Camera-to-Camera Tracking for Person Re-identification within Thermal Imagery

Student Name: T.A. Robson

Supervisor Name: T.P. Breckon

Submitted as part of the degree of MEng Computer Science to the

Board of Examiners in the School of Engineering and Computing Sciences, Durham University

***Abstract* — Context/Background**

Although use of thermal imagery currently poses significant advantages for 24/7 surveillance in terms of the visibility of human, animal and vehicle targets under all environmental conditions, a key limitation is the lack of colour. Many current approaches to the problem of cross-camera person re-identification, or the camera-to-camera tracking of pedestrian targets, rely on colour features (Gong et al., 2014). Person re-identification across multiple cameras is a key research problem within the domain of visual surveillance and a key challenge for the future deployment of thermal sensing as an autonomous sensor technology.

Aims

The aim of this project is to develop a system that would build upon and extend the range of existing thermal image detection, tracking and classification approaches. The first step for this would be to be able to detect a person within thermal video in real time, distinguish a person from other objects and track a person moving through a scene in real time. We would then implement features/attributes from the literature studied and carry out a first pass or initial framework for re-identification using these features/attributes. The system must be able to re-identify 3+ people within a test video sequence in real time using multiple cameras. This would then be extended by implementing some more advanced features/attributes and extending the re-identification ability to 6+ people.

Method

This system will be implemented in C++ using the widely used OpenCV computer vision library(Bradski, 2000), and will be run simultaneously on multiple video files, each being able to access the same array of previously identified targets. The first stage is real time person detection, which is done using Mixture of Gaussians Background Subtraction, Histogram of Oriented Gradients or Haar Cascade Person Detection and a Kalman filter per person to facilitate tracking a person and store their position, velocity and bounding box dimensions. From here, each of these people will have some of their features stored. The basic features used are Hu Moments, the Histogram of Thermal Intensity and Histogram of Oriented Gradients of the target. From here, more advanced features are implemented, these are the Thermal Correlogram and Optical Flow. To comparison and re-identification will be performed using Mahalanobis distance.

Results

Having carried out evaluation of the features, we established that Histogram of Intensities was the most effective feature, with Thermal Correlogram and Optical Flow also considered to be suitable for re-identification. We implemented a re-identification system as described in the method above, using Histogram of Intensities as our feature, and got very good performance on a single camera system. However, due to differing viewpoints, the multiple camera system did not perform as well, but still gave a good degree of accuracy in certain situations.

Conclusions

We conclude that the work carried out here is a useful starting point, as we have been able to approve or eliminate five potential thermal re-identification features and develop a classifier to use our best feature to create a functioning re-identification system over both a single camera and multiple cameras, across a range of people who have a variety of thermal signatures. Future work could look to investigate more features, try to explore the full potential of the Thermal Correlogram and Optical Flow, experiment with different camera viewpoints and improve upon the Person Detection system.

Keywords: Computer Vision, Person Re-Identification, Re-ID, Person Tracking, Features, Attributes, Thermal Imagery, Thermal Video

I INTRODUCTION

A fundamental task for a distributed multi-camera surveillance system is to associate people across different camera views at different locations and times (Gong et al., 2014). This is referred to as the Person Re-Identification Problem (Gong et al., 2014) and is an interesting and important problem within the field of computer vision, with many different approaches being taken to try and perform this process efficiently and reliably, mainly revolving around the use of features or attributes of a person (Layne et al., 2014). This has been researched widely in colour space, but there has been very little research done on solving this problem in thermal space.

There are many potential applications for this technology, but the most important in the modern world would be to support human intelligence organisations. The surveillance data that a system like this can provide would be critical for crime-prevention, forensic analysis, and counter-terrorism activities in both civilian and governmental agencies alike. While this is currently widely used by human operators, these operators have to be trained, which offsets the utility of this approach with training and staffing costs. The implementation of an automated re-identification system is therefore of great interest, as it would be very useful in supporting these human operators and enabling them to achieve better results more efficiently. This would be particularly important in thermal space, as many of the features that we propose to use are complex and mathematically grounded, making them difficult for a human analyst to observe.

From the previous work ~~that we have studied,~~ we can see that a substantial amount of research that has been done on person re-identification, but much of this relies on the colour spectrum, with attributes of the form "red shirt" (Layne et al., 2014). However, in the modern world, thermal imagery is often used for 24/7 surveillance when all environmental conditions are possible. Therefore, it is important that an effective re-identification system is developed to utilise this area of surveillance. Whilst thermal imagery has many advantages, it is not able to identify colour, making attributes that rely on colour useless. Therefore, alternative features are required to facilitate re-identification. Recently there has been an increase in the amount of research into thermal tracking, and some focus on re-identification, although mainly in robotics, rather than in a distributed camera network, as discussed in (Koenigs and Schulz, 2013) and (Ciric et al., 2013).

The intent of this project is to expand upon this existing research and answer the question *"Which features of a human target are appropriate to facilitate re-identification in thermal video?"* The challenges associated with answering this question are mostly due to the increased complexity of the features required in thermal space, as we do not have simple colour features available to us. Such colour features, as described in more detail in (Layne et al., 2014), have many applications from clothes colour to hair and skin colour. In thermal, we cannot use such simple features, ~~so must think in a more complex way.~~ The Figure below shows five different views of the same person. The features we use must be able to recognise that all of these are the same person.

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

Figure 1: Example of Thermal Views

The first thing that we shall consider as a feature to solve the re-identification problem is a measure of the size of the target person. Due to the single viewpoint of the camera, we need a measure of size that is position, scale and rotation invariant. For this purpose, we will use moments, or Hu Moments in particular, taking inspiration from (Flusser et al., 2009; Dawson-Howe, 2014). The closest we can get to a colour value per pixel in thermal is the intensity of each pixel, so a Histogram of these Intensities per target will be a key feature for us to use. We will also use the Histogram of Oriented Gradients

(HOG) descriptor as a feature for re-identification. From here, we must consider more complex features. The first of these that we will consider draws inspiration from (Jing Huang et al., 1994; Zhang et al., 2015), which suggest the use of a colour correlogram. This method gives an idea of spatial correlation of colours, and we will adapt it to get an idea of the spatial correlation of intensities. The other advance feature will be Optical Flow, which will be used to capture the gait characteristics of the person.

A Aims of the Project

The aim of this project is to investigate the above research question by developing a system that can detect a person within thermal video imagery in real time, distinguish this person from other objects in the scene and track this person moving through a scene in real time. We would then implement 3 features/attributes, and use these to carry out a first pass or initial framework for re-identification. At this stage, the system must be able to Re-Identify at least 3 people within a test video sequence in real time using multiple cameras. This would then be extended by implementing 2 more advanced state of the art features/attributes and extending the re-identification ability to at least 6 people.

