-processed images were then segmented with a standard approach for apple detection similar to the one presented in [49,53,69–71]. A classical superpixel technique (SLIC) [72] was applied followed by a simple non-supervised clustering technique, *K*-means [73], to select superpixels corresponding to apples. To keep the size of superpixel independent of the size of the attention area, we defined the number of superpixels as the ratio of

$$N = \frac{A}{S}, \tag{1}$$

where A represents the size of the attention area, and S the size of an average apple, which is equal to 900 pixels in our dataset.

To simplify the images, the tree-labels (blue in our case) and sky parts were removed by applying color thresholding (optimized on a small dataset) in the RGB color domain on the superpixel segmented attention areas, as shown in Figure 4. The number of cluster K was found optimal for $K = 2$ and was applied to feature space composed of (R, G, B, H, S) respectively for red, green, brightness, hue, and saturation from each superpixel. The cluster with the smaller size was considered as the apple cluster based on the assumption that the background occupied the largest area in the attention area. Because blue parts were withdrawn and no green apples were present, the optimal value of $K = 2$ was reasonable for our use-case of apple detection in the orchard. Indeed, the local complexities in attention areas extracted from the egocentric devices were limited to objects on a background with a contrast of color. For other use-cases, where local contrast between the object and background could depend on other features (size, texture, shape, etc.), it would be necessary to adapt this segmentation.

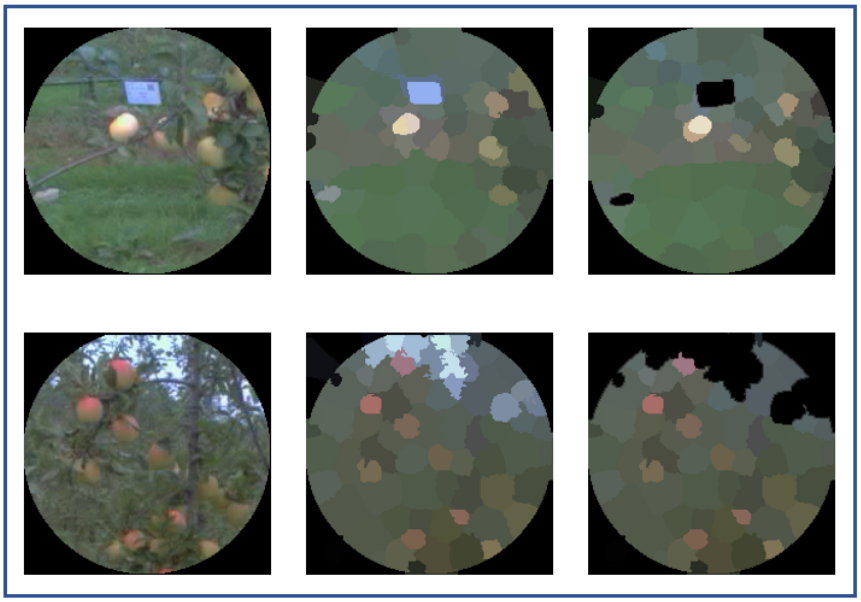

Figure 4. Color thresholding to remove blueish color belonging to the sky or blue tree-labels on superpixel segmented attention areas. Each row represents from left to right: the attention area, the superpixel segmented attention area, and the thresholded one, respectively.

Finally, the segmented apples were superimposed over the original image for qualitative assessment and localization, and compared with the manual binary ground-truth to compute the segmentation accuracy via the Dice $Dc(X, Y)$ and Jaccard index $J(X, Y)$ given by

$$Dc(X, Y) = \frac{2 * |X \cap Y|}{|X| + |Y|}, \tag{2}$$

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}, \tag{3}$$

where X and Y represent the segmented image and the ground-truth respectively.

In addition to the segmentation of apples, counting and localization were also computed in the following way. For object counting, we counted the number of connected components among detected

objects which shared sufficient overlaps with ground-truth. An empirical threshold of 75 percent was chosen for the overlap. The probability of good detection was computed as

$$PD = \frac{TP}{TP + FN}\,, \tag{4}$$

with TP number of true-positive objects and FN number of false-negative objects. We also computed the probability of true-negative rate as

$$TNR = \frac{TN}{TN + FP}\,, \tag{5}$$

with TN number of true-negative objects and FP number of false-positive objects.

In localization, the Euclidean distance between the centroid x_i of detected objects X_i and the centroid y_j of objects Y_i with a maximum intersection with ground-truth was computed as

$$d(x_i, y_j) = \sqrt{(u_{x_i} - v_{y_j})^2 + (u_{x_i} - v_{y_j})^2}\,, \tag{6}$$

with u and v, which stand for Cartesian coordinates in the images and

$$j = \arg\max_{j_0} \left| X_i \cap Y_{j_0} \right|\,. \tag{7}$$

The average distance

$$d = \frac{1}{N} \sum_{i=1}^{N} d(x_i, y_j), \tag{8}$$

was computed over all detected objects sharing sufficient overlap with ground-truth. Here again, a threshold of 75 percent of overlap was chosen. Distance d represents the average shift error of localization of apples with an egocentric device from manual ground-truth.

5. Strategies for Extracting Attention Area

In the following we mention different approaches for extracting attention area either using eye-tracking or not.

5.1. Attention Area from Eye-Tracking

In this section, we present strategies that we developed to extract attention areas from the eye-tracking devices to perform joint acquisition-annotation after passing these areas to the image processing pipeline of the previous section.

