2015 IEEE International Symposium on Multimedia

A user-based framework for group re-identification in still images

Nestor Z. Salamon, Julio C. S. Jacques Junior, Soraia R. Musse
Graduate Course in Computer Science
Pontificia Universidade Catolica do Rio Grande do Sul
Porto Alegre/RS, Brazil
nzsalamon@gmail.com, julio.jacques@acad.pucrs.br, soraia.musse@pucrs.br

***Abstract*—In this work we propose a framework for group re-identification based on manually defined soft-biometric characteristics. Users are able to choose colors that describe the soft-biometric attributes of each person belonging to the searched group. Our technique matches these structured attributes against image databases using color distance metrics, a novel adaptive threshold selection and people's proximity high level feature. Experimental results show that the proposed approach is able to help the re-identification procedure ranking the most likely results without training data, and also being extensible to work without previous images.**

***Keywords*-group re-identification; soft-biometric; image retrieval**

I. Introduction

Re-identification is the task of assigning the same identifier to all instances of a particular individual captured in a series of images or videos, even after the occurrence of significant gaps over time or space [1]. This task is still a challenge, influenced by numerous real-world factors and aspects of the human being, mainly when applied to surveillance and forensic in external and crowded environments. A feasible approach is to re-identify a person by soft-biometrics attributes such as clothing and/or hair style. While several approaches try to find methods for extract and learn attributes to match the query image and the suspects, these soft-biometric attributes can be easily described by a user who saw the suspects and wants to look for them.

In this work we propose a re-identification framework applied as a soft-biometric recognition tool, based on users' color input and single shot scenes. The main idea is to look for a person with specific attributes and other near persons with distinctive characteristics (e.g. a person with red shirt close to a guy with blue pants and black cap), once this group association could be very useful for occlusion handling, to deal with ambiguities, appearance changes and view angle variations [2].

II. Related work

During a re-identification procedure where a person is asked to describe the suspect's characteristics, commonly a composite sketch is built. But the interviewed person can also describe distinctive features related to the clothes and accessories being used by the suspect as well as characteristics from other people present in the scene.

Layne et al. [3] explored the discriminative features selecting and weighting mid-level semantic attributes such as hair-style, shoe-type or clothing-style, inspired by the operating procedures of human experts. Farenzena et al. [4] analyze people's appearance weighting salient body parts based on perceptual principles of symmetry and asymmetry. Zhao et al. [5] explored human salience for re-identification through patch matching after concluding that most methods match pedestrian images by directly comparing misaligned features, caused by variations of viewpoints and poses, which leads some approaches to consider distinctive features (such as a bag or a cap) small outliers to be removed. Those small features disregarded by matching/learning approaches can be useful to distinguish target candidates as well as other contextual information like belonging groups.

The work of Zheng et al. [2] was pioneer, in the context of re-identification, to tackle the problem of matching/associating groups of people over a large space and time, captured in multiple non-overlapping camera views. However, as mentioned by the authors, the work focused on evaluating their group descriptors and an automatic group detection is required in practice.

The proposed framework allows a user-based description of the soft-biometric attributes for groups re-identification. Users are able to select the colors of each attribute and build 2D body models representing the group of people. Hence, the framework automatically indexes possible targets, comparing them against the user defined attributes in order to detect and rank the most likely groups.

III. The model

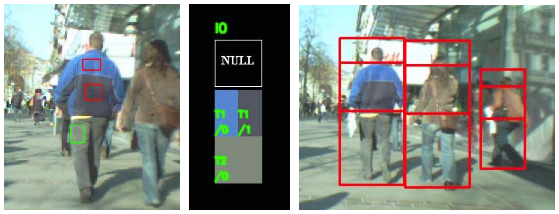
Our framework is manually initialized by a user defining the interest colors (salient or predominant) that describe the group to be searched - the color of the persons' clothes, accessories or objects they are carrying. This step can be done by selecting color samples from any source (a reference image from a gallery repository, a picture taken from the suspects or selecting from a color palette). The selected colors are then associated to a 2D body model, defined by three attributes - $head$, $torso$ and $legs$ - that will be automatically searched in the database.

978-1-5090-0379-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISM.2015.41

A. Color selection and 2D body model construction

Initially, for each person I, the user can select in a sample image up to n colors for each attribute ($head$, $torso$ and $legs$). The average of each color channel will generate the color model of the current attribute, represented by T_{km} (where $k = [0, 1, 2]$ for $head$, $torso$ and $legs$, respectively, and $m = [0, 1, ..., n-1]$ is the index number of each color associated to the attribute).

