Can You Describe Him For Me? A Technique for Semantic Person Search in Video

Simon Denman, Michael Halstead, Alina Bialkowski, Clinton Fookes, Sridha Sridharan
Image and Video Laboratory, Queensland University of Technology, Brisbane, Australia
Email: {s.denman, alina.bialkowski, c.fookes, s.sridharan}@qut.edu.au, michael.halstead@connect.qut.edu.au

***Abstract*—From a law enforcement standpoint, the ability to search for a person matching a semantic description (i.e. 1.8m tall, red shirt, jeans) is highly desirable. While a significant research effort has focused on person re-detection (the task of identifying a previously observed individual in surveillance video), these techniques require descriptors to be built from existing image or video observations. As such, person re-detection techniques are not suited to situations where footage of the person of interest is not readily available, such as a witness reporting a recent crime. In this paper, we present a novel framework that is able to search for a person based on a semantic description. The proposed approach uses size and colour cues, and does not require a person detection routine to locate people in the scene, improving utility in crowded conditions. The proposed approach is demonstrated with a new database that will be made available to the research community, and we show that the proposed technique is able to correctly localise a person in a video based on a simple semantic description.**

I. Introduction

Following a crime, it is common for a description of the alleged perpetrator to be taken. In the event that the suspect is still in the vicinity, this description is circulated to nearby personnel (i.e. police) to aid in apprehending the suspect. Typically, a description incorporating traits such as height, build, skin and hair colour, as well as the clothing worn is provided. This description can be viewed as a set of soft biometrics, features that can be used to describe, but not uniquely identify an individual [1]–[3].

To date, soft biometrics have had two main uses: as a means to improve the performance of traditional biometrics systems by incorporating soft biometrics [4]–[7]; or as a way to recognise people in surveillance footage [1]–[3]. Traits such as colour [1], [2], height [1], [3], weight [2], simplified gait [3] and gender [3] have all been proposed for use with surveillance footage.

In many respects, using soft biometrics in this manner can be seen as a form a person re-detection. Many recent person re-detection approaches have focused on colour and texture features, and attempted to extract texture features which are less view dependant. Farenzena et al. [8] proposed an appearance-based method for person re-identification using symmetry-based features consisting of the overall chromatic content, the spatial arrangement of colours into stable regions (through the extraction of MSCRs [9]), and recurrent local motifs with high entropy (i.e. recurring textures). Symmetry is used to build the model through the use of weighted colour histograms computed along the symmetric axes, and by the sampling of patches to locate local motifs along the axes of symmetry; while the axes of asymmetry are used to segment the person into head, torso and legs. Bak et al. [10] proposed appearance models based on Haar-like features and dominant colour descriptors. The most invariant and discriminative signature was extracted using the AdaBoost algorithm. Bazzani et al. [11] proposed a person descriptor that incorporates a global feature, in the form of a HSV histogram, and local features, determined through epitomic analysis [12]. Schwartz et al. [13] proposed using a co-occurance matrix to extract a dense texture representation, as well as extracting edge and colour features for subjects.

A limitation of all these techniques [8]–[13] however is that they are designed for person re-detection, i.e. recognising a person that had already been observed. If the desired task is to locate a person from a description, then such techniques are ill suited. However, soft biometrics in general, such as [1]–[3], do provide a means to conduct a visual search, as they allow a person to be described by a set of features that can be searched for and matched against. Park et al. [14] proposed extracting dominant colours, height and build (determined from the silhouette aspect ratio) to represent a subject. A query could then be submitted to the system to locate a person matching a description. Vaquero et al. [15] proposed an attribute based search to locate people in surveillance imagery. Various facial features were extracted such as facial hair (beard, mustache, no facial hair), the presence of eye wear (glasses, sunglasses, no glasses) and headwear (hair, hat, bald), as well as full body features such as the colour of the torso and legs. Queries could be formulated as a combination of these features. However a limitation of both these approaches is that they require the people in the scene to be detected and modelled so that they can be matched against a query, rather than searching the images directly based on the query. While this approach is valid, it is difficult to apply to a crowded scene where person detection itself is a challenge.

A technique aimed at preventing football hooliganism was proposed by D'Angelo et al. [16], who proposed using colour to locate regions where rival supporters were gathering, allowing authorities to intervene prior to any incident. As such, the approach of [16] is designed to work in heavily crowded scenes, and this is facilitated by the use of colour to locate regions of the scene that are likely to belong to a supporter based on the known colours of the competing teams uniforms. While this approach is less constrained than those of [14], [15],

it is not designed to localise an individual, focusing instead on groups.

Given the limitations of existing approaches, we propose a technique that can search the image directly, without requiring person detection, by building an avatar from a user-provided semantic description and using this to guide the search. A search framework that generates an avatar based on a user query, and uses this to drive a search using a particle filter is proposed. Height and clothing colour (torso and legs) are incorporated into the avatar, however additional features can be easily added. The use of a particle filter to facilitate the search allows the targets to be tracked through video, and the results of the detection to be improved through successive iterations of the filter and over multiple frames. A new database consisting of 73 test cases for a wide variety of search queries captured across six cameras is presented, and is used to evaluate the proposed technique. It is shown that the proposed technique is able to detect people within a video sequence given a semantic description.

The remainder of this paper is outlined as follows: Section II presents the proposed semantic person search technique; Section III outlines the proposed database and evaluation protocol; Section IV presents an evaluation of the proposed system; and the paper is concluded in Section V.