B Achievements of the Project

We have managed to achieve a full evaluation of all the features being examined, with Histogram of Intensities being the best, and Thermal Correlogram and Optical Flow also being considered suitable for use in re-identification. We have developed a functioning real time person detection, identification and tracking system, and have built a re-identification system on top of this using the Histogram of Intensities feature, and this has given us accurate results on a single camera. The problem becomes more complex in a multiple camera system, and although we do still see a reasonable amount of success from our system, the different viewpoints and larger number of targets reduces the accuracy of our classifications

II RELATED WORK

A Overviews of topic

When beginning to investigate this area, it is important that we first gain an overall understanding of the field before narrowing our focus to person re-identification in thermal imagery. Many re-identification approaches rely upon accurate video tracking, and therefore (Maggio and Cavallaro, 2011) was a good place to start to gain an understanding of the fundamentals of tracking, feature extraction, target representation and the methods that are commonly used to achieve this. Moving on to re-identification specifically, the work in (Gong et al., 2014) is an good resource for this purpose, as it first gives an overview of what re-identification is and why it is useful and challenging, and then proceeds to give an high level overview of some of the techniques that have been researched and implemented previously in order to find a reliable and efficient method to solve the re-identification problem.

The work in (Layne et al., 2014) expands on the information provided by (Gong et al., 2014) to focus on the gathering and use of attributes for re-identification, emphasising the typical methods for attribute gathering and how to evaluate the usefulness and relevance of the attributes. This does not suggest any particular method, or carry out new research, but instead acts as a good summary of the current state of the art when it comes to gathering and using attributes. Many of the attributes normally used are simple, appearance based attributes such as gender, shirt colour, the presence of a coat or bag and hair colour. However, to increase reliability some systems use more complex attributes, such as movement path, scene position and posture. Such attributes require more complicated algorithms with greater computational complexity, but do lead to an increase in reliability.

B Colour Re-Identification

Re-identification in colour is a well-researched area, particularly in terms of using attributes. Many of these methods are discussed in (Gong et al., 2014), and are fairly simple in nature and well known. Here we consider an important part of the current state of the art research is network layout and topology, and this is evidenced in (Martinel et al., 2016). Here, the technique of Distance Vector Routing is employed to get an idea of the relative locations of the cameras, enabling the system to prioritise the people seen most recently by the closest camera, as these are most likely to be correct. This is done by first analysing the overlap between cameras, and then computing distance vectors and probabilities of going from one camera to another, reducing the time complexity of the re-identification process in the majority of cases.

The work in (Chu and Hwang, 2014) is on a similar theme to (Martinel et al., 2016), but assumes a non-overlapping camera system. Each camera has entry and exit zones from its field of view, and if a person can get from one camera's field of view to another they are directly connected. The system can then create what is referred to as a camera link model, using a temporal, spatial and appearance relation between the entry and exit zones of the cameras. These paths are obtained from training data, but the system itself learns how to recognise people by attributes, and uses the training data to estimate where they are most likely to have gone after leaving a camera's field of view.

The authors of (Wang et al., 2014) propose a different method for feature based identification, using a feature projection matrix to project image features of one camera to the feature space of another, to effectively eliminate the difference of feature distributions between the two cameras. The feature projection matrix is obtained through supervised learning. The proposed method aims to use a simple gradient descent algorithm to accelerate and optimise the re-identification process by compensating for the inconsistency of feature distributions captured by different cameras.

The work in (Wu et al., 2012) emphasises the importance of making good use of all images and video frames captured of a target. The system proposed here creates a gallery of images of known individuals, with more images increasing the accuracy of the system. When a gallery exists for a target, this is referred to as multi-shot re-identification, and single-shot re-identification when only one image is available in both the query and the gallery. For multi-shot re-identification, the authors propose to use geometric distance in another way by collaboratively approximating the query sets using all galleries, a method known as Collaborative Sparse Approximation.

C Thermal Tracking

The work in (Padole and Alexandre, 2010) details a motion tracking system that uses only the thermal space. Developing an effective and efficient method for this is becoming more and more important as the uses of thermal imagery increase. This system uses the Wigner Distribution, which includes both time and frequency, to enable it to create an association between spatial and temporal data. This helps to alleviate one of the major problems, with tracking in thermal; that temperature is not uniform across the body. Using this in place of feature-based tracking reduces the complexity of the system and therefore allows it to perform accurate tracking faster in real time. In (Coutts et al., 2014), another system based on the relationship between spatial and temporal signatures in thermal is proposed, combining a human detection algorithm with the real-time compressive tracking algorithm. This algorithm uses bounding boxes, which surround an object to be tracked in subsequent frames. We can associate each bounding box with a set of features, both spatial and temporal in nature, and rematch this box with the target in subsequent frames. This approach gives promising results, but requires further research.

Thermal tracking and identification also has applications in robotics, as stated in (Ciric et al., 2013). This is an additional challenge due to robot movement, so it will not have a consistent view of the environment. Therefore, thermal is useful to allow detection irrespective of the light level. This movement based consideration can be seen as relevant to our problem of multiple cameras, as each of these will be trying to detect the same target from a different position. This paper also discusses the issue of overlapping signatures of multiple people standing close together, and suggests the 'fuzzy logic' approach to

solving it, which uses mathematical functions to determine how strongly an object is in the foreground or background. This approach is shown to have good results.

The work in (Koenigs and Schulz, 2013) agrees with (Ciric et al., 2013) in many respects, as it is also a person tracking and re-identification system for robotics, based on movement through a changing environment. The thermal imaging system filters out the background elements, and then uses a set of heuristics to determine if a hot-spot is a person or not. Experimentation shows that while it performs well in indoor environments, it can sometimes be confused in outdoor environments with high levels of clutter. The paper concludes suggesting a combined system of colour and thermal imaging.

The authors of (Kwak et al., 2015) present a novel algorithm for tracking in thermal space, using random ferns learning to approximate the location of a person on a frame-by-frame basis. This method uses background subtraction without the need for pre-trained detectors, along with association based tracking, a fully automatic process which associates humans with identified tracks based on motion and appearance cues. Thermal is used in this system in order to resolve any problems with the illumination level. This was shown to be an effective method of re-identification.

D Multiple Source Systems

The re-identification method proposed in (Mogelmose et al., 2013) suggests using a mixture of colour, thermal and depth based imaging to improve accuracy of re-identification. The colour channel provides a colour histogram per body part, the depth channel provides a skeletal outline, and the thermal channel provides local structural information and difference from background. Combining the three of these gives a tri-modal dataset of attributes. The system must then determine if a person is new or if they have been seen before. Nearly all new persons were identified correctly as new, but some previously seen were misidentified as new due to significant changes in one or more of the spectra.

