TRACKING PEOPLE IN 3D USING POSITION, SIZE AND SHAPE

Simon Denman, Vinod Chandran, Sridha Sridharan

Image and Video Research Laboratory
Queensland University of Technology
GPO Box 2434, Brisbane 4001, Australia

Abstract

This paper presents a prototype tracking system for tracking people in enclosed indoor environments where there is a high rate of occlusions. The system uses a stereo camera for acquisition, and is capable of disambiguating occlusions using a combination of depth map analysis, a two step ellipse fitting people detection process, the use of motion models and Kalman filters and a novel fit metric, based on computationally simple object statistics. Testing shows that our fit metric outperforms commonly used position based metrics and histogram based metrics, resulting in more accurate tracking of people.

1. INTRODUCTION

The ability to track people in video sequences is becoming increasingly important, as increasing emphasis is placed on security and surveillance. Being able to automatically monitor indoor environments such as offices, shopping centres and airports has become more vital in recent times. However, often these environments are quite crowded, and so occlusions are common place and can make tracking difficult.

In this paper, we present a person tracking system designed for use in indoor environments, in which we expect a high level of clutter and occlusions. A stereo camera is used for acquisition to allow the use of depth information. We use adaptive background segmentation [1] and a coarse people detection step to reduce locate regions of interest, and apply people detection, via head detection [2, 3] and ellipse fitting [2], to locate people in the scene. The system is able to extract additional people by segmenting occlusions using depth information. Located people are matched to people tracked by combining several computationally simple object statistics to obtain a reliable metric for matching.

Results show that our metric outperforms commonly used position metrics and histogram metrics. The use of this metric, allows our system to reliably track people at close and mid range (5m - 15m) through a variety of occlusions.

2. LITERATURE REVIEW

A wide variety of tracking systems have been developed for various purposes. Most [2, 3, 4] use some form of background segmentation as a first step to tracking. Once objects have been located, a wide variety of methods are used to maintain tracking of an object. One of the more popular methods is to try to predict the motion of the object, and thus its position in the next frame [2, 3, 5]. Use of colour is also popular [6, 7], with various forms of histogram matching and colour clustering used to maintain the tracking of objects.

Haritaoglu et al [3] developed a system for tracking objects with a single gray scale camera, using the expected motion of the object to restrict the search space, and relied on the matching of silhouettes to verify the object. Fuentes [4] formed blobs, characterised by a bounding box, centroid, width and height, from the motion image to represent tracked people. Rather than simply tracking blobs Zhao et al [2] proposed a system that used an ellipsoid shape model to locate and segment people from the motion image. Lu [7] used a colour histogram to maintain the tracking of an object. The colour histogram is unaffected by pose change or motion, and so is a reliable metric for matching after occlusion.

Systems such as [5, 6] use multiple modalities to track people. Darrell et al [6] combined the use of stereo, colour and face detection to track people, while Matsumura [5] used the region position, speed, template image and an object 'state' for tracking.

3. SYSTEM OVERVIEW

The system receives stereo image pairs as input, and for each pair (or frame), people are located within the scene and matched to people located in the previous frames. A list of currently and previous tracked people is kept, storing statistics relating to size, shape, appearance and motion models for each person. Motion detection, a two step people detection process and disparity calculations are all performed as a precursor to tracking. The system is capable of tracking

multiple people through a scene, through occlusions within the scene, and is able to resume tracking of a person when they re-enter the scene.

Motion detection [1] and coarse people detection are used to reduce the amount of data within the system. Motion detection is used to locate the areas of the scene that contain the objects, and a preliminary people detection step reduces the scene regions of motion that may contain people. Each resulting region is then analysed for people.