II. Proposed Approach

The proposed approach is intended to be able to operate independent of other detection routines (i.e. person detection), so that rather than requiring all people to be located and compared to the target query, the images can be searched directly. This has the following benefits:

- The algorithm is able to execute faster, as it does not require a detection routine (such as [17] or [18]);
- The algorithm is better suited to crowded and unconstrained environments, where people may be frequently occluded, and/or detection of individuals may be difficult.

To facilitate this, a person's appearance is defined using a set of traits, each of which is categorised into a finite set. Two types of features are considered:

1) Shape/size features, that describe the size and/or shape of the target;
2) Appearance features, that describe the appearance of a region of the target.

Shape/size features are used to determine the dimensions of the search window within the image sequences. Appearance features are used to determine how well the given window matches the target region, and motion segmentation [19] is also used to aid in the detection process. The proposed approach is outlined in Figure 1.

A. Avatar Construction

In the proposed system, we use three traits: height, dominant torso colour and dominant leg colour.

The height for a target is classified as follows:

- Very short: less than 1.6m

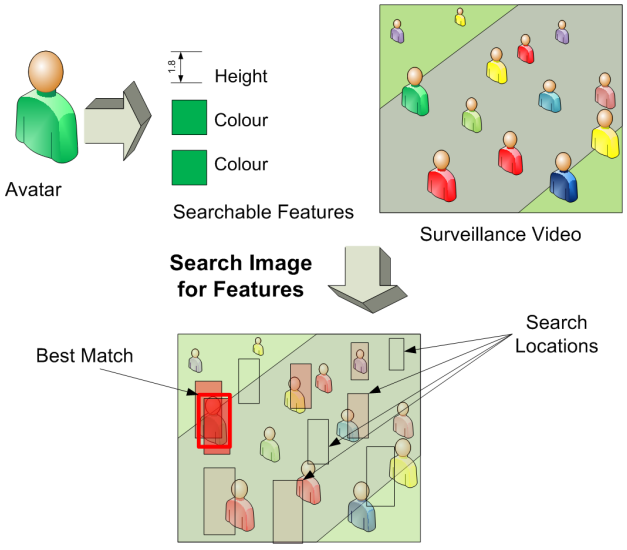


Fig. 1. Visualisation of the proposed approach - An input image or video is searched, looking for a set of simple appearance features that are determined from an avatar defined by a user. Based on the similarity of a given region to the avatar, people matching the description can be detected.

- Short: 1.5 to 1.7m
- Average: 1.65 to 1.85m
- Tall: 1.8 to 2m
- Very Tall: greater than 1.9m

Note that the height categories overlap to allow for errors in the estimation of height, either when specifying the height of the avatar, or through inaccuracies in the camera calibration.

A single dominant colour is used to represent the appearance of the torso and leg regions respectively. Rather than allow colours to be arbitrarily selected within a colour space, 'culture colours', as proposed in [16], are used to specify torso and leg colours. The 11 culture colours (black, blue, brown, green, grey, orange, pink, purple, red, yellow, white) from [16] are used. A Gaussian mixture model is trained to represent each of the culture colours. Training data is collected from surveillance footage by extracting small patches of the image which contain a single one of the culture colours. GMMs are trained in CieLAB colour space as this is found to offer the best performance in the variable lighting conditions present in the target environment. No normalisation or compensation techniques are used to cope with the variable lighting conditions, and we simply rely upon having a diverse set of training data that captures a variety of illumination conditions, as well as an appropriate colour space. From the trained models, the likelihood of a given pixel, $p(x, y)$, being a given culture colour, C, can be determined based on the colour observed at $p(x, y)$.

Given the height, torso and leg colours, the avatar is defined as shown in Figure 2. The avatar is broken into four regions vertically, which correspond to the head (H1), torso (H2), leg (H3) and feet (H4) regions. In the proposed system, the size of these regions are set to $0.25H_{avatar}$, $0.25H_{avatar}$, $0.3H_{avatar}$ and $0.2H_{avatar}$ for H1, H2, H3 and H4 respectively. Only H2 and H3 are considered when matching colours as the head and

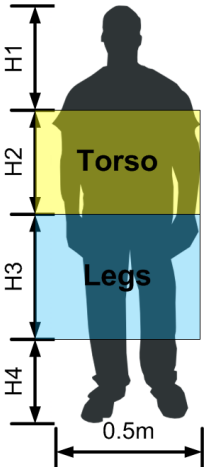


Fig. 2. The structure of the avatar - The avatar is broken down into four regions of heights H1, H2, H3 and H4. The torso colour is associated with the region H2, and the leg colour with the region H3.

feet are likely to be different colours to those specified. The person width is set to 0.5m for all subjects.

B. Searching for an Avatar

To locate an avatar in a target sequence, a condensation filter [20] is used. Particles have three dimensions:

1) the x position of the centre of the bottom edge of the bounding box in image coordinates;
2) the y position of the centre of the bottom edge of the bounding box in image coordinates;
3) the height in real-world coordinates.

The width of the person is not included as this is fixed at 0.5m for all subjects. Prior to evaluating the particle, the bounding box in image coordinates is determined using camera calibration (in the proposed approach, cameras are manually calibrated using [21]).