E More Complex Attributes

Papers (Kundegorski and Breckon, 2015) and (Kundegorski and Breckon, 2014) propose the use of photogrammetric methods to accomplish 3D localisation and tracking in (Kundegorski and Breckon, 2014) and posture estimation in (Kundegorski and Breckon, 2015). One of the major problems that (Kundegorski and Breckon, 2014) addresses is that in thermal space, bringing a 3D environment down to a 2D representation can lead to ambiguity due to overlapping thermal signatures. This is where the photogrammetry comes in, determining the 3D position using camera projection and target size. A Kalman filter is then applied to track the 3D scene trajectories of the targets. The results of this investigation show that it is an effective improvement to 3D localisation in thermal space. The assumption of human height as a constant introduces a statistical error, but this is small and is reliable in the majority of cases.

However, an issue with the approach presented in (Kundegorski and Breckon, 2014) is the assumption of posture, or that a person is always standing at full height within a scene. This would be a particularly problematic when tracking a non-cooperative target who may deliberately vary their posture to avoid detection. The work in (Kundegorski and Breckon, 2015) presents a method to address this problem. The posture of a target detected in a scene is estimated as a percentage of full height based on the use of a Histogram of Oriented Gradient feature descriptor, gathered from each detected target, and aided by Support Vector Machine based machine learning regression. Like (Kundegorski and Breckon, 2014), this system makes use of the key advantage of thermal imagery to facilitate robust target localisation. The effect of this work makes the system from (Kundegorski and Breckon, 2014) more robust and enables it to function in situations where targets are attempting to avoid detection.

F Relation to this Project

As can be seen from the prior work discussed above, the research that has been put into re-identification in the past has been focused on colour, and the state of the art is moving away from the simpler features

such as shirt colour and towards more advanced features to try and improve the accuracy of this system. However, in thermal imagery the majority of the research presented here is concerned with tracking a person or object, and not with re-identification across multiple cameras. This is where this project fits in, as we are tackling a problem that has not been effectively solved, and using some features that have rarely, if at all, been used for re-identification purposes. As our interest is primarily centred on the features we are using, we will implement fairly basic and well understood techniques for person tracking, and then implement our set of features on top of this, as will be discussed in the next section.

III SOLUTION

Having established the problem that we want to solve, we will now break down the most important elements of our solution.

A Implementation

Here we give details of the structure of the implementation and the tools used in its development.

A.1 Implementation Structure

The implementation will begin by opening each video file, or live camera feed, and concurrently running the real time target detection code on these. Each time this code identifies a person, it compares the features of this person to those of the other people that have been seen previously, and if they are deemed to be sufficiently similar to one of these people, then they are re-identified as the same person, else the system creates a new person object. Each of these person objects has an associated Kalman Feature position estimator and a set of features, and these are used to facilitate the comparison between targets, and are updated each time the target is successfully identified. This continues frame by frame until the video file or camera feed ends.

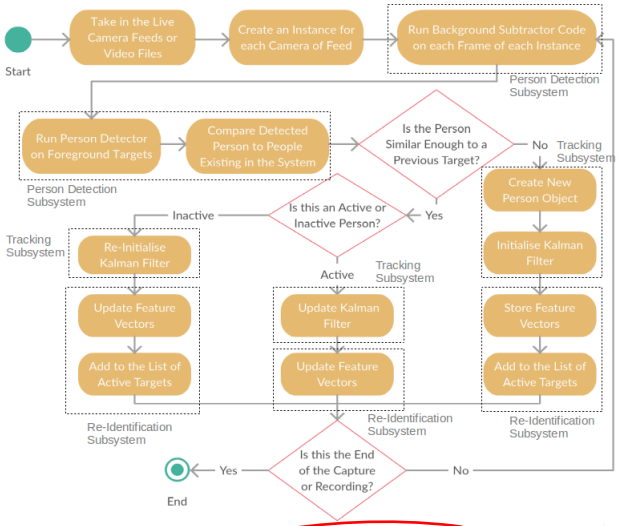

Figure 2: Process Followed by Implementation

A.2 Tools Used

Many of the computer vision techniques that will be described in the next section are complex to implement. Therefore, the OpenCV library (Bradski, 2000) has been used to allow us to use stable, well tested

code, giving confidence in the results obtained while enabling us to focus on the research and implementation of the features, which is the main purpose of this project. This library provides all of the the image processing and machine learning techniques to facilitate the first stage of this project. The implementation will be developed in C++ as it is the most suitable for the performance requirements and has a well supported implementation of OpenCV functions. A text editor will be used for code development and CMake will be used for building and deployment of the system, while GitHub has been used for version control.

B Real Time Person Detection

Before we can start concerning ourselves with feature based person re-identification, we must be able to identify whether a person is present in the image at all. This requires several stages, background subtraction, person identification and person tracking with position prediction.

B.1 Background Subtraction

The first stage of this process is background subtraction from the static camera viewpoint. We use the Mixture of Gaussians (MoG) technique to facilitate this, taking inspiration from (Zivkovic, 2004; Zivkovic and Van Der Heijden, 2005). This technique works by 'learning' a background model, modelling each of the pixels by multiple Gaussians. Using a Gaussian over the last N frames, where N is given by a parameter specifying the rate at which the background model is updated, is far more memory efficient than storing all of the pixel values across the entire video capture. This update rate is determined by the trade off between being fast enough to absorb objects that have become stationary into the background, and being slow enough to allow the detection of slow moving objects.

As the program runs through the video, during each new frame, a Gaussian for each pixel is evaluated using a simple heuristic to determine which is most likely to correspond to the background model. Pixels that do not match closely enough with these background Gaussians are classified as foreground elements, and added to a new image in the code. Once these foreground pixels are identified and built into a foreground mask, erosion and dilation are used to clean up these results. From here, we use contour detection to find the connected components, and draw bounding boxes around these contours. Figure 3 (a) shows the background model that the system has built up, the foreground of the image, the target identified and the bounding box associated with this target, taken from one frame of an example video used for testing.

B.2 Person Identification

Having identified the foreground objects, we must now determine whether they are people. This implementation has two options for this, the first of these is using Histogram of Oriented Gradients (HOG), discussed in (Dalal and Triggs, 2005). This method works by performing edge detection on the bounding box identified by the background subtraction, and calculating the gradient and magnitude of each of these edges. From here cell histograms are computed, with each of the histogram entries filled by gradient magnitudes. These histograms are then used to create overlapping block histograms of the adjacent cells. These block histograms are then combined to give a HOG descriptor, a high dimensional vector. This HOG descriptor is then passed to a pre-trained Machine Learning algorithm, in this case a Support Vector Machine (SVM). If this comes up with a positive identification, then it is classified as a person.

