Soft Biometrics in Low Resolution and Low Quality CCTV Videos

T. Semertzidis, A. Axenopoulos, P. Karadimos and P. Daras

Information Technologies Institute, Centre of Research and Technology - Hellas,
6th km Charilaou - Thermi, 57001, Thessaloniki, Greece, {theosem, axenop, p.karadimos, daras}@iti.gr

Keywords: soft biometrics, CCTV, exemplars, dataset, surveillance

Abstract

Soft biometrics are biometric traits that do not offer exact human identification, however, they can provide adequate information to narrow-down the search space and give valuable insights for the subject in question. In this work, we examine the issues that emerge by analysing CCTV videos for soft biometrics and propose a methodology for extracting soft biometrics from low-quality and low-resolution video footage taken from real, street CCTV cameras. The proposed approach is based on the concept of Exemplars, that is, to find matches of the examined subject over a labelled dataset, which is able to encode the quality, colour and light variations of the surveillance images. Experiments have been conducted in a new challenging dataset that we introduce in this paper. It has been created using real CCTV footage, enhanced with a wide range of annotations from multiple people, and a manually created segmentation mask for each detection/person. This dataset is made available to scientific community for comparison and improvement of their methodologies in real-world scenarios.

1 Introduction

Biometric traits are classified as either hard or soft, based on their ability to distinguish a subject from another or not. In non-intrusive scenarios such as video surveillance, extracting hard biometrics is difficult, or even impossible task when dealing with low-resolution and low-quality videos from real CCTV Street-Scenes. In such cases, soft biometrics have been considered strong supporting cues for narrowing down the search space [1]. The soft biometric characteristics enable the enrichment of the captured subject (person) with attributes that are not fully distinctive, however, are somehow discriminative to support or eliminate some hypotheses and reduce the data to be manually examined, e.g. by the police officers [2, 3].

Surveillance cameras are in most cases uncalibrated, which makes extremely difficult to precisely measure common soft biometrics like the height or the weight of a person in the captured scene. Another soft biometric trait that is of interest is the skin tone of the subject. Automatic skin tone detection in a typical, controlled, face dataset is an easy task with high confidence and success rate. However, in the unconstrained surveillance environment, this is not the case. The different camera specifications and color models, as well as the great variations in lighting and environmental conditions significantly affect the captured colors, which makes it extremely difficult to extract a skin tone decision. In [4], Khan et. al. performed an extensive evaluation of color spaces and color classification approaches for skin color detection. Unfortunately, in a street scene surveillance scenario such as the one we examine in this paper, accurate skin tone detection and classification cannot be easily achieved using existing methodologies due to the low resolution and quality of the available color images. A set of soft biometric traits that has been recently studied is the clothes colors of the detected subjects. Inspired by the person re-identification algorithms in surveillance scenarios [5, 6], clothes colors are used as the most solid soft-biometrics in non-intrusive configurations [7, 1, 8]. However, color estimations can vary significantly due aforementioned issues as well as cultural and semantic understanding of the color palettes by different eye-witnesses. To overcome the above issues, current research endeavours are performed using a predefined set of 11 culture colors that are commonly understood: black, white, red, yellow, green, blue, brown, purple, pink, orange and grey [9].

The aforementioned color classification issues could be overcome using an appropriate supervised learning approach and correct sampling. Following this intuition, in this work, we approach the extraction of soft biometric traits as an exemplar matching problem. Motivated by [10], our aim is to provide a fast and accurate segmentation of each detected person, to enhance the sampling quality and, subsequently, the overall accuracy of the extracted soft biometric labels. Following this simple concept and having the person segmented in each detection, we achieved performance improvements over bounding box sampling in extracting features for soft biometrics.

To better support our hypothesis, a new dataset has been prepared and introduced for the first time in this paper. It consists of images with human detections extracted from real-world CCTV footage, which has been made available by the Metropolitan Police of London (MET). Our motivation was to create a dataset that covers a wide variety of street CCTV cameras in different environmental conditions. The dataset will be made available to scientific community to test and improve their algorithms in soft biometrics.