5.1.1. Selection by Eye-Tracking Glasses

The first approach extracted attention areas via the viewer fixation computed from the egocentric eye-tracking glasses. In order to fix a threshold, a gazing position was recorded when the same fixation position was observed during an interval of 6 frames, as calculated by

$$fi = Fps \times fd\,, \tag{9}$$

where fi is the frame interval, $Fps = 25$ is the number of frames per second, and fd is the average fixation duration, which was set as 275 ms. Despite careful calibration before the acquisition, small shift errors of alignment between the front camera of the device and the gazing point of the viewer can occur. Therefore, we extended the attention area around each gazing position with a given radius to compensate for the remaining small shift error of calibration of the eye-tracker. An illustration of the creation of an attention area around a fixation point is provided in Figure 5. A systematic

analysis of the evolution of the average segmentation accuracy as a function of the radius of the attention area around each gazing position was undertaken. It is shown in Figure 6 and demonstrates a non-monotonic evolution culminating at a value corresponding to triple the size of an average apple size in our dataset. Consistently, this optimal value was also found to be very close to the maximum shift error of calibration of the eye-tracker found in the whole dataset. For attention areas that are too small, due to the shift error, apples can be missed. For overly large attention areas, due to the complexity of the scene, the segmentation process fails to detect all apples correctly in the area.

Figure 5. Construction of attention areas. (**a**) The average diameter of an average apple is 30 pixels in our dataset; (**b**) a cross indicates the center of the gaze of the annotator. There is a shift error from the apple of (**a**). The maximum distance of the gazing point with the center of the closest object was found at 169 pixels. (**c**) Chosen attention area with a size of 180 × 180 pixels.

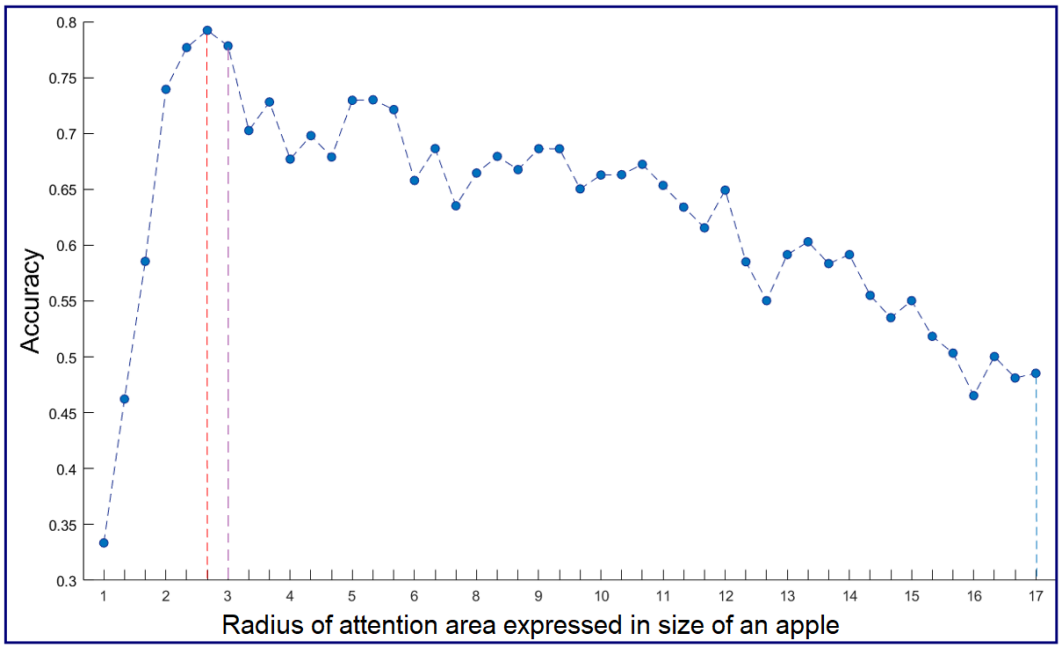

Figure 6. Apple segmentation accuracy as a function of the radius of attention area expressed in the size of apples taken as 30 pixels. Maximum accuracy achieved when the radius size of the attention map is equal to 80 (160 × 160 pixels) corresponding to the red dotted line. The purple dotted line corresponds to the maximum gaze shift error of (169 pixels) between eye-tracker and ground-truth when computed on the whole dataset.

5.1.2. Selection by Screen-Based Eye-Tracking

For comparison with the attention area created with the egocentric eye-tracker directly acquired in the orchard, we also generated an attention map from the gazing point recorded with a screen-based eye-tracker. Of course, this approach is less interesting for gain of time than the previous one with the head-mounted eye-tracker, since it does not allow a joint acquisition annotation. However, desktop eye-trackers are more accurate than head-mounted ones and thus are expected to constitute a reference serving as an upper bound in terms of quality of annotation with ego-centric vision. The experiment was performed on a screen with a resolution of 1920 × 1080 pixels while the eye movements of the viewer were recorded with an SMI binocular remote eye-tracker [74]. In this approach, for each apple tree, we peaked out one frame, which included all the apples.

The annotation protocol was the same as in the previous method. Each image was displayed to the viewer, who was asked to find the apples on the trees. The locations of the fixations of the viewer were recorded at 60 Hz. For a fair comparison, the attention area diameter around each recorded fixation was taken at the optimal value found for the eye-tracking systems embedded in glasses.

A comparison of the accuracy of the screen-based eye-tracking recording and the recording with eye-tracking embedded in glasses was conducted. Figure 7 shows that in the form of heatmap visualization of the attention of the viewer. The precision and accuracy of the produced gaze points with the screen-based eye-tracker were found to be higher than when using the head-mounted eye-tracker. The average shift error of Equation (8) was found to be 125 pixels less with the screen-based eye-tracker than with the head-mounted eye-tracker.

