Heimdallr: A Dataset for Sport Analysis

Michael Riegler[1], Duc-Tien Dang-Nguyen[2], Bård Winther[1]
Carsten Griwodz[1], Konstantin Pogorelov[1], Pål Halvorsen[1]
[1]Simula Research Laboratory & University of Oslo
[2]University of Trento
[1]{michael, griff, baardew, konstantin, paalh}@simula.no
[2]dangnguyen@disi.unitn.it

ABSTRACT

In this paper, we present Heimdallr, a dataset that aims to serve two different purposes. The first purpose is action recognition and pose estimation, which requires a dataset of annotated sequences of athlete skeletons. We employed a crowdsourcing platform where people around the world were asked to annotate frames and obtained more than 3000 fully annotated frames for 42 different sequences with a variety of poses and actions. The second purpose is an improved understanding of crowdworkers, and for this purpose, we collected over 10000 written feedbacks from 592 crowdworkers. This is valuable information for crowdsourcing researchers who explore algorithms for worker quality assessment. In addition to the complete dataset, we also provide the code for the application that has been used to collect the data as an open source software.

CCS Concepts

•Information systems → Information retrieval; Multimedia and multimodal retrieval;

Keywords

Interactive; Multimedia; Soccer; Crowdsourcing; Data Set

1. INTRODUCTION

There has been an increased interest in analyzing sports using video and sensors for tracking athletes[1] and annotating a match with *key events*, e.g., every time an athlete scores a goal in a match. To better support and maintain large quantities of data and different recording systems, sport analytic systems are created to automatize and simplify the process of event logging and data management. An automatically created summary of players' individual actions at any time during a match or training session, for example, can support both the trainer and the medical team in assessing the player data and helping them to improve their skills and fitness.

Most current vision-based sport analytic systems, e.g., Bagadus [3], provide large amounts of data that can be used for many purposes, and both event detection and summarization have attracted a lot of research. In practice, however, sports leagues are still relying on manual event logging to assess individual athlete's performance using standard, not obvious events such as *a defender loosing control over his opponent*. Semantic information about actions and poses improves the situational awareness of an analytic system, which can then be used for an improved automatic-detection of events. Actions in soccer can be annotated with a well-known dictionary (e.g., "run" or "kick"), while a player's poses can be modeled as a sequence of skeletons with joint positions. To improve algorithms for both action recognition and pose detection, we require a ground truth data set that includes a database of sequences that are annotated with both action and pose. The traditional approach for annotating this type of data is hiring experts, people who annotate every skeleton without any errors. However, this is costly. Doing the job ourselves, on the other hand, would require weeks or even months even for a single match. We have therefore built an online training tool and invited crowdworkers from around the world to use it for annotating frames. The difficulty of this approach is, of course, to obtain accurate data from non-experts. Our approach for solving this problem comprised quality control and repeated annotation of the same frame by several crowdworkers, which were then merged.

Here, we provide our dataset, named Heimdallr, which contains all data that we collected in this way. While we collected this data for pose detection, it provides also insights into crowdworker behaviour, and crowdsourcing researchers may be interested in the data set to investigate quality control and worker discard algorithms.

Similar datasets have been published before [4, 5, 7]. The datasets have similar or larger size in terms of annotated frames. Nevertheless, these datasets differ in some points:

1. Our dataset is meant to be useful for researchers looking into pose estimation, but also useful for researchers that are interested in investigating crowdsourcing itself.
2. We provide not only close-up shots of players, but also the external calibration of the camera with respect to the field, as well as x and y positions of the players. This provides a player-centered collection of views (arbitrary bounding boxes) from our original panorama view and saves a lot of manual work.