Given the bounding box for a particle, a match score for the torso and leg regions is determined. For the torso, the match score is determined as follows:

$$T(s_n) = \frac{\sum_{x,y \in H2(s_n)} P(im(x,y) = C_{torso}) \times K(x,y),}{\sum_{x,y \in H2(s_n)} K(x,y)} \quad (1)$$

where $T(s_n)$ is the torso match for particle s_n; $P(im(x,y) = C_{torso})$ is the probability that the input pixel, $im(x,y)$ is the target colour, C_{torso}; $H2(s_n)$ is the torso region of the particle, s_n (see Section II-A); and $K(x,y)$ is a weight set according the motion state $M(x,y)$. If motion is present at x,y, $K(x,y) = 1$; otherwise $K(x,y) = 0.5$. The weighting of individual pixel scores according to the presence of motion is intended to help prevent particles being matched to background regions of the scene. However, to avoid poor motion segmentation resulting in an inability to detect an object, the lower bound of $K(x,y) = 0.5$ is used.

The match score for the leg region is calculated similarly,

$$L(s_n) \frac{\sum_{x,y \in H3(s_n)} P(im(x,y) = C_{legs}) \times K(x,y),}{\sum_{x,y \in H2(s_n)} K(x,y)} \quad (2)$$

where $L(s_n)$ is the leg match for particle s_n; C_{legs} is the target colour for the leg region; and $H3(s_n)$ is the leg region of the particle (see Section II-A).

In addition to matching the target colours, we also consider if the target region is likely to contain a person. This can be done using either motion segmentation, or object detection. Using motion segmentation, the likelihood that a person is within the target is region is defined as follows:

$$O(s_n) = \frac{\sum_{(x,y) \in R(s_n)} M(x,y)}{N}, \quad (3)$$

$$O(s_n) \geq T_O : \; P(s_n) = 1, \quad (4)$$

$$O(s_n) < T_O : \; P(s_n) = \frac{O(s_n)}{2T_C} + 0.5, \quad (5)$$

where M is a binary image indicating the presence of motion in the image; $R(s_n)$ is the image region described by the particle, s_n; $O(s_n)$ is the percentage of the bounding box defined by s_n that is in motion; and T_O is a threshold that defines the minimum amount of motion that is expected within the bounding box. $P(s_n)$ is derived from $O(s_n)$ such that an occupancy over the threshold, T_O, yields a value of 1; while a $O(s_n)$ of less than T_O results in a value of $P(s_n)$ that is linearly scaled to between 1 and 0.5. This approach ensures that errors in the motion segmentation do not result in an inability to locate the target person, while still favouring regions that are in motion.

This approach can also be used with an object detection routine. If an object detection routine is used, the maximum intersection between the bounding box described by the particle and the detected regions is used to calculate $O(s_n)$, such that,

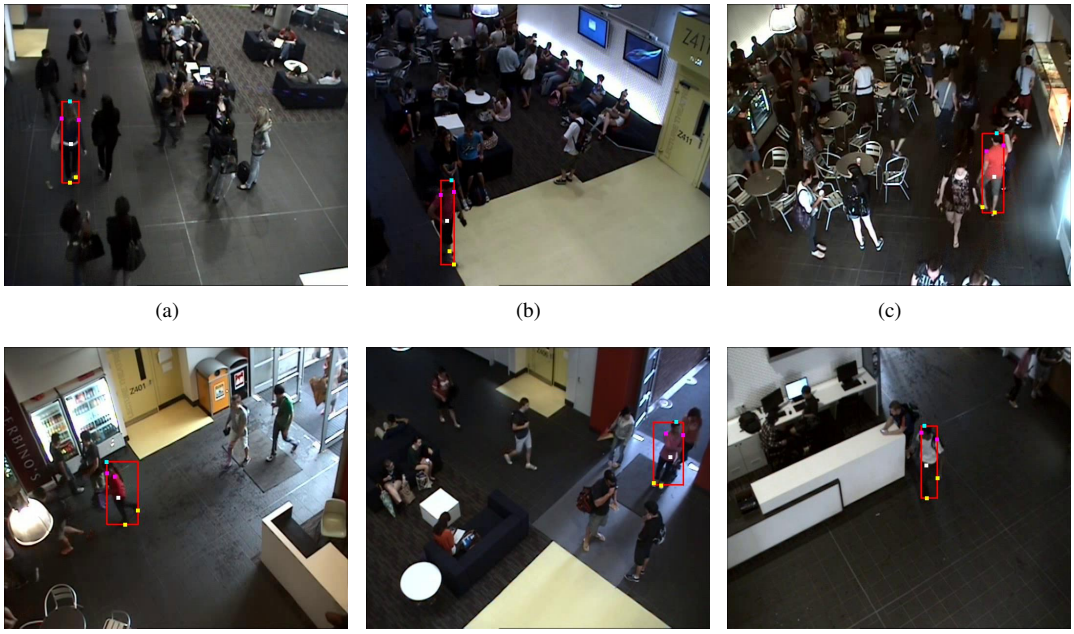
$$O(s_n) = \frac{D_{max} \cap R(s_n)}{R(s_n)}, \quad (6)$$

where D_{max} is the detected object that has the maximum overlap with $R(s_n)$. The denominator is set to $R(s_n)$ rather than the union of the two regions as object detection routines such as [18] have a tendency to return detection results that are slightly larger than the region of interest. $P(s_n)$ is then calculated from $O(s_n)$ as shown in Equations 4 and 5.

