Content-Based Image Descriptors for Enhanced Person Annotation in Personal Digital Photo Archives

Saman H. Cooray and Noel E. O'Connor
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Ireland
Email: coorays@eeng.dcu.ie

Abstract

In this paper we investigate the use of content-based image descriptors for enhancing the performance of person annotation in personal photo management applications. The descriptors examined are related to the context of person recognition through face and body-patch feature matching in personal digital photos. In order to identify the best performing content-based descriptors, we first study a number of colour and texture descriptors for body-patch matching and face recognition descriptors for face matching using a suitably chosen data set taken from typical personal photo collections. We then analyse the performance of three different fusion schemes to identify the best combination of colour, texture and face recognition descriptors. Finally, we apply those descriptors to the problem of person annotation and measure their performance using a test data set, which comprises 7 different real-life personal photo collections. The experimental results illustrate that combining body-patch feature matching with face recognition significantly improves the performance of person annotation. We further show that combining colour with texture leads to improved performance of body-patch matching. The content-based image descriptors identified in this paper show great potential for person annotation in personal photo management applications.

1. Introduction

Due to the increasing popularity of digital cameras, largely fueled by emerging advanced technologies and falling prices, the task of picture-taking has become much easier and more enjoyable for typical home users. As a result, they are taking more digital photographs than ever before, leading to significant increases in the size of their photo collections. Despite such a dramatic change in the perspectives of users, the lack of technology for automatically organising large personal photo archives remains a crucial drawback in digital photography. Addressing this problem, there has recently been significant research interest in technologies for supporting effective personal photo management [14], [15], [12], [6].

In the semi-automatic person annotation prototype system proposed by Zhang *et al.* [14], face and body-patch features are used for similarity matching from which the system generates a list of name candidates through statistical learning approaches. Zhao *et al.* [15] proposed an automated method to annotate family photos by clusters using evidence from face, body, and context information. Suh and Bederson [12] developed a semi-automatic photo annotation and recognition interface (SAPHARI) for personal photo management, enabling users to update the automatically generated metadata interactively and incrementally. They proposed hierarchical event-based clustering using time information and person-based clustering using clothing information, facilitating bulk annotation within automatically identified event groups. The EasyAlbum system proposed by Cui *et al.* [6] uses novel techniques for cluster annotation, contextual re-ranking and adhoc annotation through innovative user-interface techniques. They employ face and clothing information in combination to form clusters of the people present in the photo collection.

Whilst some success has been achieved in adopting content-based image descriptors for person annotation in various research paradigms, the basis for selecting the content-based descriptors for person recognition in this challenging application, however, remains unclear. For example, the type of body-patch and face descriptors used in such approaches varies from one approach to another. It is clear that the challenges associated with content-based analysis technologies, largely caused by the uncontrolled conditions under which the personal photos are typically taken, are the major bottlenecks in deploying personal photo management prototypes in a practically viable system. Yet, to the best of our knowledge, there has been no thorough performance evaluation in the literature on the use of content-based image descriptors to this problem. Thus, identifying the best performing content-based image descriptors for improved person annotation in personal photo archives constitutes the main focus of the work presented in this paper.

In this experimental paradigm, simulations are carried out in relation to semi-automatic person annotation where the user is provided with a list of name suggestions automatically generated by the system, enabling interactive

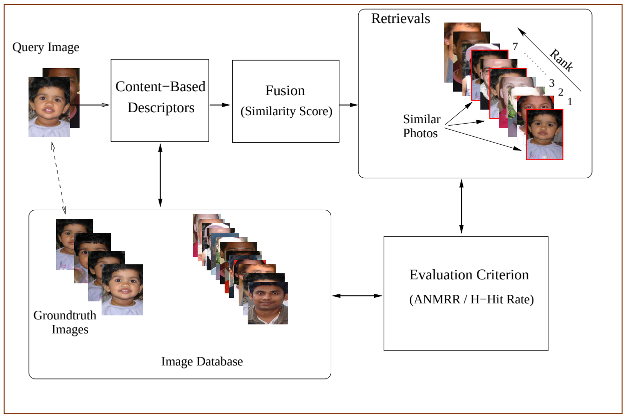


Figure 1. Test-bed environment.

