Using the MPEG-7 Colour Structure Descriptor for Human Identification in the POLYMNIA System

Andreas Kriechbaum, Werner Bailer, Helmut Neuschmied, Georg Thallinger
(Joanneum Research, Institute of Information Systems & Information Management, Austria
{firstname.lastname}@joanneum.at)

Abstract: The POLYMNIA project aims to develop an intelligent cross-media platform for personalized leisure and entertainment in theme parks or recreation venues. The visitors of the theme park are – on request – identified, tracked and recorded individually in order to create personalised photos and videos documenting the visit. One specific problem in this system is the identification of humans across different cameras and under varying environmental conditions. We use the MPEG-7 Colour Structure Descriptor (CSD) for this purpose which has been reported to perform well for this application. We propose a new distance function for the CSD, the weighted city block distance. Evaluation shows that the new matching function yields better results than the distance proposed in the standard.

Keywords: MPEG-7, Colour Structure Descriptor, CSD, distance function, matching, colour similarity
Categories: H3.1, H3.3, H5.1

1 Introduction

The POLYMNIA project aims to develop an intelligent cross-media platform for personalized leisure and entertainment in theme parks or recreation venues. The visitors of the theme park are – on her request – identified, isolated, tracked and recorded individually by a system with multiple cameras, placed at various locations within the venue, in order to provide a high quality customized souvenir, in which visitor will appear as the real protagonist. In addition personalized electronic content can be embedded into the video of visitor's activities in real time. The scenario of POLYMNIA is as follows. During her entrance in the venue or the thematic section, the visitor registers to the system and a high definition camera captures her face and body, in addition her personal data is recorded. POLYMNIA uses multiple cameras placed in the tour trail of the venue to isolate, detect, track and record the visitor continuously. The cameras within the theme park work simultaneously for multiple visitors. The visitor can notify remotely located people to enter the Internet to watch her tour in real time. The final digital product is produced when the visitor leaves the venue and can comprise still images on a PhotoCD and/or a video on a DVD.

The goals of this project pose a number of technological challenges: The first task is to recognize important content in the video streams of multiple cameras, i.e. the visitors to be identified and tracked. When the visitors are detected and located the stored image content created during registration will be matched with the detected image content to identify the visitor. In the POLYMNIA environment a combination of face recognition and separate body recognition is used, as face recognition requires high-resolution images, which are not always available. A tracking module is used to

track the occurrence of a visitor on each camera. Together with the human identification, which allows interrelating the occurrences of one visitor in the videos taken by different cameras, the visit of one person can be covered. The tracking and identification results are described using MPEG-7. A summary of the visit of one person is then generated and can be used to create the customized media products for the visitor.

This paper deals with the step of human identification in the video data. We review approaches to human identification and explain why we selected the MPEG-7 Colour Structure Descriptor. A new distance function, which is designed for the requirements in our application is proposed and evaluated together with other distance functions proposed in literature.

2 Colour Descriptors for Human Identification

The problem of identifying humans across multiple cameras can be divided into the sub-problems of segmenting the regions corresponding to the different persons and the similarity matching between a template region and one segmented from the captured image sequence. Segmentation is a hard problem, that is somewhat simplified when static cameras are used, but there remain still a number of problems such as occlusion, multiple persons grouped together, etc. [Park and Aggarwal 2002][Zhao and Nevatia 2003][Beleznai et al. 2005].

In this work we concentrate on the second step, i.e. the matching between the regions. Basically a number of different visual features could be used for this purpose. Because of the use case in POLYMNIA, which shall not require any action to be taken by the visitor in order to be identified, a number of approaches that need high-resolution images taken under good conditions such as face recognition cannot be applied. When considering low-level features, shape cannot be used as this is not a sufficiently discriminative feature for humans and we cannot rely on exact segmentation. Thus colour and texture features remain as candidates. A survey on the performance of these features for human identification can be found in [Hähnel et al. 2004]. There are of course many more approaches, but many of them need extensive training sets to ensure reliable results. In the POLYMNIA use case, it is not possible to create large training sets for each visitor in the registration phase.

