Body Part Based Re-identification from an Egocentric Perspective

Federica Fergnani* Stefano Alletto* Giuseppe Serra* Joaquim De Mira† Rita Cucchiara*

*Universita' degli studi di Modena e Reggio Emilia
*name.surname@unimore.it
†Universidade Tecnolgica Federal do Paran, Coordenao de Eletrnica
†mira@utfpr.edu.br

Abstract

With the spread of wearable cameras, many consumer applications ranging from social tagging to video summarization would greatly benefit from people re-identification methods capable of dealing with the egocentric perspective. In this regard, first-person camera views present such a unique setting that traditional re-identification methods results in poor performance when applied to this scenario. In this paper, we present a simple but effective solution that overcomes the limitations of traditional approaches by dividing people images into meaningful body parts. Furthermore, by taking into account human gaze information concerning where people look at when trying to recognize a person, we devise a meaningful way to weight the contributions of different bodyparts. Experimental results validate the proposal on a novel egocentric re-identification dataset, the first of its kind, showing that the performance increases when compared to current state of the art on egocentric sequences is significant.

1. Introduction

Thanks to the widespread of wearable cameras and recording devices, first-person videos (often referred to as egocentric, or ego-vision) are becoming more popular, and the demand for novel algorithms to elaborate such information is rapidly increasing. In particular, recent efforts have been made towards many relevant tasks such as video summarization [14, 21], daily activities recognition [18], social relationships understanding [5, 1].

The adoption of the unique first person perspective poses new challenges to the community, conditions such as poor video quality due to ego-motion, blurriness, severe changes in lighting conditions and, ultimately, a perspective completely different from what state of the art computer vision algorithms have been designed and trained on [2]. That is why it is often required to develop new methods that are designed specifically to cope with the aforementioned issues.

Figure 1. From left to right: two images from the Viper [7] and two from our egocentric dataset

Here, we focus on the task of egocentric people re-identification, i.e. assigning the same label to the same person across multiple detections in a first-person video. While this is a fairly well-studied problem in computer vision [20], to the best of our knowledge there are no available methods specifically designed to cope with the ego-vision scenario. In particular, our experiments show that directly applying a state of the art re-identification method to an egocentric video yields poor results. This is due to the fact that common re-identification benchmarks [4, 7] are acquired using environmental cameras (i.e. video-surveillance settings) and often feature lower-resolution, full-body persons. Figure 1 shows some examples taken from a common re-identification dataset and from an egocentric scenario. The lack of publicly available egocentric datasets for re-identification further shows how that this is an understudied field.

While a broad review of recent re-identification methods is out of the scope of this work, it is useful to introduce some related work on the task. In particular, color information is one of the most widely adopted features due to its discriminative power in this context [20]. Being the apparent color

subject to changes due to variations in lighting conditions, the method by Varior *et al*. [19] proposes to encode color information in a high dimensional space invariant to illumination, effectively mapping pixels of the same color closeby. Liao *et al*. [12] propose to analyze the occurrence of local features to obtain a representation robust to changes in lighting and viewpoint. In this work, the authors fuse color (HSV histograms) and texture (SILTP descriptors) information and perform metric learning to learn both a suitable subspace where to represent the data compactly and a cross-view distance metric. On a different note the work by Zhao *et al*. [22] automatically samples patches across the body and learns a set of mid-level filters that depend on how discriminative the patch is.

Inspired by the human cognitive process, Martinel *et al*. [15] propose to build a color histogram where each pixel contributes in a way that depends from its saliency, ideally reducing the information redundancy sampling only from significant patches. Combining this feature to texture information and then reducing the resulting descriptor dimensionality via principal component analysis, the method performs the re-identification of different persons. While this is a first attempt into bringing the human in the re-identification loop, it relies on visual saliency which has been demonstrated by psychological studies to be only loosely coupled with actual human gaze [8]. Thanks to the recent advancements in human eye-tracking solutions, we propose to learn how to weight the different body parts according to what humans look at when looking at people. This is in contrast to adopting bottom-up saliency techniques that are often biased towards certain low-level image features instead of accounting for the peculiarity of human gaze.

