Semantic Person Retrieval in Surveillance Using Soft Biometrics: AVSS 2018 Challenge II

Michael Halstead, Simon Denman, Clinton Fookes
School of Electrical Engineering and Computer Science
Queensland University of Technology, Brisbane, Australia
{m.halstead, s.denman, c.fookes}@qut.edu.au

YingLi Tian
Department of Electrical Engineering,
City College of New York, New York, US
ytian@ccny.cuny.edu

Mark S. Nixon
School of Electronics and Computer Science,
University of Southampton, UK
msn@ecs.soton.ac.uk

Abstract

In surveillance and security today it is a common goal to locate a subject of interest purely from a semantic description; think of an offender description form handed into a law enforcement agency. To date, these tasks are primarily undertaken by operators on the ground either by manually searching a premises or by combing through hours of video footage. Using computer vision to attempt to partially or fully automate these tasks has been gathering interest within the research community in recent years, however, to date there has been little coordinated effort to advance the field. This has motivated the challenge that is presented in this paper: the AVSS Challenge on Semantic Person Retrieval in Surveillance Using Soft Biometrics. This challenge consists of two related tasks: person re-identification from a semantic query and person search within a video from a query. In this paper, we present the publicly available data for this challenge, the evaluation framework, and the challenge results. It is our hope that the outcomes of this challenge and the availability of the data used in this challenge will expedite research and development in this societal field.

1. Introduction

Modern surveillance systems demand a solution for the challenging task of locating a subject of interest from a pure semantic description. When these tasks present there is limited capacity for automated surveillance solutions and operators are required to manually complete these searches. To date, in an effort to reduce manual requirements, researchers have focused on person re-identification methodologies to solve this complex problem, however, in circumstances where pre-search subject enrolment images are not available, these techniques fail.

Semantic search techniques, which search based on a textual query, offer an avenue to overcoming this dependence on image based enrolment; however to date there has been little coordinated effort for this task. This challenge aims to rectify this, and provide databases and protocols as a resource for researchers and to further promote research in the area. Two tasks are considered: 1) a semantic re-identification task, where the correct person must be identified from an image gallery given a semantic query; and 2) a semantic search task where the correct person must be located in a short video given a semantic query.

In the remainder of this paper, we outline the two challenge tasks in detail in Section 2, and the database that the challenge uses in Section 3. Challenge results, which will serve as a baseline for future research are outlined in Section 4, and Section 5 concludes the paper.

2. Challenge Tasks

This challenge contains two semantic search tasks that participants could complete independent of each other, or use data from one to aid in the second. The first task (see Section 2.1) can be seen as a person re-identification task, where rather than the input query being an image it is a semantic query. The second task (see Section 2.2) is a search task, where the aim is to localise a subject in a video sequence given a semantic description.

2.1. Task 1: Semantic Re-Identification

Task 1 can be seen as analogous to a person re-identification task. Given a semantic query, the challenge is to locate the correct person within a gallery. As in

a person re-identification, cumulative match characteristic curves and ranked outputs are used to assess performance.

2.2. Task 2: Semantic Search

Task 2 is intended to emulate the function of locating a person in an environment given a semantic description, and requires that given a semantic query, the person within a video who matches that query be detected. The average Intersection over Union (IoU) over a sequence is used to evaluate performance.

3. Database and Protocol

Separate data is provided for both Task 1 and Task 2, however annotation and labelling for some attributes is consistent between the two. This consistency (or near consistency) allows data from one task to serve as additional training data for the other. For both datasets, the similar attributes within the queries consist of:

- Clothing colours, defined in terms of the set of "Culture Colours" [1] (blue, black, brown, green, grey, orange, pink, purple, red, white, yellow);
- Clothing textures defined as one of: irregular (logos, pictures), plaid, diagonal plaid, plain (single colour), spots, diagonal stripes, horizontal stripes, vertical stripes;
- Clothing type for both the torso and leg regions are annotated. The labels vary slightly between the two tasks, however, the labels can be interpolated for the other task:
  - Task 1 labels both the torso and leg regions as either short (shorts or a short skirt, singlet or short sleeve shirt) or long (trousers or dress, a long sleeved shirt);
  - Task 2 has three labels for the torso region (long, short, no sleeve) and five for the leg region (long pants, dress, skirt, long shorts, or short shorts).
- Gender is defined as male, female, or unknown;
- Luggage is defined as either being in appearance or not, these labels were used to simplify the task of annotating the presence of luggage on the subject.