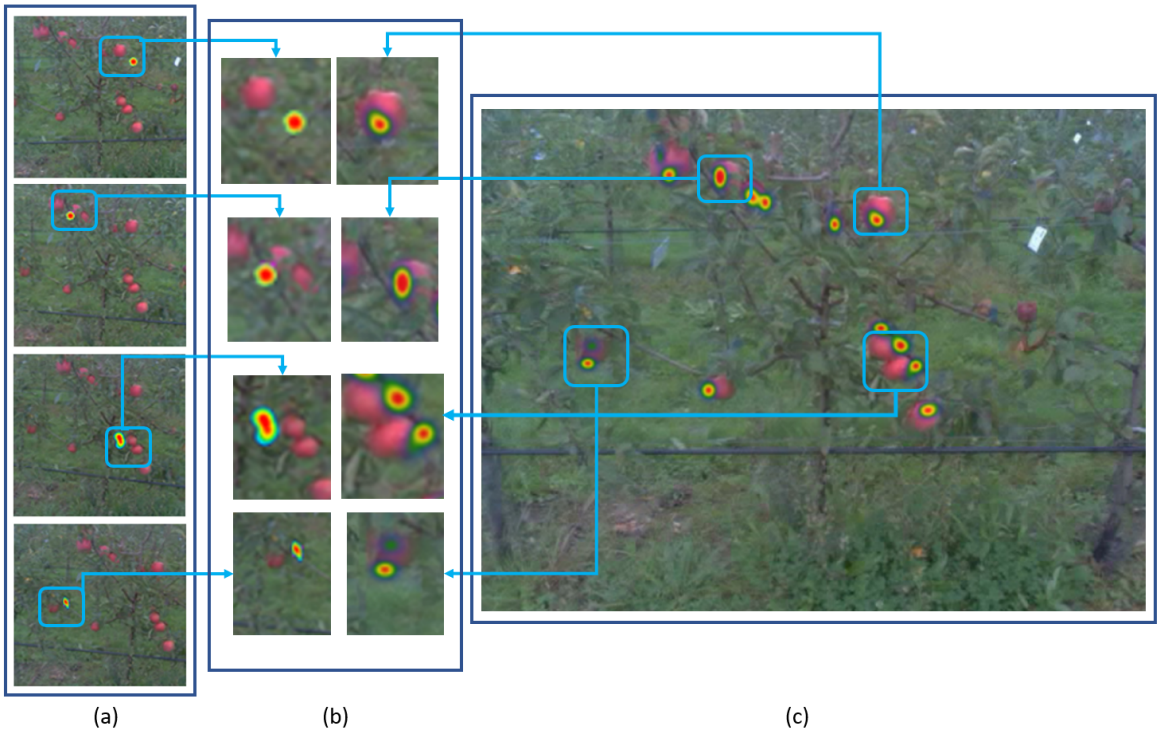

Figure 7. Heatmap visualization of the attention of the viewer captured by the head-mounted (glasses) eye-tracker (**a**) versus the screen-based eye-tracker (**c**). (**b**) Comparison of the heatmap generated by the glasses eye-tracker (left) vs. the heatmap generated by the screen-based eye-tracker (right).

5.2. Attention Area without Eye-Tracking

Other strategies were developed to extract attention areas for comparison with performances obtained with eye-tracking systems.

5.2.1. Full-Frame

In this approach, the attention map was considered as the full-frame recorded by the camera. Thus, in Figure 3, instead of a small patch of the entire original image, the full original image was directly transmitted to the superpixel segmentation. Such a choice assumes that the camera field of view is already a focus of the overall field of interest for the human annotator in charge of detecting apples.

5.2.2. Egocentric Prior

In this approach, we assumed, as is often done in egocentric vision [16], that the attention of the viewer was focused at the center of the frame. Therefore, we selected the attention area as a disk positioned at the center of the image with the size of 180×180 pixels for a fair comparison with the other approaches developed for eye-trackers.

5.2.3. Saliency Map

As the last method to compute an attention area, we turned toward a computational approach in charge of numerically identifying areas of interest. Such a concept has been developed in the computer vision literature under the name of the saliency map. Saliency acts as a local filter that enhances regions of the image which stand out relative to their adjacent parts in terms of orientation and/or gray level and/or color contrast [75]. Introduced in [76], saliency was inspired by the mechanisms of human visual attention and the fixation behavior of the observer. There are numerous computational models for salient object detection. In this study, for illustration and without any claim of optimality, we used the algorithm proposed by [77], which computes saliency map in images using low-level features and was proposed with codes included for reproducible science. Saliency maps were thresholded to binary masks following the fixed threshold procedure described in [77]. Each connected component of the binary saliency map served to produce an attention area. For a fair comparison with the other approaches, attention areas of size 180×180 pixels were chosen.

6. Results and Discussion

We are now ready to compare the results of the different approaches proposed for apple detection by extracting attention areas through egocentric vision in the perspective of a joint acquisition-annotation process. As shown in Table 1, we assessed the image annotation quality by the same image segmentation pipeline of Section 4 (depicted in Figure 3). Comparison is provided between the five different approaches presented in Section 5 for the extraction of attention areas from egocentric devices. In terms of segmentation, accuracy was estimated by the Dice Equation (2) and Jaccard Equation (3) indexes. The probability of good detection indicates the true counted apples computed by Equation (4). The true-negative rate Equation (5) represents the proportion of actual negatives that are correctly identified. The next column in Table 1 specifies the error of localization of detected apples computed by Equation (8). Time is the approximate consumed execution time (automatic annotation) acquired from each approach of the whole dataset. Finally, the time gain indicates the ratio of manual annotation time over the consumed execution time obtained from each automatic annotation approach. All these experimental results correspond to an average of over 10 different trees available in the dataset.

