Kinematic analysis of shoulder motion for diagnostic purposes

Kasper van der Hoofd, Luke de Keijzer, Vincent van den Oord, Rogier Zitman, Dr. Tony Andrioli

Abstract

- Objective: Determining whether kinematic analysis of shoulder motion is useful for diagnostic purposes.
- The data and its preparation: The original data were time frames created by the 'Flock of birds'-system that captured the location and rotation of seven parameters that were placed on patients while they executed exercises. Later in this study we cleaned all measurements by cutting out the 'noise' and splitting measurements so that they only contain one execution of an exercise.
- Research strategy & execution: Multiple classifiers were built to find the best classifier with the highest accuracy and lowest recall. All classifiers can be divided in three categories; sample-level, patient-level and exercise-level.
- Results: The classifiers can be categorized in sample level and patient level. The results used during this study are the Precision, Accuracy and Recall score on the test data.
- Discussion: It is recommended to take the results of the final classifier into account when forming the final diagnosis, yet not to base it solely on those results.
- Conclusion: Kinematic analysis of shoulder motion for diagnostic purposes is possible based on our findings.

Introduction

Before specialists can diagnose a patient's shoulder condition patients must do multiple exercises to give the specialist an impression of the shoulder mobility. The examining specialist usually wants to determine the angle where the patient experiences pain during these exercises (Flynn, Cleland, & Whitman, 2008). This angle is called the 'painful arc' and is measured with a manual instrument called the 'goniometer', which basically is a rather large protractor. Because this angle is measured manually it could be argued if this instrument is precise enough to accurately measure this painful arc compared to other measuring methods like radiography (McVeigh, Murray, Heckman, Rawal, & Peterson, 2016). The painful arc is important because it is used by specialists to diagnose a patient's shoulder condition to determine the optimal treatment. The project group to work on this study prior to us suggested to take a look at the Flock-of-Birds system, or FoB. The FoB-system uses seven sensors on the human body while a patient executes a series of motions. That team's paper (de Koning, Enthoven, van Dijk, Slemmer, & Lindeboom, 2018) suggested investigating the symmetry of the executed exercises to determine a patient's condition. This question was also incorporated in this research. It has been suggested by other researchers to investigate "whether kinematic analysis of shoulder motion is useful for diagnostic purposes" (Kolk, et al., 2017). On account of this suggestion and the suggestion of the previous team, this became the leading question of this study.

The data and its preparation

I. Original data

One of the authors/collaborators (J. de Groot) on the paper that suggests investigating "whether kinematic analysis of shoulder motion is useful for diagnostic purposes" (Kolk, et al., 2017) has access to the 'Flock of birds'-system that can accurately measure all the angles between the joints using seven sensors. This system captures the location and rotation of these seven sensors using an electromagnetic field (Krijgsman & Curet, 2009) and stores these locations and rotations in time-series. These time-series have 24 columns that are filled with the x, y, z-location and the rotation of all seven sensors that were placed on the joints of patients. In an earlier study (Meskers, de Groot, Arwert, Rozendaal, & Rozing, 2004) it has been shown that the rotation of joints can be accurately measured with sensors with 95% accuracy. During this study anonymized patient data, generated at the Leiden University Medical Center (LUMC) using this 'Flock of birds'-system, has been analyzed. The data that we received and used during this study contained measurements of patients performing five unique exercises. Because this data has been collected during four different studies in the past all measurements and therefore data hasn't been collected according to the same protocol. There could be noticeable differences in the data simply because of this reason. The patients in the dataset were grouped into four different categories by the LUMC. The type of shoulder condition, severity of the shoulder condition or the meaning each category was unknown in order to further preserve the privacy and anonymity of the patients. As this study progressed, we gradually gained more insight in the meaning of these categories.

