OBJECT REIDENTIFICATION IN REAL WORLD SCENARIOS ACROSS MULTIPLE NON-OVERLAPPING CAMERAS

Guy Berdugo, Omri Soceanu, Yair Moshe, Dmitry Rudoy, and Itsik Dvir

Signal and Image Processing Laboratory (SIPL), Dept. of Electrical Engineering, Technion - Israel Institute of Technology,
32000, Haifa, Israel
phone: + (972)4-8294746, email: {guykun,omri123,dmitry.rudoy}@gmail.com, yair@ee.technion.ac.il, itsikd@mate.co.il
web: sipl.technion.ac.il

ABSTRACT

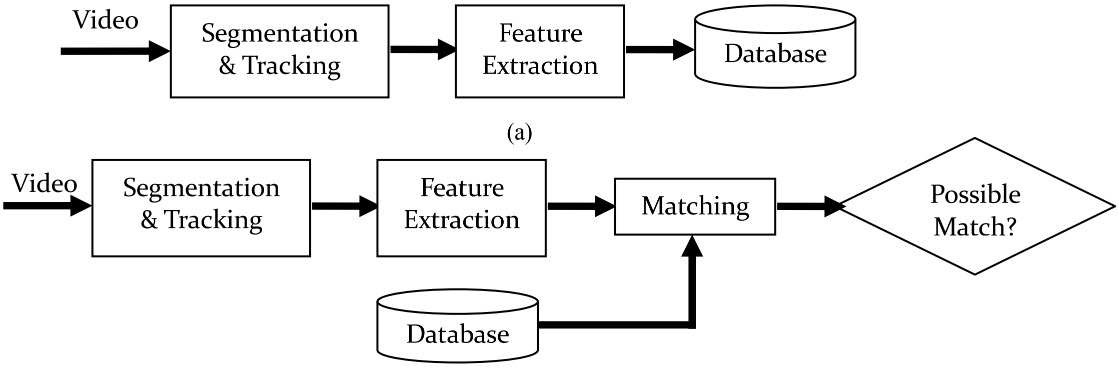
In a world where surveillance cameras are at every street corner, there is a growing need for synergy among cameras as well as the automation of the data analysis process. This paper deals with the problem of reidentification of objects in a set of multiple cameras inputs without any prior knowledge of the cameras distribution or coverage. The proposed approach is robust to change of scale, lighting conditions, noise and viewpoints among cameras, as well as object rotation and unpredictable trajectories. Both novel and traditional features are extracted from the object. Light and noise invariance is achieved using textural features such as oriented gradients, color ratio and color saliency. A probabilistic framework is used incorporating the different features into a human probabilistic model. Experimental results show that textural features improve the reidentification rate and the robustness of the recognition process compared with other state-of-the-art algorithms.

1. INTRODUCTION

Object detection and recognition is at the core of every automated surveillance system. Before dealing with higher-level tasks such as activity observation, it is crucial to find correspondence among appearances of the same object or human being on different cameras at different times. Real-world scenarios present numerous obstacles for the matching process. These obstacles include for example, lighting variations among cameras or time, different viewpoints and object poses. In addition, unpredictable trajectories combined with a lack of camera coverage exclude the possibility of location prediction. Only partial solutions have been suggested so far, dealing with some success with these real-world scenarios.

