Enhancing edge-based image descriptor models through colour unification

Dumusani Kunene
Council for Scientific and Industrial Research,
Pretoria, South Africa
Email: dkunene@csir.co.za

Vusi Skosana
Council for Scientific and Industrial Research,
Pretoria, South Africa
Email: vskosana@csir.co.za

***Abstract*—The lack of suitable robust appearance models hinders the performance of most image descriptors. Descriptors often rely on pieces of information in images called image features to discriminate the contents of images. Most successful descriptors use gradient images for determining the overall shapes of objects. Consequently, the inferred features are often susceptible to the noise caused by shadows, reflections and inner textures within the object. Significant efforts have been made towards improving the performance of image classifiers, yet generic object detection remains an open problem. In this paper, a method aimed at improving existing appearance models is proposed. The focus is on enhancing the acquired information at fundamental stages to improve the robustness of common statistical learning classifiers, as seen with the work of Holger Winnemoller et al. with human subjects.**

The selective Gaussian blur filter was applied to several human classification datasets to reduce the amount of ambiguous low-frequency noise. Experiments were then conducted to determine whether the unification of similar colours in local image regions could improve the acquired image features. The classification results that were obtained with the processed images were competitive to the results obtained with the original images, however inconclusive for demonstrating the benefits of image smoothing.

***Index Terms*—classification, image smoothing, colour unification, edge-preserving filters, feature descriptor enhancement**

I. Introduction

Object recognition is one of the fundamental goals of autonomous digital-image understanding. Over the years, various appearance models [2], [7], [17] have been developed to better discriminate the visual cues of objects in images. Appearance models often quantise large amounts of pixel data into compact vector forms that are learnable through statistical learning methods.

Image descriptors often match spatial patterns of features in images, but classifiers have to be trained with large image datasets that contain multiple examples of objects and their labels. Given large datasets, classifiers learn to predict the class labels of objects from either raw pixel information or appearance models. On most image datasets, objects of interest are often placed near the centre of the image and form at least 80% of the image. The extracted images often contain substantial background clutter that degrades the shape information of the foreground object.

Most appearance models rely on gradient images for determining the shapes of objects [2], [5], [6], [12]. The extraction of edges plays a vital role in formulating the appearance of objects for models that solely rely on the edge information. However, the extracted edges are not only formed by the contours of objects but also the changes in shading, colour, diffuse and illumination on both foreground and background regions of images. There is hardly any effort in filtering out the ambiguous edge-noise before the extraction of features, despite Holger Winnemöller et al. [25] having demonstrated with human subjects that recognition rates and memory can be improved by smoothing images.

This work studies ways of exploiting image smoothing in order to enhance the quality of edge-based features for image classification. The authors test the effects of suppressing noise in images over the human classification problem. This is a relatively difficult problem to solve as people are difficult to recognise autonomously compared to generic rigid objects.

A. Related Work

Given the focus of this work, the reviewed literature is in classification tasks, human detection and image enhancement. The human detection problem has attracted a wide range of research due to the present demand of autonomous surveillance systems, self-driving cars, crowd behaviour modelling, interactive industrial and home assisting robots and so on. The best human detection computer vision solutions thus far are based on convolutional neural networks (CNNs). The CNNs have been dominating image classification benchmarks since Krizhevsky et al. [12] considerably reduced the classification error of the ImageNet LSVRC-2010 benchmark. Unlike generic man-made features, the CNNs can automatically learn the required edge filters that can produce the ideal descriptive feature-maps for generic image classification problems. Man-made image features are often limited to representing a few classes of objects. For instance, Haar-Wavelet features only perform well on rigid objects such as faces. Scale-invariant feature transform (SIFT) features are good in cases where high contrast points can be extracted from the edges of objects.

Earlier human detection work prior to 2005 was based on template matching [21], where edge templates from reference images were used to detect people in unknown images. Such methods failed due to the non-rigid nature of people. The problem of detecting people on static images is often confused with finding moving people on video footages that are captured

Fig. 1. Sample results of a cartoonized image.

from stationary platforms. The latter problem is solvable through background subtraction and shape interpretation from motion masks. This method imposes many restrictions and is not useful when the targeted objects are stationary in the videos. Furthermore, the standard background subtraction algorithm is not effective on image sequences that are captured from moving sensors [3].

