Context-Aware Person Identification in Personal Photo Collections

Neil O'Hare and Alan F. Smeaton, *Member, IEEE*

***Abstract*—Identifying the people in photos is an important need for users of photo management systems. We present MediAssist, one such system which facilitates browsing, searching and semi-automatic annotation of personal photos, using analysis of both image content and the context in which the photo is captured. This semi-automatic annotation includes annotation of the identity of people in photos. In this paper, we focus on such person annotation, and propose person identification techniques based on a combination of context and content. We propose language modelling and nearest neighbor approaches to context-based person identification, in addition to novel face color and image color content-based features (used alongside face recognition and body patch features). We conduct a comprehensive empirical study of these techniques using the real private photo collections of a number of users, and show that combining context- and content-based analysis improves performance over content or context alone.**

***Index Terms*—Context and content, person identification, personal photo management.**

I. Introduction

TECHNOLOGIES for efficiently managing and organizing digital photos assume more and more importance as users wish to efficiently browse and search through larger and larger photo collections. It is accepted that content-based image retrieval techniques alone have failed to bridge the so-called *semantic gap* between the visual content of an image and the semantic interpretation of that image by a person. Personal photos differ from other images in that they have an associated context, often having been captured by the user of the photo management system. Users will have personal recollection about the time, place and other context information relating to the environment of photo capture, and digital personal photos make a certain amount of contextual metadata available in their EXIF header[1], which stores the time of photo capture and camera settings such as lens aperture and exposure time. GPS location information is also supported by EXIF and, although not currently captured by most commercial cameras, there are ways of "location-stamping" photos using data from a separate GPS device [1], and camera phones are inherently location-aware. Systems for managing personal photo collections could thus make use of this contextual metadata in their analysis and organization of personal photos.

One of the more important user needs for the management of personal photo collections is the annotation of the identities of people [2]. In this paper, we present the MediAssist system for personal photo management, a context-aware photo management system that includes person annotation technologies as one of its major features. The system uses context- and content-based analysis to provide powerful tools for the management of personal photo collections, and facilitates semi-automatic person-annotation in personal photo collections, powered by automatic analysis. Traditional face recognition approaches do not generally cope well in an unconstrained photo capture environment, where variations in lighting, pose and orientation represent major challenges. It is possible to overcome some of these problems by exploiting the contextual information that comes with personal photo capture. We propose a number of approaches to person identification based on analysis of both the image content and the context of their capture. We conduct a large-scale evaluation of these proposed approaches, and show that by combining context- and content-based analysis we can improve performance compared to content or context alone.

The rest of this paper is organized as follows. In the next section, we discuss related work in person identification in personal photo collections. In Section III, we describe the main features of the MediAssist photo management system. In Section IV, we propose context- and content-based approaches to person identification, as employed by the MediAssist system. After introducing an evaluation methodology in Section V, we discuss evaluation results in Section VI and finish in Section VII with some conclusions and avenues for future work.

II. Related Work

Many researchers have focused on methods of identifying people in photo collections, having first detected faces using face detection techniques [3]. Traditional face recognition technologies [4] model the faces of a database of people in order to identify unknown faces. Personal photo collections represent challenging environments for face recognition techniques due to varying lighting conditions, facial expressions, pose, occlusion etc. In spite of this, a number of researchers have proposed the use of standard face recognition techniques for identifying people in personal photo collections [5], [6]. The Riya[2] on-

Manuscript received April 14, 2008; revised October 06, 2008. . The MediAssist project was supported by Enterprise Ireland under Grant CFTD-03-216 and by Science Foundation Ireland under Grants 03/IN.3/I361 and 07/CE/I1147. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

N. O'Hare is with the Centre for Digital Video Processing at Dublin City University, Glasnevin, Dublin 9, Ireland (e-mail: nohare@computing.dcu.ie).

A. F. Smeaton is with the Centre for Digital Video Processing, and also with CLARITY: The Centre for Sensor Web Technologie, Dublin City University, Dublin 9, Ireland (e-mail: alan.smeaton@computing.dcu.ie).

Color versions of one or more figures in this paper are available at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2008.2009679

[1]http://www.exif.org

[2]http://www.riya.com/

line photo management system also uses face recognition technology in the management of personal photos.

While personal photo collections are, on one level, an unconstrained environment for photo capture with a wide variety of capture conditions, subsets of an individual's photo collection can exhibit a large degree of uniformity in capture conditions. For example, in photos taken at the same event people will tend to wear the same clothes. Since personal photographs include capture time information it is possible to exploit this characteristic, and a number of researchers have proposed a "body patch" feature to exploit this regularity [7]–[10]. Such approaches are 'contextual' in that the body patch feature is only useful within a limited context, but they do not make *direct* use of context in the sense of context outside the image's content, in particular the spatial and temporal context of photo capture. A few existing systems make direct use of context in the person classification process, such as Naaman *et al.* [11], which proposes an approach that estimates the probability of a person occurring in a given photo based on previous annotations of other photos with a similar context to the given photo. This approach is combined with body patch and face recognition features by Zhao *et al.* [12], who use these features to cluster faces for automatic person annotation. Unlike the approach proposed in this paper, they [12] do not use spatial proximity for their context-based analysis and our work also differs in that we propose alternative approaches to context-based person identification, namely a language model probabilistic approach and a nearest neighbor approach. Gallagher and Chen [13] make use of a "group prior" (similar to co-occurrence, as proposed in [11]) in combination with face recognition to improve person identification. The Mobile Media Metadata system also makes direct use of both content and context for person classification [14]. This work uses standard face recognition tools combined with spatial and temporal contextual features, along with bluetooth co-presence information and photo sharing information about the people the user shared the photo with. The approach makes limited use of temporal information, using a weekend/weekday feature, and an hour of the day feature, for example, rather than the spatial and temporal proximity used in our work, and it limits content-based analysis to face recognition alone.

