How Good is Good Enough? The Impact of Errors in Single Person Action Classification on the Modeling of Group Interactions in Volleyball

Lian Beenhakker*

Fahim Salim

l.beenhakker@utwente.nl

f.a.salim@utwente.nl

University of Twente

Enschede, The Netherlands

Dees Postma

Robby van Delden

d.b.w.postma@utwente.nl

r.w.vandelden@utwente.nl

University of Twente

Enschede, The Netherlands

Dennis Reidsma

Bert-Jan van Beijnum

d.reidsma@utwente.nl

b.j.f.vanbeijnum@utwente.nl

University of Twente

Enschede, The Netherlands

ABSTRACT

In Human Behaviour Understanding, social interaction is often modeled on the basis of lower level action recognition. The accuracy of this recognition has an impact on the system's capability to detect the higher level social events, and thus on the usefulness of the resulting system. We model team interactions in volleyball and investigate, through simulation of typical error patterns, how one can consider the required quality (in accuracy and in allowable types of errors) of the underlying action recognition for automated volleyball monitoring. Our proposed approach simulates different patterns of errors, grounded in related work in volleyball action recognition, on top of a manually annotated ground truth to model their different impact on the interaction recognition. Our results show that this can provide a means to quantify the effect of different type of classification errors on the overall quality of the system.

Our chosen volleyball use case, in the rising field of sports monitoring, also addresses specific team related challenges in such a system and how these can be visualized to grasp the interdependencies. In our use case the first layer of our system classifies actions of individual players and the second layer recognizes multiplayer exercises and complexes (i.e. sequences in rallies) to enhance training. The experiments performed for this study investigated how errors at the action recognition layer propagate and cause errors at the complexes layer. We discuss the strengths and weaknesses of the layered system to model volleyball rallies. We also give indications regarding what kind of errors are causing more problems and what choices can follow from them. In our given context we suggest that for recognition of non-Freeball actions (e.g. smash, block) it is more important to achieve a higher accuracy, which can be done at the cost of accuracy of classification of Freeball actions (which are mostly plays between team members and are more interchangeable as to their role in the complexes).

*The work presented in this paper is based on the first author's MSc Thesis.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7581-8/20/10...\$15.00

<https://doi.org/10.1145/3382507.3418846>

CCS CONCEPTS

• **Mathematics of computing** → Probabilistic representations;
• **Human-centered computing** → Empirical studies in visualization;
• **Theory of computation** → Probabilistic computation.

KEYWORDS

machine learning; behaviour analysis; action recognition; multiplayer action modeling; volleyball; social network analysis

ACM Reference Format:

Lian Beenhakker, Fahim Salim, Dees Postma, Robby van Delden, Dennis Reidsma, and Bert-Jan van Beijnum. 2020. How Good is Good Enough? The Impact of Errors in Single Person Action Classification on the Modeling of Group Interactions in Volleyball. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3382507.3418846>

1 INTRODUCTION

Multimodal sensing and analysis plays a growing role in the field of physical exercise and sports training [11, 23, 25]. Sports data may deliver insights that improve training (e.g., [1, 14, 27]), or may serve as necessary input for new interactive sports exercises (e.g., [2, 15, 26]). In the Smart Sports Exercises (SSE) Project¹ we measure data from volleyball players, automatically recognise volleyball activity on individual level and group level, and use this as input for novel interactive training exercises.

In this paper we focus on the modeling of group interactions between volleyball players, and specifically analyse how this is impacted by errors in the underlying single person action detection and classification. Our modeling of interaction among players focuses on so-called “volleyball complexes”: typical sequences of single player actions in subsequent rallies at either side of the net [10]. Automated analysis of such patterns may for example contribute to insights regarding strengths and weaknesses of athletes.

However, such analysis also depends on automatic detection and classification of single athletes' actions (such as setup, or smash). Our action recognition is done through classifiers that use as input measurement data from Inertial Measurement Units (IMUs) which integrate several sensors in one wearable package.

Errors in the action recognition will lead to deterioration of the interaction modeling and threaten the validity of any insights drawn

¹<http://smart-sports-exercises.nl>

from it. We explore the impact of such errors in the classifiers by systematically simulating various possible patterns of recognition errors on top of a manually annotated ground truth.

The interaction modeling output based on the simulated erroneous action recognition is used to derive statements about the quality required of the single person action recognition for adequate modeling of volleyball interactions. We show that not all types of classification error have the same impact in that respect.

This paper offers several contributions:

- We offer a method for automatic modeling of interaction whereas existing work so far relies on manual observations.
- Compared to [17] our model requires less (subjective) information and has a more intuitive interpretation that should be easier to understand for users.
- We present an analysis that yields fundamental insight into how different levels and patterns of errors in action recognition impact the truth and usability of the interaction models
- We thus present an analytical approach that provides an idea of how automatic modeling might fare in real life and where potential areas of improvement are found.

The rest of the paper is as follows. First, related work in modeling of actions and interaction patterns in volleyball is discussed. Next, the layered pipeline of the SSE project for automatic analysis of volleyball activities is explained, followed by our approach to volleyball interaction modeling, our systematic simulation of classification errors in the underlying action recognition, and our analysis of the impact of these on the interaction modeling. We finish by discussing implications and limitations of our work.

