Mitigating the Effects of Variable Illumination for Tracking across Disjoint Camera Views

E. D. Cheng, C. Madden, M. Piccardi, *Senior Member, IEEE*
Faculty of Information Technology
University of Technology, Sydney
{cheng, cmadden, massimo}@it.uts.edu.au

Abstract

Tracking people by their appearance across disjoint camera views is challenging since appearance may vary significantly across such views. This problem has been tackled in the past by computing intensity transfer functions between each camera pair during an initial training stage. However, in real-life situations, intensity transfer functions depend not only on the camera pair, but also on the actual illumination at pixel-wise resolution and may prove impractical to estimate to a satisfactory extent. For this reason, in this paper we propose an appearance representation for people tracking capable of coping with the typical illumination changes occurring in a surveillance scenario. Our appearance representation is based on an online K-means color clustering algorithm, a fixed, data-dependent intensity transformation, and the incremental use of frames. Moreover, a similarity measurement is proposed to match the appearance representations of any two given moving objects along sequences of frames. Experimental results presented in this paper show that the proposed methods provides a viable while effective approach for tracking people across disjoint camera views in typical surveillance scenarios.

1. Introduction

Tracking single individuals as they move under a camera network is a challenging task in computer vision since their appearance varies significantly across views. This is particularly true when cameras in the network do not provide a continuity of overlapping or quasi-overlapping views, but are instead disjoint in time and/or space to a certain extent. This last case is common in existing camera networks operated by security personnel, since humans do not need to continuously view a person to track it effectively. A similar capability would be desirable for automated computer-vision systems for efficient re-use of existing surveillance infrastructure.

In our previous paper [2], we presented our approach for the case where two disjoint tracks were acquired from the same camera at approximately the same time and under similar illumination conditions. In this work, we address the more challenging case of tracks actually disjoint in time and/or space, where illumination conditions are subject to major changes. This scenario has required significant additions which are presented in the following together with a brief description of the rest of our approach. The assumption in this work is that moving people are correctly detected and tracked within single camera views, and the goal is to find correspondences between such tracks across views.

The main references for our work are the recent paper from Javed *et al.* [1] and their previous work [3]. In their approach, Javed *et al*. propose to compensate for the different illumination conditions by estimating intensity transfer functions between each camera pair during an initial training phase. Such functions are estimated by displaying common targets to the two cameras under a significant range of illumination conditions, and modelling correspondences in the targets' color histograms. However, the authors' assumptions in [1,3] that objects are planar, radiance is diffuse and illumination the same throughout the whole field of view do not hold in real life. Illumination varies at pixel-level resolution and such variations have first-order effects on appearance. Weiss in [4] proposed an effective method to estimate illumination from a sequence of frames of the same scene. Though the method works well for static objects such as the background scene, it cannot accurately predict the illumination over 3D moving targets such as people. Accurately estimating illumination over 3D moving targets would require detailed knowledge of the position and parameters of sources of lights, reflections and shadowing. Moreover, the time-varying effects of natural light sources are hard to estimate. Therefore in this paper we propose an approach based on a fixed, data-dependent intensity transformation and color representation to cope with the typical illumination variations occurring in a surveillance scenario.

2. The Major Color Spectrum Histogram

We first introduce the concept of color distance between two color pixels in the RGB space based on a normalized geometric distance between the two pixels. Such a geometric distance is defined in (1).

$$d(C_1, C_2) = \frac{\|C_1 - C_2\|}{\|C_1\| + \|C_2\|} = \frac{\sqrt{(r_1 - r_2)^2 + (g_1 - g_2)^2 + (b_1 - b_2)^2}}{\sqrt{r_1^2 + g_1^2 + b_1^2} + \sqrt{r_2^2 + g_2^2 + b_2^2}} \quad (1)$$

C_1 and C_2 are the color vectors. (1) defines a normalized distance, i.e. the Euclidean distance between the two colors is divided by the sum of their magnitudes. This normalised distance is used to minimise the effect of the camera's ability to perceive more color variation under stronger illumination.