In each of the tasks the precise formulation of the queries using these attributes does vary and this is outlined in the following sections. The commonality of these attributes between tasks (such as colours, texture, and type). However enables them to be used together, to some degree.

In each of the tasks there are also uniquely annotated attributes, in Task 1 this includes the subjects pose with respect to the camera (front, back, 45^oC, or 90^oC), while not specifically a soft biometric trait it may aid in the classification of the other labels. Task 2 contains a number of extra labels to assist detection and classification in this more challenging domain:

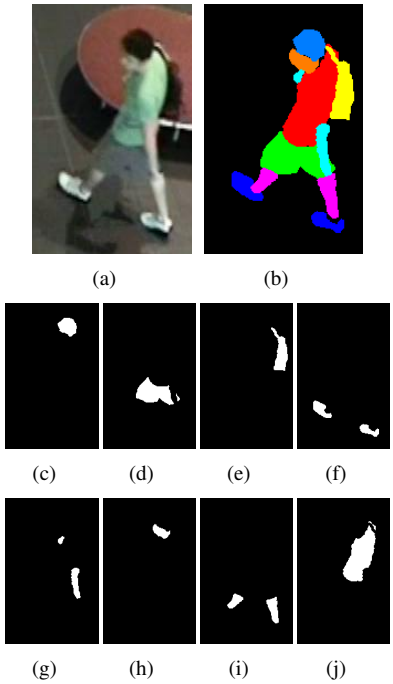


Figure 1. An example of one of annotation for one subject. The example is a male at 90^o, with short green plain textured torso clothing, short grey and plain leg clothing, carrying luggage. (a) The original image, (b) an example of the person parsed into regions; the remaining images are binary masks for the (c) hair, (d) leg clothing, (e) luggage, (f) shoes, (g) skin in the arms region, (h) skin in the head region, (i) skin in the leg region, (j) torso clothing.

- Age ($0-20, 15-35, 25-45, 35-55, 50+$);
- Height (very short, short, average, tall, or very tall);
- Build (very slim, slim, average, large, or very large);
- Skin (light, medium, or dark);
- Hair (blonde, brown, dark, red, grey, or other);
- Shoe colour.

Both of these task datasets are available for download online, as are the evaluation tools [1]. All annotations for both tasks are provided as XML files.

3.1. Task 1

3.1.1 Data

Training data for Task 1 consists of 520 images, each of which has:

