Feature Selection for Subject Ranking using Soft Biometric Queries

Emil Barbuta Cipcigan, Mark S.Nixon

University of Southampton
https://www.southampton.ac.uk

Abstract

This paper presents a feature selection model that aims to identify subjects from low-resolution surveillance images based on a soft biometric description query. The process is divided into three main stages. In the first stage, semantic segmentation is performed on the subjects, classifying and localising different parts of their bodies / accessories. The second stage extracts information from the segmentations and maps each subject to a vector in a soft biometric feature space. Last but not least, the purpose of the final stage is to find a good weighting on the features extracted in the previous step, based on the intuition that some of them are more important, more accurate or have a higher variance. It is assumed that the matching process might benefit considerably from a set of good weights. Analysis on the IEEE AVSS Challenge dataset shows encouraging performance for segmentation and subject matching with the correct subject reliably matched just outside the top ten on the training set, and just outside top 10% on the recently released test set.

1. Introduction

The ability to identify subjects from CCTV images is much needed and has been widely studied. *Soft biometrics* became a subject of interest for human re-identification due to their property of being easily describable by people which closed the gap between human and machine understanding as far as information from the images is concerned. In terms, this aids the automatic human re-identification process, avoiding the need to manually inspect many hours of surveillance footage.

This paper explores the possibility of using soft biometric features for performing human re-identification from low resolution surveillance images based on semantic queries. A pipeline that contains three stages of information processing and independently trainable modules is proposed. The first stage performs semantic segmentation on the subjects using state-of-the-art atrous convolution filters at different resolution levels in a pyramidal fashion. In the second stage, a set of experts are trained separately using machine learning classifiers on various features extracted from the segmented maps, while in the last stage the weights of the soft biometric features are tuned with respect to the score obtained in the query ranking. The remainder of this work is organised as follows. Section 2 reviews some of the previous research works done in soft biometrics and semantic segmentation. Section 3 presents the feature selection model in details. Section 4 shows and analyses the results, while the final conclusions are drawn in Section 5.

2. Related Work

2.1. Soft Biometrics

The long term objective is to develop automatic computer vision techniques that outperform human analysis. Studies in [15] reveal some of the main differences in performance between human and machine vision. The results showed that human vision finds it much easier to evaluate traits like age, race and gender and it is hardly affected by illumination conditions or the low quality of images. On the other hand, human identification by humans may be subject to memory limitations and descriptive completeness. Another aspect about the limitations of the soft biometrics is discussed in [6] which encourages the usage of as many traits as possible, because unimodal systems (systems that use only one feature) are highly sensible to noisy sensor data, low permanence, unacceptable error rates and spoof attacks. Combining evidence from multiple sources can reduce these issues.

The relative importance of the individual features has been a subject of interest in many papers ([10], [14], [15], [16]). For instance, in [10] it is remarked that weighting features based on their distinctiveness and permanence can have a high positive impact on the results. Investigations with the purpose of finding out what body features are the most relevant are also carried out in [15] and [16]. Moreover, the authors of [16] developed a study about how many of those features are correlated to each other when annotated by humans. Finding such internal correlations would be useful not only to reduce the number of features to a smaller relevant set, but also to infer the value of missing features in cases they cannot be extracted properly. In [14] the main focus was the usage of relative rather than absolute features. The work concluded that relative features clearly outperformed the categorical ones because of their ability to assign soft probabilities.

2.2. Semantic Segmentation

Semantic segmentation has been a topic of research for many years, as it helps accomplishing various tasks. Some basic proposed methods used iterative clustering to generate superpixels [2], while others performed mean-shift and re-weighting on histograms of oriented gradients [19]. Another well known approach is the usage of conditional random fields ([9], [11], [21]) that can capture both the probability of labels and the consistency between adjacent pixels.

Starting with the great achievements obtained by the deep convolutional neural networks in the large scale image classification competitions, deep nets began to receive more attention for a wide range of tasks. Semantic segmentation was of course, not an exception. Very often these models have been combined with conditional random fields in order to explicitly account for the smoothing constraint ([3], [12], [17]). A fully convolutional model based on processing images at different levels of resolution using pre-trained weights is proposed in [13]. The main observation was that intermediary coarse level representations adjust the context of the pixels and improve neighbouring consistency, while fine level representations focus on small details. The main drawback in such architectures is the need to downsample/upsample the images, which makes them lose information. And because using large convolution kernels to incorporate more context is very expensive as it drastically increases the number of parameters, researchers have come up with a novel concept called dilated or *atrous convolutions* ([4], [5], [20]). Those filters present sparse convolution kernels that have a small number of parameters, but the field of view can easily be increased. The ability to extend this field of view without overpopulating the model with parameters is what makes them such a powerful tool. It also allows for explicit control over the resolution level at which features are computed, as explained in [4].

