Locating People in Video from Semantic Descriptions: A New Database and Approach

Michael Halstead, Simon Denman, Sridha Sridharan, Clinton Fookes
Image and Vision Laboratory, Queensland Universtiy of Technology,
Email: m.halstead, s.denman, c.fookes, s.sridharan@qut.edu.au

***Abstract*—The location of previously unseen and unregistered individuals in complex camera networks from semantic descriptions is a time consuming and often inaccurate process carried out by human operators, or security staff on the ground. To promote the development and evaluation of automated semantic description based localisation systems, we present a new, publicly available, unconstrained 110 sequence database, collected from 6 stationary cameras. Each sequence contains detailed semantic information for a single search subject who appears in the clip (gender, age, height, build, hair and skin colour, clothing type, texture and colour), and between 21 and 290 frames for each clip are annotated with the target subject location (over 11,000 frames are annotated in total).**

A novel approach for localising a person given a semantic query is also proposed and demonstrated on this database. The proposed approach incorporates clothing colour and type (for clothing worn below the waist), as well as height and build to detect people. A method to assess the quality of candidate regions, as well as a symmetry driven approach to aid in modelling clothing on the lower half of the body, is proposed within this approach. An evaluation on the proposed dataset shows that a relative improvement in localisation accuracy of up to 21% is achieved over the baseline technique.

I. Introduction

A problem in surveillance and law enforcement is the location of specific targets in complex environments from personal descriptions, such as eye witness reports. To date, this task is primarily undertaken manually by security personnel combing through hours of video footage, however recent developments in soft biometrics and content based image retrieval are leading to the development of automated approaches to locate a person from a semantic description [1], [2].

Soft-biometric traits are traits that are not individually identifiable or permanent [3], [4]. While this limitation is considerable for long term identification and verification tasks there are multiple benefits to short term search and Content Based Image Retrieval (CBIR) tasks, including the ability to obtain them from a verbal description. However, semantic traits are limited by their subjective nature, necessitating a need for weighting or fusion schemes that consider their reliability as well as other external factors (i.e. imaging conditions) when performing a search based on such a description.

Although databases exist for CBIR tasks (for instance the Corel Dataset [5] or IAPR TC-12 dataset [6]), these are at present very general and not suited to the task of locating a person in video from a semantic query. Similarly, several databases exist for person re-detection [7], [8], however due to the contrasting methods used in developing the required search models, significant further annotation is required to adapt these databases to the present task of locating people in video from a semantic description. These extra annotations arise as semantic search queries are built from global instances of required traits (i.e. blue shirt, red pants and 186cm tall). Alternatively, person re-detection relies on the extraction of image features from a supplied input to generate the required appearance model before searching subsequent images or video to re-detect this previously enrolled subject.

In this paper we introduce a new, publicly available database to develop and evaluate algorithms that locate a person in a video given a semantic query. The database consists of 110 separate video clips, each of which contains a query for a single person consisting of their height, approximate age, build, clothing colour, type and texture, and gender, as well as the location of the target person throughout the video clip. In addition, we present a new method for locating a person in video based on a semantic query. In this approach we seek to overcome some of the limitations of previous approaches [2] by considering clothing type as well as colour, and by considering the uncertainty associated with each trait. In an evaluation on the proposed database, we show that the proposed approach is able to achieve superior performance to the existing state-of-the-art [2].

The remainder of this paper is organised as follows: Section II presents an overview of prior and related work; Section III outlines the new publicly available database presented in this paper, as well as the evaluation protocol; Section IV presents our proposed system; Section V presents an evaluation of the proposed system on the new database; and Section VI concludes the paper and outlines possible future work.

II. Related Work

The problem of locating a person from a semantic description can be viewed as a specific form of CBIR. However existing CBIR approaches are ill suited to the task due to the differences in data being used. For instance, CBIR approaches are typically concerned with very broad concepts (i.e. find all pictures of cows by a lake), and make use of other meta-data such as captions to learn the relationship between the image and the underlying content [5], [9]. By contrast, when locating a person a more specific and inter-related set of concepts is used, such as 'a tall male with a plain red shirt and black trousers'. This description can be viewed as a set of soft biometrics (a trait that describes, but does not uniquely identify

an individual), each of which can only have a finite number of values (i.e. gender can only be male or female; other traits such as colour or age can be quantised into a set of categories).