An alternative to HOG is the use of a Haar Cascade detector, proposed in (Rainer Lienhart and Jochen Maydt, 2002; Viola and Jones, 2001). This method works by combining successively more complex classifiers in a cascade structure. This focuses the attention of the detector on the promising reasons of the image. It is trained on a set of very simple features, which are much faster to evaluate than the HOG descriptors. The final classification is done using a combination of these classifiers in a cascade structure, beginning with those less accurate but faster classifiers to discard clearly negative areas.

The reason for the inclusion of both methods here is that there is a trade off between them. HOG is a slightly more accurate identifier, but it is slow, even within the confines of the bounding box identified. The Haar Cascade is much faster, but gets a slightly inferior classification rate. For this reason, the option exists to change between the two systems, depending on whether speed or accuracy is important. A real world example of this trade-off would be in a surveillance system, if the operator is looking to Re-Identify people from live video input they would use the Haar Cascade, but if they waned to Re-Identify people from pre-recorded video footage, they would use HOG. The output of both systems looks much the same, and is displayed in Figure 3 (b), where the red box represents the search area given by the background subtractor with additional padding, and the green box represents the identified human target from HOG or Haar Cascade.

B.3 Person Tracking and Position Prediction

Once the target has been identified as a person, we need to track each identified person. This is done by creating a new person object for each newly identified target, each with an associated Kalman Filter, explained in (Bishop and Welch, 2001). A Kalman Filter is a commonly used position estimator in such tracking problems, and is important because it makes use of current information about a target to reduce the search space required in subsequent frames.

The Kalman Filter that has been implemented here is 6-dimensional, storing the coordinates of the centre of the target, its velocity in each direction and the width and height of the associated bounding box. This is updated each time the target is detected and the implementation then calculates a predicted position for the target in the next frame of the video. This prediction is based on a transition matrix, which in this implementation keeps the velocity and bounding box dimensions the same as in the current state, and changes the position in both x and y by adding the current velocity, as this is how much its position should change in the next frame if the velocity remains constant.

With such a Kalman Filter associated with each person object, we know where each person is positioned within the distributed camera network. This can inform us as to which person any identified target is likely to be, due to their most recent position, and can speed up the comparison, as we know people cannot be in two places at once. It also helps us to reduce noise from the person detector, as we have a predicted size of the next classification of each person. Figure 3 (c) shows the output of this, with the red box showing the area identified as a foreground target, and the blue box is the predicted position of the target in the next frame. The green box from the person identification phase is omitted here to reduce the clutter in the image, but the process is still being performed to give us the measurement of the target's position in each frame.

(a) MoG Background Subtraction Output

(b) Person Detection Output

(c) Kalman Filter Output

Figure 3: Real Time Person Detection System Stages

C Features for Re-Identification

To properly associate and compare person objects, we must now consider the features of each person that we have chosen to implement in this Project.

C.1 Hu Moments

This feature enables us to describe, characterise, and quantify the shape of a target in a manner that is position, scale and rotation invariant. The method for doing this is included in the OpenCV library, inspired by (Hu, 1962), and we have taken inspiration from (Flusser et al., 2009; Dawson-Howe, 2014) for its implementation in our system. When calculating this, we first detect the contours present in our region of interest, use the largest of these contours to calculate Moments, and then use these Moments to calculate Hu Moments. This gives us a feature vector of length 7, which is normalised to give us a feature vector to represent the shape of the object.

C.2 Histogram of Thermal Intensities

The next feature that we have used is the histogram of thermal intensities. This has been done by first using edge detection methods to get just the pixels that represent the target from within the Kalman bounding box. From here, the intensity of each pixel that makes up the target is recorded, and a histogram is constructed of all pixels within certain bounds. Each frame, it builds up a histogram for the new target, and if it is sufficiently close to an existing target, then this feature will be satisfied. This is a useful feature as it can give a fair impression of clothing, as if one target is wearing a T-Shirt, for example, there will be more high intensity pixels than will be present for a target wearing a long sleeved shirt or jacket.

C.3 Histogram of Oriented Gradients

We have already discussed the process of Histogram of Oriented Gradients (HOG) when describing Person Identification in the previous section of this document. When being used for that purpose, the OpenCV function implementing the work of (Dalal and Triggs, 2005) compares the HOG descriptor to a pre-trained support vector machine to determine if the target is within acceptable bounds of being a person. Here, we use the same feature vector, but instead of passing this to the support vector machine for classification as a human or not, we save this vector for each target, and use this as a feature of comparison for re-identification. This feature vector is very large, meaning the the processing time is impractically slow, so we have had to simplify it to make it run in reasonable time.

C.4 Thermal Correlogram

For this feature, we draw inspiration from (Jing Huang et al., 1994; Zhang et al., 2015), which suggest the use of a colour correlogram. This method gives an idea of spatial correlation of colours, using a table indexed by colour pairs, where the $k-th$ entry for $< i, j >$ specifies the probability of finding a pixel of colour j at a distance$k$ from a pixel of colour i in the image. We will apply a similar principle in our implementation, replacing the colour values with thermal intensity values. Due to limitations of processing time, we cannot store the probability of each distance value for each intensity pair, so we have placed the distances into quite wide distance thresholds, creating a histogram of distances per intensity value pair, as we understand that while these thresholds are quite wide, they still give a sufficient picture of where the "hotspots" are on the target, and how large they are. Even with this simplification, this feature works slowly.

C.5 Optical Flow

The purpose of this feature is to capture the movements of the target within the bounding box. Optical Flow is the distribution of apparent velocities of movement of brightness patterns in an image (Horn and Schunck, 1981). The method employed here to calculate the Optical Flow values is the tensor based method proposed in (Farneback, 2003), simply known as the Farneback algorithm, which uses quadratic polynomials to approximate the neighbourhood of each pixel in each frame, and then estimates the change between frames based on knowledge of how these polynomials vary under translation. As suggested from this description of the algorithm, and unlike the other features we have implemented,

this feature cannot be measured over a single frame, but must instead be calculated over the course of multiple frames. As mentioned when discussing the Correlogram, this feature vector is very large, and therefore takes a lot of processing time, making this not practical in real time.