3. Approach

3.1. Dataset and Task Overview

The project uses the dataset provided by *Semantic Person Retrieval in Surveillance Using Soft Biometrics* challenge organised by the IEEE AVSS 2018 [1]. More information about the setting from which the data was captured can be found in [8]. The dataset consists of 520 surveillance images of variable size capturing one human subject at a time, and also some sets of ground truth segmentations comprising 9 classes (background, legs, shoes, torso, luggage, skin_legs, skin_arms, skin_face and hair). A soft biometrics description is provided for each subject, labelling the features in a categorical way. The goal is to build a model that ranks all the subjects based on how well they match a given query. The evaluation is done by computing the mean rank of all subjects as response to their corresponding query.

3.2. Main Pipeline

The model proposed in this paper consists of a pipeline containing three independently trained modules, as shown in Figure 1. Secions 3.4, 3.5 and 3.6 present the functionality and development of each module.

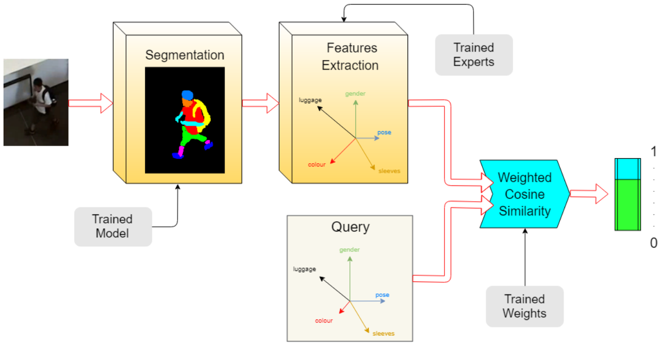


Figure 1: The Feature selection model

3.3. Data Augmentation

The size of the dataset (520 images) is relatively small, especially for training deep convolutional neural networks. For a classification task with a small number of classes, it might exhibit acceptable results, but for semantic segmentations with 9 classes, regular models would be prone to overfitting. This is why, data augmentation can be a crucial pre-processing step. Another aspect that needs to be tackled before training is the variation in the shapes of the input images. The method proposed in this work, solves both issues and it also provides an explicit technique for enforcing the subsequent trainable model to be scale invariant: a mirrored version of every image is added to the set, and for both of them n rescaled versions (excluding the original size) between some bounds are performed. In the final experiments n was set to 11. Every obtained image was padded with black pixels, so that in the end, each sample was an image of shape ($500 \times 300 \times 3$). This process led to an augmentation factor of 24 (1 original + 1 flipped + 11 scales for each of them), so the final augmented dataset contained 12,480 samples, which is a considerable increase over the initial 520. Figure 2 shows a sample of the newly obtained data. As an observation, during training small rotations have also been applied, but it turned out to be rather a shortfall than an improvement in the results regardless of the architecture of the model. The reason for this might be the fact that the rotations, even though just in the approximate range $(-15°, 15°)$, ruined the positional relations between the segmented classes and consequently just hampered the performance of training instead of reducing overfitting.

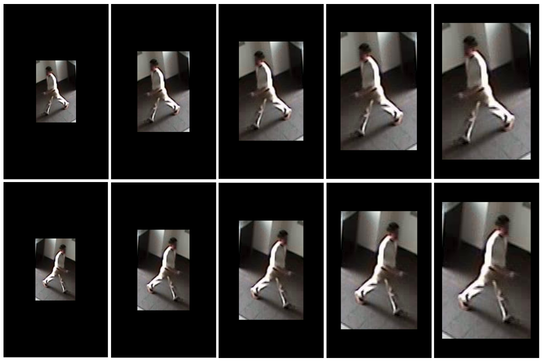


Figure 2: Data Augmentation

3.4. Semantic Segmentation (I)

The main idea of the segmentation model was to combine two key concepts implemented in several previous works in the literature. One of those concepts is processing feature maps at different levels of resolution, in a similar manner as the one presented in [13]. The assumption was that combining fine level and coarse level feature maps can capture both context and fine details. The second key concept is the usage of multiple atrous convolutions with different rates on the same feature maps, in a pyramidal fashion. This approach was first described in [5] and named *Atrous Spatial Pyramid Pooling*. The overall architecture of the semantic segmentation model is displayed in Figure 3.

