SEARCHING FOR SEMANTIC PERSON QUERIES USING CHANNEL REPRESENTATIONS

Simon Denman, Michael Halstead, Clinton Fookes, Sridha Sridharan

Image and Video Research Laboratory, Queensland University of Technology, Australia
{s.denman, m.halstead, c.fookes, s.sridharan}@qut.edu.au

ABSTRACT

It is not uncommon to hear a person of interest described by their height, build, and clothing (i.e. type and colour). These semantic descriptions are commonly used by people to describe others, as they are quick to relate and easy to understand. However such queries are not easily utilised within intelligent surveillance systems as they are difficult to transform into a representation that can be searched for automatically in large camera networks. In this paper we propose a novel approach that transforms such a semantic query into an avatar that is searchable within a video stream, and demonstrate state-of-the-art performance for locating a subject in video based on a description.

Index Terms— Semantic Search, Object Tracking, Localisation, Channel Representation

1. INTRODUCTION

A significant challenge in law enforcement and security is locating people based on a semantic description, such as may be given by an eye-witness to a crime. At present, searching for a person based on this form of description is a manual task that involves either combing through hours of CCTV footage, or having officers on the ground looking. The recent Boston Marathon bombing provides an example of how challenging this task is, as it took the FBI 3 days to search CCTV footage based on eye-witness reports and release suspect photographs to the public [1]. However recent developments in soft biometrics [2, 3] and attribute matching [4] are leading to the development of automated approaches to locate a specific person from a semantic description [5, 6], having the potential to improve security and save countless hours of labour.

Although a small number of approaches have recently been proposed, they are limited in their utility, either due to: their design as a database indexing system for post event analysis and thus an inability to support live data; their reliance on object detection methods which makes their deployment in crowded and unconstrained environments very challenging; and/or due to the simplistic region based method in which they match queries to the subject.

In this paper we propose a new approach to locate a subject from a semantic description. We show how a search query can be generated in the form of a channel representation that incorporates information about clothing colour and type, while allowing for a degree of uncertainty. State-of-the-art performance is achieved using these multi-channel models in combination with a recent single object tracking approach [7] which uses the channel representation as a template to locate the person of interest.

The remainder of the paper is organised as follows: Section 2 outlines existing work; Section 3 presents our proposed approach; Section 4 presents the evaluation and the paper is concluded in Section 5.

2. PRIOR WORK

While several approaches have been proposed to locate a person from a semantic query, the majority of these approaches are detection based [5, 8, 9, 10], in that they first require the person (or in the case of [10], the vehicle) to be detected using a computationally expensive object detection routine (such as [11, 12]). Following this detection, traits are extracted and stored in a database, allowing later processes to search for an entity within this index.

A variety of traits are used including clothing colour [5, 8, 9], gender and the presence of luggage [8], and facial attributes (i.e. bald, hair, hat) [9]; while [10] incorporates details on the vehicle size, it's location and direction of travel. While promising results are obtained for all systems, the evaluations of these approaches are piecemeal, and a rigorous evaluation using a standard database is not performed. Furthermore, these approaches are all reliant on object detection, limiting utility in crowded environments. In an unconstrained surveillance environment, person detection remains a challenging problem with state of the art approaches [13] still prone to false alarms and missed detections, whilst also being computationally expensive.

In contrast to detection and indexing based approaches, the techniques proposed in [6, 14, 15] are designed to work with a live video feed. Motivated by a desire to reduce 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.

Denman et al. [6] and Halstead et al. [15] sought to interrogate the scene directly to locate people who matched a given description. This technique allowed for a query to be searched for, within a live video feed without the use of detection routines. Colour (torso and leg) and height features are used to construct an avatar, which the system subsequently uses to search the video feed, using a particle filter to detect and track the subject. However the approaches of [6, 15] also have several limitations, largely arising from the use of a small number of traits which leads to the incorrect localisation of several subjects due to factors including incorrect trait attachment (torso as legs, and background as torso) and cases where the background is confused for an individual (i.e. the background is the same approximate colour as the target).

The approach proposed in this paper adapts and expands the query representation used in [15] by firstly combining clothing type and colour, and secondly by employing a channel representation (CR) to model the query. Through the use of a CR we are able to capture the spatial characteristics of the appearance while also allowing for a degree of uncertainty. We utilise the CR within a recent state-of-the-art single object tracker [7], and adopt a particle filter like approach to locate and track the person of interest.

