ACRONYM:
Context Metrics for Linking People to User-Generated Media Content

Fergal Monaghan, SAP Research, UK

Siegfried Handschuh, National University of Ireland, Galway, Ireland

David O'Sullivan, National University of Ireland, Galway, Ireland

ABSTRACT

With the advent of online social networks and User-Generated Content (UGC), the social Web is experiencing an explosion of audio-visual data. However, the usefulness of the collected data is in doubt, given that the means of retrieval are limited by the semantic gap between them and people's perceived understanding of the memories they represent. Whereas machines interpret UGC media as series of binary audio-visual data, humans perceive the context under which the content is captured and the people, places, and events represented. The Annotation CReatiON for Your Media (ACRONYM) framework addresses the semantic gap by supporting the creation of a layer of explicit machine-interpretable meaning describing UGC context. This paper presents an overview of a use case of ACRONYM for semantic annotation of personal photographs. The authors define a set of recommendation algorithms employed by ACRONYM to support the annotation of generic UGC multimedia. This paper introduces the context metrics and combination methods that form the recommendation algorithms used by ACRONYM to determine the people represented in multimedia resources. For the photograph annotation use case, these result in an increase in recommendation accuracy. Context-based algorithms provide a cheap and robust means of UGC media annotation that is compatible with and complimentary to content-recognition techniques.

Keywords: *Context Mining, Linked Data, Mobile Devices, Recommender Algorithms, Semantic Annotation, Social Media*

1. INTRODUCTION

"User-Created Content" (UCC) or "User-Generated Content" (UGC) (Wunsch-Vincent & Vickery, 2006) is paving the way towards a Web of "object-centred sociality" (Cetina, 1997; Bojārs, Heitmann, & Oren, 2007): a collaborative knowledge management platform built around documents or other objects of interest that goes beyond unidirectional publication and consumption. As well as user profiles, blogs, and other information manually input as text media by users, a significant proportion of UGC now consists of multimedia such as photographs, video and music. With the proliferation of cheap storage and affordable recording devices, interaction with digital multimedia has become a major activity for computer

DOI: 10.4018/jswis.2011100101

users. Furthermore, fast broadband network connections and ubiquitous, network-ready, sensor-laden mobile devices have facilitated the shift of this interaction to the global stage on the increasingly-available Internet.

Users create, store, upload, download, tag, rate, review, browse, search and share text, photograph, video and audio resources using a myriad of hardware and software tools on their personal computers, local networks, mobile devices and the Internet. This UGC multimedia is the currency of popular object-centered social network services like Flickr (photographs), YouTube (video) and Last.fm (music).

The spread of peer-to-peer distribution on the Internet and between mobile devices via Personal Area Networking (PAN) allows users to share raw multimedia directly with each other, increasing throughput over the network by cutting out centralised "middle-men" such as Web servers. However, users of social networking services often find their profiles, preferences, tags and other meta-content locked into the proprietary repository of the service that they used to create the content. Every time a user wants to join another online social network hosted by a different social network service, they must leave their existing content behind and recreate or re-upload their user profile and all other UGC. This repetitive process can be extremely tedious and leads to the current situation where social networking services hoard UGC in isolated and disjointed islands of data.

Standards can bridge the gaps between these islands by providing best-practice means of storing and sharing UGC as portable, reusable data. Advocates of data portability -such as the Data Portability project[1] -believe that users should be able to move, share, and control their identity, photographs, videos and all other forms of personal data independently. This can be done by separating the user's content from the social network service's functionality; in this way social network services can still compete for membership based on the value added to the user's content by their functionality. Data portability can be enabled by the widescale adoption of reusable and extensible standards that allow users to control, share, and move their data from one system to another. Such standards are in fact at the heart of the Semantic Web vision, which has data portability as one of its key features.

Due to the dependence of the Semantic Web on ontology-based metadata, an important question is how to support the creation of this semantic metadata. As online social networking and the Semantic Web converge, a social Semantic Web is emerging which may help kickstart this process: a Web of collaborative knowledge management which is able to provide useful information based on human contributions and which becomes more useful as more people participate.

Instead of relying entirely on automated semantics with formal ontology processing and inferencing, the idea behind the social Semantic Web is to complement the formal Semantic Web vision by adding a pragmatic approach relying on heuristic classification and tagging to create semantic metadata in standard description languages. This requires a continuous process of eliciting crucial knowledge of a domain through semi-formal ontologies and emphasises the importance of manually-created, loose semantics as a means to initialise the vision of the Semantic Web. While the Semantic Web enables integration of domain-specific processing with precise automatic logic inference computing across domains, the social Semantic Web offers a more social interface to semantics, allowing interoperability between objects of interest, actions and users.

To increase the relevancy of metadata, there have been efforts to expose, share, and connect Resource Description Framework (RDF) (Klyne & Carroll, 2004) metadata as Linked Data on the Web (Berners-Lee, 2006; Bizer, Cyganiak, & Heath, 2007). Linked Data refers to a style of publishing and interlinking structured data on the Web; the basic assumption behind Linked Data is that the value and usefulness of data increases the more it is interlinked with other data.

While much research has focused on managing and publishing existing knowledge

as portable, semantic metadata, we identify a need to support the initial supply of valuable, relevant, linked metadata to the Semantic Web. Indeed, principles, standards and tools are now emerging to support the creation of metadata that capture knowledge about community- and user-generated content (Breslin, Harth, Bojars, & Decker, 2005; Adida, Birbeck, McCarron, & Pemberton, 2008). However, most of these approaches revolve around capturing metadata from text or XHTML content; less focus on creating metadata describing UGC multimedia.

The complex activities that occur between UGC generation and sharing are receiving increased attention (management, retrieval, recommendation, etc.) (Rodden & Wood, 2003; Kirk, Sellen, Rother, & Wood, 2006; Lindley, Durrant, Kirk, & Taylor, 2009). The work reported in this paper focuses on one such activity: annotation. Annotation enhances the media collections with metadata which facilitates consequent organisation, retrieval and sharing. While multimedia resources are now the subject of major knowledge management activity for Web users, the rates of metadata creation or annotation have fallen behind the rate at which multimedia data are created:

- 95B photos on Flickr and Facebook alone (Pingdom, 2011; Mitchell, 2011).
- 152M blog posts on the Internet (Pingdom, 2011).
- 20B tweets on Twitter during 2010 (Pingdom, 2011).
- 3B photos uploaded to Facebook per month (Pingdom, 2011).
- 35hrs of new video every minute on YouTube (Pingdom, 2011).

It is further speculated that the rapid adoption of camera phones worldwide generated an additional 200 billion digital photographs in 2008 alone, with camera phones set to take 89% of the mobile market by 2009 (InfoTrends/CAP Ventures, 2004). This situation is set to further deteriorate in the future: forecasted mobile phone user and retail trends are reflected in Figure 1 (Euromonitor International, 2010; US Census Bureau, 2010).

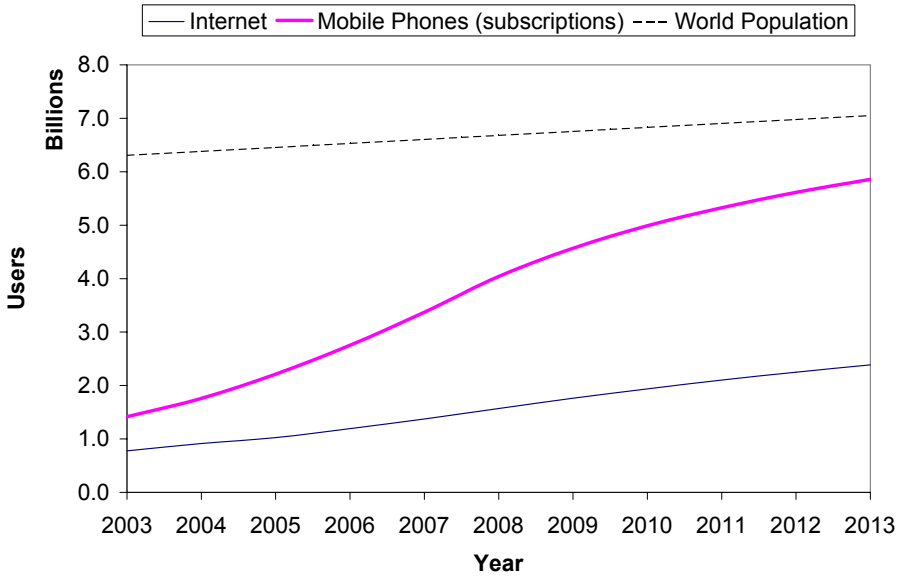
With such an enormous amount of multimedia data, a user manually looking for a particular media resource may feel like they are looking for the proverbial "needle in a haystack"; the result is that the user suffers from a particularly acute branch of information overload as their attempts to find relevant resources are frustrated. The problem presented by the gap between audio-visual data and personal knowledge has been termed the "semantic gap".

Smeulders, Worring, Santini, Gupta, and Jain (2000) first defined the semantic gap specifically for visual media as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" while (Dorai & Ventakesh, 2003) refine and generalise it to "the gulf between the rich meaning and interpretation that users expect systems to associate with their queries for searching and browsing media and the shallow, low level features (content descriptions) that the systems actually compute". Smeulders, Worring, Santini, Gupta, and Jain (2000) further elaborate that while "the user seeks semantic similarity, the database can only provide similarity on data processing".

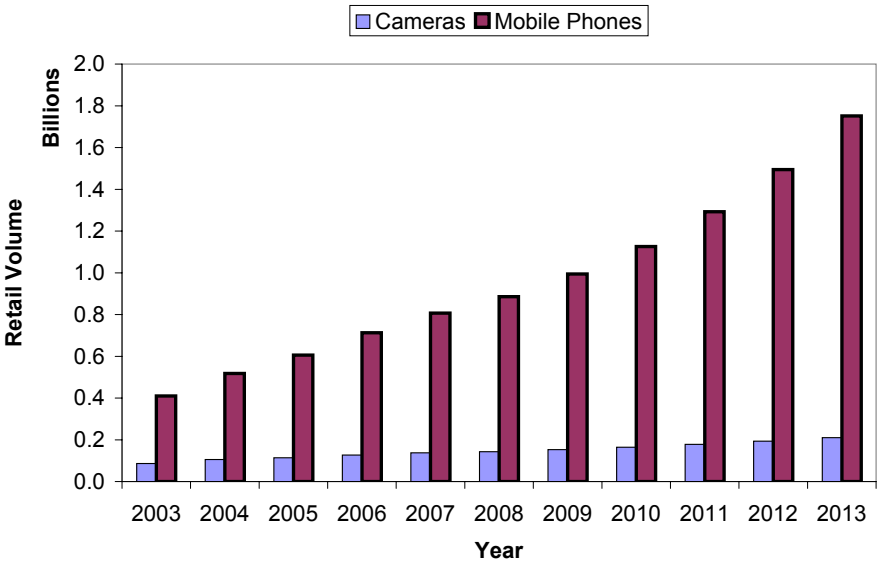
Whereas text or structured data such as XHTML may yield some explicit content that can be extracted using natural language processing or keyword extraction, it is more difficult to bridge the semantic gap with audio-visual data due to the analogue nature and high bit-rates involved. Useful metadata describing media resources bridges the semantic gap between what a resource means to a user and what it means to computers. This enables applications to perform the hard work of finding resources for the user and thereby alleviates the audio-visual data overload.

To enable the machine to retrieve media resources for the user, however, we must first analyse how users themselves mentally recall the resources: this is covered next in Section 2. Section 3 highlights key related work.

Figure 1. Forecasted user and retail volume trends 2003-2013: (a) Internet & mobile phone users; (b) Camera & mobile phone retail volumes (Euromonitor International, 2010; US Census Bureau, 2010)



(a)



(b)

Then Section 4 outlines the ACRONYM annotation framework for context-aware UGC media annotation. Since photographs are the most ubiquitous, most prolific and therefore most urgent form of UGC media to address, the initial use case for ACRONYM has been personal photograph annotation as outlined in previous work (Monaghan & O'Sullivan, 2007) and with more recent results summarised in the following Section 5. The underlying recom-

mendation algorithms used, however, may be used to annotate general UGC multimedia, as will be shown in Section 6. Section 7 presents an evaluation of the recommendation algorithms before Section 8 concludes the paper.

