Re-Identification of Visual Targets in Camera Networks A Comparison of Techniques

Dario Figueira and Alexandre Bernardino

Institute for Systems and Robotics,
Instituto Superior Técnico,
1049-001 Lisboa, Portugal
{dfigueira,alex,}@isr.ist.utl.pt

Abstract. In this paper we address the problem of re-identification of people: given a camera network with non-overlapping fields of view, we study the problem of how to correctly pair detections in different cameras (one to many problem, search for similar cases) or match detections to a database of individuals (one to one, search for best match case). We propose a novel color histogram based features which increases the re-identification rate. Furthermore we evaluate five different classifiers: three fixed distance metrics, one learned distance metric and a classifier based on sparse representation, novel to the field of re-identification. A new database alongside with the matlab code produced are made available on request.

Keywords: Re-Identification, distance metrics, pattern recognition, visual surveillance, camera network

1 Introduction

Re-identification is still an open problem in computer vision. The enormous possible variations from camera to camera in illumination, pose, color or all of those combined, introduce large appearance changes on the people detected, which make the problem very difficult to overcome.

Re-identification denotes the problem of given multiple cameras, and several people passing in front of several cameras, to determine which person detected in camera X corresponds to the person detected in camera Y.

There are a few works in the literature addressing the problem of re-identification in camera networks. [9] uses the bag-of-visterms approach, clustering SIFT [8] features into "words", and using those "words" to describe the detections, in a one to one approach to re-identification in a shopping center environment. This approach is of interest because it merges the "very high detail/specificity" of a SIFT feature with the generalization power of a cluster (a "word"). [5] uses SURF [1] features also in a one to one approach to re-identification in a shopping center environment (CAVIAR database[1]). SURF's are extracted from the image's hessian space, as opposed to SIFT's features that are extracted from the image's laplacian space. SURF's are also much faster to

[1] http://groups.inf.ed.ac.uk /vision/CAVIAR/CAVIARDATA1/

be computed, which is of note since SIFT's major drawback is its heavy computation time. Both works use a voting classifier that is already standard when using such features. Also of note is [10] for its simple features, histograms, used also in a one to one approach, to re-identify people in similar poses walking inside a train, testing different histograms and normalizations to cope with greatly varying illumination. They employ dimensionality reduction and nearest-neighbor for a classifier. [6] also uses simple histogram features but takes advantage of the appearance and temporal relationship between cameras.

Our work is similar to [6, 10]'s in term of used features (histograms), although we enrich them by considering the upper and lower body parts separately. A method to detect the waist of a person is proposed, which allows the separate representation of the colors on the upper and lower body parts. This aspect highly distinguishes our work and significantly improves the re-identification results. Also we test additional metric distances including one learned from the data. Furthermore we apply a sparse base classifier, successful in the domain of face recognition, to the re-identification problem.

In this work we not only consider the one-to-one approach, where given a person detection we want to recognize it in a database, but also a one-to-many approach, where someone is trying to find similar matches in the system to a given person's image.

We produced an indoors dataset where we evaluate several techniques and the effects of our enriched feature, in both approaches.

In the next Section we define the problem. In Section 3 we propose our new color based feature. In Section 4 we list the metrics reviewed. In Section 5 we describe the data used and show the experimental results. Finally we conclude in the last Section.

2 Problem Definition

In this Section we define our problem, while the approaches are described in the following sub-sections.

Figure 1 depicts the environment: A network of fixed cameras with non-overlapping fields of view, where people appear more than once, are detected, tracked while in camera view, and then re-identified between different cameras. While a person is in view, we track it, and extract the following feature from it:

1. Detect the person and extract its pixels with [2]'s background subtraction;
2. Normalize the color of the detection pixels with greyworld normalization [10];
3. Divide the image in two by the waist (detailed in Section 3);
4. Compute the color histograms of each part, and unit normalize them;
5. Compute the mean (μ) and covariance (Σ) for all histograms in the track sequence to obtain one point per track sequence.

Therefor each track becomes an averaged histogram feature, which will then be a point ($\mathbf{x}$ or $\mathbf{y}$) in the following formulations.

2.1 One to One Problem - Recognize - "Who is this person?"

This is the standard re-identification approach, used in [9, 5, 10], and applied in real world situations, where a surveillance operator picks out a person detection and asks

Fig. 1. Two camera views, with several detections, waist-separated, and two tracks. A correct re-identification would cluster them together or label them as the same individual from a database.

the system to recognize it. This approach is of particular interest in scenarios where you have a controlled entrance to the system. In such an entrance we can easily insert the incoming individuals into a database along with their identifications.

So in this case we always have a training set, a dataset of labeled tracks to start with. Given a test sample we compute the best match to the classes in the training set for such test sample.