[1]http://www.sloansportsconference.com/?p=5503

MMSys'16, May 10 - 13, 2016, Klagenfurt, Austria

ISBN 978-1-4503-4297-1/16/05…$15.00

DOI: http://dx.doi.org/10.1145/2910017.2910621

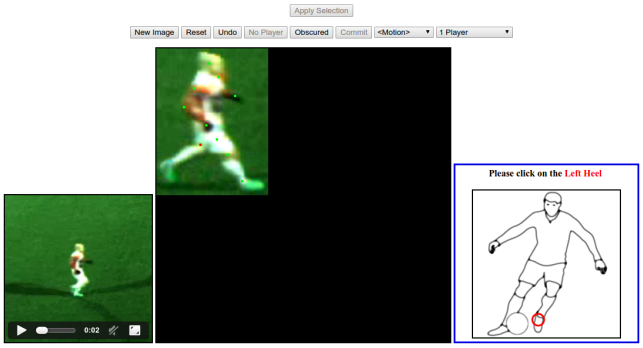

Figure 1: The `Online Training Tool` for collecting feedback and annotated skeleton for Heimdallr. From left to right is the action preview window, the annotation window and the annotation request window, respectively.

3. All annotated shots that we provide are taken by one static camera system. This removes the uncertainty about camera intrinsic parameters, focal length and angle towards the ground, and is likely to help in the verification of valid poses.

So, although the annotated dataset does not have infinite size (obviously), we provide images that are acquired under the same conditions and publish these conditions along with the dataset.

The rest of the paper is organized as follows: Section 2 presents the dataset collection. Then, we describe the details of Heimdallr in Section 3. In Section 4, we show an application of action classification to give an idea what can be done with Heimdallr. Finally, we draw conclusions in Section 5.

2. DATA COLLECTION

In this section, we describe the collection process of Heimdallr, potentially providing researchers with ideas about the possibilities to exploit this dataset. We also present the quality controls that we used during the collection of this data, which is important for researchers from the crowdsourcing community.

2.1 Human Intelligence Task

In order to collect the feedback and annotated data, we developed an `online training tool` that asks each worker to perform a Human Intelligence Task (HIT). Figure 1 shows a screenshot of the tool. The worker is provided with an image of a player, cropped from a panorama image to an initial region-of-interest (ROI) with the player at the center, and a video centered on the player, which contains the image and can be watched repeatedly. The worker can define a bounding box to select a small ROI and upscale the image, then go through the task of selecting points in the image for each of 13 joints. A very detailed description of the task, how many experts, etc., and its complexity, how outliers have been treated, etc., can be found in [8]. In this dataset paper, we will only give a brief overview about the most important aspects.

Source Frames All the video footage of soccer players used for annotation comes from an earlier archived game from the Bagadus system [6]. We provided the online training tool on the web. Workers were asked to annotate soccer

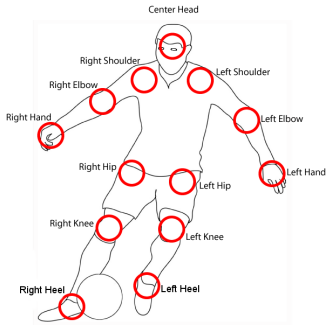

Figure 2: The joint locations used in the crowdsourcing task.

players using this tool, with each worker assigned one random frame at a time. Randomization was done for two reasons: First, to provide variation for workers, as annotating frames from the same video sequence can be too repetitive. Second, to achieve a decent overall average quality for every shot by evenly distributing results from workers performing high and low quality work over all frames. This process is done for every frame of every player for all video sequences, resulting in tens of thousands of annotated skeletons.

Bounding Box When a worker is presented with an image to annotate, this is already cropped from a larger panorama to a square ROI. Still, the worker has the option of selecting a smaller square bounding box around a soccer player. If a worker does that, the selection is scaled up, which may help the worker to annotate the joint locations with more confidence. Since these bounding boxes were not relevant for acquiring joint positions, but merely an aid to the workers, we did not keep the box size. Furthermore, the workers must also report the number of players present in the bounding box. This was done to filter out any sequences which unintentionally showed more than one player.