In this work, we contribute to the research on the first-person re-identification topic in the following manners: first, we address the lack of available benchmarks by acquiring and publicly releasing an egocentric dataset featuring several different persons, with videos acquired in different scenarios and under different conditions. Frames are manually annotated with ground-truth bounding boxes identifying people locations and with personal labels identifying different subjects. Second, we design a method that exploiting current state of the art descriptors accounts for the unique perspective of first-person camera views. This is done by extracting biological information from the subjects face (i.e. facial landmarks) and dividing the body into semantically significant zones (head, torso, legs) according to the rations learned from the face. Comparing each body parts with its counterpart and using different, specifically learned metrics, we report experimental evidence of our method outperforming current state of the art re-identification approaches when challenged with egocentric videos. Third, we propose to learn the different weights to be used when combining information from the different body parts by looking at what humans do when asked to perform re-identification of unknown subjects. Using high-end commercial Eye Tracking Glasses (ETG)[1], we perform experiments with different subjects by showing them images of people and accumulating fixation points on the different body parts into a gaze map. The re-identification process we propose hence relies on this map to combine information in a meaningful and human-inspired way.

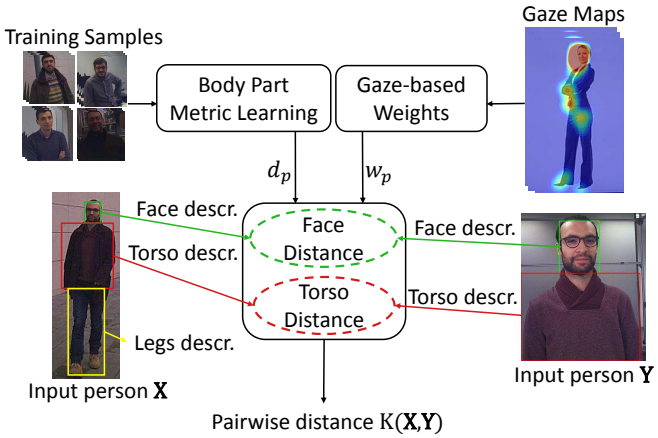

Figure 2. Proposed re-identification system

2. Proposed Method

Due to the fact that in egocentric videos a person can be either seen whole, only the upper-body or even just the head, a key insight is to divide the detected figure into different body-parts and compare them accordingly (see Section 2.1). In fact, standard methods that rely on a global description of the person often fail when challenged with the situations typical of ego-vision, which often involve comparing a upper-body query image to a full-body reference frame.

Given these premises, person re-identification can be formulated as follows. Being two persons $\mathbf{X} = \{\mathbf{x}_p\}_{p=1}^{P}$, $\mathbf{Y} = \{\mathbf{y}_p\}_{p=1}^{P}$ where $\mathbf{x}_p$ and $\mathbf{y}_p$ are the multi-dimensional feature vector presenting the appearance of the body part p with $P = \{face, torso, leg\}$, we define our kernel distance as

$$K(\mathbf{X}, \mathbf{Y}) = \frac{1}{|P|} \sum_{p=1}^{P} w_p d_p(\mathbf{x}_p, \mathbf{y}_p) \qquad (1)$$

where $|P|$ is the number of visible common body parts, $d_p(.,.)$ is the Mahalanobis distance function obtained by the metric learning component on the p body part (see Section 2.2). Since the appearance of person $\mathbf{X}$ and person $\mathbf{Y}$ can differ at the point where not every body part may be visible, our kernel distance only considers the parts in common

[1] http://www.eyetracking-glasses.com/

between the two. Another key aspect of this kernel distance is how the contributions of the different body-parts are combined, namely the parameter w_p. Accounting for the recent trend which tries to incorporate human-inspired information such as saliency in the re-identification process [15, 23], we devise a novel way to weight the body-parts by leveraging human gaze information. That is, using an ETG device we capture the fixation points of a set of people tasked with looking at images depicting different subjects and trying to re-identify them as they pass by. In fact, psychological literature [8] demonstrated that human fixations and, in a broader sense, human attention are strictly task dependent, hence the need to task the subjects with person re-identification. As a result, a fixation map for each of them is produced and by averaging the different maps it is possible to obtain a measure of where people looked at during the experiment. This averaged map is directly employed in the generation of the w_p parameter by accumulating the number of fixations in an histogram with a number of bins equal to the number of body parts and subsequently normalizing it. Figure 2 shows a schematization of our solution.