2.2 One to Many Problem - Search - "Where was this person?"

In this work we also consider the one to many approach, where a surveillance operator sees a person in a camera (identified or not), and asks the system to show him all related detections, in all cameras.

Given the track points, we compute the distances from all to all, then solve the binary classification problem of determining, given an appropriate distance metric, if a

pair of tracks belongs to a single person, or come from different people. By varying the distance threshold that determines if such a distance is small enough to belong to a pair of tracks from a single person, we output a ROC curve. Thus examining the ability for each re-identification technique to correctly cluster points.

We also study the advantage of learning a metric from data versus a standard distance metric.

3 Person Representation

Simple color histograms have been used as the appearance features in tracking across cameras [6, 10]. We enrich such features by dividing the person histogram in two parts, one above and one below the waist. We define the waist as the point that maximizes the Euclidean distance between the upper part histogram and lower part histogram. After computing the integral histogram, we do a vertical search for the waist, limiting the search to an area around the middle of the image.

- Compute Vertical Integral Histogram
  - Compute the histogram of the first horizontal line of the image; compute the histogram of the first two lines of the image; ...; compute the histogram of the whole image.
- Search for Point that Maximizes Distance Between Upper and Lower Body Parts
  - Compare line by line, the upper and lower histograms; Plot the varying distance; Find maximum.
  - Limit search to window between 35 and 60% of the image, counting from the top (maximum and minimum empirical values of position of the waist found during the manual labeling of the dataset).

4 Re-Identification Techniques

In this section we describe the methods used in the automatic re-identification system. We compared three distance metrics: Euclidean; Bhattacharya; and diffusion distance [7] with one linear metric learning method, and one recent classification method [11] developed in the face recognition field. Each track is represented by $\mathbf{x}$ or $\mathbf{y}$, as stated in Section 2.

We consider the following metrics to compute distance between the histograms.

Euclidean A simple neareast neighbor distance. $d_E(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||$.

Bhattacharya Modified Bhattacharya coefficient [3], a common choice for measuring distance between histograms.

$$d_{BHATT}(\mathbf{x},\mathbf{y}) = \sqrt{1-\sum_{i=1}^{m}\sqrt{x_i y_i}}$$

Diffusion Distance Given we use histograms for features, we looked for alternative ways to measure distances between histograms. The Earth's Mover Distance (EMD) is popular, and the Diffusion Distance [7] has been shown to have equal or better results than EMD, with the added benefit of faster computation.

$$
\begin{aligned}
& d_{DIFF}(\mathbf{x},\mathbf{y}) = \sum_{l=0}^{L} |d_l| \\
& d_0 = \mathbf{x} - \mathbf{y} \\
& d_l = [d_{l-1} * \phi(d_{l-1})] \downarrow_2 \ \ l = 1,...,L \\
& \downarrow_2 \ : \text{downsample to half-size} \\
& \phi(.) : \text{gaussian filter}
\end{aligned}
$$

Metric Learning The work of Xing in [12] linearly learns a metric of the following form:

$$d_{ML}(x,y) = ||x-y|\,|_A = \sqrt{(x-y)^T A (x-y)},$$

by solving the following optimization problem

$$
\begin{aligned}
\arg\max_{A} \quad & \sum_{(i,j)\in D} \left\|x_i - x_j\right\|_A = \sum_{(i,j)\in D} \sqrt{d^{ij^T} A\, d^{ij}} \\
\text{s.t.} \quad & \sum_{(i,j)\in S} \left\|x_i - x_j\right\|_A^2 = \sum_{(i,j)\in S} d^{ij^T} A\, d^{ij} \leq t \\
& A \geq 0
\end{aligned}
$$

where t is a scalar, S and D are square binary matrixes that represent the similar (from a same person) and dissimilar (from different people) training pair sets (in S, one if a pair is similar, and zero otherwise. Likewise in D). We implemented this optimization problem in MATlab with the CVX's optimization toolbox[2]. A training set of similar and dissimilar pairs is required.

Sparse Recognition Classifier Sparsity has been widely used in signal processing for reconstruction[4]. Here we apply it to recognition, in the form of a re-identification problem. Simply put, given a test sample $\mathbf{y}$, we solve the optimization problem

$$
\begin{aligned}
& \arg\min_{[\mathbf{i}\ \mathbf{e}]} \quad ||[\mathbf{i}\ \mathbf{e}]\,|_1 \\
& \text{s.t.} \quad [A\ I]\,[\mathbf{i}\ \mathbf{e}]^T = \mathbf{y} \\
& A = [x_1 \ldots x_T]
\end{aligned}
$$

where in the columns of A are the T training samples (*i.e.,* three random histogram vectors from three random detections from each person to be recognized). The reasoning behind this formulation is that we wish to choose from A which training class $\mathbf{y}$ belongs to. This information will be encoded in the indicator vector $\mathbf{i}$, while

[2] http://cvxr.com/cvx/

errors will be explicitly modeled in $\mathbf{e}$. Moreover, by minimizing the l_1 norm of $[\mathbf{i}, \mathbf{e}]^T$, and if the true solution is sparse, the l_1 norm minimization will output the same result as the l_0 norm [4], the sparsest solution. $\mathbf{i}$ will then be mostly zero with few large entries in the correct training set entries.
This idea has first been put forth in the field of face recognition by [11].

