IDENTIFYING PERSON RE-OCCURRENCES FOR PERSONAL PHOTO MANAGEMENT APPLICATIONS

Saman Cooray[1], Noel E. O'Connor[1,2], Cathal Gurrin[1], Gareth J. F. Jones[1], Neil O'Hare[1], and Alan F. Smeaton[1,2]

[1] Centre for Digital Video Processing
[2] Adaptive Information Center
Dublin City University
IRELAND
Email: coorays@eeng.dcu.ie

Keywords: personal photo management, face detection, similarity matching, MPEG-7 colour descriptors.

Abstract

Automatic identification of "who" is present in individual digital images within a photo management system using only content-based analysis is an extremely difficult problem. We present a system which enables identification of person re-occurrences within a personal photo management application by combining image content-based analysis tools with context data from image capture. This combined system employs automatic face detection and body-patch matching techniques, which collectively facilitate identifying person re-occurrences within images grouped into events based on context data. We introduce a face detection approach combining a histogram-based skin detection model and a modified BDF face detection method to detect multiple frontal faces in colour images. Corresponding body patches are then automatically segmented relative to the size, location and orientation of the detected faces in the image. We investigate the suitability of using different colour descriptors, including MPEG-7 colour descriptors, Color Coherent Vectors (CCV) and Color Correlograms for effective body-patch matching. The system has been successfully integrated into the *MediAssist* platform, a prototype web-based system for personal photo management, and runs on over 13000 personal photos.

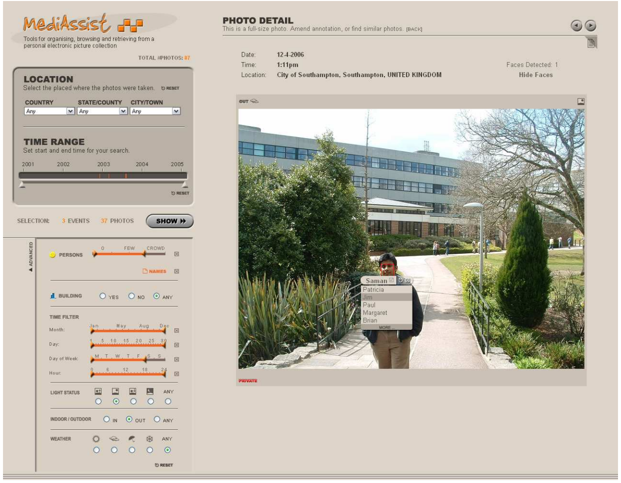
1 Introduction

Digital photo capture devices, such as cameras, webcams, mobile phones, PDAs, etc., have fueled the exponentially increasing number of digital photos the users have to deal with. However, current digital photo management technologies lack the required photo organizing capabilities, leaving a large number of users with poorly annotated photos. There is a strong need for effective digital photo management systems which can help users to easily organize their personal photo collection.

Many systems have been proposed to address the above problem using photos contexts, i.e. "when, where, what and who" in a given collection [1][2][3][4][5][6][7]. Utilizing only time and location based information (that can be easily extracted from the image EXIF data) is of limited use in photo management. However, technologies for effectively describing "what and who" in a photo collection are just beginning to emerge. Person annotation, in particular, has been a much discussed subject but it poses great challenges for content-based analysis tools. Consequently, most existing systems currently only support manual person annotation functionalities, a cumbersome and labour intensive task for the user.

A pioneering photo management prototype system, entitled FotoFile, was proposed by Kuchinsky *et al.* [1]. They employ automated content-based feature extraction tools to generate some annotation, including face detection and recognition technologies to establish a name for a person when he/she reappears. Systems proposed in [8][9] support direct annotation relying on the user's interaction to define the names for every user appearing/re-appearing in each image. Naaman *et al.* [2] proposed a semi-automated person annotation system which is based on only temporal, spatial and social contexts assuming such patterns can be learnt. Similar to our work, they also form clusters of events based on time and location information. In the semi-automated approach to face annotation proposed by Chen and Hu [3], they proposed combining face detection and recognition with CBIR and relevance feedback technologies. Like ours, their approach also involves a body-patch matching step based on colour and texture feature analysis where they use wavelet and autocorrelogram features. Their framework was then later adopted by Zhang *et al.* [4][5] with specific emphasis on the use of face recognition technologies. Suh and Bederson [6] proposed another semi-automatic image annotation approach targeting efficient bulk annotation also involving body-patch similarities to create meaningful clusters comprising similar persons in time-based events. Abdel-Motteleb and Chen proposed a system for browsing and organizing photos based on faces' arrangement and represented events with a set of composite images [7]. The event-representative image clusters are formed based on similarities of body-patches created in conjunction with face detection. Similar to the approach in [3], body-patch features are extracted using correlograms.