2.1. Body Part Identification and Representation

To automatically identify face, torso and legs of a detected person we propose to exploit proportion characteristics of human bodies. Since we are interested in re-identification of people who the camera wearer had social interactions with, it is useful to take into account their faces, which is also a key feature to estimate the of size the other body parts [3]. In fact, based on our preliminary experimentation, we can roughly define the torso and legs size as shown in Figure 3 (a). In order to detect the face we use the landmark detector presented in [10]. Its characteristics are suitable for the egocentric vision scenario since it works in real-time and achieves good performance even in presence of severe changes in illumination and blurriness. It solves the face alignment problem (of which landmark detection is a consequence) by using a cascade of regression trees. To cope with deformations or changes in lighting that would otherwise compromise the feature extraction, the method proceeds iteratively moving the image to a normalized coordinate system based on an initial estimate of the shape, sampling features and using them to update the shape until convergence. The choice of a particular set of regression functions allows to produce predictions that lie in a linear subspace of the otherwise high-dimensional and inherently non-convex problem of estimating the shape. Figure 3 (b) shows some results of our body part identification technique based on this landmark estimation method.

Once body parts are identified, they are described using color and texture information. In particular, we propose to use the descriptor presented in [12], due to its robustness to viewpoint and illumination changes that often occur in the egocentric domain. Differently from [12], in which it is used as a global descriptor, we employ it to separately represent each different body part. Color is a key feature to compare images or frames, but it is not robust to large illumination variations. Therefore, before extracting a HSV color histogram, a preprocessing step based on Retinex algorithm [9] is adopted. As texture descriptor, the Scale Invariant Local Ternary Pattern (SILTP) [13] is used. It is an improved version of the Local Binary Pattern (LBP) [17] that achieves more robustness to scale changes and noise. Finally, L2 normalization is applied to the resulting feature vector.

(a) (b)

Figure 3. (a) Estimated Body Proportions; b) Examples of body part identification.

2.2. Body Part Metric

Since Mahalanobis distance functions have been successful in traditional re-identification, we learn a distance metric based on the class of these functions for each body-part. Following the approaches [12, 11, 16] we define two classes of variations: the intra-body part variations Ω_I (corresponding to difference appearances of the same body part) and extra-body part variations Ω_E (corresponding to variations of different body parts). Considering Δ the distance between two samples ($\Delta = \mathbf{x}_p - \mathbf{y}_p$), each class can be modeled as a high-dimensional Gaussian density as:

$$P(\Delta|\Omega_I) = \frac{1}{\sqrt{2\pi|\sum_I|}} e^{-\frac{1}{2}\Delta^T \sum_I^{-1} \Delta}$$
$$P(\Delta|\Omega_E) = \frac{1}{\sqrt{2\pi|\sum_E|}} e^{-\frac{1}{2}\Delta^T \sum_E^{-1} \Delta} \quad (2)$$

where $\sum_I$ and $\sum_E$ are the covariance matrices. Since the pairwise differences are symmetric, both Ω_I and Ω_E have zero mean. By applying the Bayesian rule and the log-likelihood ratio test [11, 16], the distance function between two bodypart regions $\mathbf{x}_p$ and $\mathbf{y}_p$ is:

$$d_p(\mathbf{x}_p, \mathbf{y}_p) = (\mathbf{x}_p - \mathbf{y}_p)^T \mathbf{M} (\mathbf{x}_p - \mathbf{y}_p) \quad (3)$$

(a) (b) (c)

Figure 4. Images sampled from the dataset in the settings (a) Museum, (b) Laboratory, (c) Outdoor

where $\mathbf{M} = (\sum_I^{-1} - \sum_E^{-1})$ is the global, linear transformation of the feature space.