Skeleton Plotting After the optional bounding box selection, the worker is asked to annotate a skeleton for the soccer player, consisting of 13 joint locations: head, shoulders, elbows, hands, hips, knees and feets. Typically, researchers use more than 13 joint locations in their research work [1, 2, 10]. For two reasons, this was not meaningful in our case: First, our sequences are very long shots, and our own research challenge was actually aimed at detecting poses from a very small number of pixels. Second, crowdworkers must be able to understand the task at hand quickly and easily in spite of this small number of pixels. Annotating the neck, for example, would have been infeasible for this data.

As Figure 1 shows, we provided crowdworkers also with an easy-to-understand visual clue about the joint that they were to annotate next, hopefully reducing the effort in task understanding even further. Figure 2 shows all joint locations that we requested in our HIT. For the point in the image that a worker actually selected for a joint position, we use the term *click-point*.

Motion Labels Before starting with the annotation of skeleton points, the worker must determine the type of action that is shown in the image. This is done by selection from a catalogue. Determining the action from a single frame is impractical, and one of the reasons for providing a short video sequence including frames before and after one that is evaluated. The workers' web browser downloads

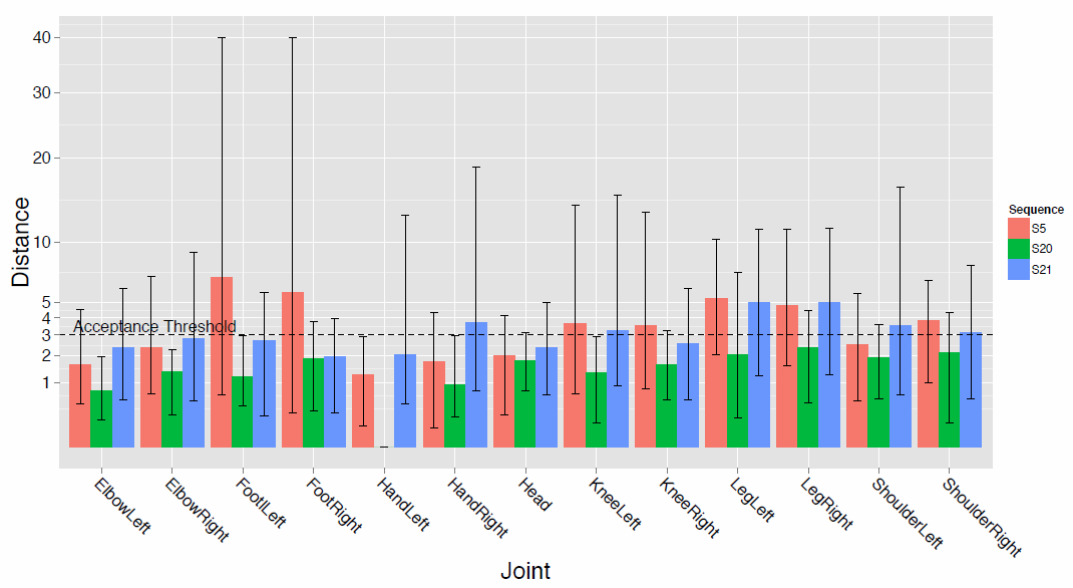

Figure 3: Mean difference between the merged crowdsourced joint placements and an expert, measured as pixels, with maximum and minimum bars. Everything below the acceptance threshold is considered acceptable, but is only valid for these sequences.

all of these frames from the server (unless they are already cached) and a video clip is created in JavaScript.

2.2 Quality Assessment

There are two primary aspects and one secondary aspect to look at when assessing the results of a crowdsourcing campaign for scientific work [9, 11]. The primary aspects consists of accuracy and efficiency, and the secondary aspect on how the crowdworkers' perception of the tasks is. The combination of these factors determines whether using a crowdsourcing platform is better or worse than to hire experts. The first and most important aspect is what we define as *accuracy*, which compares the precision and similarity between the workers and an expert. Secondly, given the accuracy provided by the workers, we evaluate the crowdsourcing platform as a viable alternative to experts, i.e., if crowdsourcing is better in terms of *Cost and Time*. Indirectly affecting accuracy and efficiency results is how the crowdworkers think about the tasks, as an engaging task is more likely to be done better.