[1] https://github.com/simondenman/SemanticSearchChallengeAVSS18

Attribute Distribution		
Labels	Training Set	Testing Set
Colour Labels	unknown, black, blue, brown, green, grey, orange, pink, purple, red, white, yellow, skin	
Torso Colour 1	4, 140, 39, 27, 21, 83, 7, 22, 21, 22, 122, 12	0, 51, 23, 12, 9, 21, 5, 7, 5, 19, 35, 9
Torso Colour 2	339, 44, 6, 7, 2, 41, 0, 3, 1, 7, 67, 3	80, 22, 5, 3, 2, 28, 0, 5, 2, 2, 43, 4
Torso Colour 3	489, 7, 0, 1, 2, 6, 0, 1, 0, 3, 6, 5	165, 8, 2, 1, 0, 5, 1, 1, 3, 2, 8, 0
Leg Colour 1	6, 210, 154, 43, 5, 61, 3, 5, 0, 3, 25, 5	0, 69, 63, 8, 7, 27, 0, 6, 2, 0, 13, 1
Leg Colour 2	477, 8, 1, 3, 0, 15, 0, 1, 0, 1, 14, 0	165, 4, 1, 1, 0, 10, 0, 1, 0, 0, 13, 1
Leg Colour 3	513, 1, 0, 0, 1, 1, 0, 0, 0, 0, 3, 1	189, 2, 0, 0, 0, 2, 0, 0, 1, 0, 2, 0
Texture Labels	unknown, irregular, plaid, diagonal plaid, plain, spots, diagonal stripes, horizontal stripes, vertical stripes	
Torso Texture	2, 106, 33, 7, 310, 4, 2, 45, 11	0, 44, 19, 3, 88, 5, 2, 32, 3
Leg Texture	3, 37, 8, 1, 457, 4, 0, 6, 4	0, 20, 3, 0, 165, 2, 0, 3, 3
Clothing Types	unknown, long, short	
Torso Clothing Type	0, 219, 301	0, 75, 121
Leg Clothing Type	0, 356, 164	0, 114, 82
Gender Labels	unknown, male, female	
Gender	3, 295, 222,	8, 98, 90,
Pose Labels	unknown, front, back, 45, 90	
Pose	0, 180, 128, 129, 83,	0, 77, 40, 58, 21,
Luggage Labels	unknown, yes, no	
Luggage 1	0, 308, 212,	9, 126, 61,

Table 1. Task 1 distribution of the different attributes available in both the testing and training datasets. Contains the label counts and the name of the labels used in the datasets.

- A cropped colour (RGB) image of the person;
- A semantic query describing that persons appearance, covering the global traits of gender, pose, and luggage. The following traits are also labelled for both the leg and torso regions: primary, secondary, and tertiary colours, clothing type, and clothing texture.
- A set of binary masks are also included for each of the following components: torso clothing, leg clothing, luggage, shoes, hair, facial skin, arm skin, and leg skin.

Figure 1 shows an example of the images for a single subject. This includes an example of the annotation being used to generate the subject's soft biometric query.

Effort has been made to include a variety of colours and textures in the data, and images are taken from two different environments on a university campus (from a total of 10 cameras) to further increase the diversity of the data.

Testing data is captured from the same cameras and consists of a further set of 196 images and their corresponding query, each of which is unique. Mask images are not provided for the test set, and will not be available upon public release. A full breakdown of the distribution of the labels for task 1 is shown in Table 1. Both the training and testing distributions are displayed, along with the associated attribute labels used to build a subject query.

3.1.2 Evaluation Protocol and Metrics

Task 1 can be seen as a re-identification task, and following the re-identification literature [9] we use cumulative match characteristic (CMC) curves and performance at Rank-1, Rank-5, Rank-10 and Rank-25 to evaluate performance.

3.2. Task 2

3.2.1 Data

Training data for Task 2 is taken from [6]. This consists of 110 short video segments, taken from 6 cameras (each annotated with Tsai's calibration scheme [11]) located within the main floor of a building on a university campus. Each sequence in the training set is labelled with a set of soft biometric traits to describe the target, where a -1 score represents either the lack of that trait or difficulty in annotating it. Along with the soft biometric traits, nine key body markers were also annotated. More details on the cameras and body markers can be found in [6].

The test set for Task 2 consists a further 41 queries taken from 4 of the 6 cameras (cameras 1 and 6 are omitted) used in collecting the training set. Collection and annotation of the testing data followed the same guidelines of the training set to ensure similarity, with at least the first 30 frames of each sequence reserved to allow the subject to fully enter the view, and for background models to be initialised. A full distribution of the training and testing soft biometric traits is outlined in Table 2.

To offer further detail on the performance of participants the testing queries were categorised into very easy, easy, medium, and hard. These labels are defined as follows:

- Very Easy: sparsely populated scene, no complicating factors, target subject clearly visible;
- Easy: scene contains multiple people, but the target is clearly distinct;
- Medium: one of the following compounding factors is present in the scene: similar subjects (i.e. partial match to query) present, occlusion of target, heavy crowding;
- Hard: two of the above mentioned compounding factors is present in the scene.

Of the 41 queries, the testing dataset contains 6 labelled as very easy, 13 as easy, 12 as medium, and 10 as hard.