2. MOTIVATION: USEFUL RECALL CUES

To organise and search through large digital multimedia collections it is necessary to extract recognisable features that can be stored as metadata and used as memory cues and filters when browsing the collections, especially as these collections grow into the tens of thousands and span dozens of years. For example, tracks in music collections may be filtered according to artist, album, genre, melody or lyrics while movies may be navigated by actors, directors, genres or parental guidance level. Meanwhile, digital image libraries of artwork such as paintings may be organised by artist, period, style or patron in the same way that catalogue images such as stock photography may be organised by dominant colour, theme or subject. Likewise, (UGC) media resources may be organised by where they were created, when they were created or who is represented.

A user study and a survey conducted by Stanford University (Naaman, Harada, Wangy, Garcia-Molina, & Paepcke, 2004) expose the most useful cue categories for recalling and finding photographs. The results of the study indicate that users recall photographs primarily by the following cues:

- Who is depicted in the photograph.
- Where the photograph was taken.
- What event the photograph covers.

The conclusions of the Stanford study generally agree with Wagenaar (1986) who stated that the most important categories for recall in general are 'who', 'where' and 'when' in that order, and with the more recent survey by Hasan and Jameson (2008). The importance of the chronological ordering of photographs had previously also been highlighted in a preceding study (Rodden & Wood, 2003). The Stanford study delved deeper into the important 'who' category and revealed that the most important cues for photographs in this category in descending order of importance were (Naaman, Harada, Wangy, Garcia-Molina, & Paepcke, 2004):

(i) The identity of subjects in the photograph.
(ii) The number of subjects in the photograph regardless of identity.
(iii) The identity of people who were present at the time of capture, but not necessarily as subjects in the photograph.

A key challenge is how to create this useful description metadata. Manual annotation of multimedia is tedious and consumes large amounts of time. Content-based approaches attempt to directly address the semantic gap by using content-based tools that try to extract semantic information from the audio-visual content of multimedia (Veltkamp & Tanase, 2002). For example, low-level visual features such as dominant colour, brightness and even simple shapes can be easily extracted from images. However, the semantic gap between identifying the low-level features and recognising important semantic themes are still wide and error-prone (Suh & Bederson, 2007).

For example, off-the-shelf software like iPhoto and Picasa make use of face recognition to recommend the identities of people present in photographs. However, traditional face recognition methods cannot satisfy the requirement of "reliable automatic face annotation" for the case of personal photographs (Choi, Yang, Ro, & Plataniotis, 2008). Robust and reliable face recognition is still not available as it depends on: (i) large training sets; (ii) direct alignment of faces to the camera; and (iii) the illumination conditions at the scene of capture (O'Toole, Phillips, Jiang, Ayyad, Penard, & Abdi, 2007).

This is especially true when considering media recorded by multi-purpose mobile devices: not only is the quality produced by the hardware lower but often media resources

are 'snapped' in unposed and active situations (Davis, Smith, Canny, Good, King, & Janakiraman, 2005). Even more complex is the ability to identify semantic themes (such as events or activities) by analysing audio-visual features.

Alternatively, context-based approaches present a lightweight, robust and scalable option to both support the abstract way in which users actually think about media resources and to compliment content-based approaches. To clarify context in the scope of this paper, we now adopt the following definition formulated by Dey and Abowd (2000).

Definition 1. *Context is any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.*

Context-based approaches attempt to automatically organise multimedia collections using context metadata provided by the recording device, e.g., the timestamps of photographs (Girgensohn, Adcock, Cooper, Foote, & Wilcox, 2003; Naaman, Yeh, Garcia-Molina, Paepcke, 2005). Context-based tools often suffer from a lack of useful context metadata from the source device. To combat this, they often supply an interface and tools for the user to also enhance and improve the organisation manually, in which case they can be considered semi-automatic (Suh & Bederson, 2007). Additionally, context can be combined with content to provide a hybrid solution, e.g., by taking into account the visual content of faces and clothes depicted over the context of a short space of time or event (Zhao, Teo, Liu, Chua, & Jain, 2006; Choi, Yang, Ro, & Plataniotis, 2008), or by mining existing sources of context (Tuffield et al., 2006).

3. RELATED WORK

There is a significant amount of research into multimedia annotation and related areas: here we limit our review to context-aware photograph annotation approaches for the sake of comparison within the photograph annotation use case.

Davis, Smith, Canny, Good, King, and Janakiraman (2005) use a statistical factor analysis approach to combine face recognition with context information such as detected Bluetooth devices, GPS coordinates and previous photograph annotations to suggest people for annotation to camera phone photographs. This previous work uses binary-valued statements about photographs as their observed variables: a disadvantage of only taking into account direct properties of a photograph is that it ignores direct relationships that instances may have with each other independent of any photographs, e.g., people may state that they know each other.

A further disadvantage arises from the use of binary data. For example, the binary model cannot use the geographic proximity of photographs, only the Boolean value of whether a photograph belongs to an artificially created geographic cluster or not. The binary approach requires photographs to be clustered geographically into an arbitrary number of clusters (100 in their case), the granularity of which may be too coarse for many applications and which does not scale well. For example, photographs taken in Ireland and Germany may be indistinguishable geographically under this model and may well both be included in the arbitrary 'Europe' cluster.

Finally, Davis, Smith, Canny, Good, King, and Janakiraman (2005) do not describe the use of any metadata standard to capture the annotations. The evaluation of their approach is further discussed in Section 7.5.

PhotoCompas combines temporal and geographic data with existing Web services to automatically provide higher level context metadata on several temporal, geographic and event recall cues (Naaman, 2005). They use GPS-enabled digital cameras to stamp photographs taken with time and geographical coordinates of capture. Given manual annotations from previous photographs in a local database, PhotoCompas can also suggest the probable identity of subjects in

consequent photographs by monitoring previous manual annotations of subjects and maintaining a history of their whereabouts (Naaman, 2005). Suggestions for subject annotation are made for subsequent photographs by checking their time and location metadata.

ZoneTag is a prototype for Nokia S60 smart-phones that allows the user to upload images to the Flickr website (Ahern, King, Naaman, Nair, & Yang, 2007; Naaman & Nair, 2008). ZoneTag leverages the location and time captured by the smartphone to find a location tag and to suggest other tags based on those previously entered by the user and their social network in similar contexts, offering valuable automation. The proprietary keyword tags employed allow minimal expressivity and limited interoperability, and since they are stored and transmitted to Flickr as separate XML data their portability is limited.

Photocopain focuses on using what the authors call "low value" information from cheaply available sources such as the user's calendar, Flickr tags, EXIF metadata and GPS logs to support domain-independent annotation automation (Tuffield et al., 2006). Content analysis provides further estimates, e.g., whether a photograph depicts natural or artificial objects. They create expressive, interoperable annotations by reusing Semantic Web standards and metadata. They plan to construct narratives to let users make better sense of digital data.

Zhao, Teo, Liu, Chua, and Jain (2006) propose an automated method to annotate people to family photographs based on face, content and social context information. They adopt an adaptive event clustering method to cluster photos based on time and then location to support "social" context-based recommendations. They report that while their social context analysis improves the recall of content recognition, it only makes minor improvements to precision. Body content information is used to identify other instances of known people and to improve face recognition accuracy as it can reject falsely recognised faces. While they state that their current body clustering technique is still preliminary and is far from ideal performance, future work includes improving its performance, as well as adding the ability to detect bodies without detected faces, and better use of social context information.

Similarly, Choi, Yang, Ro, and Plataniotis (2008) use person clustering methods based on the time and space context of photographs in combination with clothes recognition to improve face recognition performance. The two approaches do not mine or make use of existing data sources independent of the photographs nor do they generate reusable metadata. Conversely, an additional source of context that has been proposed for exploitation is the medium through which people communicate in relation to photographs: this could be the digital medium that is used to send or share a photograph (Lieberman, Rosenzweig, & Singh, 2001) or audible conversation (Barthelmess, Kaiser, & McGee, 2007).

The following section presents the ACRONYM annotation framework, which takes a robust, scalable, context-based approach towards automating UGC multimedia annotation.

4. ACRONYM ANNOTATION FRAMEWORK

The ACRONYM annotation framework enables the capture of data from the user's real-world, ambient environment via mobile device sensors and the mining of data from the user's information space where they may have already expended effort creating data about themselves. To this end, the framework essentially comprises two independent processes:

(i) Mobile device-based process to capture data from the real world (optional).
(ii) Web-based annotation process with more processing and information resources at its disposal to perform mining and further computation on the real world data.

Note that the latter may annotate any UGC media on the Web and that the former is optional

and not strictly necessary, although it increases annotation automation. The mobile-and Web-based processes each gather context data from their respective space:

(i) Ambient space: context of the user in the real world.
(ii) Information space: context of the user in the digital world.

4.1. Harvesting Context from Ambient Spaces

The first source of context data is the user's ambient space, or the space they occupy in the real world. A number of technologies may be exploited to harvest data from the user's ambient space. The inherent connectivity of mobile devices can be exploited to automate mundane tasks for the user.

Firstly, auto-synchronisation of mobile device clocks with a network provider's global clock eliminates the need for them to be manually set, regardless of power depletion and consequent cold restart.

Secondly, an increasing number of mobile devices are inherently location-aware. This location can be provided either by a GPS receiver (either built-in or via a wireless connection), via the network provider or through other means such as tracking nearby cell towers or devices. Crucially, the underlying means of location detection is becoming increasingly unimportant as it is further abstracted by middleware on phones that can accept input from various sources. Hybrid positioning systems such as Skyhook XPS[2] combine several sources to provide geographic context that is more accurate than that of any single source alone.

Thirdly, mobile devices can use their PAN connectivity to detect nearby objects. This detection provides further context as to the ambient situation in which media resources are created. Blue-tooth is practically ubiquitous amongst mobile devices and with a range of approximately 10 metres it can provide a new technological context surrounding the user (Davis, Smith, Canny, Good, King, & Janakiraman, 2005; Lavelle, Byrne, Gurrin, Smeaton, & Jones, 2007). With Wireless Local Area Network (WLAN) protocols gaining a foothold on mobile devices, we envisage that it is a matter of time until the actual underlying means of detecting nearby devices will become unimportant as technological context is abstracted in a similar way to that of geographical context.

Lastly, the user interface of a mobile device can be used for direct input from the user. Context-aware annotation systems such as ZoneTag adopt this user interface to allow the user to manually input valuable, reusable context in the form of tags (Ahern, King, Naaman, Nair, & Yang, 2007; Naaman & Nair, 2008).

4.2. Gathering Context from Information Spaces

The second source of context is the user's information space, or the space they occupy in the digital world, e.g., on their computer, social network or the Web. A number of sources may be mined to access data in the user's information space:

- The user's social network.
- The user's calendar.
- Data manually loaded or input by the user via a user interface.
- Web services that provide access to online datasets. Note that even if the user has no digital presence and no inclination to enter data manually that at least public information from Web services and linked open data sources such as GeoNames[3] are still available for mining.

5. USE CASE: PHOTOGRAPH ANNOTATION

With the ACRONYM framework's two annotation processes in mind, we implemented two applications for the use case of photograph annotation:

(i) Mobile device-based camera application.
(ii) Web-based photograph annotation application.

5.1. Mobile Device-Based Camera Application

ACRONYM's mobile device-based camera application harvests context from the ambient scene at the time of photograph capture. This was implemented as a downloadable Java Midlet[4] to be platform-independent and to take advantage of Java Microedition's abstraction interface layers that provide access to a device's services and sensors. When the user takes a photograph, this Java MIDlet runs in the background to exploit the increasingly ubiquitous sensors available in mobile devices to capture contextual metadata from the ambient environment. As illustrated in Figure 2, parallel processing threads simultaneously capture:

(i) The time of creation via the network provider
(ii) The coordinates of capture if a location service (e.g., GPS) is available
(iii) The unique physical addresses of nearby devices at the time of creation via PAN transceivers
(iv) The email address of the creator via a once-off login

Once these threads have completed gathering information from the ambient environment, an initial RDF graph describing a created photograph is created using these contextual metadata which is in turn encoded in XML. While a discussion about the nature of this graph and the ontologies used is outside the scope of this paper, further details are available in Monaghan (2008) (Ch. 4). It is suffice to say here that in the spirit of Semantic Web reuse over reinvention, the ontology used is integrated from elements of existing popular ontologies where possible. Figure 3 shows key classes and properties of the integrated ontology.