In this paper, we propose a semi-automated system that can facilitate describing "who" in a photo album by combining both content-based analysis tools and context data related to image capture. Key to this is to identify an effective



MediAssist

Tools for organising, browsing and retrieving from a personal electronic picture collection

TOTAL #PHOTOS: 87

LOCATION

Select the placed where the photos were taken. RESET

COUNTRY: Any | STATE/COUNTY: Any | CITY/TOWN: Any

TIME RANGE

Set start and end time for your search.

2001 2002 2003 2004 2005

RESET

SELECTION: 3 EVENTS 37 PHOTOS SHOW

PHOTO DETAIL

This is a full-size photo. Amend annotation, or find similar photos. [BACK]

Date: 12-4-2006

Time: 1:11pm

Location: City of Southampton, Southampton, UNITED KINGDOM

Faces Detected: 1

Hide Faces

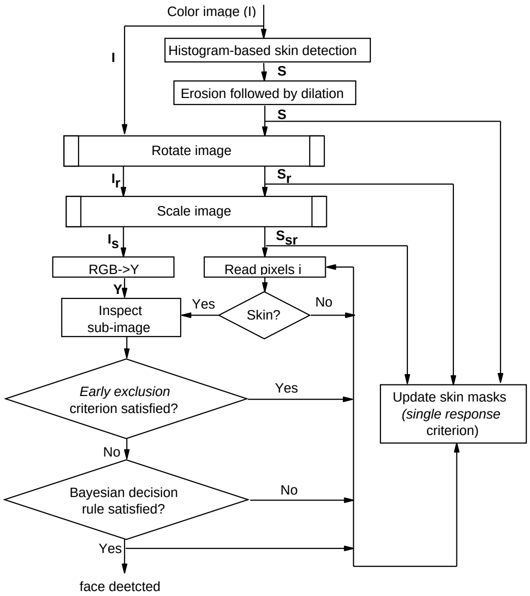
Figure 1: MediAssist person annotation in operation.

approach to human body-patch matching; an important functional element in the person annotation process. In particular, our research in this paper contributes a comparative performance analysis of body-patch matching techniques using MPEG-7 colour descriptors, colour coherent vectors and colour autocorrelograms. *The assumption made in the use of body-patch similarity features in photo management applications is that a person re-appearing within an event would be wearing the same clothing as in previous photos.*

2 The MediAssist System

MediAssist is a web-based personal photo management system built upon a novel paradigm of using content-based technologies in combination with context data related to image capture [10]. It currently consists of over 14,000 colour photographs taken by 22 users in 30 countries over a period of a few years. The MediAssist system first groups all the photos into meaningful events based on time and location information which are automatically extracted when uploading the photos to the system. Such events have proven to be very useful both in the search and indexing operations in the MediAssist system [10].

The person annotation system in MediAssist comprises automatic face detection, identification of person re-occurrences based on body-patch similarity matching, and a user-assisted annotation correction functionality. The automatic face detection approach described in section 3 is used to detect faces in the entire photo collection. We then automatically extract relevant body-patch features relative to the size, location and orientation of the detected faces. Based on our experimental results reported elsewhere, the MPEG-7 scalable colour descriptor is used for body-patch similarity matching in the current system. During the person annotation process, photos are provided to the user event by event as persons with similar clothes are more likely to occur within events than across different events. Based on the person identities initially provided by the user, the similarity match against a previously annotated set of photos is used to generate an automatic name suggestion when the user continues to annotate the photos. The user has 3 options to select, i.e. he/she can accept a suggested name, select a name from the suggested short-list or reject the suggested name in case of false occurrences. As the annotation process continues, more distinct identities are learnt and more representative body-patch features are accumulated, thereby improving the overall person annotation performance. Fig. 1 shows an example of the MediAssist person annotation system where the user is provided with a short-list of names to select from.



Color image (I)
I
Histogram-based skin detection
S
Erosion followed by dilation
S
Rotate image
I_r
S_r
Scale image
I_s
S_sr
RGB->Y
Y
Read pixels i
Inspect sub-image
Yes
Skin?
No
Early exclusion criterion satisfied?
Yes
Update skin masks (single response criterion)
No
Bayesian decision rule satisfied?
No
Yes
face deetcted

Figure 2: The system block diagram.