The best average performances (highlighted in bold in Table 1) in terms of segmentation accuracy of apples were obtained with the eye-tracking-based methods. Challenging images and resulting annotations with eye-tracking-based methods are provided in Figure 8 for qualitative assessment. Overall, the screen-based eye-tracker provided the best result but only slightly above the one obtained from the glasses eye-tracker. This embedded glasses eye-tracker, despite its substantial shift errors, had a high value since it enabled joint image acquisition and annotation. The saliency approach provided a result close to the one obtained with the baseline method (full-frame). This could certainly be improved with a systematic benchmark of other saliency methods of the literature. However,

a fundamental reason for the failure of the saliency approach, which would be common to all generic saliency maps, is that saliency is, so to say, attracted by contrasting objects which may not be apples (for example, stems, leaves, items in the background, a data matrix positioned in the field to identify trees). As a consequence, saliency creates many true-negatives in attention areas since the task of detecting apples does not specifically drive it. In contrast, human attention focuses on the apple as captured by eye-tracking systems.

Interestingly these results were consistent for the three tasks assessed: segmentation, counting, and localization. This demonstrates the robustness of the interest of eye-tracker devices for annotation. Eye-tracking systems, such as the two used in this study, can be considered as expensive devices (typically between 10,000 and 20,000 euros currently). It is interesting to see that the egocentric prior approach gave the third-best performance, and this could be accessible with any camera embedded on glasses (for 10 to 100 euros).

Table 1. Performance of apple detection with the five approaches developed for automatic apple annotation in the attention area captured by the egocentric devices. Each column corresponds to an average over the 10 trees of the dataset. Dice and Jaccard assess in percentage the quality of segmentation via Equations (2) and (3); good prediction and true-negative rate assess in percentage the quality of object detection via Equations (4) and (5); and the shift error of Equation (8) assesses in pixels the quality of good localization. The time corresponds to the approximate execution time for automatic annotation for the whole dataset in seconds. Time gain indicates the ratio of manual annotation time (6 h) over automatic annotation time obtained from each approach. Time was measured on a windows machine with an Intel Xeon CPU and 32.0 GB RAM by Matlab 2017a.

Method (Section)	Dice	Jaccard	Good Detection	True-Negative Rate	Shift Error	Time (Second)	Time Gain
Full-Frame (Section 5.2.1)	0.24 ± 0.22	0.21 ± 0.16	0.31 ± 0.20	0.17 ± 0.72	174.11 ± 34	880	24
Glasses Eye-tracker (Section 5.1)	0.78 ± 0.08	0.64 ± 0.08	0.84 ± 0.16	0.09 ± 0.07	15.97 ± 11	1960	11
Screen-based Eye-tracker (Section 5.1.2)	0.85 ± 0.09	0.77 ± 0.13	0.88 ± 0.12	0.09 ± 0.13	2.37 ± 1.86	3240	6
Egocentric Prior (Section 5.2.2)	0.46 ± 0.36	0.38 ± 0.31	0.54 ± 0.39	0.28 ± 0.23	84.82 ± 7.25	1960	11
Saliency (Section 5.2.3)	0.27 ± 0.13	0.16 ± 0.08	0.42 ± 0.45	0.51 ± 0.17	7.21 ± 8.28	2358	9

The values of the obtained results in terms of segmentation, counting, and localization were also assessed in terms of timing. As expressed in Section 3.1, acquisition time with an egocentric device is comparable with acquisition time with any standard camera. Therefore gains of time were compared regarding the annotation time only. This timing is provided in the last column of Table 1 for automatic annotation based on the image processing pipeline applied to extracted attention areas. Without any surprise, the full-frame approach, which requires no computation of attention map, is the fastest method. The second most rapid methods are the egocentric prior and glasses eye-tracker. The screen-based eye-tracker method, which gave the best performance in terms of apple detection, came with the slowest timing. However, these timings for automated annotation are to be compared with the timing requested by a human annotator to manually annotate all apples in the dataset. The estimated timing was 6 h for the 419 frames. The gain of time for all methods is presented in Table 1. Saliency, as presented here, could be criticized since many other variants of the saliency map could be tested and possibly provide better results. In terms of timing, however, we believe the performances are realistic, and it was worth mentioning them here. All in all, the glasses eye-tracker method appears to be a good trade-off between speed and annotation performance (as summarized in Table 2). For this head-mounted device, the gain in performance was about 11 times, which is smaller than what was found in the closest related work with desktop eye-trackers for object detection [8,31,32]. This difference may come from the fact that in this literature, the tasks targeted were relatively more straightforward and required less post-processing. Optimization of the code could thus increase the gain in time. We are currently investigating all those perspectives.

