Evaluation of MPEG7 Color Descriptors for Visual Surveillance Retrieval

James Annesley, James Orwell, John-Paul Renno
Digital Imaging Research Center, Kingston University, Kingston-upon-Thames, Surrey, UK.
{james.annesley, james, j.r.renno}@kingston.ac.uk

Abstract

This paper presents the results to evaluate the effectiveness of MPEG7 Color descriptors in Visual Surveillance retrieval problems. A set of image sequences of pedestrians entering and leaving a room, viewed by two cameras, is used to create a test set. The problem posed is the correct identification of other sequences showing the same person as contained in an example image. Color descriptors from the MPEG7 standard are used, including Dominant Color, Color Layout, Color Structure and Scalable Color. Experiments are presented that compare the performance of these, and also compare automatic and manual techniques to examine the sensitivity of the retrieval rate on segmentation accuracy. In addition, results are presented on innovative methods to combine the output from different descriptors, and also different components of the observed people. The evaluation measure used is the ANMRR, a standard in Content-Based Retrieval experiments.

1 Introduction

This paper presents research to evaluate Content-Based Information Retrieval (CBIR) techniques for the Visual Surveillance domain. The analysis of content underpins the key objectives of the Visual Surveillance program. The focus of this paper is the retrieval of information about the identity of subjects, observed with medium and far-view image sequence data from multiple cameras. In such scenarios, face recognition is not presently reliable, and so the subject can only be identified by other features such as their clothing, hence a short timescale is assumed. Nevertheless, there are several key applications for this technology, such as the facility for an operator to retrieve knowledge about the past or present whereabouts of a specific individual, or its use in a multi-camera system, in order to develop a coherent scene representation.

Content-Based Retrieval from image sequences is a well-established technology, with applications in many specific domains, such as entertainment, natural history and sport. In the visual surveillance domain, meta-data is processed from the input video data, stored alongside it, and used in queries to retrieve it. An overview of previous work on Content-Based Retrieval methods, especially in relation to the requirements of the Visual Surveillance domain, is provided in Section 2. In these fields, an important development has been the establishment of multi-purpose, extensible, open standards to define the meta-data. In particular, the Moving Picture Experts Group, International standards body, which is concerned with the delivery of the multimedia data through computer networks has produced the MPEG7 standard. This standard has accommodated many schemata for describing content that are potentially suitable for retrieval applications.

This paper presents results showing the effectiveness of standard MPEG7 descriptors in Visual Surveillance retrieval problems. The surveillance problems involve the retrieval of video data containing a given person, specified by an example image captured at a different time, or by a different camera. For this purpose, a test dataset was constructed of people observed on multiple occasions from multiple cameras. The metric used for evaluating the retrieval accuracy is the Average Normalized Modified Retrieval Rate (ANMRR), introduced in Manjunath [13].

A novel aspect to the work is the generation of separate meta-data for top and bottom components of each pedestrian. These are then combined in several schemata, to evaluate if the explicit segmentation of pedestrians into their principal clothing components assists in the retrieval accuracy of the system. A further investigation is conducted into the extent to which combinations of the color descriptors can be used to improve the retrieval accuracy. Finally, a comparison of automatic and manual segmentation is performed, to examine the sensitivity of the retrieval rate on segmentation accuracy.

1.1. Previous Work

CBIR is the technique of retrieving stored and indexed images based upon their signal content, rather than external attributes such as date, location or title. The signal content is the image data, commonly a raster scan of 8-bit values in a color space such as (r,g,b) or (y,u,v). To facilitate the retrieval process, these values are processed to create additional descriptive representations for its colors, shapes and textures etc. Some representations are designed to carry semantic meaning (e.g. wooden texture, or bicycle shape); others are simply alternative numerical descriptions of the data (e.g. Fourier co-efficients, or (h,s,v) values for each region).

The query by example technique compares stored data signals with a query data signal, in order to retrieve similar objects to the query subject and is outlined in [2, 7]. For instance, the query subject could be an image of an object, and its attributes, extracted through signal analysis of color, shape and texture, can be used to retrieve images with similar content. The query by attribute technique uses a semantic representation of some attribute and attempts to locate similar attributes in stored images. For instance: ‘Please retrieve images of white shoes or red hats’. The query by attribute method requires a semantic layer to bridge the “semantic gap” between high and low-level information. It turns pixel data into representations useful to human operators. Attempts to bridge this gap are made by [2, 5, 14]. Methods used when working with videos are discussed in [2, 5, 13], where preprocessing is used to organize and annotate the video sequence into classes dependent on the nature of the scene.