5 Results

In this Section, first we describe the experimental setup where the datasets were taken. We present our results in sub-sections 5.3 and 5.4, we validate our proposed feature, and we discuss the results.

5.1 Experimental Setup

We produced a database of images, extracted from an indoors camera network, completely hand labeled for ground truth. In Figure 2 we show some samples of the varying camera views from the dataset.

Indoor Dataset:

- 17388 detections;
- 275 tracks;
- 26 people (5 of which only have one track).
- 10 fixed cameras, with non-overlapping fields of view.
- Average 67 detections per track, Maximum 205 detections in a track.

Fig. 2. Indoor Dataset: Preview of some camera views.

For all experiments we computed the detections and features as described in Section 2. These combine graycolor normalization and division by the waist. Figure 3 supports the choices made.

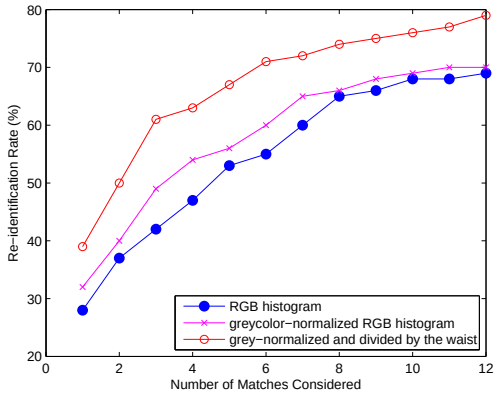

Fig. 3. Cumulative Matching Characteristic curve of several feature combinations with Euclidean metric. Re-identification results for comparing the following features: -Just RGB histogram (blue balls); Just greycolor-normalized RGB histogram (pink crosses); Greycolor-normalized RGB histogram divided in two by the waist of each person detection (red circles).

5.2 Training

For the metric learning training and testing we used 5-fold cross validation. For Sparse Recognition Classifier training, we picked 3 detection points per person to form the training class matrix (A).

5.3 One to Many Experiments

In Figure 4 we plot the ROC curves for the problem of binary classification "same/not-same pair?" described in Sub-Section 2.2. Initially, distances from all tracks to all tracks are computed. Then a threshold, that determines "similarity" is varied. For each threshold value a True Positive Rate value and a False Positive Rate is computed. All these values plot the Receiver Operating Characteristic (ROC) curves shown in the following Figure 4.

We see in Figure 4 that the baseline Euclidean distance rivals diffusion distance or bests the other techniques. Metric learning performs better as expected since it uses additional information. Learning does seem to improve re-identification and despite requiring labeled data for metric learning, this needs only be done once per system configuration.

Comparing the use of our waist-division feature (full-line) with the counterpart of a single histogram per detection (dashed-line) it is clear the positive influence our feature has in the results of all techniques.

5.4 One to One Experiments

In Figure 5 we plot the Cumulative Matching Characteristic curve for all the tecniques implemented, and also analyze the effect of our suggested improvement on the feature,

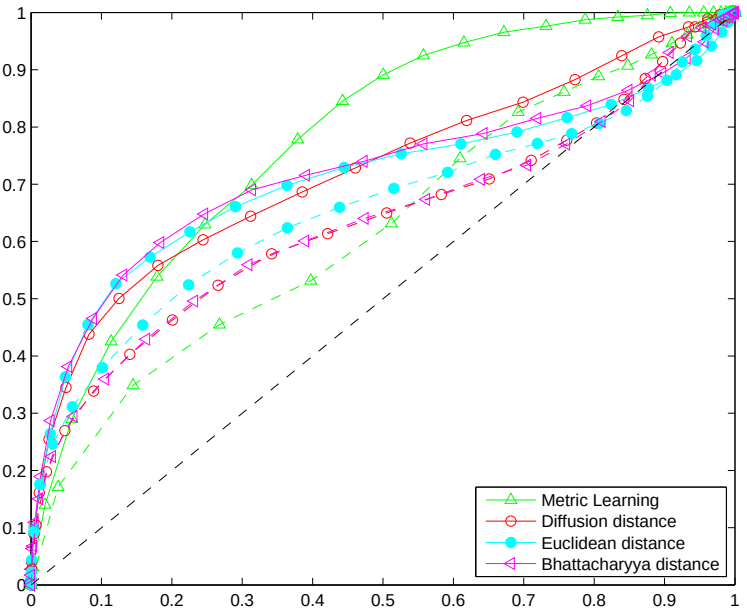

Fig. 4. ROC curves for comparing the different techniques, and confirm the improvement in the results of our suggested feature. Full line: using our waist division feature; Dashed line: not using waist division.

the waist division described in Sub-Section 3. Using our feature improves the results on all tecniques.

