Automatic analysis of activities in sports arenas using thermal cameras

Gade, Rikke; Jørgensen, Anders; Jensen, Martin Møller ; Alldieck, Thiemo; Abou-Zleikha, Mohamed; Christensen, Mads Græsbøll; Moeslund, Thomas B.; Poulsen, Mathias Krogh; Larsen, Ryan Godsk; Franch, Jesper

Published in:
2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)

DOI (link to publication from Publisher):
10.1109/SITIS.2016.94

Publication date:
2017

Document Version
Accepted author manuscript, peer reviewed version

Automatic analysis of activities in sports arenas using thermal cameras

Rikke Gade*, Anders Jørgensen, Martin Møller Jensen,
Thiemo Alldieck, Mohamed Abou-Zleikha,
Mads Græsbøll Christensen and Thomas B. Moeslund
Department of Architecture, Design and Media Technology
Aalborg University
Aalborg, Denmark
**rg@create.aau.dk*

Mathias Krogh Poulsen, Ryan Godsk Larsen
and Jesper Franch
Department of Health Science and Technology
Aalborg University
Aalborg, Denmark

***Abstract*—The demand for automatically gathered data is a societal trend quickly extending to all aspects of human life. Knowledge on the utilization of public facilities is of interest for optimising use and cutting expenses for the owners. Manual observations are both cumbersome and expensive, and they have a risk of incorrect results due to subjective opinions or lack of interest in the given task. In this paper we present the main results of a 5-year long research project revolving around the real-world application of automatic analysis of activities in sports arenas. Three topics are explored: Counting people, recognising activities, and estimating energy expenditure. The project is based on thermal image data, to preserve privacy while capturing video in public sports arenas. This paper aim to provide an overview of our published methods and results within these three topics and add a discussion of the results and perspectives of this research.**

***Keywords*-Thermal imaging, Sport, Human behaviour, Counting, Activity recognition, Energy expenditure.**

I. Introduction

Analysis of humans is one of the biggest areas within computer vision, and is still very challenging as human behaviour is highly complex. One of the interesting human activities is sports. Participation in sport is encouraged and supported by most governments, due to very positive health effects. Many sports activities require available facilities, either indoor and outdoor. Especially indoor facilities are expensive to build and maintain, and constitute a significant expense for governments and municipalities in their attempt to support sports activities for their citizens. However, the actual utilization and use of these facilities is often unknown. This is the starting point for this project; How can we automatically analyse the activities in an indoor sports arena?

For the application in public indoor sports arenas, with all types of users ranging from children, to professionals and seniors, the methods must be both non-intrusive and privacy preserving. For that reason, we are capturing data only with thermal cameras. Thermal sensors capture infrared radiation in the long-wavelength infrared spectrum ($8-14\mu m$), which is emitted by all objects with a temperature above absolute zero [1]. A thermal image represents the temperature of the scene, often using white colour for the hottest pixels, and black for the coldest pixels. An example from one of the datasets captured during this project is presented in figure 1. To cover the entire court area, three cameras with a resolution of 640×480 pixel is applied simultaneously. The three images are stitched horizontally to obtain the image shown in figure 1.

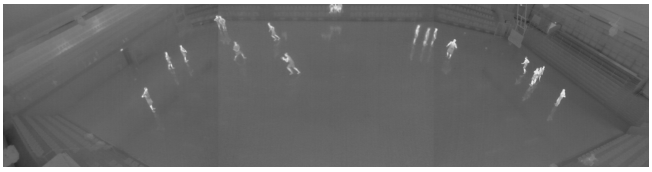


Figure 1: Example of thermal input image from a sports arena.

The purpose of this paper is to provide an overview of an ongoing research project, revolving around one real-world application. We here summarize a number of published methods developed for analysing the behaviour in sports arenas. The research is divided into three topics: Counting people (section III), Recognising activities (section IV), and Estimating energy expenditure (section V). Lastly, in section VI we compare methods and discuss the perspectives and future work in this project.

