Saliency based Person Re-Identification in Video using Colour Features

Srujy Krishna A U
PG Student
Department of Computer Science and Engineering
Federal Institute Of Science and Technology

Shimy Joseph
Assistant Professor
Department of Computer Science and Engineering
Federal Institute Of Science and Technology

Abstract

Human re-identification is to match persons observed in non-overlapping camera views with visual features for inter camera tracking. Person re-identification in a non-overlapping multi-camera scenario is an open and interesting challenge. Humans are inherently able to sample those relevant people's details that allow us to correctly solve the problem in a fraction of a second. But the task can hardly be completed by machines. Human salience is distinctive and reliable information in matching persons across disjoint camera views. So a saliency based approach can be used to re-identify persons in video. The model includes three stages: I) input processing, ii) video processing, iii) matching. The proposed person representation combines visual features either considering or not the saliency. The proposed approach has been extensively evaluated on 3DPes public datasets.

Keywords- Person re-identification, person detection

I. INTRODUCTION

Human eyes can recognize person identities based on some small salient regions. Human salience is distinctive and reliable information in matching persons across disjoint camera views. Human eyes can easily pick up one person from other candidates because of the distinctive features. These features can be reliably detected across camera views. If a body part is salient in one camera view, it usually remains salient in another view.

The existing methods for person re-identification includes mainly in two category: i) bio-metric based and ii) appearance based. In biometric based method, biological features are considered to match the persons. Appearance-based methods exploit appearance features by assuming that people do not change clothes as they walk between camera FoVs. Since the person re-identification problem can be viewed as an association problem where the goal is to track persons across camera FoVs, this is a reasonable assumption.

Appearance based methods can be further categorized into: (I) discriminative signature based methods, (ii) feature transformation based methods and (iii) metric learning based methods. Discriminative signature based methods seek for highly distinctive representations to describe a person appearance under varying conditions. This type of methods addressed the problem by using human defined representations that are both, distinctive and stable under changing conditions between different cameras. However, the exploited visual features are not invariant to the large variations that affect the images acquired by disjoint cameras. Features transformation based methods have addressed the re-identification problem by modelling the transformation functions that affect the visual features acquired by disjoint cameras. Such methods have shown to be able to capture the transformation of features occurring between cameras, however, they still face problems when large intra-camera feature variations are present. The learning process used to capture such transformation is usually highly time consuming, hence not suitable for a real deployment.

Metric learning based algorithms lie in between the two aforementioned categories. Methods belonging to such a group still rely on particular features but also advantage of a training phase to learn non-Euclidean distances used to compute the match in a different feature space. The proposed method is based on this technique. Recently, visual saliency based algorithms have been investigated for re-identification purposes. Another method is region-based methods that usually split the human body into different parts and extract features for each part. Maximally Stable Colour Regions (MSCR) is extracted, by grouping pixels of similar colour into small stable clusters. Then, the regions are described by their area, centroid, second moment matrix, and average colour. The Region Covariance Descriptor (RCD) has also been widely used for representing regions. In RCD, the pixels of a region are _rst represented by a feature vector which captures their intensity, texture, and shape statistics. The so-obtained feature vectors are then encoded by a covariance matrix.

The main contribution of this proposed method is the introduction of a simple yet effective approach to person re-identification in video using colour features. Here a multilayer approach is used.

The rest of the report organized as follows. In section II, an overview of the related works. The existing methodology and the proposed method are described in the section III. Experimental results and analysis is described in section IV. Finally, conclusion and future work are drawn in section V.

.

II. RELATED WORKS

In [1] T. Ojala et al. presents a theoretically very simple yet efficient multi resolution approach to gray scale and rotation invariant texture classification based on local binary patterns and nonparametric discrimination of sample and prototype distributions. Derive a generalized gray scale and rotation invariant operator presentation that allows for detecting the uniform patterns for any quantization of the angular space and for any spatial resolution, and present a method for combining multiple operators for multi resolution analysis.

This method starting from the joint distribution of gray values of a circularly symmetric neighbor set of pixels in a local neighborhood, then derive an operator that is by definition invariant against any monotonic transformation of the gray scale. Rotation invariance is achieved by recognizing that this gray scale invariant operator incorporates a fixed set of rotation invariant patterns. The proposed texture operator allows for detecting uniform local binary patterns at circular neighborhoods of any quantization of the angular space and at any spatial resolution. This is a theoretically and computationally simple approach.