3.2.2 Evaluation Protocol and Metrics

Metrics use the intersection over union (IoU) per [6]. An average IoU is calculated per sequence, and sequence results are averaged over all sequences to obtain a final accuracy measure.

Attribute Distribution		
Labels	Training Set	Testing Set
Colour	unknown, black, blue, brown, green, grey, orange, pink, purple, red, white, yellow, skin	
Torso 1	0, 13, 18, 6, 14, 5, 4, 10, 6, 11, 13, 10, 0,	0, 7, 3, 2, 5, 5, 1, 4, 2, 3, 6, 3, 0,
Torso 2	79, 5, 1, 1, 3, 4, 2, 1, 0, 0, 14, 0, 0,	20, 3, 1, 0, 0, 4, 0, 0, 1, 1, 11, 0, 0,
Leg 1	0, 39, 21, 18, 1, 18, 0, 3, 2, 0, 8, 0, 0,	0, 14, 13, 4, 0, 6, 0, 2, 0, 0, 1, 1, 0,
Leg 2	86, 3, 0, 1, 0, 1, 0, 0, 0, 1, 10, 1, 7,	35, 0, 0, 0, 1, 1, 0, 0, 1, 0, 3, 0, 0,
Shoes Colour	3, 39, 0, 10, 0, 15, 0, 0, 0, 1, 19, 1, 22,	1, 14, 0, 1, 0, 4, 0, 0, 0, 0, 8, 0, 13,
Texture	unknown, plain, check, diagonal stripe, vertical stripe, horizontal stripe, spots, pictures	
Torso Texture	4, 68, 5, 2, 1, 11, 2, 17,	0, 23, 3, 0, 0, 8, 0, 7,
Leg Texture	5, 93, 3, 1, 2, 2, 3, 1,	0, 35, 1, 0, 0, 2, 0, 3,
Torso Type	unknown, long sleeve, short sleeve, no sleeve	
Torso Type	0, 35, 63, 12,	0, 9, 27, 5,
Leg Type	unknown, long pants, dress, skirt, long shorts, short shorts	
Leg Type	0, 52, 14, 7, 31, 6,	0, 23, 4, 4, 8, 2,
Age labels	unknown, 0-20, 15-35, 25-45, 35-55, 50-	
Age	25, 0, 56, 22, 6, 1,	10, 1, 23, 6, 1, 0,
Gender Labels	unknown, Male, Female	
Gender	1, 64, 45,	0, 20, 21,
Height Labels	unknown, very short, short, average, tall, very tall	
Height	0, 13, 28, 36, 25, 8,	0, 8, 13, 8, 8, 4,
Build Labels	unknown, very slim, slim, average, large, very large	
Build	0, 7, 76, 26, 1, 0,	0, 3, 19, 16, 3, 0,
Skin Labels	unknown, light, medium, dark	
Skin	4, 70, 27, 9,	2, 28, 11, 0,
Hair Labels	unknown, blonde, brown, dark, red, grey, other	
Hair	0, 8, 43, 42, 3, 5, 9,	0, 5, 8, 26, 0, 0, 2,
Luggage Labels	unknown, yes, no	
Luggage	1, 77, 32,	1, 30, 10,

Table 2. Task 2 distribution of the different attributes available in both the testing and training datasets. Contains the label counts and the name of the labels used in the datasets.

4. Challenge Results

4.1. Task 1

Overall results for Task 1 are shown in Table 3 and Figure 2. The approach of [10] achieved the best performance, although [5] achieved similar accuracy. Impressively, at rank 25, [5] and [10] are able to achieve greater than 90% accuracy, and all methods achieve an accuracy above 50%.