The use of soft biometrics to describe and locate subjects in video footage has gained popularity in recent years, however re-detection based methods dominate research output and as such existing datasets are primarily aimed at achieving the goals of these systems. Techniques such as [3], [7], [8], [10], [11], [12], [13] use varying soft-biometric traits to create an appearance based model through some form of enrolment (typically from one or more images captured at a different time and/or from a different camera), and subsequently re-detect their target subject in the same or different cameras. While common databases such as [7], [8] contain large number of images, the structure and annotation within the database makes them ill suited to a semantic retrieval task.

Despite the focus on person re-detection, several systems have been proposed that locate a person from a semantic query. Motivated by a desire to reduce possible confrontations between rival sporting fans, D'Angelo and Dugelay [14] developed a system to detect situations where supporters of one team are located near supporters of a second, based on known colour quantities (i.e. jersey colours) within a crowded scene. While not specific to a single individual, the system did have some success in accurately gauging when two possibly hostile groups were converging on each other.

Other approaches have sought to locate an individual based on a query. Denman et al. [2] uses the three traits: torso colour, leg colour and height; to build a user defined avatar which is used to search for the person of interest using a particle filter. Similarly, Park et. al [1] established a "Visual Search Engine" which directly interacts with the operator of a multi camera network to narrow a search frame into a more manageable size. As with [2], this method utilises a combination of colour and height while also adding body build traits based on the aspect ratio of the bounding box around the subject.

As outlined above the primary benefit of these systems over re-detection is the absence of the need for individual(s) of interest to be located and 'enrolled' prior to the search. In each of the three mentioned systems [1], [2], [14], the primary trait used for location is clothing colour. In [14] and [2], colour is further categorized into a set of 'Culture Colours' [15][1] for colour definition. Vaquero et. al [16] also developed a person location program based on the dominant colour of the torso and leg which they segment into categorical values, however they primarily concentrate on the soft-biometric traits associated with the face including glasses, beards or hats. Other researchers have sought to model and detect different attributes [17] (i.e. male, female, wears glasses, t-shirt, etc.). While [17] is able to detect these attributes with reasonable accuracy, it requires decomposition of the target person into component parts which is typically difficult in surveillance imagery.

[1]The 11 'Culture Colours' are Black, Blue, Brown, Green, Grey, Orange, Pink, Purple, Red, White and Yellow

The limitations of using only a small number of traits is most evident in [2], where the undersized soft-biometric signature leads to the incorrect localisation of several subjects. Factors like cross subject classification, incorrect trait attachment (torso as legs, and background as torso) and cases where the background is confused for an individual (i.e. when background segmentation fails) all lead to poor performance and could be alleviated if additional traits were available. Furthermore, the ability to consider and compensate for ambiguous queries (i.e. searching for a grey shirt when the predominate background colour in the scene is grey) could further improve performance.

III. Semantic Person Search Database

This section presents a new, publicly available database[2] for the task of locating people in a video sequence from a semantic query. In Section III-A we outline the database and its structure, and Section III-B the evaluation protocol is presented. The database greatly extends that used in [2], by including additional subjects with greater complexity, and by incorporating more varied subject appearance such as clothing texture, age and build. Furthermore, annotation is more comprehensive, by including more detailed traits, more detailed pose information, and annotation for every frame, rather than every 5th.

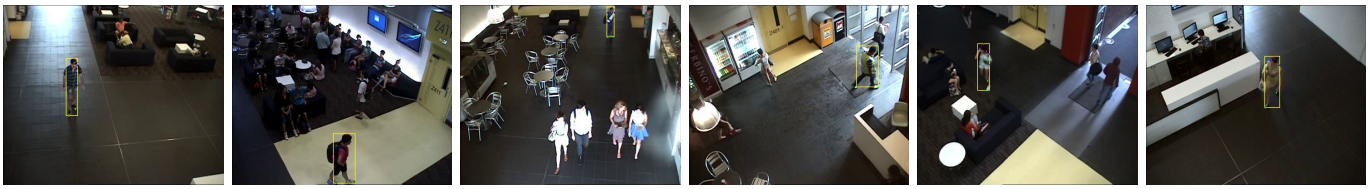
A. Data

Introduced here is a new 110 sequence database for locating people from a semantic search query. Each sequence contains a single target subject, and a query that describes the subject. A total of 6 cameras are used, with only one camera used in each sequence. All cameras are recorded at a resolution of 704×576, and have been manually calibrated using the technique described in [18]. Examples of each of these 6 camera's can be seen in Figure 1.