To investigate this feature we use both the raw flow values that we have calculated, reduced to take every eight row and column as suggested in (Dawson-Howe, 2014), and we also create a histogram of these values to enable us to better use all of the information available to us. These two approaches will be compared in the results section

D Comparison Method - Mahalanobis Distance

The method that we have employed for comparing these features and facilitating re-identification is primarily of our own design, and relies on several recognised techniques for the comparison of feature vectors. This method has several stages, the first of which being the special case for the first person. As we have nothing to compare the first identified target to in the system, we immediately create a new person object for this target. We also assume that the next frame captured by the camera is this same first person, as the way our dataset has been recorded means that this will always be the case across all four cameras. Now that we have two feature vectors associated with this person, we can then calculate the mean vector and covariance matrix, which gives us an idea of the average value of each element of the feature vector, and how much each element is likely to change. Now that we have the mean and covariance, we can compare these to each new target identified by the person detector, and calculate the mahalanobis distance.

Mahalanobis distance, originally proposed in (Mahalanobis, 1936), is a measure of the distance between one single piece of data and a distribution of this data, and relies on the mean and covariance of the distribution as has been discussed. The use of the mahalanobis distance gives an idea of how many multidimensional standard deviations the new data value is from the mean of the distribution. This multidimensional standard deviation is informed by the covariance matrix discussed above. Based on our evaluation of the features, we then determine a mahalanobis distance threshold, where the target is re-identified as one of the people already seen by the system if the mahalanobis distance is below the threshold, else a new person object is created. As anomalies are present in our dataset, we make this re-identification threshold quite wide, but then have a much tighter learning threshold. Only instances with mahalanobis distances less than this will be added to the dataset to avoid corruption of each person's data.

When this new person object is created, we only have one data value associated with it, meaning that we cannot calculate a covariance matrix, and therefore cannot calculate the mahalanobis distance. As all of the people have similar covariance matrices, until we get a second classification for this person we will use the covariance matrix from one of the other people in the system as an approximation. After a second classification for this person, we can then calculate a covariance matrix for this person and use this for all future comparisons. Now that multiple targets exist, the one with the lowest mahalanobis distance will be chosen, rather than the first one to be below the threshold to try and ensure the best possible accuracy.

IV RESULTS

A The Dataset

The data used to evaluate the features that have been implemented here was gathered with four cameras on the site of Durham University's Computer Science department, with the cameras and their fields of view being arranged as they are in the Figure below. We recorded three datasets, the first of which only contained three people, the second contained six people with multiple people in view at the same time and the third also contained six people, but introduced other targets such as cars.

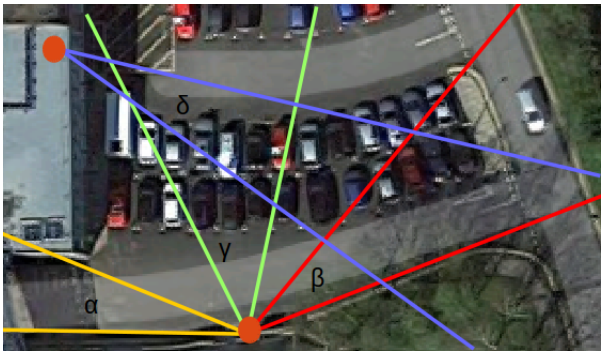

Figure 4: Setup of Cameras

B Performance of the Features

Having discussed the features that we will implement in the previous section, we will now determine which of these are sufficiently discriminative between people, and thus address our research question, **"***Which features of a human target are appropriate to facilitate re-identification in thermal video?***"**. This has been done by performing tests on one of our video files with an altered re-identification classifier to the one described above. We have a section of video taken from camera β of dataset 1 where one person is in the scene, over which the features of this person will be recorded, and each new frame will be compared against these using Mahalanobis distance as before. This person leaves the scene and a different person enters the scene. We stop recording data at this point to prevent any misclassifications corrupting the data, and just measure the mahalanobis distances between this new person and the data recorded for the previous person. A discriminative feature will be one that has a low mahalanobis distance for the right target and a high mahalanobis distance for the wrong target. These two people have fairly similar thermal signatures, and after examining our three datasets, we concluded that it was this section of video that was the best suited to the evaluation of our features. Figure 5 shows the performance of all the features from this experiment.