Visual surveillance is concerned with the construction of a representation of observed scene activity. Important to the task is the segmentation of individual objects from the background scene, this requires a good model of the background as described by [18, 19, 21]. The segmented objects can then be categorized, i.e. people and cars, described by [16], attempts to deal with problems caused by occlusions are described in [10], methods described by [3], show how the movements and associated data can be useful, and behavior analysis, to pick up on typical and atypical behavior of the tracked objects is shown in [4].

The MPEG7 standard for encoding of multimedia meta-data. It differs from the previous standards the MPEG group has provided, which address the encoding and decoding of the signal data. Multimedia data includes image, sound and video. The standard also addresses the organization and delivery of the meta-data. It uses a XML based meta-data document called a schema to organize the data, and a data definition language to extend the standard. It may be stored and transmitted using different data formats, including text, binary or a proprietary binary coding scheme. There is also reference software available to allow testing.

The use of MPEG7 to describe meta-data produced in the Visual Surveillance domain was first introduced by [1, 9]. The former work focuses on the delivery of the meta-data using the system components of MPEG7; the latter uses the Dominant Color and Contour-Based Shape descriptor standards to investigate algorithms to recognize pedestrians, also suggesting some in-house implementations of color and texture classifiers.

2 Methodology

In this section the experimental procedure is described, for storing, retrieving and evaluating MPEG7 meta-data for Visual Surveillance. The MPEG7 Experimentation Model (XM) was used to produce the MPEG7 color descriptors meta-data. It was also used to perform Query by Example (QBE) tests on the data. An in-house algorithm generated mean color meta-data, that is an average of each of the pixels in (r,g,b) color space in the image. The mean meta-data was compared with an Euclidean distance metric algorithm. A random classifier produced random results by selecting any of the meta-data items for the return results, subject to some rules.

Only the foreground regions of the images were used as meta-data, thereby reducing redundancy of the descriptors. The foreground regions were produced both manually and automatically. The automatic motion detection uses a per-pixel multi-modal background model in a (h,l,s) color space. The model is comprised of one 2-dimensional Gaussian representing the background color and 1-1D Gaussian representing the intensity. Pixel classification is performed in the usual manner, except for an additional procedure which attempts to identify the presence of shadows. See Fig. 2 for an automatic and manual segmentation comparison.

2.1 Dataset and Experimental Design

Video sequences showing 47 people entering and leaving a room were assembled as a test set. Each person is filmed by two cameras, which both observe the two movements IN and OUT, resulting in four image sequences for person *i*: A_i^{in}, A_i^{out}, B_i^{in},B_i^{out}, where A and B refer to the sequences captured from the two cameras (Fig. 1). Each of the undergraduate participants in the dataset provided written consent that the data can be used for research purposes, including publication on the Internet.

The set of experiments are designed to test the retrieval accuracy of query by example. This procedure requires a query subject (the example) to be compared with the contents of the surveillance database, the query objects. The retrieval process will select an ordered (ranked) list of the objects deemed to be similar to the query subject. The alternative query type is query by specification, for example, “select objects with white shoes”. This raises further issues such as the ontology of the specification, and is not included in this work.

There are three experiments presented. Experiment 1 presents retrieval results using the same camera, but different movement (e.g. comparing A^{in} with A^{out}). Experiment 2 compares the same movement with different cameras (e.g. comparing A^{in} with B^{in}); Finally, Experiment 3

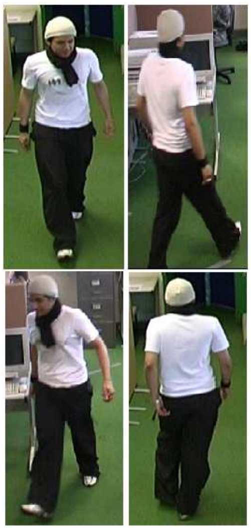

Figure 1: Example data showing the same person, from the two different cameras, each with two views. Clockwise from top left: A^{in}, B^{in}, B^{out} and A^{out}.