3 Face Detection

The block diagram of the face detection system employed is shown in Fig. 2. We extend the Bayesian Discriminating Feature (BDF) model which was originally proposed by Liu [11] for detecting frontal faces in grey-scale images. The modified algorithm effectively exploits the colour feature using a statistical skin detection model [12], whilst detection of variable sized faces with multiple rotations is facilitated by the algorithmic enhancement features described below.

Given a colour still image, the skin segmentation model first creates a skin mask which is subject to erosion/dilation morphological operations to remove noisy skin pixels while expanding the face cheek areas which are considered regions of interest in the face detection system. Both the original image and the corresponding skin mask are rotated in order to be able to detect rotated faces. The rotated image and the skin mask are then iteratively scaled by a pre-defined factor with minimum/maximum levels corresponding to the

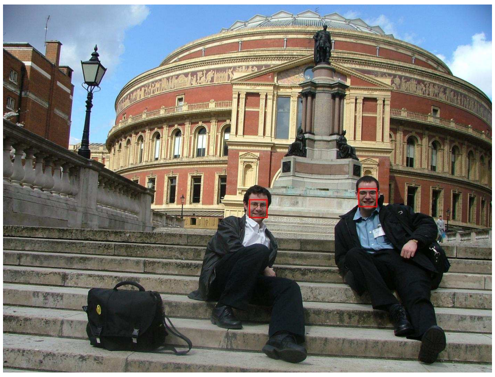


Figure 3: Face detection in a photo album scenario.

smallest/largest faces the algorithm is constrained to locate. After these processes, the skin image is read pixel by pixel and the presence of a skin pixel triggers inspection of a 16 × 16 image region for its likelihood to be a face. A modified early exclusion criterion is then applied to the sub-image. If successful, the Baysian decision rule, as defined by the original BDF model based on face/non-face error, classifies the current sub-image to be a face or not. However, the exact face location is decided by a modified single response criterion by searching for the best match sub-image within a pre-defined search space. Fig. 3 shows an example of automatic face detection results in a photo album scenario.

4 Body-Patch Matching for Person Re-occurrence Identification

Having detected faces, the corresponding body-patches are extracted relative to the location, size and orientation of the detected faces in the image as shown in Fig. 4. We consider body patches to be double the size width and height of automatically detected faces, i.e. a body corresponds to $2W \times 2W$ for a face $W \times W$. This is to ensure that inclusion of background regions are kept to a minimum. We consider that the body-patch is located d pixels below the lower level of the face (see Fig. 4). One objective of the experiments reported in this paper is to determine the best relationship between d and the face height W.

Fig. 5 shows the hierarchical approach that we introduce in this paper to perform within-event person annotation. The entire photo collection is first automatically categorized into an event hierarchy by the MediAssist system. The example shows that m number of events represented by E_1 to E_m have been defined by the system out of which the first 3 events have been already annotated by the user. So the current event, E_4, comprising n distinct persons is being annotated by the user who has already annotated k persons. Body-patch instances belonging to different persons have been stored by the system which is shown at the lowest level in the hierarchy, indicating that the persons have been annotated multiple times on re-occurrence. In this manner, when the user continues annotating photos within an event the system gathers more information about persons' appearance with their identities confirmed by the user. Thus, name suggestions for the current person P_{k+1} can be provided by comparing the current body-patch against all the available body-patch instances and taking the nearest match as the best suggestion.

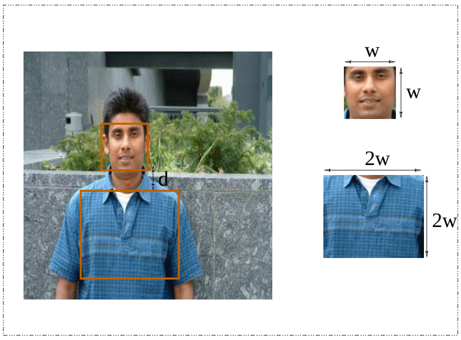


Figure 4: Body-patch extraction.