The contribution of this paper is a thorough examination of the use of context for person identification. For context-only person identification (i.e., without visual analysis) our contribution is twofold: firstly, we extend existing probabilistic approaches by using smoothing techniques to improve the statistical estimation; secondly, we introduce a nearest neighbor context-based approach as an alternative to the probabilistic approach. For person identification based on visual analysis, we introduce two new content-based features, face color and image color, which can be considered contextual in the sense that, like body patch, they assume an environment constrained by context (i.e., the same event). We conduct a large-scale evaluation of these content- and context-based approaches, used both in isolation and combined with each other, using the private personal photo collections of nine users, with an average of over 2000 photos and 50 distinct people per user collection.

III. MediAssist Context-Aware Photo Management System

The MediAssist Photo Management System is a context- and content-based photo managment system, and its main features are illustrated in Fig. 1 and outlined as follows.

- *Context-based Photo Analysis*. The system provides a number of context-based analysis tools. Photos are indexed using temporal information, such as time of the day, day of the week and month, facilitating queries such as all photos taken at the weekend during the Summer. Latitude/longidude coordinates are converted into placenames using the Geographic Names Information System[3] and the GEOnet Names Server[4]. Time and location information are used to automatically detect events, corresponding to 'bursts' of photo-capturing activity, using the approach proposed in [15]. Detecting such events allows us to efficiently summarize photo collections (indeed, this feature has recently been added to Apple's iPhoto[5]), and these events are also used to facilitate event-filtering for analysis techniques which assume photos are captured within the same event, as described later. Other context-based analysis classifies the light status at the time of photo capture as day/night/dusk/dawn using standard astronomical algorithms, determines the weather status at the time of photo capture similar to [16], and classifies photos as indoor/outdoor using EXIF metadata similar to [17]. More details of these tools can be found in [18].
- *Content-based Photo Analysis*. Content-based analysis tools provided by the system include building detection using edge orientation histogram-based features [19] and face detection using an extension of the Bayesian Discriminating Feature approach, as proposed in [20]. Person identification uses both content- and context-based features and will be described in Section IV.
- *Search, Browse and Annotation Interfaces*. MediAssist provides tools for browsing, searching and semi-automatic annotation of personal photograph collections. Searching, based on annotations created by the automatic analysis tools described above, can be carried out using a number of filters for structured searching or free-text search powered by a conventional text search engine.
- *Semi-Automatic Person-Annotation based on Content and Context*. Since automatic person identification is not yet 100% accurate, a practical alternative is *semi-automatic* person annotation, whereby the system can use automatic analysis to suggest names for unannotated faces, to be confirmed by the user [6], [7]. Some authors propose batch semi-automatic annotation to annotate multiple faces simultaneously [5], [21], [22]. MediAssist facilitates semi-automatic person-annotation by suggesting a list of candidate names for a given face, and batch annotation by suggesting, at appropriate moments as the user browses their collection, a set of faces to be annotated with a specific person name. This semi-automatic annotation is powered

[3]http://nhd.usgs.gov/gnis.html

[4]http://earth-info.nga.mil/gns/html/index.html

[5]http://www.apple.com/iphoto

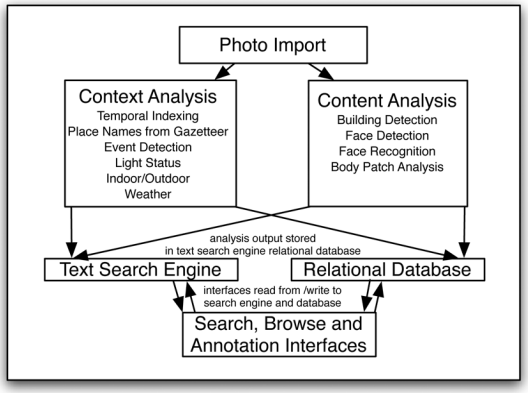

Fig. 1. Overview of the MediAssist photo management system.

by context- and content-based analysis, which is the main focus of this paper.

IV. Person Identification

In a partially annotated personal photo collection, automatic analysis of the context and content of annotated photos can suggest names for un-annotated ones. Since the system facilities two styles semi-automatic annotation, such analysis must support both of these:

- *Person Classification*. Given a face, the system can suggest a list of candidate names. We do not use the term "face recognition" because it is normally restricted to approaches based on analysis of the face.
- *Person Retrieval*. Given a specific name, the system can suggest a list of faces corresponding to that name, to facilitate batch annotation. We call this person retrieval because, as in general information retrieval, the task is to *retrieve* documents (faces) corresponding to an information need (a specific person).

In the remainder of this section we will introduce our proposed approaches to person identification. Firstly, we describe two novel context-based approaches: a probabilistic language model approach and a nearest neighbor approach. We will then describe the content-based approaches that we use: face recognition, body patch and two novel features for person identification, face color and image color. Finally, we describe how we combine context-based and content-based approaches to person identification.

A. Context-Based Person Identification

An existing context-based approach to person identification estimates the probability of a person occurring in a given photo based on the relative frequency of that person in previous annotations [11]. The simplest such estimator calculates the probability of a person given all annotations in the user's collection $(p|\mathrm{col})$, as follows: $P_{\mathrm{MLE}}(p|\mathrm{col}) = (C(p))/(N)$, where $C(p)$ is the number of annotated occurrences of person p and N is the number of annotations in the user's collection. In a similar manner, temporal proximity, spatial proximity, and co-occurrence estimators are proposed that estimate probabilities based on context. Using temporal proximity, for example, it is possible to calculate the probability of a person given all annotations within, say, one hour of the time of capture of the query photo. Co-occurrence is calculated based on how often given people co-occur in the same photo or, alternatively, within the same event. In this section we present two novel approaches to context-based person identification: a language modelling approach which extends the above idea, and a nearest neighbor approach.

1) Language Model Approach: A language model is a probability distribution that models the stochastic process behind the generation of a series of tokens in a language, such as words in text. The simplest language model, the unigram model, assumes complete independence between terms. In language modelling approaches to information retrieval, a separate language model is created for each document in the collection and, given an information need, the *query likelihood* for each document is calculated, which is the probability of the language model for that document creating the query [23]:

$$P(\mathbf{q} \mid M_d) = \prod_{i=0}^{|q|} P(q_i|M_d) \tag{1}$$

where $\mathbf{q}$ is the sequence of query terms and M_d is the language model of the document, and this *query likelihood* can then be used to rank documents. This simple model reduces the information retrieval task to the task of estimating the probabilities of the individual terms for each document.