The RDF/XML is next wrapped in an eXtensible Metadata Platform (XMP) (Adobe Developer Technologies, 2005) packet which is in turn embedded in a JPEG file along with the image data as already illustrated in Figure 2. The resulting JPEG file can be shared through normal means with another device, user or indeed uploaded to the Web while preserving its descriptive metadata. Later this initial RDF description will come into play to help automate the creation of annotations for the photograph that are actually useful to users for recalling the photograph from a query engine.

5.2. Web-Based Photograph Annotation Application

ACRONYM's Web-based photograph annotation application mines context from the user's information space, combines it with that from their ambient space and supports the reuse of the combined contextual knowledge for context-aware photograph annotation. This was implemented as a lightweight Java Servlet; an online demo is available[5]. This annotation application allows the user to login and annotate any JPEG photograph that is accessible by HTTP; the only requirement is that the photograph must at least have an EXIF timestamp. In the background, the annotation application exploits the contextual knowledge available to it to support the user in their annotation task by recommending particular annotations. These recommendations are in the form of a ranked list for each of the three most useful recall cues for photograph management:

(i) People depicted
(ii) Geographical features depicted
(iii) Events covered

Note that annotations and recommendations for each of people, places and events are placed side by side to ease comparison; these are all presented as simple "tags" with human-readable labels, familiar to users of contemporary popular photograph annotation tools such as Flickr or Facebook. The machine-readable semantics and ontology are hidden

Figure 2. Harvesting context metadata from the user's ambient space

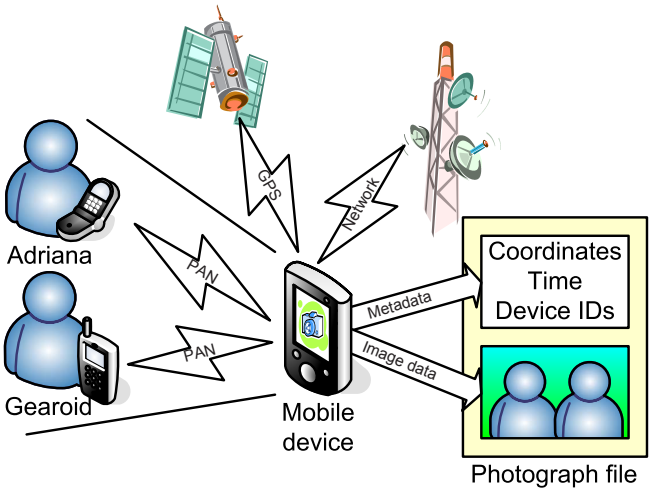
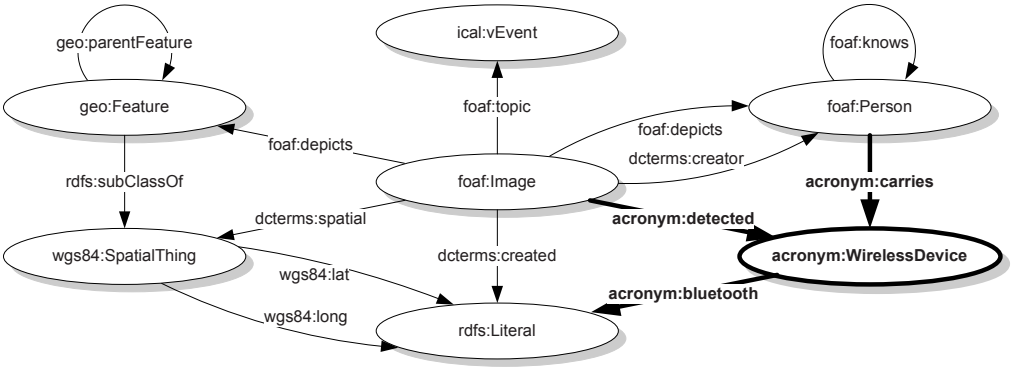


Figure 3. Key classes and properties of integrated ontology for context-aware photograph annotation



from the user, who simply selects from the list of recommendations to annotate the photograph.

The current Web-based application supports the semi-automatic annotation of places and events using fairly straightforward approaches based on spatial and temporal proximity respectively. However, a set of novel algorithms have been developed to support the annotation of the people represented, which is both one of the most useful recall cue to users as well as the most difficult to automate. Therefore, the remainder of this paper considers the place and event annotation algorithms outside of scope, and now focuses on the context-aware recommendation algorithms used to recommend people. Furthermore, while the use case in this section has demonstrated an application to personal photographs, the recommendation approach is generalised and can recommend the people represented in any UGC multimedia resource, be it a text post, a photograph, a video or an audio clip.

6. RECOMMENDATION ALGORITHMS

To inform the recommendation of the recall cues represented in UGC multimedia resources, context metrics may be defined to measure the strength of certain relationships between instances of recall cues (e.g., people, places, events) and individual UGC media resources. The hypothesis is that there exist latent relationships (social, geographical, temporal, etc.) between instances of recall cues and individual time-stamped UGC media resources in the ambient space that have observable representations in the form of contextual knowledge in the information space. The context metrics use statistical dataset analysis of these observations to provide a measure of the strength of the individual latent relationships.

In practice, the contextual knowledge is given domain-specific meaning by a media annotation ontology, e.g., the integrated ontology referred to in Section 5 for the domain of photographs. We also define combination methods that combine the measurements of several context metrics to present an aggregated result with the aim to provide more accurate and robust recommendations.

6.1. Context Metrics

While the set of context metrics may be extended without change to the combination methods, we define here several specific context metrics to specifically inform the recommendation of people represented in UGC media resources. In particular, these metrics measure the strength of several relationships between individual people and UGC media resources.

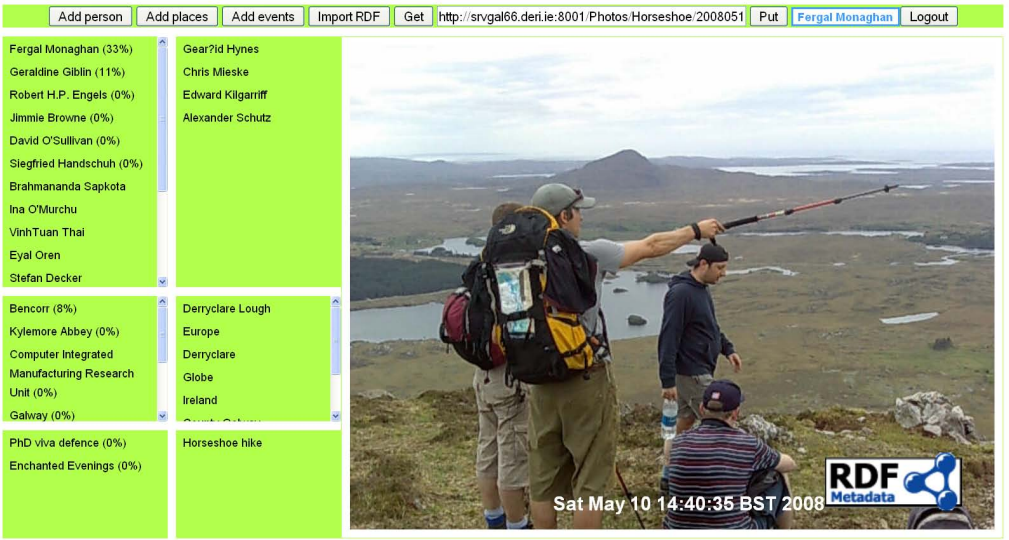
For a dataset, let P be the set of all people, R the set of all media resources and D the set of all wireless devices. Then let $p \in P$ be an individual person, $r \in R$ be an individual media resource and $d \in D$ be an individual device. Then, to enable the description of the context metrics themselves, we predefine the following important sets and properties:

- Let *rep'tations*(p) $\subset R$ be the subset of resources representing a person p.
- Let *represents*(r) $\subset P$ be the subset of people a resource r represents.
- Let *knows*(p) $\subset P$ be the subset of people a person p knows.
- Let *detections*(d) $\subset R$ be the subset of resources created with a device d detected nearby.
- Let *detected*(r) $\subset D$ be the subset of devices detected nearby when a resource r was created.
- Let *carries*(p) $\subset D$ be the subset of devices a person p carries.
- Let *created*(r) be the time when a resource r was created.
- Let *creator*(r) $\in$ P be the creator of resource r.
- Let *createdBy*(p) $\subset R$ be the subset of resources created by person p.
- Let *creations*(r) $\subset R$ be the subset of resources created by the creator of resource r, *createdBy*(*creator*(r)).
- Let *spatial*(r) be the spatial coordinates where a resource r was created.

These definitions are used by the context metrics, which each calculate a value between 0 and 1 which gives that metric's confidence that a person is represented in a given media resource. Furthermore, several of the metrics give the conditional probability that a person is represented given the existing contextual knowledge (Figure 4). To this end, they use the mathematical concept of set containment to measure what proportion of one set is contained within another (Broder, 1997). There are six context metrics based on:

(i) Spatio-temporal proximity.
(ii) Person acquaintance containment.
(iii) Person corepresentation containment.
(iv) Device carriage containment.
(v) Device copresence containment.
(vi) Creator containment.

Figure 4. Screenshot of ACRONYM web-based annotation application user interface



6.1.1. Spatio-Temporal Proximity

This metric measures how close person p is estimated to have been to the scene of creation of resource r, at the time the resource was created, $created(r)$. First, the location of the person at time $created(r)$ is estimated. Let $r_{-1} \in represents(p)$ be the previous resource - in chronological order - that p is represented in and $r_1 \in represents(p)$ be the next. The estimated location of p at $created(r)$ is then estimated by interpolating $spatial(r_{-1})$ and $spatial(r_1)$ at time $created(r)$; this interpolation assumes a spherical coordinate system such as that used on Earth (Vincenty, 1975). Note that if r is the oldest resource and therefore r_{-1} is unavailable, $spatial(r_1)$ and $spatial(r_2)$ are extrapolated instead; interpolation and extrapolation are handled identically. Likewise, in the case that r is the most recent resource and so r_1 is unavailable, $spatial(r_{-2})$ and $spatial(r_{-1})$ are extrapolated instead.

Definition 2. *Let intcrds(p, t) be the interpolated coordinates of person p at time t, assuming a spherical coordinate system.*

Definition 3. *Let gcd(coords$_1$, coords$_2$) be the great-circle (shortest path) distance between two points on a sphere with radius equal to that of the Earth's.*

This allows the distance between the interpolated co-ordinates of person p and the co-ordinates of creation of resource r to be defined:

$$dist(p,r) = gcd(intcrds(p, created(r)), spatial(r))$$

Then the final metric gives the spatio-temporal proximity of person p to resource r as a value between 0 and 1:

$$metric_1(p, r) = e^{-dist(p,r)} \quad (1)$$

For example, $metric_1(p,r)$ might give a 10% confidence that p is depicted and/or audible in video r due to their estimated spatial proximity at the time.

6.1.2. Person Acquaintance Containment

The person acquaintance containment metric is a measure of how much of a person's social net-

work is represented in a resource. The measurement taken is the containment (Broder, 1997) of the set of people that person *p* is asserted to know within the set of people represented. For resource *r*, it measures the containment of the set of *p*'s acquaintances within the set of those people represented:

$$metric_2(p,r) = \frac{\left|knows(p) \cap represents(r)\right|}{\left|knows(p)\right|} \quad (2)$$

6.1.3. Person Corepresentation Containment

Many people may not create social networking profiles or explicitly assert statements about themselves. Therefore, we also define a person corepresentation containment metric which aims to implicitly measure how often *p* is represented with those people already represented in *r*, without requiring them to assert their acquaintances (Figure 5). First the containment of the set of resources representing another person $q \in P$ within that of *p* is calculated using the $PeopleRank_{photo}$ single-resource variant (Naaman, Yeh, Garcia-Molina, & Paepcke, 2005):

$$PeopleRank(q,p) = \frac{\left|rep'tations(q) \cap rep'tations(p)\right|}{\left|rep'tations(q)\right|}$$

For example, *PeopleRank*(*q*, *p*) could calculate that *p* is mentioned in 20% of all blog posts *q* is mentioned in. Then the containment of the sets of representations of each person represented in *r* within that of *p* can be used to calculate the final person corepresentation containment metric:

$$metric_3(p,r) = \sum_{q \in represents(r)} \frac{PeopleRank(q,p)}{\left|represents(r)\right|} \quad (3)$$

For example, $metric_3(p, r)$ might give a 15% confidence that *p* is mentioned in blog post *r* due to their usually being co-mentioned with those mentioned.