II. Related Work

With the increasing demand for data and analyses of sports games, e.g., for visualizations during broadcasts and for post-game analysis of statistics, computer vision methods has become one of the popular tools applied [2], [3]. Automatic detection and tracking of sports players is a research topic important as a basis for most areas of sports analysis. Most systems are based on visual cameras. In [4] a tracking system is proposed specifically for indoor football players, while [5] proposes a tracking system for outdoor football using multiple cameras. The tracking system proposed in [6] focuses on more general sports video and it is tested on both football, basketball and hockey. Recent methods

developed for team sports suggest to include context information like Game Context Features [7] and contextual trajectory information [8], or improve tracking by modelling latent behaviour from team-level context dynamics [9]. The research presented in this paper is focused on the activities on a global level, rather than individual behaviour and performance.

The large research area regarding automatic identification of human subjects and their behaviour include both visual and thermal cameras. There exist a number of surveys and books on the subject, including [10], [11], [12], [13]. Thermal imaging has recently become popular for surveillance applications and robust detection of pedestrians, as it is independent of lighting conditions [1], [14]. In this paper we will show how thermal imaging can be used as a privacy preserving data collection method, as the basis for further behaviour analysis.

III. Counting people

The first step for analysing the activities in sports arenas is to detect the individuals and estimate the number of participants.

A. Detection

The thermal modality simplifies the task of detecting humans. The images can be segmented based on temperature, as humans have a relatively constant temperature, which is normally different from the environment. Specifically, in temperature controlled indoor arenas, humans are hotter than the background. However, a strict temperature threshold can not be applied for our recordings, as the camera automatically adjusts the gain, meaning that the correspondence between temperature value and pixel value change over time. Instead two simple segmentation approaches have been tested; automatic thresholding and background subtraction.

1) Automatic thresholding: With input being thermal images, we operate with 8-bit greyscale images. Assuming that, within the region of interest, humans are hotter than the background, it is possible to find a threshold which separates humans from the background. However, the threshold value must adapt, as the gain in the camera is automatically adjusted. For this purpose we use an automatic threshold method based on Maximum Entropy [15]. This method maximises the sum of the entropy above and below the threshold value, by iterating through every possible value. A threshold value is calculated for each image individually and applied to obtain a binary image. By evaluating the entropy value, it can be determined if the arena is empty; if the entropy is below a specified threshold the resulting image is set to black.

This method is efficient if no non-human hot objects are present within the region of interest. In other cases background subtraction may be applied.

2) Background subtraction: With static hot objects in the scene, a background image can be captured and maintained, and a binary image is obtained by subtracting the background image from each new input image. The system must be initialised with a background image, but updating the image is important to adapt to slowly changing temperatures, or adjusted gain of the image. We update the image every minute by combining a new image and the existing background, but only letting pixels classified as background contribute to the new background.

3) Post-processing methods: After binarisation each human should ideally be represented by one white blob, and everything else represented by black. However, three main challenges exist at this point:

- Occlusions, which can make several people part of the same white blob.
- Reflections, which may create ghost objects.
- Over-segmentation, which make one person consist of several smaller blobs.

These challenges are illustrated with examples in figure 2.

To handle occlusions, we define two types of groupings; tall blobs, with people standing behind each other, seen from the camera's perspective, and wide blobs, with people standing next to each other, seen from the camera's perspective. We implement two simple but effective routines aiming at splitting blobs into single persons. These routines are summarized in the following sections, details can be found in [16].

Split Tall Blobs

People standing behind each other, seen from the camera, might be detected as one blob containing more than one person. In order to split these blobs into single detections, the first step is to detect when the blob is too tall to contain only one person. If the blob has a pixel height that corresponds to more than a maximum height at the given position, found by an initial calibration, the algorithm should try to split the blob horizontally. The point to split from is found by analysing the convex hull and finding the convexity defects of the blob. Of all the defect points, the point with the largest depth and a given maximum absolute gradient should be selected, meaning that only defects from the side will be considered, discarding, e.g., a point between the legs. Figure 3 shows an example of how a tall blob containing two people will be split.