It is also possible to view the creation of personal photos as being created by a stochastic process, determined by factors such as the people present at the time of photo capture, the location etc. For person identification we can view the vocabulary of the language model as all the people who can appear in the user's photo collection. We create a language model for every photo representing the probability of occurrence of each person in the user's collection and, since we are only interested in one person at a time, a retrieval query will only have one term, namely the person we are searching for. So we use *person likelihood*, the probability of a specific person, given the language model: $P(p|M_d)$, where p is the specific person. This model can easily be extended and used to support queries for photos containing many people. This language modelling approach calculates the probability of a specific person occurring for each photo, which can then be adapted for either classification or retrieval. For classification the document is fixed, and we rank all possible person names in order of decreasing person likelihood. For person retrieval the person is fixed, and we rank all documents according to the probability, $P(p|M_d)$, of creating that person given the language model for the photo.

We calculate maximum likelihood estimates (MLE) of the language model parameters using the estimators from [11] described above: user collection, temporal proximity, spatial proximity and co-occurrence. The language modelling approach extends that approach by giving a well-understood model for retrieval in addition to classification, and allowing us to make use of smoothing techniques developed by the language modelling community to improve our probability estimates. MLE estimates are problematic because they do not deal well with

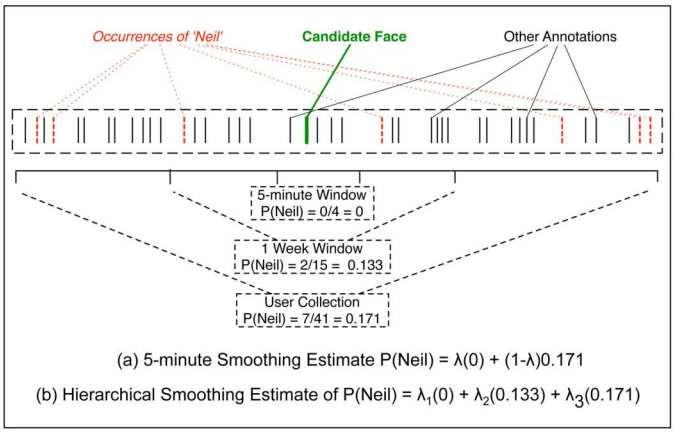

Fig. 2. Hierarchical smoothing for temporal proximity language models.

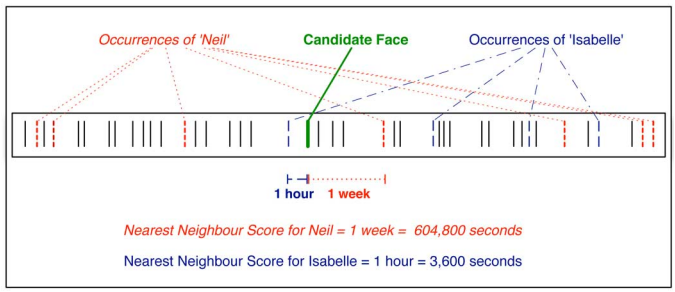

Fig. 3. Nearest neighbor classification using temporal proximity.

sparse or missing data, which will affect proximity-based estimators which use narrow windows. A five-minute temporal proximity window, for example, will often lead to a probability of zero because only a small number of photos of people will be present during such a short time span. Smoothing techniques address this problem by assigning a portion of the probability distribution to terms with zero frequency [24]. *Jelinek–Mercer* smoothing smoothes the probability distribution of the language model by interpolation with a background model which does not suffer from missing data, and this type of smoothing is illustrated in Fig. 2(a). A hierarchical smoothing technique in video retrieval can be used to model the hierarchical structure of shots and scenes of which a video is composed [25]. A similar hierarchical model can be applied to context-based photo language models with temporal proximity, for example, made of a structure composed of years, which are composed of months, which are composed of days, which are composed of hours. We propose 3-layer and 4-layer hierarchical smoothing language models to exploit this hierarchy, with a 3-layer *Jelinek-Mercer* hierarchical smoothing scheme for temporal proximity, spatial proximity and co-occurrence language models defined as follows:

$$\begin{aligned} &P(p|\text{photo}, \lambda_1, \lambda_2, \lambda_{\text{col}}) \\ &\quad = \lambda_1 \times P(p|\text{ctx}_1) + \lambda_2 \times P(p|\text{ctx}_2) \\ &\qquad + \lambda_{\text{col}} \times P(p|\text{col}) \end{aligned} \tag{2}$$

where ctx_1 is the most specific context in the hierarchy and ctx_2 is a wider context. The values for $\boldsymbol{\lambda}$ must sum to 1 and are typically learned empirically on a test collection. Fig. 2(b) shows an example of hierarchical smoothing for temporal proximity. For spatial proximity, we can create an accurate, but sparse, 1 km spatial proximity language model (based on all annotation within 1 km radius of the candidate) and then smooth this by coarser language models, such as 100 km spatial proximity, etc. We can create a 3-layer co-occurrence hierarchy by using the more accurate photo-based co-occurrence (where co-occurrence is calculated based on people appearing in the same photo) for the base estimator, and smooth this with event-based co-occurrence (where co-occurrence is calculated based on people appearing in the same event). We can expect photo-based co-occurrence to be more accurate but to be sparser, as it is a subset of event-based co-occurrence in the same way that five-minute temporal proximity is a subset of one hour or one week temporal proximity, for example.