There are two main approaches towards constructing a reidentification system: *motion prediction* based and *appearance matching* based [1]. Motion prediction based systems tend to fail when certain conditions are not met. For example, when objects reappear after a long time, the motion prediction based systems have no way of identifying them. Lack of camera coverage represents a major problem for motion prediction, resulting in particular from the inability to predict the movement of human beings between cameras. Furthermore, these systems require topological information about the camera array and the scene. For these reasons, in this paper, an appearance matching based approach is proposed. Appearance matching uses visual cues (features) derived from a given object to describe its appearance. Appearance matching is based on a process of feature extraction, which transforms the color channels of an object into another, mostly concise, representation. One of the simplest appearance features is *mean color* [2], which extracts the average color in each color channel (RGB) and describes the object using its mean color. This feature is very compact but is not discriminative enough. The *Intensity* feature [3][4][5] extracts the mean intensity of the object's pixel colors, however it is sensitive to changing light conditions. More sophisticated features include *Covariance* [2], which takes a vector comprised of (r, g, b) color values, alongside their respected (x, y) image positions and oriented gradient values and calculates the covariance of these features over a single connected component. *Dominant Color* [2][6], based on the MPEG-7 Dominant Color property, extracts the most common color of the object. *Major Color Spectrum Histogram Representation* [7] expands the dominant color feature and extracts the N-most common colors in the histogram. *Human Color Structure Descriptor* [8] which is a feature designated to human beings, attempts to incorporate structural information to the appearance feature by using a vector of three number sets. Each number set represents a different body part, and includes the mean color and the center of gravity position. The major problem with all the aforementioned features is that while they produce a compact representation of the object and allow a fast matching process, they are not discriminative enough. In addition, some of them lack robustness to illumination changes and noise, and thus fail to produce sufficient results in the most general conditions.

Lin and Davis presented a state-of-the-art reidentification system [9] that produces good results. This paper aims to improve that system in terms of robustness and performance. The system presented here is based on the framework suggested in [9], focusing on and improving the appearance modeling phase. The system uses a nonparametric multivariable kernel density method [10] to build a probabilistic appearance model for each object. This model is later used in a nearest neighbor matching process based on the Kullback-Leibler distance in order to compare two probabilistic models.



(a)

(b)

Figure 1 - Object re-identification system, (a) Appearance model database creation, (b) Re-identification process using the previously formed database.

Observations have shown that in real world scenarios most people wear similarly colored clothes e.g., blue jeans and dark shirts. For such datasets, systems that rely solely on color features tend to fail even for single camera inputs. This paper proposes a novel approach towards solving both the separability and the inter-camera variability problems. The separability of recognition was improved by incorporating textural features that provide more discriminative information. Textural features are robust since they might be independent of inter-camera variability and do not change as the color values fluctuate under illumination changes. Thus by proposing and testing a variety of textural features, a robust two-stage system for multiple cameras reidentification, as portrayed in Figure 1, is proposed. The first stage (Figure 1a) creates and stores appearance models in a database. The second stage (Figure 1b) creates an appearance model for a new given object. The model is matched against other appearance models that were stored in the database previously created.

The rest of this paper is organized as follows. Section 2 covers object modeling. Algorithm results are given in section 3 and conclusions are drawn in Section 4.

2. OBJECT MODELING

In order to efficiently store data that will be used later to build an appearance model, the raw data from the camera is going through stages of segmentation, tracking and feature extraction. In this paper we assume that a tracking process and/or object detection was already performed and an input of a bounding box containing a human figure or an object is given.

2.1 Segmentation

Testing suggests that background removal inside the bounding box improves results dramatically. Different methods exist for background removal in a bounding box, e.g.: saliency [11] or background subtraction methods [12]. In this paper background subtraction was used to roughly segment objects. A framework for dealing with still images using saliency maps [11] was used to refine the segmentation. Example segmentation results using this technique are shown in Figure 2.

2.2 Feature Extraction

An important step in object recognition is feature extraction. During this step characteristics such as color and texture are extracted from the segmented object such as the one shown in Figure 3a. Certain considerations must be taken into account when choosing features to extract from the segmented object. First, attention must be given to the discriminative nature and the separability of the feature, to achieve consistency during the matching process. Second, robustness to illumination changes is crucial when dealing with multiple cameras and dynamic environments. Finally, noise robustness and scale invariance should be taken into account. Scale invariance is obtained by resizing each figure to a constant size. Robustness to illumination changes is achieved using a method of ranking over the features, mapping absolute values to relative values. Ranking cancels any linear modeled lighting transformations, under the assumption that for such transformations the shape of the feature distribution function is relatively constant. To obtain the rank of a vector x, the normalized cumulative histogram $H(x)$ of the vector is calculated. The rank $O(x)$ is given by [9]:

$$O(x) = \lceil H(x) \cdot 100 \rceil \qquad (1)$$

Where $\lceil \cdot \rceil$ denotes rounding the number up to the consecutive integer. Using 100 as a factor sets the possible values of the