Facial detection is solvable through texture-based features but cannot be used for classifying people in images. Simply because such methods mostly work on higher resolution images where people are assumed to be facing the camera at all times. In practice, people have no consistent texture patterns from their clothing and can be photographed in different poses. The appearance of people is heavily affected by the perspective from which they are photographed. Over the years, the most promising attempts modelled the silhouette shape of people. Silhouettes describe the appearance of people better and have been shown to be ideal for handling partial occlusions [7].

Wei Gao et al. [7], have presented a method that could capture the contours of people more effectively than histogram of oriented gradients (HOG) features. This method, described as the adaptive contour feature (ACF) was aimed at improving the existing discriminative feature models before 2009. The ACF is described as a chain of square patches that model the curve of an object's contours. The ACF does this by using granules in an oriented granular space (OGS). Since this algorithm also relies on edge information, it may also benefit from the extraction of high-frequency edges. Other related models that could also benefit are described in [3], [5], [7]. The software implementation of the ACF algorithm was unavailable, studying the effects of ACF features may be part of the future work.

With the visual perception of humans, edges are vital for the neural interpretation of a scene [26]. Therefore, the objective of smoothing images in this work is to reduce noise so that the visibility of contour edges could be improved. Traditional smoothing filters like the Gaussian blur, degrade important features and dislocate edges [24]. Non-linear diffusion filters were developed to overcome this shortcoming, the Perona-Malik model [19] being the first. These kinds of filters seek to preserve and enhance semantically important information such as edges [24]. More edge preserving filters have been developed since then, they can be broadly classified into three categories namely: average-based, patch-based and optimisation-based. Average-based filters include the Perona-Malik method, the bilateral filter (BLF), bilateral textured filter and propagated image filter (PIF), guided filter (GF), rolling guidance filter (RGF). These filters are based on the weighted average of a neighbourhood to determine the output of each local pixel [16]. Patch-based filters include the region covariance (ReCov) and aim to overcome the limitation of average-based filters in distinguishing the image structure from the object textures by using a covariance matrices of an image patch [10]. Optimisation-based filters include total variation (TV), relative total variation (RTV), weighted least squares (WLS), L0 gradient minimization (L0) and mixed-domain edge-aware image manipulation (MD). Optimisation-based filters use global optimisation on energy functions, which consists of a data term and a smoothness term [16]. The data term is designed to maintain similarity between the input and output images, while the smoothness term is designed to remove details or textures.

Lin et al. [16] compared the performance of RTV [27], ReCov [11], RGF [28] and bilateral textured filter [1] with their own method. They found that the ReCov had a blurring effect and the longest execution time, while the RGF and BLF blurred small object details; only RTV and Lin et al's method were able to retain small object details. Tang et al. [22] compared their method to WLS [4], BLF [23], GF [8], RGF [28], L0 [26], MD [15], ReCov [10] and PIF [20]. The GF left much of the texture. The WLS, BLF and ReCov blurred the object edges while the MD retained sharp edges but significantly reduced image contrast. The L0 filter segmented the image well but generated colour quantisation artefacts and did not smooth out all the textures. PIF and the Lin et al. method performed well in image smoothing and execution time but required a guidance image for each smoothed image, making them impractical for a large database or real-time applications. The RTV and the Tang et al. method are both based on the TV filter and performed well in image smoothing, but did not compare well in terms of complexity and execution time.

The selective Gaussian blur (SGB) combines domain filtering with range filtering. That is, pixels are not only regarded to be similar to one another by their Euclidean distance apart but also perceptually through their colour values. Methods that

quantise colours have similar effects to colour smoothing but have certain restrictions. For example, such methods require prior knowledge of images to successfully segment colours. To quantise colour, the estimated number of dominating colours in the image, size of objects in the image, lighting and the extent of noise clutter has to be known prior. Therefore, a filtering method that could combine/smooth colours without significantly affecting edges, and require no prior knowledge would be ideal for real-time applications. The next section discusses the research methods used to test whether the reduction of noise has a positive effect on the acquisition of HOG features and learned feature maps.

B. Method

Several image pre-processing techniques that could put more emphasis on the high-frequency components while reducing low-frequency image components were considered. The first approach was in testing the use of infrared images on human detection tasks because the heat radiated from humans is less dependent on the colour patterns on their clothes. Instead, the heat radiated from objects of uniform material becomes uniform. This results in reduced shading noise within the contours of objects. As discussed in [14], the reduction of low-frequency noise clutter has the potential to improve the quality of shape descriptive features. Higher classification rates were obtained from the less cluttered infrared image datasets than colour image datasets of the same size. In continuation of this research, this work attempts to redress the same problem on colour images, by lessening the shading clutter that forms ambiguous edge-noise on gradient images. The supposition of this works is only tested with the SGB filter on several classifiers through trial and error, future work may also involve some of the filters that were discussed in related work sub section I-A.