2) Nearest neighbor Approach: An alternative nearest neighbor approach assigns a test point to the class of the closest labelled point [26]. In the semi-automatic annnotation scenario we are interested in ranking suggested person names, given a specific face, instead of assigning the face to a single class. We do this by ranking person names based on their distance from the query face. Since there are numerous annotated occurrences of each candidate name, the score assigned to a name, given a face, is the minimum of these. Let $\mathbf{b} = \{b_1 \ldots b_i\}$ be the set of all faces annotated as person p. The nearest neighbor score, $\text{NN}_{\text{score}}(a, p)$, for person p and face a, is:

$$\text{NN}_{\text{score}}(a, p) = \min_i [D(a, b_i)]. \tag{3}$$

$D(a, b_i)$ is a function which returns the distance between a and b_i. For nearest neighbor person classification, we will rank candidate names in increasing order of this score. For retrieval we rank suggested faces based on their minimum distance from a given person name. For temporal proximity, the distance is calculated as the temporal distance, in seconds, between the capture times of the photos containing the faces, as illustrated in Fig. 3. For spatial proximity, we calculate the distance in meters between two geographic coordinates. Since there is no natural nonprobabilistic distance measure of social proximity, we do not calculate a nearest neighbor equivalent of the co-occurrence language model.

B. Content-Based Person Identification

We propose two novel features for content-based person identification, *face color* and *image color*, in addition to using *face recognition* and *body patch* features, as illustrated in Fig. 4. Content-based person identification for each feature uses a nearest neighbor approach, with Manhattan distance [26] as the distance measure:

$$d_{\text{MAN}}(A, B) = \sum_i |a_i - b_i| \tag{4}$$

where A and B are two feature vectors. As we are primarily interested in the relative performance of different features for person identification, we do not explore alternative distance

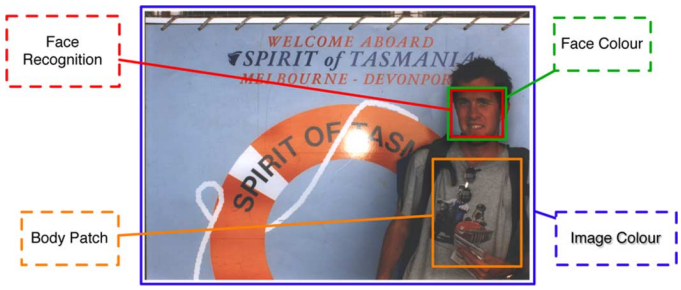

Fig. 4. Content-based features for person classification and recognition.

TABLE I
DETAILS OF EVALUATED USER COLLECTIONS

User	Total Photos	Known Faces	Photos With Faces (Known Faces)	Distinct People	Countries	States	Towns	Years Spanned
A	5,231	505	419 (341)	50	9	25	150	4
B	3,435	1,774	1,197 (1,007)	71	13	45	173	4
C	2,672	285	240 (194)	15	10	24	90	3
D	2,128	2,116	1,139 (1,080)	148	10	35	159	11
E	1,974	397	314 (258)	23	6	9	34	3
G	1,208	1554	893 (885)	45	3	8	40	4
H	1,044	705	435 (392)	40	4	12	40	2
J	753	992	488 (445)	62	5	10	22	1
L	513	519	289 (246)	45	5	7	20	1
Total	18,958	8,847	5,414 (4,848)	344	26	114	615	11
Avg	2106.4	983	601.6 (538.7)	55.4	7.2	19.4	80.9	3.67

measures in this work. For each of the color-based features, the MPEG-7 scalable color descriptor, with 256 bins, is used [27]. Some of the features make the assumption that they will be used in a highly constrained environment, such as within an event, in order to be useful for person identification. To take advantage of this assumption, we filter by event (as defined in Section III) when using content-based features, meaning that when calculating nearest neighbor distances we only compare candidate faces with annotated faces from that same event. This event-filtering can be contrasted with user-filtering, where only candidate faces within the same user's collection are considered, ignoring annotated faces from other users' collections. The content-based features used are now outlined.

1) Face Recognition: We position, scale and rotate each face to create a normalized face image, which is then used for extracting facial features for recognition. The eyes are located using principal component analysis (PCA) projections of candidate eye regions, and the independent component analysis (ICA) subspace method [28] extracts a 48-feature vector for each face.

2) Body Patch: We extract body patches corresponding to detected faces, located below the lower level of the face, as described in [8]. This feature can be used for person identification based on the observation that, within a given event, a person is unlikely to change their clothes. Not all faces have a corresponding body patch region: very large faces and faces towards the bottom of the photo will not have it.

3) Face color: In addition to using conventional face recognition techniques, we use a color-based feature to represent faces in personal photo collections. Our justification for this is that personal photo collections will often create highly constrained environments for photo capture, with similar lighting conditions and similar photo capture situations. This suggests that a simple, color-based face representation could be a salient feature, useful for person identification because faces of the same person within the same context should exhibit similar color characteristics, due to skin color or occluding objects like hair or sunglasses.

4) Image color: In personal photo collections, photos which contain the same people are often taken using the same camera setup, with the photographer often taking multiple near-duplicate photos in succession. In addition, photos taken at the same event are likely to share visual characteristics due to the fact that they were all taken in the same location under similar lighting conditions. This means that photos which contain the same people will sometimes contain the same image-level visual characteristics. Accordingly, we use an *image color* feature for person classification and retrieval, which represents each face using the global, color-based characteristics of the image containing the face.

C. Combined Approaches to Person Identification

In order to combine the various approaches to person identification presented above, we use the standard *CombSum* approach, which by summing the normalized scores from each different approach. We will also use *Weighted CombSum*, which multiplies the normalized score from each approach by a weight before summing. These standard linear interpolation based approaches have been shown to perform well in diverse fusion scenarios [29]–[31]. We normalise the scores from each individual approach as follows [32]:

$$\mathrm{nscore}_p = \frac{\mathrm{score}_p - \mathrm{score}_{\min}}{\mathrm{score}_{\max} - \mathrm{score}_{\min}} \tag{5}$$

where score_p refers to the score for person p in the classification scenario (or face p in the retrieval scenario). We will combine different context-based approaches, different content-based approaches, and finally we will combine content and context. All of these will use CombSum and Weighted CombSum, with the exception of the combined language model approach, which uses linear interpolation to combine temporal, spatial and co-occurence hierarchcial smoothing approaches to create a combined language model that estimates its parameters based on time, location and co-occurrence.