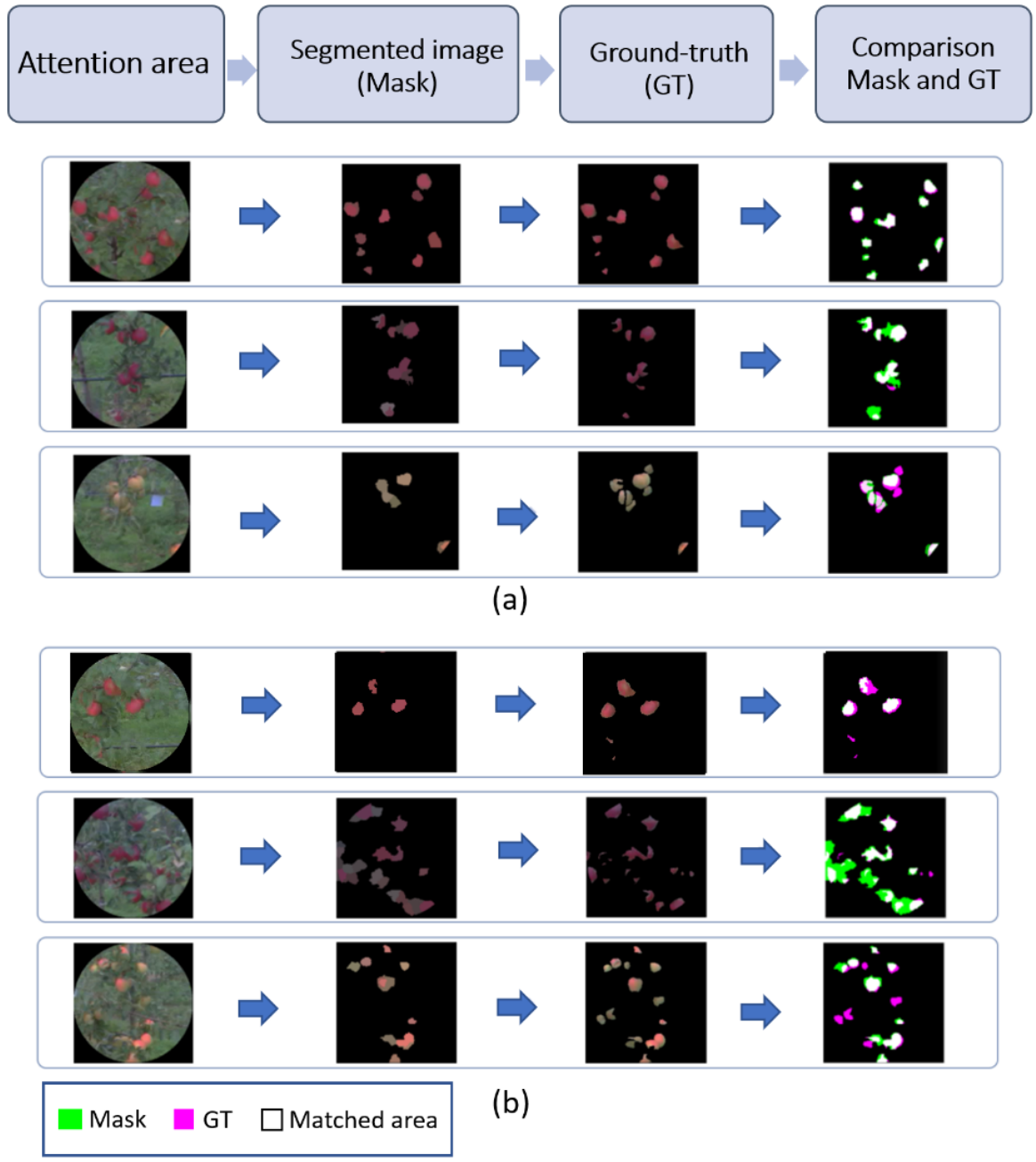

Figure 8. Qualitative assessment of results. From left to right, an example of the attention area captured by eye-tracking, automatic annotation obtained from the proposed image processing pipeline of Figure 3, ground-truth manually recorded, and comparison of manual ground-truth and automatic segmentation. (**a**) Examples of good performance; (**b**) Some challenging conditions wherein more errors were found (missed detection, false detection).

Table 2. Qualitative summary of the five uses of egocentric devices compared in this study.

Method	Joint Acquisition Annotation	Fastest Execution Time	Best Annotation	Best Counting	Best Localization
Full-Frame	+	+	-	-	-
Glasses Eye-tracker	+	-	+	+	-
Screen-based Eye-tracker	-	-	+	+	+
Egocentric Prior	+	-	-	-	-
Saliency	+	-	-	-	+

7. Conclusions

We have assessed the value of egocentric imaging devices to jointly perform acquisition and automatic image annotation. This was illustrated with apple detection in orchards, which is known to be a challenging task for computer vision applied to phenotyping or agriculture. Despite shift errors in the calibration of egocentric imaging devices, the performance of the detection of apples from the gazed recorded areas was found to be very close to the one obtained from the manual annotation. The compensation for these shift errors was obtained by applying a standard non supervised segmentation algorithm only applied in attention areas centered on the gazing positions captured by the egocentric devices. Specific interest was shown for head-mounted eye-tracking systems with an estimated gain of time in comparison with manual annotation of 11 times with non-GPU-accelerated software.

This first use of egocentric vision to speed up image annotation opens up interesting perspectives, especially in plant phenotyping. The task here was focused on apples, but the approach is in fact generic. Thus, it would be interesting to extend the applicability to other phenotyping items of interest. The non-supervised image segmentation algorithm applied in gazed areas was purposely chosen simply in this article to demonstrate the value of the eye-tracking device. It is interesting to notice that performances obtained with this simple algorithm were already interesting quantitatively and qualitatively. The literature of non-supervised image segmentation with superpixels is huge [78,79], and it would be interesting to revisit more exhaustively this literature for the segmentation of gazed areas. Specific attention could focus on the methods addressing the limitation of superpixels [80], also observed in this article, with "leakage" of boundaries in the vicinity of the targeted objects [81]. To remain on the topic of apples, this could include the determination of flowering stages or the detection of diseases. Additional technological services from egocentric vision could be tested to speed up annotation. For instance, this includes the use of sound recording, which could be coupled to automatic speech recognition for later fusion with information extracted from the captured images. The pilot study presented here is promising. For a tool to be used by technicians and engineers in the field, it would be necessary to implement an ergonomic version of the software to experiment on a large network of users the method developed to accelerate image annotation with egocentric devices. Validation of the quality of the annotation was performed at various levels, including location, object detection, and pixel-wise segmentation. Another stage of validation of the quality of the annotation would be to train a machine learning algorithm on the annotated images and compare the performance with the manually annotated data.

Author Contributions: S.S. and D.R. conceived and implemented the work. S.S. and P.R. (Pejman Rasti) performed image acquisition. G.G. and P.R. (Paul Richard) contributed to the management and administration of the study. S.S. and D.R. wrote and revised the manuscript. All authors validated the final version of the manuscript.