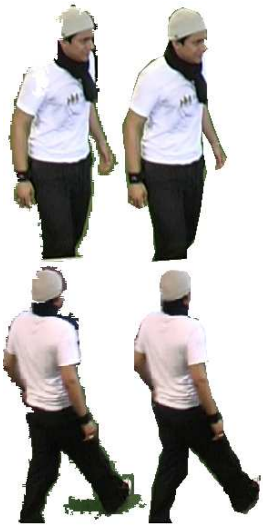

Figure 2: Automatic (left) and manual (right) segmentation of pedestrians for processing by the Color Descriptors, for the front (top) and side (bottom) cameras.

investigates the success with which sequences of an individual may be retrieved, using images from a different camera, taken at a different time (e.g. comparing A^{in} with B^{out}).

Fifteen individuals were randomly selected for the experiments. Each individual has nine separate ground truth images, and a set of three queries. Three ground truth images and one query image are used for each of the three experiments. The images taken for the ground truth data were chosen to reflect the circumstances of the experiment. The query data was taken randomly from a predetermined range of usable images and came entirely from a particular sequence. For example, the ground truth data for experiment one comes entirely from people entering the room from the front facing camera.

2.2 The Color Descriptors

The four Color Descriptors used in the experiments are outlined below: Dominant Color, Color Layout, Scalable Color and Color Structure. The experiments compare a total of eight different descriptors, since two different quantization settings are used for the last two, and in addition two validation descriptions are included to check the experimental procedure. In addition to the retrieval rate, other performance factors include the compactness of the representation, the computation required to generate the data, and the computation required to retrieve a similarity measure between two representations. The following descriptions are paraphrased from [13].

The *Dominant Color* Descriptor represents colors in an image or image region. It uses the generalized Lloyd algorithm [8] to cluster the data using the (l,u,v) color space. The cluster centers and distortion rate are calculated iteratively, and the algorithm stops when up to eight clusters are found. A connected-components algorithm joins neighborhoods of the same dominant colors and produces a global spatial homogeneity component. The clustering is optimized for similarity in human perception, where spatial resolution is less sensitive for color, than for brightness. The descriptor outputs the number of clusters, the spatial homogeneity of the image, and each cluster with a color, variance and percentage value. The color values have a resolution of 5 bits per channel, as do the percentage and spatial homogeneity components. Spatial homogeneity and variance, as included in these experiments, are optional but increase performance and computational requirements.

The *Scalable Color* Descriptor uses a Haar transformation of the color histogram, performed in (h,s,v). The output is the high and low-pass co-efficients from the transform. This representation is quantized into between 16 and 256 eight-bit values per image, depending on the required compactness - a low number of bins give a fast descriptor suitable for indexing and quick queries. The color channels are quantized unevenly, with a higher percentage taken with the hue component. The high-pass coefficients are fairly redundant and hence can be heavily compressed. The principle advantage of this descriptor is that the Haar scaling properties allow differing quantization levels can be matched to one another.

The *Color Structure* Descriptor uses a spatial structuring element when compiling the color histogram. Hence, the spatial structure in which the different colors appear is incorporated into the representation. A 4×4 kernel is passed over the image and the color channel bins are incremented if a color is present. The Color Structure descriptor uses the hue, max, min, difference (HMMD) color space, which is quantized unevenly. Different quantization levels are available, as with the scalable color descriptor, with the highest quantization levels giving the best results.

The *Color Layout* Descriptor is emphasized as a quick descriptor that is resolution independent and suitable for indexing, sketch-based retrieval and video segment identification. Extraction is performed by a discrete cosine transform (DCT) transformation in Y or Cb or Cr space. The input image is partitioned into blocks, each block changed to the mean of its color components, a DCT transform is applied to the blocks, a zigzag scanning and weighting gives binary marks. This descriptor is suitable for low-powered devices.

Techniques for similarity matching using each of the above color descriptors are given in [13]. Matching *Dominant Color* descriptors involves searching the data-set for similar distributions of colors. This can be a two-pass process, where individual colors can either be searched for individually and then combined, or where the complete descriptions are compared, as used here. The matching process calculates: the Euclidean distances between the clusters, the spatial homogeneity with a weighted difference calculation and the variance using a mixture of Gaussians measure. *Scalable Color* matching can be performed with descriptors of differing quantization and difference coefficients, due to the Haar transform. L1-norm matching can be applied to the Haar domain (sum of absolute differences) and in the histogram domain, the L1-norm degenerates to a Hamming distance because the bit-plane has been compressed to merely a sign bit. *Color Structure* matching involves equalization of query and data-set descriptors. This is more complex than histogram equalization because color quantization affects the color structure. Unlike the other descriptors, the similarity matching process is explicitly defined in the standard and involves bin unification and bin quantization stages. *Color Layout* descriptor matching uses a distance measure from the combined coefficients produced over the three color channels.