2 BACKGROUND

2.1 Action Recognition in Volleyball

In (beach) volleyball, different studies have used machine learning in combination with IMU sensors, for example in skill recognition [29], serve type recognition [6] and action recognition [8, 16, 23]. Typical features used in the latter are median, mean, sd, skewness, kurtosis, dominant frequency, amplitude of spectrum at dominant frequency, (pos of) max, (pos of) min, correlation between x and y and between x and z. Cuspinera et al. [6], in contrast, did not use features but rather used templates to recognize patterns in the data.

2.2 Interactions in Volleyball Teams

Beniscelli et al. [3] describe “interaction” in team ball sports as a level of coordination among players. Analysing interaction gives information on who can perform what actions best (e.g., receiving, setting and spiking) and thus who should perform these during a match [20]. A broader view at interaction might give information on which of the team’s strategies work (or not) [28]. Part of interaction is also the communication about tactical decisions, verbally or non-verbally [21], such as hand gestures used by the setter. Players who are not involved in the selected attack strategy can then perform fake smashes to distract the opponent. Furthermore, if one player fails to perform an action, team mates might solve this by performing a non-scripted action to still try and save the ball [3].

One specific way to look at interaction in volleyball is by dividing a rally into separate “complexes” [10] (see Figure 1). A rally in

Figure 1: Rally complexes and their relation, adapted from [17].

volleyball starts when the ball is served and ends when a point is scored (the ball touches the field). A complex describes the individual player actions on one side of the net until the ball is passed to the other side (or touches the floor, thus ending the rally).

Hileno and Buscá [10] described the “Volleyball Attack Coverage Observation System” (VACOS-1), an observational tool to methodically observe these complexes in a rally, recording information about actions and their location in the field, as well as other variables such as the number of blockers and the tempo of the attack. Each complex defines a different stage of the rally, and has its own characteristics and difficulties, requiring somewhat different skills. It is therefore worthwhile for a team to have insight into which complexes, and more specifically which actions in these complexes, they generally perform well and which are more difficult to them. Variations of the tool have successfully been applied in several studies [12, 13, 17–19] to analyse patterns in volleyball rallies.

The structural nature of complexes lends itself well to graph based analysis. In some studies, complexes are modeled as a Social Network to perform statistical analysis on the occurrence of combinations of actions / sequences of certain types, in certain locations, and combined with other factors [12, 13, 17–19]. Drikos [7] models complex K1 as a Markov chain (with the actions performed as nodes, and edges giving the likelihood to go from one action to another) and uses that to calculate scoring probabilities and to model “stabilization” after certain types of serve. Similarly, Best [4] (an MSc Thesis) used a Markov models to predict how likely a team is to score at the end of a rally depending on the actions performed.

3 SSE PROJECT RECOGNITION PIPELINE

We employ a layered approach to analysing (group and individual) volleyball activities (see Figure 2), for several reasons. The output of the lower layers are useful to other goals as well, for example, a coach can get valuable insights about which player performs certain actions such as smash or serve more often than others. Apart from that, the layered approach allows for flexibility in choosing different sensors, tools, or features, or even different modeling paradigms at the various layers independently. The fact that the lower layer recognizes actions at an individual level may also allow for greater

flexibility in accommodating changing numbers of athletes at the higher layer (which often happens in training exercises), and for future optimization of models in the lower layer with respect to individual players – something which would be precluded by a monolithic, computationally expensive approach of modeling interaction directly on top of the sensor data.

Figure 2: Our layered recognition pipeline: from movement sensors to interaction modeling of team play. The output can be shown in any form to coaches or volleyball players.

In the *sensor layer*, data is collected from volleyball players wearing Inertial Measurement Units (IMUs) on both wrists. The IMUs gather 3D acceleration, 3D angular velocity, 3D magnetometer and air pressure at a sampling frequency of 40Hz, which are pre-processed using a built-in Kalman filter.

The *action recognition layer* classifies the volleyball actions performed by each single player on the basis of their IMU data. This is done in a 2 stage process described elsewhere [8, 9, 22, 23] and briefly summarized here. First, a binary classification is performed to separate actions from non-actions. Next, the actions are classified into one of 7 classes, namely Block, Serve, ForearmPass, Smash, OneHandPass, OverheadPass, and UnderHandServe. The classifiers are trained on a manually annotated ground truth of recorded sensor data, containing unbalanced data consisting of 990 action labels which were triple coded with high reliability [23]. Performance is reported as Unweighted Average Recall (UAR) because the used data sets were unbalanced. The first algorithm identifies actions from non-actions with a UAR of 86.87%; the second algorithm classifies specific actions with a UAR of 67.87%.

The *interaction modeling layer*, finally, focuses on modeling the sequential patterns of single player actions carried out in rallies by each team, modeled through a weighted transition graph. This layer is the focus of the current paper.