V. TEST COLLECTION AND EVALUATION METHODOLOGY

The MediAssist personal photo archive contains 23 774 geotagged photos from 29 users, taken as part of their private personal photo collections. Of these, nine users have collections suitable for evaluation of person identification, with the other user collections not containing enough known people. Table I summarizes the nine individual personal photo collections used. As the focus of this evaluation is on person identification we do not want the effect of imperfect face detection to add noise to our results. Accordingly, the presence of all faces in the collection was manually annotated to give a collection of faces to be evaluated, and the names of all of these faces were manually annotated to give a ground truth for evaluation. This manual face annotation also means that nonfrontal or occluded faces that would often by missed by automatic detection techniques are included for the evaluation of person identification, giving us a a more challenging test collection.

For evaluation purposes each collection was split into a training set and a test set of equal size, modelling the situation where the user has 50% of their collection already annotated and the system is using this information to learn suggested annotations for the remaining 50% of the collection. In some of their work Naaman *et al.* [11] show that, for context-based person identification, performance improved sharply as the training set increased from 0 to 10%, but the improvement was slow and gradual as the training set was expanded from 10% to 50%. The work also showed that results using 20% training set are consistent with those using 40% training set, so we can expect that our results from using a 50% training set will be representative of what we would find with a smaller training set. Although this high ratio of prior annotation may not always be realistic in a real world scenario, it is very useful for evaluation purposes; with a smaller training set, while the proposed techniques would still work, the lack of training data would likely minimise the differences between system variants. Given this, we prefer a richer training set that allows the system to learn properly from all available features.

We use a simple approach to modelling the user annotation process, essentially assuming that the user annotates all identities in their collections in a random order. Since a user is generally only interested in annotating the most popular identities in their collection, we use the top 20-most popular people in each collection for evaluation, assuming that the user is not interested is less popular faces.

Although we propose person identification approaches that can be applied in classification or retrieval, we only present evaluation results for the classification scenario here, though elsewhere we have found that person retrieval gives broadly similar results [33]. We use the *H-hit rate* evaluation measure for semi-automatic annotation, proposed by [34], which takes a list of H suggested names for a given face, and if the correct name is present in this list then this is considered a "hit." The *H-hit rate* is the proportion of H-hits within the collection:

$$H\text{-hit} = 1/N \sum_{f \in F} \text{hit}_H(f) \tag{6}$$

where F is the set of known faces to be evaluated and N is the number of faces evaluated. $\text{Hit}_H(f)$ is 1 if f is present in the list of H suggested names, and 0 otherwise.

The weights for all language model smoothing approaches and Weighted CombSum combined approaches are learned on the test set by optimising the evaluation criterion, *5-hit rate*: we use a brute force approach to evaluate all possible values for the weights and use the values that give the best performance for *5-hit rate*. The weights are learned separately for each user collection and are biased, "oracle," weights and cannot be said to represent weights that we could expect a system to learn automatically. By learning these oracle weights, however, we can discover which smoothing schemes, and which combinations of features, are the most powerful for person identification, giving a useful upper bound on the performance of weighted approaches. This is useful for evaluation purposes because we know that there is no noise in terms of inappropriate weighting distorting the difference in performance between different features; it is also very useful for comparing weighted and unweighted approaches, as we are comparing the unweighted approaches to the best possible parameters for weighted combination.

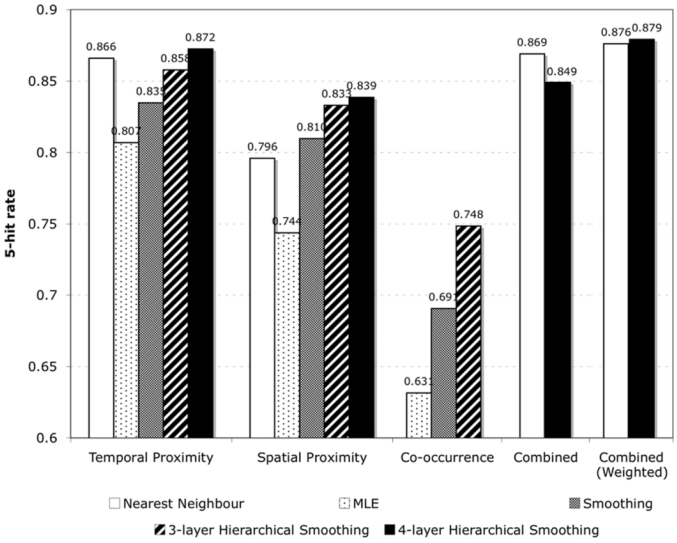

Fig. 5. Results for Context-based person classification.

We use hold-out cross-validation [35] to create a number of independent partitions between training and test set. The results we present are averaged over five such partitions. We calculate the results for each user, and calculate an average of these, ensuring that users with a larger number of test faces do not bias the results. We use randomization testing [36] to test if differences between approaches are statistically significant. The input into the tests are the *H-hit rate* results for each user, so we are testing if the differences between runs are consistent across users, or if they are due to chance variations between users. If one approach is better than another at a significance level of 0.05 we will say it is *significant*, and if it is better at a significance level of 0.01 we will say it is *highly significant*.

VI. Evaluation Results

In this section we will present our person identification evaluation results. Firstly, we will present the results of context-based personal classification, followed by content-based classification results. Finally, we present combined context- and content-based results.

A. Evaluation of Context-Based Person Classification

Fig. 5 summarizes the *5-hit rate* results for context-based person classification approaches. For MLE, Smoothing and Hierarchical Smoothing language models a large number of alternative variations were evaluated, with varying window sizes and hierarchical structures explored, and the best-performing variation for each is shown here. For spatial proximity, temporal proximity and co-occurrence, smoothing always improves performance over MLE, and this is highly significant in all cases. Also, the performance improves as we move towards deeper hierarchies, with the 4-layer hierarchical smoothing approach significantly outperforming shallower hierarchies.