Split Wide Blobs

Groups of people standing next to each other might be found as one large blob. To identify which blobs contain more than one person, the height/width ratio and the perimeter are considered. If the criteria are satisfied, the algorithm should try to split the blob. For this type of occlusion, the head of each person is often visible, and the blob can be split based on the head positions. Since the

Figure 2: Examples of challenges in the segmentation of images.

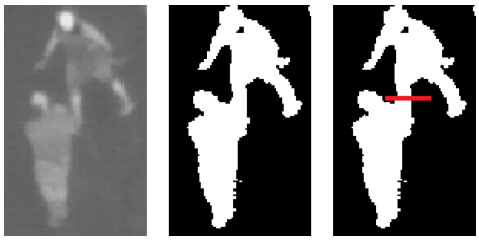


Figure 3: Example of how a tall blob, containing two people, can be split.

head is narrower than the body, people can be separated by splitting vertically from the minimum points of the upper edge of a blob. These points can be found by analysing the convex hull and finding the convexity defects of the blob. Figure 4 shows an example of how a wide blob containing two people will be split.

Figure 4: Example of how a wide blob, containing two people, can be split.

Sort Blobs

The last two challenges, reflections and over-segmentation, can be handled by sorting the blobs. The goal is to find only those blobs containing the feet of a person, as these define the position of the person on the ground.

We consider each binary blob a candidate, and generate a rectangle of standard height at the given position (calculated during calibration) and the width being one third of the height. For each rectangle we evaluate the ratio of foreground (white) pixels. If the ratio of white pixels is below 15%, the blob is discarded, otherwise the candidate is added for further processing. The second step is to check if the candidate rectangles overlap significantly, hence probably belonging to the same person. If two rectangles overlap by more than 45%, only the candidate with highest ratio of white pixels is kept as a true detection. These threshold values are chosen experimentally by evaluating 340 positive samples and 250 negative samples. Figure 5 illustrates this situation, where one person has been split into three blobs. The ultimate goal for the detection algorithm is to detect each person, and nothing else, in each frame. However, with a side-view camera angle and possibilities of a high number of people interacting, people can be fully occluded. These situations can not be solved on frame level, but rather by including temporal information.

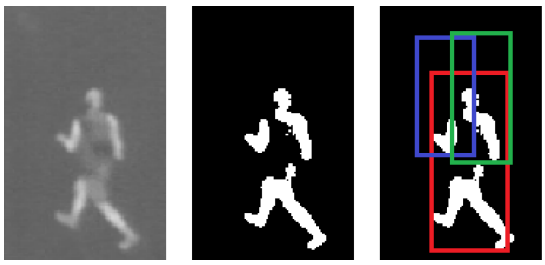


Figure 5: Example of how several candidates are generated and tested.

B. Counting

The goal of this first part of the research in analysis of activities in sports arenas is to count the number of participants. An initial estimate can be found by simply counting the number of blobs existing after the post-processing methods described in section III-A3. However, noise must be expected with both false detections and missed detections. Considering the application in sports arenas, people are mostly moving within the monitored court area, meaning that the number of people will be constant for periods of time. The idea of this counting algorithm is therefore to model the sequences with stable periods, when no people are close to the border, and unstable periods when people are close to the border; likely to leave or enter the monitored area [17]. During stable periods a number of people is

estimated per frame by counting the number of blobs. During unstable periods the number of people leaving or entering the court should be estimated by applying local tracking. Figure 6 illustrates a person crossing the border.

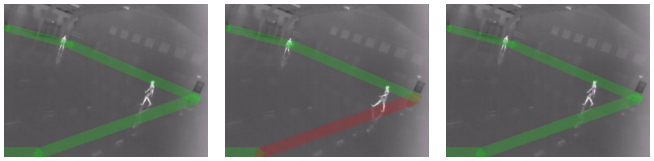


Figure 6: Illustration of the notification of a person crossing the border to the court area.