6.1.4. Device Carriage Containment

Similarly to the person acquaintance containment metric, the device carriage containment metric is a measure of how many of a person's wireless devices were detected nearby to the scene of resource creation. The measurement taken is the containment of the set of devices that person *p* is asserted to carry within the set of devices detected. For resource *r*, it measures the containment of the set of *p*'s devices within the set of those detected:

$$metric_4(p,r) = \frac{\left|carries(p) \cap \det ected(r)\right|}{\left|carries(p)\right|} \quad (4)$$

6.1.5. Device Copresence Containment

While the ACRONYM Ontology[6] allows people to explicitly assert which devices they carry, it must be assumed that many people will not. Therefore we also define a device copresence containment metric which aims to implicitly measure how often *p* is present with those devices detected near to the creation of *r*, without requiring users to state the devices they carry (Figure 6). First the containment of the set of resources for which device $d \in D$ was detected within the set of resources representing *p* must be defined:

$$prescon(d,p) = \frac{\left|detections(d) \cap rep'tations(p)\right|}{\left|detections(d)\right|}$$

Then the containment of the set of detections of each device detected in *r* within the set of representations of *p* can be used to calculate the final device copresence containment metric:

Figure 5. Corepresentation of people with other people

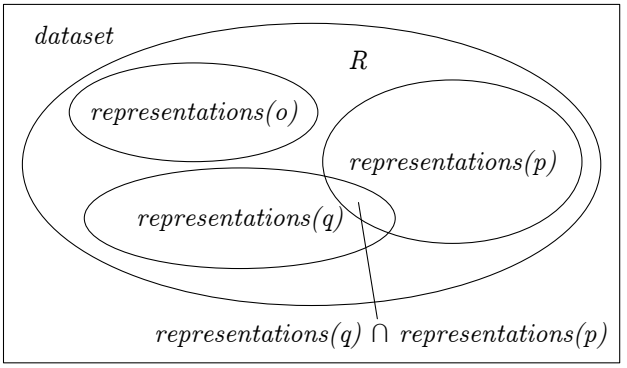
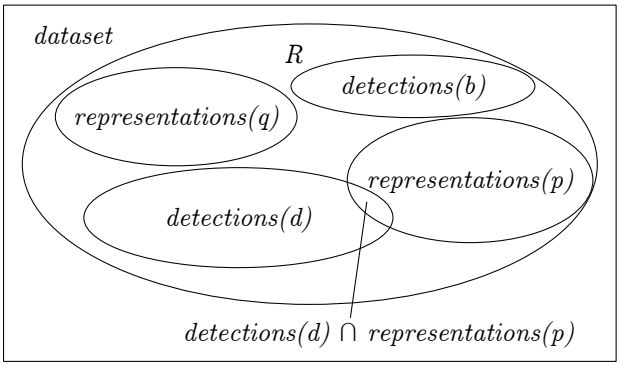


Figure 6.Copresence of people with devices



$$metric_5(p,r) = \sum_{d \in detected(r)} \frac{|prescon(d,p)|}{|detected(r)|} \quad (5)$$

$$metric_6(p,r) = \frac{|creations(r) \cap rep'tations(p)|}{|creations(r)|} \quad (6)$$

6.1.6. Creator Containment

The creator containment metric is a measure of how often a person has been represented in resources created by the creator of resource *r*. The measurement taken is the containment of the set of resources created by the creator of *r* within the set of resources that represent person *p*. For resource *r*, it measures the containment of the set of the creator's resources within the set of resources representing *p*:

For example, $metric_6(p,r)$ might give a 35% confidence that *p* is depicted in photograph *r* due to their usually being photographed by the photographer.

6.2. Combination Methods

While recommendations may be made based on any single context metric alone, if metrics are combined they may provide more accurate and more robust recommendations. This section

explores methods to combine the context metrics to provide robust recommendation algorithms. Two combination methods are presented:

(i) Correlation-weighted-mean CombSUM.
(ii) Factor analysis.

Importantly, the combination methods presented here are chosen so as to model people as individuals with unique usage/data profiles, instead of averaging over a generic profile. This means that each combination method can simultaneously handle people with widely varying usage and available data. For example, one extreme person may be very sociable -appearing in lots of photos with friends - and may travel a lot, but may not use computers, gadgets or social networking at all. For this "travelling socialite technophobe" user, the spatio-temporal proximity and person corepresentation metrics may be the most accurate at recommending them in photos they appear in, whereas the person acquaintance and device carriage metrics may be the least accurate since the user has provided no data whatsoever for them to function. Therefore this person's model should reflect a higher weighting for those more accurate metrics.

Conversely, the opposite may be true for a user on the other end of the spectrum, who may never leave home nor appear in many photos with others, but who may be a heavy user of gadgets and social networking, keeping their digital presence up to date meticulously. For this "anti-social agoraphobic technophile" user, there may not be many photos to use as historical observations, but still the explicit data provided for the person acquaintance, device carriage and device copresence metrics may provide accurate recommendations for photos they do appear in. Obviously, between these two extreme profiles (and any other combination of extremes) may lie any number of unique individuals. Both combination methods presented here model this individuality.

6.2.1. Correlation-Weighted-Mean CombSUM

The first combination method is a weighted-mean CombSUM method (Fox & Shaw, 1994) which combines the context metric confidence scores, giving a single representation confidence value between 0 and 1 for individual person-resource pairs. Once a metric has been used to calculate the confidence for the representation of p in every resource $r \in R$, these scores form a vector of confidence scores for each resource.

Definition 4. *Let $metric_i(p, r)$ denote context metric i's score for person p and resource r. Then the variable $metric_i(p)$ contains p's confidence scores for all resources $r \in R$.*

This variable is compared with the vector *rep'tations*(*p*) containing either a 0 or 1 for each previously annotated resource depending on whether p is asserted as represented or not. The Pearson correlation between these two variables is then taken as the weight to use for that metric.

Definition 5. *Let corr(x, y) denote the Pearson correlation between any variables x and y.*

$$w_i(p) = corr(metric_i(p), rep'tations(p))$$

When the weights for all metrics have been calculated for *p*, they are applied to their corresponding metric. The mean of the weighted metrics is then calculated using Equation 7 and is taken as the final confidence metric that a person *p* is represented in *r*. Let *m* be the number of context metrics being used; in our case this is six, but this may be extended:

$$confidence_{cs}(p,r) = \sum_{i=1}^{m} \frac{w_i(p) \cdot metric_i(p,r)}{m} \quad (7)$$

For example, $confidence_{cs}(p,r)$ could return 18% combined confidence that singer *p* is au-

dible in audio clip r. Once the confidences for each person have been calculated for a resource, they may be used to rank recommendations to the user to support their task of annotating people to resources. The correlation-based weighting deals with missing data for each person by assigning low weights to metrics that over the dataset do not correlate with actual depictions of that person. For example, for a user who is not part of an online social network (or indeed does not have a digital presence at all), the personalised weights would suppress the useless output of the acquaintance containment and device carriage metrics and allow the algorithm to function purely on data from the recording device and previous annotations.

6.2.2. Factor Analysis

The second combination method uses factor analysis to combine the context metric confidence scores in an unbiased way. Factor analysis refers to a variety of statistical techniques whose common objective is to represent a set of variables in terms of a smaller number of hypothetical variables. A factor analytic approach may be used to address whether there are positive correlation relationships between observed metrics that may be explained by a smaller number of latent, unobserved variables (Kim & Mueller, 1978). For example, the implicit, abstract, unobserved relationship of "social proximity" between people in the real world may cause those people to be explicitly stated as being both acquainted and corepresented in the digital world more often than not.

Such a latent variable may be detected by several metrics; hence it is a common factor shared by those metrics. If more metrics measure one latent variable than another, the purely correlation-based weights used in the CombSUM method produce biased results; in effect, they may lend more weight to certain latent variables simply because there are more metrics available to measure them. Meanwhile, factor analysis has been shown to be the most accurate method on standard collaborative filtering data (Canny, 2002).

In factor analysis, each metric is assumed to be influence by a unique component (assumed to be random noise) and one or more common factors; this is illustrated in the path model in Figure 7. In our approach we determine a separate factor model for each individual person $p \in P$. Factor analysis is described in Kim and Mueller (1978) and Davis, Smith, Canny, Good, King, and Janakiraman (2005) and we will now use notation prevalent in the literature.

X is a vector of n (partially) observed variables. F is a latent vector representing the $k \leq n$ factor variables underlying the observed variables. U is a noise function. Here W is the factor structure matrix: a 2-dimensional matrix of size $n \times k$ giving the scalar influence weight values of the latent factors on the observed variables. The factor model for a person p is then formally described as:

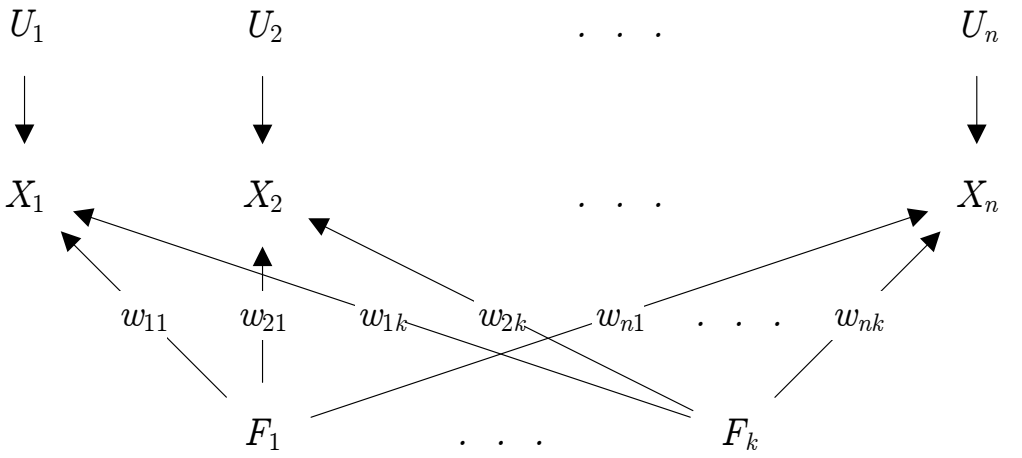
$$X = WF + U \tag{8}$$

In factor analysis, X and F are assumed to be real-valued vectors. U is assumed to be multivariate, independent Gaussian noise. F is assumed to have a Gaussian prior distribution. The $n = m + 1$ fields of the X vector encode all observed variables: there is a continuous variable (between 0 and 1) for each metric plus a discrete binary variable (either 0 or 1) stating whether p is actually represented or not.

With the general factor model for a person formally defined, the factor analytic method requires two phases. In the training phase, standard factor analysis techniques as described in Kim and Mueller (1978) are used on a training set of complete X vectors to determine the most likely values for the set of F vectors and for the factor structure matrix W for person p. The training data includes all metrics and whether p is represented or not.

In the test phase, the algorithm will receive just the metric scores (without the representation values) from a new media resource r. From these partial observations X and from the previously-trained W, F is determined for the case of r. With W and F determined, they

Figure 7. Path model for a multivariate multi-common factor model



are plugged directly back into Equation 8 to predict the missing X value. This predicted value gives a single continuous confidence value $confidence_{fa}(p, r)$ between 0 and 1 that person p is represented in resource r.

Once again, with the confidences for each person $p \in P$ calculated for resource r, they may be used to rank recommendations to the user. The factor analysis deals with missing data and allows the algorithm to function optimally with the data available.

7. EVALUATION

The recommendation algorithms were evaluated for the use case task of personal photograph annotation presented in Section 5.

7.1. Datasets

We evaluate each recommendation algorithm using training and testing datasets through a commonly applied technique of evaluating prediction of deleted values from existing data (Herlocker, Konstan, Terveen, & Riedl, 2004). The difficulty here is in obtaining a large dataset with all the test features needed for an approach that makes use of multiple context metrics (including Bluetooth, social network connections, etc.). Large US government-sponsored photo datasets are traditionally content-focused, e.g., CMU-PIE[7] and FERET[8] and do not have all the features needed for a comprehensive evaluation of context metrics. In particular, an evaluation of our context-based approach requires a dataset including the following features:

- Captured Bluetooth/wireless device addresses
- Social networking connections
- GPS co-ordinates
- Photographer (linked to OSN)
- Located and identified faces (linked to OSN)
- (Optional) Device ownership (linked to OSN)

This kind of combined dataset is only emerging now alongside the emergence of linked data on the social Semantic Web[9], and previously this kind of work would have been limited to the realm of computer vision[10]. It would require several years of community i) participation and ii) co-operation/sharing to generate large quantities of usable test data with all the context features needed. As with so much concerning Semantic Web research, the chicken-and-egg problem rears its ugly head: what should come first, large-scale data acquisition to evaluate approaches or new approaches

to generate data? Inevitably the answer is that to make progress we must attempt a bit of both.