A key consideration when collecting the database was to ensure that there was minimum ambiguity in subject description and location. Removal of this ambiguity results in the knowledge that each sequence contains only a single complete match for the target query, allowing for an accurate analysis of developed algorithms. To build a unique soft biometric query, multiple traits have been annotated in the included xml files. The following traits are annotated:

- Torso and leg primary and secondary colours, categorised into the 11 culture colours [15]
- Clothing type for the torso (short sleeve, long sleeve, no sleeve), and leg (pants, short shorts, long shorts, skirt, dress)
- Clothing texture, categorised into plain, horizontal stripe, vertical stripe, diagonal stripe, checks, spots, pictures
- Age, categorised into less than 20, $15-35$, $25-45$, $35-55$, greater than 50

[2]visit https://wiki.qut.edu.au/display/saivt/SAIVT+Semantic+Person+Search+Database



(a) Camera 1 (b) Camera 2 (c) Camera 3 (d) Camera 4 (e) Camera 5 (f) Camera 6

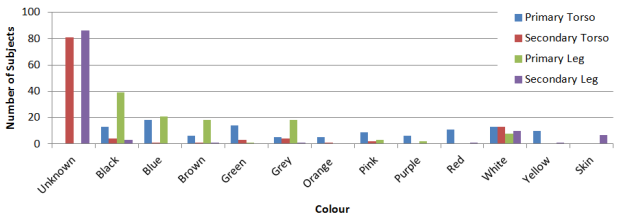
Fig. 1. A sample image from each camera, each showing an image from a different sequence. The target queries for the six images are: (a) Sequence *C1-BlackBlueStripe-BlueJeans*: 15-35 year old short caucasian male of average build with brown hair, wearing a blue striped short sleeve shirt, plain blue shorts and brown shoes, and carrying luggage; (b) Sequence *C2-RedShirt-BlackShorts*: 25-45 year old very short asian female of large build with dark hair, wearing a red short sleeve shirt, plain black shorts and white shoes, and carrying luggage; (c) Sequence *C3-DiagonalTop-HorizontalPants*: 15-35 year old short asian male of average build and height with dark hair, wearing a green and white striped short sleeve shirt, blue and white striped trousers and white shoes, and carrying luggage; (d) Sequence *C4-FluroYellowTop-BluePants*: a 15-35 year old dark skinned and dark haired male of average height and large build, wearing a yellow and grey striped sleeveless shirt, plain blue trousers and white shoes; (e) Sequence *C5-LightGreenShirt-WhiteShorts*: a 15-35 year old short caucasian female of very slim build with brown hair, wearing green sleeveless top, white shorts and white shoes, and carrying luggage; (f) Sequence *C6-YellowWhiteSpotDress*: a 15-35 year old short caucasian female of slim build with brown hair, wearing a yellow and white spotted dress, and carrying luggage. Images also show the annotated feet, waist, shoulder, neck and head positions, and the target bounding box (yellow).

- Height, categorised into very short ($1.3 - 1.6$m), short ($1.5 - 1.7$m), average ($1.6 - 1.8$m), tall ($1.7 - 1.9$m), very tall ($1.8 - 2.1$m)
- Build, categorised into very slim (aspect ratio in the range $0.10 - 0.20$), slim (aspect ratio in the range $0.15 - 0.25$), average ($0.20 - 0.30$), large ($0.25 - 0.35$), very large ($0.30 - 0.40$)
- Skin colour, categorised Caucasian, Asian, Hispanic, Dark
- Hair colour, categorised into Blonde, Brown, Dark, Red, Grey, Other
- The presence of luggage
- Gender

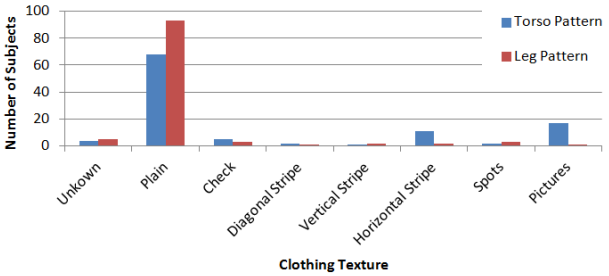
Efforts are made to collect a wide variety of subjects with different appearances, and the distributions of the colour and height traits are shown in Figure 2. To ensure further variety in subject appearance, data is captured from six different time periods with varying lighting conditions and levels of crowding.