Following this first iteration, a graph optimisation will run over the entire sequence, combining the estimated numbers from stable periods with changes during unstable periods. The graph consists of nodes, representing the weighted number of people observed during stable periods, and edges, representing the change in number between two stable periods. The weight for each node is calculated based on the probability of each detection being true, and the general uncertainty of a frame caused by occlusions and clutter. For each stable period a weighted histogram is calculated and scaled to an accumulated sum of 1. Details on the weighting algorithm can be found in [17]. Edges are weighted based on the number of crossings, which increases the uncertainty, and the weighting of each individual crossing the border. A simple example of a graph is illustrated in figure 7. Edges exist between all nodes in two consecutive periods, but to simplify the illustration they are not drawn.

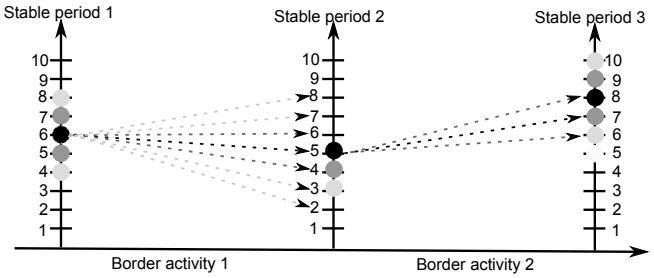


Figure 7: Example of a simple graph [17]. Nodes represents the weighted number of people observed during stable periods, and edges represents the change in number between two stable periods. Dark nodes and edges have the highest weights.

This graph optimised counting algorithm is tested on a 30 minute video sequence containing between 3 and 13 people in each frame at the monitored court area. The mean error in estimated number is 4.44 %. For comparison, using only frame based counts, without graph optimisation, results in a mean error of 8.87 % [17].

IV. Recognising activities

The second step of the analysis of activities in a sports arena is to recognise the activity type. The goal here is to be able to detect the most common sports types observed in a public indoor arena, to provide an overview of activities performed in the arena during each day. The definition of a common sports type depends on the tradition in the relevant geographical area. For this project data has been captured in sports arenas in Aalborg, Denmark. The common sports types observed here were: Team handball, volleyball, badminton, soccer, and basketball. In this section three methods for recognition of sports types will be presented and compared.

A. Heatmaps

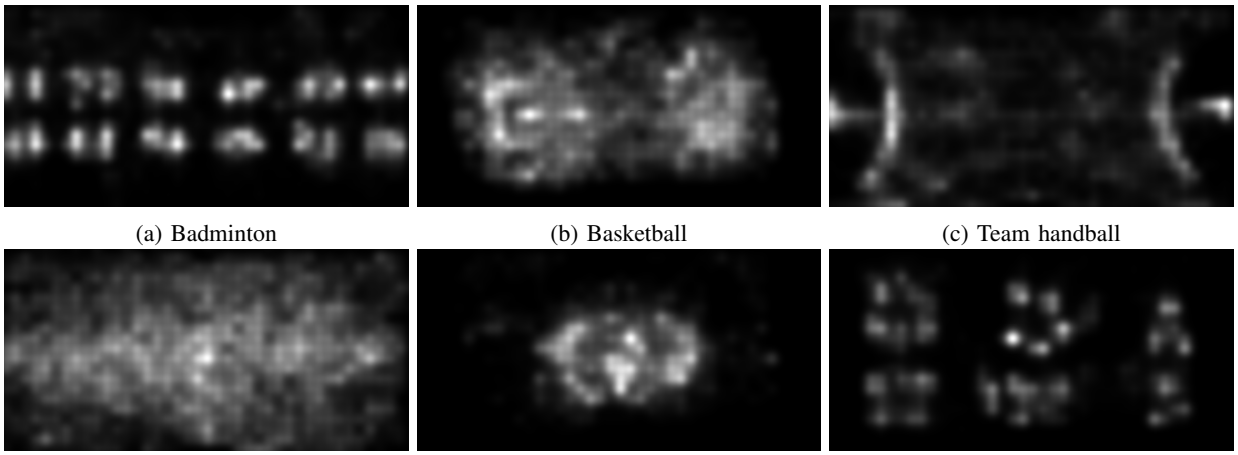
The first method developed is based on heatmaps, representing detected positions over time [18]. A heatmap is constructed by using the position of each detected person. The position must be converted from image coordinates to world coordinates, using a homography found during initialisation of the system. When summing up the positions, the area each person spans is modelled as a Gaussian distribution with a standard height of 1 and a radius corresponding to 1 metre for 95 % of the volume. Each heatmap represents a time span of 10 minutes. Examples of these heatmaps are shown in figure 8.