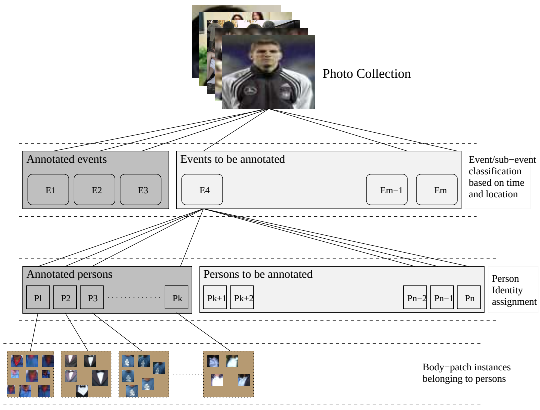


Figure 5: Within-event person annotation.

The following colour descriptors are considered in our experiments with the view to choosing the most suitable colour descriptor in the context of matching human body-patches. In addition to the four MPEG-7 descriptors, colour coherent vectors and HSV colour correlograms are considered for comparative analysis in these experiments due to their previous success in various applications.

4.1 Dominant Color Descriptor

The dominant colour descriptor (DCD) describes a fixed or arbitrary shaped image by using a number of dominant colours which can vary from image to image subject to an upper limit of 8 clusters [13][14]. The descriptor uses the generalized Lloyd algorithm in the *CIE LUV* colour space, in order to perform colour clustering by iteratively computing the distortion rate in each cluster until 8 clusters are built. An agglomerative clustering is then used to merge close clusters if the minimum distance between the two clusters is smaller than a pre-defined threshold, a process which can result in the minimum number of colour clusters present to be 1. The colour clusters are represented using dominant colour, percentage value and colour variance. The overall spatial homogeneity of dominant colours is represented by the spatial coherency measure [14].

4.2 Color Layout Descriptor

The colour layout descriptor (CLD) is a compact representation of the spatial distribution of a fixed or arbitrary shaped image which can be effectively used for CBIR applications [14]. The descriptor extraction process involves first partitioning the image into 8×8 blocks and representing them using the average colours in the YC_bC_r colour space. Applying the DCT transformation on 8×8 colour blocks produces a series of coefficients from which a few low frequency coefficients are selected and quantized. Quantizing the above low-frequency coefficients forms the colour layout descriptor. The similarity measure is defined by combining the coefficients of three colour channels.

4.3 Scalable Color Descriptor

The scalable colour descriptor (SCD) uses a Haar transformation on histograms in the HSV colour space [14]. Non-linearly quantized 256 histogram values are input to a Haar transformation, which produces a series of low-pass and high-pass coefficients with the number of coefficients being the same as that of the input. The output coefficients are then subject to a linear quantization process of different quantization levels with the H channel representing a higher percentage compared to the other two colour components. Our experiments are carried out using a 256-bin feature vector.

4.4 Color Structure Descriptor

The colour structure descriptor (CSD) embeds both the local spatial structure and the distribution of image colours in the $HMMD$ colour space [14][15]. CSD is defined using 4 different colour quantization levels, i.e. 32, 64, 128 and 256. The image is progressively scanned with a 8×8 structuring element while updating the number of occurrences of colour values encountered within the structural element. We use a 256-bin feature vector in our experiments.

4.5 Color Coherent Vector

The colour coherent vector (CCV) descriptor is a histogram enhancement technique which incorporates some spatial information of image colour distribution [16]. The spatial coherence measure is defined based on coherent/incoherent properties of image pixels in a discretized colour space where a coherent pixel is part of a sizeable region. Pixels not satisfying this property are considered incoherent. A CCV is represented by the number of coherent and incoherent pixels. The two CCVs can be compared by using a similarity measure, for example the L1 distance. We use the RGB colour space with 4 bins from each component, thereby resulting in a feature vector length 128.

4.6 HSV Autocorrelogram

Unlike histograms, the colour correlogram descriptor incorporates spatial information capturing both colour and texture characteristics in the image [17]. The pattern of change in the spatial correlation of the pair of colours (identical or distinct) with distance is captured by the correlogram descriptor. Autocorrelograms, which capture the spatial correlation between identical colours only, are a viable solution to CBIR compromising its computational complexity and performance as opposed to correlograms. Furthermore, the use of a more perceptual colour space such as HSV has been proven to be more effective in image matching [18]. In this context, we consider HSV autocorrelograms as a potential colour descriptor to body-patch matching in our experimental paradigm. We use 16 levels from Hue and 4 levels each from S and V components, and 4 distance values thereby resulting in a 256 feature vector length.

5 Performance Analysis: Body-Patch Matching

The performance evaluations reported in this paper are entirely based on retrieval accuracy, irrespective of the complexity of feature extraction, matching, etc. The two most common performance measure criteria used in CBIR systems are precision/recall and Average Normalized Mean Retrieval Rate (ANMRR). However, we used the ANMRR measure in this experiment due to the fact that different query images have varying number of result images to be compared against. This is a typical scenario in a person annotation process where different users occur different numbers of occasions.