Acknowledgments: Authors gratefully acknowledge the Région des Pays de la Loire for funding this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
2. Benoit, L.; Rousseau, D.; Belin, É.; Demilly, D.; Chapeau-Blondeau, F. Simulation of image acquisition in machine vision dedicated to seedling elongation to validate image processing root segmentation algorithms. *Comput. Electron. Agric.* **2014**, *104*, 84–92. [CrossRef]
3. Giuffrida, M.V.; Scharr, H.; Tsaftaris, S.A. ARIGAN: Synthetic arabidopsis plants using generative adversarial network. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW 2017), Venice, Italy, 22–29 October 2017; pp. 2064–2071. [CrossRef]

4. Peter, L.; Mateus, D.; Chatelain, P.; Declara, D.; Schworm, N.; Stangl, S.; Multhoff, G.; Navab, N. Assisting the examination of large histopathological slides with adaptive forests. *Med. Image Anal.* **2017**, *35*, 655–668. [CrossRef] [PubMed]
5. Giuffrida, M.V.; Chen, F.; Scharr, H.; Tsaftaris, S.A. Citizen crowds and experts: Observer variability in image-based plant phenotyping. *Plant Methods* **2018**, *14*, 12. [CrossRef] [PubMed]
6. Barth, R.; Ijsselmuiden, J.; Hemming, J.; Henten, E.V. Data synthesis methods for semantic segmentation in agriculture: A Capsicum annuum dataset. *Comput. Electron. Agric.* **2018**, *144*, 284–296. [CrossRef]
7. Douarre, C.; Schielein, R.; Frindel, C.; Gerth, S.; Rousseau, D. Transfer learning from synthetic data applied to soil–root segmentation in X-ray tomography images. *J. Imaging* **2018**, *4*, 65. [CrossRef]
8. Samiei, S.; Ahmad, A.; Rasti, P.; Belin, E.; Rousseau, D. Low-cost image annotation for supervised machine learning. Application to the detection of weeds in dense culture. In *British Machine Vision Conference (BMVC), Computer Vision Problems in Plant Phenotyping (CVPPP)*; BMVA Press: Newcastle, UK, 2018; p. 1.
9. Douarre, C.; Crispim-Junior, C.F.; Gelibert, A.; Tougne, L.; Rousseau, D. Novel data augmentation strategies to boost supervised segmentation of plant disease. *Comput. Electron. Agric.* **2019**, *165*, 104967. [CrossRef]
10. Hung, C.; Nieto, J.; Taylor, Z.; Underwood, J.; Sukkarieh, S. Orchard fruit segmentation using multi-spectral feature learning. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; pp. 5314–5320. [CrossRef]
11. Ubbens, J.; Cieslak, M.; Prusinkiewicz, P.; Stavness, I. The use of plant models in deep learning: An application to leaf counting in rosette plants. *Plant Methods* **2018**, *14*, 6. [CrossRef]
12. Fathi, A.; Farhadi, A.; Rehg, J.M. Understanding egocentric activities. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 407–414. [CrossRef]
13. Doherty, A.R.; Caprani, N.; Conaire, C.Ó.; Kalnikaite, V.; Gurrin, C.; Smeaton, A.F.; O'Connor, N.E. Passively recognising human activities through lifelogging. *Comput. Hum. Behav.* **2011**, *27*, 1948–1958. [CrossRef]
14. Pirsiavash, H.; Ramanan, D. Detecting activities of daily living in first-person camera views. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2847–2854. [CrossRef]
15. Lu, Z.; Grauman, K. Story-driven summarization for egocentric video. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Portland, OR, USA, 23–28 June 2013; pp. 2714–2721. [CrossRef]
16. Fathi, A.; Ren, X.; Rehg, J.M. Learning to recognize objects in egocentric activities. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 20–25 June 2011; pp. 3281–3288. [CrossRef]
17. Erculiani, L.; Giunchiglia, F.; Passerini, A. Continual egocentric object recognition. *Comput. Vis. Pattern Recognit.* **2019**, arXiv:1912.05029v2.
18. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef]
19. Rituerto, A. Modeling the environment with egocentric vision systems. *Electron. Lett. Comput. Vis. Image Anal.* **2015**, *14*, 49–51. [CrossRef]
20. Alletto, S.; Serra, G.; Calderara, S.; Cucchiara, R. Understanding social relationships in egocentric vision. *Pattern Recognit.* **2015**, *48*, 4082–4096. [CrossRef]
21. Betancourt, A.; Morerio, P.; Regazzoni, C.S.; Rauterberg, M. The Evolution of First Person Vision Methods: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 744–760. [CrossRef]
22. Liu, K.Y.; Hsu, S.C.; Huang, C.L. First-person-vision-based driver assistance system. In Proceedings of the 2014 International Conference on Audio, Language and Image Processing, Shanghai, China, 7–9 July 2014; pp. 239–244. [CrossRef]
23. Mayol, W.W.; Davison, A.J.; Tordoff, B.J.; Murray, D.W. Applying active vision and SLAM to wearables. In *Springer Tracts in Advanced Robotics*; Dario, P., Chatila, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; Volume 15, pp. 325–334. [CrossRef]
24. Karaman, S.; Benois-Pineau, J.; Mégret, R.; Dovgalecs, V.; Dartigues, J.F.; Gaëstel, Y. Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases. In Proceedings of the International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 4113–4116. [CrossRef]