annotation of the people in the collection. Key contributions from this research include identifying suitable colour, texture and face recognition descriptors, and investigating the effect of the identified descriptors on person annotation in real-life personal photo collections. The organisation of the paper is as follows: The test-bed environment used for this experimental study is described in Section 2. A description of the content-based image descriptors studied in this paper is given in Section 3. Experimental results are then presented in Section 4. Section 4.1 is devoted to a discussion on comparison of content-based image descriptors. Identifying a suitable fusion scheme and the best combination of colour, texture and face recognition descriptors is discussed in Section 4.2. In Section 4.3, a performance evaluation of person annotation using the identified descriptors is presented. Finally, a conclusion is given in Section 5, summarising the contribution of the paper and future research directions.

2. Test-bed Environment

The test-bed used for this experimental study comprises the modules for constructing test data sets, data fusion at similarity score level, and computing the performance figures using performance evaluation criteria, as depicted in Figure 1.

2.1. Test Data Sets

We use two different test data sets for identifying the best content-based image descriptors and evaluating the performance of those descriptors on person annotation.

2.1.1. Set-1. Test data set-1 comprises two sets of 45 query images for identifying the best-performing body-patch and face descriptors (see Figure 2 for an example set of body-patch images established in this data set). An automatic face detection technique [5] was first applied to each of the source images in order to locate faces. Body-patch images were then extracted relative to the position and size of each detected face in the image. Query images were selected in such a manner that each query image has varying numbers of similar images, which are hereafter termed “ground-truth images” in this paper. In total, there are 204 ground-truth images, implying that a given query has an average number of 4.5 ground-truth images in this data set.

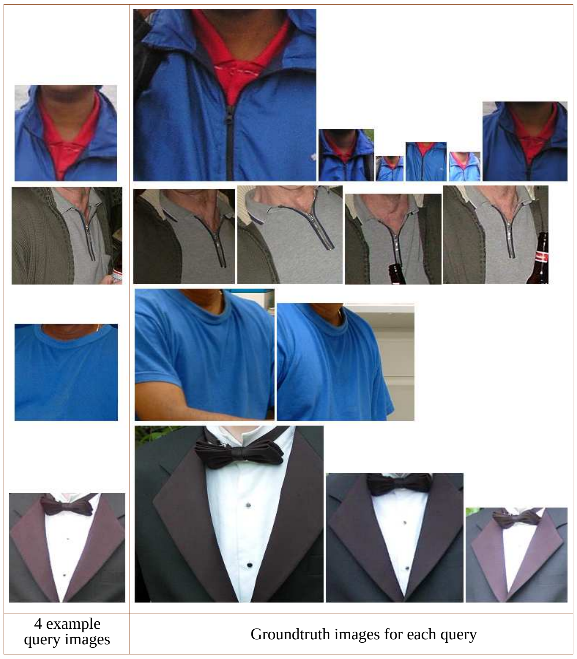
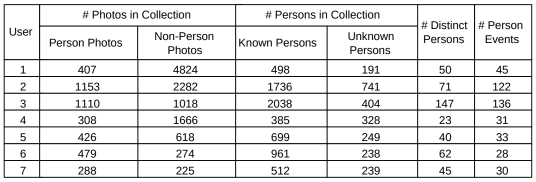


Figure 2: Example query and their ground-truth images for body-patch matching.

2.1.2. Set-2. Table 1 presents a statistical description of test data set-2 corresponding to 7 users (user 1 to user 7) of the MediAssist system [5]. These photo collections comprise different types of events, such as birthday parties, meetings, family gatherings, graduation ceremonies, weddings, etc. Table 1 also describes the characteristics of the photo collection that each user has donated in terms of the number of photos that contain people (Face Photos) and that do not contain people (Non-face Photos), the number of known and unknown faces (Known Faces, Unknown Faces) in each collection, the number of distinct faces (Distinct Faces), and the number of person events (Person Events). The number of distinct faces in a collection corresponds to the number of known people that possess unique identities whereas the number of person events corresponds to the number of events formed using the photos that contain people. All automatic face detection results were carefully inspected to ensure that they are labeled correctly. In particular, all profile faces were labeled manually since the automatic face detection technique [5] employed in our system is limited to detecting only frontal faces.

