J. Vis. Commun. Image R. 58 (2019) 439–452

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Multimodal clothing recognition for semantic search in unconstrained surveillance imagery ☆

Michael A. Halstead [a,*], Simon Denman [a,*], Sridha Sridharan [a], YingLi Tian [b], Clinton Fookes [a]

[a] Queensland University of Technology, 2 George Street, Brisbane 4000, Australia
[b] The City College, City University of New York, New York, NY 10031, United States

ARTICLE INFO

Article history:
Received 18 December 2017
Revised 12 July 2018
Accepted 2 December 2018
Available online 7 December 2018

Keywords:
Soft biometrics
Dempster-Shafer theory
Surveillance
Semantic person search

ABSTRACT

To date, surveillance based person search has focused on locating a person of interest from an image query, distinct from the law enforcement task of locating a person from a description.

In this paper, we introduce a novel probabilistic framework that combines multiple traits whilst incorporating their uncertainty to tackle the emerging challenge: locating a person from a semantic query. In addressing this, we improve clothing texture recognition by leveraging Dempster-Shafer theory against an ensemble of support vector machines; achieving state-of-the-art performance for high and low resolution clothing textures.

Our proposed person search framework combines information from clothing texture and colour in the torso and leg regions to produce a probabilistic match between unknown subjects and the designated target query. Results are presented on a newly created 520 subject surveillance dataset which is made available to researchers. This multi-modal person search technique achieves promising results for locating target subjects, without the requirement of pre-search target enrollment.

1. Introduction

In surveillance and security applications it is often necessary to search for and locate a subject of interest from a textually supplied subject description. At present, searches using these forms of queries are predominantly performed manually and often ineffectively, via physical searching of a premise or watching countless hours of video footage.

While the use of semantic descriptions for person search has become more prevalent in recent research [14,8], enrollment-based re-identification methods continue to dominate the research [7,10,6,1]. These methods exclude however, instances where the subject has not previously been located and enrolled from the footage. The ability to transfer a human describable query into a genuine image/video search would be invaluable to multiple domains and applications including query-based image retrieval, clothing fashion recognition, post-event video analysis, and surveillance and security.

In search tasks based on soft biometrics [15,26], clothing attributes are a commonly used search due in part to their dominance of an individuals outward appearance (see Fig. 1), leading to the ease in which it can be detected, and relevant information extracted by a human operator. The common use of these traits becomes further evident when viewing *offender description forms*,[1] where it should be noted that the possible trait descriptors are constrained to a subset of those humanly describable.

In this paper we introduce a new method for locating individuals of interest in surveillance based images from the following subset of soft biometric descriptors: clothing colour, and clothing pattern. The proposed approach utilises a multi stage classification technique where textually supplied target queries (i.e. torso clothing colour, leg clothing colour, torso texture, leg texture) are used as search parameters. Each of the modalities, in the supplied textual query, are constrained to a subset of those humanly describable, similar to that seen on a *offender description form*. We then combine clothing colour and texture; recognised using a pixel-wise Gaussian mixture model (GMM) classification technique and a novel approach utilising an ensemble of support vector machine (SVM) classifiers respectively. Finally, the individual modality scores are fused using the independent method of combination contained in

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding authors.
E-mail addresses: m.halstead@qut.edu.au (M.A. Halstead), s.denman@qut.edu.au (S. Denman), s.sridharan@qut.edu.au (S. Sridharan), ytian@ccny.cuny.edu (Y. Tian), c.foookes@qut.edu.au (C. Fookes).

[1] https://www.ems.edu.au/images/pdf/Offender-Identification-Form-V-1.02.pdf.

https://doi.org/10.1016/j.jvcir.2018.12.001
1047-3203/

Fig. 1. Comparison of images from low resolution surveillance type situations ((a)-(c)), with higher resolution fashion or photography based datasets ((d)-(f)).

Dempster-Shafer theory (DST). This method of trait fusion creates a probabilistic output, providing a probabilistic score to the initial target query.

Performance of the proposed texture descriptor is evaluated on high and low resolution images and compared to the state-of-the-art, where we achieve new state-of-the-art results at both resolutions. Evaluation of the person search framework is completed on a newly created surveillance dataset. To the best of our knowledge, evaluation from semantic descriptions achieves state-of-the-art performance in this challenging domain.

Our novel contributions in this paper are as follows:

- Clothing texture classification in both high and low resolution domains using four hand selected features. Classification is performed using an ensemble of SVMs with DST classifiers.
- Novel calculation of the uncertainty term contained in the DST classifiers is used to discriminate against poor performing models (low cross validation accuracy), and non distinct classification results. Feature fusion incorporates the use of the DST *method of independent combination* to produce state-of-the-art results.
- Inclusion of the clothing texture classification technique into a surveillance framework, evaluated on a newly created dataset consisting of 520 subjects with full pixel-wise annotation for coarse body regions. This dataset, which is publicly available to researchers,[2] was created to overcome limitations within existing datasets, such as occlusions, lacking trait availability, and the requirement of first running clothing/person parsing techniques which can introduce other errors due to poor segmentation. Incorporation of this extra trait (clothing texture) in semantic search provides greater utility in a surveillance framework as existing approaches have predominantly concentrated on clothing colour alone.

The remainder of this paper is structured as follows: Section 2 provides an overview of the prior work related to this task; Section 3 outlines the proposed approach including, feature mapping, classification and use of DST for fusion; Section 4 describes the experimental setup and results are outlined in Section 5. Finally, Section 6 concludes the paper and outlines future work.

2. Background and related work

Soft biometrics as a concept has existed since Bertillon [3] considered a method for criminal identification in the late 19*th* century. His work persists today in the form of criminal "mug shots". After being largely superseded by traditional biometrics due to their limitations, lacking permanence and discriminatory ability [15,6]; they have recently found renewed interest in computer vision research.

Jain et al. [15] re-introduced soft biometrics to offset traditional biometrics due to the ease in which they could be extracted (i.e. eye colour with retina scan). Categorising body measurements into semantically describable groups, the work by Samangooei et al. [26] began to outline their strength in multimodal forms. In this short period of time, soft biometrics have gained much traction in the research community where they are showing promise in various areas including: fashion parsing [32], content based image retrieval [16], surveillance [8], shopping applications [20], simplistic replication of human verbal descriptions of images [18], and recently as an image labeling tool for surveillance with a deep neural network (DNN) [24].

From a security standpoint, closing the "semantic gap" is an essential step towards creating usable and accurate soft biometric localisation techniques. To this goal, Reid et al. [25] outlined the challenges associated with the subjective manner by which humans identify and describe soft biometric traits, and the link to associated computer vision trait extraction procedures. This concept is directly transferable to a surveillance framework and the requirement of sufficient and effective novel traits, to create a distinct target signature. However, as seen in law enforcement practices such as the *offender description forms*, description options may be limited in depth to aid in closing the semantic gap.

From a computer vision and surveillance context, *person re-identification* [7,10,6,1] and *person search* (semantic/attribute search) [29,12,9,11] are two distinct methods of locating a subject of interest from (one or more) soft biometric traits. Reid et al. [25] outlined the differences between these similar techniques: re-identification directly interrogates a supplied image for feature extraction, while semantic search obtains the trait values through verbal or textual descriptions (i.e. no direct image interrogation).

2.1. Semantic search

The search for an individual of interest without the use of direct scene interrogation for soft biometric signature construction can be considered a constrained form of scene labelling. From an extensive list of possible traits, Samangooei et al. [26] declared that age and gender are two of the most easily identified and described traits by human witnesses. Extraction of the required information for accurate classification of these traits in low resolution images (Fig. 1 ($a-c$)) can be a difficult proposition. By dominating the outward appearance of a subject, clothing is also a trait which is commonly identifiable and simplistically relayed between operators with relative invariance to image resolution (see Fig. 1).

Early semantic search approaches [29,23] used simple colour descriptors for regions such as the torso and legs; while Vaquero et al. [29] also combined these with facial features. Creating a

[2] Data is publicly available on request from the authors.

probabilistic search technique for locating individuals of interest, Thornton et al. [28] proposed the use of clothing attributes (categorised colour) and luggage, where the clothing area is extracted using a moving pedestrian algorithm and image chips. Using heuristically segmented regions of interest Halstead et al. [12] calculate the dominant colour (using culture colours [2]) in torso and leg region using GMMs, and allow for different clothing types in the leg region by using asymmetry driven methods adapted from Farenzena et al. [10] to detect clothing bounds. Improving on this, Denman et al. [9] introduced a channel based approach to attribute detection, that used learned appearance templates for each trait to build a searchable query.

Satta et al. [27] blurred the lines between re-identification and textual search by using model based learning and dissimilarity-based appearance descriptors in a re-identification framework. Primarily utilising facial tracking and soft biometric facial features, Feris et al. [11] show promise across various datasets for detecting subjects of interest using semantically describable queries. Torso and leg regions were segmented using bounded regions guided by the face detection and can detect 13 different colours, providing ranked outputs using similarity. Interestingly they proved their search framework on a "real world" application, the unfortunate Boston Marathon Bombing event.

Generally using segmented bounding boxes, convolutional neural network (CNN) based techniques such as Zhu et al. [34] have outlined the benefits of using a single trained classifier to detect multiple soft biometric traits such as clothing and gender. However, one of the benefits of individual soft biometric traits is that they remain computationally inexpensive. The advantage of this is that traits can potentially be added to existing techniques in a modular fashion. For instance, the addition of gender or age traits using temporal sequencing and Gabor filters [13]. Where existing techniques classify clothing colour, clothing texture could be incorporated by exploiting existing approaches in alternate fields such as those outlined by Yang et al. [33] for assisted vision, or correlation filters [30] used in face recognition.