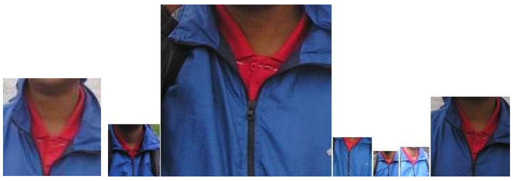
5.1 Test Data

Our test data comprises 50 subjects (i.e. 50 queries) with varying numbers of result images per query. The query and result-image data sets were established both manually and automatically. There are 4 data sets in total. The first 3 sets correspond to body-patches extracted by the system by varying d, i.e. $0.25W$, $0.5W$ and $0.75W$ (see Fig. 4) following automatic detection of faces. The fourth set corresponds to manually established query and result images through human inspection. Table 1 shows the corresponding number of result images for each query $q1 - q50$. In this experiment, the test data sets were established in such a way that they imposed real-life challenges on the descriptors used in this paper. Fig. 6 shows an example of 6 result images corresponding to the query

Table 1: Statistics on query and result images used in the experiments, with 50 query subjects and their 239 result images.

q 1	3	q 11	2	q 21	2	q 31	5	q 41	2
q 2	4	q 12	3	q 22	3	q 32	5	q 42	3
q 3	6	q 13	3	q 23	2	q 33	4	q 43	11
q 4	2	q 14	5	q 24	1	q 34	5	q 44	2
q 5	4	q 15	4	q 25	5	q 35	3	q 45	1
q 6	3	q 16	3	q 26	5	q 36	3	q 46	4
q 7	4	q 17	6	q 27	4	q 37	3	q 47	3
q 8	5	q 18	3	q 28	5	q 38	1	q 48	4
q 9	2	q 19	3	q 29	9	q 39	4	q 49	2
q 10	3	q 20	4	q 30	6	q 40	8	q 50	2

image, illustrating variations including spatial size, colour, deformations, inclusion of background regions, etc.



(a) query (b) result images

Figure 6: Example test data illustrating some discrepancies between the query and its result images.

5.2 ANMRR Performance Measure Criterion

ANMRR, a recently introduced and a vastly used performance evaluation measure in core MPEG-7 experimental evaluation systems [14], takes into account not only the number of retrieved images for the query but also the rank in retrieval results. This is defined as the average of Normalized Modified Retrieval Rank (NMRR) taken over a number of queries. NMRR is defined by:

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

where $NG(q)$ denotes the number of ground-truth images for the query q with rank $Rank(k)$. K is defined as $K = min(4 \cdot NG(q), 2 \cdot GMT)$, where GMT is the maximum value of ground-truth images for all queries. ANMRR always lies in the range $[0, 1]$, and the lower values indicate the better retrieval accuracy.

6 Results

We present the performance evaluation results for 6 colour descriptors: DCD, CLD, SCD, CSD, CCV and HSV autocorrelograms. Table 2 illustrates the performance of the descriptors in 4 categories. The first 3 columns indicating $d = 0.25W$, $d = 0.5W$, $d = 0.75W$ correspond to query and result images using different d values. The objective of evaluating 3 different data sets is to determine the optimum distance between the face and the body-patch region when extracting such body-patches automatically. The fourth column corresponds to the manually extracted query and result images where body-patches were more accurately established through human inspection, thereby keeping the possibilities of inclusion/exclusion of background/foreground regions to a minimal. This is considered as the ground truth in our experiments. Table 2 shows the average ANMRR figures for 50 queries when evaluated against 189 result images given in Table 1. As expected, ground truth data proved to be more accurate than any of the automatically established data. Based on this set of results, we can conclude that the scalable colour descriptor performs best with ANMRR 0.231, and the HSV autocorrelogram, CSD, CLD, CCV, DCD following that. Considering the 3 categories of automatic data sets, $d = 0.5W$ proves to be the best option when extracting body-patches relative to the face detection results. For SCD, its performance with ANMRR 0.240 is only 0.009 below its ground truth (ANMRR 0.231).

Table 2: Average ANMRR performance over 50 queries.

