Evaluation of a Body Photo Browser*

Bastiaan Weijers 3256669
Colin Smits 4075390

April 9, 2015

Abstract

A computer vision systems recognising gestures in order to operate a photo browser is reviewed. The system evaluation focuses mainly on the accuracy of the system. We see that some gestures are confused, limiting the system's overall performance. The user evaluation is carried out by usability questionnaires. An overall evaluation is given using the PSSUQ, which shows that the system does not meet the standards set for a computer vision system. Parts of the system are evaluated using ASQ and SEQ, which show difference between tasks within the system. Linking both, we see that the way the system is running short of recognising gestures influences the user's perception of the system.

Keywords: Computer vision, gesture recognition, system evaluation, user evaluation

1 Introduction

The ease of using systems has become more important through the years. In order to check the usefulness, we need to evaluate the system. In this paper, we will discuss the usability of a body photo browser system, as will be defined later this section. The need for this is quite clear: the system is built for a user who wants a fluent experience, as well as a clear view of how to use the system. Taking this into consideration, we will discuss the system both from a system's perspective and the user's perspective.

The purpose of this evaluation is mainly to identify pitfalls in the system's accuracy, as well as recognising limiting factors for a satisfying user experience. We want to know what adaptations must be made to the current system in order to create an overall better performing system in terms over ease, satisfaction and accuracy.

In section 2, the system's reliability concerning robustness and accuracy will be discussed. After that, the usability of the system regarding ease of

*Word Count:2957

use and intuitiveness will be discussed in section 3. Last but not least, a conclusion will be drawn based upon the evaluations in sections 2 and 3.

1.1 The system

The system is a body photo browser. Using the system, one is able to scroll through and rotate pictures in the Windows Photo Gallery software The following gestures are recognised by the system:

Swipe Left: Using one hand (in view), going from right to left in the view. This will give the previous photo.

Swipe Right: Using one hand (in view), going from left to right in the view. This will give the next photo.

Rotate left: Using two hands (in view), going up with the right hand, and simultaneously going down with the left hand. This mimics the idea of rotating a picture counter clockwise.

Rotate Right: Using two hands (in view), going up with the left hand, and simultaneously going down with the right hand. This quite mimics the idea of rotating a picture clockwise.

The system works by recognising skin colour, then tracking the hands blobs through the screen, and depending on the total offset measured, a gesture might or might not be recognised and handled.

2 System Evaluation

In this chapter we will discuss how the system evaluation was performed and what the results were of this evaluation.

2.1 Methods and Environment

To quantify the performance of the Body Motion program, we setup a test environment to measure how well the system can translate user interaction to concrete system actions. The test setup is contained in an ecologically sound environment, in this case a bedroom with closed curtains was used with as a primary light source three led bulbs lighting the room. There was a single participant used to create quantifying data, whereas the particular user was well introduced to the workings, capabilities and limitations of the system. An analysis of how the system performs when used by lesser informed participants can be found later on in Section 3 - User Evaluation. The software was used with a YCrCb spectrum at [.. ...] and a minimum delay of $70ms$ between actions, further settings were left at program-default.
For each action performed in the video recording, a chart was updated with the value the system interpreted it to be. Possible predicted and actual actions were: Next, Previous, CW rotation, CCW rotation and Nothing. Whereas "Nothing" is represented as an idle action wherein nothing happens. It is worth noting that this particular action can only be actively observed on the actual outputs of the system, as predicting idle actions is not possible in our system.

		Predicted action				
		Next	Previous	CW	CCW	Nothing
Actual action	Next	34	0	0	0	0
	Previous	1	32	0	0	2
	CW	0	0	22	4	6
	CCW	0	0	0	14	0
	Nothing	2	5	16	21	0

Table 1: Testresults System Evaluation, $n = 159$

2.2 Results

To achieve accurate and representing results, we have recorded a total of 159 actions that we will measure its interpretation by system off, this accounts to roughly 40 actions per possible move. The number was picked as such because a $n > 30$ will give us a normal distribution over our data [ref needed]. All actions are recorded in a total of two video's, this allows the system to run in a continuous mode, without having to reinitialize after every action. This continuous mode should yield results that are more ecologically sound as opposed to having system global variables reinitialized with every action.
The evaluated output of a the actions will be tested against their expected outputs, the result of this can be found in Table 1. The data from the table shows the Predicted actions against the Actual actions delivered by the system for every possible action. From the table one can read how many predictions actually translated to correct behavior of the system, and how many did something unexpected.
Table 1 can be parsed to a so-called Confusion Matrix, which is depicted in Table 2.2. This matrix is a further generalization and will help us in the question on how well our system performs. Stated before (????, dit is wel in de presentatie gezegd maar nemen we 't op in de paper?), a preferred bias exists in having low false positives as this might be experienced by end-users as undesired behavior. From looking at the Confusion matrix we strongly suspect that we have satisfied this condition, but to confirm this we need to look at the Precision and Recall and establish a PR-curve. hier een stuk over precision, recall en pr-curve.
moeten we nog een H0 opstellen voor de confusion matrix? We kunnen misschien proeven dat false positive ¡ false negative?
Verder zien we uit Tabel 1 al dat de PR-curve voor swift-left/right er heel anders uit zal zien dan de PR-curve voor rotation. Die laatste is gewoon beroerd, terwijl de precision en recall voor left/right vrij goed zullen zijn. Dit apart testen denk ik in een subhypothese?