Perhaps unsurprisingly, all approaches for Task 1 were based around deep learning, although with key differences. [12], [5] and [10] utilised similar pipelines whereby a pre-trained DCNN (or multiple in the case of [10]) was adapted to the attribute classification task. [12] sought to segment the subject from the background first, and then used local (upper and lower body regions) and global classifiers whose results were averaged. [5] and [10] eschew segmentation and simply pass the images of the subjects through DCNN to predict attributes. All of [12, 5, 10] perform classification of attributes jointly. However different distance measures have been proposed. [12] uses the Hamming distance between the query and the classified attributes; [5] and [10] both investigated different measures (including the Hamming distance) and settled on using a productory between the query and classification results [5] and the Euclidean distance [10]. These results suggest that delaying the hard decision that is forced by using the Hamming distance is beneficial. It is important to note that [10] uses an ensemble of four DCNNs, and their combined results from all four networks is substantially higher than for either network individually; while [5] uses a single network.

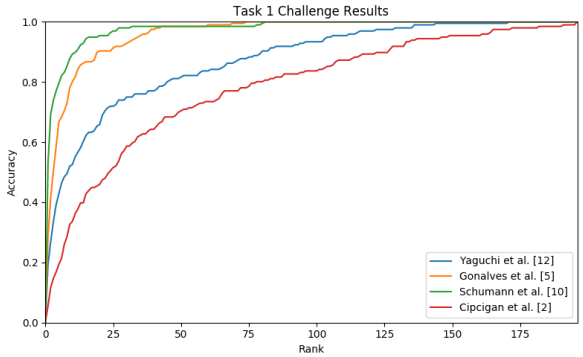


Figure 2. CMC Curves for Task 1

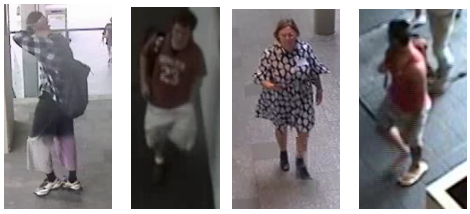
Finally, [2] utilised semantic segmentation via a DCNN to locate body parts and subsequently used hand-crafted features derived from the segmentation to classify traits. [2] was the only method to not be completely based around a DCNN, and achieved the poorest performance.

Approach	Rank 1	Rank 5	Rank 10	Rank 25
[12]	0.194	0.429	0.526	0.719
[5]	0.286	0.668	0.801	0.913
[10]	**0.531**	**0.796**	**0.893**	**0.969**
[2]	0.056	0.194	0.337	0.515

Table 3. Summary of Performance for Task 1

Although overall performance for Task 1 was high, there were a small number of queries that proved difficult. Of the 196 queries, 10 were identified with an average rank (i.e. average rank returned by all participants) of 60 or more. A sample of these are shown in Figure 3.

From this, it can be seen that these queries contain a mix of poor lighting (b and d), unusual textures (a, in particular the pants, and c) and ambiguous or possibly erroneous annotation (a and d). For (a), the query incorrectly specifies no luggage; while for (d), the target description is for a pink long sleeved shirt. However the garment may be a pink singlet of a similar colour to the target's skin, making it hard to tell where one starts and the other finishes. These examples illustrate some of the on-going challenges in this task. The ambiguity in annotation, whether caused by poor lighting, low resolution or human errors; alongside the diverse nature of appearance and the difficulty in categorising some features (such as the clothing texture in Figure 3 c)



(a) Q-20, R-72 (b) Q-111, R-66 (c) Q-125, R-60 (d) Q-138, R-66

Figure 3. A selection of hard to identify queries for Task 1. Q-X denotes the query number, and R-Y indicates the average rank the query was recognised at. Queries are (a) Male, long sleeve black, while and grey checked shirt, long black and grey irregular patterned pants, no luggage; (b) Male, short sleeved brown and white shirt, grey shorts, with luggage; (c) Female, short sleeved while and brown irregular patterned dress, with no luggage; (d) Male, long sleeved pink shirt, grey and white shorts, no luggage.

will likely always pose a challenge. One interesting avenue to help address this problem is comparative labels, however such annotation was beyond the scope of this challenge.