However, for all of the advancement in semantic person search techniques many limitations persist, including the heuristic measures of obtaining the bounded regions by which the traits are calculated, and the basic fusion techniques used to produce the final similarity calculations. Finally, in all of these techniques one major limitation is witnessed, and that is the lack of measured traits, providing a limited soft biometric signature to match against.

One solution to heuristically obtaining bounded regions is the use of clothing detection and parsing. Detecting clothing and clothing styles from images is an important component of a fully operational person search technique using attribute based detection. With accurate clothing detection not only is the clothing type itself able to be included in the search modalities, but more accurate detection of the clothing minutia is possible, such as colour and texture.

An early technique to parse clothing, in a fashion sense, was that of Yamaguchi et al. [31], where they were able to parse high resolution fashion images into 53 clothing labels. The garments were parsed using conditional random fields, exploiting the relationship between clothing (and super-pixels) and body pose. To improve the performance of fashion parsing Yamaguchi et al. [32] introduced a retrieval based approach using templates that fit the clothing attributes. Their approach combined pre-trained global clothing models, local clothing models learned dynamically, and transferred parse-masks. To parse the image, similar to their earlier approach, they rely on an estimated pose mask of the subject with features extracted around each point. Tag prediction is based on K-nearest neighbours with a KD-tree to index samples, and iterative label smoothing to produce the final parsed results.

Also relying on pose, Chen et al. [5] proposed a novel clothing attribute detector that uses images with people in unconstrained settings. Features including SIFT, maximum response filter, CieLab colour information, and skin probabilities which are extracted based on the detected pose. SVM classification is used with conditional random fields to parse twenty-three binary attributes and three multi-class attributes.

Using a deep CNN, Liang et al. [19] proposed a contextualised CNN which produced pixelwise attribute labels in an end to end fashion. Their proposed network comprised five main components: local-global–local, global image-level content, semantic edge content, local superpixel content, and attribute output. Their technique was able to predict the presence of eighteen attributes, output on a full resolution image containing the pixelwise annotation.

In each of the above mentioned techniques one major issue arises when applied to low resolution surveillance imagery, they are all trained and evaluated on high resolution fashion style images. While techniques do mention the ability to perform "across domain" this does not extend to the very low resolution surveillance domain. For person search in surveillance settings, techniques that are able to perform in lower resolution are required.

One technique which aimed to reduce this problem was the deep decompositional network described by Luo et al. [21]. This approach reduced the fashion based problem down to a coarse parsing problem where semantic regions including: hair, head, torso, legs, and arms. Based on a HOG representation of the pedestrian image, they create the parsed representation using three network components: occlusion estimation, completion, and decomposition layers. The network then produces a set of masks that segment the image into the desired regions.

Clothing parsing has progressed significantly in a short period of time and shows promising results. While techniques exist to parse a low resolution image, in general techniques have concentrated on high resolution images to reduce image noise due to the reliance on fine grained details to ascertain the appropriate class.

Unfortunately these techniques are still erroneous when segmenting images which introduces further noise to a surveillance image. While potentially not being as severe as errors introduced by heuristic or static bounded regions, these errors in segmentation can be detrimental to the performance of subsequent processes (i.e. colour detection in a region).

3. Approach

In this section we outline a probabilistic framework to locate subjects of interest in surveillance imagery. Fig. 2 represents the pipeline used in the semantic search of target subjects resulting in their similarity to a supplied target query.

In Section 3.1, the generation of a target query is composed from cues that capture semantically describable clothing attributes: colour, and pattern. The process for clothing attribute recognition into the semantic traits is outlined in Section 3.2. Finally, the use of DST for the fusion of features and traits is outlined in Section 3.3.

3.1. Query generation

Regardless of camera resolution, clothing dominates the outward human appearance and as such clothing traits will comprise the target query. The colour and texture queries for both the leg and torso regions are selected from the descriptors outlined in Table 1.

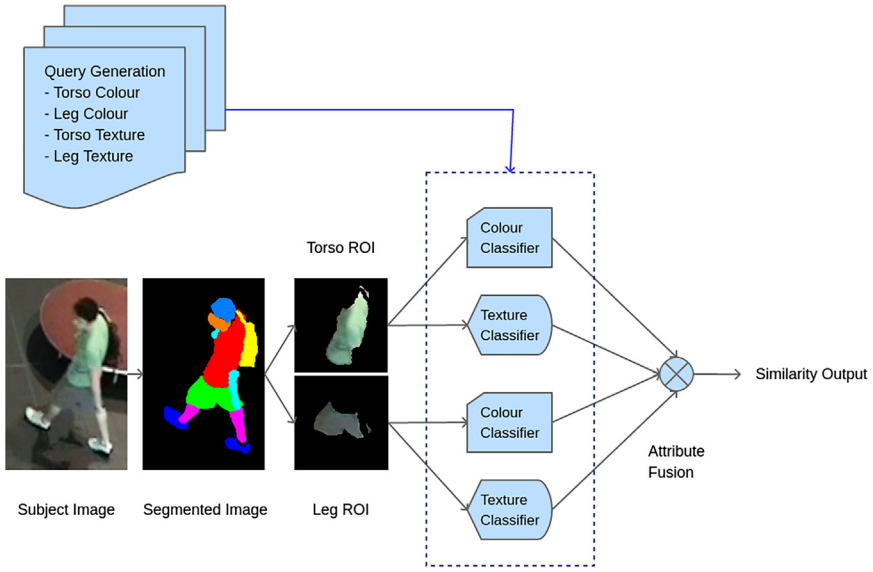


Fig. 2. Overall system pipeline, initially a target query is designated from semantically describable attributes. The target image with associated segmentation mask is used to create the regions of interest within the image. Based on these ROIs and the query classification for texture (expanded in Fig. 3) and colour is completed. The texture and colour information is then combined using DST based techniques to supply a final similarity score to the query. Note that in this paper we consider only hand segmented semantic masks to remove any errors associated with automated semantic parsing approaches, however these could be used in place of this.

Table 1
Clothing attributes for target query generation.

Attribute	Categories
Colour	Black, Brown, Blue, Gray, Green, Orange, Pink, Purple, Red, White, Yellow
Pattern	Irregular, Plain, Plaid, Diagonal Plaid, Spots, Diagonal Stripes, Horizontal Stripes, Vertical Stripes

Selected colours are taken from the culture colours of Berlin and Kay [2], and shown in Table 1. For texture we use 8 textures that we believe best encompassed the majority of available clothing textures, while also expanding on the textures used by Yang et al. [33].

3.2. Attribute classification

The attribute classification component probabilistically matches a region of interest (ROI) to the clothing cues described in Section 3.1. For classification of the texture snippets, the entire snippet is considered the ROI as no occlusions or deformities are present in the dataset.

To classify pedestrians in surveillance footage the masks designating the ROI are more complex. These masks need to allow for both the dynamic nature of pedestrians and the body component to be classified (i.e. torso and leg region). The query based classification technique undertaken here helps to constrain this problem. While there exist techniques to automatically segment masks for clothing attributes in both high and low resolution settings, all of these introduce additional errors into the pipeline. These can be associated with incorrect pixel labels which can act to reduce the performance of the attribute classifiers. Similarly, the expansive nature of the attributes being segmented with existing parsing techniques are unnecessary for this task. The attribute classification component here only relies on the presence of the torso clothing mask and leg clothing mask as holistic components, in this evaluation we do not require that the torso region be broken down into multiple articles of clothing (i.e. shirt, vest, jacket).

To ensure that the attribute classifiers themselves are being evaluated without possible biases caused by a parsing technique, all the pedestrians in the dataset are hand segmented into pixel-wise masks. An example of these masks is shown in Fig. 8 where we only consider the torso clothing an leg clothing masks. When annotating a region of clothing, if there are multiple items of clothing within the region, all articles were combined into one such that an open jacket with a shirt underneath becomes a single mask component.

3.2.1. Colour attribute

Colour classification uses 11 colours (see Table 1), modelled using GMMs trained on ancillary colour data supplied in Halstead et al. [12]. We train our GMMs in the cieLab colour space using the expectation maximisation algorithm and a maximum of 3 mixtures selected using the Bayesian information criterion, as it provided the most consistent results when compared against the colour snippets reserved for testing.

The similarity score for the ROI for the selected colour, c_i, of the i_{th} trait is,

$$S_{c_i} = \sum_{I \subseteq ROI} \frac{G_c(I)}{ROI}, \tag{1}$$

where S is the similarity score of the current target image (I) over the ROI. The probability map, G, is scaled from the sum of all densities in the GMMs, normalised to 1 such that,

$$G_c = \frac{GMM_c}{\sum_{j=1}^{C} GMM_j}, \tag{2}$$

where C is the total number of colour GMM's (11).