Blue nodes represent layers that are static (frozen or no parameters), while the orange ones contain trainable parameters. The model begins by extracting features using the pre-trained VGG network [18] (initially trained for image classification in the ImageNet competition). It extracts feature maps from 3 different layer levels, preserving all the operations performed in the original VGG. This is why the size of the image is reduced when extracting from layers 4 and 7. The next step applies an *Atrous Pyramid Module* which is inspired from the last layers proposed in [5], but slightly changed. The configuration of this module is shown in Figure 4a.

After applying 4 parallel atrous convolutions with rates 1, 2, 4 and 6, a set of convolution layers with 16 filters is applied pixelwise on each individual feature set, which are finally concatenated into a feature map with 64 channels. It should be noted that the atrous VGG layers are all initialized to the weights from the original subsequent VGG layer. For instance, for the module that extracts features from Layer 7, the following atrous filters (Figure 4a) are all initialized to the weights from VGG layer 8, but are modifiable in the training process.

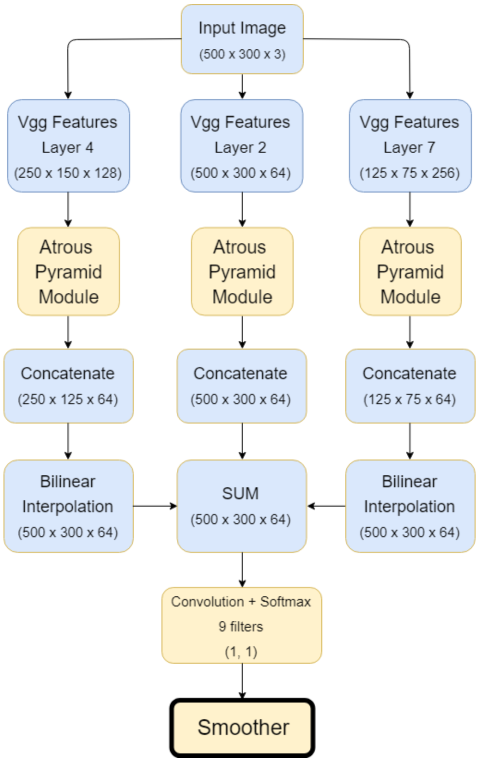
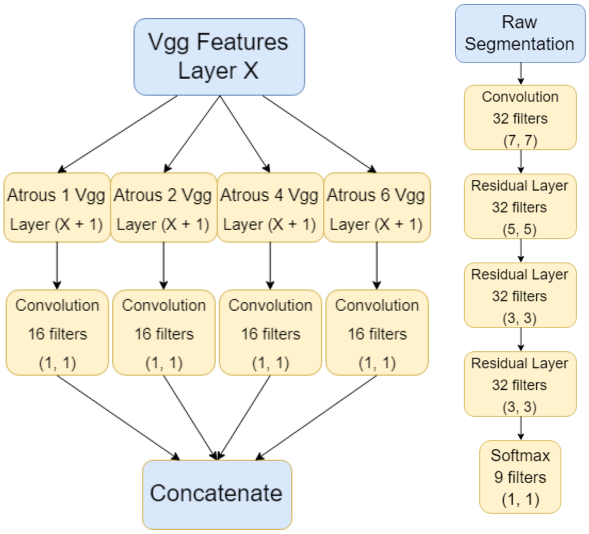


Figure 3: Segmentation model architecture

The model comprises interpolation of the low resolution maps and summation to reduce noise and inconsistencies between adjacent pixels. A general suggestion ([3], [4] and [12]) is to add a CRF to enforce a smoothing constraint over the final resulting map. This model on the other hand, proposes appending another module called the *Smoother*, as it can be observed in Figure 3. This entity is basically an additional set of convolutional and residual layers (as shown in Figure 4b). The choice of using residual layers is motivated by the fact that the task of this component is to get a raw segmentation as input and correct it by accounting for smoothness as well as for the accuracy of pixel predictions. Every residual layer has the task of getting a segmentation map and adding a small refinement to it. The whole network is trained in two main steps. First the initial part excluding the smoother is trained to minimize the cross-entropy loss until it settles at a point from where it no longer decreases for a period of time. After that, the smoother starts to train by keeping the rest of the network frozen. Dividing the training process in those 2 stages avoids creating false correlations between the first inference part and the second smoothing part, and reduces the chances of causing an internal covariate shift. The output of the smoother is considered the output of the segmentation stage. A batch size