Ellipse fitting [2] is used to locate people within each coarse region. Heads are located and ellipses are fitted at the heads [2]. We use a two step ellipse fitting process. The first pass works directly on the motion image, removing the detected candidate from the motion image. The second pass reinstates the removed candidates at a fraction of their original weight, allowing objects that are partially occluded to still be detected by using some of the 'weight' from the overlapping object. Ellipses are fitted to the valid heads at several different aspects, starting from the most square (this allows the system to handle different body shapes and poses more effectively). A fit is achieved when the area enclosed by the ellipse meets an occupancy threshold.

Disparity is calculated for each candidate person. Sparse depth maps are used; a set of points within the interior of each object are selected and the disparity is calculated for these points. Outliers are removed and the mean of the remaining points is taken as the disparity of the object. The sparse map is translated back into a full size object image to allow segmentation in later stages.

Located people are tracked from frame to frame, by observing their position, size and shape. These statistics are combined into a metric that measures the likelihood of a match between an existing tracked person and one that has been located in the current frame. In circumstances where an obvious fitting person cannot be found, additional criteria such as the colour of the person is used. Tracked people that cannot be matched for a frame are assumed to be occluded, and their position is estimated according to a motion model and Kalman filter. Located people that could not be matched are either matched to people that were previously present within the system (lost), or used to create a new person for tracking.

4. PERSON MATCHING

For each candidate person that is detected, the system will attempt to match them to a tracked person in the system. If a suitable match cannot be found, the object is compared to people that were previously tracked by the system. If there is still no match, the person is assumed to be new, and is added to the list of tracked people.

4.1. Determining A Match

To match located objects to tracked people, a criteria that describes the likelihood of a match is needed. A 'fit' criteria, that combines various aspects of the objects shape and size was developed, that results in a numerical value which can be used to determine the most likely match for each object to person. The 'fit' of a person to an object, is made up of four components, the distance from the persons expected median to the median of the object; the difference in shape and size of the of the person and the object; the difference in the expected disparity of the person and that of the object and the expected direction of travel from one frame to the next.

The distance between the expected and actual position is calculated using the manhattan distance, meaning a small value indicates a good fit.

$$Fit_{Position} = |x_{Exp} - x_{Act}| + |y_{Exp} - y_{Act}| \quad (1)$$

The difference in shape and size is itself made of up four separate components. The area, perimeter and aspect ratio of the objects bounding box, and the number of motion pixels within the object are combined to provide a measure describing the similarity in shape of the tracked and located object. Each of these describes a slightly different aspect of the objects shape, allowing us to be more certain about a shape change. Fit values are obtained by calculating the ratio of the last known value and the value of the located candidate object. The ratios are averaged and squared, ensuring that each criteria has an equal impact.

The four ratios are then combined, such that a strong shape fit will have a value very close to 1, while a poor fit will approach 0.

$$Fit_{Shape} = \left(\frac{Fit_{Area} + Fit_{Perim} + Fit_{Aspect} + Fit_{Pixels}}{4} \right)^2 \quad (2)$$

The disparity fit criteria is very simple, taking the absolute difference between the expected and actual disparities.

$$Fit_{Disparity} = |Disparity_{exp} - Disparity_{act}| \quad (3)$$

This could be incorporated into the position fit, but as typical position error can be up to 30 pixels for a correct fit and the entire range for disparity within the system is no more than 40 pixels (an error of more than 3 for a correct match is uncommon), incorporating the disparity directly into the position fit calculations would result in the disparity being 'lost' within the metric. Like the position fit, a value close to 0 indicates a good fit.

Finally a direction fit is calculated. The expected direction of movement for the three directions (x, y and disparity) is calculated from the last known position and the next predicted position. The actual direction of movement for the

tracked object to arrive at the location of the located object is then calculated. If the tracked object has moved as expected then, the two sign vectors for these directions will be equal. The error in movement is defined as the number of corresponding signs within the sign vectors that do not match. This error is then applied to the fit as follows.

$$Fit_{Direction} = 2^{DirectionError-1} \quad (4)$$

For objects that are traveling in the wrong direction, the fit is increased further to decrease the likelihood of a match. For objects that are moving in the correct direction, the fit will be halved. Having one direction in error will not result in any penalty.