So in the meantime, since we are unaware of any publically available evaluation datasets that include all of the above features, we undertook to collect data with all the features needed to prove the concept. This has obvious budgetary issues (devices, people, time, etc.) which limit its size, compared to the aforementioned government-sponsored datasets of faces. We use several datasets gathered from three users (one of which is an author) who each annotated their own personal photograph collections using the ACRONYM Web-based annotation application outlined in Section 5:

- Datasets 1, 2 and 3 were each annotated by their respective owners; all photographs in 2 and 3 and the majority in 1 were taken with regular digital cameras.
- Dataset Σ is the result of merging and consolidating datasets 1, 2 and 3; this simulates a situation where several users of a social network may benefit from each other's annotations.
- Dataset 1a is a subset of dataset 1 that consists only of photographs captured using the ACRONYM mobile device-based camera application outlined in Section 5.

An anonymised version of dataset Σ (which contains all data and may be split by creator to recreate the other datasets) is publically available online[7].

Due to the private nature of the personal photographs and their metadata to the evaluation users, the dataset had to be anonymised before publication to remove "human-readable" personal names and faces. In this regard, public research is faced with the same restrictions that large OSNs like Facebook encounter: they also do not publish private user data, instead allowing each user to choose their own personalised privacy settings.

While we cannot fit a research paper on multimedia annotation and simultaneously, within the same short article, solve the general problem of data acquisition and publication that somehow sidesteps but does not break privacy laws, in this article we do our utmost to ensure and encourage the repeatability, comparability and extensibility of our work in Section 7.6. There's a whole other branch of "Web Science" emerging on the social, legal, etc., side dealing with this in general. That's a story for another article, and more likely another entire journal issue full of articles.

The anonymisation procedure has two main consequences. Initially, the actual photographic content has been removed. Furthermore, all attempts have been made to purge the metadata of personally-identifying strings as well as URIs to linked data about people: these have been replaced by unique but meaningless and non-dereferenceable strings. All other non-person concepts, e.g., geographical locations retain their identifying strings and original URIs dereferencing more information and integrating them into the Linked Data Web.

Figure 8 illustrates the structure of each evaluation dataset along with the numbers of resources and predicates present in the dataset for collection Σ. Table 1 elaborates with a summary of the number of resources and predicates by class in each collection's dataset; note that depicts predicates that are used to state that people are depicted are denoted with $depicts_p$ while those used similarly for features are denoted with $depicts_f$.

Table 2 gives further statistics for the numbers of people annotated as depicted in some way in each photograph (whether their faces appear or not). The first column gives the total number of photographs in the dataset that actually depict people. The remaining columns give the minimum, maximum, mean and median statistics for the number of people depicted in each photograph. For example, it can be seen that of the 329 photographs in dataset Σ that depict people, each depicts between 1-6 people. The mean is 1.97 people per photograph. Figure 9 shows the frequency distribution of people depictions for all datasets.

Additionally, the 3D scatter plot in Figure 10 gives the distribution of photographs in datasets 1, 2 and 3 in both space (horizontal

Figure 8. Resources and predicates in evaluation dataset Σ

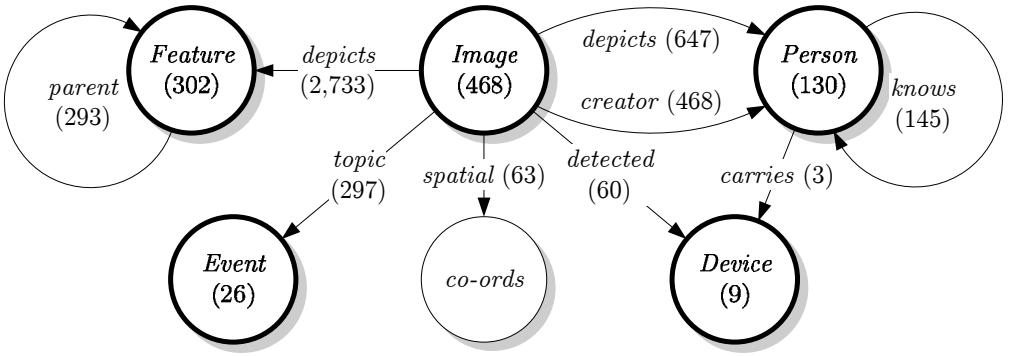
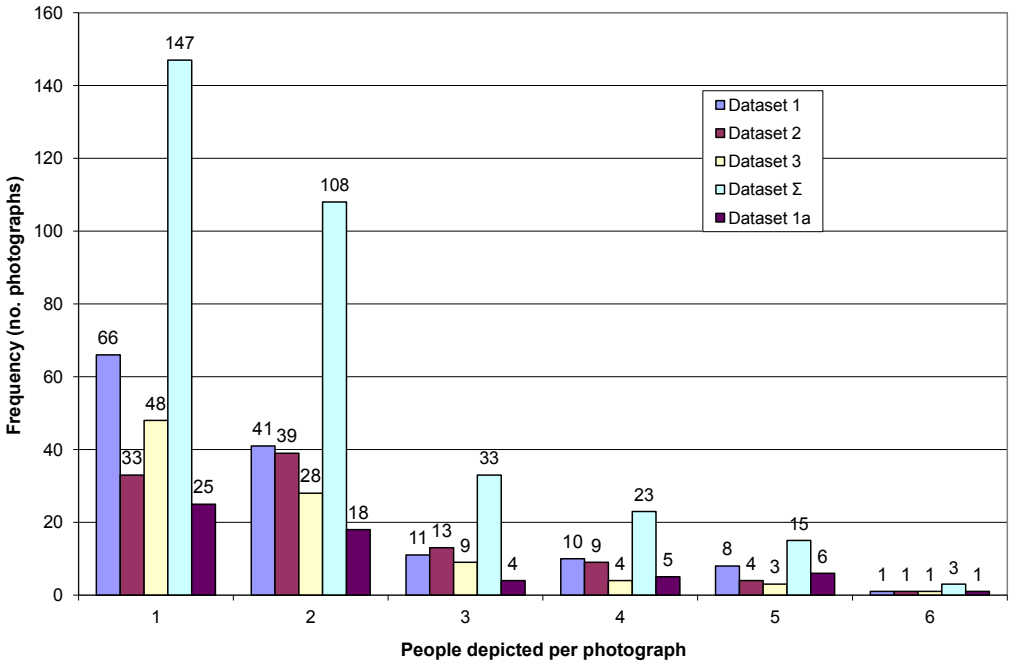
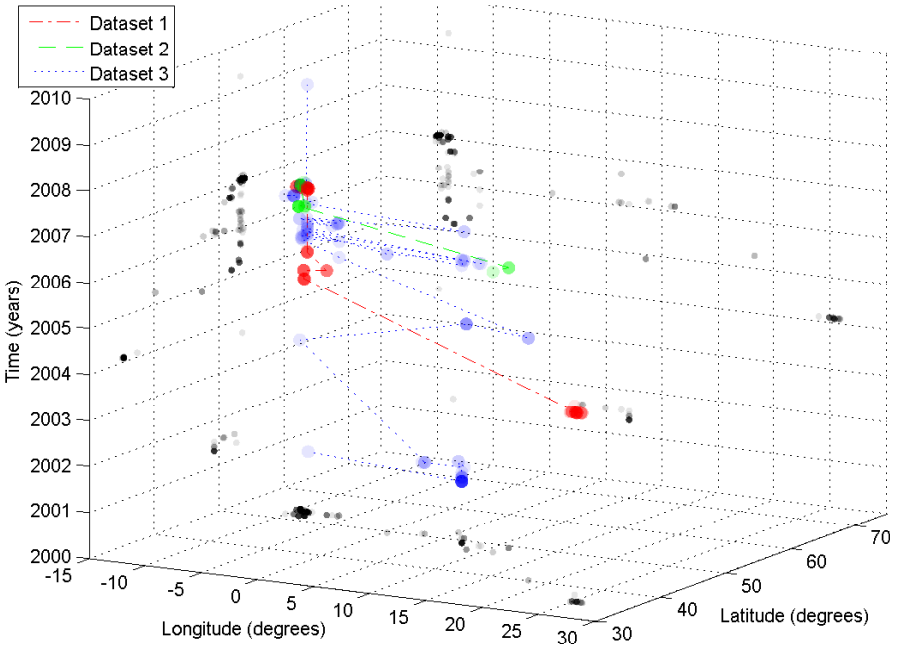


Figure 9. Frequency distribution of people per photograph



axes) and time (vertical axis). The transparency of each data point is set to 0.05 opacity, so that closely clustered photographs are illustrated as overlapping clouds of increased opacity (i.e., clusters of 20+ overlapping photographs appear fully opaque). The size of the data points was aesthetically chosen for visibility and holds no meaning. The data points in each dataset are connected by a line that shows their ordering in time. It can be seen that the "bursty" nature of personal photographs Naaman (2005, Ch. 3) holds true for datasets 1 and 2 which have photographs densely clustered around discrete spatio-temporal events with empty gaps between events. Dataset 3 presents an alternate scenario where the photographs

Figure 10.Spatio-temporal distribution of photographs



are sparsely distributed with many events scattered throughout space and time with few photographs per event (several events are represented by a single photograph), made evident by the voluminous but low-density clouds of blue data points.

7.2. Experimental Setup

In the evaluation, we emulated the process of users annotating their photographs (Naaman, 2005). Having obtained the annotations in advance, we do not require interaction with human subjects. Rather, we "hide" the depicted people annotations from the algorithm; the process of users annotating photographs is simulated by revealing the annotations to the system. In other words, we have a "virtual user" who adds annotations to the system. In this paper we consider a "casual user" mode for the virtual user only (Naaman, 2005): for an extended evaluation please see Monaghan (2008, Ch. 6).

Casual users (Naaman, 2005) annotate a certain percentage of the people represented in media resources: this helps to evaluate how accurate an algorithm's recommendations are when only a fixed percentage of represented people have been annotated. Thus, in the beginning of the process, the algorithm is trained by randomly selecting resources and pre-annotating some of the people that are represented, such that after initialisation 75% of the annotations are known to the algorithm (e.g., for dataset Σ we randomly pre-annotate 485 of its 647 person depiction statements).

After initialisation, the casual user selects one random media resource to fully annotate; at each annotation step, the casual user then annotates one person to the current resource, supported by the algorithm's recommendations. At each annotation step, however, the algorithm starts with the same initial state and the casual user randomly selects a person to annotate to the resource who was not selected before, until all the people represented in that resource have been selected.

The casual user then repeats this for all media resources that were not fully annotated

Table 1. Resources, predicates and faces in each evaluation dataset

Dataset	Image	Person	Feature	Event	Device	creator	depictsp	depictsf	topic	spatial	knows	carries	detected	|T|	|Ttrn|
1	267	64	97	10	9	267	267	1,695	191	63	61	3	60	260	63
2	101	21	47	11	1	101	212	476	101	0	18	1	0	208	74
3	100	67	188	5	0	100	168	562	5	0	66	0	0	168	93
Σ	468	130	302	26	9	468	647	2,733	297	63	145	3	60	636	295
1a	82	64	97	10	9	82	129	468	9	63	61	3	60	125	25

Table 2. Statistics for number of people depicted in each photograph

Dataset	Depict	Minimum	Maximum	Mean	Median
1	137	1	6	1.95	2
2	99	1	6	2.14	2
3	93	1	6	1.81	1
Σ	329	1	6	1.97	2
1a	59	1	6	2.19	2

during initialisation, until all resources have been selected. Due to the dependence on the random nature of the initialisation, the entire evaluation run is then repeated 10 times to yield 10-set cross-validated results.

The primary evaluation technique is prediction of deleted values. However, a requirement of recommender systems is to rank recommendations according to their relevance instead of just returning them as a set of results (Herlocker, Konstan, Terveen, & Riedl, 2004). Since in practice not all recommendations can be displayed or will be considered by the user, for each annotation step we analyse at which rank position the removed statements are recommended.