The literature on human identification using colour and texture descriptors suggests the use of the MPEG-7 Colour Structure Descriptor (CSD) [Annesley et al. 2005] [Örten et al. 2005]. The CSD [Messing et al. 2001] seems to be the most appropriate global visual feature descriptor for body recognition because it not only compares colour histograms of two regions like many others. This MPEG-7 descriptor extracts colour histograms and spatial information of the colours in the body region of the people to be identified.

3 MPEG-7 Colour Structure Descriptor (CSD)

Part 3 (“Visual”) [ISO 2001a] of the MPEG-7 standard proposes a number of descriptors for the low-level visual features colour, texture, motion and shape. These descriptors are designed for use in multimedia applications such as similarity matching. The descriptors can be applied to whole images or to regions thereof. It has

to be noted, that only the representation of the feature descriptors is normative according to part 3, while the methods proposed for feature extraction and matching in part 8 of the standard [ISO 2001b] are only informative. The MPEG-7 colour and texture descriptors have been presented in [Manjunath et al. 2001]. Among them is the Colour Structure Descriptor (CSD), which is described in more detail in [Messing et al. 2001].

The CSD is based on the approach of colour histograms and their benefits, namely the invariance to geometrical transformations of the image and to noise. Also similarity matching between histograms can be done efficiently and there exist approaches for matching histograms with differing number of bins and quantisation. The drawback of the invariance of histograms to geometric transformations is that all spatial information about the colour distribution in the image is lost. To keep some of this information, a structuring element can be used when creating the histogram. Instead of counting each pixel value independently, the structuring element slides over the image and the counts of the bins corresponding to colours lying inside the structuring element are increased. This leads to different histograms depending on the homogeneity of the spatial distribution of a colour.

3.1 Extraction and Representation

The CSD is extracted from an image in the HMMD colour space [Manjunath et al. 2001], which has been introduced in the MPEG-7 standard. A structuring element with 64 samples is used. While the number of samples is fixed, the spatial extent of the structuring element is adjusted to the image size. The accumulated bin values are normalized by the number of locations of the structuring element and non-linearly quantized. The histogram size is variable and may be 32, 64, 128 or 256.

3.2 Matching

As the standard allows different histogram sizes, it also proposes an approach for re-sizing descriptors of different size to make them comparable. Once they have the same size, the L^1-norm (city block distance) is used as matching function [Messing et al. 2001]:

$$dist(h_s, h_s') = \sum_{i=1}^{N} \left| h_s(i) - h_s'(i) \right|$$

where h_s and h_s' are the colour structure histograms being compared and N is the number of bins. According to [ISO 2001b], the L^1-norm has been found to perform best among all L^p-norms. However, we have found that the L^1-norm used here has some drawbacks for CSD comparison. It is sensitive to slight shifts between the bins of the two histograms that may be caused by global variations such as change of the lighting conditions of the images from which the descriptors have been extracted. Another problem is that large deviations of only sparsely represented colours may have a severe influence on the overall matching result. This causes problems when applying the descriptor to automatically segmented regions, that may have inexact borders introducing small amounts of background colours, which are not present in the other descriptor.

4 CSD Matching Approaches

Recently, a number of different distance functions for the CSD have been proposed and evaluated, for example in [Rubner et al. 2000], [Rubner et al. 2001] and [Eidenberger 2003]. In the evaluation of [Eidenberger 2003] it turned out that the city block distance does not achieve the best results. The best overall distance measure was the Meehl index. In [Rubner et al. 2000] the Jeffrey distance and the EMD lead to the best results, in [Rubner et al. 2001] the evaluation showed that there is no distance which leads to best overall results, but EMD is found to perform well. However, the results are highly data dependent.