The product of these three components is taken as the final particle weight,

$$w_n = P(s_n) \times L(s_n) \times T(s_n). \quad (7)$$

Given the weighted particle set that is the output of the condensation filter, we determine the final position for the detected object. The localised position is given as the weighted average of all particles within a radius, r (set to $0.5m$ in the proposed approach), of the highest weighted particle. This approach is chosen as the distribution that is output by the condensation filter can be multi-modal, and so the weighted average of all particles may not accurately reflect the location of any one object.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Examples of the six camera views and the ground truth annotation. For each of the six cameras used, an example ground truth bounding box is shown in red. Points for the head, shoulders, waist and feet are marked in cyan, purple, white and yellow respectively. Note that across the six cameras, there are significant variations in the lighting conditions and the typical pose of people.

III. Semantic Person Search Database

To evaluate the proposed technique, a new database is proposed [1]. The proposed database consists of 73 short video clips (54-300 frames long), each taken from one of six cameras (see Figure 3) which have been manually calibrated using [21].

For each video clip, a target query is specified for a person who is known to appear in the video. The target query consists of the dominant torso and leg colours, specified as one of the 11 culture colours (see Section II-A), and the height, defined as one of the ranges listed in Section II-A. For all clips, only one person matching the target query is contained within the clip, although there may be other people that provide a partial match (i.e. have the correct torso colour, but the incorrect leg colour).

The first 30 frames of each video is reserved for learning the background model and initialising the search. Following this, 5-30 frames (depending on the length of the video clip) are annotated with the location of the person of interest. Every fifth frame in the sequence is annotated, although frames where the person is significantly occluded are omitted. The head, both shoulders, waist and both feet are annotated in each frame of ground truth. A bounding box for the person is determined based on these locations. Examples of the annotated ground truth locations are shown in Figure 3.

Using this ground truth, the localisation accuracy of the system can be measured. The localisation accuracy for a given frame is determined as follows,

$$L_t = \frac{D_t \cap GT_t}{D_t \cup GT_t}, \tag{8}$$

[1]Please contact the authors for details on obtaining the database

where D_t is the detection result at time t, GT_t is the ground truth annotation at time t and L_t is the localisation accuracy.

From Equation 8, two measures of performance are defined:

- The average localisation accuracy across the entire database,
- The number of frames in which a minimum localisation accuracy according to a threshold, T_{loc}, is achieved.

In our evaluation, we calculate both measures across each clip, as well as the entire database.

IV. Results

We present two evaluations:

1) An evaluation of the culture colour models to demonstrate their suitability for matching colours in surveillance imagery (Section IV-A).
2) An evaluation of the proposed semantic person search technique using the proposed database (Section IV-B).

A. Culture Colour Classification Evaluation

Culture colour models are trained using image patches extracted from the surveillance network used to capture the data set (the footage the colour patches are selected from is separate to that used for the database). Between 70 and 210 patches are extracted for each colour. The number of patches selected varies due to both the frequency of colours occurring (i.e. black and blue and much more common and thus there is more data for both) and the variation within the colour. 10 patches for each colour are held out of training and are used to test the models. A confusion matrix showing the performance of the trained models with this small test database is shown in Figure 4.

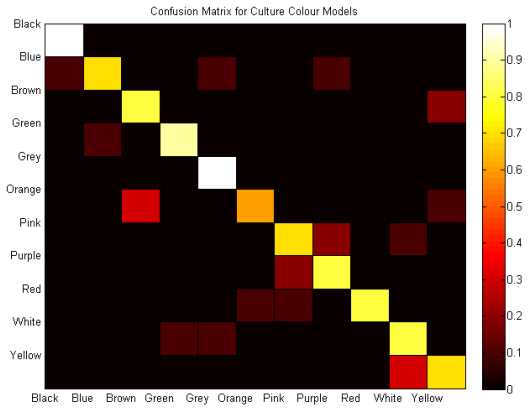


Fig. 4. Confusion matrix for classification of culture colours.

Overall, the trained models achieve a correct classification rate of 80%, with the majority of the errors made being between similar colours (i.e. confusing purple with pink, confusing red with orange and pink). These errors arise from the inherent ambiguity in classifying these colours, and this is also illustrated by the example classified images shown in Figure 5.

B. Person Search Evaluation

We evaluate the proposed technique on the database presented in Section III. We evaluate four different systems:

1) Using motion segmentation to determine if a person is present, and re-initialising the particle set in each frame (i.e. re-detect in every frame).
2) Using motion segmentation to determine if a person is present, and retaining the particle set for successive frames (i.e. tracking).
3) Using person detection to determine if a person is present, and re-initialising the particle set each frame.
4) Using person detection to determine if a person is present, and retaining the particle set for successive frames.

For the two person detection systems, we use a histogram of orientated gradients detector [18] to locate 'head and shoulders' regions. This is used instead of a full body detector as it performs better in the cluttered environment which the system is evaluated in. As a 'head and shoulders' detector is used, the overlap between the detected regions and the $H1$ region of the particles' bounding box (see Section II-A is considered (rather than the entire bounding box described by the particle) when evaluating Equation 6. T_O is set to 0.3 for the motion segmentation based systems (i.e. 30% of the region should be in motion for a person to be present), and 1.0 for the person detection based systems (i.e. a head and shoulders region should be present).

Average results over the entire database for the four configurations are shown in Table I. From Table I, it can be seen that using motion detection is preferable to relying on object detection, and that, as expected, performance improves when we track objects rather than re-detect them in every frame. Sample output from the system is shown in Figure 6 (all output shown is for the first configuration, motion segmentation and single frame detection). It can be seen that the system is able to detect a person in a variety of lighting and scene conditions. The system is able to cope with a crowded scene, however the person does need to be mostly un-occluded to be detected properly.