II. Cleaned data

Besides using the original data that we received for this study we also inspected and cleaned the data after discovering that there was a lot of noise and that sometimes exercises were done multiple times during a single measurement. The noise that we cut out at

the beginning and ending of measurements were (according to us) unnecessary frames or frames that didn't seem to be a part of the exercise. These frames were usually very still where the angles didn't change much or sets of frames where the angles were very random. Our assumptions were that when the frames were very still that the specialist started recording a measurement some time before a patient executed an exercise. We also assumed that when the frames were very random (before and after we saw the graph of an exercise) something strange happened with the sensors before or after the recording of an exercise. We assumed that this could be the adjusting of some sensors or a patient experimenting with the sensors and behaving strangely. We wanted to know whether filtering out the noise and splitting the measurements in separate exercise would improve the accuracy and lower the recall of all classifiers. This cleaning of the data was done manually, and we did this cleaning and splitting by looking at the graph of the angles of the sensor that was placed on the Humerus. The graph of the Humerus was used because it appeared that this graph of the Humerus always behaves differently at each exercise. As said earlier the cleaning was done by looking at the graph and trying to identify a pattern that could have been an exercise. The unnecessary still and very random frames were deleted. When we thought that an exercise could have been repeated multiple times in a single measurement we split the measurement and deleted the unnecessary frames on two or more newly split measurements. The picture (1) below is an example of how the data was cleaned.

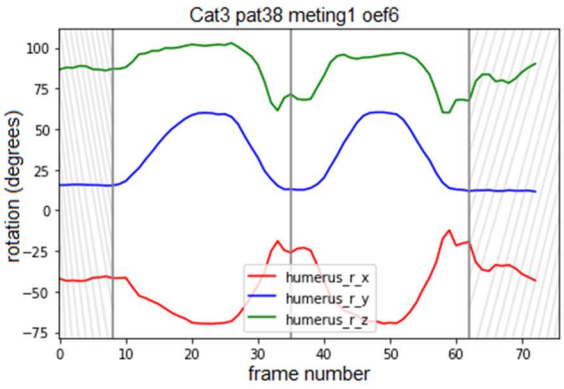

Figure 1. Cleaning and splitting a measurement

At the beginning all classifiers were trained and tested with the entire data that was provided to us by J. de Groot from the Leiden University Medical Center. Later on we used the new cleaned dataset for the training of some classifiers.

Research strategy & execution

The goal of this study is to investigate whether kinematic analysis of shoulder motion is useful for diagnostic purposes. Since the data that was available to us didn't contain any information about specific shoulder conditions our goal was investigating the possibility to determine to which category a patient belongs based only on the data. Because this is a classification problem the goal is finding the optimal classifier that can accurately predict to which category a patient belongs by providing it either a single frame, or a time-series of a measurement. The execution of this research can be split into two parts. During this first part the classifiers were trained and tested on single data 'frames'. The measurements were originally time-series. All these measurements contained around 40-400 frames of data that contained the location and rotation of all sensors. For this first part all measurements were split in single 'frames' that we trained our (sample-level) classifiers were trained on. We tried to determine the category of patients by grouping these single 'frames'. During the second part of this study the classifiers we built were patient-level classifiers. We didn't train these classifiers on parameters that were true for single 'frames' but we trained them on parameters that were true for patients.

I. Method of testing

After receiving the data for this study we set aside 10% of all data. This was done so that the classifiers could be tested with data that they weren't trained with. We randomly chose 10% of the patients from each category and set all their measurements apart. The splitting of the data is done on patient level, meaning the there is no data from one patient in both datasets. The remaining data (90%) was used for training (80%) and testing (20%) all classifiers.

II. Sample-level classifiers

The first classifier that was built during this study was a simple classifier that used logistic regression. It compared one category against the remaining three categories. This classifier (no. 1 - 4) was run four times and every category was compared to all three other categories (1, 2, 3, 4). After we did this we grouped the amount of correctly predicted samples per category for every patient. Then we trained a (5in1) classifier (no. 5) on these results and it predicted a category that each patient belongs to. After building classifier no. 5 we built a k-NN (k-nearest neighbor) classifier. We ran this classifier four times (no. 6 - 9) and tried to classify one category while training it on all four categories. Each time we ran this classifier we ran it 10 times (each time with a different k) so that we would find the

optimal k for every category so to get the optimal accuracy and recall. After trying our chances with the k-NN classifiers we were curious if our classifiers would perform better if they were trained with data from just two categories. For this we used the simplest linear regression classifiers we had built so far and trained them with the data samples from just two categories. After this we tested how many samples it had predicted correctly. We ran this classifier 12 times (no. 10 - 21) because we compared all categories against all other categories. After running these classifiers, we gained some useful insights about the data and the challenges we face. Because the accuracy of the classifier that was trained on category two and three had the lowest accuracy (67%) we decided we wanted to try improving this accuracy by manipulating the input parameters. To accommodate for non-linear regression boundaries, we simply modified the entire dataset that this classifier was trained on. The first time (no. 22) we ran it we squared the input parameters, the second time (no. 23) we manipulated the input by taking the root of all values, and the third and last time (no. 24) we ran this classifier we took the natural logarithm of all input parameters.

