OCCUPANCY ANALYSIS OF SPORTS ARENAS USING THERMAL IMAGING

Rikke Gade, Anders Jørgensen and Thomas B. Moeslund
Visual Analysis of People Laboratory, Aalborg University, Aalborg, Denmark

Keywords: Thermal Imaging, Image Processing, Human Detection.

Abstract: This paper presents a system for automatic analysis of the occupancy of sports arenas. By using a thermal camera for image capturing the number of persons and their location on the court are found without violating any privacy issues. The images are binarised with an automatic threshold method. Reflections due to shiny surfaces are eliminated by analysing symmetric patterns. Occlusions are dealt with through a concavity analysis of the binary regions. The system is tested in five different sports arenas, for more than three full weeks altogether. These tests showed that after a short initialisation routine the system operates independent of the different environments. The system can very precisely distinguish between zero, some or many persons on the court and give a good indication of which parts of the court that has been used.

1 INTRODUCTION

In the modern world jobs are becoming ever more sedentary and less physically demanding. This leads to higher demands for activities in people's spare time, which puts a still growing pressure on the sports arenas. From 1964 to 2007 the number of athletes has quadrupled with a steady increase (Pilgaard, 2009). Surveys also show that people are dropping the classic club sports in favour of more flexible sports (Brixen et al., 2010). This calls for a better and more optimal use of the existing sports arenas to keep up with this growing trend.

In order to improve the utilisation of a sports arena, its existing use must be examined. This includes examining the number of users using the arena at the same time and the occupancy of the court. Administrators are especially interested in whether the arena is empty, used by a few people or full and the time for when the occupancy changes. The position of the users is also important as they might only use half a court, which means the other half could be rented out to another group. Manual registration of this would be cumbersome and expensive and an automatic approach is therefore needed. For such a system to work in general it should be independent of the size of the court, lighting conditions and without any interaction with the users. This can be obtained with a camera.

Detecting people with a camera raises some privacy issues though. Not all people like surveillance and the fear of being observed could keep some people out of the arenas. This work therefore proposes an automatic method to analyse the occupancy of a sports arena using thermal imaging. One of the advantages of thermal cameras is that the persons recorded cannot be identified, which is an important factor if the system is to be accepted by the users of the sports arena. On top of that, thermal cameras are invariant to lighting, changing backgrounds and colours, which make them more desirable for a general application.

2 RELATED WORK

Automatic detection and tracking of sports players is a research area important for all sports analysis. Most systems are using visual cameras. In (Needham and Boyle, 2001) a tracking system is proposed specifically for indoor football players, while (Saito et al., 2004) proposes a tracking system for outdoor football using multiple cameras. The tracking system proposed in (Xing et al., 2011) focuses on more general sports video and it is tested on both football, basketball and hockey.

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 (Ko, 2008), (Turaga et al., 2008), (Wei and Yunx-

Gade R., Jørgensen A. and B. Moeslund T..
OCCUPANCY ANALYSIS OF SPORTS ARENAS USING THERMAL IMAGING.
DOI: 10.5220/0003843202770283
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 277-283
ISBN: 978-989-8565-04-4

iao, 2009) and (Moeslund et al., 2011).

Thermal cameras measure the amount of thermal radiation, which lies in the long-wavelength infrared spectrum (8-15 μm). All objects with a temperature higher than the absolute zero emit thermal radiation. The intensity and dominating wavelength depends on the temperature.

Thermal cameras have a clear advantage over visual cameras in night conditions, therefore the main focus for systems using thermal cameras have been on security applications and trespasser detection. A few papers with the purpose of detecting trespassers include (Wong et al., 2009) and (Wong et al., 2010).

Other work using thermal cameras include systems for pedestrian detection and tracking. In (Wang et al., 2010) a pedestrian detection method is presented based on the Shape Context Descriptor with the Adaboost cascade classifier framework. (Bertozzi et al., 2003) proposes the pedestrian detection as part of a driver assistant system while (Davis and Sharma, 2004) proposes a people detection system for different environments based on contour analysis.

Most vision systems, including the systems mentions above, are only tested on very short video sequences, proving the concept in one or few conditions. In this work the most important issue is stability over a long time period and under different conditions. Therefore the system will be tested over three weeks and in five different arenas. The main results will be average values showing the tendency of occupancy for hours or days.