2.1. Major color representation

In the RGB color space, using the concept of color distance given in (1), it is possible to cluster colors into a limited number of "major colors" without losing much accuracy in representing an object. Several methods have been proposed in the literature for color clustering [5-8]. Since we aim at real-time applications, clustering speed is a major requirement. In our approach, we choose to use clusters of a single size within the normalized distance (1) space. This tends to keep clustering (and associated color comparisons) fast as optimizations are based upon cluster placement, and not optimizing the number, or size of the clusters. The number of clusters is chosen dynamically so that a minimum of 90 percent of the pixels are represented by them. We call our representation the Major Color Spectrum Histogram representation (MCSHR). An online clustering algorithm is proposed to calculate MCSHR as described in the next section [9]. The first iteration of the algorithm is described below.

A color distance threshold is chosen as the radius of sphere-shaped clusters in the normalized RGB space. The object's pixels are then scanned in row-major order. As the first pixel appears, it is set as the centre of the first cluster. For each following pixel, if such a pixel is within the distance threshold from an existing cluster's centre, the pixel count for that cluster is increased by one; otherwise, a new cluster is created, centred on that pixel. This procedure is similar to that proposed by Li *et al*. to calculate principal colors [7].

2.2. Major color clustering algorithm

The above step provides an initial set of the most significant, or major color clusters. Given the simple cluster creation procedure, a cluster's centre may be significantly displaced with respect to the cluster's centroid. In our experiments, we found that this may affect, in turn, the following comparisons between object representations. For this reason, we decided to refine the set of clusters by an online K-means algorithm. A K-means algorithm iteratively alternates two steps: *a*) calculating the membership of pixels to clusters and *b*) re-computing the clusters' centres based on their member pixels. In traditional versions of the K-means algorithm, each step is applied in turn over the whole population of pixels. In its online versions, steps *a* and *b* are computed in sequence for each pixel. K-means algorithms converge to local optima which often prove adequate solutions. In particular, this is true in our case since the algorithm is allowed to start from a reasonable initial solution. Our online K-means major color clustering algorithm works as follows: for each pixel, the closest cluster centre is computed (step *a*). Then, the position of such a cluster is updated as:

$$\begin{cases} \mathbf{R}_{\mathbf{ClusterCenter}}(\mathbf{i}) = \mathbf{w(i)R(i) + (1 - w(i))R}_{\mathbf{ClusterCenter}}(\mathbf{i - 1}) \\ \mathbf{G}_{\mathbf{ClusterCenter}}(\mathbf{i}) = \mathbf{w(i)G(i) + (1 - w(i))G}_{\mathbf{ClusterCenter}}(\mathbf{i - 1}) \\ \mathbf{B}_{\mathbf{ClusterCenter}}(\mathbf{i}) = \mathbf{w(i)B(i) + (1 - w(i))B}_{\mathbf{ClusterCenter}}(\mathbf{i - 1}) \end{cases} \quad (2)$$

where i is the current number of pixels in the cluster and $w(i) = 1/i$ the current weighting coefficient (step *b*). We can see that with the increase in the number of pixels falling into a same cluster, the weighting coefficient decreases, meaning that changes in position tend to gradually slow down. Iterations are necessary until all centres are stabilized. In our experiments, between 80% and 90% of pixels are already member of their final cluster after just one iteration. An example of an Ore Gold Rose picture (rich in color tones) is shown in Fig. 1. The corresponding major color spectrum histogram representation calculated by using the online K-means clustering algorithm is shown in Fig. 2.

Figure 1: An Ore Gold Rose

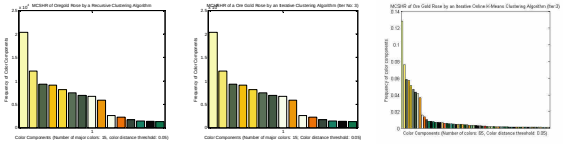

Figure 2: Major Color Spectrum Histograms with the proposed color clustering algorithm (a) Iteration 1, top 15 Major Colors (b) Iteration 3, top 15 Major Colors (c) Iteration 3, all 65 Major Colors

Figures 1 and 2 show that the MCSHR calculated by using the online K-means clustering algorithm on the example Ore Gold Rose image is visually accurate and no obvious improvement was made by increasing the number of iterations from one to three. With a 90% cut-off threshold, only a limited number of major colors is retained.