2.2.1 Accuracy

Testing the accuracy of the workers is best done by comparing the click-points against an expert's. Firstly, three sequences was chosen: (i) Sequence $S5$ with action "run", (ii) Sequence $S20$ with action "side-jump", and (iii) Sequence $S21$ with action "kick". More details of the dataset organization will be introduced in Section 3. Secondly, the filtered and merged set of the workers click-points are used instead of that from individual workers. This should ensure the best annotations the workers can collectively provide and is also more correct, because the joint positions obtained here are the ones actually used in the annotated database for skeleton re-projection.

Figure 3 shows the mean difference between the expert's and the crowdsourcers' joint placements for the three sequences $S5$, $S20$ and $S21$. The y-axis displays the mean Euclidean distance, with maximum and minimum bars. The mean is computed for the non-obscured click-points in a sequence for a particular limb (which can be seen with *HandLeft* in sequence $S20$, which is zero because all click-points for this limb are obscured). Additionally, an acceptance threshold is also present to define what would be an acceptable accuracy, with anything above this line being unacceptable. The threshold is set to three pixels, a value determined by inspecting the video sequences' individual images and measure the area under which a click-point can be considered correct.

Overall, the mean accuracy of the workers is rather good and a lot better than anticipated. A small variance between the expert and the workers is expected, and everything below three pixels is considered correct. The results are also highly dependent on the level difficulty of annotating a player, with "side-jump" ($S20$) being the simplest as the player is always facing the camera with all limbs clearly visible and indistinguishable. When limbs are harder to tell apart, the maximum distance increases. The most notable example is $S5$'s feet, which suffers from a mix-up of which one is left and which one is right, causing a large difference between the workers and expert. Actually, most of the maximum error distances measured are from either disagreement between the workers themselves or hard to determine joint locations. For example, in Figure 4, the workers disagree on the exact location of the right leg, making the merged joint location more of a random guess rather than an actual estimate. On the other hand, the minimum shows that it is possible to have great accuracy, if not even better than what the expert plotted, as shown in Figure 5.

To summarize, the accuracy of the crowdworkers are rather decent compared to an expert. Most notably is the left-right annotation ambiguity that makes the distances from the experts far too large and would greatly increase the overall accuracy for all joints and sequences if it was not the case. With the total mean deviance from expert being **2.66** pixels for the three sequences makes it barely adequate enough to attempt using it in pose estimation.

Figure 4: Workers attempt at plotting an ambiguous right leg (more precisely the hip) with the individual click-points marked with orange points.

Figure 5: Skeleton obtained after Majority Vote Filtering, with blue points marking joints and red lines marking connecting limbs.

Table 1: Comparison between a crowdworker and an expert. *Images* are for the 1898 frames in the database and an *image* for what can be done during an hour. Day and completion estimates are based on eighth hour work days, with the results being approximate.

	Worker	Expert
USD/hr	2.20	$25 - 128$
USD/image	0.13	$0.50 - 2.56$
images/hr	$1 - 48$	50
images/day	$1 - 384$	400
Completion Estimate hrs	$1898 - 40$	38
Completion Estimate days	$1898 - 6$	5

2.2.2 *Efficiency*

To determine if a crowdsourcing platform is a viable alternative to experts for annotating skeletons depends not only on accuracy, but also efficiency. We define efficiency as being both faster and less expensive compared to an expert.

Time For reference, an expert can annotate 30 to 60 images per hour (depending on difficulty) with an average of about 50 images per hour. In addition, an expert is expected to get paid 25 USD to 128 USD per hour. This section evaluates the crowdworkers performance in both of these categories.

In our crowdsourcing campaign, a worker is paid 1.10 USD per task, with each task consisting of 24 images and an expected completion within 30 or 40 minutes. A summary is given in Table 1 which shows how a single worker or expert is estimated to perform. The table also includes an estimation of how much time a single worker or expert would need to completely annotate all the frames in the campaign, called completion estimate.