3.2.2. Texture attribute

The texture component comprises a three phase classification process (see Fig. 3), with the initial probabilistic classification being performed by an ensemble of SVMs. Using the extended texture dataset outlined in Section 4.1 we train our SVMs using libsvm [4] with a radial basis function for feature mapping. We use 5−fold cross validation, with parameter selection (C and γ) completed using an exponential grid methodology between the ranges $[-40, 40]$, with probability mapping enabled. The training of our SVMs differs from Yang et al. [33], where they use a single SVM

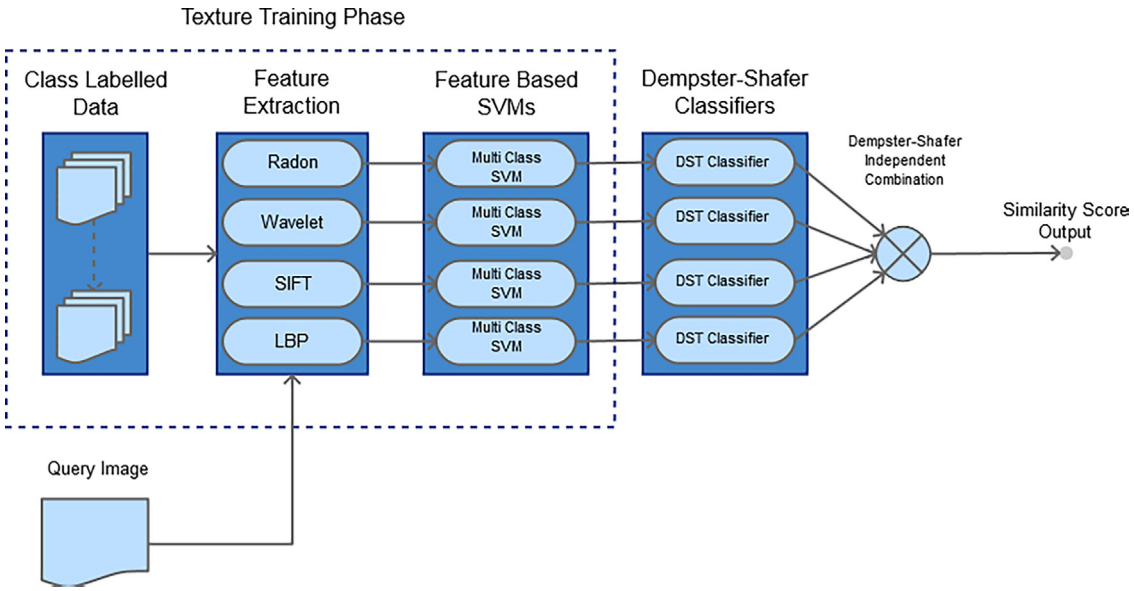


Fig. 3. The Texture classifier comprises 3 layers. An image has features extracted using the Radon transform, Wavelet sub-band statistics, dense SIFT descriptor, and local binary pattern. During training features are used to train individual SVMs, while during evaluation the features are fed to the SVMs for probabilistic classification. The independent SVM results are fed forward to individual DST classifiers, and the cross validation accuracy of each SVM is used as an input to the DST classifiers. The final stage uses the DST method of independent combination to fuse the outputs producing a final similarity score.

with all features concatenated into a single vector. In our novel technique we compute a SVM for each feature, creating an ensemble of classifiers. This ensemble methodology is adopted to provide discriminatory ability against the individual features performance in the DST classifiers.

The Radon transform (RT) signature (the variance of the RT) provides global rotational features based on the Sobel edge profile. Similarly providing global features, statistical measures including the variance (*Var*), energy (*Ene*), uniformity (*Uni*), and entropy (*Ent*) are calculated in each of the wavelet transform sub-bands (wavelet subband statistics - WSS), creating $16 \times L$ features where L is the depth of the sub bands. Yang et al. [33] provides an in depth look into the RT and WSS, and the same approach is adopted here.

SIFT and LBP are used to capture local properties. SIFT extraction follows Yang et al. [33] where a densely sampled SIFT descriptor is fed to a bag-of-words (BOW) classifier, reducing dimensionality. To emphasise more discriminative words, we weight the BOW output with a term frequency-inverse document frequency (tf-idf) methodology. In all evaluations we use 100 words in our BOW classifier, and perform $L2-$normalisation on the resulting feature vector. For the local binary pattern (LBP) of the texture snippet, the cell size is set to the size of the image snippet itself, producing a histogram of the local textures. Computing the quantized LBP of the image snippet produces a feature vector of length 58.

We train a multi-class SVM for each feature descriptor (see Fig. 3). Each SVM is trained with a mapped probability output and the cross validation accuracy is retained to provide classifier accuracy based on the supplied training data.

Finally, we create the DST classifiers (see Fig. 3). Following common DST notation, we define our *frame of discernment* (sample space), Θ, as a set of mutually exclusive hypothesis,

$$\Theta = [Gn, Im], \tag{3}$$

where *Gn* (genuine) is the desired class probability from the SVM, *Im* (imposter) is the summed probability of the remaining classes for each of the independent features (f),

$$Gn = SVM_f(T), \tag{4}$$

$$Im = \sum_{t \notin T} SVM_f(t), \tag{5}$$

and T is the desired texture. The power set (2^{Θ}) is then used to map all possible combinations of Θ and the empty set ($\emptyset$),

$$2^{\Theta} = [\emptyset, Gn, Im, \Theta], \tag{6}$$

defining the inclusion of the *frame of discernment* (Θ) in 2^{Θ}. Θ can be thought of as a subjective term that maps the uncertainty within the classifier itself. The strength of DST is the ability to measure uncertainty within the data.

Next, the mapping of the entire power set is required, where degrees of belief, $m(A_i) : [0, 1]$, are assigned by a *Basic Assignment Function* such that,

$$\sum \left(m(A) | A \subseteq 2^{\Theta} \right) = 1, \tag{7}$$

$$m(\emptyset) = 0, \tag{8}$$

ensuring the power set sums to one. It can be seen that as the uncertainty (Θ) decreases to zero, DST reverts to Baysian probability theory. One of DST's strengths is the assignment of the *measure of belief*, denoted $bel(A_i)$; and the *measure of plausibility*, denoted as $pl(A_i)$, where $bel(A_i) \leqslant pl(A_i)$, further illustrated in Fig. 4. The belief of A_i is the sum of the *basic assignment functions* of all subsets of A_i (in our case *Gn* and *Im* maintain their values),

$$bel(A_i) = \sum (m(B) | B \subseteq A_i), \tag{9}$$

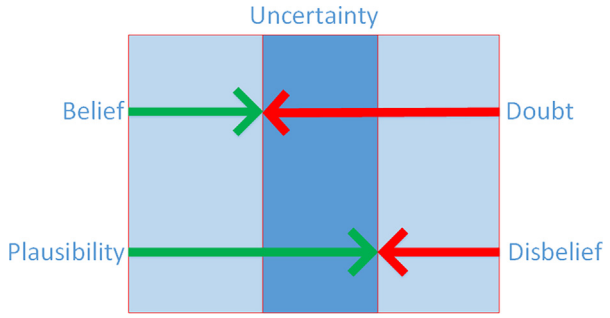


Fig. 4. Dempster-Shafer mapping of the full sample space, including the uncertainty effect on belief and plausibility [17].

where B represents each of the subsets of A_i. The *measure of plausibility* is calculated similarly using either of,

$$pl(A_i) = \sum(m(B)|B \cap A_i) = 1 - \sum bel(\bar{A}_i). \quad (10)$$

This benefit of DST also presents one of its major challenges: the creation of the uncertainty term, and assigning values to the *Basic Assignment Function*. This assignment can prove challenging in computer vision problems, and Nguyen et al. [22] outline some of the problems witnessed in other applications of DST including situations where the uncertainty is set to zero, or the uncertainty is set to *Im*, with *Im* becomming 0.

To calculate our uncertainty term we exponentially scale a ratio of the SVM output, which is then further scaled by the SVM classifiers cross validation accuracy (*CV*). A ratio (R) is calculated as the ratio of the two highest probability scores from the current SVM regardless of class, providing a metric on the distinctiveness of classification in the current feature's SVM. Then exponential weighting is calculated,

$$\widetilde{U} = e^{\tau \times R}, \quad (11)$$

where τ is a weighting factor. For the evaluation of the texture classifier we linearly normalise $\widetilde{U}$ into the range $[0, 1]$, while in our localisation system we attempt to maintain a higher value of uncertainty by removing this normalisation. We then further weight the uncertainty using the cross validation accuracy,

$$U = (1 - CV) \times \widetilde{U}, \quad (12)$$

creating an uncertainty value that incorporates the distinctiveness of classification (R) and the validity of the model (*CV*). This discriminates against poor classifiers, allowing more dominant classifiers to maintain high scores. Finally, prior to DST calculation we perform $L1-$normalisation of the genuine, imposter and uncertainty components, such that,

$$U + Gn + Im = 1. \quad (13)$$

From Eq. (13), it can be seen how normalised or unnormalised uncertainty values will impact the masses of *Gn* and *Im*. Weighting with higher values of τ without subsequent normalisation between $[0, 1]$, and depending on the initial value of R, we would expect that the final values for the *Gn* and *Im* to be scaled to a smaller value. However, this gives way to the plausibility of the DST classifier to return greater values.