At each annotation step the algorithm being evaluated returns a ranked recommendation list of people for which we calculate the accuracy using precision and recall. Regular precision and recall are measures for the entire list: they do not account for the quality of ranking the recommendations in the list. Relevance ranking is measured by computing precision at different cut-off points (Manning, Raghavan, & Schütze, 2008, Sec. 8.3; therefore if the first correct recommendation is returned at position n in the list, we calculate the precision at n. Similarly we calculate the recall at n: the actual number of relevant recommendations up to the first correct recommendation in the list. Plotting the trend of precision over recall for the list then yields the precision-recall plot. Averaged over all annotation steps during the virtual user's annotation of a given dataset, this plot gives an impression of information retrieval accuracy.

7.3. Comparison with Face Recognition

The evaluation includes comparison and combination with a face recognition method. While Choi, Yang, Ro, and Plataniotis (2008) shows that a Bayesian face recognition method (Moghaddam, Jebara, & Pentland, 2000) is shown to perform slightly better than the Principal Component Analysis (PCA/eigenfaces) method (Turk & Pentland, 1991) on the manually selected face images from the CMU-PIE and FERET face datasets, (Davis, Smith, Canny, Good, King, & Janakiraman, 2005) has shown PCA to be the more robust method on real-world personal photographs. These images may be noisy, may not have well-aligned faces nor favourable illumination conditions and may have been captured on lower-resolution capture devices like camera phones. Subsequently, while there has been recent enhancements to state-of-the-art face recognition attempting to circumvent the pose, illumination and expression problem, as well as to use body and clothing recognition to boost performance (Zhao, Teo, Liu, Chua, & Jain, 2006; Anguelov, Lee, Gokturk, & Sumengen, 2007; O'Hare & Smeaton, 2009), PCA face recognition is chosen here as a reasonable baseline content analysis approach for comparison and combination on the task of user-generated photograph annotation.

To this end, we reuse the face distance metric used by Davis, Smith, Canny, Good, King, and Janakiraman (2005) as our face proximity content metric. During training for a given evaluation run, the face proximity metric uses all known training faces of a given person to

Figure 11. Frequency distribution of faces per person

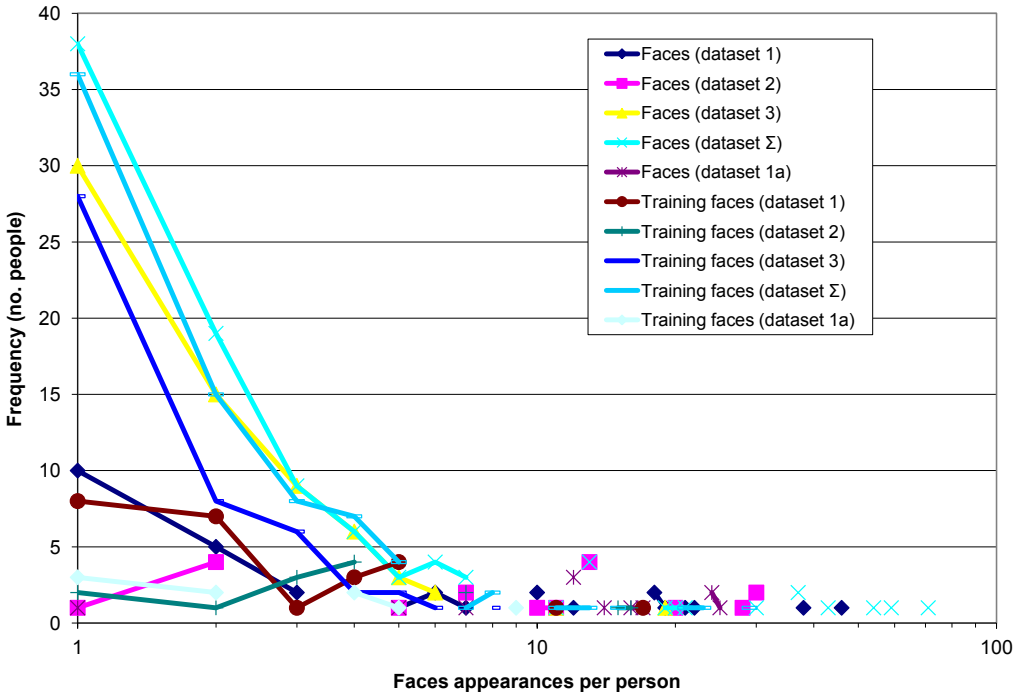


Figure 12. Precision-recall plot for recommendations to casual annotator of dataset Σ

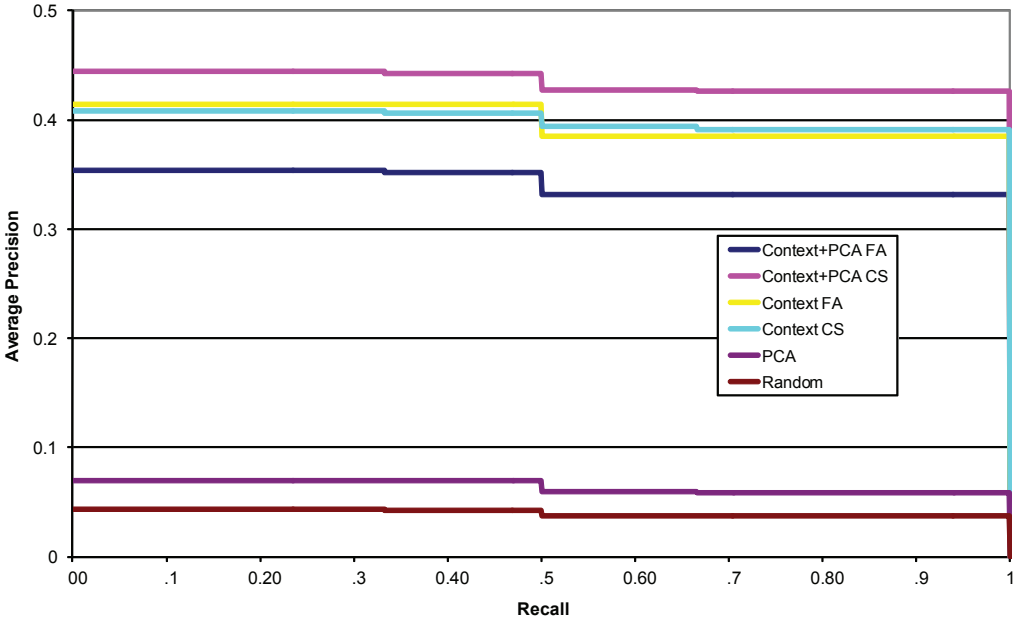


Table 3. Statistics for number of face appearances for each person

Dataset	Depict	No face	Min.	Max.	Mean	Median
1	32	7(1)	0(0)	17(46)	2.66(8.25)	2(2.5)
2	18	7(0)	0(1)	16(30)	4.11(11.61)	3(10.5)
3	67	4(0)	0(1)	8(19)	1.40(2.51)	1(2)
Σ	98	17(1)	0(0)	29(71)	3.13(6.54)	1(2)
1a	12	3(1)	0(0)	9(24)	2.42(10.58)	1.5(9.5)

build a single face class for them, as described in Turk and Pentland (1991). During testing, we first assume perfect detection. This is due to contemporary face detection (as opposed to recognition) being well-solved, accurate and trainable on generic human faces elsewhere (Viola & Jones, 2001). We therefore give the face proximity metric the location of all faces in each photograph during testing, only hiding the identities of the located faces to test recognition alone. The face proximity metric handles multiple faces identically to that of Davis, Smith, Canny, Good, King, and Janakiraman (2005): since there may be multiple faces in a photograph - each giving a distance measurement from a given trained face class - we use the minimum of those distances when considering that face class for recommendation.

For the evaluation of the face proximity metric, each test collection owner manually identified the faces present in their personal photographs by dragging squares and identities across them. They also manually indicated which of those faces were clear enough to be used as profile pictures for training purposes. Table 1 gives the size $|T|$ of the set of faces T and the size of the subset of those that are also training faces $T_{trn} \subset T$ for each dataset.

Table 3 gives further statistics for face appearances of people in each dataset. The first column gives the number of distinct people whose faces are actually depicted in at least one photograph in the dataset. The second column gives the number of those annotated who have no training faces visible, with the number of those who also have no face visible whatsoever given in brackets. The remaining columns give the minimum, maximum, mean and median statistics for the number of training faces per person, with the figures for the total number of faces per person in brackets.

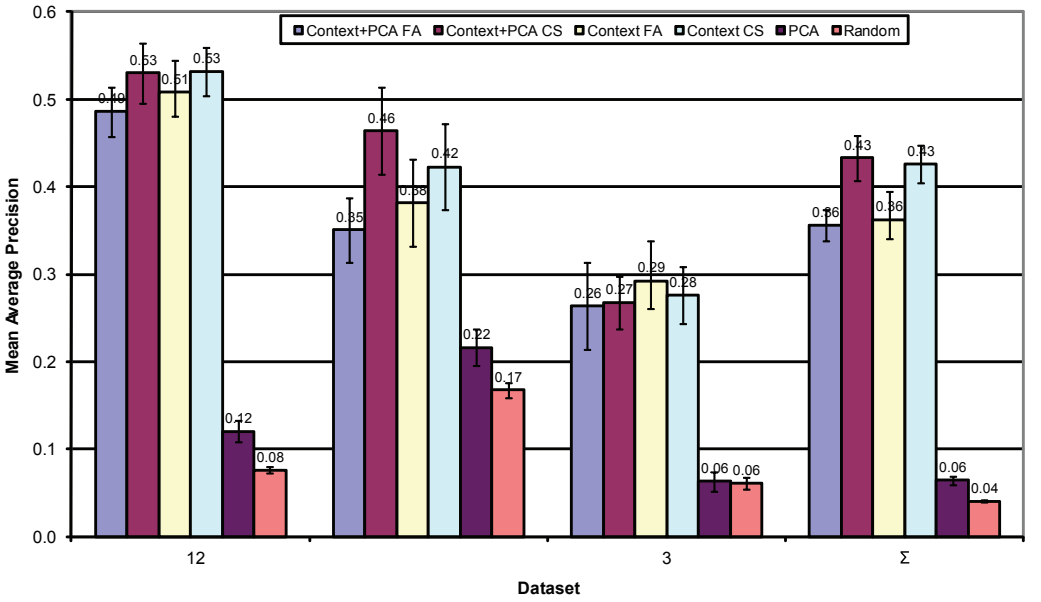
For example, it can be seen that each of the 98 distinct people annotated to photographs in dataset Σ has their face appear between 0-71 times: 1 person present in the photographs does not have their face appear at all. The mean is 6.54 face appearances per person. Each person has between 0-29 training faces: 17 annotated people had no training face at all. The mean is 3.13 training faces per person. Figure 11 shows the frequency distribution of all face appearances for all datasets. The face proximity content metric was then added to the set of context metrics for both comparison and combination during the evaluation.

7.4. Results

Figure 12 shows the trend of precision over recall of recommendations given to a casual user for the consolidated dataset Σ. The results for each algorithm are averaged across all annotation steps of each evaluation repetition, and averaged again across all 10 random repetitions. The plot shows the results for the following algorithms - the random metric presents recommendations in random order and acts as a control to show how "easy" each dataset is to recommend for:

(i) Random metric (Random)
(ii) PCA face proximity content metric (PCA)
(iii) Context metrics combined by CombSUM (Context CS)

Figure 13. Mean average precision of recommendations to casual annotator across datasets



(iv) Context metrics combined by factor analysis (Context FA)
(v) Context and PCA metrics combined by CombSUM (Context+PCA CS)
(vi) Context and PCA metrics combined by factor analysis (Context+PCA FA)

It can be seen that while PCA performed rather poorly on the data, Context FA and Context CS performed similarly, both with initial average precision above 40%. PCA has boosted Context+PCA CS while pulling Context+PCA FA down: this shows that, for this selection of metrics at least, CombSUM can perform at least as well if not better than factor analysis, even if for metrics in general it may be a biased approach.

This disparity in performance could be due to the number of factors selected by the factor analytic approach for each person. While factor analysis texts warn against trying to label the common factors (Kim & Mueller, 1978) at a glance two factors loosely emerged for most people: one with strong "social" relationships to the corepresentation and creator metrics, the other with a strong "geographic" relationship to the spatio-temporal metric. On the other hand, some people with stronger "technological" device copresence relationships had three. However, these comments are not founded on rigorous analysis and in fact contravene the intention of factor analysis by attributing too much meaning to the emergent factors: this is identified as an area that deserves further study.