Hence, we obtain the Mahalanobis distance function correspond to estimate the covariance matrices $\sum_I$ and $\sum_E$. In our approach we adopt the solution proposed by [12], in which the problem is formulated as a generalized Rayleigh Quotient and the closed-form solution can be efficiently obtained by the generalized eigenvalue decomposition.

3. Experimental results

To evaluate the proposed approach, we create a dataset in real world scenarios and unconstrained environments (different illumination conditions, pose changes, different backgrounds, occlusions, etc.). Concerning image acquisition, several videos are recorded in different scenarios using the frontal camera of an eye-tracking device. These videos are recorded at a resolution of 960×720 pixels, 30 fps and with a duration ranging from 5 to 10 minutes. The dataset is composed of images from three different scenarios: a museum, laboratory and outdoor. Figure 4 displays some examples extracted from the dataset. Each scenario contains images of 8 persons with around 100 images of each one under different viewpoints (orientation) and in different times. The images of the dataset were obtained from the bounding box of the person returned by the DPM detector [6] applied on the video sequences. The dataset is available at the project website[2]. To perform the experiments our dataset is divided in two parts: half of the people in the training set and the other half in the test set.

Table 1 shows the results in terms of MAP (Mean Average Precision) obtained from the experiments, first displaying the overall results (obtained sampling images from every scenario) and then reporting the performance of each scenario individually. Results shown in the Table are obtained via a 10 iterations cross-validation loop, where in each iteration different people are used to training and test

Results are separately displayed for the torso and face descriptors automatically extracted by our method as in Section 2.1 (referred in the Table as *Torso descr.* and *Face descr.*). Table 1 also reports results obtained by using our approach that combines body parts with both uniform and gaze-derived weights. Comparing the results of torso and face, we can observe that the face part is more discriminative than torso with an improvement of 7%, a situation that is reflected by the human behavior during the ETG experiment. By combining different body parts, the performance increases about 7% and 10% by adopting uniform and gaze-based weights respectively. This confirms the fact that, while humans deem the face more discriminative, including information about other body parts results in better performance.

Furthermore, the state of the art re-identification method by Liao *et al*. [12] is included as comparison. In particular, the proposed method improves over this state of the art approach by 17% and 20% when adopting uniform and gaze-based weights, respectively. This improvement is mainly due to the fact that traditional approaches such as the one by Liao *et al*. are designed to cope with images acquired from environmental cameras. In particular, such images feature appearance characteristics far from the ones obtained in unconstrained ego-vision scenarios.

These differences are further remarked by the influence of metric learning on the results. In fact, when training on individual scenarios, the euclidean distance achieves better performance than using metric learning. This is due to the fact that metric learning does not have enough training data and thus cannot extract discriminative information to increase the performance for the re-identification task. When considering the full dataset (combining every scenario), the performance with metric learning obtains the best results in all the evaluated approaches but the one by Liao *et al*. This confirms the fact that the method is not suitable for egocentric vision. On the other hand, the results show that in re-identification from first-person camera views metric learning can be useful, provided that a suitable image description is employed.

To qualitatively evaluate results in terms of whether a person is correctly recognized or not by assigning to it a discrete label, re-identification can be seen as a binary classification problem in a one-versus-all paradigm. To implement a simple decision step that given the scores for a person decrees whether it is the same person wrt. a reference, we threshold the confidence score. Figure 5 shows the ROC curve of each approach considered in the evaluation. It can be seen that the best performance is achieved by the proposed approach using different weights, confirming the results of Table 1.