2.3 Combining Descriptors

In this section, two methods are described that require the similarity output from the above descriptors to be combined, to generate a joint ranking. Firstly, to exploit the frequent segmentation in pedestrian clothed appearance, a method is proposed by which the foreground mask for each person is split into two separate regions, giving top and bottom only meta-data. The experiments use data split automatically, where foreground region is split half-way down, and manually, where the two outer layers of top and bottom items of clothing are segmented (Fig. 3).

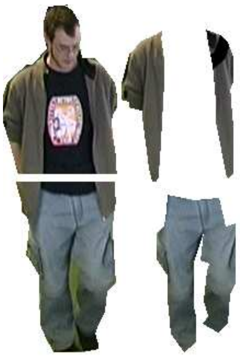

Figure 3: Example of automatic (left) and manual splitting (right) to produce Top and Bottom data.

Although these data are used separately, the intention is to combine them, to give a retrieval process that jointly uses both top-half and bottom-half meta-data (while maintaining an explicit distinction between these two, to exploit the assumption that pedestrians generally stay the same way up). This method may also have application where where people are occluded from e.g. the waist down. In the paper, this is called a ‘Spatial Combination’. Secondly, it may be the case that the different color descriptors have complementary characteristics, which, if combined appropriately, could improve the overall retrieval rate. Below, this is called a ‘Descriptor Combination’.

For both Spatial and Descriptor Combinations, there are several methods by which the individual results are combined. In either case, a potential difficulty is the incompatibility of outputs from the two Descriptors, which may have completely different units and scales of output. Although there are solutions to this problem, such as converting each into a Mahalanobis distance, this is not without complications, and so the rank output from each Descriptor was selected as the most appropriate input to combine into a joint descriptor. Four different operators were tested to combine the ranks: *addition*, *multiplication*, *minimum* and *maximum*. (The *minimum* operator selects the best (lowest) rank from the two, while the *maximum* selects the worst (highest) rank. The output from these operations are then re-ranked accordingly, and the resulting ranks are used as the final retrieval answer. Existing work on pedestrian recognition exists in [11], which also suggest combinations of descriptors.

2.4 Evaluation of Retrieval Accuracy

The Average Modified Normal Retrieval Rate (ANMRR) metric is used to evaluate the performance of the Color Descriptors. This metric is introduced in Manjunath [13], used in [6, 12, 20] and analyzed by [15]. The purpose of the metric is to allow an evaluation of different descriptors that is unbiased with respect to different sample and ground truth sizes, and correlates well with perceptual judgment about the retrieval success rate [15]. Scores are based upon the rank of results and not their value. The rank of each retrieved ground truth data is counted and penalties are issued if any of the items comes after a threshold, K. The same penalty applies to all items after K, i.e. the procedure penalizes low-ranking ground-truth items, no matter how low-ranking. The size of the ground truth set determines the rank at which the threshold is placed. The rule of thumb suggested in [13] is that K is set at twice the ground value. However, they also suggest a practical minimum of three ground truth items, for the rule to apply. Each retrieval operation is assigned an NMRR, the Modified Normal Retrieval Rate: this is averaged over all operations in the set, to produce the ANMRR:

$$NMRR(q) = \frac{MRR(q)}{1.25 \cdot K(q) - 0.5 \cdot [1 + NG(q)]}$$

where K = relevant rank mark, NG = number of ground truth data elements and q = query.

3 Results

As discussed above, a number of different CBIR techniques were evaluated in three scenarios, designed to provide different grades of difficulty. In the first scenario, the goal is to retrieve images of a person walking towards the camera; providing, as an example, an image of the person walking away from this same camera. The data-set comprised four images each of fifteen people. Six different MPEG7 descriptors were evaluated in this scenario, alongside two simple control descriptors: the mean and random descriptors, all described above. These alternatives are evaluated using the ANMRR: here, a value of 0.0 indicates perfect retrieval, and 1.0 corresponds to no retrieval at all.