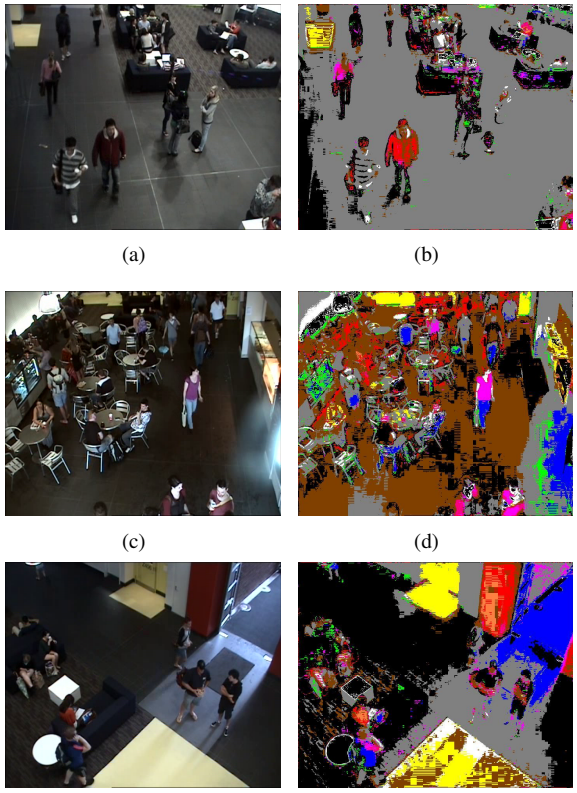
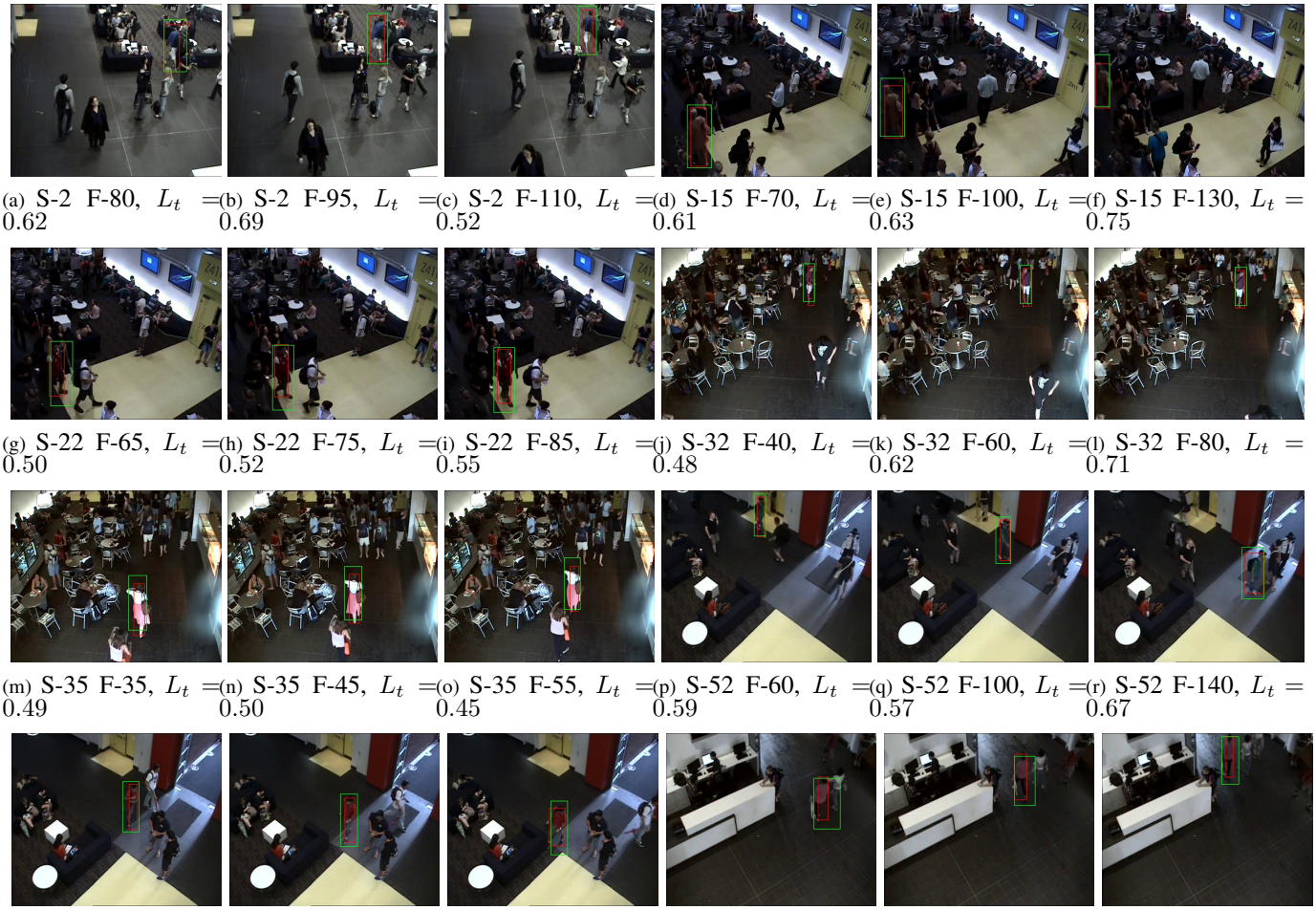


Fig. 5. Sample classification output using the culture colour models. The best matching culture colour for each pixel in the input image is selected. The ambiguity of the classification is clearly evident in regions such as the red jacket in (b), which is classified as both red and brown; the pink shirt in (b) which is classified as pink and purple; and the red pillar in (f), which is classified as red, orange and brown. It can also be seen that the floor, which is a common surface through all three cameras is variously classified as black, brown and grey depending on the ambient lighting.