4.2. Task 2

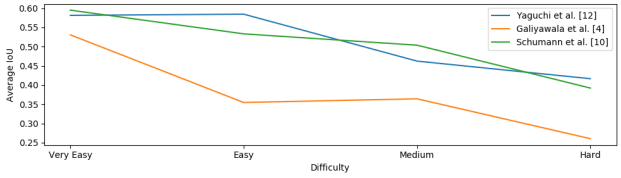
Overall results for Task 2 are shown in Table 4, including a comparison to the baseline system of [3]. From this we can see that all systems outperformed the baseline, and [12] and [10] achieved similar levels of performance.

Approach	Average IoU	% w IoU > 0.4
[12]	**0.511**	0.669
[4]	0.363	0.522
[10]	0.503	**0.759**
Baseline [3]	0.290	0.493

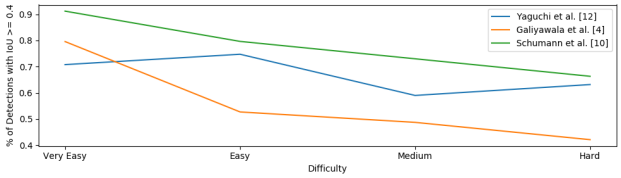
Table 4. Summary of Performance for Task 2

Considering the methodologies used, [12] and [10] extend their Task 1 approaches, using deep learning based multi-class classification methods that deploy a DCNN to detect people in the scene. [4] also uses a DCNN for detection, and follows this with a cascade of classifiers that aims to whittle down the detected people to leave only the target. The baseline of [3] is a non-deep learning approach that uses an avatar that is constructed from the query to drive a gradient descent search. It is worth noting that both [4] and [3] use only a subset of the available traits, while [12] and [10] use the full set, likely greatly aiding performance.

To further analyse performance, we break results down by difficulty (Figure 4) and into individual sequences (Figure 5). From Figure 4 it is clear performance decreases with difficultly. The "Very Easy" sequences typically contain only a single moving person, and all systems perform well on these sequences. As difficulty increases, [4] suffers a greater performance decrease than [12] and [10]. Of interest is the difference in performance between [12] and [10]. Both achieve similar average IoU's, but [10] obtains



(a)



(b)

Figure 4. Task 2 performance broken down by sequence difficulty

more detections with an IoU $>= 0.4$. This is likely due to the tracking approach of [10], which helps overcome any intermittent detection errors that may impact non-tracking approaches such as [12]. This also suggests that [12] has a tendency to either detect a subject very accurately, or very poorly, with less of a middle ground than the approach of [10].

From Figure 5 it is clear that performance varies across sequences, with each system performing best for a single sequence at some point. A number of sequences pose a challenge to two or all three of the systems, and Figure 6 shows an example from some such sequences.

Interestingly, two of these sequences (20 and 39) contain minimal crowding, but are complicated by having another subject of similar appearance being present. The other two sequences contain higher levels of crowding (27 and 40). In the instances of high crowding, we see [12] perform worst, despite offering better performance in the "hard" sequences as shown in Figure 4 (a). This may be a result of less robust object detection and/or noise in the masks used by [12], or perhaps the tracking approach of [10] which propagates a high quality detection forward and back through time.

5. Conclusions

The AVSS challenge on Semantic Person Retrieval in Surveillance Using Soft Biometrics has presented two tasks: a semantic re-identification task; and a semantic search (i.e. localisation) task. Both tasks required participants to map a trait based description capturing appearance features such as clothing types, colours, patterns and other soft biometrics to the image domain.

As part of this challenge, new data has been publicly released. Data from Task 1 contains 520 training images and 196 testing images, where semantic segmentation is provided as part of the training set. With each image across both the training and testing sets, manually annotated labels is provided. Task 2 data expanded on the previously