While the average precision of all algorithms stays fairly flat across recall, there are noticeable drops in precision where recall is 1/6, 1/5, 1/4, 1/3, 2/5, 1/2, etc. This is intuitive due to there being between 1-6 people depicted in each image in the dataset (see Table 2, Section 7.1). The most severe drop at 1/2 recall has significance since the sharpest accuracy drop is likely to occur between the first and second recommendations. This drop also has a trivial element since there are significantly more photographs with one or two people in the dataset than three or more (Figure 9, Section 7.1), and therefore most incorrect (as well as most correct) recommendations were made on these photographs.

Figure 14. Breakdown of the mean average precision of the individual context metrics for dataset 1a

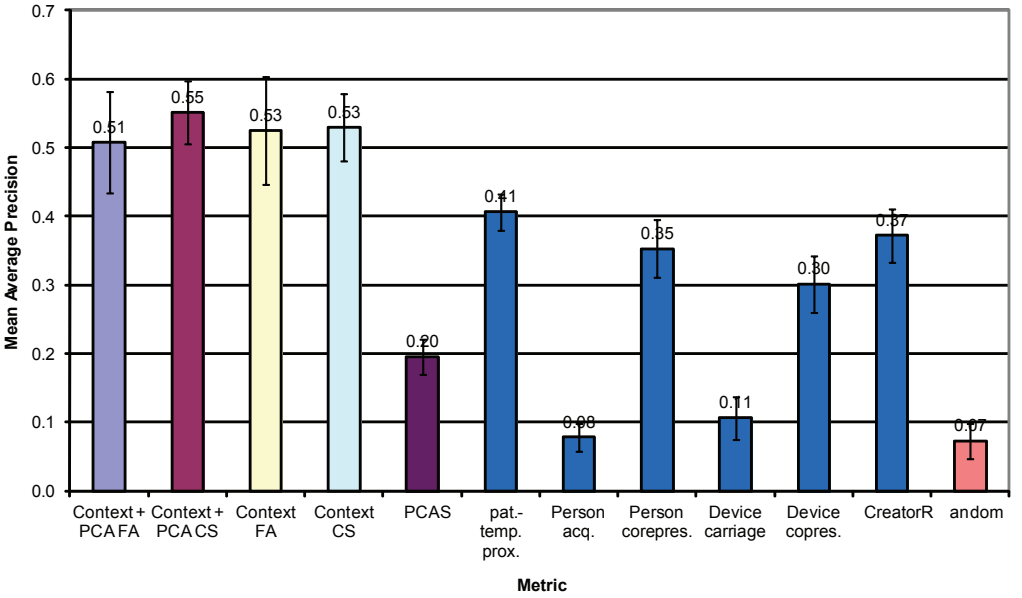


Figure 13 further summarises the results for all recommendation algorithms for the remaining evaluation datasets. The figure displays Mean Average Precision (MAP) of each evaluation run; this is simply the area under each precision-recall plot, as seen in Figure 12 for the plots for dataset Σ. As well as the mean, the standard deviation of the average precision across the 10 repetitions is given by the thin errors bars.

In general, CombSUM slightly outperformed factor analysis. Dataset 1 gave the best results; as seen in Section 7.1, this dataset contained the most observations (photographs), and therefore a larger pre-annotated training set as described in Section 7.2. Dataset 3 gave the worst results; as seen in Section 7.1, this dataset consists of photographs split over several distinct events, many events having just a single photograph. This demonstrates that the context-aware recommendation algorithms perform poorly in the absence of a certain amount of training data from each event. The results for the largest, consolidated dataset Σ suggest that overall and on a larger scale CombSUM outperformed factor analysis and that PCA did not significantly improve either algorithm's accuracy.

Finally, we analysed which of the individual metrics were the most accurate when used on their own. This evaluation may additionally help implementers to choose which individual metrics to use given constrained resources. To provide an unbiased view of the performance of the device carriage and device copresence metrics, this evaluation was only carried out on photographs from dataset 1a; these were taken by a camera phone with the ACRONYM mobile device-based camera application installed and so support the full range of context metrics. The breakdown of the MAP and standard deviation for each metric is compared in Figure 14 along with the combinations and random control for comparison.

The MAP results for individual context metrics suggest that the spatio-temporal proximity, person corepresentation and creator based ones are the most accurate. It is noteworthy that the implicit person corepresentation and device copresence metrics were more accurate than their explicit person acquaintance and device carriage counterparts; this may well be due to

a lack of social networking and device ownership statements in the dataset. Either way this demonstrates the robust nature of these implicit, real-world metrics over their explicit, digital world counterparts. The results also suggest that both combinations of context metrics are indeed more accurate than any single context metric alone. These results suggest that CombSUM is again the slightly more accurate combination method when ambient space context is available to the given context metrics and face recognition is used, with a MAP of 55% compared to 51% for factor analysis.

7.5. Discussion

Factor analysis has already been used to combine context-based and content-based cues to support the prediction of faces in photographs (Davis, Smith, Canny, Good, King, & Janakiraman, 2005). The major difference is that this previous work uses binary-valued statements about photographs as their observed variables while we use continuous-valued metrics taking into account relationships between various contextual recall cues. As detailed in Section 3, their approach has 2 disadvantages. First, by only considering direct properties of a photograph, it ignores direct relationships that instances may have with each other independent of any photographs, e.g., people may state that they know each other. By using context metrics derived from the user's information space, we take advantage of the rich contextual knowledge that is available for mining on the social Semantic Web.

The second disadvantage of the approach of Davis, Smith, Canny, Good, King, and Janakiraman (2005) arises from their binary modelling of data. For example, the binary model cannot use the geographic proximity of photographs, only the boolean value of whether a photograph belongs to an artificially created geographic cluster or not. The binary approach requires photographs to be clustered geographically into an arbitrary number of clusters (100 in their case), the granularity of which may be too coarse for many applications and which does not scale well. For example, photographs taken in Ireland and Germany may be indistinguishable geographically under this model and may well both be included in the arbitrary 'Europe' cluster.

Importantly, the authors of Davis, Smith, Canny, Good, King, and Janakiraman (2005) report some significantly more accurate results than this paper: they report 60% initial precision for a combined PCA+context method. As well as using a larger training set, it is important to note that in their evaluation the authors report that they manually tailored the training set in several ways with constraints that introduce several assumptions (Davis, Smith, Canny, Good, King, & Janakiraman, 2005, Sec. 4): "For each face, 8 photos were taken at random and 4 photos were selected manually for the training set. Manual selection was done to insure a sufficient number of visible faces in the training set. We will automate this process in future work. Each photo contained images of 1 to 4 people. The training gallery contained 2-4 images of each subject on average."

These assumptions would tend to boost the accuracy of all results, but specifically face recognisers. While the dataset used in Davis, Smith, Canny, Good, King, and Janakiraman (2005) is not publically available, the boosts provided by the larger size and manual selection of the training set can be measured by comparing the results of the PCA recogniser: while this algorithm was included in both evaluations, it achieved 43% accuracy in Davis, Smith, Canny, Good, King, and Janakiraman (2005) but a maximum of only 22% in the evaluation presented here (dataset 2). Since the PCA method is identical, this 20%+ boost in performance may be attributed to the training/test data. This assumption of manually chosen training data is outside of the original algorithm, not repeatable in the real-world without universal user labour and contributes to a smudging of the performance of the algorithm with the performance of the data. Since these assumptions cannot be guaranteed in actual use, we did not introduce similar assumptions to our evaluation but rather allowed the algorithms to be compared on a robust and realistic playing field of randomly

selected, 10-set cross-validated training sets, as described in Section 7.2

Furthermore, it should be noted that throughout their evaluation, Davis, Smith, Canny, Good, King, and Janakiraman (2005) only give the best results numbers: initial precision. They do not average their results for entire lists of recommendations using MAP, as we do here. Finally, to conclude a direct comparison with Davis, Smith, Canny, Good, King, and Janakiraman (2005), their context-only approach (without content/face recognition) achieves sub-50% MAP (initial precision 50%, dropping off for further hits), whereas our context-only approach achieves 53% MAP (dataset 1a).

Similarly to Davis, Smith, Canny, Good, King, and Janakiraman (2005), the context analysis approaches of Naaman (2005) have been combined with content analysis and extended by O'Hare and Smeaton (2009). In particular, they also extend the PeopleRank algorithm, and their spatial and temporal proximity metrics take a similar manually-chosen window approach to the spatial proximity method of Naaman (2005) and Davis, Smith, Canny, Good, King, and Janakiraman (2005). While the naming of their separate temporal proximity and spatial proximity metrics may appear similar to our single spatio-temporal proximity metric, they in fact operate quite differently. As detailed in Section 6.1.1, we use continuous spatial interpolation over time to calculate a single proximity value estimating how close a person was at a given time, whereas Naaman (2005) and Davis, Smith, Canny, Good, King, and Janakiraman (2005) use time and space windows of arbitrarily-chosen granularity to come up with two different scores which they later attempt to combine.

O'Hare and Smeaton (2009) extend this latter approach with Jelinek-Mercer smoothing over layered windows. For example, for spatial proximity, they use 100km windows layered on top of 1km windows layered on top of cooccurrence within single photographs. Our patented spatio-temporal proximity approach is shown in our evaluation to be our most successful individual metric, which complements the conclusion of O'Hare and Smeaton (2009) where their best performing individual metric was the one that incorporated temporal proximity.

Of note, O'Hare and Smeaton (2009) report relatively better results than reported here for face recognition when compared against context-based approaches (reporting 1-hit rates of 45% compared to 55% respectively in their approach while we report MAP of 20% compared to 53% in our approach). However, this is only after a substantial amount of dataset and evaluation method tweaking. Initially, in their face recognition approach they "position, scale, and rotate each face to create a normalized face image" before running PCA/ICA (O'Hare & Smeaton, 2009, Sec. IV.B.1). We do not, reusing the approach of Davis, Smith, Canny, Good, King, and Janakiraman (2005) to let PCA perform on its own on the original faces as seen in the photographs. This would contribute significantly to any relative disparity in results reported for face recognition.

Furthermore, while the dataset used in O'Hare and Smeaton (2009) is not publically available, they describe several major ways in which they manually tailor it for the evaluation of their approach. Since these would again tend to blur the line between dataset and evaluation method and significantly impact any evaluation results, we detail each below. It should be noted that since our approach is intended to work at Web scale with UGC data, and not on curated data, we must be able to deal with (and therefore should evaluate on) real untampered data. For a detailed discussion of the non-trivial problems with data cleanliness on the Semantic Web, see Hogan, Harth, Passant, Decker, and Polleres (2010).

Firstly, and similar to Davis, Smith, Canny, Good, King, and Janakiraman (2005) as described, (O'Hare & Smeaton, 2009, Sec. V) manually tailor their dataset to be a sample of good training data from the real-world population: "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 1 summarizes the nine individual personal photo collections used."

Secondly, they further manually tailor the dataset by only including in the test set those people with the most available training data, introducing the following unfounded and uncited assumption: "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 the user is not interested in less popular faces.". Based on our own observations, this would have a major impact on performance. Since it is completely unsupported by evidence from the real world, and is non-repeatable in actual use, we cannot and do not make such an assumption.

Thirdly, they continue to bias their results with weights learned specifically for each user's dataset, never providing an evaluation of their approach over their dataset as a whole (O'Hare & Smeaton, 2009, Sec. V): "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." On the other hand, our approach is intended from the ground up to automatically learn its weights for each person represented in multimedia on the social Semantic Web, and does not require an artificial ringfence to be erected around any one users' collection.

Fourthly, in their evaluation of their smoothing methods, (O'Hare & Smeaton, 2009, Sec. VI.A) make trial-and-error manual optimisations in the arbitrary sizing of their smoothing layer window sizes and in the arbitrary number of layers. This is outside of their presented algorithms and again is not repeatable in actual use without human intervention: "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."

When considering these latter points on manual blurring of the actual approach with the dataset and with the evaluation method, it can be reasonably inferred that the approach of O'Hare and Smeaton (2009) would fare significantly worse on an even playing field of untampered data from the real world using a consistent and repeatable evaluation method. In particular, this would help to explain any relative disparity in results for content-based approaches - which require plenty of good training data (Section 2) reported between this paper and O'Hare and Smeaton (2009).