C. Selective Gaussian Blur

The SGB filter takes a step further than the original BLF [23] by introducing an additional edge-protection mechanism that limits or completely suppresses the low-pass Gaussian filter whenever edges are found. For this reason, the SGB filter was favoured in this work. The aim is to unify similar colours in local regions of objects. However, combining similar colours in natural scene images becomes challenging because the colours within foreground regions often blend with the background colours. To do so, the images were converted to their cartoon-like appearance with the use of the SGB filter in the cross-platform image editor GIMP. The SGB outputs better results than the BLF. Figure 2 shows the extent to which noise can be reduced through cartooning images. The gradient image of the cartooned image had remarkably less noise than that of the original image. Other examples of cartooned images can be seen in Figure 1. The two figures show how the SGB filter produces a cartoon-like appearance of images. It is clear why the SGB filter was selected, as this work aims at evaluating image features when the noise caused by ambiguous edges is reduced.

Algorithm 1 Selective Gaussian Blur

procedure CALCULATEGAUSSIAN($I, radius$) ▷ Given image I
  `rad = radius`
  `Convolve a filter` I_f `of size` $[rad \times rad]$
  for filterLocation pos in source **do**
    $gaussI_f \leftarrow$ Gaussian within I_f
    `weights` $= gaussI_f \times I_f$
    Consider center pixel $centrPx$ of I_f
    Let $localPx \in I_f \cap localPx \neq centrPx$
    for each $localPx \in I_f$ **do**
      $diff$ = $centrPx$ - $localPx$
      if $diff \geq maxDelta$ **then**
        ignore this pixel
      else if $diff \leq -maxDelta$ **then**
        ignore this pixel
      else
        $accumulatedWeights \mathrel{+}= weights \times localPx$
        $weightCount \mathrel{+}= weights$
      end if
    end for
    if $weightCount == 0.0$ **then**
      $destinationImage_{pos} = \frac{accumulatedWeights}{weightCount}$
    else if $diff \leq maxDelta$ **then**
      $destinationImage_{pos} = I_{pos}$
    end if
  end for
  return $destinationImage$ ▷ Blurred image
end procedure

A simplified algorithm of the SGB function is shown in Algorithm 1. From the algorithm, it is clear to see that the average distributed weights of pixels is formed from neighbouring pixels. This results in cases where the SGB filter is applied to some pixels, applied to a limited extent with others and not applied to the rest.

The results appear to be better when the images are first passed through the original Gaussian blur filter (implemented in GIMP using `pdb.plug_in_gauss_iir()`, and thereafter passed to a colour normalisation procedure before applying the SGB filter. The assumption made, is that no prior knowledge of images is available. Therefore, the same parameters of the SGB filter have to be used for all images in the datasets. The final parameters were as follows, the initial Gaussian blur pass has a small 2x2 filter size to try and not affect sharp edges in images. This filter is applied to ease the process of smoothing colours on the latter stage. The standard colour normalisation algorithm was applied using the `pdb.plug_in_normalize()` function. The following parameters were selected for SGB filter, the `SelGauss_radius = 5`, `max_delta = 15`. The filter was applied using the GIMP function `pdb.plug_in_sel_gauss()`. This multi-step smoothing process (Referred to as type-B cartooning from here onwards) will be compared to the simpler smoothing step of just apply-

Fig. 2. Comparisons of cartooned images, the images are shown in consecutive pairs of the original image(Left) and the cartooned image(Right).

ing the SGB filter alone (Referred to as type-A cartooning).

D. Classification Datasets

For the experiments, the HOG features were extracted in preparation of training the support vector machines (SVMs) and extreme learning machines (ELMs) for binary classification, CNNs are trained with the raw images. The effects of smoothing colours are tested on both colour and infrared images of the same size. The INRIA and NICTA datasets were selected for colour images. These datasets are commonly used for human detection research [2], [5], [18]. The infrared images were selected from the SIGNI dataset [13].

The work by [12] showed that having large datasets with millions of images per class does improve the generalisation of classifiers. In our case, the datasets only contain a few thousands of images. Consequently, a limited level of data augmentation was performed on the datasets, where a horizontal mirror of each image was computed. For the positive samples, 3401 images of upright people from the INRIA, NICTA and SIGNI dataset were used. Over 5400 distinct negative visual samples were extracted from the background images of the INRIA dataset. The results of the original images from each dataset are compared to the results of their pre-processed counterparts.