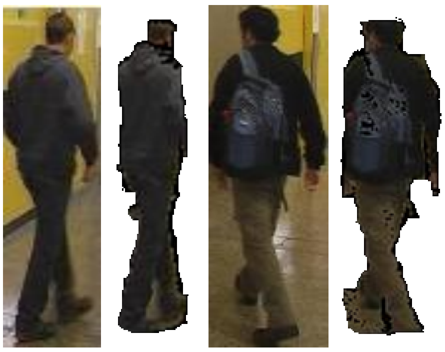


Figure 2 – Human figures before and after background removal as described in [11].

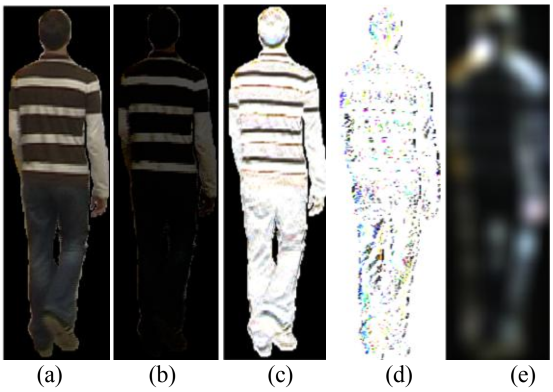


Figure 3 – Features extracted from a human figure before and after ranking, (a) The figure after segmentation, (b) Color Rank, (c) Color Ratio, (d) Oriented Gradients, (e) Saliency.

ranked feature to $[x]$ and sets the values of $O(x)$ to the percentage values of the cumulative histogram.

The proposed ranking method is applied on the chosen features to achieve robustness to linear illumination changes. The proposed features are of two types, *color* features and *textural* features. For color we use the *Color Rank* feature [13], as can be seen in Figure 3b. Color rank values are obtained by applying the ranking process on the RGB color channels using Eq. (1). Another color feature, proposed in [13] is the *Normalized Color*. The feature values are obtained using the following color transformation:

$$(r,g,s)=\left(\frac{R}{(R+G+B)},\frac{G}{(R+G+B)},\frac{(R+G+B)}{3}\right) \qquad (2)$$

where R, G and B denote the red, green and blue color channels of the segmented object respectively. r and g denote the chromaticity of the red and green channel respectively and s denotes the brightness. Transforming to the rgs color space separates the chromaticity from the brightness resulting in illumination invariance [13].

Using only color features might be insufficient when dealing with similarly colored objects or with figures with similar clothing colors, e.g., a red and white striped shirt compared with a red and white shirt with a crisscross pattern. In order to solve this problem, we suggest several textural features. Textural features obtain values in relation to their spatial surroundings. Information is extracted from a region rather than a single pixel. Thus a more global point of view is obtained.

We propose several novel textural features. First is the *Ranked Color Ratio* feature, in which each pixel is divided by its upper neighbor. This feature is derived from a multiplicative model of light and a principle of locality. This operation intensifies edges and separates them from the plain regions of the object, as can be seen in Figure 3c. For a more compact representation, as well as rotational invariance around the vertical axis, an average can be calculated over each row. This results in a column vector corresponding to the spatial location of each value. Finally, the resulting vector or matrix is ranked by applying Eq. (1).

Another textural feature is the *Ranked Oriented Gradients*, based on *Histogram of Oriented Gradients* [14]. First gradients are calculated on both the horizontal and the vertical directions. The gradient's orientation of each pixel $\theta_{(i,j)}$, is calculated using:

$$\theta_{(i,j)}=\arctan\left(\frac{dy_{(i,j)}}{dx_{(i,j)}}\right) \qquad (3)$$

where $dy_{(i,j)}$ is the vertical gradient and $dx_{(i,j)}$ is the horizontal gradient in pixel (i,j). Instead of using a histogram, the matrix form is kept in order to maintain spatial information regarding the location of each value, as can be seen in Figure 3d. Then, ranking is performed using Eq. (1) for quantization.