In [2] C. Koch et al. propose a method for bottom up visual saliency. A Graph-Based Visual Saliency (GBVS) is proposed. It consists of two steps: first forming activation maps on certain feature channels, and then normalizing them in a way which highlights

Conspicuity and admits combination with other maps. Feature map is computed by linear filtering followed by some elementary nonlinearity. Activation map is computed using Markova approach. Normalization is done through concentrating mass on activation maps

The model is simple and biologically plausible insofar as it is naturally parallelized. This model powerfully predicts human fixations on 749 variations of 108 natural images, achieving 98% of the ROC area of a human-based control. That is GBVS predicts human fixations more reliably than the standard algorithms. The model exploiting the computational power, topographical structure, and parallel nature of graph algorithms to achieve natural and efficient saliency computations. The model is more reliable and Highlights salient region away from object border. But Object boundaries are not clear in this methodology, Reduce spatial frequencies in the original image and computationally quite expensive

In [3] E. O. Postma et al. proposes a dimensionality reduction technique. Here a comparative study of the most important linear dimensionality reduction technique (PCA), and twelve front ranked nonlinear dimensionality reduction techniques are presented. The Experimental result of this approach shows that, local techniques for dimensionality reduction perform strongly on a simple dataset such as the Swiss roll dataset but on some natural datasets, the classification performance of proposed classifiers was not improved by performing dimensionality reduction. The result of this approach is that nonlinear techniques for dimensionality reduction do not yet clearly outperform traditional PCA.

In [4] P M Roth et al. proposes a pairwise metric learning approach taking advantage of the structure of the data. The model consists of three stages: (1) feature extraction, (2) metric learning, and (3) classification. During training the metric between two cameras is estimated, which is then used for calculating the distances between an unknown sample and the samples given in the database. This model introduces the metric learning technique but require more computational cost.

In the first stage person image representation is create using HSV and Lab color channels as well as Local Binary Patterns. The features are extracted from 8x16 rectangular regions sampled from the image with a grid of 4x8 pixels. Then mean values per color.

Channel and histogram of LBP codes is generated. These values are used to form feature vector and by concatenating feature vectors of all regions, whole image representation is created. Metric learning step is done by calculating Mahalanobis metric during training.

This metric is used to calculate the distance between two samples. In the third stage, calculating the distances between the probe image and all gallery images using the learned metric, and returning those gallery images with the smallest distances as potential matches.

In [5] R Zhao et al. proposes an approach that use transferred metric learning technique for person re-identification. Five types of low-level visual features are used here. They include dense color histograms, dense SIFT, HOG, Gabor and LBP. This approach use adaptive metric learning but different visual metrics should be optimally learned for different candidate sets. This approach includes

Two key steps: searching and weighting nearest training samples for each candidate; and learning an adaptive metric for each candidate set. Given a large training set, the training samples are selected and reweighted according to their visual similarities with

The query sample and its candidate set. A weighted maximum margin metric is online learned and transferred from a generic metric to a candidate-set-specific metric.

In [6] M Kostinger et al. proposes a metric learning technique. They introduce a simple though effective strategy to learn a distance metric from equivalence constraints, based on a statistical inference perspective. This method is motivated by a statistical inference.

Perspective based on a likelihood-ratio test. The resulting metric is not prone to over-fitting and very efficient to obtain. But have to learn a Mahalanobis metric.

Proposed method considers two independent generation processes for observed commonalities of similar and dissimilar pairs. The dissimilarity is defined by the plausibility of belonging either to one or the other. From a statistical inference point of view the optimal statistical decision whether a pair is dissimilar or not can be obtained by a likelihood ratio test. A high value of

likelihood ratio means that null hypothesis is validated. In contrast, a low value means that null hypothesis is rejected and the pair is considered as similar.

In [7] K Zhang et al. proposes a simple yet effective and efficient tracking algorithm with an appearance model based on features extracted from the multi-scale image feature space with data-independent basis. This model employs non-adaptive random projections that preserve the structure of the image feature space of objects. A very sparse measurement matrix is adopted to efficiently extract the features for the appearance model. The tracking task is formulated as a binary classification via a naive Bayes classifier with online update in the compressed domain. The proposed method perform well in accuracy, robustness and speed. But it is complex in implementation. Object tracking remains a challenging problem due to appearance change caused by pose, illumination, occlusion, and motion, among others. Tracking algorithms can be generally categorized as either generative or discriminative based on their appearance models. Generative tracking algorithms typically learn a model to represent the target object and then use it to search for the image region with minimal reconstruction error. Discriminative algorithms pose the tracking problem as a binary classification task in order to find the decision boundary for separating the target object from the background. This model is generative as the object can be well represented based on the features extracted in the compressive domain. It is also discriminative because we use these features to separate the target from the surrounding background via a naive Bayes classifier. Features are selected by an information-preserving and non-adaptive dimensionality reduction from the multi-scale image feature space based on compressive sensing theories. This model use a very sparse measurement matrix that satisfies the restricted isometric property, thereby facilitating efficient projection from the image feature space to a low-dimensional compressed subspace. For tracking, the positive and negative samples are projected with the same sparse measurement matrix and discriminated by a simple naive Bayes classifier learned online.