An expert can manage to annotate 400 images in the course of an 8 hour work day. While this number is significantly larger than a single worker average of about 192, it does not represent the actual images per day correctly. One of the key components in using crowdworkers is that they are a crowd of people ready to solve problems. Figure 6 shows the number of images annotated per day since the start of the campaign, with the crowdworkers greatly outperforming the expert in terms of annotated images, but

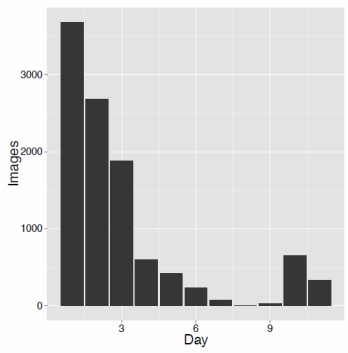

Figure 6: Images annotated by crowdworkers per day since start of the campaign.

not by that much. Because every image is annotated several times results in the actual annotated images per day are only a fraction of the actual count. In our case, it would be only one-fifth of the actual count, as each image is tasked an average of five times. Even then, though, they still exceeds the expert performance. Righteously, more experts can be hired, but they are far more difficult to hire, making the crowdsourcing platform both faster and easier to get the images annotated compared to experts.

Cost Despite the large number of annotated images per day and low time consumption for the crowdworkers, they are actually a rather cheap workforce. With a total of 1937 solved tasks and 1267 accepted tasks in the campaign, and a total campaign cost of 1393.70 USD, means that it is costing only slightly more than the cheapest expert ($25 \text{ USD} * (1989/50) = 950$ USD). This makes crowdsourcing a good alternative, especially considering the most expensive expert coming closer to 4900 USD in salary. And, if not for the cost, at least for the completion time.

2.2.3 *Worker Feedback*

Feedback from the crowdworkers are indirectly tied to the resulting accuracy and efficiency for the annotated skeletons, as worker who enjoy or like the task are more likely to form proper work. Based on the workers who provided feedback, there is apparently a great interest in our campaign: The workers found the task to be original, interesting and even calling it “a game". Moreover, the workers understood the concept of using low resolution and noisy images, but they would still prefer higher resolution to make annotation easier and less ambiguous. In short, the crowdworkers are more than happy to annotate skeletons, making it possible to continue use of a crowdsourcing platform for this type of tasks.

3. DATASET DETAILS

Heimdallr contains **42** video sequences, over **3000** fully annotated frames and over **10.000** crowdsourced feedbacks. Among these 42 sequences, **27** of them were also annotated by experts (see Table 2 for the list of these sequences). The complete dataset is released as a *mysql* database, the sequences as images and several scripts that help to sort, clean up and parse data. The database consists of four main tables that contain the data and some additional views. The SQL queries for the views are released as part of the dataset. The main tables are *trainingdata*, *crowdworker*, *feedback* and *goldLog* (see Figure 7).

Table 2: The list of sequences annotated by both experts and crowdsourced workers.

Sequence	Motion	Frames	Sequence	Motion	Frames	Sequence	Motion	Frames
S0	run	36	S9	run	115	S18	run	44
S1	run	154	S10	walk	66	S19	kick	18
S2	sprint	57	S11	run	48	S20	side-jump	32
S3	walk-backwards	60	S12	walk	163	S21	kick	25
S4	walk-backwards	88	S13	side-jump	47	S22	run	54
S5	run	56	S14	run	63	S23	kick	30
S6	sprint	49	S15	run	90	S24	run-backwards	51
S7	walk	168	S16	walk	131	S25	run-backwards	46
S8	sprint	52	S17	side-jump	29	S26	walk	126