For the position, shape and disparity fits, a loose limit must be met, which prevents fits being calculated for objects that cant possibly match. A similar limit cannot be applied to the direction metric as it is acceptable for an object to totally change direction. Provided the loose limits are met, the fit of the tracked object to the located object is calculated as follows:

$$Fit_{Object} = \frac{Fit_{Position}}{Fit_{Shape}} \times Fit_{Disparity} \times Fit_{Direction} \quad (5)$$

This equation will yield small values for strong fits and large values for poor fits, simplifying thresholding, as a fit of zero represents an ideal match, with increasing values indicating an increasingly poor match.

Thresholds are applied to classify the fit into one of four categories: Strong Fit; Tracked Fit; Occluded Fit; or No Fit. A strong fit indicates that the objects are very likely to match, size, position and disparity are all very similar. A tracked fit indicates that the objects possibly match, most of the criteria are close, but not exact. An occluded fit is only accepted for occluded objects, as they have been unsighted for a period of time, they are likely to have greater errors, necessitating a looser fit. If the fit value is greater than the occluded fit threshold, then there is no fit for that object to person. If multiple objects satisfy the fit criteria, then the best fitting object is taken.

One problem that can occur when using this metric is that two objects that may not be particularly close to one another can record a 'strong fit' for the same object if they are moving at the same depth. To counter this, an additional constraint is placed on the fit calculations whereby if the fit is a 'strong fit', but the position is not within half the required distance for a positional fit, the the object cannot have a 'strong fit', and the fit is set slightly above the 'strong fit' threshold.

4.2. Matching Located Objects

When matching objects, the system compares every object against every person, and assigns them in order of best fit. There are two special cases when matching objects: two or more tracked objects match a located object with a strong fit; two or more objects have a positional fits (they meet the positional criteria of the initial fit check, and potentially no others) and one full fit.

For these situations, the system will attempt to locate additional objects by segmenting the disparity map. In the first case, the system will segment the depth map into regions above and below the mean of the occluding objects expected disparity, resulting in an object for each region. For the second case, parts of the depth map within a tolerance value of the main objects disparity will be located and assigned to that object. Regions outside of this tolerance will be segmented and added to the list of possible candidates, possibly resulting in several new candidates.

4.3. Matching Lost Objects

Located people which do not match any currently tracked objects may be previously tracked people that have become lost. Being lost, their position and size is no longer valid, and so they must be compared based on texture models [3].

The problem with relying purely on textural models is that if the person re-enters at an orientation that is different to that which was used when the model was last updated, a match is unlikely to be made. To overcome this, the colour histogram of the incoming person is compared with the colour histogram of the texture model. This match score is combined with the match value from the texture map comparison to obtain an overall match. If a match is successful, the Kalman filter and motion model for the matched object are reset.

When people enter the scene, it can be difficult to compare them against lost people as their entire body may not be visible yet, so matching against a model that is built (primarily) using images of the body in full view can be difficult. To overcome this, located people are matched against lost people at two separate times; when they first appear and when they have fully entered the scene.

If a match is successful, the two people are merged into the original person, with the newer track for the person being deleted. Motion models and Kalman filters from the new track are kept, while the old tracks image models are retained with the current images incorporated.

5. RESULTS

The system was tested using data acquired in house. Data was capture from a stereo camera at between 15 and 25 frames per second, at a resolution of 320x240. Due to space restrictions, at most three people were within a single scene. As a result of these space restrictions, scenes contain several

Metric	Total	Swap	Loss	Assignment
Our Metric	0	0	0	0
Pos	11	6	3	2
Pos & Disp	5	1	3	1
PAPD	5	1	3	1
Hist	10	4	6	0
Pos & Hist	14	5	7	2

Table 1: Performance of Tracking Metrics on Occlusions

Metric	Total	Swap	Loss	Assignment
Our Metric	0	0	0	0
Pos	5	2	1	3
Pos & Disp	4	1	1	2
PAPD	2	0	1	1
Hist	3	2	1	0
Pos & Hist	4	2	1	1

Table 2: Performance of Tracking Metrics on Scenes

occlusions and people are rarely visible within the scene for more than a couple of hundred frames.