The data as presented can be further parsed into a confusion matrix which is shown in Table 2.2.

		prediction outcome		
		p	**n**	**total**
actual outcome	$\mathbf{p}'$	True positive 102	False negative 44	P'
	$\mathbf{n}'$	False positive 5	True negative 0	N'
	total	P	N	

Table 2: Confusion Matrix

2.3 Discussion

Hier discussie over de resultaten. Hoewel de resultaten nog ontbreken, weten we waarschijnlijk al wel waar de resultation heengaan. Overall scores zullen 'mediocre' zijn at-best. Tenzij we de hypotheses opsplitsen en apart kijken naar swipe-left/right en rotatie. Dan kunnen we misschien iets zeggen over het verschil daarin. We kunnen dan ook "recommendations" maken.

3 User Evaluation

Now that we have had a look from a system point-of-view, we will now evaluate the system from a user perspective. First, the used methodology will be discussed. After that, the results and discussion will follow.

3.1 Methods and Environment

In our user evaluation, we want to see whether our system is intuitive and easy to use. In order to do that, we need to have a number of people testing our system. First of all, we need to define the environment in which we will be testing. Using the notion of ecological validity, which states that the surroundings in which the research takes place must resemble a real-world setting, the tests will take place in normal rooms in terms of for example lighting and background. A laboratory would not be suitable as the user will react differently. The tests are then carried out as follows.

First of all, we ask the user to take a look at the system. We have defined four different gestures, and for each of the different options, we ask the user what would be his first opinion on how to perform the task. The user will give his own gesture which he would use in order to perform the action. That way, we can compare the thoughts of users to our own

Scenario	Task
1	Navigate to the next photo
2	Navigate to the previous photo
3	Rotate the photo clockwise
4	Rotate the photo counter-clockwise
5	Navigate to the second next photo (twice to the next)
6	Rotate the previous photo clockwise
7	Rotate the photo and put it back up
8	Navigate to the photo containing [something]
9	Put the current (rotated) photo back up
10	Navigate to the rotated photo and put it back up

Table 3: *Specified scenarios in the user tests*

throughts on the used gestures. After that, we set up 10 scenarios, as shown in Table 3. The first four scenarios contain each of the different gestures in order to make the user comfortable with the system. After that, we start combining the gestures so that the user should be thinking about what gestures he must use in which order. Finally, we do not specify the gestures any more, but we describe the task by specifying pictures the user must recognise. In this way, we can test whether the user can perform the right actions while still thinking of what to do. The last part mainly shows how intuitive and easy our system can be used, as the user will be distracted from the gestures.

Secondly, in order to get measurements, we set up a questionnaire. According to Sauro and Lewis [1], and Tullis and Albert [2], there are different standardized questionnaires. For our system, we want to know whether the user is satisfied with the system as a whole. As defined by Sauro and Lewis, the PSSUQ (Post-study System Usability Questionnaire) is "designed to access users' perceived satisfaction with computer system applications" [1]. Looking at our goal, we think this questionnaire fits our research. Therefore, we incorporated this questionnaire at the end of the test, after all the scenarios have been completed. We chose the PSSUQ over the SUS (System Usablity Scale), which can also be used as a measure of the usability of a system, due to an increased reliability [1].

In addition, we thought that the user should be able to tell us what he perceived during each scenario. For that, we set up a combination of the ASQ (After Scenario Questionnaire) and the SEQ (Single Ease Question). The AQS was developed by Lewis in order to describe the user's satisfaction with completing a certain task [1][3]. This can be perfectly used for our research goal. Moreover, the ASQ is linked to the PSSUQ, and also describes the user's satisfaction with a system, in this case a certain task. Furthermore, as the name suggests, the SEQ is used to describe a user's view on the ease of completing a task. The combination is used to get more reliable results.

3.2 Results

To start off, we look at the user's thoughts on the gestures to be used. In Table 4, we show the percentage of user's defining a certain gesture for an action. We can see that for the navigation actions, the users differ in view, whereas for rotations, all the user say the same.