Before we moved on to building patient-level classifiers we also experimented with a few neural networks. We simply changed most of the classifiers that were mentioned earlier (no. 1 – 24) to MLP-classifiers (neural networks). Because none of the results were significantly better than the original classifiers no. 1 – 24 none of the results were incorporated in table 1.

III. Patient-level classifiers

As explained earlier patient level classifiers are classifiers that are trained on values that are characteristic for entire patients. The first patient-level classifier we built was a simple logistic regression classifier. We used the maximum values of all exercises per patient as input/training parameters. The maximum was the only possible way to display all exercises in one value (per parameter). We ran this classifier 12 times (no. 25 - 36) because we compared all categories against all other categories. After running this classifier we ran it again 12 times but instead of using the maximum we used the absolute maximum. These classifiers (no. 37 - 48) gave similar but different results. After comparing individual categories to all other categories we built an multivariate linear regression classifier (no. 49) that was trained on the same training set as classifiers no. 37 - 48 (the absolute maximum). After running these classifiers we wanted to build an classifier that was trained on something different than simply the locations and rotation of all sensors. To take in account more than simply the movements of the measurement, a specific energy related variable was used as well. This variable was calculated from the graph that showed the angle of the movement VS the acceleration with which the angle changed per sample. A linear regression classifier (no. 50) was trained on this 'energy'-variable. This classifier was tested on the cleaned data. After building these classifiers we built our final classifier of this study. This multivariate classifier operates on patient level and contains lines that say something about a patient as a whole. To fully cover a patient, multiple lines are used in this classifier. The training data is generated by taking the data from the five unique exercises , each consisting of; the energy variable for the X-, Y- and Z-axis, the extreme values of the sensor placed on the Humerus for also the X-, Y- and Z-axis and five samples in the theta/sample graph taken at 10%, 25%, 50%, 75% and 90% of the graph on every one of the 24 variables . Then, all this data is cross-joined for the five unique exercises resulting in a csv file of 49.510 lines, of which 43998 (88,8%) are used for training and the other 5515 (11,2%) lines are set aside to serve as the testing dataset. The cross-joining was done because some patients had multiple measurements for the same exercises. Because we didn't know which of all measurements had to be selected as measurement for an exercise we cross-joined them so that all possible combinations of measurements for all exercises would be included in the dataset. After seeing the results we decided to test this classifier with the data that was set aside at the beginning of this study. The results that are given in table 1 for classifiers no. 51 – 58 are the results of this 10% data that was set aside at the beginning of this study. To see whether our cleaning of the measurements had any effect we ran the same classifier with the original data (no. 51 – 54) and the data that was cleaned by us (no. 55 – 58). For these classifiers (no. 51 – 58) both the original and cleaned data has been manipulated in the same way (as described above).

Results

As shown in the table below (table 1. Overview of classifiers and their performance.), the first 24 classifiers operate on sample level and the other 34 on patient level, meaning the input data consists of each sample individually or each individual patient respectively.

For every classifier where the value in the "Additional explanation" column is "Category X vs Y", means that the samples/patients that are correctly

categorized as category X are seen as the 'true positives'.

Furthermore, the results of the classifiers no. 50 and 55-58 are generated by using the cleaned dataset.