Figure 3 shows how the output of the interaction modeling layer is analysed. Data was collected from a friendly game of 23 minutes in which ambitious amateurs performed a total of 398 actions that were annotated by hand. In this game, players did not perform the UnderHandServe, which was left out of subsequent analyses, but did perform a TipBall, which was included. On the basis of the ‘true’ annotated actions, a second layer of ‘erroneous’ annotations was created by randomly modifying labels to get data with simulated recognition errors. The two annotations are used to derive interaction models (cf. section), yielding a ‘true’ and an ‘erroneous’ interaction graph. These are compared to show how the type of error simulation impacts their difference and to reflect on implications for using the error graph for analysing team performance.

Figure 3: Flowchart depicting our approach to analysing the output of the interaction modeling layer.

Figure 4: Actions performed in the five complexes extended with the names of the nodes in the model. Original figure from [10].

4 THE SSE MODEL OF VOLLEYBALL INTERACTIONS

We model group level interaction in volleyball as a weighted labelled graph that indicates how the athletes in the teams proceed through the various complexes of [10]. The model is like a Markov model, but with a few differences. As in a Markov model, *nodes* are labelled, in our case with the different stages the rally can be in. However, in a Markov model, *edges* represent probabilities to go from one node to another. In our model, edges are labelled with various volleyball actions and we keep track of absolute occurrences of the transitions. This model can function as an automaton keeping track of how a rally progresses. Figure 4 shows the idealized stages of a rally: a complex starts with an action of the opposing team (serve, attack, counter-attack, block, or freeball – a non-attacking ball that passes over the net); this is ideally followed by a sequence of 3 transitions (defence, setup, attack) culminating in the ball being played back over the net in a new attack, or by other combinations of transitions.

4.1 Structure of the model

Figure 5 shows some details of the weighted transition graph for complex 1. The names of the nodes follow from the complexes as shown in Figure 4, which shows that for each complex three

nodes exist: 1) Start of the complex based on the last action of the opponent, 2) Defensive state 3) SetUp state. In a volleyball match, K1 is the first complex after serve, and ideally, a team wants to defend with a ForearmPass (Neutral_Serve to K1_ServeReception) to send the ball to the setter. The setter can then play an OverHeadPass (K1_ServeReception to K1_SetUp) so that an attacker can Smash (K1_SetUp to Neutral_Attack). It is also possible that the rally ends after the smash and depending on its successfulness, either K1_RallyWon or K1_RallyLost is the final state.

However, this is an ideal situation and at any point of these steps, actions might not be performed as wanted. For example, if the defense with the ForearmPass goes wrong, it could be that the SetUp also has to be performed with a ForearmPass. This can even cause the attacker to perform an OverHeadPass, so that the opponent gets a Freeball (start of K5, Neutral_Freeball). Of the available actions, a Serve can only occur in K0, as it is the starting action. A Block causes the start of K4 (Neutral_Block) and the FP, OHP, TB and 1HP are so-called Freeball actions that cause the start of K5 (Neutral_Freeball) when they are played as attack.

Figures 5(a), 5(b) and 5(c) show parts of the rally model as described above. These figures also demonstrate how actions form transitions between states. The resulting model with all states and transitions consists of 28 different states and 301 transitions.

4.2 Traversing the Model to Count Occurrences of Certain Sequences

Above, the *structure* of possible actions in rallies is described. To keep track of who performs which actions, each edge is weighted with an occurrence counter for the team and counters for each individual player. To this end, the model is traversed following the actions that occur during actual rallies in match or training (as recognized automatically, or annotated by hand). Every time an action occurs while the model is in a certain state, the counters for that transition are increased and the model state is updated to the end point of that transition.

If an action is detected that is impossible in the current state (e.g., a Smash during K1 is wrongly classified as a Serve, which cannot be represented in the created model), the model transitions to an ErrorState and remains there until a new rally starts.

Once all traversing is done (e.g., the match is over), a “rally graph” is created (see Figure 6), showing what actions are performed most often. Such graphs can be used as a multi-purpose output (e.g. to coach, players, or others interested to further analyse the match).

5 IMPACT OF ERRONEOUS ACTION RECOGNITION ON INTERACTION MODELS

The model described above can be traversed using the actions performed by players. These actions can either be annotated by hand or recognized from sensor data by an action classifier. The first option leads to a “true” graph representation in which the rally graph of Figure 6 fully represents what happened in the rally. For the second option, it can be that some actions are recognized incorrectly, leading to a weighted graph that does not fully represent the occurrences of sequences that happened during the rally. This “erroneous rally graph” can differ from the true graph in three ways:

- the weight of transitions within and between complexes may change (i.e., traversal of a certain transition may be over-counted or under-counted to a great extent)
- a large number of traversals of transitions from a complex to the ErrorState may be counted (i.e., actions were detected that should not possibly happen in a certain state)
- a large number of traversals of transitions within the ErrorState may be counted (i.e., the traversal getting stuck within the ErrorState for some time)

We hypothesize that not every type of action recognition error has the same impact on the resulting rally graph. To investigate such a possible impact further, as a use case we investigate how various forms of *simulated erroneous action recognition* lead to different impact on interaction modeling on the basis of our data.