A further experimental parameter is the Spatial Combination method of the data submitted to the Color Descriptors: there are a total of seven configurations: the whole region can be submitted, or top only, or bottom only; or these last two can be combined in four different ways, as discussed in Section 2.3.

First of all, the experiment is conducted with *manual* segmentation of each person in the image. Fig. 4 shows manually segmented data with automatically split top and bottom halves. This virtually eliminates contamination with background. The results for automatic segmentation and splitting are shown in Fig. 5. The results for the manually segmented top and bottom halves are shown in Fig. 7 which represent the best possible segmentation.

For all configurations of data, the random classifier gives a result of roughly 0.9, in line with theory, providing one useful validation point for the experimental procedure, and a point of reference by which the other methods may be judged. A second reference point is provided by the simple (r,g,b) mean descriptor: here, the top and whole configurations provide a retrieval rate of about 0.54. Unsurprisingly, it is less reliable to retrieve an individuals identity using only their bottom half (0.68). However, all four Combination methods, operating on Top and Bottom retrieval ranks significantly improve the performance of the mean classifier, to a rate of around 0.40.

Rank (1-K)	Gnd. Truth Retrieval (n)	NMRR (0-1)
1,2,3,x,x,x	3 in top 6	0
1,2,x,x,x,x	2 of 3 in top 6	˜0.25
1,x,x,x,x,2	2 of 3 in top 6	˜0.5
x,x,x,x,1,2	2 of 3 in top 6	˜0.75
x,x,x,x,x,x	0 in top 6	1

Table 1: Example retrieval scenarios, together with the resulting Normalized Modified Retrieval Rate (NMRR). There are three ‘true’ items to be retrieved; all the rest are ‘false alarms’.

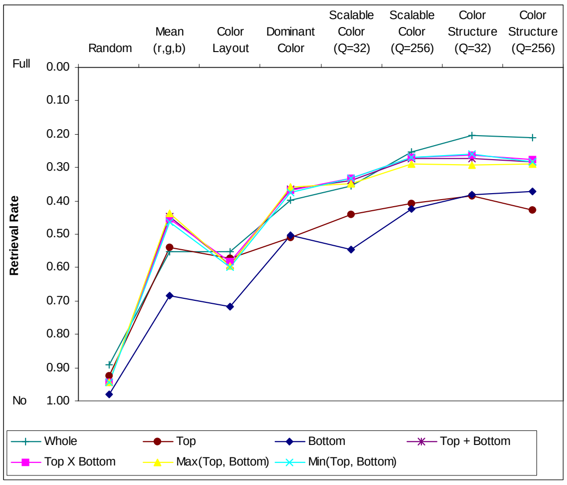

Figure 4: Retrieval Rate for Experiment 1 (same camera, different direction of motion) with manual segmentation. Compare with the results for automatic segmentation (Fig. 5) and manually segmented top and bottom (Fig. 7). The experiment suggests the most effective descriptor in this scenario is Color Structure.

The best retrieval performance is displayed by the Color Structure Descriptor. It demonstrates an ANMRR of around 0.21, on experiment 1, with manual segmentation. Where the top and bottom halves are manually segmented (Fig. 7), an improvement over the whole is noted, with Color Structure 256 combined with max operator, providing the best result (0.15).

The same techniques show similar results when process-

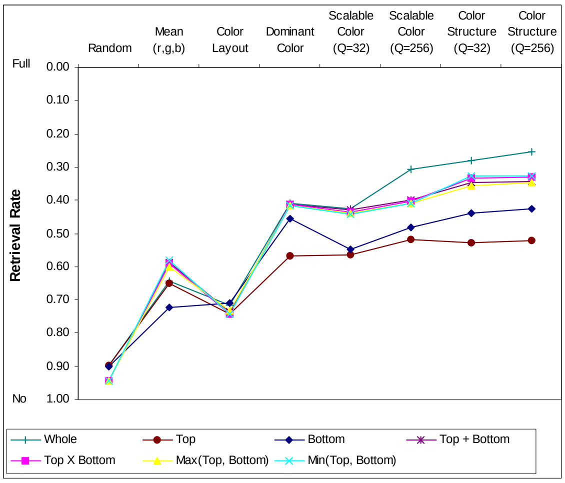

Figure 5: Retrieval Rate for Experiment 1 (same camera, different direction of motion) with automatically segmented foreground data.