Figure 1(a) exemplifies two tones of blue selected for the $torso$ region (jacket, T_{10} and T_{11}) as well as one predominant color for the $legs$ (pants, T_{20}); none color was select for the $head$ (T_0). Figure 1(b) illustrates the generated 2D body model, built with the average color of each selection. Due to the flexibility of the framework, this step could also be achieved by selecting colors from a palette or any other gallery set, which allows to search groups without previous images.



(a) (b) (c)

Figure 1. Illustration of the color selection (a) and construction of the 2D body model to be searched (b). (c) people detection and body part division illustration in one target scene of the database.

B. People detection and body part division

Given the image database where the 2D models will be searched, the framework fills a list of the candidate persons P (all possible targets), detecting every person in the scenes using a feature detector based on HOG [6].

For each target person P in the list, the framework proportionally divides their body structure in three parts, according to the defined attributes: i) $head$ features represents 17% of the body height, ii) $torso$ represents 33% and iii) $legs$, 50%, as illustrated in Figure 1(c). These measures were obtained analyzing the fitting structure in the adopted datasets where people are viewed from a lateral camera.

C. Color Segmentation

The kernel of the framework relies on this color segmentation step: the base data generation for targets and attributes comparison. Once the color space usage is very dependent on the application, we opt to use the CIE1976 (Lab) in our model because it has colors more uniformly spaced than in RGB and HSV [7] and the Lab comparison metrics take in consideration the human perception. The ΔE_{94} method was used in order to give a color distance inside the LCh color space (being the LCh's Ch component derived from Lab's ab). Given two Lab colors, (L_1, a_1, b_1) and (L_2, a_2, b_2), the ΔE_{94} distance is defined by Equation 1:

$$\Delta E_{94} = \sqrt{\left(\frac{\Delta L^*}{k_L S_L}\right)^2 + \left(\frac{\Delta C^*_{ab}}{k_C S_C}\right)^2 + \left(\frac{\Delta H^*_{ab}}{k_H S_H}\right)^2}, \quad (1)$$

where S_L, S_C, S_H are weighting functions that adjust the CIE differences depending upon the location of the standard CIE1976 (being $S_L = 1$, $S_C = 1+K_1C^*_1$, $S_H = 1+K_2C^*_1$), and k_L, k_C and k_H are application specific parameter, defined as used in textiles applications: $k_L = 2$, $k_C = 1$, $k_H = 1$, $K_1 = 0.048$ and $K_2 = 0.014$ [8].

Finally, for each person in the list P of targets, all pixels inside each attribute ($head$, $torso$ and $legs$) are confronted - using ΔE_{94} distance - to their respective body part (T_{km}). Each body part k and selected color m will generate a distance map D_{km}. Pixels with distance smaller than a predefined threshold (defined by Th^*_{km}) are kept; otherwise they are ignored. We observed that the image quality can be very related to the adopted threshold, making its choice determinant to the success or failure of the segmentation step. Hence, to automate the threshold selection and to avoid the usage of different thresholds manually chosen, we propose a novel approach to compute Th^*_{km} by analyzing the histogram of D_{km}.

D. Adaptive threshold selection

The adaptive threshold approach here proposed is used to segment the computed distance map D_{km}. It assumes that the object to be segmented is the one with the lowest distance, the one whose color is most similar to the color model T_{km}.

The implementation is based on the work of Jacques Junior et al. [9], in which a histogram-based model is used to compute the desired threshold, given a reference color and a search region. One drawback of the original approach occurs when the region used to compute the histogram is large enough to include several pixels with small distances but a little different from the ones we want to segment, making almost the entire search region be segmented as the region of interest.

To contour such problem we propose to subdivide the distance map into small cells and then to compute a threshold value for each cell using the histogram-based approach as described in [9]. In order to divide the region of interest (each attribute) into small cells, we used the SLICO Superpixel algorithm [10]. The idea is to compute a threshold for each cell using its distances and also those from the cells connected to it, retrieving the smallest value inside each map D_{km}, as defined by Equation 2. Our hypothesis is that when the distance map is divided into small cells, there will be at least one cell in which the desired distance is isolated, generating a peak close to the origin of the computed histogram - related to the best threshold.

$$Th^{*}_{km} = K \min_{i=1 \text{ to } p_k} Th_{km}(i), \quad (2)$$

where p_k is the number of generated cells and K is the adopted scale factor, used to give some flexibility to the proposed adaptive threshold.