3.3. Fusion

Fusion is required at both a trait and system level, and at all levels we assume complete trait and classifier independence. In instances where the trait has multiple annotations attached, specifically during colour classification in the person search evaluation, a method of fusing the independent scores together is required. Using Fig. 4 as a reference, we perform trait level fusion of multiple attached annotations in the following manner,

$$A_{doubt} = \prod_{m=1}^{M} B_{doubt}(m), \quad (14)$$

$$A_{disbelief} = \prod_{m=1}^{M} B_{disbelief}(m), \quad (15)$$

$$A_{uncertainty} = A_{doubt} - A_{disbelief}, \quad (16)$$

where A is the fused DST classifier; and M is the number of DST classifiers, B, to be fused. Calculating the disbelief and doubt in this manner reduces their overall impact on the new classifier, increasing the expected values for belief and plausibility (as expected when fusing two accurate classifiers).

In both the texture classification and person localisation there is a requirement for accurate DST fusion: at feature level (RTSig, WSS, SIFT, LBP) for the texture descriptor; and at trait level (torso colour, torso texture, leg colour, leg texture) fusion for the localisation. In each of these the assumption of independence allows the use of the DST *Independent Method of Combination*, defined as,

$$K = 1 - \sum(m(A_i)|A_i \cap B_i = \emptyset) \times (m(B_i)|A_i \cap B_i \neq \emptyset), \quad (17)$$

$$m(C_i) = \frac{\sum m(A_i|A_i \cap B_i \neq \emptyset) \times m(B_i|A_i \cap B_i \neq \emptyset)}{K}, \quad (18)$$

where the new DST classifier C, is created from existing DST classifiers, A and B. This rule derives the *Basic Assignments* of the new variable by calculating the sum of the accumulated evidence in A_i and B_i and dividing by the sum of the conflict. At the feature level this creates a final DST classifier that incorporates the information of each of the 4 features. The trait level fusion incorporates information from all of the traits creating a final DST classifier, producing a final similarity belief (i.e. how similar the image is to the query).

4. Experimental setup

Two evaluations are contained in this paper: firstly, the proposed texture classification technique is compared to the state-of-the-art; and secondly, a person search framework evaluated to illustrate the performance of the texture and colour classifiers in a surveillance setting. The texture and person search datasets used in these experiments are outlined in Sections 4.1 and 4.2 respectivley; and experiment details and parameters are given in Sections 4.3 and 4.4 for the texture classification and person search tasks respectivley.

4.1. Texture dataset

We recast and extend the clothing texture dataset of Yang et al. [33] to include the additional traits: diagonal plaid, spots, diagonal stripes, horizontal stripes, and vertical stripes; giving a more comprehensive and challenging dataset. All data was collected from the internet via manual search, and inter and outer class variations were included.[3] An example of each trait, and the full breakdown of the dataset can be seen in Fig. 5.

Texture evaluation is performed across two versions of the dataset: at the original resolution (140×140 pixels) and at 25×25 pixel resolution. Use of the original resolution allows a comparison to Yang et al. [33], while the lower resolution snippets better reflect a surveillance environment. Four examples of the high and low resolution snippets are shown in Fig. 6, outlining the extra challenge expected from the lower resolution task.

4.2. Person search dataset

For the evaluation of our novel person search technique a newly developed dataset was required due to a number of factors including: the need for robust pixelwise annotation of subject appearance; avoidance of occlusion due to clothing and/or luggage; and a descriptive annotation of the clothing colour and texture for each subject.

This dataset contains 520 subjects (examples shown in Fig. 7) collected in a surveillance setting, with varied lighting and resolution. In addition to the clothing attribute traits, the gender (male or female), pose (front on, rear, 45°, and 90° to the camera), and clothing type (long or short) for the torso and leg regions are annotated.[4]

[3] Data is publicly available on request from the authors.
[4] Data is publicly available on request from the authors.

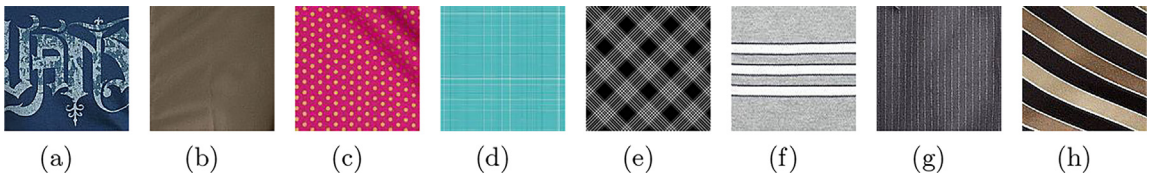


Fig. 5. An example of each of the 8 classes contained in the texture dataset, expanded from that of Yang et al. [33], and available upon request. (a) irregular pattern (156 in dataset), this class also includes all forms of logos, (b) patternless or plain (156 in dataset), (c) spots (156 in dataset), (d) plaid or check (159 in dataset), (e) diagonal plaid or check (157 in dataset), (f) horizontal stripes (154 in dataset), (g) vertical stripes (154 in dataset), and (h) diagonal stripes (154 in dataset). Note also the other variations within the data including the crease in (c) and the differing width of the stripes in (d)-(g).

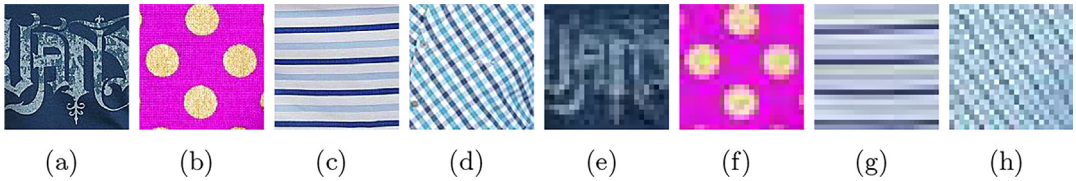


Fig. 6. Comparison examples between the two resolution types: (a) to (d) show high resolution for irregular, spots, horizontal stripes, and plaid. (e) to (h) show the corresponding patches in low resolution.

Fig. 7. Examples of four subjects from the newly created database. (a) Male wearing a short sleeve plain green top and short plain grey pants. (b) Female wearing a short sleeve plain black top and short plain white pants. (c) Male wearing a long sleeve plain grey top with long plain black pants. (d) Male wearing a short sleeve horizontal striped grey, blue, and white shirt with short plain brown pants. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For each article of clothing, the subject is annotated for a single texture based on those outlined in Section 3.1. Due to the inherent nature of some textures multiple colours are visible, to allow for this a maximum of three colours were annotated in each of the leg and torso regions. Finally, each subject is pixelwise annotated for the following regions: torso clothing, leg clothing, hair, shoes, luggage, arm skin, leg skin, facial skin. An example of a parsed subject with their respective binary masks is shown in Fig. 8. For the person search evaluation in this paper only the torso clothing and leg clothing regions are utilised. These masks were used to avoid introducing errors associated with automatic parsing techniques, and to better reflect the overall performance of the proposed approach.

4.3. Texture evaluation setup

The texture evaluation is performed across both high and low resolution snippets. To extract the information for the RT, the rotational properties of the transform are performed over the range: $\delta = [0 : \frac{\pi}{60} : \pi]$. These values of δ were selected to provide global rotational properties without saturating the information being extracted. The Sobel operator component of the RT requires a threshold for edge detection, and this is set to one half of the signal to noise ratio of each individual snippet. This threshold best

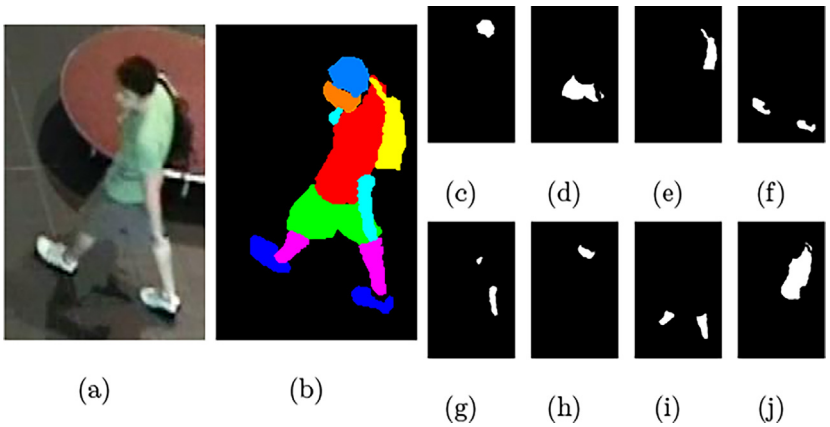


Fig. 8. An example of one of the subjects broken down into their binary masks. (a) The original image, (b) an example of the person parsed into regions, (c) represents the binary mask of the hair, (d) represents the binary mask of the leg clothing, (e) represents the binary mask of the luggage, (f) represents the binary mask of the shoes, (g) represents the binary mask of the skin in the arms region, (h) represents the binary mask of the skin in the head region, (i) represents the binary mask of the skin in the leg region, (j) represents the binary mask of the torso clothing.

reflected the results seen in Yang et al. [33], which were determined after empirical evaluation.

The WSS are constructed from a standard Haar wavelet transform, with a depth of $L = 3$ for both resolutions, similar to Yang et al. [33].

The SIFT descriptor varies between the high and low resolution domains. For the high resolution snippets the descriptor uses spatial resolutions of 4, 8, and 16, while at low resolution spatial resolutions of 2, 3, 4 are used. The disparity is due to the resolution of the image snippets, and the need to maintain SIFT as a local information extractor.

LBP requires no changes between the different resolutions. For both, the LBP is extracted in a 3×3 neighbourhood and then aggregated into a histogram based on the size of the current snippet, creating a histogram based on 58 uniformly quantized patterns.