The first 30 frames of each sequence is reserved for initialising the system and components such as a background model (if used). Following these initial frames, each frame that includes the subject is manually annotated to include both feet, both shoulders, the top of the head, as well as the points of furthest distance from the centre of the body at the waist and neck regions (as seen in Figure 1). While the evaluation protocol outlined in Section III-B requires only the construction of a bounding box based on the maximum feet and head position as well as the largest x and y direction distances, the additional annotations potentially enable more advanced evaluation protocols in the future.

In cases where a trait or a body marker cannot be determined (i.e. if the trait can not be clearly seen or accurately approximated) the trait is labelled unknown (denoted -1 in the annotation), and annotation continues. Including these missing traits and body markers in this manner allows for any system to handle the lacking annotation without subjective annotation reducing the output accuracy. Overall there is an average of 102 annotated frames per subject with a maximum of 290 and a minimum of 21 annotated frames. This large number of annotated frames allows for accurate checking over a wide range of scenes, pose and positions within camera view planes.



(a) Clothing Colour Distribution



(b) Clothing Pattern Distribution

Fig. 2. Distribution of torso and legs colours (a) and clothing patterns (b) across the 110 annotated subjects.

In addition to the 110 annotated subjects, 4 approximately 5 minute clips from each camera have been captured to train background models; and a collection of image patches for the 11 culture colours and skin has been collected to train colour models. This data is also be provided with the main database.

B. Evaluation Protocol

Performance is evaluated by computing the localisation accuracy for all frames with annotated ground truth. A ground truth bounding box is constructed from the annotated points. Left and right bounds are taken from the feet, shoulder, waist or neck annotations, while the head and feet provide the top and bottom bounds. If any of these bounds were not located

then that frame is deemed to be un-annotated, and is not used in the evaluation.

The bounding box computed by the proposed algorithm is compared to the ground truth using,

$$L_{x,i} = \frac{D_{x,i} \cap GT_{x,i}}{L_{x,i} \cup GT_{x,i}}, \quad (1)$$

where $D_{x,i}$ is the output of the algorithm for subject x, frame i, and $GT_{x,i}$ is the corresponding ground truth bounding box; to measure the localisation accuracy.

The accuracy over an entire sequence is given by the average accuracy for each frame in the sequence,

$$L_x = \frac{1}{N} \sum_{i=1}^{i=N} L_{x,i}, \quad (2)$$

where N is the total number of frames for subject x; and the overall performance is given by the average of each sequence accuracy,

$$L = \frac{1}{M} \sum_{x=1}^{x=M} L_x, \quad (3)$$

where M is the total number of subjects (110). Sequences are not weighted by their length when computing the average to ensure that all queries are given equal weight, and long sequences do not dominate the overall performance metric.

To replicate a human operators ability to recognise an individual based on the detected regions, results are presented based on 4 accuracy thresholds. Absolute misses are represented by $L_x < 1\%$, majority misses where the subject could be recognised dependent on a precise region are $L_x < 5\%$, and $L_x < 10\%$, and finally substantial matches are given by $L_x > 50\%$.

IV. Proposed System

To localise a subject in video, we propose an approach that builds a target query for each video sequence based on the following semantic descriptors: primary torso colour, primary leg colour, subject height and build, and the leg clothing type. This is a significant difference from re-detection systems as no prior registration of the desired target subject is required. To model the subjective nature witnessed primarily within the colour based traits, a weighting system based on the confidence of the observation is proposed. To allow for an efficient search in video, a particle filter is used as in [2].

A. Search Query Formulation

As with [2], [3], colour selection is categorised into the 11 culture colours. Our approach also incorporates skin, resulting in a total of 12 colours. A Gaussian mixture model is trained for each colour in the LabCie colour space, using colour snippets obtained from the same camera network as the main database (see Section III). No normalisation or illumination correction is made and each model relies solely on the range of colour snippets available, and the expected illumination invariance achieved through using the LabCie colour space. Leg clothing type is also further categorised from the annotation (see Section III-A) to one of three available choices: unknown, full (pants) and not full (a combination of long shorts, short shorts, skirt and dress).