Configuration	Average L_t	% $L_t \geq 0.2$	% $L_t \geq 0.6$
Mo-Seg, Single Frame	0.30	0.58	0.12
Mo-Seg, Track	0.31	0.59	0.13
P-Det, Single Frame	0.20	0.38	0.09
P-Det, Track	0.20	0.39	0.08

TABLE I
PERFORMANCE OF THE PROPOSED SYSTEM WHEN SEARCHING FOR A TARGET PERSON. 'MO-SEG' DENOTES MOTION SEGMENTATION, AND 'P-DET' DENOTES OBJECT DETECTION. EACH CONFIGURATION IS RUN FIVE TIMES AND THE AVERAGE RESULTS ARE SHOWN. ALL CONFIGURATIONS USE 500 PARTICLES, AND THREE ITERATIONS OF THE CONDENSATION FILTER ARE USED EACH FRAME.



(a) S-2 F-80, L_t = 0.62 (b) S-2 F-95, L_t = 0.69 (c) S-2 F-110, L_t = 0.52 (d) S-15 F-70, L_t = 0.61 (e) S-15 F-100, L_t = 0.63 (f) S-15 F-130, L_t = 0.75

(g) S-22 F-65, L_t = 0.50 (h) S-22 F-75, L_t = 0.52 (i) S-22 F-85, L_t = 0.55 (j) S-32 F-40, L_t = 0.48 (k) S-32 F-60, L_t = 0.62 (l) S-32 F-80, L_t = 0.71

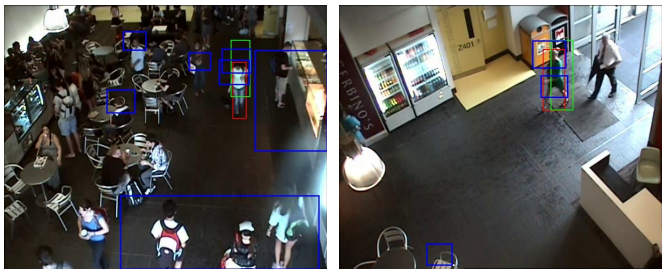
(m) S-35 F-35, L_t = 0.49 (n) S-35 F-45, L_t = 0.50 (o) S-35 F-55, L_t = 0.45 (p) S-52 F-60, L_t = 0.59 (q) S-52 F-100, L_t = 0.57 (r) S-52 F-140, L_t = 0.67

(s) S-60 F-100, L_t = 0.46 (t) S-60 F-120, L_t = 0.50 (u) S-60 F-140, L_t = 0.57 (v) S-69 F-50, L_t = 0.33 (w) S-69 F-70, L_t = 0.52 (x) S-69 F-90, L_t = 0.41

Fig. 6. Example output for the proposed system. The red bounding box indicates the ground truth localisation, and the green bounding box is the localisation result returned by the proposed approach. The test sequence (denoted by S), frame number (denoted by F) and localisation score for each image are shown. The search queries for the sequences shown are: *Average Height, Blue Torso, White Legs* for sequence 2; *Short, Brown Torso, Brown Legs* for sequence 15; *Average Height, Red Torso, Grey Legs* for sequence 22; *Short, Purple Torso, White Legs* for sequence 32; *Short, White Torso, Pink Legs* for sequence 35; *Short, Green Torso, Blue Legs* for sequence 2; *Short, Red Torso, Grey Legs* for sequence 60; and *Average Height, Pink Torso, Black Legs* for sequence 68.

Average performance over each sequence for both the motion segmentation (MS) and person detection (PD) configurations (both detecting over a single frame) is shown in Figure 7. It can be seen that in the majority of sequences, the MS configuration outperforms the PD configuration. The PD configuration is hindered by the poor performance of the person detection, which is prone to both false and missed detections (see Figure 8 for examples of the detection output). The large amount of clutter present in the scene as well as the highly variable pose of the people presents additional challenges for the person detection which contributes to the poor performance. The PD configuration also has one further disadvantage over its motion segmentation counterpart, in that it runs at 1.2 fps (for 500 particles), compared to 4.1 fps when motion segmentation is used (for a single threaded implementation running on an Intel Xeon E5-2600).

The proposed approach does however struggle in some situations, as shown in Figure 9. Errors in detection are typically caused by one or more of the following:



(a) (b)

Fig. 8. Performance of the 'head and shoulders' detection. Detected regions are shown in blue. The ground truth locations are shown in red, and the detected locations are shown in green. It can be seen that there is a large number of missed detections, and several false detections as well.