Action	**Users opinions**	%
Next photo	Hand left to right	75
	Hand right to left	**25**
Previous photo	**Hand left to right**	**25**
	Hand right to left	75
Rotate left	Rotate hands in circle counter-clockwise	100
	Left hand down, right hand up	**0**
Rotate Right	Rotate hands in circle clockwise	100
	Left hand up, right hand down	**0**

Table 4: *Comparison of the implemented gestures (in* **bold***) with the user's opinions on the gestures to be used.* $n = 4$

After that, we need to examine the results from the questionnaires. We will start by analysing the scenario questionnaires, the SEQ and ASQ. Firstly, the SEQ results are shown in Table 5. On first glance, we can see that for the scenarios containing only navigating to photos, the overall score is low (< 3), as opposed to those only containing rotations (>4). Furthermore, the more complicated tasks show a broad confidence interval, meaning that there is much variation between users as far as ease of use is concerned.

[ASQ results will follow in the final paper]

Now we need to study the user's satisfaction overall. In Table 6, the results from the PSSUQ questionnaire are shown.

3.3 Discussion

From the results obtained for the user's opinion on the gestures, we can conclude that the navigation is something that matters from person to person. The intention of the implemented gesture was to simulate a swipe on, for example, a tablet. To go to the next photo, one must then swipe from right to left, seemingly pushing the photo roll to the left. However, users perceive that it could also be seen as the direction to where one wants to go. In this case, the meaning of a swipe to the left is that the

Scenario	Description	Lower	Mean	Upper
1	Navigate to the next photo	0.58	1.5	2.42
2	Navigate to the previous photo	0.58	1.5	2.42
3	Rotate the photo clockwise	2.45	4.5	6.56
4	Rotate the photo counter-clockwise	2.45	4.5	6.56
5	Navigate to the second next photo	0.73	2.25	3.77
6	Rotate the previous photo clockwise	0.40	3.0	5.60
7	Rotate the photo and put it back up	1.25	3.25	5.25
8	Navigate to the photo containing [something]	1.25	3.25	5.25
9	Put the current (rotated) photo back up	0.10	3.0	5.90
10	Navigate to the rotated photo and put it back up	0.70	2.0	3.30

Table 5: *Results of the SEQ, showing the mean and 95% confidence intervals for each of the scenarios.* $n = 4$

Question	Lower	Mean	Upper
1	1.95	2.75	3.55
2	1.23	2.75	4.27
3	1.95	2.75	3.55
4	0.75	3	5.25
5	1.23	2.75	4.27
6	1.73	3.25	4.77
7	3.36	5.75	8.14
8	2.70	4	5.30
9	1.74	4.5	7.26
10	2.75	5	7.25
11	2.36	4.75	7.14
12	2.75	5	7.25
13	0.75	3	5.25
14	2.46	3.25	4.05
15	2.58	3.5	4.42
16	1.73	3.25	4.77
Avg. 1-6	1.47	2.88	4.28
Avg. 7-12	2.61	4.83	7.05
Avg. 13-15	1.92	3.25	4.57
Avg. 1-16	2.00	3.70	5.40

Table 6: PSSUQ results showing the mean for each question compared to the 95% confidence intervals. $n = 4$

photo to the left should be shown, thus the previous. This is something that has to be worked out further. Another striking observation is that users all have the same opinion on rotation. However, this differs from our

own implementation. Looking closely to the implementation, however, we see that the gesture implemented resembles a rotating movement without circulation. The circular limitation still hinders smooth use of the system, as can also be seen in the SEQ.

The SEQ shows us that the system's easy lies in the navigation, and that the rotation is more difficult. As stated before, the cause of this might be the implementation of the gestures as is. The navigation issue is resolved due to the fact that the right gestures are explained after the first questions (Section 3.1.

From the PSSUQ, we can conclude that the system's performance is not good enough, when compared to the PSSUQ norms [1].

Overall, for the user evaluation, we can generally say that when only navigation is incorporated, the user is able to do tasks with ease. On the contrary, when rotation is included, the user will start to struggle. Nonetheless, we still have to be careful concluding from our result set, as the sample size is much too low ($n = 4$) in order to get a clear representation.

4 Discussion

All in all, when we link the system evaluation to the user evaluation, we can see one striking property already. Due to problems with the implementation of the rotating gestures, the system performs poor and the user cannot use the system according to plan.

In the system evaluation, we have seen that rotating gestures are easily confused with the other rotation, and that the system sometimes does not recognise any gesture at all when rotating. In the user evaluation, we have seen these properties return in the SEQ, where scenarios involving rotation were seen less satisfactory to the user. This implies that for further research, the system's implementation of rotation must be reviewed and edited, in order to get a more accurate and more satisfying system.

References

[1] J. Sauro, J. R. Lewis: *"Quantifying the User Experience"*, Elsevier, 2012

[2] T. Tullis, W. Albert: *"Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics" (2nd Edition)*, Elsevier, 2013

[3] J. R. Lewis: *"IBM computer usablity satisfaction questionnaires: Psychometric evaluation and instructions for use"*, International Journal for Human Computer Interactions, 7 (1995), 57-78

[4] M. Brewer: *"Research Design and Issues of Validity"*. In: H. Reis, C. Judd (eds), *"Handbook of Research Methods in Social and Personality Psychology"*, Cambridge University Press, 2000