The rest of the paper is organised as follows: Section 2 describes the real street-scene dataset introduced in this paper, while Section 3 analyses the proposed methodology for soft

biometric traits extraction. Experiments are reported in Section 4 and conclusions are drawn in Section 5.

2 Data and Preprocessing

Although several publicly available datasets with pedestrian images in low resolution exist, such as the Person Re-ID (PRID) 2011[1] and the Viewpoint Invariant Pedestrian Recognition (VIPeR)[2], they do not entirely reflect the real-world case. The PRID 2011 images have been acquired using only 2 cameras located in static positions. VIPeR is a more challenging dataset, where images are taken from arbitrary viewpoints under varying illumination conditions. However, in real street-scene CCTV footage, apart from the different camera positions and lighting conditions, there is great diversity in the quality and specifications of the sensors and a wide range of environmental conditions. This makes detection of soft biometrics, especially skin and clothes colors, a very challenging task. In Figure 1 (a), the Hue histogram of LFW faces dataset[3] is given, while Figure 1 (b) depicts the respective histogram in the set of faces extracted from the dataset provided by the MET. In the second dataset, apart from the dominant Hue (the one that better reflects the average skin colour), there are also other values of Hue with high occurrence, e.g. a blue component, which correspond to cameras with different parameters, making this dataset particularly challenging. Some example faces (blurred versions are depicted in the manuscript) of the second dataset are depicted in Figure 2. In the first row, faces are extracted from cameras with a decent colour profile, while the second and third row correspond to cameras with high contrast (over exposure) and increased blue component (cool colours), respectively.

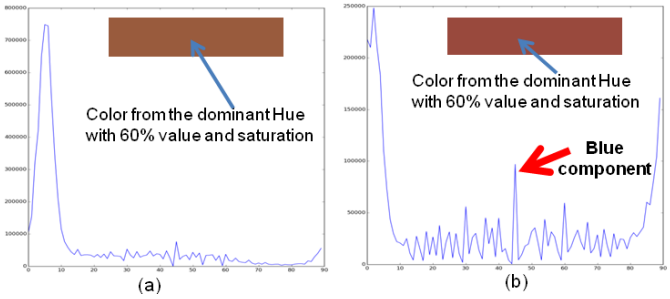

Figure 1. (a) Hue histogram of LFW faces dataset. The dominant Hue reflects the dominant skin colour of the LFW dataset. (b) Hue histogram of the set of faces extracted from the dataset provided by the MET. Note that apart from the dominant Hue, there is also a blue component with high occurrence, which reflects those images captured by CCTV with increased the blue component.

In an attempt to better address the problems described above, a new dataset has been created and introduced in this paper, based on real-world CCTV footage. The original footage

[1]https://lrs.icg.tugraz.at/datasets/prid/
[2]https://vision.soe.ucsc.edu/node/178
[3]http://vis-www.cs.umass.edu/lfw/

Figure 2. Example faces extracted from the real CCTV footage provided by the MET.

that was made available by the MET consists of a few terabytes of video sequences from a large number of CCTV cameras. The footage was preprocessed using a set of object detectors and trackers to make it searchable. The original set, contain some videos with relatively good resolution and with cameras mounted not too far from the street. However, the majority of videos are of low resolution videos with low color quality, from Pan-Tilt-Zoom (PTZ) cameras that move fast and blur the subjects, cameras that observe from far away as well as cameras the are highly affected by light variations from day to night. These facts make the dataset extremely challenging.

The annotations set has been obtained through crowdsourcing, to enhance the validity of the labels. Ten human annotators were provided with a web interface showing one single image at a time. For each image, the user could select the dominant color for the upper body and the dominant color for the lower body of the human detection as well as one or more attributes from a predefined list using check boxes. Finally, a text area enabled the users to add tags for other attributes they observed in the subject image. Images were presented randomly to the annotators. An important parameter of the experiment is that each image should have annotations from at least three different users. As soon as the experiment is over, the annotators' input is processed and each image maintains those attributes selected by the majority of users. The only annotation that performed by a single person was the segmentation mask that separates the foreground from the background for each image. This dataset can be made available for scientific purposes upon request[4].