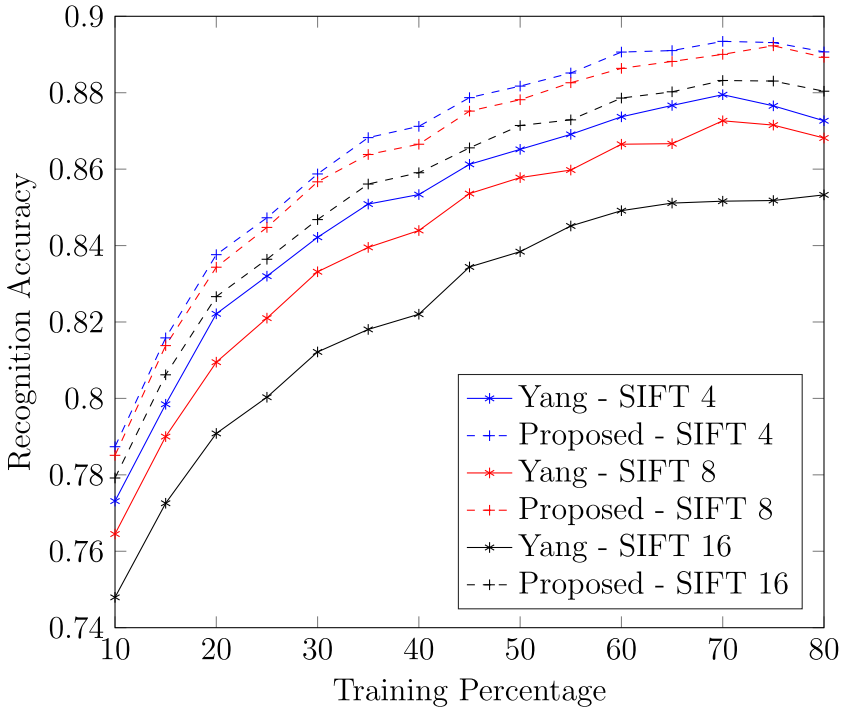


Fig. 9. Comparison between Yang et al. [33] and the proposed technique using the full range of descriptors: RT, WSS, SIFT, and LBP, as an ensemble of classifiers. Evaluation is computed over three SIFT resolution sizes (4, 8, and 16), and evaluated on the high resolution (140×140 pixels) image snippets.

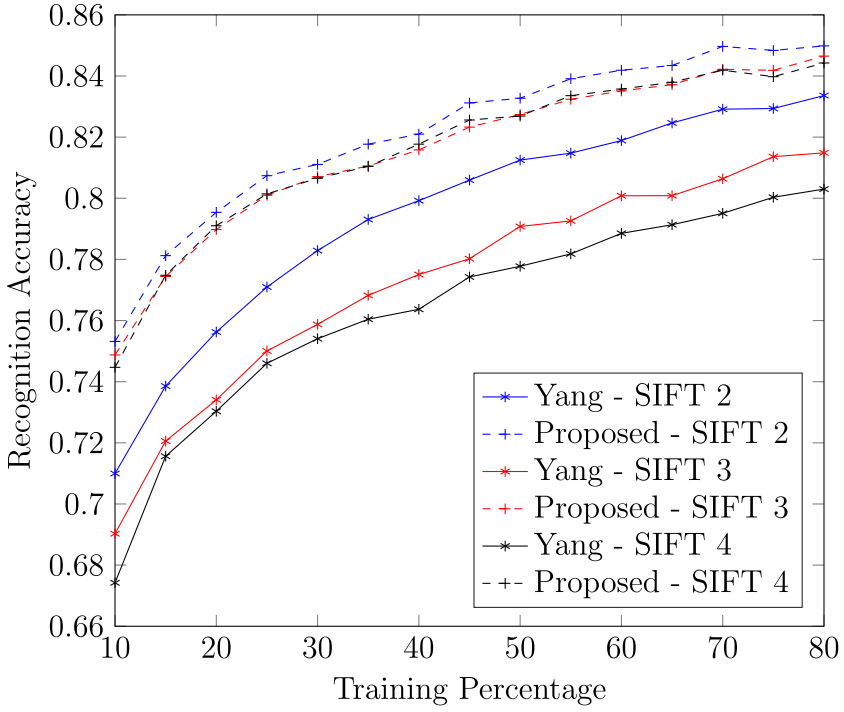


Fig. 10. Comparison between Yang et al. [33] and the proposed technique using the full range of descriptors: RT, WSS, SIFT, and LBP, as an ensemble of classifiers. Evaluation is computed over three SIFT resolution sizes (2, 3, and 4), and evaluated on the low resolution (25×25 pixels) image snippets.

Finally, to set up the DST classifiers in the texture classification framework, the accurate construction of the uncertainty term is critical to system performance. The primary parameter to be set is the weighting component τ. For both resolutions τ is set to 1.

Evaluation of the clothing texture classifier is completed in a similar manner to Yang et al. [33]: the training and testing split is computed over a range of training percentages, starting at 10% training data and increasing in 5% increments to 80%, creating 15 splits per resolution. At each split we create 50 random allocations of the image snippets to the training and testing sets. Presented results are an average measure of recognition accuracy over the entire training/testing set.

4.4. Person search evaluation setup

To calculate texture features for person localisation we need to allow for non-uniform ROIs obtained for the torso and leg clothing (this is of less concern for colour as it can be classified pixelwise). For texture, models are trained using all low resolution snippets. Each snippet contains 625 pixels, creating separation from the person search dataset as ROIs are distorted due to factors including pose and occlusion. To compensate, image patches are extracted based on the boundaries of the binary mask. Patches are padded with zeros to obtain a square patch of minimum size (25×25). Finally, the patch is resized such that the foreground pixels in the binary mask are approximately equivalent to the number of pixels in the texture snippets (i.e. ≈ 625 pixels).

Allowances are also made for individual features in this setting. RGB images are filtered using the mask, such that background pixels in the mask are set to the background value in the RGB image. This is problematic for the RT due to its' usage of the Sobel operator: the Sobel operator will detect an edge profile on the border of the foreground and background. As such, once the edge map is calculated we perform edge suppression between the foreground and

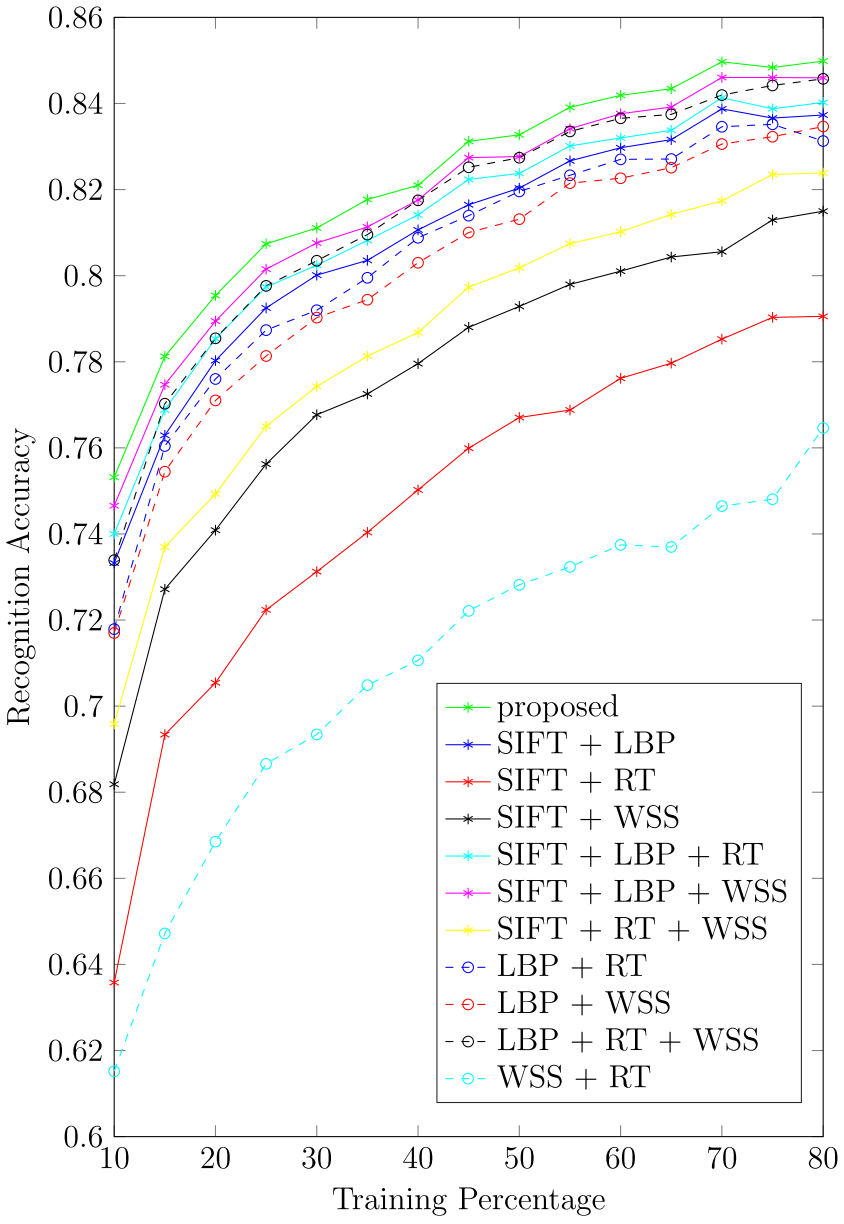


Fig. 11. Recognition accuracy comparison of the various permutations of the modalities as well as the fully proposed technique, the SIFT descriptor is evaluated with a spatial resolution of two and the WSS with a depth of three. From approximate weakest to strongest: RT and WSS; SIFT and RT; SIFT and WSS; SIFT, RT, and WSS; LBP and WSS; LBP and RT; SIFT and LBP; SIFT, LBP, and RT; LBP, RT, and WSS; SIFT, LBP, and WSS; proposed technique.

background, allowing the edge profile to be utilised with the RT component outlined in Section 3.2. The WSS also requires statistical allowances to ensure the obtained values maintain viability. We calculate the probability mass function over only the specified *ROI* (designated by the binary mask), such that the mean of the sub-band is calculated as:

$$\mu = \sum_{i=0}^{I-1} x_i \times P_{ROI}(x_i). \tag{19}$$

Both the densely sampled SIFT and LBP use the binary mask to aggregate features at valid locations. For both, if any of pixels in the current neighbourhood are background the neighbourhood is ignored. Extraction then proceeds as outlined in Section 3.2.