Vgg Features Layer X
Atrous 1 Vgg Layer (X + 1)
Atrous 2 Vgg Layer (X + 1)
Atrous 4 Vgg Layer (X + 1)
Atrous 6 Vgg Layer (X + 1)
Convolution 16 filters (1, 1)
Convolution 16 filters (1, 1)
Convolution 16 filters (1, 1)
Convolution 16 filters (1, 1)
Concatenate
Raw Segmentation
Convolution 32 filters (7, 7)
Residual Layer 32 filters (5, 5)
Residual Layer 32 filters (3, 3)
Residual Layer 32 filters (3, 3)
Softmax 9 filters (1, 1)

(a) Atrous Pyramid Module

(b) Smoother architecture

Figure 4: Expanded components

of 4 images was used and the loss was minimized using the Adam optimizer with a polynomial decay on the learning rate that starts at 0.001 and saturates at 0.0001 after 10,000 iterations.

3.5. Local Feature Experts (II)

The next step is to extract information from the segmented maps and infer the value of the features that can be present in the queries. The annotations from the training set have been used as labels. Each feature was evaluated using an individual *expert*, and each expert may use different input information that is fed to a machine learning algorithm. For choosing the most appropriate features and classifier, a series of empirical tests were carried out and the best performing ones were kept. A short description of each expert is given in the following:

Colour (both torso and legs): the area of interest was extracted from the RGB image using the segmentation and a historgam of values was build for each channel. An RBF-kernel SVM was then used to classify the colours.

Gender: the ratio between the area of each class and the total segmented area was computed and considered as a feature for a subsequent Random Forest.

Luggage: the ratio between the area classified as luggage and the rest of the segmented area was considered alone and fed to a Linear SVM with hinge loss.

Type (both torso and legs): the ratio between the area classified as clothing and the area classified as skin was taken as the sole feature and fed to a Linear SVM.

Pose: the ratio between the area of each class of the segmentation and the total segmented area was computed and considered as a feature for a subsequent Gradient Boosting.

Texture (both torso and legs): texture information has not been used as it was considered a difficult problem that might turn out to be detrimental.

An important remark that has to be made here is that all classifiers gave their response as a soft probability over the possible classes. This was achieved by taking advantage of the specific mathematical properties of each classifier, like distances to the boundary lines of SVMs for instance .

3.6. Global Feature Weighting (III)

A prior weight depending on the size of the feature descriptor vector can be assigned (for example luggage presence is binary, so its only node has weight 1, while pose has a 4 node categorical representation, each node weighting 1/4) in order to normalize things, but this procedure still does not capture the usefulness of the features. One approach to find an optimal set of weights is to perform a grid search. But having 7 features, and assigning only 4 possible values for each would result in evaluating the score for the queries $4^7 = 16,384$ times, which is impractical.

Instead, this paper proposes a *Discrete Gradient Descent* (Algorithm 1) that proceeds as follows. It starts by assigning equal weights to all features. At each iteration it tries to see the direction of the ”gradients” for each feature, and then changes the weights in proportion to those gradients. The *counter* is used to signal when many consecutive iterations do not show improvement, meaning that the algorithm converged to a solution. The $update\_decay$ ensures updates are getting smaller and smaller over time, in a similar fashion to the learning rate decay in deep neural networks.

4. Results and Analysis

Figure 5 shows segmentation results both on training (top 3 rows) and on testing (bottom 3 rows) images. The examples on the left side are good quality samples, while on the right side there are cases where the results appear to be rather poor. The model tends to give worse results in images where illumination varies (row 1), edges are very difficult to detect (row 2) or multiple humans appear in the frame (row 3).

The numerical results representing the mean intersection over union of the predictions of different models are displayed in Table 1. *Static* versions are those that disallowed the atrous layers to be trained, *L2* stands for adding euclidean norm penalty to the loss (0.001 in our experiments), and *-E* indicates an extension to (3, 3) kernels instead of (1, 1) in the convolutions that follow the atrous layers. The data was split in 75% training, 15% validation and 10% testing.