Color Descriptor	ANMRR			
	d=0.25*W	d=0.5*W	d=0.75*W	manual
Dominant Color Descriptor	0.355	0.336	0.358	0.32
Color Layout Descriptor	0.313	0.337	0.385	0.283
Scalable Color Descriptor	0.251	0.240	0.285	0.231
Color Structure Descriptor	0.274	0.270	0.296	0.268
Color Coherent Vector	0.371	0.367	0.375	0.319
HSV Autocorrelogrm	0.334	0.301	0.389	0.245

7 Conclusion

We have presented an approach for semi-automated person annotation comprising automatic face detection and body-patch similarity matching techniques within the framework of *MediAssist* photo management system. The key research area reported in the paper is on identifying a suitable body-patch matching technique using the MPEG-7 colour descriptors, colour coherent vectors and colour autocorrelograms. The above descriptors were each tested using a test data set which were subject to lighting variations, orientation variations, spatial size variations, partial occlusions, etc. The results illustrate that, of the descriptors studied, the scalable colour descriptor performs best in the body-patch matching process. We also conclude that $d = 0.5W$ is the optimum distance between the face and the body-patch region when automatically extracting such body-patches.

In order to carry out further investigations on robust body-patch matching techniques for person re-occurrence identification, we intend to compare the effectiveness of some texture descriptors in the future and to combine the colour descriptors in some way. We also intend to use colour segmentation for identifying more accurate body-patch regions as opposed to a fixed-shape region.

Acknowledgments

Part of this material is based on the MediAssist project funded by the Enterprise Ireland. The support by Science Foundation Ireland under Grant Nos. 03/IN.3/I361, Adaptive Information Cluster (AIC) and K-Space are also acknowledged.

References

[1] M. L. Creech D. Freeze B. Serra A. Kuchinsky, C. Pering and J. Gwizdka. Fotofile: A consumer multimedia organization and retrieval system. In *CHI'99*, pages 496–503, 15-20 May 1999.

[2] 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, Denver, Colarado, USA, 2005.

[3] L. Zhang M. Li L. Chen, B. Hu and H. Zhang. Face annotation for family photo album management. *International Journal of Image and Graphics*, 3:1–14, 2003.

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

[5] L. Zhang, M. Li, W. Ma, , and H. Zhang. Efficient propagation for face annotation in family albums. In *ACM Conference on Multimedia*, New York, October 2004.

[6] B. Suh and B. B. Bederson. Semi-automatic image annotation using event and torso identification. In *Technical Report 2004*, Computer Science Department, University of Maryland, College Park, MD, 2004.

[7] Abdel-Motteleb M. and Chen L. Content-based photo album management using faces' arrangement. In *IEEE Intl. Conf. on Multimedia and Expo (ICME)*, pages 2071–2074, 2004.

[8] H. Kang and B. Shneiderman. Visualisation methods for personal photo collections: Browsing and searching in the photofinder. In *IEEE Intl. Conf. on Multimedia and Expo (ICME2000)*, pages 1539–1542, 2000.

[9] Adobe photoshop album, adobe systems inc., www.adobe.com/products/photoshopalbum/.

[10] N. O'Hare, C. Gurrin, G. Jones, and A. Smeaton. Combination of content analysis and context features for digital photograph retrieval. In *EWIMT*, 2005.

[11] C. Liu. A bayesian discriminating features method for face detection. *IEEE Tran. on PAMI*, 25:725–740, June 2003.

[12] S. Cooray and N. O'Connor. A hybrid technique for face detection in color images. In *IEEE Conf. on Advanced Video Surveillance (AVSS'05)*, Italy, Sep. 15-16 2005.

[13] Y. Deng, B. S. Manjunath, C. Kenney, M. S. Moore, and H. Shin. An efficient color representation for image retrieval. *IEEE Tran. on Image Processing*, 10:140–147, January 2001.

[14] 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, June 2001.

[15] D. S. Messing, P. V. Beek, and J. H. Errico. The mpeg-7 colour structure descriptor: Image description using colour and local spatial information. In *IEEE Intl. Conf. on Image Processing (ICIP'01)*, pages 670–673, 2001.

[16] Pass G. and Zabih R. Histogram refinement for content-based image retrieval. In *IEEE Workshop on Applications of Computer Vision*, pages 96–102, 15-16 October 1996.

[17] J. Huang, S. R. Kumar, M.Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *CVPR'97*, pages 762–768, 1997.

[18] T. Ojala, M. Rautiainen, E. Matinmikko, and M. Aittola. Semantic image retrieval with hsv correlograms. In *12th Scandinavian Conf. on Image Analysis*, pages 621–627, Norway, 2001.