3. Compensating for illumination variations across disjoint views

The greatest challenge for the matching of moving objects from disjoint camera views is in the different and varying illumination causing great differences in their appearances. This work assumes that footage is taken in a typical indoor environment where lighting is mainly from artificial white light sources or sunlight through glass windows or doors. The light can be considered fairly constant in some areas; however, illumination changes such as clouds filtering the sunlight can be expected to impact the scene illumination over time. As explained above, the computation of an exact transformation justifying the changes in appearance can prove impractical. For this reason, we propose to use a fixed, data-dependent intensity transformation based on cumulative color histograms local to each moving object.

First, moving objects are extracted based on a background subtraction procedure. Let us call *A* the set of pixels in a generic object and *N* its size. Let us also call *B* a second set of *N virtual* pixels, with a histogram that is equalized in their R,G,B, components. A cumulative histogram transformation (or equalization) is then computed on their union, A ∪ B, for the R,G,B components separately. The three resulting transforms (T_r, T_g and T_b) are applied to remap the R,G,B components of the histogram of A, the moving object's pixels. The rationale for our approach is that of applying a form of "controlled equalization" to the object to compensate for local illumination without requiring neither impractical training nor knowledge/estimation of any parameters. An example is shown in Fig. 3. The original histogram for the R component is shown in Fig 3.f. After the proposed color transformation, the histogram (see Fig. 3.h) qualitatively retains its original shape, but extends over the available spectrum thus compensating for local illumination. The effect of the proposed transformation is not to be confused with that of full equalization, which would result in a flat histogram. Some analogy may instead be seen with histogram stretching algorithms. However, such algorithms are very sensitive to the stretching parameters (original starting and ending bins and/or position and number of modes) while the proposed transformation requires no parameters.

After computing MCSHR for a sequence of frames, we consider each frame and integrate MCHSR over the last *K* frames (*K* was set to 3 in the experiments reported in this paper). In this way, we further compensate for small differences in appearance caused by slight pose variations and limited segmentation errors. We call this augmented representation the *incremental* MCHSR (IMCSHR). The combination of this representation and the proposed color transformation proved tolerant to illumination changes of typical surveillance scenarios.

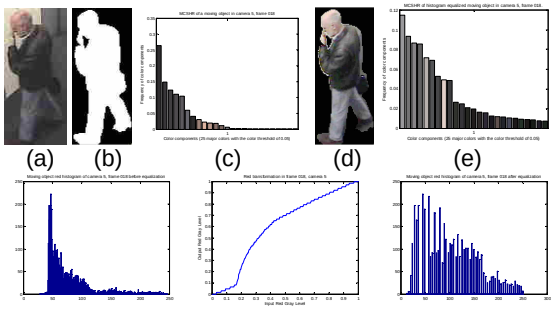

Figure 3: Effects of the proposed color transformation: (a) original object; (b) its mask and (c) major colors; d) the object after color transformation and (e) its major colors; the 'red' histograms (f) before and (h) after the transformation, and the corresponding transform (g). (h) makes better use of the available spectrum than (f), and "normalises" the histogram with respect to local illumination.

4. Similarity measurement and matching

Once given the appearance representation, a similarity measurement is needed to quantify the overall similarity between any two moving objects. We assume that there exist *M* major colors in the spectrum of moving object A, which can be represented as:

$$MCSHR(A) = \{ C_{A_1}, C_{A_2}, \cdots, C_{A_i}, \cdots, C_{A_M} \} \quad (3)$$

where $C_{A_i}, i = 1,2,\cdots,M$ is a major color (RGB) in object A. Object A's major color bin counts can be represented as:

$$p(A) = \{ p(A_1), p(A_2), \cdots, p(A_i), \cdots p(A_M) \} \quad (4)$$

Object B can be represented similarly over N colors by *MCSHR(B)* and *p(B)*. In order to define the similarity between two moving objects, a subset of $MCSHR(B)$ is firstly defined as:

$$MCSHR^{'}(B \mid C_{A_i}, \sigma) = \{C_{B_1^{'}}, C_{B_2^{'}}, \cdots, C_{B_L^{'}}\} \quad (5)$$

where the distance between $C_{B_j^{'}}, j = 1,2,\cdots,L$ and C_{A_i} is less than a given cluster distance threshold, σ.