The final component of the person search is creating the DST classifiers for colour and texture. In Section 3.2 we outlined the general approach to uncertainty calculation. For colour classification, the exponential weighting with τ is removed, creating uncertainty values solely using the ratio of the top two probabilistic outputs. This approach is used as only the distinctiveness of classification was required to accurately model the subjectivity within the data. For the classification of texture and the subsequent calculation of uncertainty the exponential weighting is employed. τ was set to 10 to provide higher values of uncertainty within the data due to disparity in the classification distinctiveness.

To evaluate the person search approach, two different evaluation techniques are used. Initially, performance is evaluated using direct and independent comparisons of each subject's similarity to their personal target query. To produce the final similarity score, both the belief and plausibility components of the fused DST classifier are extracted. The belief information of a subject contains only the knowledge that directly matches the target subject to the query. Alternatively, the plausibility score has the benefit of including any overall uncertainty attached to the DST classifier. The second method emulates an surveillance identification task, where there is a single target query to be matched against a corpus. The set of target queries is limited to unique queries only, avoiding multiple evaluations on the same query. Each query is compared to the corpus to calculate a ranked output of the subjects that match. In each evaluations only fully annotated subjects are considered (at least one colour in the torso and leg region, and texture in each region). Omiting subjects with only partial annoation, we have a total of 509 out of 520 subjects available.

5. Results

The system evaluation is completed in two distinct parts, first we evaluate our novel texture classifier on the newly expanded texture dataset described in Section 4.1 and the parameters outlined in Section 4.3. The second evaluates the novel texture classifier and a colour classifier in a person search setting, using the dataset outlined in Section 4.2 and the parameter selection described in Section 4.4.

5.1. Texture evaluation

We evaluate our novel texture classifier on both full and low resolution image snippets. The full resolution (140×140 pixels) allows a direct comparison to the state-of-the-art Yang et al. [33]. As shown in Fig. 9, at a SIFT spatial resolution of 4, we are able to achieve a recognition accuracy of 89.07% compared to 87.27% for Yang et al. [33]. As can be seen, lower spatial resolutions of the dense SIFT descriptor offer superior performance due, in part, to the ability to provide a richer description of the local texture.

In Fig. 10 we compare the low resolution snippets performance to Yang et al. [33]. Once again we see the benefit of lower spatial resolutions of the SIFT classifier, providing a richer texture representation. At a training split of 80% we are able to outperform the state-of-the-art by 1.4% (84.75% to 83.36%), at a SIFT spatial resolution of 2.

Finally, the texture classifier was evaluated across all the possible permutations of the modalities using the lower resolution snippets. This evaluation stands as proof that the proposed technique requires all four modalities for optimal performance. From Fig. 11, while a number of the modes in combination compare favourably with the full technique, the complete combination of the RT signature, WSS, LBP, and SIFT descriptors are best able to classify the texture snippets at low resolution.

5.2. Person search evaluation

Performance of the proposed person search technique is evaluated using the techniques outlined in Section 4.4. In these evaluations there is no subject enrollment, as seen in person re-identification techniques, meaning that attribute classification requires the global models described in Section 3.2.

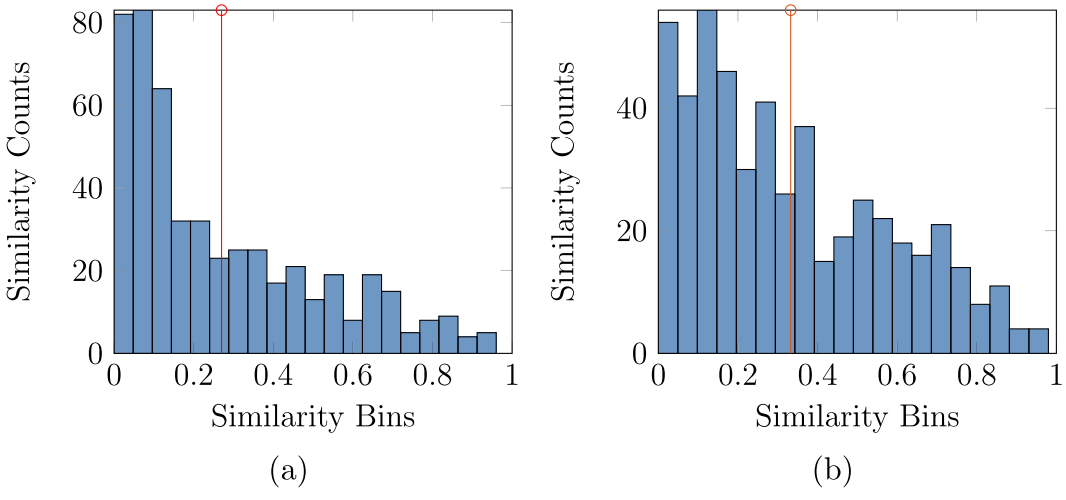
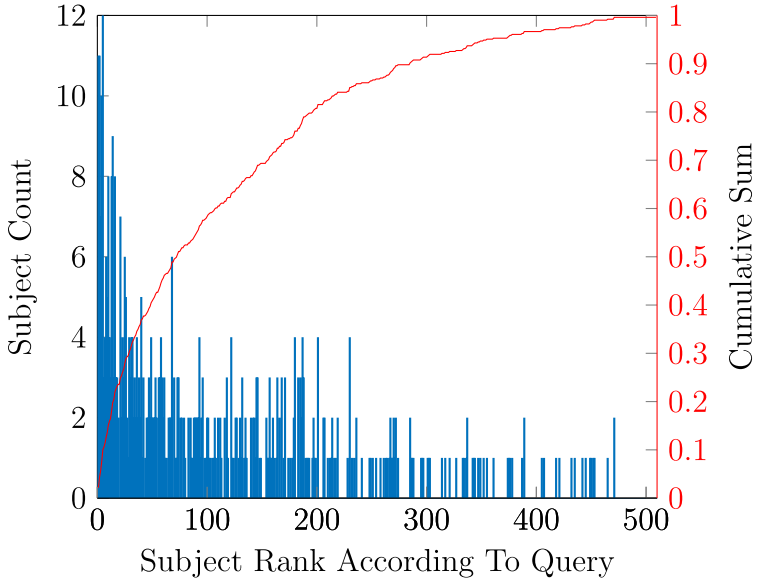


Fig. 12. Similarity histogram of the query to subject evaluation, where the red line indicates the mean similarity. (a) Outlines the belief of the Dempster-Shafer theory classifier with a mean similarity of 0.270, and (b) displays the plausibility of the Dempster-Shafer theory classifier with a mean similarity of 0.332.

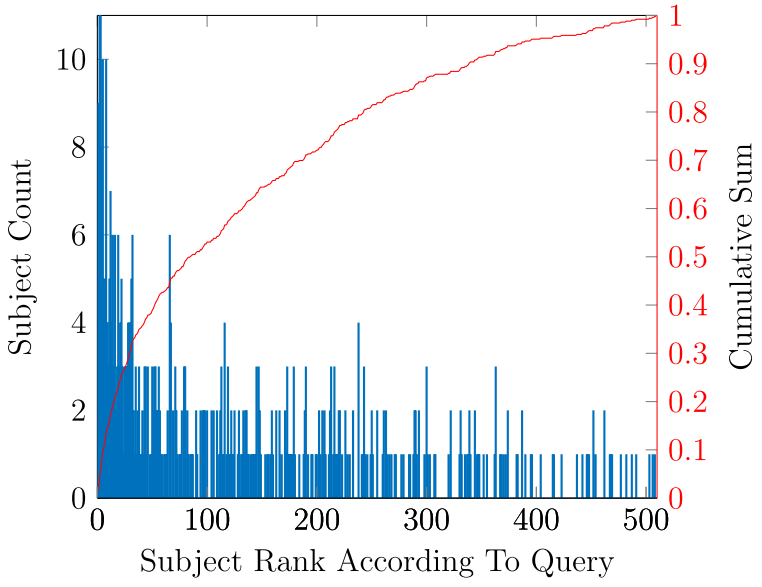
The first evaluation, subject to query, calculates the fused similarity score based on the torso clothing colour and texture, and the leg clothing colour and texture. The similarity metrics are based on the output from the fused DST classifier for each of the belief and plausibility properties. System performance for the subject to query is displayed in Fig. 12.