II. Experimental Work

In this section, experiments are carried out to determine whether cartooning images improve the acquired HOG features for the classification of humans in still images. To do so, three different classifiers (SVMs, ELMs and CNNs) were trained with 80% of the images in each dataset. The remaining 20% is reserved for testing the generalising performance of the classifiers. The two shallow learning classifiers (SVMs and ELMs) were tested against the deep learning convolutional neural networks. The randomisation function that randomly selects the training and testing partitions of the datasets was reset whenever the classifiers were tested. This ensures that the same images are tested to effectively compare the results between the different classifiers. The randomisation function ensures that consecutive images are never selected in the dataset. This becomes useful, as most datasets often contain images that were captured at similar environments organised together through file naming conventions.

A. Support Vector Machine experiments

The first classifier to be trained was the SVMs. The SVMs have been highly successful in solving binary classification tasks. Given a distinct sample, the SVMs map data into a higher dimensional space to make it linearly separable. The data is separated by a decision function that uses a hyperplane to optimise the margin between opposite data samples.

The primary task when training a classifier is to use an existing collection of known observations to find the optimal parameters that can be used to learn a function $f(x) : X \rightarrow Y$, such that $f(x)$ is a good approximation of any unobserved sample x. Where known observations are pairs of input samples and their labels. The parameters are highly dependent on the size of the datasets used, the type of image features used, the feature vector size and so on. Such parameters are often found through trial and error during preliminary experiments.

The data is often projected into higher-dimensional spaces through kernel functions. The linear, polynomial, radial basis and sigmoid functions are traditionally used as kernel functions for the SVMs. During preliminary experiments of this work, the RBF kernel performed better than the linear, polynomial and sigmoid kernel functions. As a result, the RBF Kernel was selected as the kernel of choice for the SVM classifier. A cross-validation procedure was done over random permutations of each dataset to find the optimal sigma to use for the RBF kernel. The SVMs performed better when values of sigma were between 25 and 45.

The results shown in figures 3, 4 and 5 appear in pairs, where the blue bars represent the classification rate obtained from the unprocessed images, the green bars are values obtained with type-A cartooning and the red bars are the values obtained with type-B cartooning. It is worth noting the scale of the graphs ranges between 90% to 100% to outline the differences in the classification rates.

Inconsistent results were observed from the SVM classifier on the INRIA and SIGNI datasets. For instance, the cartooning methods were beneficial to the INRIA but lead to poorer results on the NICTA. The type-A cartooning method performed better on the INRIA than the type-B method. The poorer results on the NICTA could be attributed to the dataset's lower resolution images. The original NICTA images had dimensions of 32×80. The images had to be up-scaled to

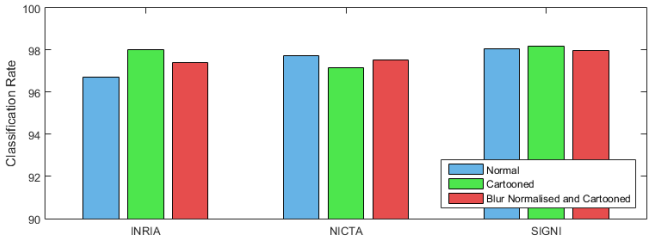

Fig. 3. Classification results of the SVM classifier.

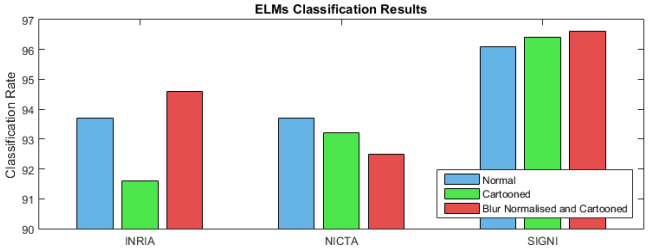

Fig. 4. Classification results of the ELM classifier.