5.1 Method: Simulating Erroneous Recognition

We simulate erroneous action recognition on top of the hand annotated ground truth of volleyball actions for a collection of rallies, as also shown in Figure 3. First, we created the “true rally graph” for this ground truth. Next, we systematically modified some of the ground truth action labels according to an error model for action recognition and used the resulting erroneous data to create an additional “erroneous rally graph”. For example, if in the ground truth 100 ForearmPasses occur and the error model assumes that the algorithm classifies ForearmPasses correct 90% of the time, randomly, ten of the ForearmPasses are changed into another action before the data is used to traverse the model. By repeating this process 100 times, an average weight and standard deviation was calculated for each transition in the erroneous rally graph.

Different error models can be simulated by changing both the ratio of misclassifications, and changing the likelihood with which an action label is changed into each of the other labels. We express these models in the form of confusion matrices. The next subsection gives various confusion matrices with a recall of 95% for actions. Simulations are also done with a recall of 90% and 80% for a given subset of actions. The full set of confusion matrices are provided in the supplementary file to this paper.

5.2 Material: Multiple Grounded Error Models

We simulated various error models. First, we were interested in the influence of overall accuracy with random errors. Next, we focussed on the impact of errors only between Freeball actions, only between non-Freeball actions, or the combination of the two. The final simulation scenario was based on typical classification errors that we achieved in our own action classification experiments to be published elsewhere [24].

In this section, we will go into detail for several of these confusion matrices. The full set of confusion matrices and their results are given in the supplementary file to this paper.

Random classification errors simulations: The first simulation that we discuss here concerns one in which all actions have the same recall, with errors evenly divided over all other actions. This is illustrated with the confusion matrix in Table 1.

Confusions within Freeball actions: As described earlier, Freeball actions are those that are generally played *within* a team, but

Figure 5: Snippets of the full K0 and K1 graphs show the most salient parts. The snippets show unidirectional transitions to next states within a complex, or to the next complex.

Figure 6: A Rally Graph is the rally model updated with weights according to the actions taking place in a series of rallies. It can be seen that the most common path starts at 'StartRally' with Serve, followed by a ForearmPass, OverHeadPass and a Smash in K1. There are also transitions that go to the ErrorState.

can also fly over the net to the opponent (who then proceed in K5). Since these can be played at any time, it is likely that confusion within Freeball actions will not lead to the traversal getting significantly more stuck in the ErrorState. This type of error is shown in the confusion matrix in Table 2.

Confusions within Non-Freeball actions: Similar to the above, Table 3 illustrates erroneous classification that only confuses within Non-Freeball actions. This type of error is expected to lead to the traversal getting stuck more in the ErrorState than confusion with Freeball actions.

Combined Freeball and non-Freeball confusion: In Table 2 and 3 it is assumed that confusion only happens within that subset of actions (Freeball or non-Freeball), whereas all other actions are recognized perfectly. The two confusion matrices are combined into one in which the specific subsets are confused at the same time. This results in the confusion matrix in Table 4.

Simulating real classification errors: The final simulation is based on the typical patterns of error in our own automatic action recognition work [24]. Table 5 shows the confusion matrix to simulate those error patterns. This confusion matrix is a little adjusted

from the paper as the earlier work did not include TipBall as action, but instead UnderHandServe. The adjusted confusion matrix has UnderHandServe removed and TipBall added, assuming perfect recognition as no information on this action is available.

Compared to the previous confusion matrices, the values are given in absolute numbers to keep the recognition ratio between different actions; for the number of TipBall actions we took the average of the other actions. These absolute numbers follow from the total number of actions performed in the volleyball session that we used for our action classification experiments [24].

Table 1: Confusion matrix to simulate incorrect recognition of actions with an overall accuracy of 95%.

	B	FP	1HP	OHP	Serve	Smash	TB
B	95	0.83	0.83	0.83	0.83	0.83	0.83
FP	0.83	95	0.83	0.83	0.83	0.83	0.83
1HP	0.83	0.83	95	0.83	0.83	0.83	0.83
OHP	0.83	0.83	0.83	95	0.83	0.83	0.83
Serve	0.83	0.83	0.83	0.83	95	0.83	0.83
Smash	0.83	0.83	0.83	0.83	0.83	95	0.83
TB	0.83	0.83	0.83	0.83	0.83	0.83	95

Table 2: Confusion matrix to simulate incorrect recognition of Free-ball actions with a recall of 95%.

	B	FP	1HP	OHP	Serve	Smash	TB
B	1	0	0	0	0	0	0
FP	0	95	1.67	1.67	0	0	1.67
1HP	0	1.67	95	1.67	0	0	1.67
OHP	0	1.67	1.67	95	0	0	1.67
Serve	0	0	0	0	1	0	0
Smash	0	0	0	0	0	1	0
TB	0	1.67	1.67	1.67	0	0	95

Table 3: Confusion matrix to simulate incorrect recognition of non-Freeball actions with a recall of 95%.

	B	FP	1HP	OHP	Serve	Smash	TB
B	95	0	0	0	2.5	2.5	0
FP	0	1	0	0	0	0	0
1HP	0	0	1	0	0	0	0
OHP	0	0	0	1	0	0	0
Serve	2.5	0	0	0	95	2.5	0
Smash	2.5	0	0	0	2.5	95	0
TB	0	0	0	0	0	0	1

Table 4: Confusion matrix to simulate incorrect recognition of Free-ball actions and incorrect recognition of non-Freeball actions at the same time, both with a recall of 95%.