2.1 The ITI-MET dataset

The dataset consists of 6104 images of person detections scaled in 48×128 pixels resolution. The dataset accompanied with manually extracted masks for foreground/background segmentation. Moreover, an annotations file is holding the dominant color for the upper and lower parts of the person as well as a

[4]The dataset is made available upon request to the authors by email. The requester will have to sign a form to comply with the terms of use and other legal constrains that come with the data.

	black	blue	brown	green	grey	orange
upper	44,36%	6,67%	3,47%	3,75%	9,99%	0,77%
lower	56,90%	17,87%	1,11%	0,59%	15,79%	0,16%
	pink	purple	red	white	yellow	
upper	1,77%	2,05%	5,36%	20,49%	1,31%	
lower	0,10%	0,18%	0,69%	6,54%	0,07%	
bag	front	phone	police	shorts	sleeves	stripes
37,35%	52,20%	3,69%	0,92%	6,63%	19,18%	2,21%

Table 1. Color labels distribution for upper and lower parts and number of images in other attributes

set of other attributes that can be extracted by a human annotator, like "holding bag", "shorts", "short sleeves", "clothes with stripes". Figure 3 presents a sample of the images and their masks while Table 1 presents the set of the attributes and the number of images that have each label.

Figure 3. Sample images of the MET dataset with their ground truth segmented masks.

3 Methodology

The core idea of the proposed approach is to have a fast and robust segmentation of each detection that will improve the sampling and the feature extraction process and thus the overall performance of the system. The focus of this work is mainly on the labelling of the subjects clothes colors, however, we believe that the approach may be applied to improve also other soft biometric traits in similar conditions.

3.1 Body segmentation

A small set of manually segmented body shapes/poses is required to transfer their segmentation masks to the matched detections. The approach starts by selecting these exemplar images that will create the library of segmentation masks for the online process.

It is easy to collect a large set of people detections from a video collection, however, to build an effective poses library the selection process requires a diversification approach. A possible solution to this would be a two phase clustering approach using a typical K-Means algorithm to get a set of diverse samples from the detections dataset. However, a critical point is that one should provide the K parameter i.e. the number of clusters to be formed and the number of exemplars to be selected. An easy way is to use a relatively large K value to be sure that all poses are represented. This approach may be effective but not so efficient if a manual segmentation task has to be applied to all K cluster representatives. Thus, to avoid increased manual effort, a small number of exemplar items should be selected that are still well representing the different body poses and camera angles of our sensors. For this purpose we use the Affinity Propagation clustering algorithm [11]. The Affinity propagation (AP) clustering is based on the concept of message passing between data points. The characteristics that make AP suitable is that the algorithm itself is selecting the "exemplar" data points that are the most representative of the given data without the need to define the number of clusters.

Given the set of selected exemplar poses, each exemplar image is split in 3 overlapping horizontal stripes where Histogram of Oriented Gradients (HOG) features [12] are extracted and concatenated in a single vector $\mathbf{x}$ to build a dataset X. The rest of the people detections are also processed in the same way to create a dataset Y of negative samples. As it is described in [10], each of the exemplar vector $\mathbf{x}$ train a unique Linear Support Vector Machine (SVM) classifier with one positive sample and all other vectors in X and Y as negatives, to have at the end as many exemplar SVM classifiers as the representatives extracted in the AP clustering step. We select a large parameter C of the SVM to build very strict classifiers that separate the positive sample from the others even with a small margin in the selected hyperplane. Moreover, we follow the Platt scaling [13] to generate probabilities from the decision of each classifier, that we later use to rank the classifiers' results to find the best matches. An aggregated mask that uses the best n exemplars and combine their masks in a weighted voting process is generated as:

$$A = \frac{1}{n} \sum_{i=1}^{n} w_i M_i \tag{1}$$

where M_i is each exemplar mask and w_i the weight of each mask that equals to the probability of the classification. Finally, the A aggregated mask is thresholded to keep a binary mask as foreground.