Finally, for a more global point of view, a novel feature is proposed based on saliency maps, *Ranked Saliency Maps*. In neuroscience, an object that attracts the attention of the eye for any of many reasons is considered as salient. A saliency map sM is obtained, as suggested by Soceanu et al. [11], for each of the RGB color channels by:

$$\phi(u,v)=\angle F\left(I(x,y)\right) \qquad (4)$$

$$A(u,v)=\left|F\left(I(x,y)\right)\right| \qquad (5)$$

$$sM(x,y)=g(x,y)*\left|F^{-1}\left[A^{-1}(u,v)\cdot e^{j\cdot\phi(u,v)}\right]\right|^2 \qquad (6)$$

where $F(\cdot)$ and $F^{-1}(\cdot)$ denote the Fourier Transform and Inverse Fourier Transform, respectively. $A(u,v)$ represents the magnitude of the color channel $I(x,y)$, and $\phi(u,v)$ represents the phase spectrum of $I(x,y)$. $g(x,y)$ is a 8x8 Gaussian filter. The result can be seen in Figure 3e. Each of the saliency maps are then ranked according to Eq. (1).

In order to represent all the aforementioned features in a structural context, spatial information is stored by using a *height* feature. The height feature is calculated using the normalized y-coordinate of the pixel. The normalization ensures scale invariance.

2.3 Building a Probabilistic Model

The features' values of each pixel are represented in an n-dimensional vector where n denotes the number of features extracted from the image. Feature values for a given person or object are not deterministic and vary among frames. Hence a stochastic model which incorporates the different features is used. *Multivariate kernel density estimation* (MKDE) [10] is used to construct the probabilistic model as suggested in [9].

Given a set of feature vectors $\{s_i\}$,

$$s_i=(s_{i1},...,s_{in})^T, \qquad i=1...N_p \qquad (7)$$

$$\hat{p}(\mathbf{z})=\frac{1}{N_p\sigma_1\cdot...\sigma_n}\sum_{i=1}^{N_p}\prod_{j=1}^{n}\kappa\left(\frac{z_j-s_{ij}}{\sigma_j}\right) \qquad (8)$$

$\hat{p}(\boldsymbol{z})$ is the probability of obtaining a given feature vector z with the same components as s_i. $\kappa(\cdot)$ denotes the Gaussian kernel, which is the kernel function used for all channels. N_p is the number of pixels sampled from a given object and σ_j

are parameters denoting the standard deviation of the kernels which are set according to empirical results.

2.4 Matching

In order to evaluate the correlation between two appearance models, a distance measure is defined. The measure should be robust and produce separable results. One such distance measure is the *Kullback-Leibler* distance [15] denoted as D_{KL}. The Kullback-Leibler distance presents a robust information gain tool, quantifying the difference between two probabilistic density functions:

$$D_{KL}\left(\hat{p}^{A} \mid \hat{p}^{B}\right)=\int \hat{p}^{B}(\mathbf{z}) \cdot \log \frac{\hat{p}^{B}(\mathbf{z})}{\hat{p}^{A}(\mathbf{z})} d\mathbf{z} \qquad (9)$$

where $\hat{p}^{A}(\boldsymbol{z})$ and $\hat{p}^{B}(\boldsymbol{z})$ denote the probability to obtain the feature value vector $\boldsymbol{z}$ for appearance model B and A respectively.

The transformation into a discrete analysis is the same as in [9]. Appearance models from a dataset are compared with a new model using the Kullback-Leibler distance measure. Low D_{KL} values represent small information gains corresponding to a match of appearance models based on a nearest neighbor approach.