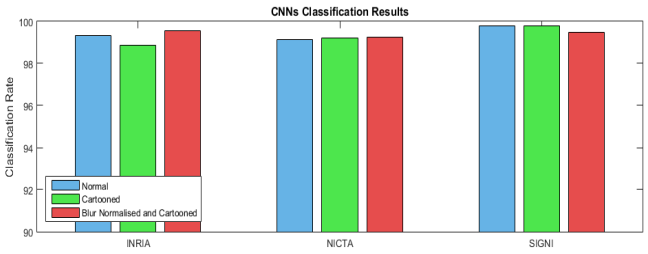

Fig. 5. Classification results of the CNNs classifier.

the standard 64×128 resolution that was used for the HOG feature descriptor. This implies that the proportions of the people were distorted. Up-scaling images can cause aliasing artefacts, especially when the new resolution is not a multiple of the original. It is possible that the cartooning method worsened the conditions of the images on the NICTA instead of improving them. These may be the reasons for the poorer performance on the NICTA. With the infrared dataset, the performance on the three datasets was not much separable. Only slight improvements were observed with the type-A cartooning method. Whilst the type-B cartooning method had the weakest score.

B. Extreme Learning Machine experiments

It is important that the outcomes of cartooning images prior to extracting HOG features are tested on different classifiers. Therefore, similar experiments that were performed with the SVMs were also carried out with ELMs. The ELMs are a variant of feed-forward neural networks. The ELMs were introduced by Huang et al. [9] in 2006, as a new learning model that could be trained faster than the pre-existing traditional neural networks. Instead of tuning all the weights on the entire network iteratively, only the output layer of the ELMs needs to be trained. More information about ELMs is available in [9], [14].

The standard ELMs are often optimised by iteratively searching for the right number of hidden neurons that works best for a specific dataset and by comparing several activations functions for the neurons. For this work, the ELMs were initially trained with 500 neurons on each dataset. Thereafter, the ELMs were iteratively trained with increasing number of hidden layer neurons. This was done to determine the optimal network-size that obtains the best classification accuracy. The iterative steps increased the number of hidden neurons with an additional 200 neurons until a maximum of 10 000.

The overall results were not harmonious with the ELMs. The type-B cartooning method performed well on the INRIA and SIGNI datasets. Poor results were expected on the low-resolution NICTA dataset as with the SVMs. The edge preserving filter appears to perform poorly on low-resolution images. Moreover, the ELMs were very sensitive to the varying depth sizes of the hidden layer and also the datasets used. For instance, adding neurons would improve the performance of the original images at times, other times type-A or type-B images. Thus leading to inconclusive outcomes. Favourably to the supposition, ELMS performed well on both cartooned images compared to the original images from the SIGNI dataset.

C. Convolutional Neural Networks experiments

Lastly, the supposition was tested on the third classifier, the CNNs. The CNNs are known to classify images better than classifiers that rely on man-made features. A thorough description of this classifier can be from [6], [12]. In our case, the AlexNet and the LeNet 5 architectures were tested. After several attempts of training the AlexNet model, high classification results could not be obtained. Perhaps the number of parameters that needed to be trained were relatively large for the small datasets used, even when employing a model that was pre-trained on the ImageNet dataset [12]. Consequently, the LeNet-5 architecture was selected. The architecture was similar to the one used in [13]. The only difference was the solver used, as the adaptive gradient (AdaGrad) solver was selected instead of the stochastic gradient descent.

Unlike the previous classifiers, the CNNs can learn their own feature maps from raw images. The choice of man-made features for this classification problem may have lead to a negative bias for the shallow learning classifiers. Because the outcomes of the CNNs were more consistent on several attempts, see Figure 5. On each attempt, the type-B SGB filter was more favourable on the colour datasets than the infrared dataset. The type-B filtering appears to degrade the quality of the grayscale infrared. Perhaps the single-channel images, are easily susceptible to the edge dislocation from the Gaussian blur step.

D. Results Analysis

For the SVMs, across all datasets, the average results for processed images (type-A and type-B) were 97.72%, compared to the 97.51% for normal images, in favour of the supposition by 0.21%. On the ELMs, the average results for processed images were 94.15%, a disadvantage of 0.35% to supposition. Then for the CNNs, the processed images obtained 99.36%, marginally in disagreement with the supposition by 0.06%. Across all classifiers, the processed images obtained 97.07% versus the 97.14% for normal images. The performance of processed images was competitive with normal images. When only considering the SVMs and CNNs, the performance becomes in favour of the supposition by 0.07%.

III. Conclusion

This work proposed a method for enhancing the quality of extracted edge features. Experiments were conducted with several classifiers to test whether the proposed method can improve the quality of edge-based features. The results show that larger datasets with better image quality and better smoothing filters are required to sufficiently test the supposition. The fact that results on the smoothed images were competitive to those obtained from the original images is encouraging for further investigation on this supposition.