A manual segmentation step for the exemplar images is required for higher accuracy of the results as the segmentation mask of each exemplar image will be transferred to each matched detection to extract the foreground pixels.

3.2 Upper and lower body clothes colors

After extracting the masks, each detection is divided in upper and lower body parts using a naive split in the middle of the foreground object. This simplified approach is followed to evaluate the impact of the exemplar based segmentation without any other algorithmic optimisations. For each body part, a Hue-Saturation-Value (HSV) histogram is extracted. The binning of the HSV values were experimentally selected to 12 bins in Hue and 6 bins for Value and Saturation channels. The low

Figure 4. Example segmentations for three images. The first column is the original image, the second column the aggregated mask, the third column is the thresholded mask and the last column is the extracted foreground image. The first two rows show successful segmentation results while last row shows a bad match that even in this case the segmentation is acceptable.

dimensionality of the descriptor was selected to fit to the 11 culture colors that we aim to extract, since higher dimensionality descriptors were introducing large variations and were decreasing the performance of the system. Next, a linear SVM classifier with parameter $C = 2$, was trained as a multi-class classification problem with 11 classes to be the 11 culture colors. This approach aims to encode in the classifier all the color variations and environmental conditions for each color and ensure that the possible color drifts (e.g. blue seen as black or white as grey) be also modelled appropriately.

4 Experiments and Discussion

In our experiments we evaluated the quality of the segmentations using the Dice similarity between our aggregated exemplar based masks and the ground truth masks. The AP algorithm selected 109 exemplar images and these where used to train 109 exemplar based classifiers. Having as baseline the simple bounding box with 59% of average Dice similarity, the average Dice similarity for the best match exemplar was 76% and the aggregated mask reached the score of 82% of average Dice similarity. Similar results were reached with K-Means clustering for selection of the exemplars having $K = 600$ exemplar images. Figure 4 presents three example detections

	Upper body	Lower body
Bounding box	59%	62%
Proposed approach (Aggr. mask)	72%	71%

Table 2. Color Classification Accuracy using the exemplars segmentation compared to a simple bounding box.

	black	white	red	yellow	green	blue	brown	purple	pink	orange	grey
black	2543	27	7	0	12	35	19	9	3	0	**53**
white	90	1005	2	1	6	36	15	3	5	1	**87**
red	27	5	264	0	0	0	3	8	13	6	1
yellow	7	7	3	47	3	0	9	0	0	1	3
green	40	10	1	0	138	13	6	1	0	2	11
blue	**129**	50	4	0	13	172	1	8	0	0	**30**
brown	**67**	16	10	3	1	1	87	0	4	2	**21**
purple	**50**	8	5	0	0	12	2	22	10	0	16
pink	10	**19**	23	0	0	1	5	5	39	1	5
orange	6	3	4	0	0	3	12	1	4	14	0
grey	**225**	**133**	1	1	5	8	**21**	1	3	0	212

Table 3. The confusion matrix for the upper body colors with the proposed method

with their aggregated exemplar matches, the thresholded mask and the final foreground segmentation. The reader should note that the improvement is greater than it appears from the Dice similarity scores, since the detections give an already aligned person at the center of each image. The improvement permits a much clearer segmentation of the head, hands and legs of each detection and thus permit higher level soft biometrics to be extracted.

Next, a five-fold cross validation process was followed and the extracted experimental results for the upper and lower body colors are presented in Table 2 both for the simple bounding box sampling and the proposed approach. The results show an improvement both for the upper and the lower body parts over the baseline approach. An interesting result though is presented in Table 3. The Table 3 holds the confusion matrix of the upper body colors classification for the proposed method. The rows represent the ground truth colors while the columns represent the predicted colors. So, the first row presents the number of blacks that were classified as blacks as well as other colors from the classifier.