The robustness of the appearance model is improved by matching key frames from the trajectory path of the object, rather than matching a single image. Key frames are selected using the Kullback-Leibler distance along the trajectory path. The distance between two trajectories $L^{(I,J)}$ is obtained using:

$$L^{(I,J)} = \underset{i \in K^{(I)}}{median}\left[\min_{j \in K^{(J)}} D_{KL}\left(p_i^{(I)} \mid p_j^{(J)}\right)\right] \qquad (10)$$

where $K^{(I)}$ and $K^{(J)}$ denote the set of key frames from the trajectories I and J respectively. $p_i{}^{(I)}$ denotes the probability density function based on a key frame i from trajectory I. First, for each key-frame i in trajectory I the distance from trajectory J is found. Then, in order to remove outliers produced by segmentation errors or object entrance/exit in the scene, the median of all distances is calculated.

3. RESULTS

In order to evaluate the performance of the proposed system and to compare the results using different feature extraction methods, a video dataset was created. Two video cameras shot two hallways under different lighting conditions, different viewpoints and at different times. The dataset was manually annotated for easier ground-truth testing, and contains 6000 frames of more than 30 different human figures that appear and reappear in both cameras between 1-8 times with an average blob size of 70x150 pixels. The dataset presents real world scenarios, such as people with similar color schemes, as shown in Figure 4. The results of the tests were compared against results produced using the state-of-the-art system suggested in [13]. Testing was performed using the leave-one-out procedure over the database. Various features were tested for their ability to reidentify figures that reappear in the same camera at various time and in various cameras at a different time. The latter is of course harder to accomplish due to the differences in lighting and viewpoint. Combinations of color and textural feature were tested in order to determine which textural feature provides the best discriminative information. Since the saliency feature provides a very high level perspective, tests show that the saliency feature improves results only when combined with another textural feature. Only then can it identify both lower level and higher level details as seen in Table 1. The problem with such a feature is that it has a high dimension and as such drastically slows down the matching process. Comparing the combination of the color ratio feature and a color feature to the combination of the oriented gradients feature and a color feature, one notices that the oriented gradients feature combination surpasses the color ratio combination in its results. This may be a result of the accuracy of the floating point representation, i.e. subtraction of pixel values provides better distinction then pixel values division. The results of the tests using the conventional features as proposed in the state-of-the-art system [13], and the best novel feature combination can be seen in Table 1.

As predicted, single camera reidentification produces better results than multiple camera reidentification for most cases. For some features multiple camera reidentification had produced slightly better results, due to some variability between the two datasets. Nevertheless, it is clear from the results that the use of textural features in combination with color features improves results over the state-of-the-art color based system described in [13]. Table 1 shows that the best results were obtained by using a combination of the *Normalized Color* feature and the *Oriented Gradients* feature. Using this combination, 90% of the human figures were identified correctly when they reappeared at the same camera, and 66.7% of the figures were identified correctly when they reappeared at a different camera. These results, as reflected in the dataset we use, suggest 10% and 9% increase over the state-of-the-art system described in [13] in a single and dual viewpoints matching scenarios respectively.

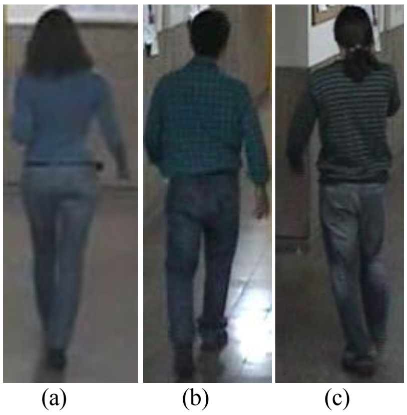


Figure 4 – Examples from the GBSEO dataset of human figures with similar color schemes. (a), (b), (c) all have blue shirts and blue jeans and can only be differentiated by the pattern of their shirts - (a) plain blue, (b) checkers and (c) stripes.