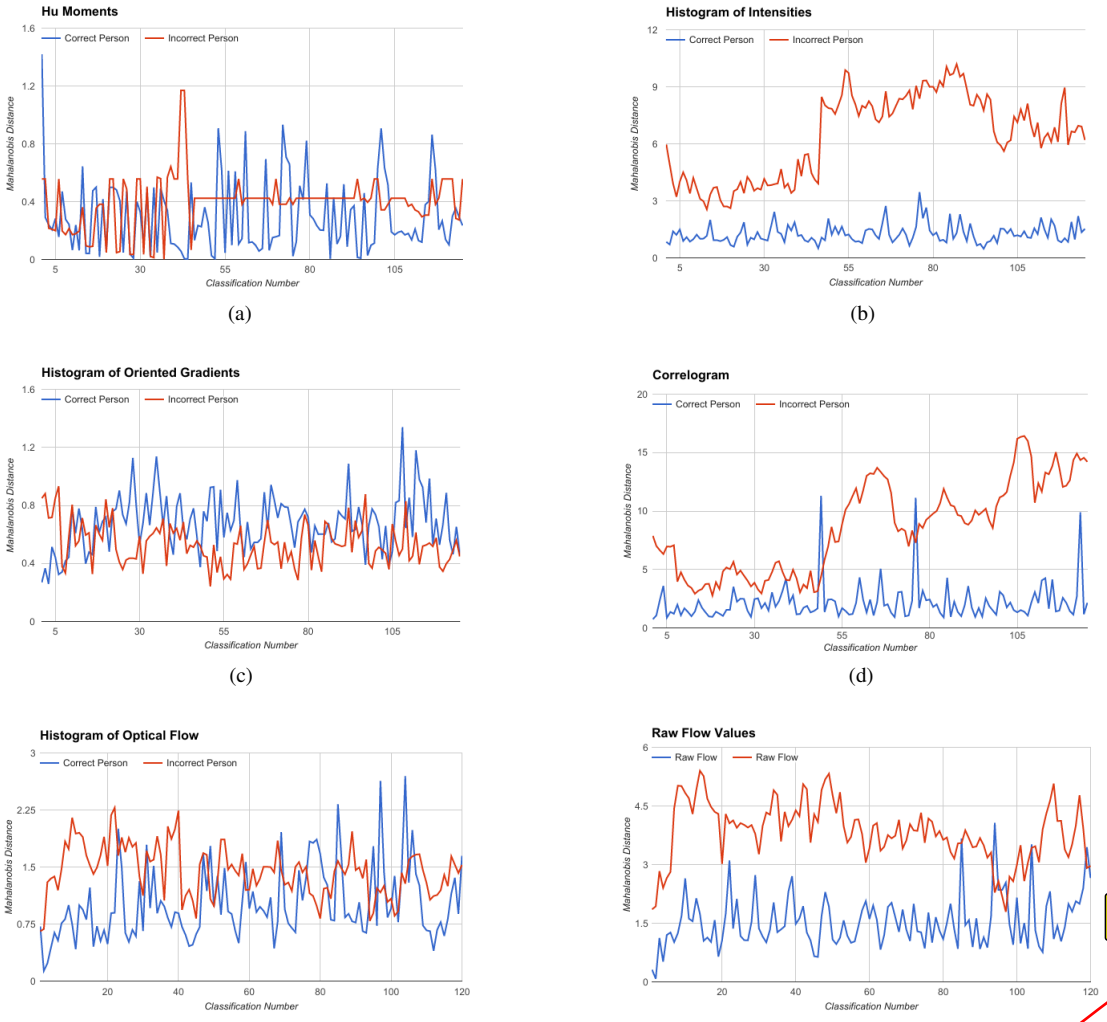

Figure 5: Performance of the Features

B.1 Hu Moments

Hu Moments proved to not be a very discriminative feature for the purpose of re-identification, as seen graph (a) of Figure 5, as the mahalanobis distances for both correct and incorrect targets are very similar. The first entry also has a comparatively large mahalanobis distance relative to the others, which we believe was due to the covariance matrix not yet allowing sufficient variation, as this feature seems to vary to quite a large extent. This feature is therefore not recommended for use in re-identification, and will not be used in the final system.

B.2 Histogram of Thermal Intensities

As can be seen in graph (b) of Figure 5, the Histogram of Intensities was shown to be a very good feature, as the line for the incorrect person is always higher than the line for the correct target, and we can therefore determine a threshold that will be correct for the vast majority of classifications. This will be a very useful feature for re-identification, and will be used in the final system.

B.3 Histogram of Oriented Gradients

The Histogram of Oriented Gradients is another feature that is not very discriminative for re-identification. All of the mahalanobis distances are very low, as shown in graph (c) of Figure 5, but there is very little difference between the correct and incorrect target. We have simplified this feature to improve runtime, but testing with more complex HOG parameters did not improve the performance of this feature enough to make it usable for the re-identification. The low mahalanobis distance for all classifications and the fact that this feature cannot effectively discriminate between different people validates its use as a generic person detector earlier in the system.

B.4 Thermal Correlogram

This feature has proved to be quite discriminative, as graph (d) of Figure 5 shows. However, there are some notable anomalies, which we believe are due to certain body parts being obstructed, such as a hand disappearing behind the persons back for a frame. Despite this, the performance of this feature is good even with the vastly simplified nature of the feature in this implementation, so we can expect that if our runtime constraints did not exist, it would perform even better, as these anomalies would probably not occur.

B.5 Optical Flow

The Histogram of Flow is slightly discriminative feature, as the correct person has a smaller average mahalanobis distance than the incorrect target, but the two lines of graph (e) of Figure 5 are close together and vary highly, meaning that it would be very hard to determine an effective re-identification threshold for using this feature, so it is therefore of very limited use to us when considering which features to use in our final solution.
However, when just using the raw Optical Flow values that we have calculated from our region of interest, we have a far more discriminative feature than when we used these values to construct a histogram, as can be observed in graph (f) of Figure 5. Even though we are only using every eighth row and column to enable this feature to run in an acceptable amount of time, these results show that it is one of our more discriminative features. If we did not have this runtime limitation, it would be reasonable to expect this feature to perform better.

C The Re-Identification System

Having evaluated the features, we now move on to the re-identification system itself. We need to decide on the best feature, or combination of features, to use with a re-identification threshold in order to get the best possible accuracy from the system. The best feature that we have is the Histogram of Intensities, but the Thermal Correlogram and Raw Optical Flow values are also discriminative features. We have therefore experimented with different combinations of these three features. The combination of all three gives good results, but runs so slowly that it would be impractical to use, as it does not run anywhere close to real time. Due to the size of the Correlogram, which runs slowly even when not combined with another feature, all of the combinations that involve this feature run slowly. All of these combinations of the discriminative features give good results, but they all introduce anomalies into the system, which would lead to the creation of new targets, which would then cause a split in the data stored for the correct person, or could be closer to the data of the incorrect person, leading to cross-pollution of the data.