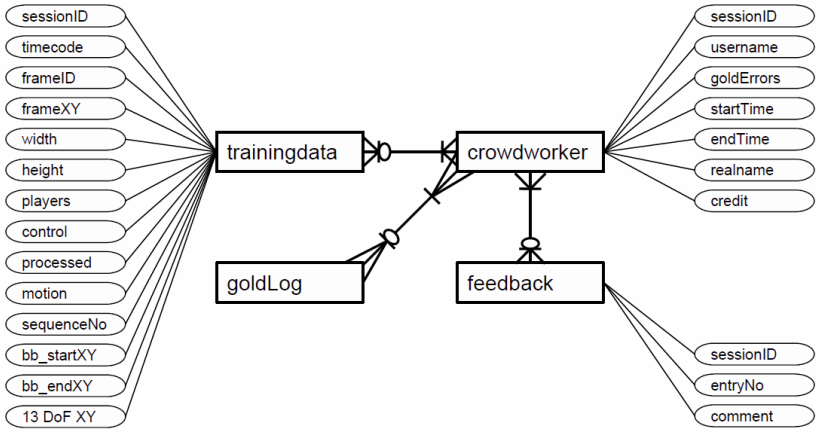

Figure 7: ER model of tables used in Heimdallr.

Table **trainingdata** contains the workers feedback for each image for all sequences. It consists of *sessionID*, used by the system to group different tables together, also indicate the session for training; *timecode* which is the timecode the frame is at in the video; *frame* name (generated from DataGenerator); *frameX* position of upper left corner of this frame section (which is part of a full raster image), i.e., where the image starts given the timecode and full-raster frame; *frameY* same as frameX, but for Y (coordinates starts in upper left corner). All coordinates are stored as relative coordinates. The dataset contains scripts for translating these in coordinates that can be used by everyone; *width* and *height* of the frame; *players* how many players are present in the image; *control* used to indicate a gold sample image; *processed indicate* if and how many times this image has been trained; *motion* which contains the action class; and *sequenceNo* that provides the sequence number (generated from DataGenerator). Further, the table contains bounding box pointers where 1 is the start (upper left corner), 2 is the end (bottom right corner). Finally, it contains the position data for all joints. The exact positions are stored as **<location>_<(l)eft|(r)ight><x|y>**. Possible locations are *h* for head; *s* for shoulder; *e* for elbow; *n* for hand; *l* for leg; *k* for knee; and *f* for foot.

The **crowdworker** table contains every important information about the crowdsourcing workers: *username* contains the name of the worker for the crowdsourcing platform. *goldErrors* contains the numbers of gold errors by this worker. The *startTime* and *endTime* contains the time that the worker needed to fulfill the task. *Realname* contains the real name of the worker, and *credit* contains information about if the worker wanted to be credited for her contribution or not.

The **feedback** table contains all feedbacks that we got from the workers. This might be interesting for researchers that want to use the dataset for crowdsourcing related research.

Finally, the **goldLog** table contains all the data collected by a trusted expert, and has the same information as the **trainingdata** table. This can be used as gold standard to compare the output of the crowdsourcing workers (as discussed in the quality assessment) or for control algorithms. We used it to decide if we want to use submitted data or not for further experiments like for example pose estimation.

The complete dataset containing all data and code required to use it plus a detailed documentation can be downloaded at goo.gl/lkHouo. Furthermore, we also provide the complete code for the crowdsourcing online training tool at bitbucket.org/mpg_code/bagadus-humanactionretrieval as an open source software.

4. APPLICATION OF THE DATASET

In this section, we present a conducted experiment exploiting Heimdallr: Action classification. The action classification algorithm is designed as a pipeline which is shown in Figure 8. The pipeline requires that the query sequence is normalized and has its optical flow vectors computed. Then, the annotated in Heimdallr is used to reproject skeletons into query sequences. The rest of the pipeline starts with taking the input flow vectors and transforming them into a matrix format. It then uses the transformed matrices to obtain frame-to-frame similarity of all frame combinations

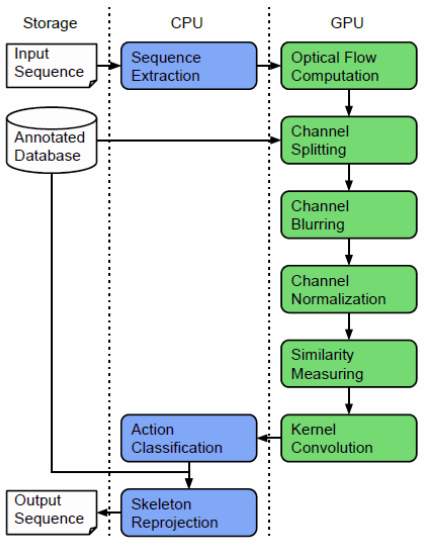

Figure 8: The complete pipeline for action classification and skeleton reprojection, with data in white, CPU stages in blue and GPU stages in green. The prerequisites are also included in here, which consists of the `Sequence Extraction` and `Optical Flow Computation` stages.