TABLE I
HUMAN REIDENTIFICATION RESULTS

One Camera	Success Rate
Lin et al [9] - Color Rank, Height	50%
Color Rank, Oriented Gradients, Saliency, Height	53.8%
Lin et al [9] - Normalized Color, Height	80%
Normalized Color, Oriented Gradients, Height	**90%**
Two Cameras	Success Rate
Lin et al [9] - Color Rank, Height	58%
Lin et al [9] - Normalized Color, Height	58%
Normalized Color, Oriented Gradients, Height	**66.7%**

4. CONCLUSION

A human reidentification system is proposed based on the system described in [13]. Novel appearance features were incorporated into the system in order to accommodate changing conditions and to overcome problems created by similar color schemes. The use of textural features, alongside with the conventional color features, proves to be a decisive factor in the correct matching of similarly colored figures. Often the difference between a correct and a false match of a recurring figure would be minute. By adding a textural feature one improves the separability of the results for similarly colored figures thus improving the reidentification success rate. Furthermore the incorporation of textural features is shown to produce better viewpoint and light invariance due to its independence from absolute color values. A manually annotated video dataset that has been created for this research activity is freely available for non commercial comparing similar recognition systems. On this dataset, the proposed system achieves 9% increase over the state-of-the-art previously described system described in [13].

5. ACKNOWLEGMENTS

We would like to thank Gal Braun for his advice on running time optimization and our friends for their participation as actors in creation of the video database. We would also like to thank the staff of the Signal & Image Processing Lab (SIPL) for their support, and especially its head, Prof. David Malah. The work was supported in part by Mate Intelligent Video http://www.mate.co.il . A special thanks is sent to Dan Raudnitz from Mate for encouraging this activity.

6. APPENDIX

The annotated video dataset (GBSEO – Ground-truth Bounding - boxes for Surveillance Evaluation and Optimization) can be downloaded from the Signal & Image Processing Lab (SIPL) website at the following link:
http://sipl.technion.ac.il/GBSEO.shtml

REFERENCES

[1] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," Computer Vision and Image Understanding, vol. 104, no. 2-3, pp. 90-126, November 2006.

[2] A. Colombo, J. Orwell, and S. Velastin, "Colour constancy techniques for re-recognition of pedestrians from multiple surveillance cameras," in *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications* (M2SFA2 2008), October 2008, Marseille, France.

[3] K. Jeong, C. Jaynes, "Object matching in disjoint cameras using a color transfer approach," *Special Issue of Machine Vision and Applications Journal*, vol. 19, pp 5-6, Oct. 2008.

[4] F.M. Porikli, A. Divakaran, "Multi-camera calibration, object tracking and query generation," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Baltimore, MD, July 6-9, 2003, vol. 1, pp. 653-656.

[5] O. Javed, K. Shafique, M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 20-25. 2005, vol. 2, pp 26-33.

[6] V. Modi, "Color descriptors from compressed images", in *CVonline: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision*. Retrieved December 30, 2008

[7] C. Madden, E.D. Cheng, M. Piccardi, "Tracking people across disjoint camera views by an illumination-tolerant appearance representation" in *Machine Vision and Applications*, vol. 18, pp 233–247, 2007.

[8] S.Y. Chien, W.K. Chan, D.C. Cherng, J.Y. Chang, "Human object tracking algorithm with human color structure descriptor for video surveillance systems," in *Proc. of 2006 IEEE International Conference on Multimedia and Expo*, Toronto, Canada, July 2006, pp. 2097-2100.

[9] Z. Lin, L. S. Davis, "Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance," in *Proc. of the 4th International Symposium on Advances in Visual Computing*, Lecture Notes in Computer Science, Vol. 5358, pp. 23-24, 2008.

[10] C. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[11] O. Soceanu, G. Berdugo, D. Rudoy, Y. Moshe, I. Dvir, "Where's Waldo? Human figure segmentation using saliency maps," in *Proc. ISCCSP 2010*, Limassol, Cyprus, March 3-5. 2010.

[12] T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis" *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90-126, Nov. 2006.

[13] Y. Yu, D. Harwood, K. Yoon, and L. S. Davis, "Human appearance modelling for matching across video sequences," in *Machine Vision and Applications*, vol. 18, no. 3-4, pp. 139-149, Aug. 2007.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. International Conference on Computer Vision*, Beijing, China, October 17-21. 2005, pp. 886-893.

[15] S. Kullback, *Information Theory and Statistics*. John Wiley & Sons, 1959.