Then, $C_{B_j|A_i}$ is defined as the most similar color to C_{A_i} in subset $MCSHR^{'}(B)$ satisfying:

$$C_{B_j|A_i} : j = \underset{k=1,\cdots,L}{arg\,min}\{d(C_{B_k^{'}}, C_{A_i})\} \quad (6)$$

The portion of C_{A_i} in object A can be simply calculated as:

$$p_{norm}(A_i) = \frac{p(A_i)}{\sum_{i=1,2,\cdots,M} p(A_i)} \quad (7)$$

Similarly, the portion of $C_{B_j|A_i}$ in object B can be calculated as:

$$p_{norm}^{[A_i]}(B_j) = \frac{p^{[A_i]}(B_j)}{\sum_{j=1,2,\cdots,N} p(B_j)} \quad (8)$$

where $p^{[A_i]}(B_j)$ is the count of $C_{B_j|A_i}$. Then, we define the similarity of colors C_{A_i} and $C_{B_j|A_i}$ as:

$$Sim(C_{A_i}, C_{B_j|A_i}) = min\{p_{norm}(A_i), p_{norm}^{[A_i]}(B_j)\} \quad (9)$$

Then, the similarity of the whole objects A and B in the direction from A to B is then defined as:

$$Sim(A,B) = \sum_{i=1}^{M} Sim(C_{A_i}, C_{B_j|A_i}) \quad (10)$$

Sim(B,A) is computed similarly, with its value obviously differing from that of *Sim(A,B)*. In order to derive a symmetric similarity measurement, an adequate minimum and maximum are defined as:

$$Sim_{min}(A,B) = min\{Sim(A,B), Sim(B,A)\} \quad (11)$$

$$Sim_{max}(A,B) = max\{Sim(A,B), Sim(B,A)\} \quad (12)$$

If $Sim_{min}(A,B)$ is less than a given discrimination threshold, $\eta_{discrim}$, the similarity of objects A and B is simply defined as:

$$Similarity(A,B) = Sim_{min}(A,B). \quad (13)$$

The rationale is that in this case the two object similarities between A and B, (11) and (12), are either very asymmetric or both low and for this reason we decide to bound them by their lowest value. Instead, if $Sim_{min}(A,B)$ is above or equal the discrimination threshold, we define:

$$Similarity(A,B) = 1 - \frac{Sim_{max}(A,B) - Sim_{min}(A,B)}{Sim_{max}(A,B) + Sim_{min}(A,B)} \quad (14)$$

In this case, we are confident that the two visual objects are possibly a same physical one. As a further verification, we choose to check the difference between the maximum and minimum similarities in a ratio form. In (14), the bigger the difference between maximum and minimum similarity, the less similar are considered the two objects. Such a definition aims to prevent asymmetric, partial matches between two objects. Eventually, matching (a 0/1 decision) is assessed if *Similarity(A,B)* is above a set appearance similarity threshold. Moreover, all the measurements above are computed over IMCHSR values.

For matching single objects in two disjoint views, we consider two available tracks, one for each view (by "track", in this paper we mean the state information for the moving object, such as position, mask and texture, versus time). Similarity measurements and matching decisions are then repeatedly computed over pairs of frames from the two tracks in frame order, and matching decisions integrated. Alternatively, one can integrate the similarity measurements, thus making the appearance similarity threshold unnecessary. Eventually, the integrated value is compared against a set post-integration threshold, leading to the final decision. For this algorithm, we opted for the frame order as an arbitrary choice to keep our algorithm of linear computational complexity in the number of frames. Complexities above linear would not suit real-time applications.

5. Experimental results and analysis

In this section, we report example results from six typical tracks from four real disjoint video surveillance cameras installed in the Faculty of Information Technology building, University of Technology, Sydney, where two moving objects have been detected and tracked. Experimental results are shown in the following sub-sections.