3 METHODS

3.1 System Overview

The desired system should take a thermal grey scale image as input and find every person in the image. In order to analyse the nature of the problems related to this work, five different sports arenas were selected and used to develop and test the system. During the initial investigations some typical difficulties to obtain the result were registered. Some of these difficulties were occlusions and reflections from both persons and other warm objects, e.g. lamps, on the floor. These typical difficulties must all be addressed in order to make a general system.

As the intention is to monitor the long term use of a sports arena, the system should always be operating. Therefore data should be processed in soft real-time to avoid data pile ups. The output of the system should be, for a given time, the number of users on the court and their position.

The system should be independent of the camera's viewing angle in relation to the court, as long as the camera can observe the court and is placed in a sufficient height to avoid users covering each other too much. The users' size, level of activity and posture should not have an effect on the measurement either. Figure 1 shows a diagram of the system structure.

The overall idea is to develop a system that can detect persons in a thermal image and, with the inputs from the initialisation, find the persons' positions at the court. As mentioned in the introduction, the data should be categorised after the occupancy level into zero, some and many people and presented in a timetable.

3.2 Initialising the System

The initialisation routine must be conducted for each new mounting of the camera. This routine handles the adjustments necessary to fit the system to the layout of the actual sports arena. First the court must be found in the image to avoid that cold or hot objects outside the court influences the system. As it is only wanted to measure the players at the court, spectators should also be removed. Defining the court in the images gives the opportunity to remove all objects standing outside the court. During the initialisation the corners of the visible part of the court should be marked in the image, and lines connecting them define the outline of the court. To find the persons' positions at the court a mapping from the image to the court must be found. Using at least four corresponding points in image and world coordinates, a homography matrix **H** can be calculated (Criminisi, 1997).

After initialising this matrix the mapping between image coordinates and world coordinates is calculated as $P_w = Hp_i$, where P_w are the weighted world coordinates $[P_X\ P_Y\ W]^T$ and p_i are the image coordinates $[p_x\ p_y\ 1]^T$. The real world coordinates are found by dividing P_w with the weight W.

At least four corresponding points must be used in order to calculate **H**, but tests of the homography show that using more points increase the precision. This is due to nonlinearity in the mapping, as the lens has some barrel effect. Therefore it is desirable to use as many points in the initialisation as possible. In this work a two-dimensional grid with steps of 5 metres is used to mark the points at the court. A hot or cold object is necessary to detect the grid points in the image.

3.3 Run-time

This continuous loop receives an image from the ther-

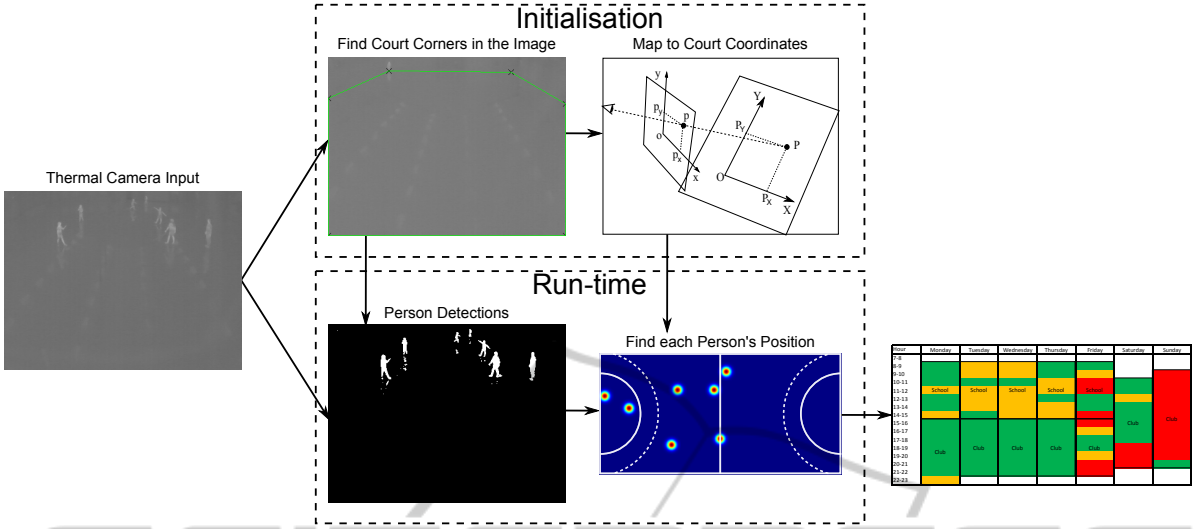


Figure 1: Diagram showing the system steps.

mal camera and after a number of functions it delivers a set of regions each containing one person. First the thermal camera captures a grey scale image. The warm objects (persons) are bright while the surroundings are dark grey. After capturing a frame the first step is to extract the warm objects. For this an automatic threshold method based on Maximum Entropy is used (Kapur et al., 1985). This method maximises the sum of the entropy above and below the threshold value s, by iterating through every possible value.