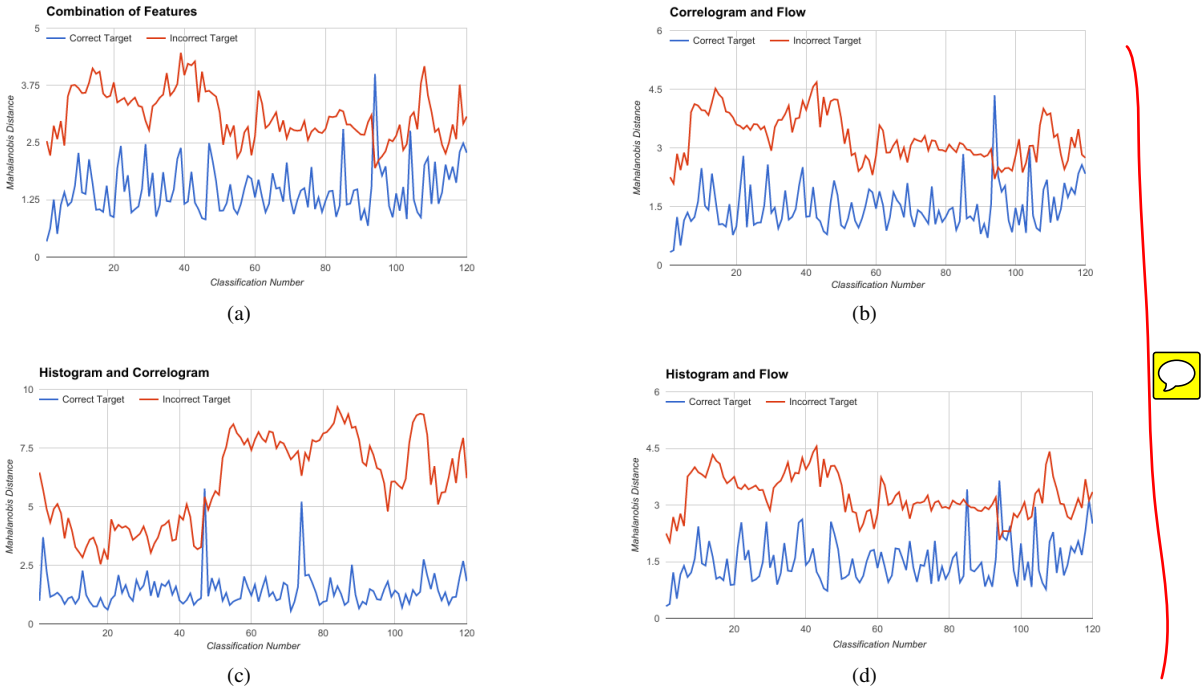

Figure 6: Combinations of Discriminative Features

In view of these anomalies and the issues that they may cause, the most discriminative feature that has the fewest anomalies and that runs in real time is using the Histogram of Intensities alone. Having decided upon using this feature, we will now implement this within the final re-identification system, using the full comparison method described previously in the Solution section, determining an effective re-identification threshold and learning threshold based on the results that we have collected for this feature. When this system is run on a single camera, the results that it gives are very good, as Figure 7 shows. Subfigures (a), (b) and (c) show different identification of different targets, while subfigure (d) shows the situation where two targets have crossed over. For the majority of this video, the system has a perfect re-identification rate for every target that it recognises as a person, until the first mistake occurs towards the end of the video. Despite this, across all our video files, the system has a long period of effective re-identification before any mistakes are made.

Figure 7: Single Camera Re-Identification

To enable the multiple-camera functionality, we will pass multiple files to the system and run each camera on a different thread, using locks to make sure that the global targets array is not corrupted, and a wait command to ensure that all cameras move through frames at the same rate, whether they have pro-

cessing to do or not. We then have a post processing stage where all the camera views are combined into one video, to allow the results to be observed more easily. However, as the multiple-camera dataset introduces very different viewpoints, people observed from different cameras have differing characteristics, which means they are sometimes similar enough to be re-identified, and sometimes not. It is also more difficult to differentiate between people from some viewpoints, with camera δ being the main source of this problem. We also rarely get a detection in camera γ, as the people are too close to the camera, and parts of them are cut off, meaning that the person detector does not pick them up. The result of these problems is many person objects loosely associated with each person, as if all of them are within the re-identification threshold, it will change often between them as individual classifications are by nature closer together than others. Figure 8 shows our setup, as well as an instance of this problem of multiple targets associated with a single person.

As the videos proceed, and more people enter the scene, we begin to get a huge amount of targets in the system, and the output becomes very hard to gain any useful insight from. Figure 9 shows a small instance of this, as this person is far away and difficult to re-identify, so a new target is created for them for every frame they are identified in. This person takes the total number of targets in the system up a huge amount, and from here the system struggles, as with so many targets, each frame is bound to be closer to a different one of them, or none of them, and will create a new target. This was from dataset 2, and at the end of the run we have over 200 targets in the system. Therefore, we can see that this system, while effective on a single camera, is ineffective over the four cameras used in this dataset. This figure also shows the problems with identifying targets at distance in the other cameras, and the lack of detections from camera γ.

(a) (b)

Figure 8: Multiple Camera Re-Identification

(a) (b)

Figure 9: Rapid Target Creation

V EVALUATION

We will now evaluate the strengths, weaknesses and limitations of the research that has been presented here.

A Real Time Person Detection System

As the primary focus of this project was to investigate the features and their viability for re-identification, we have implemented a fairly simple person detection system to give our features a platform to work on. While this performed well in the majority of cases, there were some issues associated with these methods, such as the detector cutting parts off people, giving us very different results, as if the head is not part of the classification, a significant thermal hotspot has been missed. In order to mitigate this problem, we have forced acceptable classifications to be larger. This removes many of these partial detections, although some do still occur, but it also means that we cannot identify people far from the camera. We have attempted to find a balance between these erroneous classifications and being able to classify targets at a distance, but this has not been completely successful. We also rarely get detections of people in camera γ because the edges of the people are too close to the edges of the frame, meaning that this camera is providing little to the multiple camera system due to a weakness in the person detector. This should be addressed by future work.

B Features

As the research that we have conducted here is in an area that has previously been largely unexplored, our first objective was to answer the research question *"Which features of a human target are appropriate to facilitate re-identification in thermal video?"*. This has been extensively explored in our results section, and we have effectively proved that Hu Moments and HOG are not suitable to be used for re-identification, and the Histogram of Intensities is suitable. We have also effectively established that while our more advanced features, the Thermal Correlogram and Optical Flow, are reasonably discriminative features in the form that we have implemented here, but have such a high level of complexity that they would need a more powerful computation system than the one used to perform this work if they are to be used to their full re-identification potential. We state in our aims for this project that the final system must run in real time on the video files that we are using, or at least reasonably close to real time. In order to achieve this aim we have had to greatly simplify these features, by only using every eighth row and column of the optical flow matrix and breaking the correlogram down to wide intensity and dis-

tance thresholds. However, with these simplifications these features perform worse than the Histogram of Intensities, so this is the feature we use for re-identification.

C Re-Identification System

We have developed an effective classifier to measure the effectiveness of these features, by using the Mahalanobis distance between each new frame and the distribution of previous identifications of each target. For our effective features this classifier produces a clear difference between correct and incorrect classifications in the majority of cases. The problems that our classifier encounters are due to the anomalies that the features generate. The reasons for this have already been discussed, but on the majority of correct detections from the person detection system and using the Histogram of Intensities as our feature, our classifier functions very well.

However, a major weakness of our system is that fact that miss-classifications cause major problems for the re-identification system. For example, if we have a succession of classifications for a person, stored in the associated person object, and then a miss-classification leads to the data of another person being added to this person object, then the data is no longer exclusively the features of the original person. We refer to this as cross-pollution of the data, and if it occurs then the target that has become polluted is now useless to us, as it has data from the wrong person in it, so may continue to be matched with either of the two people whose data it contains. Also, now that this person object does not contain the data for just one person, the two people that have been misidentified as this object may now be closer to other unpolluted person objects, which will lead to further cross-pollution and to the data becoming meaningless. Therefore, the re-identification system is only effective until it begins making mistakes, and these mistakes become more and more prevalent when the system is applied to multiple cameras, as was discussed in the results section, with many of the misclassifications being traceable to camera δ.

D Limitations of Equipment

A limitation that we faced in this project is associated with the infrared cameras that we used to record the data. As these are uncooled cameras, they require a thermal reset at regular intervals, to enable the cameras to calibrate themselves to correct for changes in background heat levels. While this calibration occurs, new footage is not recorded, and this leads to jumps in the video and some frames being repeated. To mitigate this, we compare every frame that comes in from the camera to the previous frame, and if they are identical then we ignore it. This stops the data per person from becoming skewed by repetitions of the same data value, but it does lead to some anomalies in the results, which can be seen as spikes in the graph, due to the sudden change in position of the person being tracked.