4. Conclusions

In this paper we presented a method for person re-identification in ego-vision scenarios. Exploiting the natural proportions of human body we obtain a lightweight division

[2] http://imagelab.ing.unimore.it/ego-reidentification

	All Scenarios		Museum		Laboratory		Outdoor	
	BPM	**No BPM**	**BPM**	**No BPM**	**BPM**	**No BPM**	**BPM**	**No BPM**
Torso descr.	0.531	0.478	0.584	0.777	0.536	0.612	0.405	0.485
Face descr.	0.603	0.593	0.658	0.738	0.624	0.718	0.486	0.628
Our approach (uniform weight)	0.675	0.613	0.726	0.819	0.566	0.601	0.531	0.534
Our approach (gaze weights)	0.703	0.623	0.643	0.846	0.662	0.711	0.521	0.597
Liao *et al*. [12]	0.496	0.536	0.469	0.664	0.630	0.643	0.426	0.558

Table 1. Comparison between the results of our method when dealing with individual body parts (Torso descr., Face descr.), when combining them with uniform and gaze-derived weights and the method by Liao *et al*. Results are reported both when adopting body part metric learning (BPM) and when disregarding it.

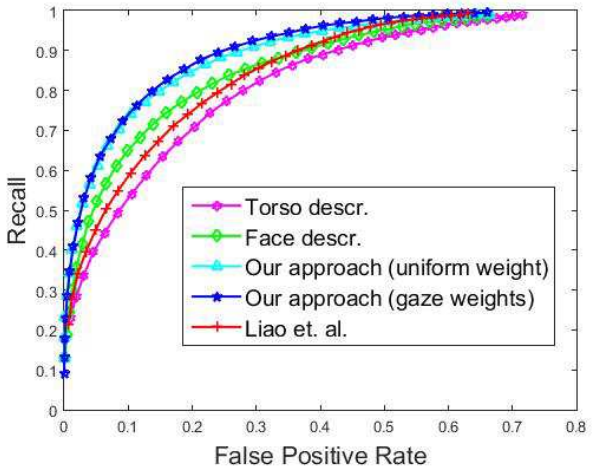

Figure 5. ROC curve showing the performance of the different methods under varying recognition thresholds.

of body parts that only relies on a real-time estimation of facial landmarks and does not require complex deformable part models. Through this division, the proposed method is able to cope with the fact that in first-person videos subjects are often seen from very different viewpoints and at different scales. Furthermore, human gaze information is exploited to weight the contributions of the different body parts according to what people look at when trying to re-identify people, improving performance in a semantically meaningful way. Finally, a new egocentric dataset for people re-identification is released, which to the best of our knowledge is the first of its kind.

References

[1] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara. Understanding social relationships in egocentric vision. *Pattern Recognition*, 48(12):4082–4096, 2015. 1

[2] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: A survey. *IEEE Transactions Circuits on and Systems for Video Technology,*, 25(5):744–760, 2015. 1

[3] B. Bogin and M. I. Varela-Silva. Leg length, body proportion, and health: a review with a note on beauty. *International Journal of Environmental Research and Public Health*, 7(3):1047–1075, 2010. 3

[4] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *Proc. of BMVC*, 2011. 1

[5] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Proc. of CVPR*, 2012. 1

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 4

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. of ECCV*. 2008. 1

[8] J. M. Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003. 2, 3

[9] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7):965–976, 1997. 3

[10] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc. of CVPR*, 2014. 3

[11] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Proc. of CVPR*, 2012. 3

[12] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *PRoc. of CVPR*, 2015. 2, 3, 4, 5

[13] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Proc. of CVPR*, 2010. 3

[14] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proc. of CVPR*, 2013. 1

[15] N. Martinel, C. Micheloni, and G. L. Foresti. Saliency weighted features for person re-identification. In *Computer Vision-ECCV 2014 Workshops*, pages 191–208. Springer, 2014. 2, 3

[16] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000. 3

[17] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on fea-

tured distributions. *Pattern recognition*, 29(1):51–59, 1996. 3

[18] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Proc. of CVPR*, 2012. 1

[19] R. R. Varior, G. Wang, and J. Lu. Learning invariant color features for person re-identification. *arXiv preprint arXiv:1410.1035*, 2014. 2

[20] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 46(2):29, 2013. 1

[21] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proc. CVPR*, 2015. 1

[22] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–151, 2014. 2

[23] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by saliency learning. *arXiv preprint arXiv:1412.1908*, 2014. 3