Due to the nature of the SRC algorithm, that outputs a sparse solution of one training class per test sample, it is not possible to plot a Cumulative Matching Characteristic curve. Nevertheless SRC gives the best results on the nearest-neighbor level.

6 Conclusions

In this work we addressed the re-identification problem. We not only consider the classical one to one problem, but also the one to many case that may be of interest for practical surveillance applications. In this case someone tries to find similar matches in the system of a given person's image; and the one-to-one approach, where given a person detection we recognize it in a database.

We built upon previous work [10], enriching the feature used by considering upper and lower body parts separately, thus improving the re-identification results.

Metric learning showed promissing results in the one-to-many approach, expectedly better than the other distances since it makes use of more information (labeled similar-

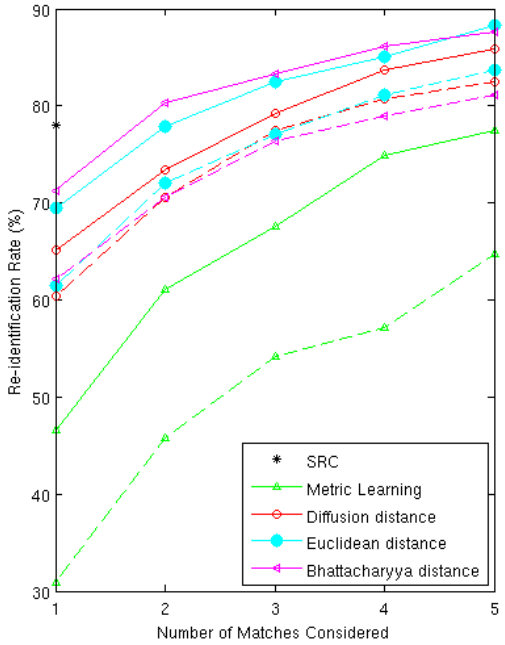

Fig. 5. Cumulative Matching Characteristic curves for comparing the different techniques, and confirm the improvement in the results of our suggested feature. Full line: using our waist division feature; Dashed line: not using waist division.

ity/dissimilarity pairs). In the one-to-one approach SRC reports the best re-identification rates.

Simple Euclidean distance rivaled or bested the other techniques for re-identification in both approaches.

6.1 Future Work

In the future, we will further enrich the set of features by either adding further body-part selection or integrating SIFT or SURF features to form a multi-modal feature. We will also take advantage of spatiotemporal constraints and appearance correlations between cameras, to limit the search space of re-identification, reducing errors.

We make available by request the matlab source code of the methods developed in this paper as well as the image database produced.

Acknowledgements

This work was supported by project the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds, partially funded with grant SFRH/BD/48526/2008, from Fundação para a Ciência e a Tecnologia, and by the project CMU-PT/SIA/0023/2009 under the Carnegie Mellon-Portugal Program.

References

[1] Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: In ECCV. pp. 404–417 (2006)
[2] Boult, T.E., Micheals, R.J., Gao, X., Eckmann, M.: Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings. In: Proceedings Of The IEEE. vol. 89, pp. 1382–1402 (October 2001)
[3] Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. vol. 2, pp. 142 –149 vol.2 (2000)
[4] Donoho, D.L.: For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Comm. Pure Appl. Math 59, 797–829 (2004)
[5] Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. pp. 1 –6 (sep 2008)
[6] Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking across multiple cameras with disjoint views. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. pp. 952 –957 vol.2 (2003)
[7] Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 246–253. IEEE Computer Society, Washington, DC, USA (2006)
[8] Lowe, D.G.: Distinctive image features from scale-invariant keypoints (2003)
[9] Teixeira, L.F., Corte-Real, L.: Video object matching across multiple independent views using local descriptors and adaptive learning. Pattern Recognition Letters 30(2), 157 – 167 (2009), video-based Object and Event Analysis
[10] Truong Cong, D.N., Achard, C., Khoudour, L., Douadi, L.: Video sequences association for people re-identification across multiple non-overlapping cameras. In: ICIAP '09: Proceedings of the 15th International Conference on Image Analysis and Processing. pp. 179–189. Springer-Verlag, Berlin, Heidelberg (2009)
[11] Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(2), 210 –227 (feb 2009)
[12] Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning, with application to clustering with side-information. In: Advances in Neural Information Processing Systems 15. pp. 505–512. MIT Press (2002)