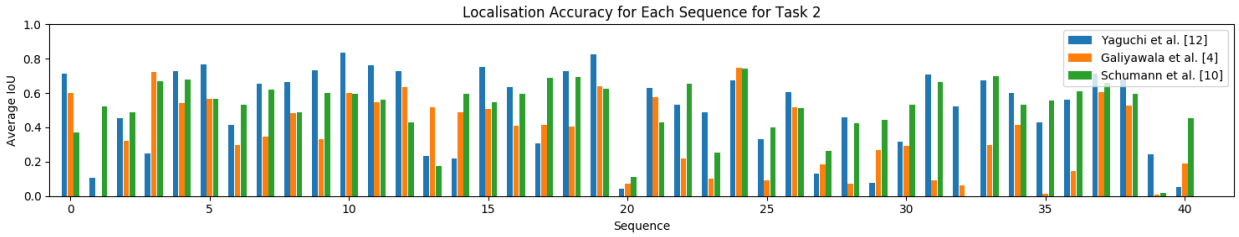


Figure 5. Per Sequence Performance for Task 2

(a) Sequence 20 (b) Sequence 27

(c) Sequence 39 (d) Sequence 40

Figure 6. Example frames from challenging sequences for Task 2. Correct target locations are indicated with a red bounding box.

released data of [6] (training set), this new data forms a testing set of 41 queries. For each new subject a soft biometric signature is labelled in a similar manner to training set. The new data contains varying levels of crowd flow and density along with instances of soft biometric signature similarities, making it a complex search task.

Participants used a variety of techniques, with a common theme being deep neural networks. For task 1, [10] achieved best performance with a rank 1 recognition rate of 0.531, and a rank 25 recognition rate of 0.969. For task 2, [12] achieved the best performance with an average IoU of 0.511, slightly ahead of [10] at 0.503.

While impressive performance was achieved for both tasks, challenges still remain. In particular, for Task 1 we see that poor lighting and/or ambiguities in traits remain a challenge. For Task 2 we note that high levels of crowding, or having multiple people that partially match the query continue to pose a challenge to techniques. Within Task 2 the value of a tracking was clearly demonstrated by [10], who propagated detections forward and back in time and was able to locate the target in over 75% of frames with an IoU > 0.4, compared to almost 67% for [12]. However, while this highighted the value of tracking for such a task, it is important to acknowledge that the method of [10] is non causal, and would require changes to operate in a live setting.

Through this challenge, the state-of-the-art in soft biometric semantic search has been advanced within the research community. Recent advances from the broader computer vision field have been utilised, notably a number of DCNN methods such as Mask R-CNN [7] and DenseNet [8], to improve performance in a field dominated, to date, by hand-crafted features and piecemeal classifiers. It is the organiser's hope that coupling the newly available datasets and protocols with the advances made by the participants, will contribute to further advancements to this societal field in the coming years.

References

[1] B. Berlin and P. Kay. *Basic color terms: Their universality and evolution*. Univ of California Press, 1991.

[2] E. B. Cipcigan and M. Nixon. Feature selection for subject ranking using soft biometric queries. In *IEEE AVSS Challenge 2*, 2018.

[3] S. Denman, M. Halstead, C. Fookes, and S. Sridharan. Searching for people using semantic soft biometric descriptions. *Pattern Recognition Letters*, 68:306–315, 2015.

[4] H. Galiyawala, K. S. Shah, V. J. Gajjar, and M. Raval. Person retrieval in surveillance video using height, color and gender. In *IEEE AVSS Challenge 2*, 2018.

[5] G. R. Gonalves, A. C. Nazare, M. Diniz, L. E. L. Coelho, and S. W. R. Soft biometric retrieval using deep multi-task network. In *IEEE AVSS Challenge 2*, 2018.

[6] M. Halstead, S. Denman, S. Sridharan, and C. B. Fookes. Locating people in video from semantic descriptions: A new database and approach. In *Proceedings of the 22nd International Conference on Pattern Recognition*, pages 4501–4506. IEEE, 2014.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988. IEEE, 2017.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[9] A. Khatun, S. Denman, S. Sridharan, and C. Fookes. A deep four-stream siamese convolutional neural network with joint verification and identification loss for person re-detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1292–1301. IEEE, 2018.

[10] A. Schumann and A. Specker. Attribute-based person retrieval and search in video sequences. In *IEEE AVSS Challenge 2*, 2018.

[11] R. Y. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 364–374, 1986.

[12] T. Yaguchi and M. Nixon. Transfer learning based approach for semantic person retrieval. In *IEEE AVSS Challenge 2*, 2018.