The simple nearest neighbor approaches perform very well, with temporal proximity nearest neighbor outperforming all

temporal proximity smoothing approaches except 4-layer hierarchical smoothing, and it is not significantly outperformed by that. It is not necessary to learn optimal parameter weights for the nearest neighbor approach, which gives it an advantage over smoothing approaches. Also, it should not need as many annotations to achieve reasonable performance because it does not use annotation frequencies to estimate probabilities. Spatial proximity nearest neighbor does not perform quite so well, and we believe this is because many photos will have identical or near-identical locations (on the other hand every image has a unique time-stamp), meaning that ranking by distance from a given location will not always give useful information.

Temporal proximity is the best performing context-based modality, highly significantly outperforming spatial proximity for both language model and nearest neighbor approaches. Combining temporal proximity information with other context-based approaches gives only a small improvement in performance, even if we learn optimum weights, suggesting that the other context-based features tend to be somewhat redundant in the presence of temporal proximity information. Although Zhao *et al.* [12] do not include full details of their context-based approaches and so a direct comparison is not possible, we can consider their approach to be broadly similar to our temporal proximity 3 layer hierarchical smoothing: it is clear from these results that this is outperformed by temporal proximity 4 layer hierarchical smoothing and nearest neighbor proximity hierarchical smoothing, and by our combined approaches that make additional use of location information.

B. Evaluation of Content-Based Person Classification

Fig. 6 shows the *5-hit rate* results for content-based person classification. For user-filtering, the global image color feature significantly outperforms all of the region-level features for this coarse evaluation measure. This is surprising, but we believe the reason for this is that user-filtered image color classification will often rank photos from the same event highly because similar lighting conditions and similar locations mean that photos within the same event should be most similar to the query image, naturally performing event-filtering that region-based approaches do not achieve. Event-filtering always improves performance over user-filtering, a difference that is statistically significant for face recognition and face color. Face color performs surprisingly well, showing that within the constrained environment of personal photo collections it can be a useful feature for identity classification due to its ability to model skin tone and other color-based variations between identities (e.g., color of hair occluding face, sunglasses).

Fig. 7 shows the *H-hit rate* for event-filtered content-based approaches and combined content-based approaches for various values for h. Event-filtered body patch is highly significantly better than all other approaches for *1-hit rate* and significantly better than all other approaches for *2-hit rate*, showing that the simple, color-based, body patch feature is more powerful than other region-based features for identity classification in highly constrained environments. The fact that not all faces have a corresponding body patch region, in addition to the occasional occlusion of the torso and the fact that sometimes multiple people wear very similar clothes at the same event, are factors which inhibit the performance of *body patch* when evaluating with a larger *H-hit rate*.

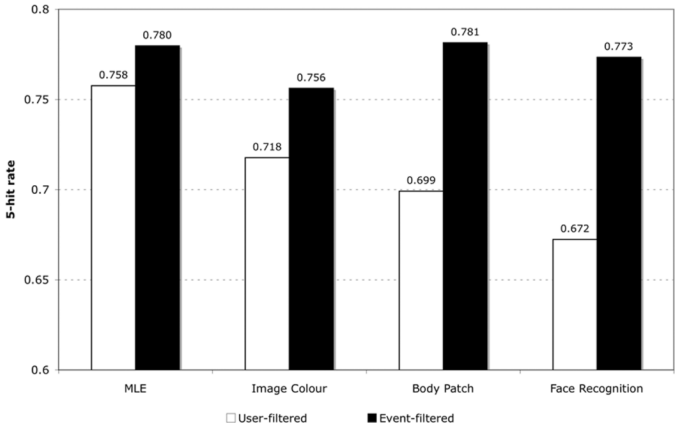

Fig. 6. User-filtered versus event-filtered content-based person classification.

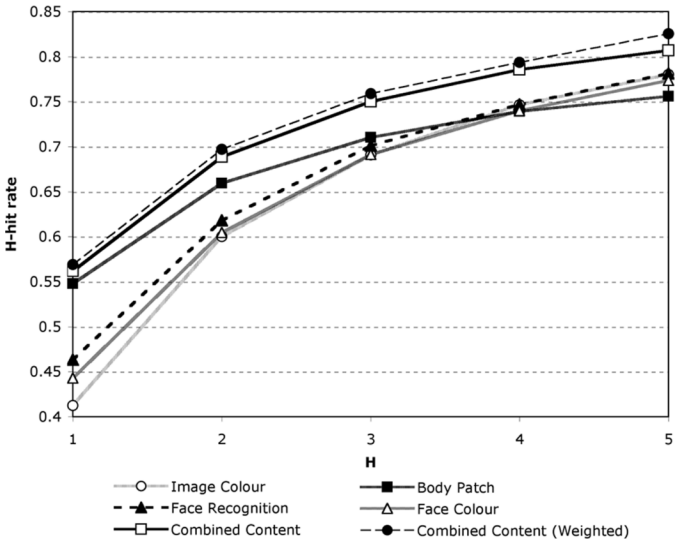

Fig. 7. Results for content-based person classification.

Weighted and unweighted combined content-based approaches give a highly significant improvement over all individual approaches for *3-hit rate*, *4-hit rate* and *5-hit rate*, and a significant improvement over all individual approaches except body patch for *1-hit rate* and *2-hit rate*. Although both improve over body patch for *1-hit rate* or *2-hit rate*, this is not statistically significant. There is no significant difference between weighted and unweighted combination for *1-hit rate*, *2-hit rate* and *3-hit rate*, suggesting that there is little benefit to be gained from weighting when combining these content-based approaches to person classification.