Each particle describes a candidate location (x and y position, height and aspect ratio), and particle heights and aspect ratios are distributed within the ranges outlined in Section III-A according to the target query (i.e. when searching for a 'tall' person, the height will be uniformly distributed between 1.7m and 1.9m).

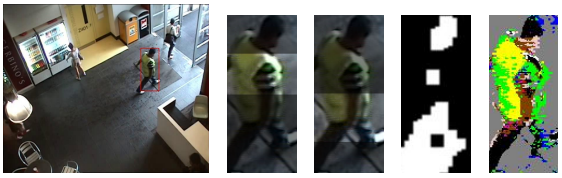
B. Searching for a Target

The proposed approach has 4 stages: particle filtering, trait similarity calculation, trait weight filtering and finally particle weighting. The particle filter is initialised with a randomly generated 500 particle set, and 3 iterations are performed per frame to avoid convergence on incorrect subjects.

The similarity of the primary torso colour and primary leg colour are computed using a method similar to that outlined in [2]. A similarity score (S_t for the torso, and S_l for the legs) is produced based on the normalised probability of the target colour appearing in the desired area, such that for the torso,

$$S_t = \frac{\sum_{x,y \in t} P_{x,y,c} \times Mo_{x,y}}{\sum_{x,y \in t} \times Mo_{x,y}}, \quad (4)$$

where the summation of the likelihood, $P_{x,y,c}$ of the target colour, c appearing at each pixel location, x, y, is completed over the limits of the torso region, t. $Mo_{x,y}$ indicates the presence of motion at x, y, and is set to $Mo_{x,y} = 1$ if motion is present, and $Mo_{x,y} = 0.5$ otherwise, resulting in regions that contain motion (and are thus more likely to contain the target) receiving a higher weight, and regions that are likely to correspond to the background (and are thus less likely to match the target colour) being diminished. S_l is calculated in a similar manner. Torso and leg regions are selected as in [2], such that areas likely to correspond to the head and feet (and thus likely to have a different colour) are discarded. Figure 3 displays an overview of the two primary regions of interest for which the torso and leg similarity scores are calculated.



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

Fig. 3. An image to be searched, and a particle location (red bounding box) are shown in (a). Given this particle location, regions of interest for the torso (b) and legs (c) are defined. Within each of these regions, the motion (d) and colour (e) are considered and used to determine how well the candidate bounding box matches the query (note that when considering the colour, a soft decision approach is used rather than the hard decision indicated by (e)).

A limitation of [2] is that all traits are assumed to be equally reliable. To overcome this, we seek to capture the quality of each trait such that observations that have a lower quality receive a diminished weight, helping the particle filter converge on areas of greater quality and confidence.

To model the quality of the trait, we use a combination of motion segmentation (the approach of [19]) and a pixel-wise histogram that captures the colour distribution at each pixel over time. The use of motion segmentation allows uncertainty around the location to be captured (confidence is reduced when there is little to no motion), while the colour history allows the confidence of the query to be assessed (a query colour that is the same as the dominant historic background colour has greater uncertainty).

The motion segmentation quality component, Q_{Mpr}, is calculated in a two step process. Initially the proportion of motion within the bounded region is obtained Mpr. If this proportion is above 0.5,

$$Q_{Mpr} = 1 - log_2(Mpr), \tag{5}$$

is used to calculate the final quality score for motion segmentation. However if $Mpr < 0.5$, the quality of the motion segmentation component is set to zero.

A pixel-wise histogram model that captures individual pixel colour classification over a period of time is used to capture uncertainty associated with the colour query. Additional captured video sequences (see Section III-A) are classified using the trained GMM's to capture the distribution of colours at each individual location over time to create the histogram model, V_{PH}. The colour quality, Q_{PH}, is then calculated as follows,

$$Q_{PH} = \frac{\sum_{x,y \in r} 1.0 - V_{PH}(x,y,c)}{R}, \tag{6}$$

where r is the region of interest (i.e. the target torso or leg region), c is the target colour and R is the size of the region, r. Thus, Q_{PH} represents the proportion of the background that has historically been a different colour to the target colour.