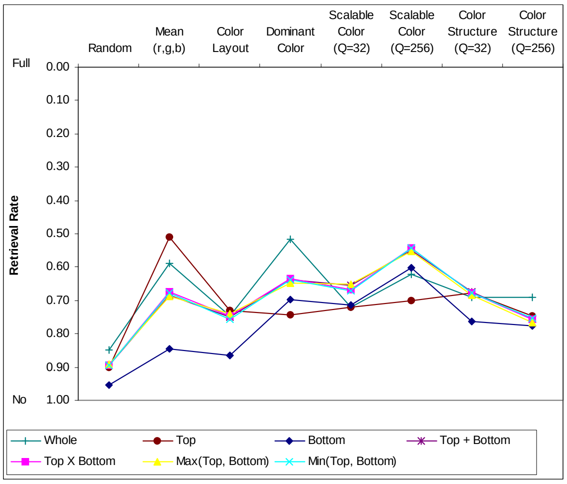

Figure 6: Retrieval Rate for Experiment 3 (different camera, different direction of motion) with automatic segmentation.

-	DC	SC 256	SC 32	CS 256	CS 32
DC	-	0.307	0.369	0.272	0.244
SC 256	0.305	-	0.408	0.248	0.290
SC 32	0.347	0.357	-	0.292	0.331
CS 256	0.278	0.252	0.317	-	0.268
CS 32	0.256	0.292	0.335	0.258	-

Table 2: Descriptor combination experiments with Experiment 1 automatically segmented data (see Fig. 5), combined with the min operator. The results suggest an improvement over a single descriptor.

ing automatically segmented foreground data. These regions will exhibit a higher incidence of missing components and background contamination, than the manually segmented data. The performance of the Color Structure Descriptor degrades moderately, from 0.21 to 0.27. The Color Layout Descriptor shows most sensitivity to noisy data: its retrieval rate drops from 0.58 to 0.73. Using the proposed system, the same high retrieval rate is not maintained if the source camera for the query image is different to the source camera for the stored dataset. In experiment 3 (Fig. 6), this retrieval rate is plotted, for the same spread of Color Descriptors. None of these demonstrated a higher retrieval rate than what was obtained for a simple (r,g,b) meta-data (and corresponding Euclidean distance measure). Top and Whole segmentation strategies worked equally well (ANMRR=0.50) using this simple mean descriptor.

To help improve the performance between the cameras, an implementation of the Gray World algorithm for color constancy [17] was used but did not improve the results. Finally, experiments designed to compare performance of different combinations of Color Descriptors showed that small but significant improvements were possible. Table 2 plots retrieval rates obtained in Experiment 1 for automatic segmentation policy. Combining Color Structure and Dominant Color with a *min* operator gives a retrieval rate of 0.244, compared to 0.253 or 0.424, when these are used separately.

4 Conclusions

An experimental methodology has been described and used to evaluate the effectiveness of MPEG7 Color Descriptors for Content Based Retrieval of Surveillance Data. The results clearly illustrate the relative performance of these descriptors in retrieving images of people in an indoor environment. Although there are cases for which a segmentation into Top and Bottom improved results, especially when manually segmented, the best results are obtained (for a single camera) when all foreground data is input into the Color Structure Descriptor.

The ANMRR provides a useful, unbiased, bounded indication of the performance of the retrieval process. However, it fails to adequately address certain issues familiar to Visual Surveillance researchers. For example, the rank ordering method cannot in itself provide evidence that a given query example does not appear in a dataset: there will always be one element of the data set most similar to the example. Similarly, probabilistic estimates of identity, for incorporation with other uncertain cues, are not easily deduced from the rank method. One challenge is the unification of retrieval metrics across the research communities.

For multiple camera datasets, the MPEG7 Color Descriptors do not outperform the simple (r,g,b) mean descrip-