	B	FP	1HP	OHP	Serve	Smash	TB
B	95	0	0	0	2.5	2.5	0
FP	0	95	1.67	1.67	0	0	1.67
1HP	0	1.67	95	1.67	0	0	1.67
OHP	0	1.67	1.67	95	0	0	1.67
Serve	2.5	0	0	0	95	2.5	0
Smash	2.5	0	0	0	2.5	95	0
TB	0	1.67	1.67	1.67	0	0	95

Table 5: Confusion matrix with absolute values to simulate incorrect recognition of actions. Ratios between incorrect recognition of actions follow from our earlier work [24].

	B	FP	1HP	OHP	Serve	Smash	TB
B	9	2	2	10	1	0	0
FP	14	169	5	14	2	3	0
1HP	4	9	6	1	0	6	0
OHP	4	7	2	103	2	0	0
Serve	1	11	0	2	69	10	0
Smash	1	7	2	6	5	88	0
TB	0	0	0	0	0	0	96

5.3 Measure: Quantifying the Impact of Errors

Using the confusion matrices above, the different erroneous rally graphs were obtained, for recall percentages of 95%, 90% and 80%. For each of these erroneous rally graphs, the transition weights (and their standard deviation) were obtained and compared to the weights from the true rally graph. Z-tests were used as a means to quantify the change in transition weights between the erroneous graphs and the true graphs. Transition weights that were significantly different at an alpha level of 0.05 were scored and expressed as a percentage of the total number of transitions for that confusion matrix. With transition weights (average and standard deviation) being derived from a sample of $n = 100$ simulated rounds of erroneous classification, the Central Limit Theorem holds, meaning that the sampling distribution approximates the normal distribution, regardless of the population distribution being sampled. Results are reported separately for the three types of transitions described earlier at the start of this section (within complexes, from complex to ErrorState, and within ErrorState).

5.4 Results: Impact on Interaction Models

In Figure 7, the percentage of transitions that were different from the true rally graph at an alpha level of 0.05 is given per confusion matrix at the various levels of simulated UAR recognition rate. Results are shown, split out by type of transition.

In general, when the recall of actions improves (from 80% to 95%), the number of transitions that changed because of erroneous recognition drops. This pattern is seen for all types of confusion matrices. The highest degree of change of transition weights is observed when actions are randomly changed into any of the other actions ('all') and the lowest when only Freeball actions ('FBact') are confused with one another.

The results of two separate confusion matrices (FBact, Table 2, and nonFBact, Table 3) can be compared to the result of the combined confusion matrix (Table 4), which shows that the degree of change within transition weights for the combined confusion matrix is less than the sum of the two separate matrices.

Furthermore, when using the confusion matrix from our own classification experiments with an UAR of 67.87% [24], less transitions are changed through recognition errors than when using the confusion matrices with a recall of 80% and 90% for 'all' actions (with a UAR of 80% and 90%).

6 CONCLUSION AND DISCUSSION

The results of this paper show how action recognition errors might propagate to higher level interaction modeling, and show this for

Figure 7: Bar graph showing the percentage of transitions that were different from the true rally graph at an alpha level of 0.05 per confusion matrix. The left-most bar (tagged *) is the reference (i.e. true rally graph), showing the proportion of the three types of transition in the data. On the x-axis, the confusion matrices are set out; the numbers refer to the simulated recall of the actions and the names to the different confusion matrices

a realistic error pattern grounded in the real classification errors of our action recognition layer. In this final section we discuss the contributions and limitations of our work, look at implications for use of action recognition and interaction modeling in volleyball monitoring systems, and present an outlook to future work.

6.1 Contributions

The work presented in this paper offers several contributions.

Firstly, the Rally Graphs introduced in this paper offer **a method for modeling interaction in volleyball**. Rally graphs are weighted, labeled transition graphs that model the structure of rallies in volleyball matched and training sessions in terms of so-called volleyball complexes. By traversing this structure, counting actual actions taking place in a match or training, the rally graphs can furthermore show which sequences occur more, or less, for a team or for individual players. This could be turned into insight about strengths and weaknesses of a team and individuals, showing who is able to perform which actions best (e.g. in terms of defending/attacking).

Secondly, we presented **a method to investigate the impact of errors in the lower level action recognition on this interaction modeling**. We described our approach as a multi-level process using a hand annotated ground truth, real and theoretically grounded error models, and simulations of these error models in the action recognition.

Thirdly, the results presented in Figure 7 show **insights into how different kinds of error propagate differently** to the interaction modeling layer, and what kind of trade-offs can be made for improving further iterations of action recognition models for volleyball monitoring systems.

The results suggest that not all errors are equal: when all actions are equally confused, the three types of transitions seem to be changed to a greater degree than for the other error models, even the

one derived from the error patterns shown by our actual classifiers. Simply reporting overall accuracy (UAR) of an action recognition layer is thus not a very good measure to say something about the (potential) influence of errors in the *action recognition layer* on the model of the *interaction layer*.