25. Doherty, A.R.; Hodges, S.E.; King, A.C.; Smeaton, A.F.; Berry, E.; Moulin, C.J.; Lindley, S.; Kelly, P.; Foster, C. Wearable cameras in health: The state of the art and future possibilities. *Am. J. Prev. Med.* **2013**, *44*, 320–323. [PubMed]
26. Li, Y.; Fathi, A.; Rehg, J.M. Learning to predict gaze in egocentric video. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3216–3223. [CrossRef]
27. Li, C.; Kitani, K.M. Pixel-level hand detection in ego-centric videos. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3570–3577. [CrossRef]
28. Bambach, S.; Lee, S.; Crandall, D.J.; Yu, C. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; Volume 2015, pp. 1949–1957. [CrossRef]
29. Ma, M.; Fan, H.; Kitani, K.M. Going Deeper into First-Person Activity Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1894–1903. [CrossRef]
30. Tatler, B.W.; Baddeley, R.J.; Gilchrist, I.D. Visual correlates of fixation selection: Effects of scale and time. *Vis. Res.* **2005**, *45*, 643–659. [CrossRef]
31. Walber, T. Making use of eye tracking information in image collection creation and region annotation. In Proceedings of the 20th ACM International Conference on Multimedia (MM 2012), Nara, Japan, 29 October–2 November 2012; ACM Press: New York, NY, USA, 2012; pp. 1405–1408. [CrossRef]
32. Lucas, A.; Wang, K.; Santillan, C.; Hsiao, A.; Sirlin, C.B.; Murphy, P.M. Image Annotation by Eye Tracking: Accuracy and Precision of Centerlines of Obstructed Small-Bowel Segments Placed Using Eye Trackers. *J. Digit. Imaging* **2019**, *32*, 855–864. [CrossRef]
33. Parrish, E.A.; Goksel, A.K. Pictorial Pattern Recognition Applied To Fruit Harvesting. *Trans. Am. Soc. Agric. Eng.* **1977**, *20*, 822–827. [CrossRef]
34. D'Grand, E.; Rabatel, A.G.; Pellenc, R.; Journeau, A.; Aldon, M.J. Magali: A self-propelled robot to pick apples. *Am. Soc. Agric. Eng. Pap.* **1987**, *46*, 353–358.
35. Whittaker, A.D.; Miles, G.E.; Mitchell, O.R.; Gaultney, L.D. Fruit Location in a Partially Occluded Image. *Trans. Am. Soc. Agric. Eng.* **1987**, *30*, 591–596. [CrossRef]
36. Slaughter, D.C.; Harrell, R.C. Color vision in robotic fruit harvesting. *Trans. ASAE* **1987**, *30*, 1144–1148. [CrossRef]
37. Sites, P.W.; Delwiche, M.J. Computer Vision To Locate Fruit on a Tree. *Trans. Am. Soc. Agric. Eng.* **1988**, *31*, 257–263, 272. [CrossRef]
38. Rabatel, G. A vision system for Magali, the fruit picking robot. In Proceedings of the International Conference on Agricultural Engineering, Paris, France, 2–5 March 1988.
39. Kassay, L. Hungarian robotic apple harvester. In Proceedings of the ASAE Annual Meeting Papers, Charlotte, NC, USA, 21–24 June 1992.
40. Ceres, R.; Pons, J.; Jimenez, A.; Martin, J.; Calderon, L. Agribot : A Robot for Aided Fruit Harvesting. *Ind. Robot.* **1998**, *25*, 337–46. [CrossRef]
41. Jiménez, A.R.; Ceres, R.; Pons, J.L.; Jimenez, A.R.; Ceres, R.; Pons, J.L. A survey of computer vision methods for locating fruit on trees. *Trans. Am. Soc. Agric. Eng.* **2000**, *43*, 1911–1920. [CrossRef]
42. Zhou, R.; Damerow, L.; Sun, Y.; Blanke, M.M. Using colour features of cv. 'Gala' apple fruits in an orchard in image processing to predict yield. *Precis. Agric.* **2012**, *13*, 568–580. [CrossRef]
43. Song, Y.; Glasbey, C.A.; Horgan, G.W.; Polder, G.; Dieleman, J.A.; van der Heijden, G.W. Automatic fruit recognition and counting from multiple images. *Biosyst. Eng.* **2014**, *118*, 203–215. [CrossRef]
44. Sa, I.; Ge, Z.; Dayoub, F.; Upcroft, B.; Perez, T.; McCool, C. Deepfruits: A fruit detection system using deep neural networks. *Sensors* **2016**, *16*, 1222. [CrossRef]
45. Boogaard, F.P.; Rongen, K.S.; Kootstra, G.W. Robust node detection and tracking in fruit-vegetable crops using deep learning and multi-view imaging. *Biosyst. Eng.* **2020**, *192*, 117–132. [CrossRef]
46. Wang, Q.; Nuske, S.; Bergerman, M.; Singh, S. Automated Crop Yield Estimation for Apple Orchards. In *Experimental Robotics*; Springer Tracts in Advanced Robotics; Springer: Berlin/Heidelberg, Germany, 2013; pp. 745–758. [CrossRef]