C. Evaluation of Context- and Content-Based Person Classification

Fig. 8 shows combined content-only, combined context-only and combined context and content event-filtered classification results, evaluated using *H-hit rate* for different values of H, with both the best weighted and the best unweighted combinations shown. For weighted combination we use combined hierarchical smoothing as the context-based approach, while we

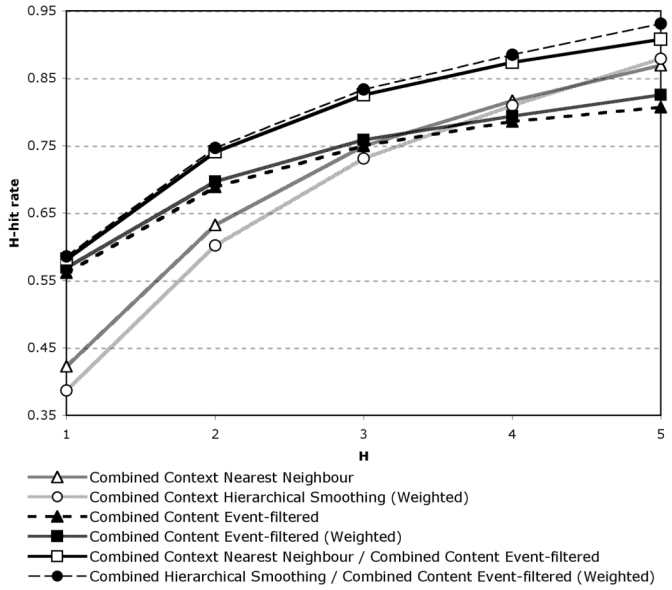

Fig. 8. Combined content- and context-based person classification.

use combined nearest neighbor approaches for evaluation of unweighted combination. Comparing content-based approaches with context-based approaches, we can see that content-based approaches outperform context-based approaches for *1-hit rate* and *2-hit rate*, with weighted and unweighted content-based approaches both outperforming context-based approaches for these evaluation measures. For *4-hit rate* and *5-hit rate*, however, context-based approaches perform better, with the difference statistically significant for *5-hit rate*. The better *4-hit rate* and *5-hit rate* of the context-based approaches reflects the fact the event-filtered content-based approaches sometimes suffer when there are no annotations of the correct person name in the current event.

Combined context-based and content-based approaches outperform content-only and context-only approaches for all values of H. The weighted combined hierarchical smoothing language model combined with content highly significantly outperforms all content-only and context-only approaches for *2-hit rate*, *3-hit rate*, *4-hit rate* and *5-hit rate*. The unweighted combination of context and content performs significantly better than unweighted content-only and context-only approaches for all values of H, and although it is outperformed by the weighted language model approach, this difference is quite small and is only significant for *4-hit rate* and *5-hit rate*, encouragingly suggesting that, for the scenario presented in this paper and using the features we have extracted, this simple, unweighted, approach, can achieve equivalent performance to an approach tuned with optimum weights.

Our results show that it is possible to improve the results from content-based approaches to person identification with the additional use of context. While acknowledging that the face recognition technique used here is relatively simple and that more sophisticated approaches would yield improved performance, there will always be cases where content-based approaches fail, for example due to occlusion, nonfrontal faces etc, and we can expect contextual information to be useful in these cases. Also, for technologies that can detect people where the face is not necessarily visible [37], additional contextual cues can aid identification.

VII. Conclusion

We have proposed context- and content-based approaches to person identification in partially annotated personal photo collections, and have conducted a large scale evaluation of these approaches for the person classification scenario, where the task is to suggest a list of names given an un-annotated face. The proposed context-based smoothing and nearest neighbor approaches perform well, outperforming the MLE approach. The best-performing context-based modality is temporal proximity, and spatial proximity and co-occurrence tend to be redundant in combination with this feature. The proposed content-based features, image color and face color, have been shown to be useful for person identification and, while body patch is the best performing feature in isolation, it is possible to achieve a significant improvement by using a simple unweighted combination with other content-based features. Most importantly, we have shown that it is possible to improve performance by combining context and content, and it is not necessary to learn weights to achieve this. Although not reported here, person retrieval yields similar results.

Possible directions for future work in this area include applying the techniques to automatic annotation of people, without prompting the user for confirmation, reserving such automatic annotation for those cases with the highest confidence. Other batch annotation approaches automatically create clusters of similar faces, which can be annotated in a batch manner: the context- and content-based features presented in this paper could improve the quality of the clusters for such approaches. In large online photo-sharing communities such as Flickr[6], friends' annotations, if known to be from the same event, can be used to support the techniques presented here when none of the user's own photos from that event have been annotated. Other sources of context for person identification, such as bluetooth co-presence, have been proposed elsewhere (e.g., [14]), and could be integrated with our approach.

We believe our results show that there is a place for context-based person identification techniques alongside traditional techniques such as face recognition, even as these more traditional techniques continue to improve. The presence of contextual information can reinforce content-based information, and will always be able to provide good annotation suggestions in situations where these content-based approaches fail.

Acknowledgment

The work of G. Jones, C. Gurrin, S. Cooray, B. Uscilowski and H. Lee on the MediAssist project is gratefully acknowledged, as is the insightful advice from Prof. S. Boll and the comments and suggestions of the anonymous reviewers.

References

[1] K. Toyama, R. Logan, and A. Roseway, "Geographic location tags on digital images," in *MULTIMEDIA'03: Proc. Eleventh ACM Int. Conf. on Multimedia*, New York, Nov. 2003, pp. 156–166.