The number of generated Superpixels cells p_k, for each body part attribute k, is a fraction of its area ($p_k = A_k 0.015$, where A_k is the area of each attribute k, adjusted to deal with different image resolutions). Moreover, to deal with noisy images, illumination changes and a relative small number of samples, we considered a point as a maximum peak if it has the maximal value and was preceded (to the left) by a value lower by δ (based on experiments, the parameters were defined as $K = 2$ and $\delta = 0.5$). We also propose to ignore the Superpixel cells (and its pixels) connected to the vertical borders once the attributes are usually not connected to it, as highlighted in yellow in Figure 2(c). The segmentation result obtained using the proposed adaptive threshold is illustrated in Figure 2(d).

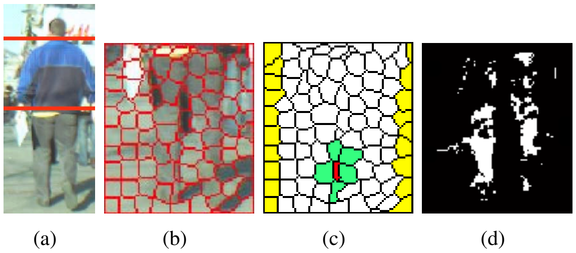


Figure 2. An overview of the proposed adaptive threshold model. (a) analyzed image divided into attributes; (b) result of the SLICO Superpixel for the $legs$ attribute; (c) the cell with minimum computed threshold (in red) and the ones connected to it (in green); (d) segmentation result using the proposed adaptive threshold approach.

In addition, to prevent problems caused by poor segmentation results, we consider a $NULL$ segmentation when the area of the segmented pixels is smaller than 1% of the area (A_k) of the attribute under analysis - too small to represent an interest feature in our experiments. With the segmentation result, the framework is able to compute person's metrics to evaluate the group re-identifications.

E. Person comparison metrics

The output of the segmentation stage is a distance map D_{km} for each body part attribute k and selected color m. From this distance map, the error value E_{km} is computed as an average distance, considering the segmented pixels inside the map. The average error E_{km} will be used to calculate the total measured error $S(I, P)$ for a target person P compared to a reference person I, as described in Equation 3.

$$S(I,P) = \sum_{s=0}^{s'-1} E_{0s} + \sum_{t=0}^{t'-1} E_{1t} + \sum_{u=0}^{u'-1} E_{2u} + \sum_{k'=0}^{k-1} W_{k'}, \quad (3)$$

where s', t' and u' are the number of colors selected by the user for each body part attribute k' and $W_{k'}$ is a penalty value computed for each body part attribute when some selected color is not found (e.g. $NULL$ segmentation result), as defined below. Such penalty value is equal to zero ($W_{k'} = 0$) for a specific body part attribute k' when the user did not assign any color to it or if all selected colors were found.

$$W_{k'} = \begin{cases} 2v\mu_{k'} & \text{if B = true} \\ 2n\mu_{k'} & \text{if B = false,} \end{cases}$$

where

$$\mu_{k'} = \frac{1}{z'} \sum_{z=1}^{z'} Th^{*}_{k'}(z), \quad (4)$$

z' is the number of adaptive thresholds, $B = true$ means the user selected m colors (up to n) for a specific body part attribute and at least one color was found, and v is the number of missed colors; $B = false$ means that the user selected m colors for a specific attribute and no one was found. Thus we make sure that when no color is found for a specific attribute, the error will be greater than any other situation where at least one color is found. These individual $S(I, P)$ metrics will be used to compute the group ranking.

F. Group detection, metrics and ranking

We define a group as a pair of individuals in a scene with distance d smaller than a threshold cTg, where d is the Euclidean distance between them (using bounding-box center as reference), Tg is set to be the smallest width between the bounding-boxes under analysis and c is the scale factor ($c = 2$) to deal with sparse or crowded scenes.

To search for a group, the user must select the desired colors of two individuals as reference (I_1 and I_2). These colors will be used to compute S_{g1} and S_{g2} error values for each detected group, using the Equations 5 and 6.

$$S_{g1} = S(I_1, P_1) + S(I_2, P_2) \quad (5)$$

$$S_{g2} = S(I_1, P_2) + S(I_2, P_1), \quad (6)$$

where $S(I_1, P_1)$ is the measured error assigned to person I_1 as the target P_1, following the same idea for $S(I_1, P_2)$, $S(I_2, P_1)$ and $S(I_2, P_2)$ in order to allow position swap. Finally, the pair of individuals with smallest S_g value is that most similar to the group we are looking for.