E Appraisal of Project Organisation

In the implementation of this system we have used a modular approach, implementing each major component separately. This was a useful process to use as it enabled us to fully test and optimise each part of the system as much as possible, and eliminate as many of the errors and limitations as we could as they came up in each subsystem. Notably, we encountered difficulties with the Kalman filter, as although this is a well understood and often used technique, it was one of the elements of the OpenCV library that changed between version 2 and 3, meaning that much of the existing work that we looked at for this part of our solution functioned in a different way to how our system would, as we have used OpenCV version 3. We also had difficulties in the implementation of the comparison system, as we initially intended to use a Bayesian Classifier in a similar manner to (Fukunaga, 1990), but this did not function correctly with our data, and much time was used trying to get this to function before we decided that using the pure Mahalanobis distances was a superior solution.

VI CONCLUSION

We will now conclude this paper, giving an overview of what we have accomplished and how this could be extended by future work.

A What has been Accomplished?

The research that has been performed in this project has enabled us to effectively answer our research question, as we have shown that Hu Moments and Histogram of Oriented Gradients are not suitable for use in re-identification, Histogram of Intensities is suitable and performs fairly well, and Thermal Correlogram and Optical Flow are moderately effective in a simplified form, and need more computation resources if they are to use their full discriminative potential. We have used this information to create a functional re-identification system using the Histogram of Intensities feature and Mahalanobis distance as a comparison method. This system has proven to be very good on a single camera system, and even on multiple cameras with similar viewpoints. A useful insight that our dataset provided was that while Histogram of Intensities is a very discriminative feature over the range of distances in the section of video that we used to measure the performance of the features, it is not as effective as a feature over more of a range of viewpoints, as shown by the issues with camera δ in Figures 8 and 9.

B Extensions

There are many more potential features that could be used for re-identification, and therefore future work and extensions of this project should explore more of these features. It would also be worth exploring the full possibilities of the advanced features that we have discussed here, the Thermal Correlogram and Optical Flow, if we had sufficient computational resources to make use of the full detail that these features can supply. If we made use of multiple cameras in the same position, or stereo cameras, we will be able to get measurements of depth, and this will most likely be a better feature than Hu Moments proved to be, which was our approximation of size. We could also consider the possibilities of using a different classifier, perhaps storing more information per each target, or incorporating the previous position of the target to improve the accuracy. We could also improve the Person Detection system to enable it to function at distance and not pick up small parts of people as different detections.

References

Bishop, G. and Welch, G. (2001). An Introduction to the Kalman Filter. *SIGGRAPH*, Course 8.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Chu, C.-T. and Hwang, J.-N. (2014). Fully Unsupervised Learning of Camera Link Models for Tracking Humans Across Nonoverlapping Cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6):979–994.

Ciric, I., Cojbasic, Z., Nikolic, V., and Antic, D. (2013). Computationally intelligent system for thermal vision people detection and tracking in robotic applications. In *2013 11th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services, TELSIKS 2013*, volume 2, pages 587–590. IEEE.

Coutts, F. K., Marshall, S., and Murray, P. (2014). Human detection and tracking through temporal feature recognition.

Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection.

Dawson-Howe, K. (2014). *A Practical Introduction to Computer Vision with OpenCV*. Wiley Publishing, 1st edition.

Farneback, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. *Lecture Notes in Computer Science*, 2749(1):363–370.

Flusser, J., Zitova, B., and Suk, T. (2009). *Moments and Moment Invariants in Pattern Recognition*. Wiley Publishing.

Fukunaga, K. (1990). Statistical Pattern Recognition. *Pattern Recognition*, 22(7):833–834.

Gong, S., Cristani, M., Loy, C. C., and Hospedales, T. M. (2014). The Re-Identification Challenge. ... *Re-Identification*, pages 1–21.

Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *Information Theory, IEEE Transactions on*, 8:179–187.

Jing Huang, Kumar, S., Mitra, M., Wei-Jing Zhu, and Zabih, R. (1994). Image Indexing using Color Correlograms. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 191(3-4):762–768.

Koenigs, A. and Schulz, D. (2013). Fast Visual People Tracking using a Feature-Based People Detector and Thermal Imaging. In *2013 IEEE International Symposium on Safety, Security, and Rescue Robotics ~~(SSRR)~~*, pages 1–6. IEEE.

Kundegorski, M. E. and Breckon, T. P. (2014). A Photogrammetric Approach for Real-time 3D Localization and Tracking of Pedestrians in Monocular Infrared Imagery. *In Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence, SPIE, pp. 1-16, 2014.*

Kundegorski, M. E. and Breckon, T. P. (2015). Posture estimation for improved photogrammetric localization of pedestrians in monocular infrared imagery. *Proceedings of SPIE - The International Society for Optical Engineering*, 9652.

Kwak, J. Y., Ko, B., and Nam, J. Y. (2015). Multi-person tracking based on body parts and online random ferns learning of thermal images. In *Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015*, pages 41–46. IEEE.

Layne, R., Hospedales, T. M., and Gong, S. (2014). Attributes-based Re-Identification. ... *Re-Identification*.

Maggio, E. and Cavallaro, A. (2011). *Video Tracking - Theory and Practice.*

Mahalanobis, P. (1936). on the Generalized Distance in Statistics.

Martinel, N., Foresti, G. L., and Micheloni, C. (2016). Person Reidentification in a Distributed Camera Network Framework.

Mogelmose, A., Bahnsen, C., Moeslund, T. B., Clapes, A., and EscalerA, S. (2013). Tri-modal person re-identification with rgb, depth and thermal features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 301–307.

Padole, C. N. and Alexandre, L. A. (2010). Wigner distribution based motion tracking of human beings using thermal Imaging. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 9–14. IEEE.

Rainer Lienhart and Jochen Maydt (2002). An Extended Set of Haar-like Features for Rapid Object Detection. *ICIP*.

Viola, P. and Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In *ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2001*.

Wang, Y., Hu, R., Liang, C., Zhang, C., and Leng, Q. (2014). Camera compensation using a feature projection matrix for person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(8):1350–1361.

Wu, Y., Minoh, M., Mukunoki, M., Li, W., and Lao, S. (2012). Collaborative sparse approximation for multiple-shot across-camera person re-identification. In *Proceedings - 2012 IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2012*, pages 209–214. IEEE.

Zhang, D., Li, Y., Xu, J., and Shen, Y. (2015). Multiple-shot People Re-identify based on Feature Selection with Sparsity. *International Journal of Hybrid Information Technology*, 8(1):27–34.

Zivkovic, Z. (2004). Improved Adaptive Gaussian Mixture Model for Background Subtraction.

Zivkovic, Z. and Van Der Heijden, F. (2005). Efficient adaptive density estimation per image pixel for the task of background subtraction.