1) Ambiguous colours in the target subject;

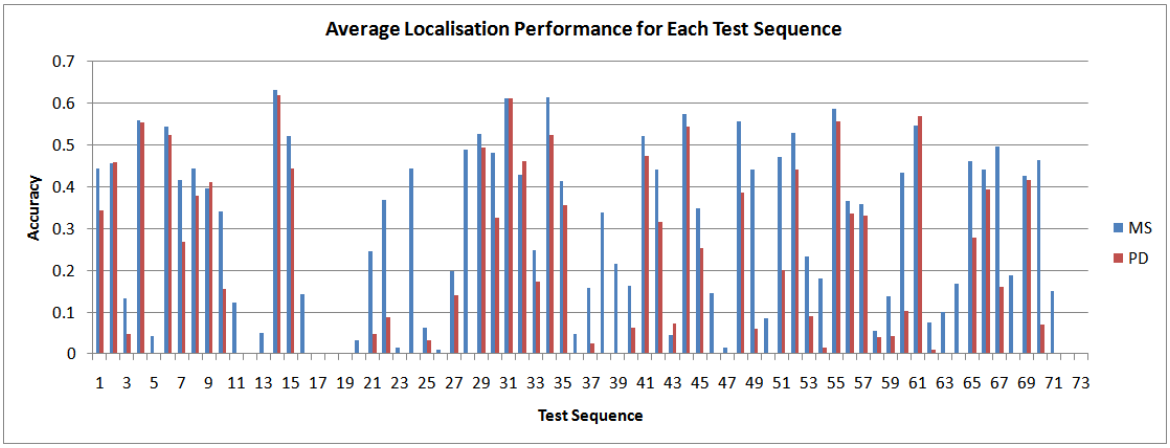
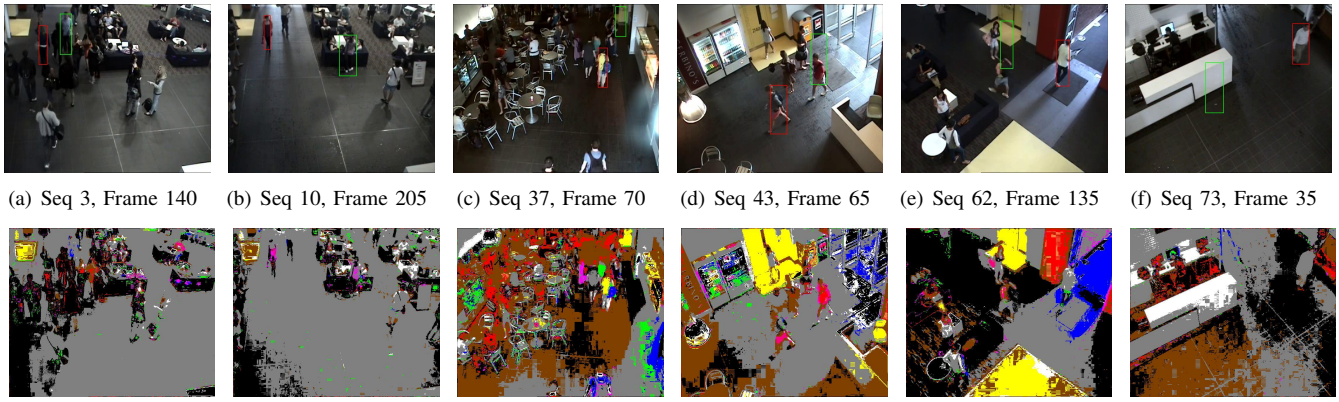


Fig. 7. The average localisation for each test sequence in the database for the MS and PD configurations. For the MS configuration, of the 73 sequences 6 sequences record a score of 0, while 45 record an average localisation greater than 0.2.



(a) Seq 3, Frame 140 (b) Seq 10, Frame 205 (c) Seq 37, Frame 70 (d) Seq 43, Frame 65 (e) Seq 62, Frame 135 (f) Seq 73, Frame 35

(g) Seq 3, Frame 140 (h) Seq 10, Frame 205 (i) Seq 37, Frame 70 (j) Seq 43, Frame 65 (k) Seq 62, Frame 135 (l) Seq 73, Frame 35

Fig. 9. Example output for the proposed system showing errors made by the system. The red bounding box indicates the ground truth localisation, and the green bounding box is the localisation result returned by the proposed approach. The search queries for the sequences shows are *Short, Blue Torso, Black Legs* for sequence 3; *Very Short, Purple Torso, Black Legs* for sequence 10; *Short, Yellow Torso, Brown Legs* for sequence 37; *Tall, Green Torso, Pink Legs* for sequence 43; *Tall, Yellow Torso, Black Legs* for sequence 62; and *Average Height, White Torso, Brown Legs* for sequence 73.

2) The target being the same colour as large portions of the background;
3) Errors in motion segmentation or object detection.

While errors in either the motion segmentation or the person detection processes also contribute to missed detections (as shown by the difference in performance between the motion segmentation and person detection configuration in Table I), errors in these processes alone will not result in a complete detection failure without uncertainty in the colour matching as well. Examples of typical errors, and the corresponding colour classifications for each image, are shown in Figure 9.

The classification errors present in sequences 3 and 10 are caused by a similar subject with a similar appearance being present in the scene. Errors in sequences in 37, 62 and 73 are all caused by regions of the background matching the target torso and leg colours. In such sequences, when there is also a misclassification in the target region (in sequence 62, the target yellow shirt is classified as grey, while in sequence 73 both the white shirt and brown trousers are classified as grey), the proposed approach may incorrectly detect the background region as being the target. In sequence 43, the localisation errors are caused by a combination of the above mentioned two factors, the pink shirt of another person is mistaken for the pink shorts of the target, while the grey floor is mistaken for the targets shirt.