More specifically, when confusions are only between Freeball actions, no transitions to the ErrorState are made. This is because they are interchangeable in terms of transitions between within and between complexes (cf. Figures 5(a), 5(b) and 5(c) for illustration). Non-Freeball actions are not as interchangeable, and are more “complex specific”. Confusions between Non-Freeball actions therefore have a greater impact on the interaction modeling layer.

Fourthly, our results also allow us to say something about the **suitability of the results in our own action recognition experiments [24] for the purpose of interaction modeling**. Errors in the classifiers are certainly not equally distributed across all combinations of labels: In Figure 7 the bar of the confusion matrix from the ‘Paper’ is lower than the bars with overall confusion of all actions while it has worse UAR than those other bars. Yet, it is equally clear that the classifiers underlying our action recognition layer should improve, given how many transitions change by applying an equivalent measure of errors to the ground truth data. Given that the impact of confusion between Freeball actions is far less severe than that of misclassifying Non-Freeball actions, improvement should focus on the recall for non-Freeball actions, even at cost of the recognition of the Freeball actions. This can be done using a loss function penalizing the incorrect recognition for non-Freeball actions more than for Freeball actions [5].

6.2 Limitations

There are several limitations to our approach that should be addressed in follow-up work.

First, in earlier work we see that actions can be recognized from IMU data in two steps. Actions can be separated from non-actions [8] after which the actions are recognized as specific actions [23]. In our study we only look into this second step, to see what consequences are of possible recognition errors in our own classifiers. However, it is also possible that errors are made in the first step, when separating actions from non-actions. This should be taken into account in future work: Even though the UAR of the first step is 86.87%, the precision for the action class is only 25.4% [8]. This means that a lot of automatically recognized actions are actually non-actions that may lead to more transitions to the ErrorState than what we found in this paper. To get a more complete view on the impact of recognition errors on interaction modeling, this first step should also be taken into account.

Second, the discussed impact of a recurring propagating ErrorState also shows a weakness of our implementation and perhaps even fundamentally in the approach. From related work in other domains we know that using statistics of the likeliness of sequences may help to re-classify the lower level assigned action if that makes the overall sequence more likely. In our use case, for example, a sequence starting with Serve is more likely, when 10 actions are following, then starting it with Smash. So although fundamentally we showed advantages of a multilevel approach, reconsidering the lower level classification via the results in a higher layer might be used to improve results overall, also in a team sports setting.

Third, to quantify the impact of errors at the action recognition layer on the interaction modeling layer, we used Z-tests to identify changes in transition weights at an alpha level of 0.05. As such, the Z-tests in our current approach rather served as a threshold for counting occurrences of change than a formal procedure for testing significant differences between transition weights of different models (i.e. erroneous graphs vs true graphs). In future work, we aim to go beyond this indication of change and uncover the specific mathematical relations that describe how different error models affect model performance at the level of complexes.

Finally, the data set that we used was based on adult, relatively good players (although not elite). At other levels (higher and lower) and age groups, results might be different, also because the *a priori* distribution of action types will certainly not be the same. This deserves further exploration as well, with other data sets.

6.3 Implications for Volleyball Exercise Technology

In the larger context of our research, we aim to use the automatic action recognition and interaction modeling for volleyball monitoring systems and for novel forms of digital physical volleyball training. Insights in complexes, and how a team typically progresses through sequences of actions in these, towards a score or a loss of ball, can be very helpful either for decision making in training, or as input to new forms of interactive training exercises. And again there, not every type of error has the same impact in either type of application. For example, the system can be made to provide feedback to the player. If this feedback is general statistics about which complexes are performed how often, it is most important that non-Freeball actions are recognized correctly as those determine which complexes are reached. If the feedback is about whether players perform e.g.

bump, set, spike correctly, all these actions should be recognized at the right moment. On the other hand, if we want to show a graph with average performance afterwards, it is of less importance that specific instances of actions are recognized correctly, as long as it shows the right trend (so mostly ‘symmetrical’ confusions are still acceptable, between actions and between complexes reached). Still, it remains important the non-Freeball actions are recognized correctly more often, because then less transitions to the ErrorState are made and more information is gathered.

6.4 Future Work

Besides addressing shortcomings and other suggestions made above, for our work on interaction modeling in volleyball we foresee several possibilities for interesting future work.

The model as described above follows from the concept of complexes in a volleyball rally [10]. However, in our other work with volleyball we have seen that *practice exercises* do not always consist of regular rallies. It might be interesting to develop *exercise model*, with a different structure, that reflect the typical patterns of the training exercise, different from full rallies. These could be built manually on the basis of the exercise design, or developed bottom up (see below) based on actual recordings of the exercise being carried out. Following up on that, one could completely automatically create the model structure, for rallies as well as for exercises. This would however probably require better levels of action recognition than we currently have available.

Furthermore, now that we are able to derive and present these rally graphs based on data, it will be good to start exploring how exactly these graphs provide insight for trainers and athletes in volleyball training practice. This requires understanding the decision making that goes into modifying training sessions, and might require insight into what is the best possible way to visualise the knowledge embedded in these graphs.