The results are displayed as a histogram of the similarity scores of both the "belief" and "plausibility" measures to their respective queries. Investigating the performance under this metric outlines the strength of the plausibility as a similarity metric. A mean similarity score of 0.332 is achived for the plausibility, compared to 0.270 for the belief, outlining that we achieve higher similarities between the subject and their query when using plausibility. This gain due to the larger value of τ being used without linear normalisation, creating larger values of uncertainty. As outlined in Section 3.2, higher values of uncertainty result in the scaling of the genuine and imposter variables, however, the plausibility remains unchanged due to its inherent use of the uncertainty component.

Many techniques, including traditional biometrics, may require definitive matching to a target query, and these results illustrate strengths of DST in this field. In a surveillance person search situation the plausibility contains the added benefit of incorporating



(a)



(b)

Fig. 13. Ranked performance of the subject location technique in a query to corpus setting, with a cumulative plot (red). (a) Outlines the performance of this technique as a measure of the ranked performance using the belief score from the DST classifier. (b) Outlines the performance of this technique as a measure of the ranked performance using the plausibility score from the DST classifier. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

all uncertainty contained within the query, trait, and classifier itself. Utilising this property provides a higher chance of detecting the desired target subject within the current scene.

However in surveillance and security situations it is rare that a target query would be compared to only a single subject. In such situations it is more likely that hundreds of potential matches will be compared to. The ranked performance of our query to corpus metric is displayed in Fig. 13, where each unique query is compared to all subjects. Both the belief and plausibility scores from the DST classifier are investigated to ascertain the best performing metric. In Fig. 13 it appears that both scores display equivalent performance. This similarity is further evidenced when considering the cumulative sum over various locations: At rank 1 we achieve a belief of 0.022 and a plausibility of 0.018; rank 20 achieves a belief of 0.236 and plausibility of 0.238; and at rank 100 we achieve a belief of 0.583 and a plausibility of 0.530.

In contrast to the subject to query evaluation where the plausibility outperformed belief, here the incorporation of uncertainty fails to have the same positive impact on the overall performance. This could be due, in part, to the fact that subjects are all being weighted through the use of the uncertainty, negating its overall impact.

In this difficult domain the proposed methodology is still able to produce the majority of matches by rank 100 for both the plausibility and belief. This ability to narrow the search field has benefits in security and surveillance situations where a query is entered as a textual parameter and then the best matches are returned for operator review. Search and evaluation in this manner would provide operators at the scene with a much more manageable list of subjects to visually inspect when searching for the desired subject.

Fig. 14 shows four examples of matching different queries to the corpus. The top two rows outline successful searches, the bottom two show failures. Success requires a subject annotated with the correct query (outlined in green) be returned within the top four matches.

With three of the top four matches being correct, row one outlines a successful attempt. While the fourth subject is of similar appearance to the query, the subject must match the entire query for it to be correct. In row two a similar situation is presented, with many of the subjects being similar to the query where the green shirt dominates the results. In this case, while a single subject completely matches the query, other returned subjects have a partial (yet strong) match. Row three represents an example of colour confusion. The search parameters are a pink shirt and grey pants, and in this instance the red is confused with pink creating incorrect matches. Finally, in row four the target is a horizontal striped white and grey top with blue pants. Although this search fails, two of the subjects returned are wearing the correct texture, but with varying colours, indicating that while we fail to detect the correct subject it does utilise the texture trait correctly.

Figs. 13 and 14 show that while promising results are achieved, there are still considerable limitations. To investigate these, a deeper evaluation of the traits at a class level is performed. Fig. 15 shows the ranked output for each individual trait, allowing an evaluation of each independently.

From Fig. 15 it is clear that system is highly reliant on the colour in the torso and leg regions. The strength of the colour is most obvious when evaluating rank one performance of each of the traits. We see here that torso colour achieves a rank one performance of 57.8% as the top scoring trait, leg colour scores 45.9%, torso texture scores 28.8%, and as the leg texture is the least accurate with a rank one performance of only 17.4%.

Additionally, investigation into the distribution of the texture shows that both the torso and leg regions are predominantly

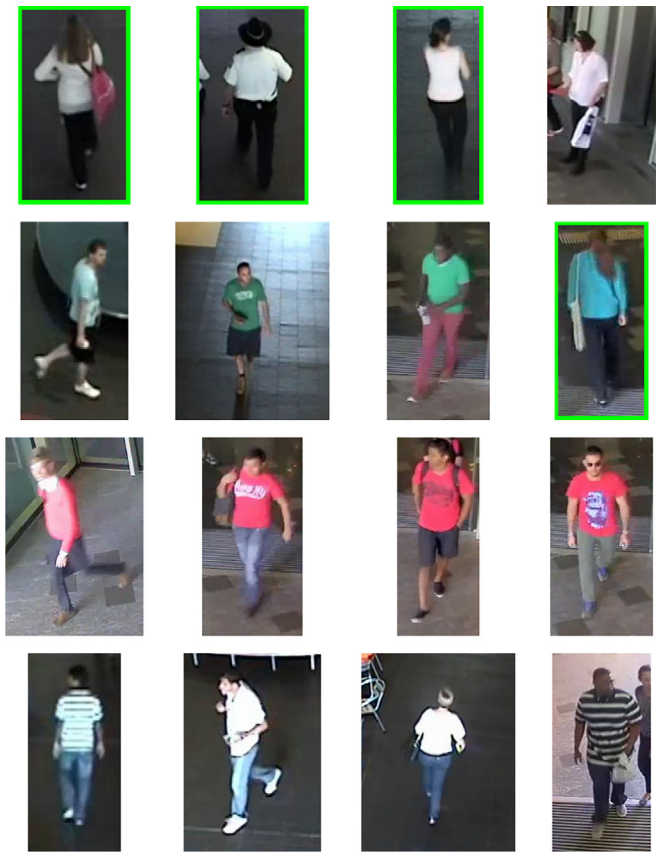


Fig. 14. Ranked comparison of various successes and failures of the proposed technique, each row is a different query and displays the top ranked subject down to the fourth ranked subject for each. Each ranks search annotation is: row (1) plain white top with plain black leg clothing; row (2) plain green top with plain black leg clothing; row (3) plain pink top with plain grey leg clothing; and row (4) horizontally striped white and grey top with plain blue leg clothing. In each row the subjects bounded in green indicate that they have been annotated with the description that is being searched for. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

labelled as 'plain', resulting in a disparity in the evaluation of the traits as their general lack in appearance equates to an inability to adequately evaluate the classifiers. While care was taken to ensure minimal occlusions through bags and multiple items of clothing, these factors still exist which account for some of the reduced performance across the colour and texture classification. We also found that when considering the SIFT and LBP components, the leg region obtained about 50% smaller neighbourhoods, primarily due to the pose of the legs compared to the centralised location of the torso.

While producing promising results in a low resolution surveillance setting, there are still some limitations in the person search technique, particularly with regards to texture. The state-of-the-art results on the texture snippets did not directly transfer to the person search domain, however, the use of DST for fusion was able to somewhat mitigate this and promising results were still obtained. The scaled uncertainty component associated with the clothing texture modalities created the ability to discriminate against these modes while allowing for stronger colour classification where appropriate. The higher resulting uncertainty of the texture classifier (due to higher values of τ) proves the viability of DST as an intelligent fusion technique for soft biometric modalities.

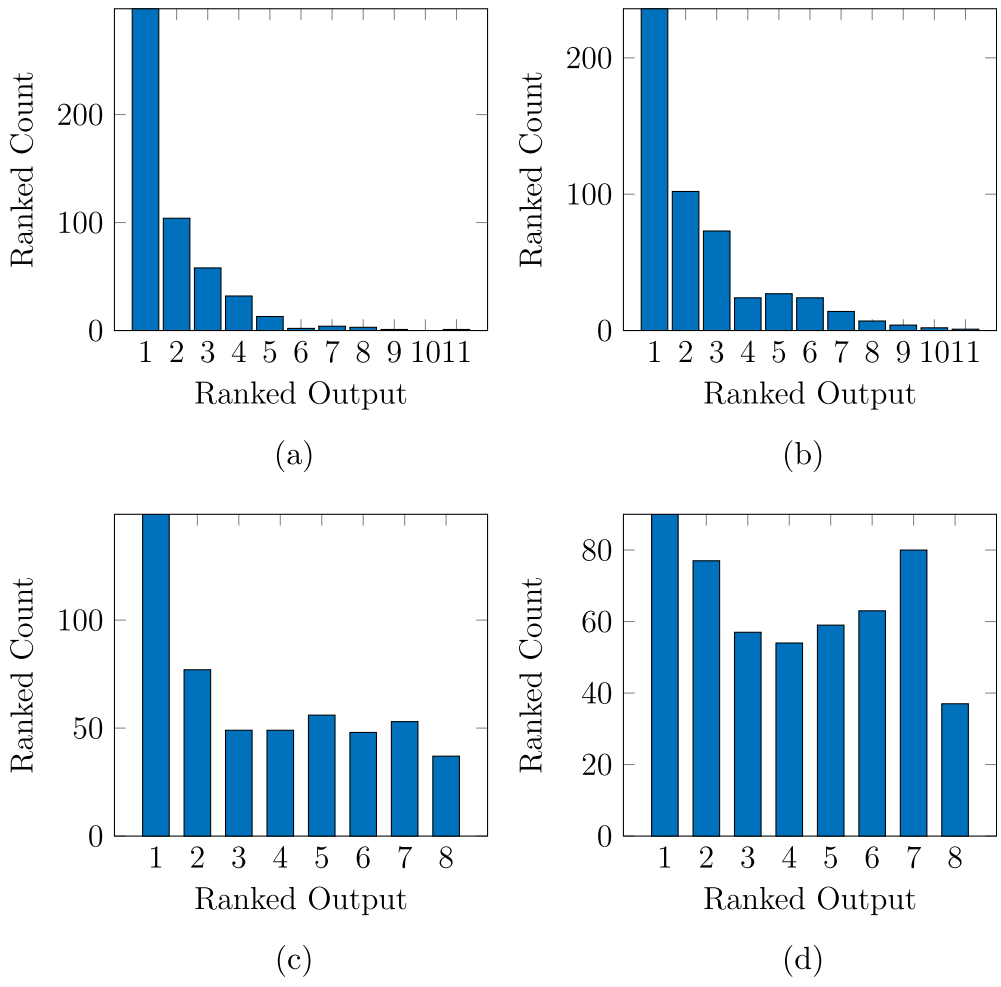


Fig. 15. Ranked output of the different traits evaluated on a class level. (a) torso clothing colour performance, (b) leg clothing colour performance, (c) torso clothing texture performance, and (d) leg clothing texture performance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6. Conclusion