The threshold function is only run for the pixels inside the court area, to avoid disturbance from spectators. The result is a binary image where ideally the persons are white and anything else is black. If the maximum entropy is below a specified threshold TH there are no persons on the court and the frame can be discarded.

The white regions are now found using the contour finding algorithm described in (Suzuki and Abe, 1985). If there are no valid regions i.e. regions larger than a specified minimum area the frame is discarded.

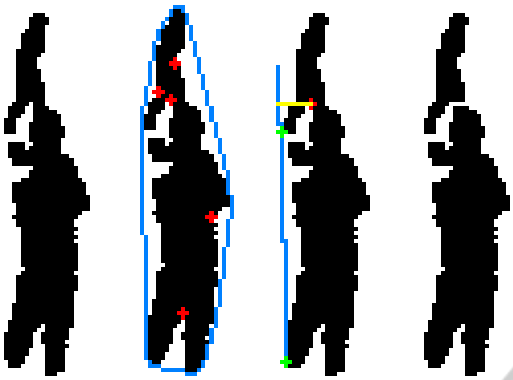
3.3.1 Split Tall Regions

People standing behind each other, seen from the camera's point of view, can often be found as one tall region as shown in figure 2(a). In order to split such regions into the right number of people, it must be investigated when a region is too high to contain only one person. Using the camera's height c, the vertical resolution r_v and vertical field-of-view f_v of the camera, the height in pixels can be found as a function of the person's height p and distance to the camera x:

$$y_p = \frac{r_v \cdot \left(tan^{-1}\left(\frac{x}{c-p}\right) - tan^{-1}\left(\frac{x}{c}\right)\right)}{f_v}$$

Statistics show that only 0.26% of Danish conscripts were taller than 2 metres (DST, 2006), therefore 2 metres is chosen as the height limit. So for each region found in the image the distance to the camera is calculated using the homography and if the pixel height corresponds to more than 2 metres, the algorithm should try to split the region horizontally. This is done by finding the convex hull and the convexity defects of the contour, as shown in figure 2(b). The point selected to split from is the defect point with the largest depth and a maximum absolute gradient of 1.5. The gradient is calculated for the line from the defect point perpendicular on the line between the convexity defect start and end points (green points and yellow line in figure 2(c)). The defect point between the legs of the person has the largest depth, but is discarded because the gradient is too high. Also the defect point should not be in the top or bottom fourth of the region, to avoid e.g. feet or head to be split from the body. As shown in figure 2(d) the region is split horizontally from the selected point.

The algorithm starts with the defect point with largest depth and continues until a point with an acceptable gradient and location is found. If no accepted points are found, the region will not be split. If the region has been split, the algorithm will start over and examine the height of the resulting regions.

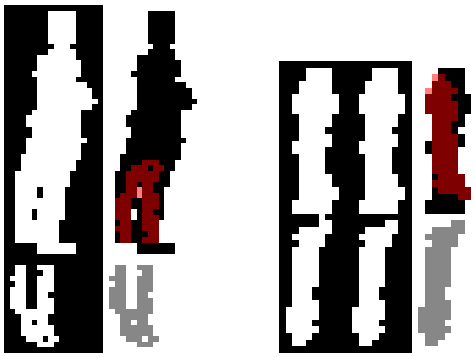


(a) (b) (c) (d)

Figure 2: Example on division of tall regions. Note that black and white colours are reversed for better visibility. The blue line is the convex hull, red marks indicate convexity defects and the yellow line the orthogonal depth of the defect.

3.3.2 Remove Reflections

Just as visible light, infrared waves are also reflected in glossy surfaces, but as the infrared reflections are created by the persons themselves they are always pointing towards the camera. Therefore the mirror axis will be roughly horizontal, and reflections could be removed by trying to mirror them in a region above. An example can be seen in figure 3(a)(Left).



(a) (b)

Figure 3: Persons having their reflection removed. The red areas mark the reflections after they have been mirrored and translated. In (b) the reflection is first split from the person by the algorithm splitting tall regions.