III. PROPOSED WORK

The existing methodology can apply only on the images from two disjoint cameras. In order to extend the system into person re-identification in video, a saliency based person re- identification method is proposed. The architecture of the proposed methodology is shown in fig. 3.1

A. Input Processing

There are two features are used in this methodology, colour space and saliency.

1) Saliency Computation

Saliency computation is an important task in this methodology. For saliency kernalized graph based saliency detection method is used. KGBVS is a kernalized form of graph based visual saliency. Kernelized Graph-Based Visual Saliency include following steps: (i) Salient image points are detected by means of a Markov chain approach in which the transition probabilities are proportional to the features dissimilarity. (ii) Neighbouring image points having high dissimilarity are grouped together using a Markov chain approach. (iii) The final saliency master map is computed as the weighted sum of the saliency maps obtained for the different feature. (iv) Different kernels are used in the computation of transition probabilities and (v) the saliency computation benefits from a visual saliency prior related to the person localization and shape.

Let $I \epsilon R^{m\,X\,n}$ be the image of a person and let assume that the silhouette stands somewhere in the center of it. Also, let $F \epsilon\ R^{m\,X\,n}$ be a feature map such that an element $F_{x,y} = \pi(I, x,\ y)$, where $\pi(\,)$ is a feature extraction function (e.g., wavelet transform, filter response, edge detector, etc.). Then, an activation map $A\ \in R^{m\,X\,n}$ is computed such that an element $A_{z,y}$ has high value if $(\,x, y)$ is

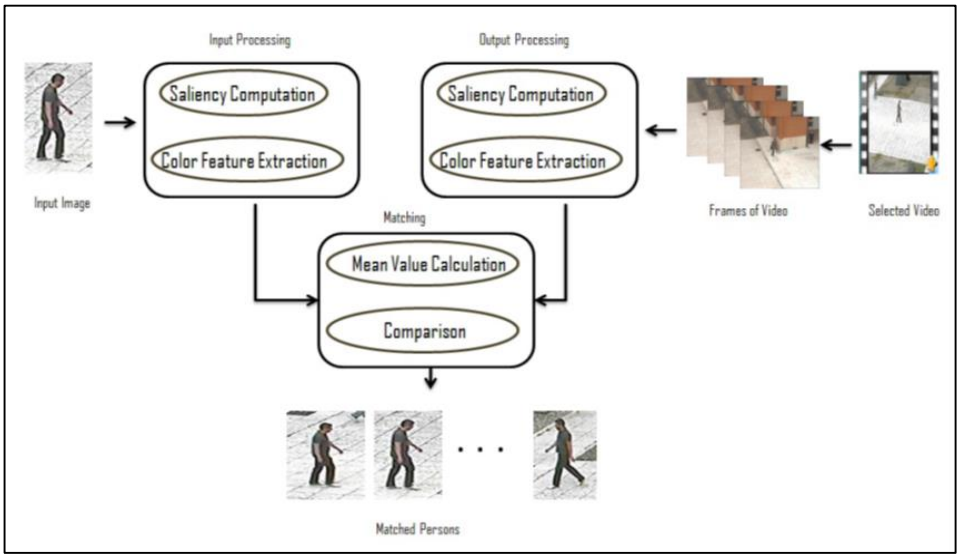

Fig. 3.1: Proposed system architecture

In the center of the image and neighboring values of $F_{x,y}$are different one to each other. Let $GF = (V, E)$ be a fully-connected directed graph where $V = (x;\ y)|x = 1\ ;x, y = 1, \ldots, n$ is the set of vertices. The weight of a directed edge $w((x;\ y);\ (p;\ q))F \in E$ is computed as

$$\mathrm{w((x;\ y);\ (p;\ q))F} = \left|\log\left(\frac{\mathrm{Fx;\ y}}{\mathrm{Fp;\ q}}\right)\right| \mathrm{KF\ ((x;\ y);\ (p;\ q))}$$