In this paper, we propose a person search technique to match a textually supplied query to a target subject in imagery, using clothing colour and texture. For clothing colour, we train Gaussian mixture models (GMM) from colour snippets extracted from a surveillance setting. A novel texture model is trained using a suite of features (Radon transform signature, wavelet statistics, SIFT, LBP) and an ensemble of support vector machines (SVM), the output of which is fused using Dempster-Shafer theory (DST) to incorporate uncertainty.

As a pure texture classifier, evaluations were completed on an eight class database using images at the full resolution of 140×140 and a down sampled resolution of 25×25, representing what would be expected in a low resolution surveillance situation. Results at the lower resolution showed an improvement of 1.4% over a baseline approach, with the proposed approach scoring a recognition accuracy of 84.75% and the baseline scoring 83.36%. The texture descriptor was also evaluated as part of a person search technique using a newly created and fully annotated surveillance database, which is publicly available to researchers, allowing direct evaluation of the colour and texture classifiers without the problems caused by parsing or other pre-processing errors.

We evaluated the person search technique by attempting to retrieve matches to a query from the entire corpus. In this paper we utilised hand segmented surveillance images to reduce errors associated with automated semantic segmentation techniques. This evaluation showed, that while limitations exist, our ability to match to a target query using just two soft biometric descriptors was promising. In this metric we demonstrate that at rank 100 (out of 520) we were able to produce 58.3% chance of detecting the desired subject. Investigation into the matching power of the individual traits reveals a heavy reliance on colour, as the individual texture performance in a person search setting was less than optimal. In part this is due to the imbalanced nature of the data and the inconsistent nature of the patches, however encouraging results are still obtained.

The evaluation of the full system and individual trait performance also outlines the strength of DST as a method of fusion. The ability to weight traits based on their uncertainty allowed the system to achieve promising results for person location. Similarly, the strengths of DST would allow for the modular addition of further traits, potentially providing stronger soft biometric signatures, and ultimately better localisation.

Conflict of interest

None.

Acknowledgments

This research was supported by an Australian Research Council (ARC) Linkage grant LP140100282. We would also like to acknowledge High Performance Computing and Research Support for supplying the computing resources.

References

[1] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: CVPR, June 2015, pp. 3908–3916.
[2] B. Berlin, P. Kay, Basic Color Terms: Their Universality and Evolution, University of California Press, Berkeley, 1969.
[3] A. Bertillon, Signaletic Instructions Including the Theory and Practice of Anthropometrical Identification, The Werner company, 1896.
[4] C.-C. Chang, C.-J. Lin, Libsvm: A library for support vector machines, Trans. Intell. Syst. Technol. 2 (3) (2011) 27.
[5] H. Chen, A. Gallagher, B. Girod, Describing clothing by semantic attributes, in: ECCV, Springer, 2012, pp. 609–623.
[6] A. Dantcheva, C. Velardo, A. DAngelo, J.-L. Dugelay, Bag of soft biometrics for person identification: New trends and challenges, Multimedia Tools Appl. 51 (2) (2011) 739–777.
[7] S. Denman, C. Fookes, A. Bialkowski, S. Sridharan, Soft-biometrics: unconstrained authentication in a surveillance environment, DICTA (2009) 196–203.
[8] S. Denman, M. Halstead, C. Fookes, S. Sridharan, Searching for people using semantic soft biometric descriptions, Pattern Recogn. Lett. 68 (2015) 306–315.
[9] S. Denman, M. Halstead, C.B. Fookes, S. Sridharan, Searching for semantic person queries using channel representations, in: ICASSP, April 2015, pp. 1568–1572.
[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: CVPR, 2010, pp. 2360–2367.
[11] R. Feris, R. Bobbitt, L. Brown, S. Pankanti, Attribute-based people search: Lessons learnt from a practical surveillance system, in: International Conference on Multimedia Retrieval, 2014, p. 153.
[12] M. Halstead, S. Denman, S. Sridharan, C.B. Fookes, Locating people in video from semantic descriptions: a new database and approach, in: ICPR, 2014, pp. 4501–4506.
[13] M. Hu, Y. Wang, Z. Zhang, Maximisation of mutual information for gait-based soft biometric classification using gabor features, IET Biometrics 1 (1) (2012) 55–62.
[14] E. Jaha, M. Nixon, From clothing to identity; manual and automatic soft biometrics, Trans. Informat. Forensics Sec. PP (99) (2016) 1.
[15] A.K. Jain, S.C. Dass, K. Nandakumar, Soft biometric traits for personal recognition systems, in: International Conference on Biometric Authentication, 2004, pp. 717–738.
[16] J. Johnson, R. Krishna, M. Stark, L.J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Image retrieval using scene graphs, in: CVPR, June 2015, pp. 3668–3678.
[17] R.U. Kay, Fundamentals of the dempster-shafer theory and its applications to system safety and reliability modelling, Reliab.: Theory Appl. 3 (2007) 173–185, special Issue.
[18] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Babytalk: Understanding and generating simple image descriptions, Trans. PAMI 35 (12) (2013) 2891–2903.
[19] X. Liang, C. Xu, X. Shen, J. Yang, J. Tang, L. Lin, S. Yan, Human parsing with contextualized convolutional neural network, Transactions on PAMIDOI: 10.1109/TPAMI.2016.2537339, 2016.
[20] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, S. Yan, Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set, in: CVPR, 2012, pp. 3330–3337.
[21] P. Luo, X. Wang, X. Tang, Pedestrian parsing via deep decompositional network, in: ICCV, 2013, pp. 2648–2655.
[22] K. Nguyen, S. Denman, S. Sridharan, C. Fookes, Score-level multibiometric fusion based on dempster–shafer theory incorporating uncertainty factors, Trans. Human-Machine Syst. 45 (1) (2015) 132–140.
[23] U. Park, A. Jain, I. Kitahara, K. Kogure, N. Hagita, Vise: Visual search engine using multiple networked cameras, in: ICPR, vol. 3, 2006, pp. 1204–1207.
[24] H.A. Perlin, H.S. Lopes, Extracting human attributes using a convolutional neural network approach, Pattern Recogn. Lett. 68 (Part 2) (2015) 250–259.
[25] D.A. Reid, M.S. Nixon, S.V. Stevenage, Soft biometrics; human identification using comparative descriptions, Trans. PAMI 36 (6) (2014) 1216–1228.
[26] S. Samangooei, M. Nixon, B. Guo, The use of semantic human description as a soft biometric, in: Biometrics: Theory, Applications, and Systems, September 2008.
[27] R. Satta, F. Pala, G. Fumera, F. Roli, People search with textual queries about clothing appearance attributes, in: Person Re-Identification, Springer, 2014, pp. 371–389.
[28] J. Thornton, J. Baran-Gale, D. Butler, M. Chan, H. Zwahlen, Person attribute search for large-area video surveillance, in: Conference on Technologies for Homeland Security, Nov 2011, pp. 55–61.
[29] D.A. Vaquero, R.S. Feris, D. Tran, L. Brown, A. Hampapur, M. Turk, Attribute-based people search in surveillance environments, in: Workshop WACV, 2009.
[30] Q. Wang, A. Alfalou, C. Brosseau, New perspectives in face correlation research: a tutorial, Adv. Opt. Photon. 9 (1) (2017) 1–78.
[31] K. Yamaguchi, M. Kiapour, L. Ortiz, T. Berg, Parsing clothing in fashion photographs, in: CVPR, June 2012, pp. 3570–3577.
[32] K. Yamaguchi, M.H. Kiapour, L.E. Ortiz, T.L. Berg, Retrieving similar styles to parse clothing, Trans. PAMI 37 (5) (2015) 1028–1040.
[33] X. Yang, S. Yuan, Y. Tian, Assistive clothing pattern recognition for visually impaired people, Trans. Human-Machine Syst. 44 (2) (2014) 234–243.
[34] J. Zhu, S. Liao, D. Yi, Z. Lei, S.Z. Li, Multi-label cnn based pedestrian attribute learning for soft biometrics, in: International Conference on Biometrics, IEEE, 2015, pp. 535–540.