The interesting finding is that for this content quality, there are many instances of the dataset that drift to another color with consistence. These "color drifts" explain greatly the classification results as well as the crowd based annotations that we collected. In such a low resolution and low quality images, colors deteriorate to other colors and human observers (e.g. eyewitnesses) can't tell even these simple 11 colors. The table has bold values for the interesting and expected color drifts such as "black to grey" or "blue to black". This finding guide us to work further on query expansion and other directions [14, 15] to link such "typical" color drifts and improve the soft biometrics labelling for e.g retrieval purposes. Finally, another future direction is to combine the exemplar matching process to transfer also other labels.

5 Conclusions

The proposed exemplar based segmentation process found to be fast and robust to create rough segmentation masks for person detections in surveillance scenarios. The improvements over bounding box sampling for color classification supported our intuition. The methodology is simple to implement and to maintain. Moreover, with a wide number of camera sensors and from different viewing angles, a library with exemplars may be easily build and expanded. However, with the previously presented experiments we conclude that in these extreme cases of very low resolution and low color quality the color binning in 11 culture colors are some times too many. In our experiments we show that even the eye-witnesses (i.e. the dataset annotators) were drifting in between colors and thus for a classification process it is very difficult to follow.

The proposed methodology has been demonstrated in a specific soft biometric feature, the clothes colour (upper and lower body). In a similar manner, it can be easily adapted to cover a wider variety of soft biometric traits, such as skin colour, hair colour, wearing glasses/no glasses, wearing hood/no hood, even ethnicity. In all these cases, the same exemplar-based framework can be applied, focusing on the appropriate selection of the exemplars of the dataset. For example, in the case of skin colour detection, selected exemplars for each colour tone should cover all diverse colour variations as well as camera parameters of the dataset (e.g. high-contrast, cool colours with increased blue component, etc.).

Our future work will follow these findings and focus on the encoding of such "color drifts" to "query expansion" and "label expansion" schemes to support retrieval scenarios in surveillance video footage.

Acknowledgements

The work presented in this paper was supported by the European Commission under contract FP7-607480 LASIE. The authors would like to thank the Metropolitan Police of London, UK, for providing the CCTV video footage and for giving permission to process and share the dataset with the research community.

References

[1] E. S. Jaha and M. S. Nixon, "Soft biometrics for subject identification using clothing attributes," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pp. 1–6, IEEE, 2014.

[2] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Biometric Authentication*, pp. 731–738, Springer, 2004.

[3] A. Dantcheva, C. Velardo, A. Dangelo, and J.-L. Dugelay, "Bag of soft biometrics for person identification," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 739–777, 2011.

[4] R. Khan, A. Hanbury, J. Stttinger, and A. Bais, "Color based skin classification," *Pattern Recognition Letters*, vol. 33, no. 2, pp. 157 – 163, 2012.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2360–2367, IEEE, 2010.

[6] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[7] E. S. Jaha and M. S. Nixon, "Viewpoint invariant subject retrieval via soft clothing biometrics," in *Biometrics (ICB), 2015 International Conference on*, pp. 73–78, IEEE, 2015.

[8] V. Lovatsis, A. Dimou, and P. Daras, "Introducing context awareness in multi-target tracking using re-identification methodologies," *Imaging for Crime Detection and Prevention*, 2013.

[9] A. Dangelo and J.-L. Dugelay, "A statistical approach to culture colors distribution in video sensors," *Proceedings of VPQM*, 2010.

[10] T. Malisiewicz, A. Gupta, A. Efros, *et al.*, "Ensemble of exemplar-svms for object detection and beyond," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 89–96, IEEE, 2011.

[11] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[12] 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*, vol. 1, pp. 886–893, IEEE, 2005.

[13] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pp. 61–74, MIT Press, 1999.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[15] A. Joly and O. Buisson, "Logo retrieval with a contrario visual query expansion," in *Proceedings of the 17th ACM international conference on Multimedia*, pp. 581–584, ACM, 2009.