Where the kernel function K_F returns values inversely proportional to the distance of the input points. The ratio between the two feature values represents the standard definition of dissimilarity. The absolute of the log allows to reach the lowest dissimilarity when the ratio is 1, while it returns higher values when the ratio is either lower or higher than 1. Once the graph is constructed, a Markov chain approach is exploited to detect the most dissimilar points of the image. The equilibrium distribution computed on such a Markov chain effectively reveals the set of points that are most dissimilar from the others. Such a distribution defines the activation map A. A fully connected graph G_A is exploited to concentrate the mass of each as into nodes with high activation values. The weight of the direct edge between two nodes (x, y) and (p, q) is computed as

$$\mathrm{w\big((x,y),(p,q)\big)A} = \mathrm{Ap,qKA\big((x,y),(p,q)\big)}$$

Where K_A is a kernel function substituting the similarity measure. Let A_j be the activation map computed for the j -th feature map F_j , for j = 1, 2, . . . , J, then the final saliency master map is defined as

$$\Omega = \mathrm{p}(\mu, \Sigma) + \sum_{\mathrm{j=i}}^{\mathrm{J}} \alpha_{\mathrm{j}}\, \mathrm{A^j}$$

Where α is a vector of weights and

$$\mathrm{p}(\mu, \sigma) \sim \exp\left(-\left(\frac{\mathrm{x} - \mu_{\mathrm{x}}}{\sigma_{\mathrm{x}}} + \frac{\mathrm{y} - \mu_{\mathrm{z}}}{\sigma_{\mathrm{y}}}\right)\right)$$

is a non-isotropic Gaussian kernel prior centered at $\mu = [\mu_x,\ \mu_y]^T$ with$\sigma = [\sigma_{x,}\ \sigma_y]^T$, which accounts for silhouette location and shape.

2) Colour Space

The proposed system is mainly depend on colour. Four colour spaces are used here, RGB, HSV, YUV, LAB. For all these colour spaces find the mean value of each dimensions for k patches and store as a vector. These mean values are used for matching.

B. Video Processing

The selected video is divided into no of frames and detect persons from each frame. For that vision object called people detector is used. The people detector object detects people in an input image using the Histogram of Oriented Gradient (HOG) features and a trained Support Vector Machine (SVM) classifier. The object detects un-occluded people in an upright position.
For each person detected from video do the same procedure explained in A.

C. Matching

For matching all the values calculated are used. Find the sum of all the mean values. In the following equation mean1 can rgb, hsv, yuv, lab or saliency mean of input image and mean2 can be rgb, hsv, yuv, lab or saliency mean of all person's image detected from video.

$$\mathrm{A(i)} = \mathrm{sum(abs(mean1 - mean2))}$$

Then find the sum these four mean sums and choose the highest value as the similar person. In below equation A1 stands for sum of difference between mean values of lab colour space of input image and persons detected from video. Likewise A2 A3 and A4 are yuv, hsv and saliency values.

$$\mathrm{D = A1 + A2 + A3 + A4}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The method was implemented in a MATLAB 2014 prototype and tested with randomly selected images from 3DPeS dataset. The image processing was performed on a desktop PC with the following characteristics: Intel Core i3 CPU, 2.30 GHz, 2 GB RAM.

The person re-identification is demonstrated by an example selected from 3DPeS database. Graph based visual saliency of the input image is found in the first stage. And the saliency map is divided into equal patches and find the mean value of each patch. Then the colour features are extracted. Initially the input image is converted into Lab colour space. Then the lab image is divided into 8 patches and find the mean value of each patch. Similarly, yuv and hsv color spaces are treated. Input image first converted into Yuv/hsv colour space. Then the output image is divided into patches and find the mean of each patch.