The two quality components are then simply combined by calculating their geometric mean,

$$Q = \sqrt{Q_{Mpr} * Q_{PH}}. \tag{7}$$

Quality measures are calculated for both the torso and leg regions (Q_t and Q_l respectively), and are subsequently used to scale the similarities (S_t and S_l for the torso and legs respectively). The similarities are then combined using the geometric mean.

C. Assessing Clothing Type

A limitation of the approach in Section IV-B is that for torso and leg regions, when the subject is wearing clothing that does not cover the entire region (i.e. shorts rather than long trousers), a significant portion of the region will be skin colour which will falsely diminish the weight. To help alleviate this problem, we propose an approach to modify the bounds of the target leg region to remove areas that do not represent the article of clothing. Although the annotation conducted allows for a wide variety of clothing types to be incorporated, for simplicity we utilise the following subset [Unknown, Long, Not Long], where 'Not Long' is considered to be everything that does not fall explicitly into the other two categories.

The motivation behind this method is to rebound the leg region by decreasing the total area, and in turn increasing the desired target colour presence in the target region. We utilise the asymmetry driven chromatic and spatial operators described in [11] (which were proposed to segment a body into head, torso and legs and find the dominant plane of symmetry in the head and leg regions), and apply these to a skin mask, Sk, rather than a motion mask. The chromatic operator measures the difference in two adjoining colour patches, while the spatial operator measures the difference in two adjoining mask regions. The skin mask is computed as follows,

$$Sk_{x,y} = P_{x,y,c} \times Mo_{x,y}, \tag{8}$$

where $P_{x,y,c}$ is the probability of a given colour, c, in this case skin colour; and $Mo_{x,y}$ is a scaling factor based on the motion at pixel x, y, such that $Mo_{x,y} = 1$ if motion is present, otherwise $Mo_{x,y} = 0.5$. As with the region similarity computation in Equation 4, motion pixels are emphasised as they are less likely to represent background regions.

The chromatic, $C(i,\delta)$, and spatial, $S(i,\delta)$ operators from [11] are then applied as follows to rebound the region of interest,

$$A_i = argmin((1 - C(i,\delta)) + S(i,\delta)) \tag{9}$$

where A_i corresponds to the image index associated with the minimum score achieved through the summation of the spatial and chromatic operators resulting in a newly bounded leg region, and is intended to represent the lower boundary between the item of clothing (i.e. shorts) and skin.

V. Experimental Results

We evaluate the proposed approach and compare it to the baseline system of [2], using the new database proposed in Section III. The proposed system is evaluated in two stages: *Proposed 1* denotes the proposed system without compensating for leg clothing outlined in Section IV-C; and *Proposed 2* denotes the system outlined in Section IV. As both the proposed system and the baseline are non-deterministic due to the use of a particle filter, 10 evaluations are run across the complete database and results are averaged per subject. Average localisation is reported as outlined in Section III-B.

The baseline system achieves an average localisation accuracy of 24.28%, while *Proposed 1* achieves an improved localisation accuracy of 29.38%, and *Proposed 2* achieves a slightly lower accuracy (though still above the baseline) of 27.05%.

However as Figure 4 shows, despite a lower overall accuracy than *Proposed 1*, the complete system of *Proposed 2* does obtain the fewest subjects for which very poor localisation is achieved. Furthermore, it is clear that incorporating quality into the system results in significant performance gains. The improved performance on challenging subjects is most noticeable on subjects wearing short shorts, where the modifications to the region of interest are pronounced. However for subjects wearing longer shorts, the changes to the region of interest are often inconsistent, possibly due to the smaller amount of skin

visible, which leads to a reduction in performance. However, further investigation is still required.

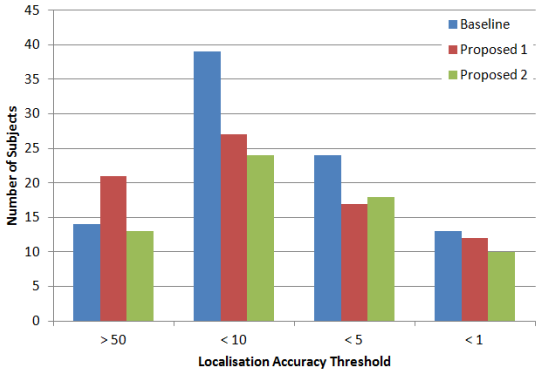


Fig. 4. The number of subjects with localisation accuracy greater or below a given threshold. The thresholds shown are (from left to right): greater than 50% localisation accuracy (i.e. very good localisation), less than 10%, less than 5% and less than 1% (i.e. failure to localise)

Despite the improvement over the baseline, overall performance is still limited. Limitations are due to a variety of factors, including errors in motion segmentation, skin classification, and incorrect detection of the correct leg region using the method proposed in IV-C.