Table 1. A statistical description of the test data used for the evaluation of person annotation performance.



User	# Photos in Collection		# Persons in Collection		# Distinct Persons	# Person Events
	Person Photos	Non-Person Photos	Known Persons	Unknown Persons		
1	407	4824	498	191	50	45
2	1153	2282	1736	741	71	122
3	1110	1018	2038	404	147	136
4	308	1666	385	328	23	31
5	426	618	699	249	40	33
6	479	274	961	238	62	28
7	288	225	512	239	45	30

2.2. Performance Measure Criteria

In this paper, we use two performance evaluation criteria, namely Average Normalised Mean Retrieval Rate (ANMRR) and H-Hit rate, which have been successfully employed in numerous research paradigms in the past [4], [9]. The ANMRR measure is used to identify the best-performing content-based image descriptors for person recognition whereas the H-Hit rate criterion is used to evaluate the performance of person annotation.

2.2.1. ANMRR Measure. ANMRR takes into account not only the number of correct items retrieved for a given query but also how highly they are ranked in the list of retrieved items. It is defined as the average of NMRR (see Equation 1) values taken over a range of queries.

$$NMRR(q) = \frac{\sum_{k=1}^{NG(q)} \frac{Rank(k)}{NG(q)} - 0.5 - NG(q)/2}{K + 0.5 - NG(q)/2} \qquad (1)$$

where $NG(q)$ denotes the number of ground-truth items marked as the result images for the query q, Rank(k) is the ranking of the ground-truth item in the list of retrieved items. K is defined as $min(4 \cdot NG(q), 2 \cdot GMT)$, where GMT is the maximum value of ground-truth items for all queries. The values of ANMRR always lie in the range [0,1], with smaller values representing better retrieval performances.

2.2.2. H-Hit Rate. H-Hit rate is a performance evaluation method that has been used in previous person annotation paradigms [4], [10]. A “Hit” is said to happen if the true name of the person is present in the predicted name-list. Assuming that the entire collection is divided into two sub-collections: training (C_1) and test (C_2) with N_1 and N_2 persons in them, H-Hit defines the prediction accuracy for a given query with H indicating the number of candidates in the list:

$$H - Hit = \frac{1}{N_2} \sum_{f \in C_2} hit_{H,C_1}(f) \qquad (2)$$

where $hit_{H,C_1}(f)$ is 1 if f is included in the suggested list of H names taken from C_1, and 0 otherwise.

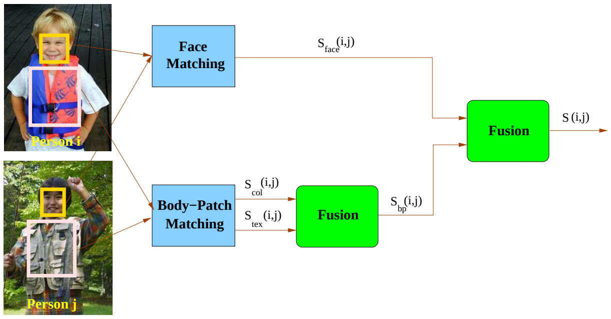


Figure 3: Fusion of colour, texture and face recognition descriptors at similarity score level.