3. PROPOSED APPROACH

The recent single object tracking approaches of [7, 16] have demonstrated state-of-the-art performance using a simple distribution field [16] or channel representation [7]. Furthermore, the template used by these approaches is well suited to being generated from a description rather than an image patch. The nature of these template representations is such that they model an estimated appearance, incorporating uncertainty about the exact colour, or location. In this section we outline our proposed approach: the process of generating the avatar as a channel representation is outlined in Section 3.1; the search process is presented in Section 3.2; and details on how scale is accounted for are given in Section 3.3.

3.1. Generating an Avatar

We generate an avatar from a set of characteristics that describe the target subject. In this approach, we consider the following three traits and categories:

- Torso clothing type: long sleeve shirt, short sleeve/sleeveless shirt.
- Leg clothing type: long trousers/skirt/dress, shorts (or short skirt/dress).
- Torso and leg clothing colour: black, blue, brown, grey, green, orange, pink, purple, red, white, yellow.

We denote these four components as T_{type} and L_{type} for the torso and leg clothing type; and T_{colour} and L_{colour} for the torso and leg colour.

To generate an avatar, we sum a set of learned components that relate to the selected traits. The average appearances for different body regions and clothing types are learned using a set of manually annotated silhouettes (see Section 4.1 for details). Silhouettes are resized to the same height, and edges are zero padded to the same width. Following this, all examples belonging to a given class (i.e. torso regions for 'long sleeve shirt') are used to learn the appearance by simply averaging the examples,

$$A_c = \frac{1}{N} \sum_{n}^{N} a_c(n), \tag{1}$$

where A_c is the average appearance of component c; $a_c(n)$ is the nth example image; with N examples in total. In this manner, we learn the appearance of the two torso clothing types and corresponding torso skin regions; the two leg clothing types and leg skin regions; and the head skin regions.

Colour models are trained for each of the target colours and skin (i.e. 12 colours in total) using the colour patches provided by [15]. GMMs with up to 12 components (determined using the BIC) are trained in the LAB colour space. A confusion matrix is also computed using the test patches provided in [15], which is used when generating the expected colour of a region.

Using the learned appearances and colour models, we then generate an avatar that describes the target subject's appearance. An occupancy mask that indicates the likelihood of a person being at a location is generated by combining the learned silhouettes for the target modes,

$$A_{skin} = A_{ts,T_{type}} + A_{ls,L_{type}} + A_h, \tag{2}$$

$$A_{torso} = A_{tc,T_{type}}, \tag{3}$$

$$A_{legs} = A_{lc,L_{type}}, \tag{4}$$

where $A_{ts,T_{type}}$ and $A_{tc,T_{type}}$ are the appearance of the torso skin (ts) and clothing (tc) regions for the selected torso type, T_{type}; $A_{ls,L_{type}}$ and $A_{lc,L_{type}}$ are the appearance of the leg skin (ls) and clothing (lc) regions for the selected leg type, L_{type}; and A_h is the average appearance of the head.

A set of colour masks, C_n, that correspond to a channel representation are then generated as follow,

$$\begin{aligned} c(x,y,n) = A_{torso}(x,y) \times P(n|T_{colour}) \\ + A_{legs} \times P(n|L_{colour}) + A_{skin} \times P(n|S_{colour}), \end{aligned} \tag{5}$$

$$C_n(x,y) = \frac{c(x,y,n)}{A_{torso}(x,y) + A_{legs}(x,y) + A_{skin}(x,y)}, \tag{6}$$

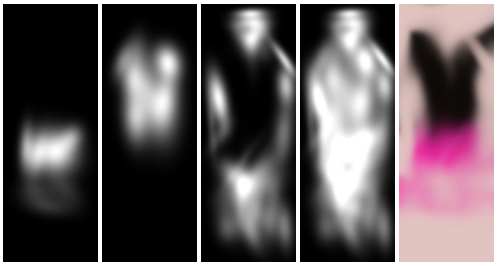
where $C_n(x,y)$ is a pixel, x, y, in the nth channel (i.e. representing the nth colour) of the channel representation; T_{colour}, L_{colour} and S_{colour} are the selected torso, leg and skin colours; $P(n|T_{colour})$ is the likelihood of colour n being the target colour, T_{colour}, and is estimated using the earlier computed confusion matrix. Note that when generating the

channel representation we normalise using the sum of the masks (i.e. $A_{torso}(x,y) + A_{legs}(x,y) + A_{skin}(x,y)$) to ensure that it sums to 1. This process will generate a mask for each colour, with the mask's content determined by the likelihood of the colour being observed at each pixel. This likelihood is driven by two factors: the colour in the target query, and the confusability of that colour.