No.	Additional explanation	Level	Accuracy	Recall	Precision
1	Category 1 vs 2,3,4	Sample	0.86	0.419	0.378
2	Category 2 vs 1,3,4	Sample	0.463	0.177	0.68
3	Category 3 vs 1,2,4	Sample	0.698	0.489	0.448
4	Category 4 vs 1,2,3	Sample	0.975	0.427	0.535
5	5 classifiers combined	Sample	0.667	-	-
6	k(2)-NN classifier category 1	Sample	-	0.808	0.844
7	k(3)-NN classifier category 2	Sample	-	0.524	0.569
8	k(4)-NN classifier category 3	Sample	-	0.643	0.694
9	k(2)-NN classifier category 4	Sample	-	0.958	0.963
10	Category 1 vs 2 (True vs False)	Sample	0.67	0.621	0.439
11	Category 1 vs 3	Sample	0.858	0.745	0.837
12	Category 1 vs 4	Sample	0.888	0.986	0.877
13	Category 2 vs 1	Sample	0.67	0.69	0.822
14	Category 2 vs 3	Sample	0.645	0.722	0.687
15	Category 2 vs 4	Sample	0.915	0.987	0.921
16	Category 3 vs 1	Sample	0.858	0.92	0.868
17	Category 3 vs 2	Sample	0.645	0.538	0.579
18	Category 3 vs 4	Sample	0.839	0.936	0.879
19	Category 4 vs 1	Sample	0.888	0.609	0.941
20	Category 4 vs 2	Sample	0.915	0.392	0.806
21	Category 4 vs 3	Sample	0.839	0.339	0.506
22	Classifier no. 18 with manipulated parameters (^2)	Sample	0.493	0.52	0.661
23	Classifier no. 18 with manipulated parameters (sqrt)	Sample	0.467	0.469	0.648
24	Classifier no. 18 with manipulated parameters (e-log)	Sample	0.465	0.476	0.642
25	Category 1 vs 2 .MAX	Patient	0.6	0.571	0.8
26	Category 1 vs 3 .MAX	Patient	0.889	0.833	1
27	Category 1 vs 4 .MAX	Patient	1	1	1
28	Category 2 vs 1 .MAX	Patient	0.6	0.667	0.4
29	Category 2 vs 3 .MAX	Patient	0.727	0.857	0.75
30	Category 2 vs 4 .MAX	Patient	0.857	0.857	1
31	Category 3 vs 1 .MAX	Patient	0.889	1	0.75
32	Category 3 vs 2 .MAX	Patient	0.727	0.	0.667
33	Category 3 vs 4 .MAX	Patient	1	1	1
34	Category 4 vs 1 .MAX	Patient	1	1	1
35	Category 4 vs 2 .MAX	Patient	0.857	0	0
36	Category 4 vs 3 .MAX	Patient	1	1	1
37	Category 1 vs 2.MAXABS	Patient	0.7	0.857	0.75
38	Category 1 vs 3 MAX.ABS	Patient	0.889	0.833	1
39	Category 1 vs 4 MAX.ABS	Patient	1	1	1
40	Category 2 vs 1 MAX.ABS	Patient	0.7	0.333	0.5
41	Category 2 vs 3 MAX.ABS	Patient	0.727	0.857	0.75
42	Category 2 vs 4 MAX.ABS	Patient	0.857	1	0.857
43	Category 3 vs 1 MAX.ABS	Patient	0.889	1	0.75
44	Category 3 vs 2 MAX.ABS	Patient	0.727	0.5	0.667
45	Category 3 vs 4 MAX.ABS	Patient	1	1	1
46	Category 4 vs 1 MAX.ABS	Patient	1	1	1
47	Category 4 vs 2 MAX.ABS	Patient	0.857	0	0
48	Category 4 vs 3 MAX.ABS	Patient	1	1	1
49	Multivariate classifier MAX.ABS	Patient	-	0.5	0.781

50	Energy classifier (Cleaned data)	Patient	-	0.247	0.151
51	Final classifier (Original data) cat 1 vs 2, 3, 4	Patient	-	0.833	0.394
52	Final classifier (Original data) cat 2 vs 1, 3, 4	Patient	-	0.828	0.860
53	Final classifier (Original data) cat 3 vs 1, 2, 4	Patient	-	0.825	0.283
54	Final classifier (Original data) cat 4 vs 1, 2, 3	Patient	-	0.683	0.884
55	Final classifier (Cleaned data) cat 1 vs 2, 3, 4	Patient	-	0.983	1
56	Final classifier (Cleaned data) cat 2 vs 1, 3, 4	Patient	-	1	0.993
57	Final classifier (Cleaned data) cat 3 vs 1, 2, 4	Patient	-	1	0.965
58	Final classifier (Cleaned data) cat 4 vs 1, 2, 3	Patient	-	0.977	1

Table 1. Overview of classifiers and their performance.