2.3. Content-Based Descriptor Fusion

In this paper, fusion of colour, texture and face recognition descriptors is studied at similarity-score level using three different fusion methods, namely the *weighted average*, *similarity score product* and *max score*, which are expressed by Equations 3, 4 and 5 respectively. In order to study the behavior of these three fusion schemes as well as to identify the optimal combinations of the best-performing descriptors, experiments are conducted as a two-stage cascade approach illustrated in Figure 3.

$$S_{avg} = \alpha(1 - d_1) + (1 - \alpha)(1 - d_2) \qquad (3)$$

$$S_{product} = (1 - d_1)^{\alpha}.(1 - d_2)^{1/\alpha} \qquad (4)$$

$$S_{max} = max[\alpha(1 - d_1), (1 - \alpha)(1 - d_2)] \qquad (5)$$

where S represents the final similarity score, d_1 and d_2 are the normalised distance values corresponding to two feature descriptions, and α is the weighting factor.

Having determined the weighting value for each descriptor, similarity between person i and j can be computed using:

$$S(i,j) = \alpha_{col} S_{col} + \alpha_{tex} S_{tex} + \alpha_{face} S_{face} \qquad (6)$$

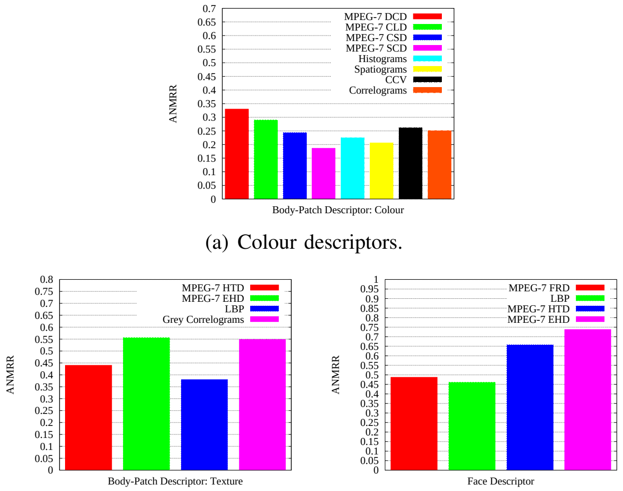
3. Descriptors Studied for Person Recognition

In order to identify the best performing content-based image descriptors, we study some of the prominent colour, texture and face recognition descriptors from those available in the literature. We choose colour histograms [13], colour coherent vectors (CCV) [11], colour correlograms [8], colour spatiograms [3], MPEG-7 dominant colour descriptor (DCD) [9], MPEG-7 colour layout descriptor (CLD) [9], MPEG-7 colour structure descriptor (CSD) [9] and MPEG-7 scalable colour descriptor (SCD) [9] as the potentially useful colour descriptors for this study. The texture descriptors chosen include the MPEG-7 edge histogram descriptor (EHD) [9], MPEG-7 homogeneous texture descriptor (HTD) [9], local binary pattern (LBP) descriptor [2], and grey correlograms [8]. The MPEG-7 face recognition descriptor (FRD) [1], LBP descriptor, MPEG-7 EHD and MPEG-7 HTD are included in the study for identifying a suitable face recognition descriptor.

4. Experiments and Results

4.1. Comparison of Content-Based Descriptors

To identify the best performing colour, texture and face recognition descriptors for person recognition in personal digital photo archives, we first carry out a performance evaluation of a selected number of content-based descriptors using the test data described in Section 2.1.1. The performance figures of each of the colour, texture and face recognition descriptors is measured for all 45 query images following which the best-performing descriptor is identified as the one outputting the highest of the average performance figures. We use the ANMRR measure as the retrieval-performance evaluation criterion in this experiment. In all cases where conventional histogram-based descriptors are used, i.e. histograms, spatiograms, CCVs and correlograms, we ensure that the resulting feature vector length is kept at 256 or close to that level. In using the MPEG-7 descriptors, which are still relatively short in length at their best retrieval accuracy, our basis for experiments was made in such a manner that no compromise is made to balance out the levels of retrieval accuracy against descriptor size, but to extract a descriptor that allows us to get the maximum retrieval accuracy. We have also performed experiments using different colour spaces such as RGB, HSV and CIE-LAB, for histograms, spatiograms, CCVs and correlograms, however the results presented in this section depict only the best result retrieved. In the performance evaluation of texture and face recognition descriptors, we have included the LBP descriptor as a potentially powerful candidate based on its proven success in the past [2]. We also followed the strategy proposed by Hadid *et al.* [7] in applying this descriptor to relatively low resolution images in such a manner that the images are represented by a concatenation of a global and a set of local LBP histograms. We empirically identified a suitable configuration for representing both body-patch and face images using "$LBP_{4,1}$ on 9 overlapping local histograms + $LBP^{u2}_{8,2}$ global histogram".