The classification system aims at classifying each heatmap as one of the five well defined sports types, or as miscellaneous. The heatmaps being 200×400 pixels can be represented as a sample in an 80,000-dimensional space, considering each pixel as a feature. In cases of such high dimensionality, it is beneficial to start with dimensionality reduction. It is, however, important for classification purposes to seek the dimensions which best separates the classes. For this purpose we choose to apply a method appropriate for pixel data, originally proposed for face recognition, called Fisherfaces [19]. This method starts by reducing the number of dimensions using PCA, after which a new projection of data is found using Fisher's Linear Discriminant, which seeks the directions that are efficient for discrimination between classes.

Tested on a total of 386 manually annotated heatmaps, captured over 12 days, the classification result is a true positive rate of 89.64 % [18].

B. Tracklets

The method based on heatmaps has a few limitations, which include the dependency on scale, direction, and location on the field. To overcome these limitations, the second method presented is based on local features from motion, which are invariant to the position and direction of play [20]. Based on trajectories (tracklets) from each player, motion features are extracted and used for classification.



(a) Badminton (b) Basketball (c) Team handball

(d) Soccer (e) Volleyball (1 court) (f) Volleyball (3 courts)

Figure 8: Examples of heatmaps representing each sports type.

Tracklets are produced using a multi-target tracking scheme based on the Kalman filter [21]. In the thermal modality, re-identifying targets after occlusions is very difficult, due to a lack of distinct features. For this reason, instead of trying to construct long trajectories, with a high risk of switching between targets, we aim to construct short, but reliable tracklets, from which we can estimate motion features. Examples of tracklets from five sports types is presented in figure 9.

The motion features chosen should be invariant to the size and direction of the court, the position of the players on the court, and to the direction of play. The features must be robust to noisy detections and tracking errors as well. Based on these criteria, four features are selected: Lifespan [frames], total distance [m], distance span [m], and mean speed [m/s] [20]. These features are extracted for each tracklet in each 2 minute video sequence and combined by the mean value, such that each sequence is represented by a 4-dimensional feature vector. For classification both linear and quadratic discriminant functions (LDA and QDA) are tested, of which QDA shows best performance [22].

For evaluating this method five hours of video was captured and annotated; one hour, corresponding to 30 2-minute sequences, per sports type. A correct classification rate of 94.5 % was obtained [20].

C. Audio-visual features

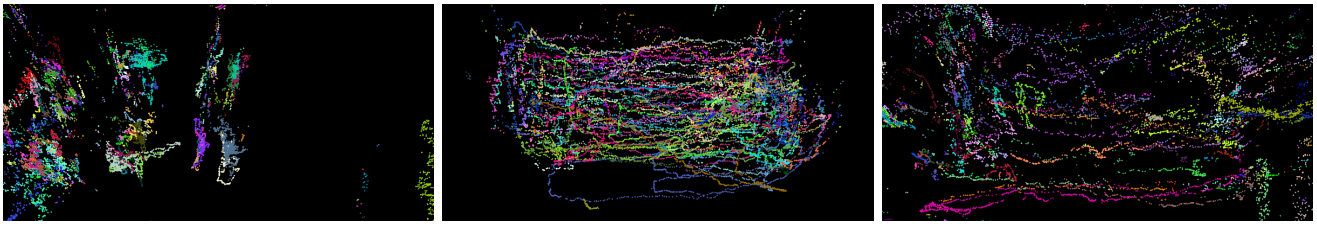
In order to further improve the classification performance, the last method presented in this section combines motion features, as described in section IV-B, with audio features [23]. For audio-features a perceptual time-frequency representation, Mel Frequency Cepstral Coefficients (MFCC) [24], is used. MFCC features are considered one of the main features used for audio signal processing applications such as speech and speaker recognition, and emotion recognition in music and speech. 25 MFCC features plus the log of the energy is extracted, from which also the first and second derivatives are computed. The analysis window is set to 25 ms, with overlaps of 10 ms. As a result, a 78×600 feature matrix is extracted per one minute sequence. These are then summarised by estimating the mean and variance. The result is a 78×2 feature matrix which is then converted to a vector of 156 features. Finally, PCA is performed to reduce the feature space from 156 to 10 features.