Finally, a combined mask that represents the likelihood of each pixel belonging to the person is created,

$$A_P = min(1.0, A_s + A_{tc,T_{type}} + A_{lc,L_{type}}), \quad (7)$$

where min is an operator that takes the minimum of two values. Importantly in this combined mask, we do not use the normalised masks for the components, as we wish to ensure that pixels that are unlikely to contain the target have a low weight. This mask is used as an auxiliary channel in the CR to weight the contribution of each pixel to the overall similarity. An example avatar is shown in Figure 1.



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

Fig. 1. An Example Avatar for the query 'Tall, pink shorts, black short sleeve shirt': (a), (b) and (c) show the masks for the leg clothing, torso clothing and skin regions respectively. These masks are generated based on the type of clothing selected, and are combined to create (d) and (e). (d) shows the overall occupancy mask, and (e) shows the expected colour at each location. Note that the most likely colour for every pixel is given in (e), even though some pixels are very unlikely to belong to the person as shown by (d).

3.2. Searching for a Target

A particle filter-like approach is used to search for the target. A small number of particles are created at random locations within the field of view, and the single object tracking approach of [7] is used to refine the location of each. The similarity of each refined particle location to the target model and a stored background model is computed, based on which the particle set is refreshed and the current position of the target is estimated.

We first transform input images into a channel representation where the channels are given by the culture colour space,

$$I_n(x,y) = P(n|I(x,y)), \quad (8)$$

where I_n is the nth channel of the channel representation; $P(n|I(x,y))$ is the likelihood of the pixel $I(x,y)$ being the nth colour, and is computed using a lookup table that contains the likelihoods for the whole colour space computed using the learned GMMs [1]. As per [7], each channel is then convolved with a Gaussian filter.

A gradient descent search is then performed to find the particle location (X_p, Y_p) within a local region that minimises

$$S_F = \sum_{n,x,y}^{N,X,Y} |C_n(X_p + x, Y_p + y) - I_n(x,y)| \times A_P(x,y), \quad (9)$$

where S_F is the similarity of the foreground to the template; $C_n(x,y)$ is the nth channel of the template (i.e. search query) and X and Y are the width and height of the template, and N is the number of channels. Note that A_P is used to weight the summation such that pixels that are more likely to belong to the person are given a greater weight. Although this search will converge on an optimal location, it is only guaranteed to be locally optimal [16]. To overcome this, multiple particles are refined using this search process, after-which particles are filtered in preparation for the next input frame.

As the particles move during the search process, we do not require dense collections of particles around target locations as is the case for a particle filter. We use a two stage filtering process, where first particles which record a poor match are removed, after-which particles which are located nearby another are removed.

One of the limitations observed in systems such as [6, 15] is that locations in the background can be well matched to the target. To overcome this we take inspiration from the notion of universal background models as used in biometrics. We store an average channel representation for the background which is progressively updated every frame,

$$B_n(x,y,t) = \alpha \times B_n(x,y,t-1) + (1-\alpha) \times I_n(x,y), \quad (10)$$

where $B_n(x,y,t)$ is the n channel in the background channel representation at time t; and α is the learning rate. Using this stored background model, we compute the similarity of the template to the target location in the background,

$$S_B = \sum_{n,x,y}^{N,X,Y} |B_n(X_p + x, Y_p + y) - T_n(x,y)| \times A_P(x,y), \quad (11)$$

and use this similarity to determine if the particle is a better match to the background or the foreground,

$$S_R = S_B / S_F. \quad (12)$$

If $S_R \leq 1$, the particle is discarded, as it is more like the background than the foreground. Remaining particles are then filtered such that if two particles lie within a distance, d, the particle with the lower value of S_R is removed. The location of the target subject is given by the particle that yields the highest value of S_R.

[1] The lookup table is pre-computed for computational efficiency

3.3. Compensating for Scale

An additional parameter for the search query is the target height. To incorporate this and to account for perspective distortion with the images, we utilise camera calibration information. We use Tsai's camera calibration [17], which is provided with the dataset of [15]. A resolution parameter, R, in pixels per meter, is used to generate the template and resized target patches when searching. The target template is set to a height (in pixels) of $R \times H$, where H is the target query height in meters. When comparing the template to an image, the local region around the initial estimated position (i.e. X_p, Y_p) is resized to the same resolution.

4. EVALUATION

4.1. Database and Evaluation Protocol

We use the database and evaluation protocol outlined in [15]. This database consists of 110 video clips with semantic queries and the correct match annotated at every frame. As in [15], we evaluate our proposed approach by measuring the localisation accuracy using the ratio of the logical AND of the detected and ground truth regions over the local OR of the same regions. Avatar models are trained using a set of person images extracted from the background videos in [15] (i.e. they do not overlap with the queries). In total, we use images of 103 people from the same 6 camera network[2]. An example of the annotation is shown in Figure 2.