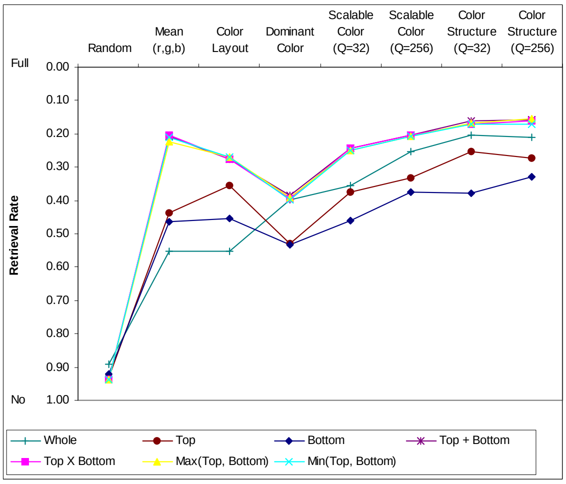

Figure 7: Retrieval Rate for Experiment 1 (same camera, different direction of motion) with manual segmentation and manual segmentation of top and bottom clothing.

tion of foreground data. It is suspected that the lack of color constancy is responsible for this degradation of performance. This clearly indicates a useful direction for future work, i.e. specification of a preprocessing method through which the multi-camera retrieval rate can be enhanced.

References

[1] W. P. Berriss, W. G. Price, and M. Z. Bober. The use of mpeg-7 for intelligent analysis and retrieval in video surveillance. In *IEE Symposium of Intelligence Distributed Surveillance Systems*, page 8/1, UK, 2003. Mitsubishi Electric.

[2] A. Del Bimbo. *Visual information retrieval*. Morgan Kaufmann Publishers, San Francisco, 1999.

[3] J. Black, T. J. Ellis, and D. Makris. A hierarchical database for visual surveillance applications. In *IEEE International Conference on Multimedia and Expo (ICME2004)*, volume 3, page 1571, 2004 2004.

[4] F. Cupillard, F. Brmond, and M. Thonnat. Behaviour recognition for individuals, groups of people and crowds. In *IEE Proceedings of the IDSS Symposium - Intelligent Distributed Surveillance Systems*, 2003.

[5] A. Dorado, J. Calic, and E. Izquierdo. A rule-based video annotation system. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):622–633, MAY 2004.

[6] A. Doulamis and N. Doulamis. Performance evaluation of euclidean/ correlation-based relevance feedback algorithms in content-based image retrieval systems. In *IEEE International Conference on Image Processing*, volume 1, page 737, 2003.

[7] H. Eidenberger. Query model based content-based image retrieval. In *ACM Multimedia Conference*, 2000.

[8] A. Gersho and R. M. Gray. *Vector quantisation and signal compression*. Kluwer Academic Publishers, 1992.

[9] K. Grant, A. T. Lindsay, M. Mainds, and A. Perrott. Retrieve: Realtime tagging and retrieval of images eligible for use as video surveillance. In *IEE Symposium on Intelligent Distributed Surveillance Systems*, 2003.

[10] D. Greenhill, J. Renno, J. Orwell, and G. A. Jones. Occlusion analysis: Learning and utilising depth maps in object tracking. In *Proceedings of British Machine Vision Conference*, pages 467–476, 2004.

[11] M. Hahnel, D. Klunder, and K. F Kraiss. Color and texture features for person recognition. In *IEEE International Joint Conference on Neural Networks*, volume 1, page 652, 2004.

[12] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, JUN 2001.

[13] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7*. John Wiley and Sons, Ltd., San Francisco, 2002.

[14] P. Muneesawang, H. S. Wong, J. Lay, and L. Guan. *Learning and adaptive characterization of visual contents in image retrieval systems*, chapter 11. Handbook of Neural Network for Signal Processing. CRC Press, 2002.

[15] P. Ndjiki-Nya, J. Restat, T. Meiers, J. R Ohm, A. Seyferth, and R. Sniehotta. Subjective evaluation of the mpeg-7 retrieval accuracy measure (anmrr). Technical Report M6029, 2000.

[16] J. Renno, J. Orwell, and G. A. Jones. Learning surveillance tracking models for the self-calibrated ground plane. In *Preceedings of the British Machine Vision Conference*, 2002.

[17] G. Schaefer. How useful are colour invariants for image retrieval? In *Int. Conference on Computer Vision and Graphics*. Kluwer Academic Publishers, 2004.

[18] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In

IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2, page 252, 1999.

[19] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Proceedings of IEEE International Conference on Computer Vision*, pages 255–261, 1999.

[20] K. Wong and L. Po. Mpeg-7 dominant color descriptor based relevance feedback using merged palette histogram. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, page 433, 2004.

[21] M. Xu and T. J. Ellis. Illumination-invariant motion detection using colour mixture models. In *British Machine Vision Conference, BMVA*, pages 163–172, 2001.