47. Hung, C.; Underwood, J.; Nieto, J.; Sukkarieh, S. A feature learning based approach for automated fruit yield estimation. In *Springer Tracts in Advanced Robotics*; Mejias, L., Corke, P., Roberts, J., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 105, pp. 485–498. [CrossRef]
48. Bargoti, S.; Underwood, J. Image classification with orchard metadata. In Proceeidngs of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; Volume 2016; pp. 5164–5170. [CrossRef]
49. Bargoti, S.; Underwood, J.P. Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards. *J. Field Robot.* **2017**, *34*, 1039–1060. [CrossRef]
50. Zhong, G.; Huang, K. *Semi-Supervised Learning: Background, Applications and Future Directions*; Nova Science Publishers, Inc.: Hauppauge, NY, USA, 2018.
51. Pise, N.N.; Kulkarni, P. A Survey of Semi-Supervised Learning Methods. In Proceeidngs of the 2008 International Conference on Computational Intelligence and Security, Suzhou, China, 13–17 December 2008; Volume 2, pp. 30–34.
52. Zhu, X.J. *Semi-Supervised Learning Literature Survey*; Technical Report; University of Wisconsin-Madison Department of Computer Sciences: Madison, WI, USA, 2005.
53. Roy, P.; Kislay, A.; Plonski, P.A.; Luby, J.; Isler, V. Vision-based preharvest yield mapping for apple orchards. *Comput. Electron. Agric.* **2019**, *164*, 104897. [CrossRef]
54. Goldberger, J.; Gordon, S.; Greenspan, H. An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures. In Proceedings of the 9th IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 1, pp. 487–493.
55. Blignaut, P. Fixation identification: The optimum threshold for a dispersion algorithm. *Atten. Percept. Psychophys.* **2009**, *71*, 881–895. [CrossRef]
56. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **1998**, *124*, 372–422. [CrossRef]
57. Jacob, R.J.K. What You Look at is What You Get: Eye Movement-Based Interaction Techniques. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 90), Seattle, WA, USA, 1–5 April 1990; Association for Computing Machinery: New York, NY, USA, 1990; pp. 11–18. [CrossRef]
58. Irwin, D.E. *Eye Movements and Visual Cognition: Scene Perception and Reading*; Springer: New York, NY, USA, 1992; pp. 146–165. [CrossRef]
59. Jacob, R.J.K. Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces. *Adv. Hum. Comput. Interact.* **2003**, *4*, 151–190.
60. Salvucci, D.D.; Goldberg, J.H. Identifying Fixations and Saccades in Eye-Tracking Protocols. In Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA'00), Palm Beach Gardens, FL, USA, 6–8 November 2000; Association for Computing Machinery: New York, NY, USA, 2000; pp. 71–78. doi:10.1145/355017.355028. [CrossRef]
61. Manor, B.R.; Gordon, E. Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *J. Neurosci. Methods* **2003**, *128*, 85–93. [CrossRef]
62. Duchowski, A. *Eye Tracking Methodology*; Springer: London, UK, 2007. [CrossRef]
63. Shic, F.; Scassellati, B.; Chawarska, K. The incomplete fixation measure. In Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, Savannah, GA, USA, 26–28 March 2008; ACM Press: New York, NY, USA, 2008; p. 111. [CrossRef]
64. Spakov, O.; Miniotas, D. Application of Clustering Algorithms in Eye Gaze Visualizations. Available online: https://pdfs.semanticscholar.org/b016/02b60a1fcb1ca06f6af0d4273a6336119bae.pdf (accessed on 21 June 2020).
65. Zagoruyko, S.; Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. *arXiv* **2016**, arXiv:1612.03928.
66. Safren, O.; Alchanatis, V.; Ostrovsky, V.; Levi, O. Detection of green apples in hyperspectral images of apple-tree foliage using machine vision. *Trans. ASABE* **2007**, *50*, 2303–2313. [CrossRef]
67. Gené-Mola, J.; Vilaplana, V.; Rosell-Polo, J.R.; Morros, J.R.; Ruiz-Hidalgo, J.; Gregorio, E. KFuji RGB-DS database: Fuji apple multi-modal images for fruit detection with color, depth and range-corrected IR data. *Data Brief* **2019**, *25*, 104289. [CrossRef] [PubMed]

68. Häni, N.; Roy, P.; Isler, V.; Hani, N.; Roy, P.; Isler, V. MinneApple: A Benchmark Dataset for Apple Detection and Segmentation. *IEEE Robot. Autom. Lett.* **2020**, *5*, 852–858. [CrossRef]
69. Kang, H.; Chen, C. Fruit detection and segmentation for apple harvesting using visual sensor in orchards. *Sensors* **2019**, *19*, 4599. [CrossRef]
70. Liu, X.; Jia, W.; Ruan, C.; Zhao, D.; Gu, Y.; Chen, W. The recognition of apple fruits in plastic bags based on block classification. *Precis. Agric.* **2018**, *19*, 735–749. [CrossRef]
71. Liu, X.; Zhao, D.; Jia, W.; Ji, W.; Sun, Y. A Detection Method for Apple Fruits Based on Color and Shape Features. *IEEE Access* **2019**, *7*, 67923–67933. [CrossRef]
72. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2281. [PubMed]
73. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979**, *28*, 100. [CrossRef]
74. Sahin, A. SensoMotoric Instruments launches SMI Eye Tracking. Available online: https://en.wikipedia.org/wiki/SensoMotoric_Instruments (accessed on 21 June 2020).
75. Achanta, R.; Estrada, F.; Wils, P.; Süsstrunk, S. Salient region detection and segmentation. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; ICVS 2008; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5008; pp. 66–75. [CrossRef]
76. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
77. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1597–1604. [CrossRef]
78. Wang, C.; Chen, J.; Li, W. Review on superpixel segmentation algorithms. *Appl. Res. Comput.* **2014**, *31*, 6–12.
79. Wang, M.; Liu, X.; Gao, Y.; Ma, X.; Soomro, N.Q. Superpixel segmentation: A benchmark. *Signal Process. Image Commun.* **2017**, *56*, 28–39. [CrossRef]
80. Stutz, D.; Hermans, A.; Leibe, B. Superpixels: An evaluation of the state-of-the-art. *Comput. Vis. Image Underst.* **2018**, *166*, 1–27. [CrossRef]
81. Levinshtein, A.; Stere, A.; Kutulakos, K.N.; Fleet, D.J.; Dickinson, S.J.; Siddiqi, K. Turbopixels: Fast superpixels using geometric flows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2290–2297. [CrossRef]