Fig. 4.1: Input image and its saliency map

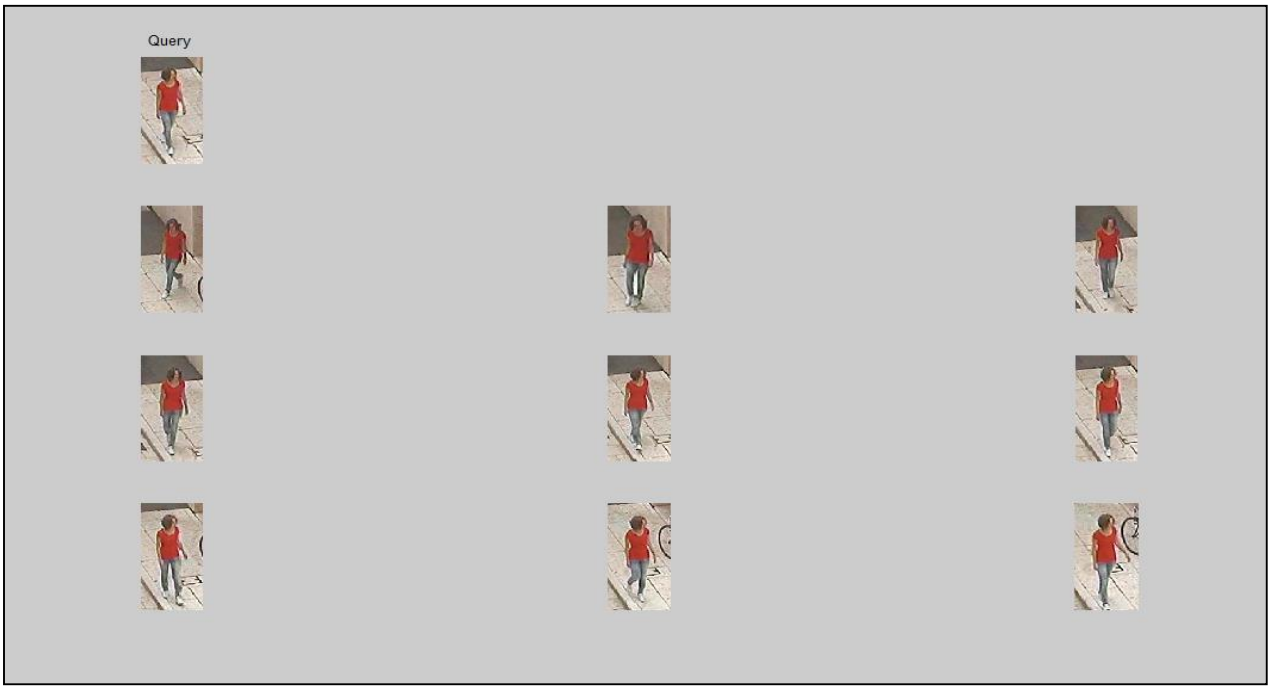

Fig. 4.2: Matched persons from video

V. CONCLUSION

This existing approach is proposed to address the re-identification problem in images by introducing a novel algorithm able to identify the salient regions of a person. The saliency detection is used to find the salient region. The computed saliency is used as a weight in the feature extraction process which also combines other features that do not consider it. The manifold where the extracted features lie is learned through PCA, and the resulting coefficients are input to the proposed pairwise-based multiple metric learning framework. The obtained metrics are exploited to learn the coefficients of a linear combination used to compute the dissimilarity between image pairs. This system does not applicable in video. So a saliency based person re- identification in video using color feature method is proposed to extend the system into video.

Here the input query is in image and it will compare with video. A color based comparison is used. It is a simple method to find the person re-identification. It take more time when the video is used as input query. So this can be done as future work.

REFERENCES

[1] T. Ojala, M. Pietikainen, and T. Maenpaa, \Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971987, Jul. 2002.
[2] J. Harel, C. Koch, and P. Perona, \Graph-based visual saliency", in Proc.NIPS, 2007, pp. 552554.
[3] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, \Dimensionality reduction: A comparative review," J. Mach. Learn. Res., vol. 10, pp. 141, Feb. 2009.
[4] M. Hirzer, P. M. Roth, M. Kstinger, and H. Bischof, \Relaxed pairwise learned metric for person re-identification", in Proc. ECCV, 2012, pp. 780793.
[5] W. Li, R. Zhao, and X. Wang, \Human reidentification with transferred metric learning", in Proc. ACCV, 2012, pp. 3144.
[6] M. Kstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, \Large scale metric learning from equivalence constraints," in Proc. IEEE Conf. CVPR, Jun. 2012, pp. 22882295.
[7] K. Zhang, L. Zhang, and M.-H. Yang, \Real-time compressive tracking," in Proc. ECCV, 2012, pp. 864877.
[8] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, \Local Fisher discriminant analysis for pedestrian re-identification", in Proc. IEEE Conf. CVPR, Jun. 2013, pp. 33183325.
[9] R. Zhao, W. Ouyang, and X. Wang,\ Person re-identification by salience matching", in Proc. IEEE ICCV, Dec. 2013, pp. 25282535
[10] R. Zhao, W. Ouyang, and X. Wang, \Unsupervised salience learning for person re-identification", in Proc. IEEE CVPR, Jun. 2013, pp. 35863593
[11] Niki Martinel, Christian Micheloni, and Gian Luca Foresti , \Kernelized Saliency-Based Person Re-Identification Through Multiple Metric Learning", IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 24, NO. 12, DECEMBER 2015.