The errors associated with matches to background regions could be reduced by placing further emphasis on either the motion segmentation or object detection results when assessing particles. However while such an approach will improve performance in many of the situations shown in Figure 9, additional errors will arise in situations where the detection algorithms are performing poorly. A more appropriate option may be to add additional traits, and alter the way in which traits are combined when assessing a particle such that traits with a greater uncertainty (i.e. if the shirt colour is the same as the floor, there would be a high degree of uncertainty for

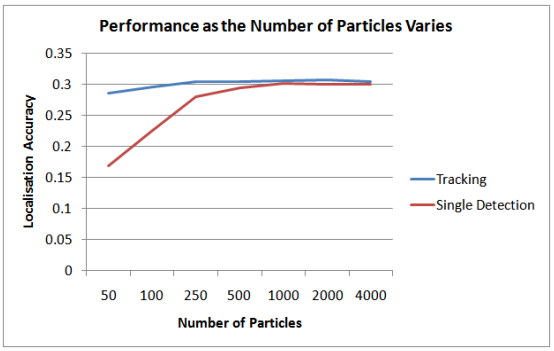


Fig. 10. Performance as the number of particles used by the condensation filter varies. For all configurations, three iterations of the filter are used each frame.

the trait) are given a lower weight. Such an approach could result in a more reliable traits (i.e. an unusual colour, or the detection results) being given a higher weight, resulting in improved localisation.

Finally, we consider the system performance as the number of particles used varies. Figure 10 shows the performance for different numbers of particles. When operating the particle filter as a tracker, performance is very consistent, even when very small numbers of particles are used. When detecting the subject each frame, performance improves sharply up to 500 particles, at which point performance plateaus and approximately matches that of the tracking variant. Notably, when only 50 particles are used, the system is capable of operating at 11.2 frames per second (as a single threaded implementation on an Intel Xeon E5-2600).

V. Conclusion

In this paper, we have presented a novel technique to search for a person in video footage given a semantic query (height, torso and leg colours). A new database for evaluating this type of algorithm has been presented, and using this database we have demonstrated that the proposed approach can effectively locate a target person. Future work will focus on incorporating additional traits such as build and ethnicity (hair and skin colour), as well as improving the way in which traits are combined when evaluating candidate locations by incorporating the uncertainty associated with a given trait. The proposed database will also be extended to include the additional traits, as well as additional test cases including multi-camera sequences.

Acknowledgment

This research forms part of the work undertaken by the project "Airports of the Future" (LP0990135) which is funded by the Australian Research Council Linkage Project scheme. More details on "Airports of the Future" and its participants can be found at www.airportsofthefuture.qut.edu.au.

References

[1] S. Denman, C. Fookes, A. Bialkowski, and S. Sridharan, "Soft-Biometrics: unconstrained authentication in a surveillance environment," *Digital Image Computing: Techniques and Applications (DICTA)*, pp. 196–203, 2009.

[2] A. Dantcheva, C. Velardo, A. DAngelo, and J.-L. Dugelay, "Bag of soft biometrics for person identification: New trends and challenges," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 739–777, 2011.

[3] Y. Ran, G. Rosenbush, and Q. Zheng, "Computational approaches for real-time extraction of soft biometrics," in *IEEE Int. Conf. On Pattern Recognition*, 2008, pp. 1–4.

[4] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *International Conference on Biometric Authentication*, Hong Kong, 2004, pp. 731–738.

[5] H. Ailisto, E. Vildjiounaite, M. Lindholm, S. Makela, and J. Peltola, "Soft biometrics–combining body weight and fat measurements with fingerprint biometrics," *Pattern Recognition Letters*, vol. 27, no. 5, pp. 325–334, Apr. 2006.

[6] G. Marcialis, F. Roli, and D. Muntoni, "Group-specific face verification using soft biometrics," *Journal of Visual Languages and Computing*, vol. 20, no. 2, pp. 101–109, Apr. 2009.

[7] K. Niinuma, P. Unsang, and A. K. Jain, "Soft biometric traits for continuous user authentication," *Information Forensics and Security, IEEE Transactions on*, vol. 5, no. 4, pp. 771–780, 2010.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2360–2367.

[9] P.-E. Forssen, "Maximally stable colour regions for recognition and matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, june 2007, pp. 1 –8.

[10] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using haar-based and dcd-based signature," in *2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems, AMMCSS 2010, in conjunction with 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS*. AVSS, 2010.

[11] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by hpe signature," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 1413–1416.

[12] N. Jojic, B. Frey, and A. Kannan, "Epitomic analysis of appearance and shape," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, oct. 2003, pp. 34 –41 vol.1.

[13] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, 2009, pp. 322–329.

[14] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita, "Vise: Visual search engine using multiple networked cameras," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3, 0-0 2006, pp. 1204 –1207.

[15] D. Vaquero, R. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *2009 Workshop on Applications of Computer Vision (WACV)*, dec. 2009, pp. 1 –8.

[16] A. D'Angelo and J.-L. Dugelay, "Color based soft biometry for hooligans detection," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 30 2010-june 2 2010, pp. 1691 – 1694.

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *International Conference on Computer Vision & Pattern Recognition*, C. Schmid, S. Soatto, and C. Tomasi, Eds., vol. 2, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005, pp. 886–893.

[19] S. Denman, C. Fookes, and S. Sridharan, "Improved simultaneous computation of motion detection and optical flow for object tracking," in *Digital Image Computing: Techniques and Applications (DICTA)*, Melbourne, Australia, 2009.

[20] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[21] R. Y. Tsai, "An efficient and accurate camera calibration technique for 3d machine vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, 1986, pp. 364–374.