(a) Colour descriptors.

(b) Texture descriptors.

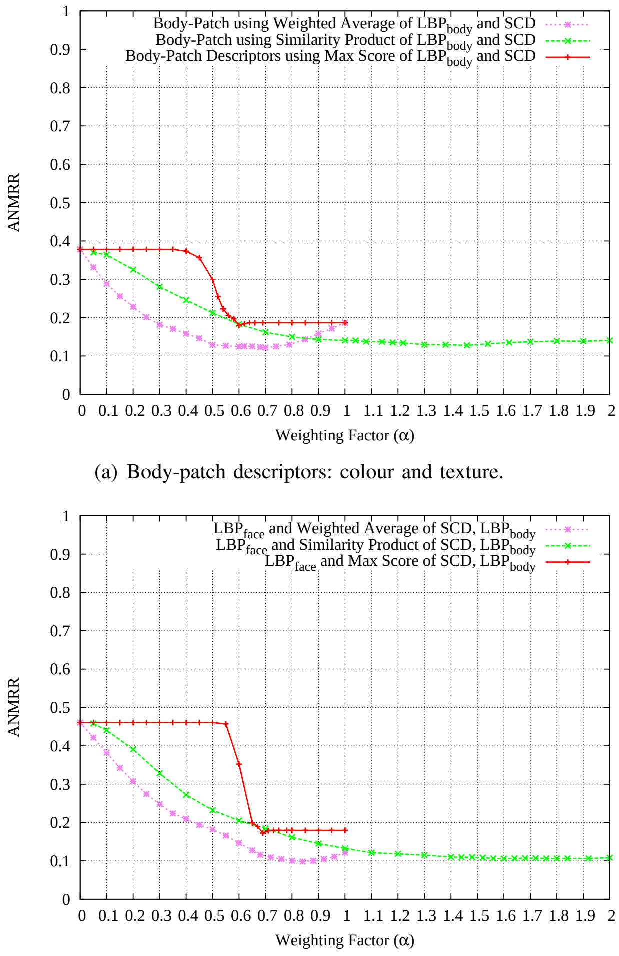
(c) Face recognition descriptors.

Figure 4: A performance comparison of colour, texture and face recognition descriptors.

Figure 4(a) shows the ANMRR performance figures of the 8 different colour descriptors studied in this paper. The results show that the MPEG-7 SCD with an ANMRR figure of 0.187 proves to be the best colour descriptor while spatiograms with 0.206 falling second in the ranked retrieval set. Surprisingly, simple histograms seem to have performed better than some of the histogram enhancement techniques, such as CCVs and correlograms. In fact, the MPEG-7 SCD can be considered as yet another simple histogram operated in the HSV colour space when adopted as a descriptor of maximum feature vector length, i.e. 256 coefficients, with zero bit-planes discarded. However, further experiments revealed that the non-linear quantisation step in the MPEG-7 SCD makes it a more robust descriptor than a conventional histogram descriptor. Another interesting outcome arising from our experiments is that spatiograms, in fact, perform better than any other histogram descriptor if a 256-bin histogram representation is used when creating them. However, due mainly to the fact that a spatiogram feature vector becomes much larger than that of any other descriptor, we limited the number of bins in spatiograms to 128 for this study. The results given in Figure 4(b) depict the ANMRR performance figures of the 4 texture descriptors studied in this paper. As the results show, the best-performing texture descriptor can be identified as the LBP descriptor, with MPEG-7 HTD, grey correlograms and MPEG-7 EHD following in that order. Similarly, Figure 4(c) shows a performance comparison of the face recognition descriptors, depicting that the LBP descriptor is the best out of the 4 descriptors studied in this paper.