In a similar task on a comparable dataset, [7] set the baseline for locating people in video frames using semantic descriptions at a mean_iou of 44%. The results in Table 1 show an improvement of 19% (considering best vs best) over this

Algorithm 1 Discrete Gradient Descent

```
$feature\_weights \leftarrow (1.0, 1.0, ..., 1.0)$
$counter \leftarrow 0$
$update\_decay \leftarrow 0.95$
$update\_magnitude \leftarrow 1$
while counter < COUNTER_THRESHOLD do
    $score$ ← evaluate the score on all queries (mean ranking)
    if $score$ is minimum so far then
        save current $feature\_weights$
        $counter \leftarrow 0$
    end if
    $counter \leftarrow counter + 1$
    for each feature weight $f$ do
        make a slight increase and decrease and re-evaluate the score on those
        choose the operation that improves the score and store the value of the improvement $f\_improvement$
    end for
    normalize all $f\_improvement$ across features
    for each feature weight $f$ do $f \leftarrow f + update\_magnitude * f\_improvement * f$
    end for
    $update\_magnitude \leftarrow update\_magnitude * update\_decay$
end while
```

Model	Training	Validation	Test
Dynamic	64.85%	51.54%	48.22%
Static	60.83%	47.73%	47.93%
Dynamic + L2	61.84%	51.58%	49.38%
Static + L2	64.92%	51.40%	**51.97%**
Dynamic-E	64.44%	49.56%	**51.33%**
Static-E	63.09%	50.98%	**52.25%**
Dynamic-E + L2	49.04%	45.25%	45.42%
Static-E + L2	49.23%	45.82%	46.62%

Table 1: Segmentation models mean-IOU

baseline, which proves that the model proposed for semantic segmentation in this research project has great potential for tackling this kind of tasks.

Table 2 displays the results obtained for each individual expert, while Table 3 shows the mean and median ranking obtained for ranking all subjects from their queries. When optimizing for the best feature weights using Algorithm 1, the Dynamic model was used for extracting segmentation maps (on the training images). The table shows the scores when testing this set of weights not only on the true segmentations, but also on the ones obtained using the deep convolutional model. The reason why the median is always much smaller than the mean is because most of the subjects are ranked in the top 5% or better, but some samples are heavily misranked due to very poor segmentations.

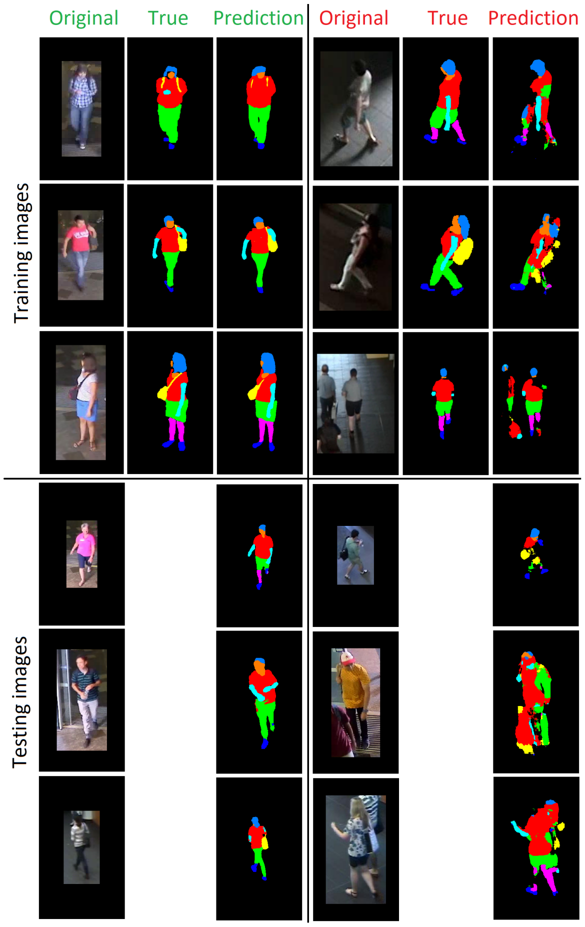


Figure 5: Segmentation results

Feature	Classes	Classifier	Accuracy
Pose	4	Gradient Boosting	59.04%
Gender	2	Random Forest	79.30%
Luggage	2	Linear SVM	91.35%
Torso colour	11	Kernel SVM	48.64%
Leg colour	11	Kernel SVM	56.61%
Torso type	2	Linear SVM	90.96%
Leg type	2	Linear SVM	96.54%