Table 3: Summarized classification accuracy for different kernel sizes and sigma values.

$N \backslash \sigma$	0.0	0.2	0.5
9	76%	70%	70%
13	76%	72%	72%
15	74%	72%	70%
21	78%	72%	70%

and a kernel is then applied to get the temporal information encoded into the similarity. Based on a Nearest Neighbor algorithm, the highest scoring sequence is found and is used to obtain the action label and pose for the query sequence.

To determine the accuracy of the classifier presented in this section, we exploited all the 42 sequences in Heimdallr. We ran the classifier with a combination of kernel sizes sigma values. Table 3 displays an approximately accuracy measure, calculated from *true_positives/total_elements*, obtained from the different kernel configurations. With this approach, **78%** of all sequences were correctly classified and up to pixel-perfect poses were estimated (i.e., reprojected). This result can be considered as a baseline for other approaches that will exploit Heimdallr.

5. CONCLUSION

We presented Heimdallr, a dataset that can be useful for two different fields of research: crowdsourcing and computer vision. The dataset allows to address tasks such as action classification, worker discarding, worker quality estimation and pose estimation, etc. We presented how to collect annotations of soccer players, by building an online training tool that can register and store user inputs. A quality assessment was carefully performed, showing that with the total mean deviance from expert being 2.66 pixels makes Heimdallr barely adequate enough to be used in pose estimation or action classification. With tens of thousands of feedback, crowdsourcing researchers can exploit Heimdallr to develop algorithms for assessing worker quality. The application that was used to collect the data is provided as an open source software.

6. ACKNOWLEDGEMENTS

This work is funded by the FRINATEK project "EONS" (#231687).

References

[1] T.-J. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. In *IEEE Computer Vision and Pattern Recognition*, pages 2239–2245, 1999.

[2] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H. P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Computer Vision and Pattern Recognition*, pages 1746–1753, 2009.

[3] P. Halvorsen, S. Sægrov, A. Mortensen, D. K. C. Kristensen, A. Eichhorn, M. Stenhaug, S. Dahl, H. K. Stensland, V. R. Gaddam, C. Griwodz, and D. Johansen. Bagadus: An integrated system for arena sports analytics: A soccer case study. In *ACM Multimedia Systems*, pages 48–59, 2013.

[4] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler. Learning human pose estimation features with convolutional networks. *In Proc. arXiv*, 2013.

[5] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *IEEE Computer Vision and Pattern Recognition*, pages 1465–1472. IEEE, 2011.

[6] S. A. Pettersen, D. Johansen, H. Johansen, V. Berg-Johansen, V. R. Gaddam, A. Mortensen, R. Langseth, C. Griwodz, H. K. Stensland, and P. Halvorsen. Soccer video and player position dataset. In *ACM Multimedia Systems*, pages 18–23, 2014.

[7] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *IEEE Computer Vision and Pattern Recognition*, pages 3674–3681, 2013.

[8] B. Winther, M. Riegler, L. Calvet, C. Griwodz, and P. Halvorsen. Why design matters: Crowdsourcing of complex tasks. In *ACM International Workshop on Crowdsourcing for Multimedia*, pages 27–32, 2015.

[9] O. F. Zaidan and C. Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1220–1229, 2011.

[10] Z. Zhang, H.-S. Seah, C. K. Quah, and J. Sun. Gpu-accelerated real-time tracking of full-body motion with multi-layer search. *IEEE Transactions on Multimedia*, 15(1):106–119, 2013.

[11] Y. Zhao and Q. Zhu. Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, 16(3):417–434, July 2014.