Evaluation of a Body Photo Browser

Bastiaan Weijers   Colin Smits
3256669   4075390

April 8, 2015

Abstract

Here comes the abstract

keywords:

1 Introduction

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.
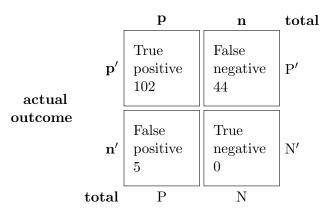
prediction outcome

		p	n	total
actual outcome	**p′**	True positive 102	False negative 44	P′
	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

4   Discussion

References