VI. Conclusion

In this paper we have presented a new, publicly available, unconstrained 110 subject database for the purpose of person location from a semantic description. The database includes full subject annotation for both the semantic query consisting of gender, height and build, skin and hair colour, as well as clothing colour, type and texture. The annotation makes use of traits easily recognised and described by the general population. Overall, in excess of $11,000$ target locations are annotated, with between 21 and 290 locations annotated for each subject. In addition, data for training background models for each of the 6 cameras used, as well as data to train colour models using the culture colour categories is also provided.

Finally, we propose and evaluate a novel approach for person localisation from a semantic description that extends the baseline system of [2]. We incorporate a measure of quality to indicate observation confidence, and use symmetry driven operators to better localise the target region in the presence of large regions of skin. Through the incorporation of quality measures we achieve a relative localised accuracy improvement of 21% over the baseline system. Baseline system improvement is also achieved through the incorporation of these quality measures and asymmetry driven operators to rebound the leg region, however the primary benefit is witnessed in the reduced number of poorly localised sequences.

Future work will aim to develop and incorporate more robust methods of clothing detection for both the upper and lower body, and investigate improved methods for incorporating quality such as through the Dempster-Shafer theory of evidence. Other traits will also be investigated such as clothing texture and patterns, as well as gender and age.

References

[1] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita, "Vise: Visual search engine using multiple networked cameras," in *ICPR*, vol. 3, 2006, pp. 1204–1207.

[2] S. Denman, M. Halstead, A. Bialkowski, C. Fookes, and S. Sridharan, "Can you describe him for me? a technique for semantic person search in video," in *DICTA*, 2012.

[3] 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, p. 38, 2011.

[4] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *ICBA*, 2004, pp. 717–738.

[5] G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognition*, vol. 46, no. 1, pp. 188 – 198, 2013.

[6] H. J. Escalante, C. A. Hernández, and et al., "The segmented and annotated iapr tc-12 benchmark," *CVIU*, vol. 114, no. 4, pp. 419–428, 2010.

[7] D. Gray and T. H, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, Marseille, France, 2008, pp. 262–275.

[8] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, and C. B. Fookes, "A database for person re-identification in multi-camera surveillance networks," in *DICTA*, 2012.

[9] J.-E. Lee, R. Jin, A. Jain, and W. Tong, "Image retrieval in forensics: Tattoo image database application," *MultiMedia, IEEE*, vol. 19, no. 1, pp. 40–49, 2012.

[10] S. Denman, C.Fookes, A. Bialkowski, and S. Sridharan, "Soft-biometrics: unconstrained authentication in a surveillance environment," *DICTA*, p. 7, 2009.

[11] M. Farenzena, L. Bazzani, A. Perina, and et al., "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010, pp. 2360–2367.

[12] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching." ICCV, 2013.

[13] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu, "Human re-identification by matching compositional template with cluster sampling." ICCV, 2013.

[14] A. D'Angelo and J.-L. Dugelay, "Color based soft biometry for hooligans detection," in *ISCAS*, 2010, pp. 1691–1694.

[15] B. Berlin and P. Kay, *Basic Color Terms: Their Universality and Evolution*. Berkeley: University of California Press, 1969.

[16] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *WACV*. IEEE, 2009.

[17] J. Joo, S. Wang, and S.-C. Zhu, "Human attribute recognition by rich appearance dictionary." ICCV, 2013.

[18] R. Y. Tsai, "An efficient and accurate camera calibration technique for 3d machine vision," *CVPR*, pp. 364–374, 1986.

[19] S. Denman, C. Fookes, and S. Sridharan, "Improved simultaneous computation of motion detection and optical flow for object tracking," in *DICTA*, Melbourne, Australia, 2009.