[6]http://www.flickr.com

[2] K. Rodden and K. R. Wood, "How do people manage their digital photographs?," in *CHI'03: Proc. SIGCHI Conf. on Human Factors in Computing Systems*, Fort Lauderdale, FL, 2003, pp. 409–416.
[3] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, 2002.
[4] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.
[5] A. Girgensohn, J. Adcock, and L. Wilcox, "Leveraging face recognition technology to find and organize photos," in *MIR'04: Proc. 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, New York, Oct. 2004, pp. 99–106.
[6] A. Kuchinsky, C. Pering, M. L. Creech, D. Freeze, B. Serra, and J. Gwizdka, "FotoFile: A consumer multimedia organization and retrieval system," in *CHI'99: Proc. SIGCHI Conf. on Human Factors in Computing Systems*, Pittsburgh, PA, May 2001, pp. 496–503.
[7] L. Zhang, L. Chen, M. Li, and H. Zhang, "Automated annotation of human faces in family albums," in *MULTIMEDIA'03: Proc. Eleventh ACM Int. Conf. on Multimedia*, Berkeley, CA, 2003, pp. 355–358.
[8] S. Cooray, N. O'Connor, C. Gurrin, G. Jones, N. O'Hare, and A. F. Smeaton, "Identifying person re-occurrences for personal photo management applications," in *Int. Conf. on Visual Information Engineering*, Bangalore, India, Sep. 2006, pp. 144–149.
[9] J. Sivic, C. Zitnick, and R. Szeliski, "Finding people in repeated shots of the same scene," in *Proc. British Machine Vision Conf.*, Edinburgh, U.K., 2006, pp. 909–918.
[10] D. Anguelov, K. Lee, S. B. Gokturk, and B. Sumengen, "Contextual identity recognition in personal photo albums," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, Jun. 2007.
[11] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke, "Leveraging context to resolve identity in photo albums," in *JCDL'05: Proc. 5th ACM/IEEE-CS Joint Conf. on Digital Libraries*, Denver, CO, Jun. 2005, pp. 178–187.
[12] M. Zhao, Y. Teo, M. Zhao, M. Zhao, Y. Teo, S. Liu, T.-S. Chua, and R. Jain, "Automatic person annotation of family photo album," in *5th Int. Conf. on Image and Video Retrieval (CTVR 2006)*, Tempe, AZ, 2006, pp. 163–172.
[13] A. C. Gallagher and T. Chen, "Using group prior to identify people in consumer images," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, Jun. 2007.
[14] M. Davis, M. Smith, F. Stentiford, A. Bambidele, J. Canny, N. Good, S. King, and R. Janakiraman, "Using context and similarity for face and location identification," in *Proc. IS&T/SPIE 18th Annu. Symp. Electronic Imaging Science and Technology*, San Jose, CA, 2006.
[15] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd, "Time as essence for photo browsing through personal digital libraries," in *JCDL'02: Proc. 2nd ACM/IEEE-CS Joint Conf. on Digital Libraries*, Portland, OR, Jul. 2002, pp. 326–335.
[16] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke, "Context data in geo-referenced digital photo collections," in *MULTIMEDIA'04: Proc. 12th Annu. ACM Int. Conf. on Multimedia*, New York, Oct. 2004, pp. 196–203.
[17] M. R. Boutell and J. Luo, "Beyond pixels: Exploiting camera metadata for photo classification," *Pattern Recognit.*, vol. 38, no. 6, pp. 935–946, 2005.
[18] N. O'Hare, H. Lee, S. Cooray, C. Gurrin, G. Jones, J. Malobabic, N. O'Connor, A. F. Smeaton, and B. Uscilowski, "MediAssist: Using content-based analysis and context to manage personal photo collections," in *CTVR2006—5th Int. Conf. on Image and Video Retrieval*, Tempe, AZ, 2006, pp. 529–532.
[19] J. Malobabic, H. LeBorgne, N. Murphy, and N. E. O'Connor, "Detecting the presence of large buildings in natural images," in *Proc. CBMI2005—4th International Workshop on Content-Based Multimedia Indexing*, Riga, Latvia, Jun. 2005.
[20] S. Cooray and N. O'Connor, "A hybrid technique for face detection in color images," in *AVSS—Int. Conf. on Advanced Video and Signal Based Surveillance*, Como, Italy, 2005, pp. 253–258.
[21] B. Suh and B. B. Bederson, "Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition," *Interacting with Computers*, vol. 19, no. 4, pp. 524–544, 2007.
[22] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang, "A face annotation framework with partial clustering and interactive labeling," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, 2007, IEEE Computer Society.
[23] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. 21st Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'98)*, Melbourne, Australia, 1998, pp. 275–281.
[24] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press, 1999.
[25] T. Westerveld, A. de Vries, and A. van Ballegooij, "CWI at the TREC-2002 video track," in *The Eleventh Text REtrieval Conf. (TREC-2002)*, Gaithersburg, MD, 2003, pp. 207–216.
[26] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2000.
[27] 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*, vol. 11, no. 6, pp. 703–715, 2001.
[28] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Netw.*, vol. 13, no. 6, pp. 1450–1464, Nov. 2002.
[29] J. H. Lee, "Analyses of multiple evidence combination," in *Proc. 20th Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'97)*, Philadelphia, PA, 1997, pp. 267–276.
[30] W. B. Croft, *Combining Approaches to Information Retrieval*. Berlin, Germany: Springer, 2002, pp. 1–36.
[31] K. Mc Donald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *CIVR 2005—Int. Conf. on Image and Video Retrieval*, Singapore, 2005, pp. 61–70.
[32] E. A. Fox and J. A. Shaw, "Combination of multiple searches," in *Proc. Third Text REtreival Conf. (TREC-1994)*, Gaithersburg, MD, 1994, pp. 243–252.
[33] N. O'Hare, "Semi-Automatic Person-Annotation in Context-Aware Personal Photo Collections," Ph.D. Dissertation, Sch. Computing, Dublin City University, Dublin, Ireland, 2007.
[34] L. Chen, B.-G. Hu, L. Zhang, M. Li, and H. Zhang, "Face annotation for family photo album management," *International Journal of Image and Graphics* vol. 3, no. 1, pp. 81–94, 2003 [Online]. Available: http://ejournals.wspc.com.sg/ijig/03/0301/S0219467803000920.html, [Online]. Available
[35] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London, U.K.: Prentice-Hall, 1982.
[36] O. Kempthorne and T. E. Doerfler, "The behaviour of some significance tests under experimental randomization," *Biometrika*, vol. 56, no. 2, pp. 231–248, 1969.
[37] N. Dalal and B. Triggs, S. S. Cordelia Schmid and C. Tomasi, Eds., "Histograms of oriented gradients for human detection," in *CVPR'05: Proc. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 2005, vol. 2, pp. 886–893.

Neil O'Hare is a Postdoctoral Researcher in the Centre for Digital Video Processing in Dublin City University, Dublin, Ireland. His research interests are in the area of multimedia information management, in particular the management of personal photos using a combination of context- and content-based analysis.

Alan Smeaton (M'05) is a Professor of Computing at Dublin City University, Dublin, Ireland, where he is the Director of the Centre for Digital Video Processing and Deputy Director of the CLARITY CSET. His research interests cover information management of all kinds, including managing video archives and managing information gathered from sensor networks.