4.2. Identifying a Suitable Fusion Scheme

In order to identify a suitable fusion method, we analyse the performance of three fusion schemes using the same test data set used in Section 4.1. Figure 5 shows the performance of the three fusion schemes, plotted as the graphs of ANMRR against weighting factor α. Results from fusion of colour and texture descriptors are depicted in Figure 5(a) whereas the results from fusion of colour, texture and face descriptors are shown in Figure 5(b). These plots show the performance variation of the three fusion schemes



(a) Body-patch descriptors: colour and texture.

(b) Body-patch and face descriptors.

Figure 5: A performance comparison of different fusion schemes.

against α, allowing us to trivially identify the best weighting factors in each fusion method. Concerning the behavior of the three fusion methods, we can clearly observe that the "weighted average" fusion scheme performs better than the "similarity score product" and "max score" methods. In summary, the ANMRR performance figures of the best-performing individual and combined descriptors are shown in Table 2. As can be seen in the results, fusion of body-patch colour and texture descriptors based on the "weighted average" fusion scheme result in improved person recognition performance with an ANMRR performance figure of 0.122 compared to their individual performance figures of 0.187 and 0.378 respectively. The results also show that, upon fusion of face and body-patch descriptors using the same fusion scheme, the overall performance of person recognition can be improved to 0.098. Table 2 gives a clear comparison of the ANMRR performance figures, illustrating that the "weighted average" fusion scheme is the best out of the three fusion schemes studied in this paper.

Table 2. Comparison of three fusion schemes.

Descriptor(s) Used	ANMRR			
	Individual	Weighted Average	Similarity Score Product	Max Score
Body-Patch Colour	0.187	-	-	-
Body-Patch Texture	0.378	-	-	-
Face	0.460	-	-	-
Body-Patch Colour and Texture	-	0.122	0.128	0.179
Body-Patch Colour, Texture and Face	-	0.098	0.106	0.172

4.3. Person Annotation: Testing the Performance of Content-Based Descriptors

In order to examine the performance of colour, texture and face recognition descriptors together with their combinations identified in Section 4.1 and 4.2, we present a performance evaluation of those descriptors on person annotation using a test data set comprising 7 different real-life personal photo collections. We use a 30% initial annotation of each user's collection as a reasonable choice to begin the annotation task and the nearest-neighbor classifier to infer the identity of all remaining persons in the collection throughout these experiments. Based on our experience with real users in the MediAssist system, we believe that 30% is typical of the amount annotation would expect to perform and would be willing to perform. The H-Hit rate figures are computed by comparing the classification result with the true label of the person. As the annotation process continues, knowledge from all the previous annotations is applied to recognise people for subsequent annotations.

Figure 6 shows the person annotation results in terms of hit-rate figures against H as a performance comparison of individual and combined content-based descriptors. The hit-rate figures are computed as the average figures of all 7 users in the test set. It can be observed that body-patch matching using colour results in higher person annotation performance as compared to using texture. However, combining colour with texture improves the performance of body-patch matching, in which case a hit-rate figure of 0.58 can be noted against 0.55 and 0.45 using colour and texture alone at H=1. We can also observe that body-patch feature matching results in higher person annotation performance compared to face recognition alone, where a hit-rate figure of 0.58 using combined colour and texture proves to be a significantly better result compared to a hit-rate figure of 0.47 from face recognition at H=1. Interestingly, using combined body-patch and face features leads to the best content-based descriptor configuration in this experimental system. A hit-rate figure of 0.62 at H=1 compared to a

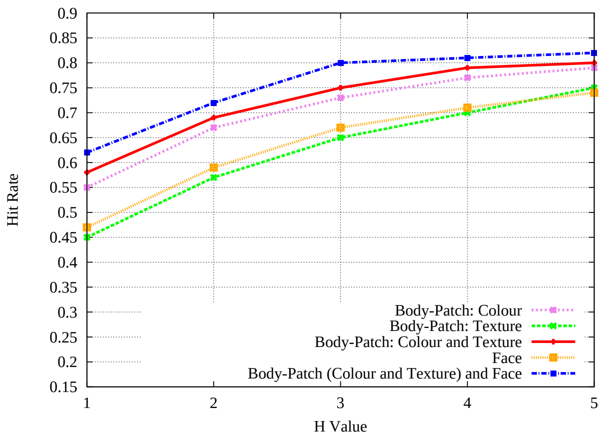


Figure 6: Person annotation results.

figure of 0.58 shows the relative advantage of employing face recognition in combination with body-patch feature matching in this challenging problem.