Finally, we can explore possible transfer to other domains to generalise beyond volleyball. In basketball and Futsal, for example, there are also structural patterns of attack, defense, and counter attack that depend on “typical sequences of individual player actions”. However, in contrast to volleyball, these sports do not have the strict three-actions-per-team-per-time rule, so the equivalent of complexes for these sports might look much more complex and would require a modified approach to build the “rally graphs”. This would be a major complexification of the work presented in this paper, but would also give the concepts presented here a vastly wider applicability.

ACKNOWLEDGMENTS

This work was carried out as part of the Smart Sports Exercises project funded by ZonMw Netherlands. We would like to thank the students and colleagues who helped annotate data and the athletes and coaches whose data we were allowed to collect.

REFERENCES

- [1] David Altimira, Florian “Floyd” Mueller, Jenny Clarke, Gun Lee, Mark Billingham, and Christoph Bartneck. 2016. Digitally Augmenting Sports: An Opportunity for Exploring and Understanding Novel Balancing Techniques. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI ’16). Association for Computing Machinery, New York, NY, USA, 1681–1691. <https://doi.org/10.1145/2858036.2858277>

- [2] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. 2017. Social Scene Understanding: End-to-End Multi-person Action Localization and Collective Activity Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3425–3434. <https://doi.org/10.1109/CVPR.2017.365>
- [3] Violeta Beniscelli, Gershon Tenenbaum, Robert Joel Schinke, and Miquel Torregrosa. 2014. Perceived distributed effort in team ball sports. *Journal of Sports Sciences* 32, 8 (2014), 710–721. <https://doi.org/10.1080/02640414.2013.853131>
- [4] Spencer Best. 2013. *Using Markov Chains to Analyze a Volleyball Rally*. Master's thesis. Carthage College.
- [5] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.
- [6] L. Ponce Cuspinera, Sakura Uetsuji, F. J. Ordonez Morales, and Daniel Roggen. 2016. Beach Volleyball Serve Type Recognition. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers (Heidelberg, Germany) (ISWC '16)*. Association for Computing Machinery, 44–45. <https://doi.org/10.1145/2971763.2971781>
- [7] Sotirios Drikos. 2019. Complex 1 in Male Volleyball as a Markov Chain. In *Proceedings of MathSport International 2019 Conference (Athens, Greece)*, Dimitris Karlis, Ioannis Ntzoufras, and Sotiris Drikos (Eds.). 80–85.
- [8] Fasih Haider, Fahim A Salim, Sena Busra, Yengce Tasdemir, Vahid Naghashi, Kubra Cengiz, Dees B W Postma, Robby Van Delden, and Dennis Reidsma. 2019. Evaluation of Dominant and Non-Dominant Hand Movements For Volleyball Action Modelling. In *ICMI 2019*. ACM, Suzhou.
- [9] Fasih Haider, Fahim A Salim, Dees BW Postma, Robby van Delden, Dennis Reidsma, Bert-Jan van Beijnum, and Saturnino Luz. 2020. A Super-Bagging Method for Volleyball Action Recognition Using Wearable Sensors. *Multimodal Technologies and Interaction* 4, 2 (2020).
- [10] Raúl Hileño and Bernat Buscà. 2012. Observational tool for analyzing attack coverage in volleyball. *Revista Internacional de Medicina y Ciencias de la Actividad Física y del Deporte* 12 (10 2012), 557–570.
- [11] Yu-Liang Hsu, Shih-Chin Yang, Hsing-Cheng Chang, and Hung-Che Lai. 2018. Human daily and sport activity recognition using a wearable inertial sensor network. *IEEE Access* 6 (2018), 31715–31728.
- [12] Marta Hurst, Manuel Loureiro, Beatriz Valongo, Lorenzo Laporta, T. Pantelis Nikolaidis, and José Afonso. 2016. Systemic Mapping of High-Level Women's Volleyball using Social Network Analysis: The Case of Serve (K0), Side-out (KI), Side-out Transition (KII) and Transition (KIII). *International Journal of Performance Analysis in Sport* 16, 2 (2016), 695–710. <https://doi.org/10.1080/24748668.2016.11868917>
- [13] Marta Hurst, Manuel Loureiro, Beatriz Valongo, Lorenzo Laporta, Pantelis Nikolaidis, and José Afonso. 2017. Systemic Mapping of High-Level Women's Volleyball using Social Network Analysis: The Case of Attack Coverage, Freeball, and Downball. *Montenegrin Journal of Sports Science and Medicine* 6 (03 2017), 57–64.
- [14] Mads Møller Jensen, Majken K. Rasmussen, Florian "Floyd" Mueller, and Kaj Grønbaek. 2015. Keepin' It Real: Challenges When Designing Sports-Training Games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (Seoul, Republic of Korea) (CHI '15)*. Association for Computing Machinery, 2003–2012. <https://doi.org/10.1145/2702123.2702243>