Discussion

We made to types of classifiers, the first are operate on sample level and the others on patient level.

The results from classifier no. 10-21 suggest that the shoulder problems increase in severity with the categories. Simply because category one and four differ rather much from each other in comparison to any other category. And because category two and three seem to differ not so much from each other.

Classifier 25-48 add to this suggestion. Note that the testing set for category four contains only one patient. Which is most likely why it scores perfect most of the times on accuracy and recall.

The final classifier (no. 51 – 58) performs by far the best of all the other classifiers. And its accuracy, precision and recall suffice in every situation according to the standards used in the medical world when it comes to distinguishing the categories from each other. Therefore, we cannot recommend to conclude the severity of the shoulder injury solely on the results of this classifier. Nevertheless, we do recommend taking it into account when forming the final diagnosis.

At last, the cleaning of the data had a very positive influence on the performance of the classifiers. And manipulating the value of different parameters did not.

Conclusion

According to our findings we think that it is possible to use kinematic analysis of shoulder motion for diagnostic purposes. We are basing this conclusion on the results of our final classifier (no. 55 - 58) that can classify a patient with a precision of 97.7% - 100%. An important step for achieving these results is the cleaning of the data and can conclude that this improves the performance of classifiers. Regardless of these results we think that it would be wise to continue research in this area and investigate a few topics that we couldn't tend to ourselves. Looking into these topics could improve the results that we achieved during this study and solidify our conclusion.

I. Manipulating parameters

During this study we briefly investigated whether parameter manipulation is useful for our classification problem. We quickly moved on and did not investigate this subject thoroughly enough so we cannot say whether it is useful or not based on our results. We suggest looking into this in the future.

II. Amount of data

At various points during this study we encountered problems because of our relatively small dataset that contained data of 144 patients. We often encountered these problems while testing our classifiers because we couldn't set enough data aside for testing because we had to use it for the training of the classifiers. Using a bigger dataset (and therefore training-set) would probably result in results (accuracy, recall, precision) that are more accurate. We had also quite a lot of problems with neural networks because of this small dataset. Because of the size the neural networks overfitted very easily and often it was very hard solving those issues. An easy fix of this problem is training them with more data (if it were available). We suggest obtaining a larger dataset and investigating neural networks again. A larger dataset would have made this entire study significantly easier.

III. Cleaning the data

During this study we cleaned the original data that we received by cutting out the 'noise' and splitting measurements when we thought an exercise was repeated multiple times during a single measurement. We cleaned this data quite intuitively and tried to stay as objective as possible. Because this cleaning was done manually by us it could be biased. We suggest investigating the cleaning of the measurements and finding the optimal and automatic way to clean the data that is generated by the 'Flock of birds'-system.

IV. Usage of parameters

For most of our classifiers we simply used all data from all seven sensors that were used during the recording of exercises. We haven't investigated which sensors are the most important and whether it is necessary to use all the data that we used for the training of our classifiers. Researching the impact and importance of every joint could be useful for future studies.

References

de Koning, C., Enthoven, B., van Dijk, H., Slemmer, B., & Lindeboom, R. (2018, 1). Schouderprobleem detectie met behulp van de Kinect.

Flynn, T., Cleland, J., & Whitman, J. (2008). *The Users' Guide to the Musculoskeletal Examination: Fundamentals for the Evidence-Based Clinician.* Louisville: Buckner.

Kolk, A., Henseler, J., de Witte, P., van Zwet, E., van der Zwaal, P., Visser, C., . . . de Groot, J. (2017, 4 11). The effect of a rotator cuff tear and its size on three-dimensional shoulder motion. *PubMed*.

Krijgsman, M., & Curet, P. (2009). *Final Report Flock of Birds System.* TUDelft.

McVeigh, K., Murray, P., Heckman, M., Rawal, B., & Peterson, J. (2016, 1 20). Accuracy and Validity of Goniometer and Visual Assessments of Angular Joint Positions of the Hand and Wrist. *PubMed*.

Meskers, C., de Groot, J., Arwert, H., Rozendaal, L., & Rozing, P. (2004). Reliability of force direction dependent EMG parameters of shoulder muscles for clinical measurements. *PubMed*.

scikit-learn. (2018). *sklearn.neural_network.MLPClassifier*. Opgehaald van scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html