5.1. Matching of a same moving person in disjoint camera views

The test data reported here are from the same person recorded from two disjoint video surveillance cameras (camera *3a*, frames 001-019, and camera *5*, frames 300-318), with some of the frames shown in Figure 4. The two cameras are significantly disjoint in both space and time and the person's appearance in the two tracks could not be matched trivially. Moreover, illumination also varies significantly with the object's position within each camera view (unlike assumptions in [1,3]). However, our representation proves capable of coping with such variations in appearance. IMCSHR matching results are reported in Table 1 showing that the same moving object in the two disjoint camera views is reliably matched.

Figure 4: Moving objects from camera 3a, frames 001-009 and camera 5, frames 300-308.

Table 1: Results of IMCSHR matching – same person

Test Case	Frame No	Camera	Similarity	Matching Results
1	001-005	*3a*	0.9817	1 (Yes)
	300-304	*5*		
2	003-007	*3a*	0.9758	1 (Yes)
	302-006	*5*		
3	005-009	*3a*	0.9772	1 (Yes)
	304-308	*5*		
4	007-011	*3a*	0.9856	1 (Yes)
	306-310	*5*		
5	009-013	*3a*	0.9452	1 (Yes)
	308-312	*5*		
Integration	001-019	*3a*		100% (Match)
	300-318	*5*		

Note: with a major colors cut off level of 90%; color distance threshold = 0.05; cluster distance threshold = 0.05; discrimination threshold = 0.4; appearance similarity threshold = 0.8; and post-integration threshold = 80%.

5.2. Matching of two different people from two disjoint camera views

The test data reported here are from two different people recorded from the same video surveillance cameras (camera *3a*, frames 001-019, and camera *5*, frames 010-022), with some of the frames shown in Figure 5. The IMCSHR matching results are reported in Table 2 showing that the two different moving objects are correctly discriminated since the integrated matching rate is only 40% (much lower than the set 80% post-integration threshold).

Figure 5: Moving objects from camera 3a, frames 001-009 and camera 5, frames 010-018.

Table 2: Results of IMCSHR matching - two different people

Test Case	Frame No	Camera	Similarity	Matching Results
1	001-005	*3a*	0.3538	0 (No)
	010-014	*5*		
2	003-007	*3a*	0.7588	0 (No)
	012-016	*5*		
3	005-009	*3a*	0.7224	0 (No)
	014-018	*5*		
4	007-011	*3a*	0.8348	1 (Yes)
	016-020	*5*		
5	009-013	*3a*	0.8075	1 (Yes)
	018-022	*5*		
Integration	001-019	*3a*		40% (No match)
	010-022	*5*		

Note: Parameters are the same as those used for Table 1.

5.3. Comprehensive Matching Tests on Disjoint Camera Views

Five sets of additional track matching results from disjoint camera views, disjoint either in space or time, or both, are shown in Table 3.

Table 3: Comprehensive results of IMCSHR matching

Test Case	Camera	Typical frame similarity (IMCSHR)	Integrated matching rates
1 (Same object, time disjoint)	*3_0*	0.9785	80% (4 out of 5 matched)
	3a		
2 (Same object, space disjoint)	*3a*	0.9817	100% (5 out of 5 matched)
	5		
3 (Different objects, time and space disjoint)	*4*	0.3696	20% (4 out 5 discriminated)
	5		
4 (Same object, time and space disjoint)	*3_0*	0.8410	100% (5 out of 5 matched)
	5		
5 (Different objects, space disjoint)	*4*	0.3696	20% (4 out 5 discriminated)
	5		

Note: Parameters are the same as those used for Table 1.