The tracking metric was tested using eighteen clips showing a variety of occlusions involving two and three people; and two longer sequences, one with a single person and one with two people, showing the people moving into and out of the scene several times. The metrics used were the metric described in this paper; position (x, y) only (Pos), position and depth (Pos & Disp); position, depth, area and perimeter (PAPD); colour histogram (Hist); and colour histogram with position (Pos & Hist).

Comparison of the metrics was based on the number of errors within the tracking for each metric. We define three classes of errors, listed from most severe to least severe:

1. Swap - Occurs when two objects swap identities
2. Loss - A tracked object cannot be matched to a candidate, and thus is lost
3. Assignment - The tracked object is matched to a candidate that has been created as result of an error in the people detection

As is shown, our metric was the only one not to produce any errors on the sample data. The simpler metrics tended to suffer more severely than the more complex ones, with the histogram metrics suffering particularly badly. Reduced versions of our own metric (such as the position and depth and position, depth, area and perimeter) performed well.

Occlusions that could be resolved into the individual people by depth map analysis proved to be the easiest for the various metrics, with all metrics handling these correctly in the majority of instances. Occlusions where people became totally obstructed, even if only for a few frames, proved more challenging. The histogram metrics were less susceptible to these problems, but performed badly in test cases where the subjects were wearing similar coloured clothing. Another deficiency that became apparent with the full scene analysis was the swapping of tracks when an object entered at the same time that another object was leaving at a similar position. The only metric besides our own that was able to handle this was the PAPD metric, a reduced version of ours.

6. CONCLUSION

We have presented a tracking system capable of tracking people at close to mid range (5m - 15m) accurately. The system is able to deal with occlusions by either segmenting the occluding objects from the depth map, using the ellipse fitting people detection to locate the people from the motion image, or by predicting the persons position until they are sighted again. We have demonstrated the ability of our matching metric, which outperforms simpler metrics of a similar style and histogram metrics, and which only uses information that can be gathered directly from the motion image and people detection. The increase in performance from position, to position and depth, PAPD and our metric suggests that use of additional simple features improves overall performance.

Future work will be done to evaluate the system at a longer range, in a larger, more crowded environment. It is anticipated that as long as the baseline of the stereo camera is sufficiently large, the system will perform well. Face detection will also be added, with the ultimate goal to be able to recognise a person within the crowd.

7. REFERENCES

[1] D Butler, S Sridharan, and V. M Bove Jr, "Real-time adaptive background segmentation," in *ICASSP '03*, 2003.

[2] Tao Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1208–1221, 2004.

[3] I.; Haritaoglu, D.; Harwood, and L.S.; Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809 – 830, 2000.

[4] L. M. Fuentes and S. A. Velastin, "Tracking people for automatic surveillance applications," in *Pattern recognition and image analysis*, F. J. Provider Oclc Perales, Ed., Puerto de Andratx, Spain, 2003, pp. 238–245, Berlin; Springer; 2003.

[5] A.; Matsumura, Y.; Iwai, and M.; Yachida, "Tracking people by using color information from omnidirectional images," in *41st SICE Annual Conference*, 2002, vol. 3, pp. 1772 – 1777.

[6] T.; Darrell, G.; Gordon, M.; Harville, and J.; Woodfill, "Integrated person tracking using stereo, color, and pattern detection," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 175–185, 2000.

[7] Wenmiao; Lu and Yap-Peng; Tan, "A color histogram based people tracking system," in *2001 IEEE International Symposium on Circuits and Systems*, 2001, vol. 2, pp. 137 – 140.