In the following we discuss the distance functions proposed in the literature and a new function we have developed for our application. The *city block distance* is the distance function proposed in the MPEG-7 standard [see Section 3.2]. This function yields good precision, but the recall is poor for the classification of humans in different environments and under changing lighting conditions. Further candidates for distance measures are the *Euclidean distance*, the *Jeffrey distance* [Rubner et al. 2000], which reduces the influence of noise and the size of histogram bins, the *Chi-square distance* and the *histogram intersection*, which allows to handle partial matches of histograms. The *earth movers distance* (EMD) [Rubner et al. 2000] is an approach that is based on the minimal cost that has to be paid to transform one distribution to another one. It is more robust because it avoids quantization and other binning problems.

We propose the *weighted city block distance,* which is a new distance function that has been designed to better fit the requirements of our application. The modification made to the city block distance is that we introduce weights for every bin difference in the two different colour histograms. The weights depend on the number of pixels assigned to a bin in each of the images. This correlates the influence of the difference between corresponding bin values to the support area for the bins in the images. The weights are defined as

$$w_{h_s h_s'}(i) = \frac{h_s(i) + h_s'(i)}{\sum_{i=1}^{N} | h_s(i) + h_s'(i) |}$$

and the distance function is thus

$$dist(h_s, h_s') = \sum_{i=1}^{N} w_{h_s h_s'}(i) \cdot | h_s(i) - h_s'(i) |$$

This modification optimizes the recall and increases the robustness to small changes in illumination and pose.

5 Evaluation

An evaluation of different distance functions has been performed with the test environment and results described in the following.

5.1 Test Environment

In the evaluation two databases are used where each database contains images of fourteen persons. The first database contains one frontal image of every person. This frontal image is used as the registration image in respect to the POLYMNIA system.

The second database contains three images of every person to evaluate if the recognition results improve if more than one image of a person is used. These three images per person include a frontal image, an image from the side and an image from the back. The motivation is to achieve a pose-independent matching. In this test, we assume that the segmentation of the person yielded a reasonable result (without showing too much background or cutting parts of the person), which is realistic in an application scenario that uses static cameras.

In both databases the ground truth data has been collected manually. For every person there are at least three matches to test different poses of the persons and under changing lighting conditions.

5.2 Results

The evaluation results show that recognition rates of up to 90 percent can be reached. The highest recognition rate was achieved by the weighted city block distance we have proposed. This supports the assumption that colours sparsely represented in the image influence the recognition results. A further result is that multiple images of one person do not always improve the recognition rate. The results of the evaluation on the two databases are presented in Figure 2. The earth movers distance yields lower recognition rates than other distance functions if the images being compared are still quite similar. The large improvement of the performance of the earth movers distance when using three instead of one image per person is based on the properties of this distance measure. If matching between images showing different persons, but with similar colour values and distributions is performed, the EMD yields a relatively low distance value, resulting in false positives. If more

Figure 1: On the left example images of the first database, on the right example images from the second database.

images per person are used, this effect is reduced. However, when the images differ more, e.g. due to lighting changes, the distance increases less than with other functions.

6 Conclusion

Identification of humans across multiple cameras and under varying environment and lighting conditions is an important part of the workflow in the POLYMNIA system. Based on previous work described in the literature, we have chosen to use the MPEG-7 Colour Structure Descriptor (CSD) for matching candidate regions, as it has been reported to perform best.

However, in our applications we encounter conditions under which the distance function proposed in the MPEG-7 standard, the city block distance, does not perform satisfactorily. These conditions are lighting changes and patches with background colour introduced due to segmentation errors. We have defined a new distance function, the weighted city block distance, in order to overcome these problems. The evaluation of this function and some others proposed in the literature on a set of real-world data from POLYMNIA shows that the weighted city block distance performs best. The evaluation also shows that using multiple templates per person does not always improve the recognition performance.

Acknowledgements

The authors would like to thank Herwig Rehatschek and Werner Haas as well as several other colleagues at JOANNEUM RESEARCH, who provided valuable feedback. This work has been funded partially under the 6th Framework Programme of the European Union within the IST project ”POLYMNIA“ (IST-2-004357, http://www.polymnia-eu.org).