Table 3 shows that the proposed method is capable of correct matching. The differences in matching rates for same and different individuals are always high and allow easy discrimination. In test case 1, a same person is viewed under a same camera in the morning and afternoon. In the morning view, there is a significant amount of natural light in the right part of the scene (with resemblances to a typical outdoor view), while artificial illumination is predominant in the left and central parts. In the afternoon, the whole view is dominated by artificial illumination, with slight changes in chromaticity. Variations in the intensity of

the R,G,B components for the moving object across and between such views are in the order of 25-30%. With the proposed method, the object is successfully matched with an integrated matching rate of 80%. The other test cases cover a variety of disjointedness in time and space. In test cases 2 and 4, the same objects are successfully matched with an integrated matching rate of 100%. In test cases 3 and 5, two different objects are successfully discriminated thanks to an integrated matching rate of only 20%.

6. Conclusions

In this paper, we have proposed a method for matching two objects based on their appearance along their tracks from disjoint camera views. Such views are challenging in that illumination conditions can be very different and the appearance of single objects vary correspondingly. Computing an exact transformation to compensate for such appearance changes is impractical since the appearance of moving objects depends on a number of illumination parameters which cannot be fully retrieved from videos, even with an initial training stage. The major contributions of this paper are a special cumulative histogram transformation mitigating the varying illumination conditions across the disjoint views and a K-means online clustering algorithm to accurately represent a moving object in its most frequent colors. Our integrated object matching approach provided reliable matching of single objects whilst at the same time discrimination between objects of differing appearance remained high. Results from experiments reported in this paper can be summarised as:

1) the special cumulative histogram transformation makes the appearance of a single object reasonably invariant across disjoint camera views while different objects are easy to discriminate;
2) the proposed K-means online clustering algorithm provides an effective appearance representation;
3) the incremental major color spectrum histogram representation (IMCSHR) copes with small view changes occurring over a window of successive frames;
4) post-matching integration of the decision made between two frames along the objects' tracks improves the reliability of object matching. The matching algorithm also has linear complexity with the number of frames.

The proposed object matching procedure can provide video surveillance applications with the ability of tracking single objects based upon color across disjoint camera views, which are predominant in existing surveillance camera networks. Such ability is useful to track assigned individuals (a "watch list") from entry to exit of a building in real time, or as a forensic tool to automatically back-track movements of people from an assigned point in time and space (such as an event of interest). At the moment, we are working on automated estimate of the method's various parameters and adding further features such as localised appearance features and various shape factors for more general object matching. In any case, color appearance features are a fundamental clue for people tracking over disjoint views and their accurate representation is a focal issue.

Acknowledgment

This research is supported by the Australian Research Council, ARC Discovery Grant Scheme 2004 (DP0452657).

References

[1] O. Javed, K. Shafique, M. Shah, "Appearance Modeling for Tracking in Multiple Non-overlapping Cameras," IEEE CS Conf. on Computer Vision and Pattern Recognition 2005, vol. 2, pp. 26-33.

[2] M. Piccardi and E. D. Cheng, "Track Matching Over Disjoint Camera Views Based On An Incremental Major Color Spectrum Histogram", IEEE Int. Conf. on Advanced Video and Signal based Surveillance 2005, Como, Italy.

[3] O. Javed, Z. Rasheed, K. Shafique, M. Shah, "Tracking Across Multiple Cameras With Disjoint Views," Ninth IEEE Int. Conf. on Computer Vision, vol. 2, pp. 952-957, 2003.

[4] Y. Weiss, "Deriving intrinsic images from image sequences," Eight IEEE Int. Conf. on Computer Vision, vol. 2, pp. 68-75, 2001.

[5] Y. Rubner, C. Tomasi, L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, November 2000, pp. 99-121.

[6] Y. Chen and E. Wong, "Augmented image histogram for image and video similarity search," SPIE Storage and Retrieval for Image and Video Databases, pp. 523-532, 1999.

[7] Liyuan Li, Weimin Huang, I.Y.H. Gu, K. Leman, Qi Tian, "Principal Color Representation for Tracking Persons," IEEE Int. Conf. on Systems, Man and Cybernetics 2003, vol. 1, pp. 1007-1012.

[8] Z. Zivkovic and B. Krose, "An EM-like algorithm for color-histogram-based object tracking," 2004 IEEE CS Computer Vision and Pattern Recognition, vol. 1, pp. 798-803.

[9] B. Mirkin, *Mathematical Classification and Clustering* Kluwer Academic Publishers, 1996.