- [15] Raine Kajastila and Perttu Hämmäläinen. 2015. Motion Games in Real Sports Environments. *Interactions* 22, 2 (Feb. 2015), 44–47. <https://doi.org/10.1145/2731182>
- [16] Thomas Kautz, Benjamin H. Groh, Julius Hannink, Ulf Jensen, Holger Strubberg, and Bjoern M. Eskofier. 2017. Activity recognition in beach volleyball using a Deep Convolutional Neural Network. *Data Mining and Knowledge Discovery* 31, 6 (nov 2017), 1678–1705. <https://doi.org/10.1007/s10618-017-0495-0>
- [17] Lorenzo Laporta, José Afonso, and Isabel Mesquita. 2018. Interaction network analysis of the six game complexes in high-level volleyball through the use of Eigenvector Centrality. *PLOS ONE* 13, 9 (sep 2018), 1–14. <https://doi.org/10.1371/journal.pone.0203348>
- [18] Lorenzo Laporta, Pantelis Nikolaidis, Luke Thomas, and José Afonso. 2015. The Importance of Loosely Systematized Game Phases in Sports: The Case of Attack Coverage Systems in High-Level Women's Volleyball. *Montenegrin Journal of Sports Science and Medicine* 4 (03 2015), 19–24.
- [19] Manuel Loureiro, Marta Hurst, Beatriz Valongo, Pantelis Nikolaidis, Lorenzo Laporta, and José Afonso. 2017. A Comprehensive Mapping of High-Level Men's Volleyball Gameplay through Social Network Analysis: Analysing Serve, Side-Out, Side-Out Transition and Transition. *Montenegrin Journal of Sports Science and Medicine* 6, 2 (Sept. 2017), 35–41. <https://doi.org/10.26773/mjssm.2017.09.005>
- [20] Ana Paulo, Frank T.J.M. Zaal, Ludovic Seifert, Sofia Fonseca, and Duarte Araújo. 2018. Predicting volleyball serve-reception at group level. *Journal of Sports Sciences* 36, 22 (2018), 2621–2630. <https://doi.org/10.1080/02640414.2018.1473098>
- [21] Gaetano Raiola and Alfredo Di Tore. 2012. Non-verbal communication and volleyball: A new way to approach the phenomenon. *Mediterranean Journal of Social Sciences* 3, 2 (2012), 347–356. <https://doi.org/10.5901/mjss.2012.v3n2.347>
- [22] Fahim Salim, Fasih Haider, Sena Busra Yengce Tasdemir, Vahid Naghashi, Izem Tengiz, Kubra Cengiz, Dees Postma, Robby van Delden, Dennis Reidsma, Saturnino Luz, and Bert-Jan van Beijnum. 2019. A Searching and Automatic Video Tagging Tool for Events of Interest During Volleyball Training Sessions. In *2019 International Conference on Multimodal Interaction (Suzhou, China) (ICMI '19)*. ACM, New York, NY, USA, 501–503. <https://doi.org/10.1145/3340555.3358660>
- [23] Fahim Salim, Fasih Haider, Sena Busra Yengce Tasdemir, Vahid Naghashi, Izem Tengiz, Kubra Cengiz, Dees Postma, Robby van Delden, Dennis Reidsma, Saturnino Luz, and Bert-Jan van Beijnum. 2019. Volleyball Action Modelling for Behavior Analysis and Interactive Multi-modal Feedback. In *eNTERFACE'19 (Ankara, Turkey)*.
- [24] Fahim A Salim, Fasih Haider, Dees BW Postma, Robby van Delden, Dennis Reidsma, Saturnino Luz, and Bert-Jan van Beijnum. accepted for publication. Towards Automatic Modelling of Volleyball Players' Behavior for Analysis, Feedback and Hybrid Training. *Journal for the Measurement of Physical Behaviour* (accepted for publication).
- [25] Graham Thomas, Rikke Gade, Thomas B. Moeslund, Peter Carr, and Adrian Hilton. 2017. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding* 159 (2017), 3–18. <https://doi.org/10.1016/j.cviu.2017.04.011>
- [26] Javier Vales-Alonso, David Chaves-Diéguez, Pablo López-Matencio, Juan J. Alcaraz, Francisco J. Parrado-García, and F. Javier González-Castaño. 2015. SAETA: A Smart Coaching Assistant for Professional Volleyball Training. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 8 (2015), 1138–1150. <https://doi.org/10.1109/TSMC.2015.2391258>
- [27] Robby van Delden, Alejandro Moreno, Ronald Poppe, Dennis Reidsma, and Dirk Heylen. 2014. Steering Gameplay Behavior in the Interactive Tag Playground. *Ambient Intelligence* 1 (2014), 145–157. <https://doi.org/10.1007/978-3-319-14112-1>
- [28] Jan van Haaren, Horesh Ben Shitrit, Jesse Davis, and Pascal Fua. 2016. Analyzing Volleyball Match Data from the 2014 World Championships Using Machine Learning Techniques. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA) (KDD '16)*. Association for Computing Machinery, 627–634. <https://doi.org/10.1145/2939672.2939725>
- [29] Yufan Wang, Yuliang Zhao, Rosa H. M. Chan, and Wen J. Li. 2018. Volleyball Skill Assessment Using a Single Wearable Micro Inertial Measurement Unit at Wrist. *IEEE Access* 6 (2018), 13758–13765. <https://doi.org/10.1109/access.2018.2792220>