In order to remove reflections the system searches for regions that are below a larger or equally sized region. If such a region is found it is mirrored up in the upper region to see if it fits in the person region. If it does not, the reflection is translated one pixel horizontally and checked again. This continues up to three pixels in all directions. If more than 90 % of the reflection is within the person region it is marked as a reflection and removed. Figure 3(a)(Right) shows a situation where 77 out of 79 pixels are within the person region resulting in a coverage of $\approx 97\%$.

In some cases the reflection is connected to the person who created it. See figure 3(b)(Left). In these situations the region should first be split by the function splitting tall regions. Figure 3(b) shows a situation where a region is first split and secondly the reflection can be removed. Here 72 of 74 ($\approx 97\%$) reflection pixels are within the person.

3.3.3 Split Wide Regions

People standing close to each other will often form one large region. In order to count the people correct such regions must be divided into regions containing only one person. For groups of people standing side by side, seen from the camera's point of view, it will often be possible to separate them based on their head position. Since their heads are narrower than the body they can often be separated by cutting vertically from the minimum points of the upper edge.

As it is not desired to split regions containing only one person, two criteria for the regions must be satisfied before looking for a minimum point to split from. Measuring the features of several regions gives the criteria that to contain more than one person the height of the bounding box must be less than five times the width and the contour of the region must be longer than the bounding box perimeter. If these criteria are satisfied and a minimum point can be found at the upper edge of the region, the region will be divided.

The points are now found as convexity defects in the same way as described for the tall regions. Instead of measuring the angle this method uses the y-coordinates of the points. The found point must be located on the upper edge of the region and have a y-value greater than both the convexity defect's start and end point to make it a minimum point.

As for splitting the tall regions, the algorithm will continue until no more regions are split.

3.3.4 Sort Regions

The final step is to sort the regions. After the regions have been split and the reflections have been removed, the remaining regions are now investigated before they are counted as a person. If a region's area does not match its distance to the camera it is removed. This could be a small region which is found in the foreground where persons typically would be larger. This step also calculates the person's position

on the court, using the homography from the initialisation. This is done for the lowest middle pixel of the accepted regions, which will be the position on the floor.

3.4 Occupancy of the Court

A user's position will be given as a x,y coordinate with multiple decimals. In order to examine the occupancy a method must be found that preserves the position, but also mimics the size of a person. Therefore every found region is represented as a 3D gaussian distribution with a height of 1 and $\sigma = 5$, equivalent to a radius of 1 metre for 95 % of the volume. This is also roughly the radius of a person. An example for one frame can be seen in figure 4 where 8 persons have been found.

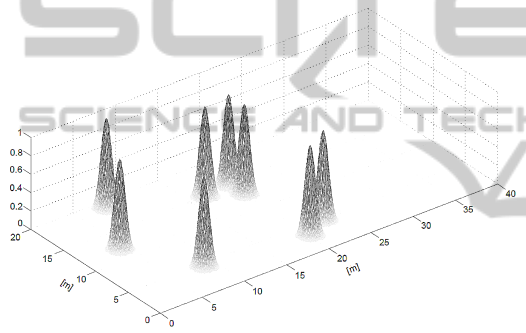


Figure 4: A single frame where 8 persons have been found.

For longer periods these frames can be summed to show the occupancy of the court during e.g. an hour.

4 RESULTS

4.1 Objective

As described in section 2 a very important parameter for the system is the stability in changing conditions and in changing set-ups. Therefore the system is tested in five different arenas, capturing more than three weeks altogether. For all tests in different arenas the same parameters of the system has been used. Only the initialisation of the system, described in section 3.2, depends on the arena. By measuring the entropy of a number of frames with and without people, the entropy threshold TH is chosen to be 4.1. The thermal camera used in the test is an AXIS Q1921-E, with a resolution of 384×288 pixels and a horizontal field-of-view of 55°.

4.2 Annotation of Data

Capturing three weeks continuously with 30 fps gives a total of 54,432,000 frames, which would be nearly impossible to manually annotate. Therefore it is chosen to manually annotate 54,000 frames, resulting in 30 minutes of video. This will be used for calculating the precision of the system. But as this test does only evaluate the system during one specific activity an additional test will be conducted. A period of 36 consecutive hours will be sampled and manually annotated with 0.04 fps (1 frame per 25 seconds). This is covering two days with different sports activities and a night. Even though the frame rate here is low this will still give a good evaluation of the system and ensure that it is tested with both a varying number of people and different types of sports. The data from the full test period of more than three weeks will be evaluated by random checks against the videos.

4.3 30 Minutes Test

The results for the 30 minutes period are calculated as a mean number for every five minutes. The automatic results compared to the ground truth (manually annotated data) are shown in figure 5 with black and red.