5. Conclusion

In this paper, we have investigated the use of content-based image descriptors towards enhancing the performance of person annotation in personal photo management applications. First, we carried out an empirical study to identify the best-performing colour, texture and face recognition descriptors for person recognition using a suitably chosen test data set. Second, we examined the performance of three different fusion methods to identify a suitable scheme for combining colour, texture and face recognition descriptors. Finally, we evaluated the performance of those descriptors using real-life personal photo collections as the test data. Our experiments prove that combining body-patch matching with face recognition significantly improves the performance of person annotation, in addition to the fact that texture is a complementary feature to colour-based body-patch matching. Based on these observations, we have shown that the identified content-based image descriptors are of great importance to enhancing the performance of person annotation in real-life scenarios. Future work will focus on studying other potential face recognition techniques and then combining with body-patch features to further improve the robustness of person recognition in personal digital photo archives. Additionally, we will also examine how the effectiveness of person annotation can be improved using automated person clustering techniques.

Acknowledgment

This research is supported by Science Foundation Ireland under grant 07/CE/I1147.

References

[1] ISO/IEC 15938-3. Information technology — Multimedia content description interface — Part 3: Visual. (ISO/IEC 15938-3), 2002.

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. on PAMI*, 28:2037–2041 (2006).

[3] S. T. Birchfield and S. Rangarajan. Spatiograms vs Histograms for Region-Based Tracking. In *CVPR'05*, pages 1153–1163 (2005).

[4] L. Chen, B. Hu, L. Zhang, M. Li, and H. Zhang. Face Annotation for Family Photo Album Management. *Intl. Journal of Image and Graphics*, 3:1–14 (2003).

[5] S. Cooray, N. O'Connor, and *et al.* Identifying Person Re-occurrences for Personal Photo Management Applications. In *VIE 2006*, pages 144–149 (2006), September.

[6] J. Cui, F. Wen, and *et al.* Easy Album: An Interactive Photo Annotation System Based on Face Clustering and Re-ranking. In *CHI'07*, pages 367–376 (2007), USA.

[7] A. Hadid, M. Pietikainen, and T. Ahonen. A Discriminative Feature Space for Detecting and Recognizing Faces. In *CVPR'04*, pages 797–804 (2004).

[8] J. Huang, S. R. Kumar, and *et al.* Image Indexing using Color Correlograms. In *CVPR'97*, pages 762–768 (1997).

[9] B. S. Manjunath, J.-R. Ohm, and V. V. Vasudeven. Color and Texture Descriptors. *IEEE Tran. on Circuits and Systems for Video Technology*, 11:703–715 (2001), June.

[10] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. In *JCDL'05*, pages 178–186 (2005), Denver, Colarado, USA.

[11] G. Pass and *et al.* Comparing Images using Color Coherent Vectors. In *ACM Multimedia*, pages 65–73 (1996).

[12] B. Suh and B. B. Bederson. Semi-Automatic Photo Annotation Strategies using Event based Clustering and Clothing based Person Recognition. *Interacting with Computers*, 19:524–544 (2007).

[13] M. J. Swain and D. H. Ballard. Color Indexing. *Intl. Journal of Computer Vision*, 7(2):11–32, 1991.

[14] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. In *ACM Conference on Multimedia*, pages 335–338 (2003), Berkeley, November.

[15] M. Zhao and *et al.* Automatic Person Annotation of Family Photo Album. In *CIVR'06*, pages 163–172 (2006), USA.