IV. Experimental Results

The experiments to evaluate our group re-identification are conducted using a subset of the ETHZ dataset [11], [12], built with sparse frames in which pairs of individuals captured on camera B (right, the gallery set) were detected by our group detection approach.

This procedure resulted in a database with a total of 141 detected groups, encompassing 213 individuals in 72 images of each camera. The images in the subset were also analyzed to look for individuals and groups that appear repeatedly on different scenes - simulating the user acceptance when the same person/group appears on different cameras/scenes - finding 29 individuals and 11 groups (13.61% of the individuals and 7.8% of the groups appear at least twice).

In our evaluation protocol, the scene where each individual of the detected groups appears in the camera A (left, probe) was presented to the users. The users should select at least one color for $torso$ and one color for $legs$ - the $head$ attribute was defined as optional. This step could also be achieved using a color palette, but in order to measure our results, the ground truth was defined by the input images.

Each group's signature, composed of the mean colors of each body part attribute for each individual, was then confronted to each group captured on camera B. The total error (S_g) of each target group was calculated and a ranking was built with the correct findings. To demonstrate the improvement when comparing to a person re-identification approach, we computed the average rank for each group member individually re-identified against the group rank position. Table I summarizes the results and presents the comparison, showing that two individuals can be better ranked when searched as a group.

Table I
ETHZ DATASET RESULTS: TOP RANKED MATCHING RATES (IN %) WITH 141 GROUPS AND 213 INDIVIDUALS, COMPARING GROUP RE-IDENTIFICATION IMPROVEMENTS $versus$ PERSON RE-IDENTIFICATION ($average$, WITHOUT GROUP INFORMATION)

Approach / Rank	Rank1	Rank2	Rank3	Rank4
Group re-identification	82.26	92.90	96.45	98.58
Group members individually	70.92	85.10	93.61	97.12

Despite the ETHZ dataset being widely used, several approaches (e.g. [4]) use cropped images, excluding group contextual information. To evaluate our group approach, such context was needed, requiring us to assemble a subset with the complete scene images. Other works that re-identify groups (e.g. [2]) use temporal information for background subtraction, which is not feasible in a single shot approach. Hence, comparisons to other works were not possible.

As seen in this experiment, our framework applied to group re-identification achieved a matching rate of 82.26% in the top rank with a simple color description as the search query, improving the person re-identification through the group contextual information.

V. Conclusion

In this paper we proposed a group re-identification framework applied as a soft-biometric recognition tool. The approach utilizes manually inputted color information, using sample patches/colors from any source to create a semantically organized signature of the persons inside the group to be searched. The possible matches are ranked according to the smallest color differences of their group attributes. The differences are calculated with a color distance metric improved through the novel adaptive threshold here proposed. Experimental results showed that the proposed group re-identification technique performs well in non-trivial images based on user's color inputs.

Acknowledgements

Authors would like to thank Brazilian agencies FAPERGS and CAPES for their financial support.

References

[1] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Computing Surveys*, vol. 46, no. 2, pp. 29:1–29:37, Dec. 2013.

[2] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proceedings of the British Machine Vision Conference*, London, UK, 2009, pp. 23.1–23.11.

[3] R. Layne, T. Hospedales, and S. Gong, "Person re-identification by attributes," in *Proceedings of the British Machine Vision Conference*, Guildford, UK, 2012.

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

[5] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *IEEE International Conference on Computer Vision*, Sydney, Australia, Dec 2013.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, San Diego, CA, USA, June 2005, pp. 886–893.

[7] J. Cai and A. Goshtasby, "Detecting human faces in color images," *Image and Vision Computing*, vol. 18, no. 1, pp. 63–75, 1999.

[8] A. K. R. Choudhury, *Principles of Colour and Appearance Measurement: Visual Measurement of Colour, Colour Comparison and Management*. Woodhead Publishing, 2014.

[9] J. C. S. Jacques Junior, L. Dihl, C. Jung, M. Thielo, R. Keshet, and S. Musse, "Human upper body identification from images," in *17th IEEE International Conference on Image Processing*, Hong Kong, China, Sept 2010, pp. 1717–1720.

[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. S usstrunk, "SLIC Superpixels Compared to State-of-the-art Superpixel Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, 2012.

[11] A. Ess, B. Leibe, K. Schindler, , and L. van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, June 2008, pp. 1–8.

[12] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *11th IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007.