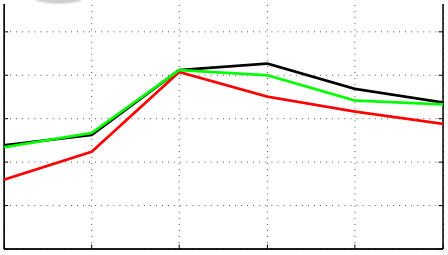


Figure 5: Manual (red) and automatic (black) result for six five minutes periods, sampled with 30 frames per second, and automatic results (green) sampled with 0.04 frames per second.

Calculating the error of the automatic system, sampled with 30 fps, for each five minutes period gives an average error of 20.5 %. Comparing the green line to the black line shows that in 4 out of 6 periods the automatic results with different sample rates are nearly the same, while for the last two periods the difference is about 0.5 person. From this it is concluded that even with a sample rate decreased to 0.04 fps the results will still be reliable.

4.4 Two Days Test

For the 36 hours, sampled with 1 frame per 25 seconds, a mean error is found for every five minutes, and stated as a mean error for each hour, since the activities in the arena are typically the same for at least an hour. This method is used for both the error measured in persons and per cent. The hours are then categorised by the maximum number of people, to investigate the relation between the error and the number of persons. See the results in table 1.

Table 1: Error categorised by the maximum number of people during the hour.

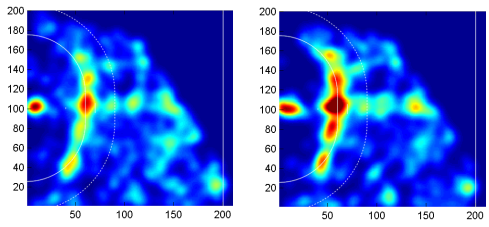
# persons	# hours	Mean error	Mean error (%)
0	12	0.0017	0.17 %
1-2	15	0.0428	7.35 %
7-15	9	0.5100	11.76 %

For the nine hours with maximum 7-15 persons on the court the error for each hour lies from 4-20 %, with an average of 11.76 % as stated in table 1. It is clear that the error for detecting empty arenas is very low, and the error increases with the number of persons. This will typically be due to occlusions. As mentioned in section 3.1 the occupancy level should be categorised to zero, some or many users, which means that the precise number of people is not critical for this application.

The 30 minutes test described in section 4.3 showed an error of 20.5 %, which is equivalent to the maximum error found during this two day test. The video of 30 minutes had a high activity level and a highly varying number of people on the court, with up to 14 people in each frame. Therefore it is also expected that it should have a higher error than the average videos.

4.5 Evaluation of Positions

The calculated positions of the persons will be evaluated by visually comparing the manually marked positions with the automatic found positions. This is done for the 36 hours sampled with 1 frame per 25 seconds. An example of one hour showing a handball match can be seen in figure 6. The upper image shows the positions found by the system and the lower image shows the true position found manually for the same period. There are found more people manually than automatic during this hour, resulting in darker colours in the bottom image, but it is evident that there is a high correlation between the two images and the overall picture is the same. Note that the camera could only see the left half of the court.



(a) (b)

Figure 6: Positions of users during a handball match. Left: Automatic. Right: Ground truth. Note that there are found more people manually than automatic, resulting in darker colours in the bottom image.

As mentioned in section 1 the position should be used to examine whether the entire court is being used or only part of it. Therefore the main point in evaluating the found positions is not to examine the position of each person, but to ensure that the overall picture of the occupancy during a booking is correct. This correlation between automatic and manually found positions is found to be very high for all 36 hours.

4.6 One Week Evaluation

The main objective for this system is to analyse the use of the sports arenas. Most sports arenas in Denmark have a booking system, where the local schools and sports clubs book their hours in the arena. To evaluate the use of the arenas the bookings should be considered. Seven consecutive days in one sports arena has been chosen, and the use is here measured as a mean number of persons per hour. The number is categorised as zero, some or many persons to describe the level of occupancy. Table 1 showed that the precision of the system depends on the number of persons, the error increases when the number increases. As the error is very low for detecting empty arenas and few people on the court, the error of the exact number will not have a visible effect on the categorisation.

Finally the utilisation is compared to the booking as shown in figure 7. White areas are not booked, red areas are booked but never used, orange areas are booked and used by two or less persons in average, the green area are booked and used by more than two persons in average while the blue areas are used by more than two persons, but not booked. During this test a frame rate of 1 fps has been used.