(a) Hair (b) Clothing (Legs) (c) Clothing (Torso) (d) Skin

Fig. 2. Example of the annotation of a subject into their constituent parts. Note that in the skin mask we annotate the head (red), arms (blue) and legs (green) separately.

4.2. Results

Performance of the proposed approach is shown in Table 1 for a number of parameter settings. By default, we set the number of particles to 20, the resolution, R, to 20 (i.e. 1 meter is scaled to 20 pixels), and Gaussian blur kernel width and standard deviation to 3 and 1.5 respectively. Table 1 shows the effect of varying each of these parameters individually, and it is clear that with the exception of extreme values (i.e. very low resolution or particle numbers, no Gaussian blurring), performance is consistent.

Comparing the proposed approach to that of [15], the proposed approach achieves state-of-the-art results with a relative performance improvement of 31% (a best average localisation of 0.38 compared to 0.29 for [15]). Figure 3 shows the number of sequences that achieve above a given threshold and it can be clearly seen that the proposed approach is able to better locate the majority of queries. For instance, for the proposed approach 56% of queries are located with an average accuracy greater than 0.4, while only 32% can be located with the same level of accuracy for [15].

[2]Please contact the authors for this annotated data

Parameter	Value	Average Localisation
Number of Particles	5	0.30
	10	0.35
	20	0.37
	40	0.37
Resolution (R)	5	0.25
	10	0.36
	20	0.37
	40	0.34
Gaussian Blur Kernel (Width, Std. Dev.)	N/A	0.26
	2,1	0.35
	3,1.5	0.37
	7,3	0.38

Table 1. Performance of the proposed approach when parameters are varied.

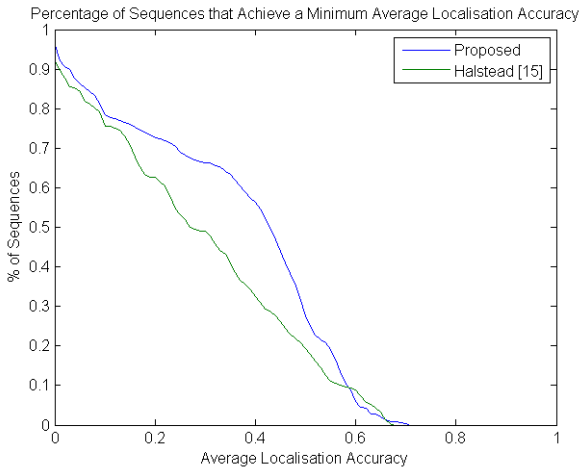


Fig. 3. The percentage of sequences above a given accuracy level.

5. CONCLUSIONS

In this paper we have proposed a new approach to locate a person in video from a semantic query. This approach generates a searchable query in the form of a channel representation, which can then be used by a gradient descent process to locate the target subject; and achieves state-of-the-art performance on the dataset of [15]. Future work will consider how other traits such as texture and luggage can be incorporated into this process, along with ways to incorporate pose and multiple views to further improve performance.

6. REFERENCES

[1] T. McKelvey and K. Dailey, "Boston marathon bombings: How notorious bombers got caught," *BBC News Magazine*, 2013.

[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. 38, 2011.

[3] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *International Conference on Biometric Authentication*, 2004, pp. 717–738.

[4] Naman Turakhia and D Parikh, "Attribute dominance: What pops out," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1225–1232.

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

[6] Simon 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 *Digital Image Computing Techniques and Applications (DICTA)*, 2012, pp. 1–8.

[7] Michael Felsberg, "Enhanced distribution field tracking using channel representations," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*. IEEE, 2013, pp. 121–128.

[8] J. Thornton, J. Baran-Gale, D. Butler, M. Chan, and H. Zwahlen, "Person attribute search for large-area video surveillance," in *IEEE International Conference on Technologies for Homeland Security (HST)*, Nov 2011, pp. 55–61.

[9] 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.

[10] Rogerio Feris, Behjat Siddiquie, Yun Zhai, James Petterson, Lisa Brown, and Sharath Pankanti, "Attribute-based vehicle search in crowded surveillance videos," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 2011, p. 18.

[11] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *In CVPR*, 2005, pp. 886–893.

[12] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008.

[13] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel, "Efficient pedestrian detection by directly optimize the partial area under the roc curve," in *ICCV*, 2013, pp. 1057–1064.

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

[15] M. Halstead, S. Denman, C. Fookes, and S Sridharan, "Locating people in video from semantic descriptions: A new database and approach," in *ICPR*, 2014.

[16] Laura Sevilla-Lara and Erik Learned-Miller, "Distribution fields for tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1910–1917.

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