Table 2: Feature experts results

Many versions of models can be compared in terms of performance, but two of them which attracted a lot of interest have their cmc ranking curves plotted in Figure 6. The curve *DA_weights* represents a model that trained the feature weights on segmentations predicted by the Dynamic model, whereas *GT_weights* used the ground truth segmen-

Segmentation	Rank Mean	Rank Median
Ground Truth	7.510	1
Dynamic	38.162	11
Static	45.423	16
Dynamic + L2	41.373	13
Static + L2	39.042	11
Dynamic-E	38.967	11
Static-E	42.277	13
Dynamic-E + L2	58.560	25
Static-E + L2	57.981	22

Table 3: Ranking scores on the training set

tations as input. As for segmenting the test images, both of them used the Dynamic model. The results show that DA_weights clearly outperform GT_weights. Hence, it has been proved that accounting for the noise in the segmentation stage when optimizing the feature weights for the matching stage offered a considerable improvement in the final ranking results. Additionally, DA_weights got a ranking mean of **43.14** and a ranking median of **24**, while the GT_weights scored **49.51** and **29.5** respectively.

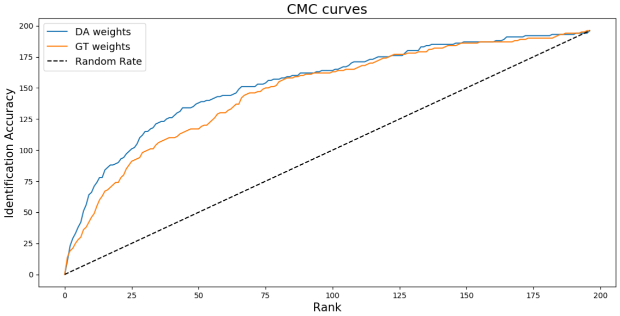


Figure 6: CMC curves

5. Conclusions and Future work

A novel method for ranking subjects based on soft biometric feature selection was presented. The methodology shows how the main problem can be divided into independently solvable subproblems. Finding better models, or gathering more data could improve the quality of the segmentation maps and this can ease the feature extraction stage. Furthermore, using soft predictions on the maps like in [7] and implementing a texture extractor can offer a boost. Another method that might be beneficial is integrating all feature experts into one single entity that would be able to also take into account correlations between the attributes. Nevertheless, the overall approach presented promising results and offered insights into the usage of soft biometric traits for human identification.

References

[1] IEEE AVSS : Semantic person retrieval in surveillance using soft biometrics. *(https://semanticsbsearch.wordpress.com)*. Accessed: 2018-08-21.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012.

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[5] L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[6] A. Dantcheva, C. Velardo, A. D'Angelo, and J. Dugelay. Bag of visual soft biometrics for person identification: New trends and challenges. In *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages xii–xii. Oct 2015.

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

[8] 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.

[9] X. He, R. S. Zemel, and M. A. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*

[10] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In D. Zhang and A. K. Jain, editors, *Biometric Authentication*, pages 731–738. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[11] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *CoRR*, abs/1210.5644, 2012.

[12] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. *CoRR*, abs/1509.02634, 2015.

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] D. Martinho-Corbishley, M. Nixon, and J. N. Carter. Super-fine attributes with crowd prototyping. pages 1–1, 2018.

[15] S. Samangooei, M. Nixon, and B. Guo. The use of semantic human description as a soft biometric. In *Biometrics: Theory, Applications, and Systems*. September 2008.

[16] S. Samangooei and M. S. Nixon. On semantic soft-biometric labels. In V. Cantoni, D. Dimov, and M. Tistarelli, editors, *Biometric Authentication*, pages 3–15. Springer International Publishing, Cham, 2014.

[17] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. *CoRR*, abs/1503.02351, 2015.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[19] Y. Socarrás Salas, D. Vázquez Bermudez, A. M. López Peña, D. Gerónimo Gomez, and T. Gevers. Improving hog with image segmentation: Application to human detection. In *Advanced Concepts for Intelligent Vision Systems*, 2012.

[20] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015.

[21] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*.