This method is evaluated on an audio-visual dataset containing footage from three different sports types, one hour from each sports type; Basketball, soccer, and volleyball. The videos are then divided into 1-minute sequences, which results in a total of 180 sequences to classify. The result from a 10-fold cross validation using a kNN classifier (k=9) is a correct classification rate of 96.11 % [23]. The kNN classifier is chosen for best performance among six applicable standard classifiers tested in WEKA [25].

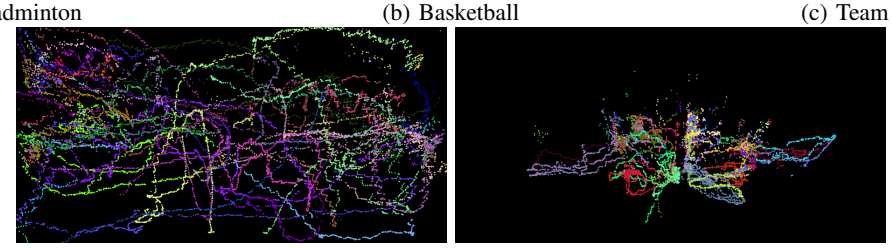
The results presented in this section on activity recognition are directly comparable in terms of image types and arena. However, different datasets were used for each work as new data requirements were added for each method; higher framerate to enable tracking, and audio data, for extraction of audio features.

V. Estimating energy expenditure

The last part of this work concerns the estimation of energy expenditure during sports activities. Current methods applied within health and sports science require direct measurements of oxygen uptake by metabolic carts (wearing respiratory masks) [26] or indirect measurements using heart rate monitors, accelerometers [27], or GPS data [28]. In this work we introduce a non-intrusive method based on thermal video, which is evaluated against traditional measurements of oxygen uptake [29]. The ultimate goal is to be able to



(a) Badminton (b) Basketball (c) Team handball



(d) Soccer (e) Volleyball

Figure 9: Examples of tracklets from a 2-minute period of each sports type. Each tracklet is assigned a random colour.

estimate energy expenditure of all individuals participating in a sports activity, but the first research in this direction will start with a controlled environment using a treadmill experiment. Fourteen endurance-trained test participants performed 4-minute intervals walking and running at the treadmill at velocities of 3, 5, 7, 8, 10, 12, 14, 16, and 18 km/h. Heart rate, oxygen exchange, and mean accelerations of ankle, thigh, wrist and hip were measured for each participant throughout the test. Thermal video was captured with a frame rate of 30 fps from a side view. An example of an input image and the results of histogram normalization and the following threshold segmentation is shown in figure 10.

Figure 10: Input frame, histogram normalization, and segmentation.

Energy values are extracted from optical flow estimations, representing the local movement. First, to stabilize the scene, horizontal global movement is removed, by tracking the upper body with a KLT tracker [30]. Local movement is then estimated by calculating the Large Displacement Optical Flow [31] of the body limbs, quantized into an $N \times M$ grid. A linear correlation between oxygen uptake and optical flow is proved, illustrated in figure 11 with colours representing individual test subjects, and black regression line for pooled data from all subjects.