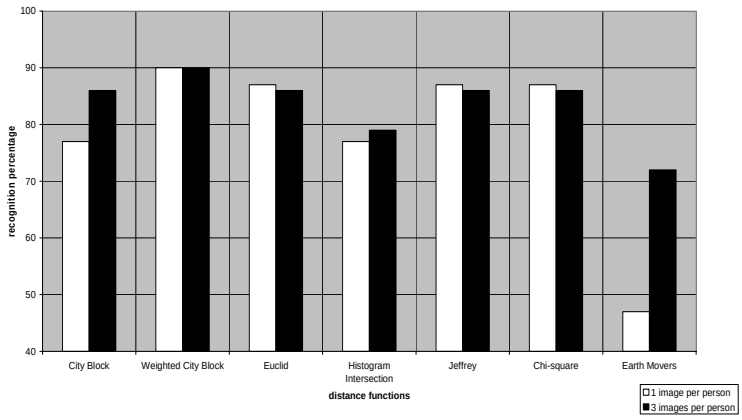

Figure 2: Evaluation results of the proposed matching algorithms for one and three images per registered person. The diagram shows the percentage of the correctly identified persons of all occurrences of this person in the database.

References

[Annesley et al. 2005] J. Annesley, J. Orwell, J. P. Renno, Evaluation of MPEG-7 color descriptors for visual surveillance retrieval, 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation, Oct. 2005.

[Beleznai et al. 2005] C. Beleznai, B. Frühstück, H. Bischof, W. Kropatsch, Model-Based Occlusion Handling for Tracking in Crowded Scenes, Proc. Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition, 2005.

[Eidenberger 2003] H. Eidenberger, Distance measures for MPEG-7-based retrieval, ACM Multimedia Information Retrieval Workshop, 2003.

[Hähnel et al. 2004] M. Hähnel, D. Klünder, K. F. Kraiss, Color and Texture Features for Person Recognition, International Joint Conference on Neural Networks IJCNN 200 , vol. 1, pp. 647-652, July 2004.

[ISO 2001a] Information Technology—Multimedia Content Description Interface, Part 3: Visual. ISO/IEC 15938-3, 2001.

[ISO 2001b] Information Technology—Multimedia Content Description Interface, Part 8: Extraction and use of MPEG-7 Descriptions. ISO/IEC 15938-8, 2001.

[Manjunath et al. 2001] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, A. Yamada, Color and texture descriptors, IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, nr. 6, pp. 703-715, Jun. 2001.

[Messing et al. 2001] D. S. Messing, P. van Beek, J. H. Errico, The MPEG-7 color structure Descriptor: image description using colour and local spatial information, Proc. IEEE International Conference on Image Processing, pp. 670-673, Thessaloniki, GR, Oct. 2001.

[Ojala et al. 2002] T. Ojala, M. Aittola, E. Matinmikko, Empirical evaluation of MPEG-7 XM color descriptors in content-based retrieval of semantic image categories, ICPR02, 2002.

[Örten et al. 2005] B. B. Örten,, M. Soysal, A. A. Alatan, Person Identification In Surveillance Video By Combining MPEG-7 Experts, WIAMIS'05 (Workshop on Image Analysis for Multimedia Interactive Services), April 2005.

[Park and Aggarwal 2002] S. Park, J.K. Aggarwal, Segmentation and tracking of interacting human body parts under occlusion and shadowing, Proc. Workshop on Motion and Video Computing, 2002, pp. 105- 111, Dec. 2002.

[Rubner et al. 2000] 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, pp. 99–121, 2000.

[Rubner et al. 2001] Y. Rubner, J. Puzicha, C. Tomasi, J. Buhmann, Empirical Evaluation of Dissimilarity Measures for Color and Texture, Computer Vision and Image Understanding vol. 84, pp. 25-43, 2001.

[Wang et. al. 2003] S. Wang, L. T. Chia, D. Rajan, Efficient image retrieval using MPEG-7 descriptors, Proc. IEEE International Conference on Image Processing 2003, vol. 2, pp. 509-512 vol. 2, 2003.

[Zhao and Nevatia 2003] T. Zhao, R. Nevatia, Bayesian Human Segmentation in Crowded Situations, Proc. IEEE Conference on Computer Vision and Pattern Recognition, Jun., 2003.