Therefore, future work entails the implementation and study of more filters like the propagated image filter (PIF) and the relative total variation (RTV). These filters seem to perform well in preserving small object features and could overcome the challenge in smoothing low-resolution images. The results from the ELMs were unreliable and inconclusive. Therefore, future work will focus on comparing CNN and SVM based classifiers. Larger datasets like the ImageNet should be acquired for training denser CNNs and for improving the generalisation of classifiers. Transfer learning on some of the recent CNN architectures such as the Inception-ResNet and Inception V3 should also be considered.

Acknowledgment

The authors would like to thank Armscor for funding this work.

References

[1] H. Cho, H. Lee, H. Kang, and S. Lee. Bilateral texture filtering. *ACM Transactions on Graphics (TOG)*, 33(4):128, 2014.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[3] J. W. Davis and M. A. Keck. A two-stage template approach to person detection in thermal imagery. In *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, volume 1, pages 364–369. IEEE, 2005.

[4] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. In *ACM Transactions on Graphics (TOG)*, volume 27, page 67. ACM, 2008.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[6] H. Fukui, T. Yamashita, Y. Yamauchi, H. Fujiyoshi, and H. Murase. Pedestrian detection based on deep convolutional neural network with ensemble inference network. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 223–228. IEEE, 2015.

[7] W. Gao, H. Ai, and S. Lao. Adaptive contour features in oriented granular space for human detection and segmentation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1786–1793. IEEE, 2009.

[8] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE transactions on pattern analysis & machine intelligence*, (6):1397–1409, 2013.

[9] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1):489–501, 2006.

[10] L. Karacan, E. Erdem, and A. Erdem. Structure-preserving image smoothing via region covariances. *ACM Trans. Graph.*, 32(6):176:1–176:11, Nov. 2013.

[11] L. Karacan, E. Erdem, and A. Erdem. Structure-preserving image smoothing via region covariances. *ACM Transactions on Graphics (TOG)*, 32(6):176, 2013.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] D. Kunene. SIGNI infrared pedestrian dataset, electronic dataset. Google Drive Storage, 2017. https://goo.gl/ugbevV.

[14] D. Kunene and H. Vadapalli. Better feature acquisition through the use of infrared imaging for human detection systems. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists*, page 21. ACM, 2017.

[15] X.-Y. Li, Y. Gu, S.-M. Hu, and R. R. Martin. Mixed-domain edge-aware image manipulation. *IEEE Transactions on Image Processing*, 22(5):1915–1925, 2013.

[16] T.-H. Lin, D.-L. Way, Z.-C. Shih, W.-K. Tai, and C.-C. Chang. An efficient structure-aware bilateral texture filtering for image smoothing. In *Computer Graphics Forum*, volume 35, pages 57–66. Wiley Online Library, 2016.

[17] M. S. Nixon and A. S. Aguado. *Feature extraction & image processing for computer vision*. Academic Press, 2012.

[18] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson. A new pedestrian dataset for supervised learning. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 373–378. IEEE, 2008.

[19] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on pattern analysis and machine intelligence*, 12(7):629–639, 1990.

[20] J.-H. Rick Chang and Y.-C. Frank Wang. Propagated image filtering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10–18, 2015.

[21] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *European Conference on Computer Vision*, pages 700–714. Springer, 2002.

[22] C. Tang, C. Hou, Y. Hou, P. Wang, and W. Li. An effective edge-preserving smoothing method for image manipulation. *Digital Signal Processing*, 63:10–24, 2017.

[23] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846. IEEE, 1998.

[24] J. Weickert. A review of nonlinear diffusion filtering. In *International Conference on Scale-Space Theories in Computer Vision*, pages 1–28. Springer, 1997.

[25] H. Winnemöller, S. C. Olsen, and B. Gooch. Real-time video abstraction. In *ACM Transactions On Graphics (TOG)*, volume 25, pages 1221–1226. ACM, 2006.

[26] L. Xu, C. Lu, Y. Xu, and J. Jia. Image smoothing via l 0 gradient minimization. In *ACM Transactions on Graphics (TOG)*, volume 30, page 174. ACM, 2011.

[27] L. Xu, Q. Yan, Y. Xia, and J. Jia. Structure extraction from texture via relative total variation. *ACM Transactions on Graphics (TOG)*, 31(6):139, 2012.

[28] Q. Zhang, X. Shen, L. Xu, and J. Jia. Rolling guidance filter. In *European Conference on Computer Vision*, pages 815–830. Springer, 2014.