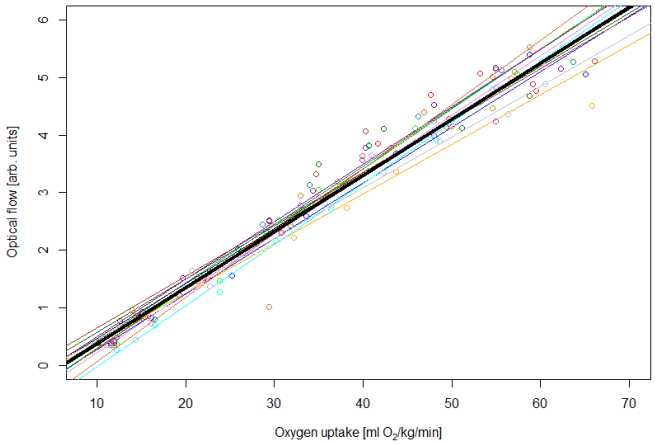


Figure 11: Individual oxygen uptake measurements in relation to optical flow during walking and running. Colours represent individual test subjects, and the black regression line uses pooled data from all subjects.

This work concludes that energy expenditure of a person walking or running at a treadmill can be estimated from a video sequence of 3-4 seconds. Future work in this field will investigate the possibilities of extending the method to free movement in a sports arena.

VI. Discussion

The demand for automatically gathered data is a societal trend quickly extending to all aspects of human life. Knowledge on the utilization of public facilities is of interest for optimising use and cutting expenses for the owners. Manual observations are both cumbersome and expensive, and they have a risk of containing incorrect results due to subjective opinions or lack of interest in the given task.

In this paper we have presented the main results of a 5-year long research project revolving around the real-world application of automatic analysis of activities in sports arenas. The choice of thermal cameras has been essential throughout the project, primarily due to privacy issues. The amount of previous work on analysis of humans in thermal images has been limited, and we believe that we with this project have contributed to this field. We do also support the progress within this field, by making multiple thermal and multi-modal datasets publicly available[1].

The first topic of this project, concerning the task of counting people, has today reached a stage where it is applied commercially. The methods and setup have been tested in a large number of different arenas, and a compromise between precision, hardware costs, and installation time has been found, using a single thermal camera mounted near one of the corners of the court.

Recognising activities is an interesting task, not only used for sports types, but generally for automatic labelling of video data. Within this topic we have proposed three different methods with different potentials and limitations. Classification of heatmaps is the simplest solution, requiring only single frame position data. The limitation that follows is that it is tied to the location and direction of play on the court. These limitiations are eliminated using tracklets instead. This method shows a higher classification rate, however, it has higher requirements for temporal coherence between frames. The last method presented on activity recognition is based on both audio and visual features. The addition of audio features again increased the classification rate. However, capturing audio data might cause privacy concerns in some locations, which is also the reason that the dataset captured for this purpose is smaller than datasets of only thermal video data.

The last part of this paper presents an initial study for a non-intrusive method of estimating energy expenditure during sports. Using a treadmill experiment we show a linear correlation between optical flow measurements and oxygen uptake, which is the traditional way of measuring energy expenditure. This work will be continued for investigating the possibilities of extending the method for analysis of free movement in a sports arena. For this purpose, the influence of several challenges will have to be further researched; Going from treadmill running to free movement will result in changing pose and different viewing angles of the body, which has not been considered in the method yet. Furthermore, when observing several people participating in an activity, occlusions between people need to be considered. To limit the amount of occlusions, the camera can be mounted at a higher location. However, this will result in a non-perpendicular angle to the bodies. Possible solutions for both setup and methods will be the next research topic related to this project.

The work presented in this paper has mainly considered the global behaviour observed in a sports arena. With our latest research in energy expenditure we move towards analysis of the individual people participating in sports activities. From here, we start to reach an area which might be interesting to both athletes, coaches, and broadcasters of professional sports, as well as professionals working with exercise intensity during physical education.

[1] http://www.vap.aau.dk/dataset/

REFERENCES

[1] R. Gade and T. B. Moeslund, "Thermal cameras and applications: A survey," *Machine Vision & Applications*, vol. 25, no. 1, pp. 245–262, 2014.

[2] T. B. Moeslund, G. Thomas, and A. Hilton, Eds., *Computer Vision in Sports*. Switzerland: Springer, 2014.

[3] C. B. Santiago, A. Sousa, M. L. Estriga, L. P. Reis, and M. Lames, "Survey on team tracking techniques applied to sports," in *International Conference on Autonomous and Intelligent Systems (AIS)*, June 2010, pp. 1–6.