Figure 7 indicates that during the measured seven days 21.2 % of the booked hours are not used, while 23.4 % are used by an average of two or less persons, which either means that the arena has only been used for a very short period of the hour, or there have been only one or two people at the court. One hour are used but not booked, which could also be a problem,

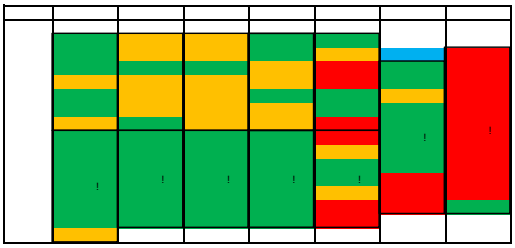


Figure 7: Table of utilisation compared to the booking. White areas are not booked, red areas are booked but never used, orange areas are booked and used by two or less persons in average, green areas are booked and used by more than two persons in average and blue areas are used by more than two persons, but not booked.

depending on the policy for the administration of the arena.

5 CONCLUSIONS

This work presented an approach for automatic detection of persons using thermal cameras. For the intended application in sports arenas the privacy issue is important, therefore a thermal camera is chosen.

The system shows very satisfactory results, with only a short initialisation it works independently of the changing conditions in different arenas. The system can easily distinguish between an empty arena, few or many people. The work will continue with further tests of the system and work on improving the segmentation of people. This could be by including temporal information or by using a more detailed human template for comparison with the found regions. For future work there are a lot of possibilities for developing new features, including analysis of the activity level, activity type and user type.

ACKNOWLEDGEMENTS

We would like to thank Aalborg municipality for support and for providing access to the sports arenas.

REFERENCES

Bertozzi, M., Broggi, A., Grisleri, P., Graf, T., and Meinecke, M. (2003). Pedestrian detection in infrared images. In *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, pages 662 – 667.

Brixen, S., Larsen, K. H., Lindholm, J. V., Nielsen, K. F., and Riiskjær, S. (2010). *Strategi 2015: En Situationsanalyse (Strategy 2015: A Situation Analysis)*. DGI.

Criminisi, A. (1997). Computing the plane to plane homography. Technical report, University of Oxford.

Davis, J. and Sharma, V. (2004). Robust detection of people in thermal imagery. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 713 – 716 Vol.4.

DST, D. S. (2006). Tabel 44: De værnepligtiges højde (conscripts' height in 2006). http://www.dst.dk/aarbogstabel/44.

Kapur, J., Sahoo, P., and Wong, A. (1985). A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*, 29(3):273 – 285.

Ko, T. (2008). A survey on behavior analysis in video surveillance for homeland security applications. In *Applied Imagery Pattern Recognition Workshop, 2008. AIPR '08. 37th IEEE*, pages 1 –8.

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

Needham, C. J. and Boyle, R. D. (2001). Tracking multiple sports players through occlusion, congestion and scale. In *British Machine Vision Conference*, pages 93–102.

Pilgaard, M. (2009). *Sport og Motion i Danskernes Hverdag (Sport and Exercise in the Everyday Life of Danish People)*. Idrættens Analyseinstitut.

Saito, H., Inamoto, N., and Iwase, S. (2004). Sports scene analysis and visualization from multiple-view video. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, volume 2, pages 1395 –1398 Vol.2.

Suzuki, S. and Abe, K. (1985). Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32 – 46.

Turaga, P., Chellappa, R., Subrahmanian, V., and Udrea, O. (2008). Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473 –1488.

Wang, W., Zhang, J., and Shen, C. (2010). Improved human detection and classification in thermal images. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 2313 –2316.

Wei, W. and Yunxiao, A. (2009). Vision-based human motion recognition: A survey. In *Intelligent Networks and Intelligent Systems, 2009. ICINIS '09. Second International Conference on*, pages 386 –389.

Wong, W. K., Chew, Z. Y., Loo, C. K., and Lim, W. S. (2010). An effective trespasser detection system using thermal camera. In *Computer Research and Development, 2010 Second International Conference on*, pages 702 –706.

Wong, W. K., Tan, P. N., Loo, C. K., and Lim, W. S. (2009). An effective surveillance system using thermal camera. In *Signal Acquisition and Processing, 2009. ICSAP 2009. International Conference on*, pages 13 –17.

Xing, J., Ai, H., Liu, L., and Lao, S. (2011). Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling. *Image Processing, IEEE Transactions on*, 20(6):1652 –1667.