[4] C. J. Needham and R. D. Boyle, "Tracking multiple sports players through occlusion, congestion and scale," in *British Machine Vision Conference*, 2001, pp. 93–102.

[5] H. Saito, N. Inamoto, and S. Iwase, "Sports scene analysis and visualization from multiple-view video," in *IEEE International Conference on Multimedia and Expo*, vol. 2, june 2004, pp. 1395 –1398 Vol.2.

[6] J. Xing, H. Ai, L. Liu, and S. Lao, "Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1652 –1667, june 2011.

[7] J. Liu, P. Carr, R. T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[8] T. Zhang, B. Ghanem, and N. Ahuja, "Robust multi-object tracking via cross-domain contextual information for sports video analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[9] J. Xiao, R. Stolkin, and A. Leonardis, "Multi-target tracking in team-sports videos via multi-level context-conditioned latent behaviour models," in *Proceedings of the British Machine Vision Conference*, Sept. 2014.

[10] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," in *37th IEEE Applied Imagery Pattern Recognition Workshop*, oct. 2008, pp. 1 –8.

[11] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473 –1488, nov. 2008.

[12] W. Wei and A. Yunxiao, "Vision-based human motion recognition: A survey," in *Second International Conference on Intelligent Networks and Intelligent Systems*, nov. 2009, pp. 386 –389.

[13] T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, *Visual Analysis of Humans - Looking at People*. Springer, 2011.

[14] E. S. Jeon, J. H. Kim, H. G. Hong, G. Batchuluun, and K. R. Park, "Human detection based on the generation of a background image and fuzzy system by using a thermal camera," *Sensors*, vol. 16, no. 4, p. 453, 2016.

[15] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273 – 285, 1985.

[16] R. Gade, A. Jørgensen, and T. B. Moeslund, "Occupancy analysis of sports arenas using thermal imaging," in *Proceedings of the International Conference on Computer Vision and Applications*, feb. 2012.

[17] ——, "Long-term occupancy analysis using graph-based optimisation in thermal imagery," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[18] R. Gade and T. B. Moeslund, "Sports type classification using signature heatmaps," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2013.

[19] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *PAMI*, vol. 19, no. 7, pp. 711 –720, jul 1997.

[20] R. Gade and T. B. Moeslund, "Classification of sports types from tracklets," in *KDD Workshop on Large-Scale Sports Analytics*, aug. 2014.

[21] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[23] R. Gade, M. Abou-Zleikha, M. G. Christensen, and T. B. Moeslund, "Audio-visual classification of sports types," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec 2015, pp. 768–773.

[24] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[25] "WEKA, University of Waikato, New Zealand," http://www.cs.waikato.ac.nz/ml/weka/.

[26] A. V. Hill and H. Lupton, "The oxygen consumption during running," *J Physiol.*, vol. 56, pp. xxxii–xxxiii, 1922.

[27] B. W. Fudge, J. Wilson, C. Easton, L. Irwin, J. Clark, O. Haddow, B. Kayser, and Y. P. Pitsiladis, "Estimation of oxygen uptake during fast running using accelerometry and heart rate," *Med Sci Sports Exerc*, vol. 39, no. 1, pp. 192–198, 2007.

[28] S. Costa, D. Ogilvie, A. Danton, K. Westgate, S. Brage, and J. Panter, "Quantifying the physical activity energy expenditure of commuters using a combination of global positioning system and combined heart rate moniters," *Preventive Medicine*, vol. 81, pp. 339–344, 2015.

[29] M. M. Jensen, M. K. Poulsen, T. Alldieck, R. G. Larsen, R. Gade, T. B. Moeslund, and J. Franch, "Estimation of energy expenditure during treadmill exercise via thermal imaging," *Medicine and Science in Sports and Exercise*, 2016, e-pub ahead of print.

[30] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proceedings of the seventh international joint conference on Artificial intelligence*, vol. 2, pp. 674–679, 1981.

[31] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *IEEE Trans Pattern Anal Mach Intell.*, vol. 33, no. 3, pp. 500–513, 2011.