Similarly, Zhao, Teo, Liu, Chua, and Jain (2006) take several measures to circumvent the face pose, illumination and expression problems for their content-based approach: "Face detection is first applied to detect the near frontal faces. To alleviate the pose problem, eye detection is used to rotate the faces so that the two eye[s] are horizontal. The eye detector is trained with AdaBoost, which has been used successfully in face detection. To overcome the illumination problem, we employ the generalized quotient image for delighting. Finally to tackle the pose problem, we use translation and rotation to generate more training faces for three views, i.e., left-view, front-view and right-view, for each person." They then use pseudo 2DHMM models for each of the three views to recognise faces. They also add body recognition to their content-based approach.

Zhao, Teo, Liu, Chua, and Jain (2006) also support their content-based approach with social context using 1) the PeopleRank algorithm amongst their own algorithms based on 2) global popularity (a rude probability for each person, identical for each photo), 3) event co-occurrence (based on clustering of events) and 4) temporal reoccurrence (based on temporal clustering of photos). While they do not provide an evaluation of their context-based approach standing alone, they report that it improves the recall of their content-based approach when used in combination, but it does not improve precision noticeably.

Finally, while Zhao, Teo, Liu, Chua, and Jain (2006) report 70% initial precision results for their combined Face+Body(A)+Context approach, they also encounter sharp drop-offs across recall. They do not report overall performance numbers for MAP, but these can still be estimated from the PR plot in their Figure 4. They achieve about 30%, 20% lower than our best context-based approach's MAP results. This can be roughly calculated by taking the area under their plot: 0.70 (precision) * 0.82 (recall) / 2 (rough triangle shape) = 0.29.

7.6. Repeatability

This section summarises a strategy to help the community have a reusable and extensible suite of data so that evaluations are repeatable and comparable. The strategy aims to convince that this, related and future work can be evaluated with relevance to Web-scale (in the context of UGC). In the spirit of linked open data we have gone above and beyond the state of the art in making our data as well as our approach available for others. Below, we clearly address the issue of repeatability of our approach and evaluation giving our best view on:

- How the data is reusable
- How others can compare the results
- How others can build upon or participate in developing a larger evaluation set

Data Reusability

We have made every attempt possible to publish as much data as we are legally allowed to with the permission of the data owners. We provide our entire dataset of metadata, minus "human-readable" personally identifying content like names and faces as is required practice for personal data. This goes beyond efforts from state of the art related work which mostly (including those cited) do not even attempt to share the data used for repetition. The data retains all other strings, e.g., placenames and all anonymised *Person* concepts retain unique identifiers as well as links to other concepts. This dataset is valid RDF, classifies as Linked Data as it reuses URIs dereferencable to existing data out there and is ready to go into any RDF store as-is.

We provide a detailed description in Section 7 of the makeup of the dataset including a detailed statistical breakdown of concepts and relationships, along with distribution of faces per person and photo. This includes graphical representations of the data. This gives the reader deep insight into the dataset, before they have even clicked the link to download it into their RDF store.

Result Comparability

Our zero-assumption evaluation approach is completely repeatable without human intervention, and involves no tailoring of the data for our or any one approach. We also use the completely repeatable and directly-comparable evaluation technique of Mean Average Precision. See Section 7.5 for a discussion and comparison with related work in this regard.

Data Extensibility

We have reused and integrated popular existing ontologies like FOAF, GeoNames and iCal where possible. See Section 5.1 for an overview and Monaghan (2008, Ch. 4) for a detailed description of the integrated ontology used. We also go beyond any related work and actually provide free of charge the same mobile device app we use to capture sensor data in Section 5.1. A detailed description of our algorithms is provided for anyone who wants to repeat or extend our work.

8. CONCLUSION

We have proposed an approach to media annotation that is social Semantic Web-aware and which taps the rich contextual data available thereon. This is a novel dimension beyond existing approaches that comes with its own

implications, and has resulted in the following contributions of this paper.

Robust recommendation algorithms: novel context-aware recommendation algorithms that require no manual tweaking or dataset tailoring.

Factor analytic combination method: unique application of personalised, continuous-valued factor analysis to combine metrics without bias. This alternative approach to handling sparse media annotation data automatically tailors itself to annotated individuals and is an innovation beyond Davis, Smith, Canny, Good, King, and Janakiraman (2005) or any other work we are aware of.

Media agnostic: our approach is designed from the ground up to be media-agnostic, and is generically reusable beyond any one media (e.g., photographs) and applicable to UGC multimedia at large (video, text posts, e.g., tweets, audio clips, etc.)

Dataset of context-aware photo metadata: the first publically available RDF dataset for context-aware media annotation evaluation.

Evaluation of algorithms for reuse: an analysis highlighting the most useful of the metrics and combination methods presented for the personal photograph annotation use case.

The automation provided by these contributions to multimedia annotation can alleviate the audiovisual data overload on users searching for or managing UGC multimedia. Face recognition and other content-recognition techniques can be supported by context-aware, data mining and social techniques. More specifically, the social Semantic Web domain provides powerful tools for mining and integrating distributed information to create new knowledge on the context surrounding multimedia resources and the people represented by them.

Mobile devices can be used to gather ground truth data about the ambient environment at the time of capture which can be reused to infer higher level contextual information. This information can help automate the annotation of media resources with cues useful for knowledge recall by users. This paper defines several context metrics which measure the strength of relationships between some of the key recall cues people use for managing UGC multimedia.

In particular, the context-aware ACRONYM approach can support the semi-automatic annotation of the people depicted in personal photographs with 53% mean average precision based on context alone. The same approach can increase the accuracy of state of the art face recognition from 20% to 55%. Additionally, it has been shown that implicitly measured context allows robust recommendations to be made when explicitly asserted statements are sparse. Furthermore, the context metrics can be combined to make recommendations that are more accurate than any that could be made by a single metric alone. The knowledge represented in completed annotations can then be integrated into the Semantic Web as portable, linked data describing the ever-increasing amount of UGC content.

ACKNOWLEDGMENTS

The authors would like to thank Robert Engels for his key input to the context metrics. The work presented in this paper has been funded in part by Science Foundation Ireland under Grants No. SFI/02/CE/I131 (Líon) and SFI/08/CE/I1380 (Líon-2) and the European Union under Grants No. FP6-2004-IST-NMP-2-17120 (AmI4SME) and FP6-027705 (Nepomuk).

REFERENCES

Adida, B., Birbeck, M., McCarron, S., & Pemberton, S. (2008). *RDFa in XHTML: Syntax and processing*. Retrieved from http://www.w3.org/TR/rdfa-syntax/

Adobe Developer Technologies. (2005). *Extensible Metadata Platform (XMP) specification.* San Jose, CA: Adobe Systems Incorporated. Retrieved from http://partners.adobe.com/public/developer/en/xmp/sdk/XMPspecification.pdf

Ahern, S., King, S., Naaman, M., Nair, R., & Yang, J. H.-I. (2007). ZoneTag: Rich, community-supported context-aware media capture and annotation. In *Proceedings of the Mobile Spatial Interaction Workshop at the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, CA. New York, NY: ACM.

Anguelov, D., Lee, K.-C., Gokturk, S. B., & Sumengen, B. (2007). Contextual identity recognition in personal photo albums. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN (pp. 1-7). Washington, DC: IEEE Computer Society. Retrieved from http://robotics.stanford.edu/~drago/Papers/cvpr2007.pdf

Barthelmess, P., Kaiser, E., & McGee, D. R. (2007). Toward content-aware multimodal tagging of personal photo collections. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, Nagoya, Aichi, Japan (pp.122-125). New York, NY: ACM. Retrieved from http://home.comcast.net/~pbarthelmess/ Publications/Photos/icmi259-barthelmess.pdf

Berners-Lee, T. (2006). *Linked data* (Tech. Rep.). Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Cyganiak, R., & Heath, T. (2007). *How to publish linked data on the web* (Tech. Rep.). Berlin, Germany: Freie Universität Berlin. Retrieved from http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/20070727/

Bojārs, U., Heitmann, B., & Oren, E. (2007). A prototype to explore content and context on social community sites. In *Proceedings of the International Conference on Social Semantic Web* (pp. 47-58). Retrieved from http://www.eyaloren.org/pubs/cssw2007.pdf

Breslin, J. G., Harth, A., Bojars, U., & Decker, S. (2005). Towards semantically interlinked online communities. In A. Gómez-Pérez & J. Euzenat (Eds.), *Proceedings of the 2nd European Semantic Web Conference: Research and Applications*, Heraklion, Greece (LNCS 3532, pp. 71-83). Retrieved from http://sw.deri.org/2004/12/sioc/index.pdf

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Proceedings of the International Conference on Compression and Complexity of Sequences*, Salerno, Italy (pp. 21-29). Washington, DC: IEEE Computer Society. Retrieved from http://www.cs.princeton.edu/courses/archive/spr05/cos598E/bib/broder97resemblance.pdf

Canny, J. (2002). Collaborative filtering with privacy via factor analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland (pp. 238-245). New York, NY: ACM. Retrieved from http://www.cs.berkeley.edu/~jfc/'mender/sigir.pdf

Cetina, K. K. (1997). Sociality with objects: Social relations in postsocial knowledge societies. *Theory, Culture & Society*, *14*(4), 1–30. doi:10.1177/026327697014004001

Choi, J.-Y., Yang, S., Ro, Y. M., & Plataniotis, K. N. (2008). Face annotation for personal photos using context-assisted face recognition. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, Vancouver, BC, Canada (p. 44). New York, NY: ACM.

Davis, M., Smith, M., Canny, J., Good, N., King, S., & Janakiraman, R. (2005). Towards context-aware face recognition. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, Singapore (pp. 483-486). Retrieved from http://fusion.sims.berkeley.edu/GarageCinema/pubs/pdf/pdf_89FB89A7-2534-412F-A815230DFD-B32CDC.pdf

Dey, A. K., & Abowd, G. D. (2000). Towards a better understanding of context and context-awareness. In *Proceedings of the Workshop on The What, Who, Where, When, and How of Context-Awareness at the Conference on Human Factors in Computing Systems*, The Hague, The Netherlands.

Dorai, C., & Ventakesh, S. (2003). Bridging the semantic gap with computational media aesthetics. *IEEE MultiMedia*, *10*(2), 15–17. doi:10.1109/MMUL.2003.1195157

Euromonitor International. (2010). *Global Market Information Database (GMID)*. Retrieved January 19, 2010, from http://www.portal.euromonitor.com

Fox, E. A., & Shaw, J. A. (1994). Combination of multiple searches. In *Proceedings of the 2nd Text Retrieval Conference TREC2 NIST SP 500215* (pp. 243-252). Gaithersburg, MD: NIST.

Girgensohn, A., Adcock, J., Cooper, M., Foote, J., & Wilcox, L. (2003). Simplifying the management of large photo collections. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction*. Amsterdam, The Netherlands: IOS Press.

Hasan, T., & Jameson, A. (2008). Bridging the motivation gap for individual annotators: What can we learn from photo annotation systems? In *Proceedings of the 1st Workshop on Incentives for the Semantic Web at the 7th International Semantic Web Conference*, Karlsruhe, Germany. Retrieved from http://km.aifb.uni-karlsruhe.de/ws/ insemtive2008/ hasan2008insemtive.pdf

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. doi:10.1145/963770.963772

Hogan, A., Harth, A., Passant, A., Decker, S., & Polleres, A. (2010). Weaving the pedantic web. In *Proceedings of the Linked Data on the Web World Wide Web Workshop*, Raleigh, NC. Retrieved from http://sw.deri.org/~aidanh/docs/pedantic_ldow10.pdf

InfoTrends/CAP Ventures. (2004). *Mobile imaging: Technology trends, consumer behavior, and business strategies* (Tech. Rep.). Retrieved from http://www.capv.com/home/Multiclient/ MobileImaging.html

Kim, J.-O., & Mueller, C. W. (1978). *Introduction to factor analysis: what it is and how to do it* (10th ed.). Thousand Oaks, CA: Sage.

Kirk, D., Sellen, A., Rother, C., & Wood, K. (2006). Understanding photowork. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 761-770). New York, NY: ACM.

Klyne, G., & Carroll, J. J. (2004). *Resource Description Framework (RDF): Concepts and abstract syntax*. Retrieved from http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

Lavelle, B., Byrne, D., Gurrin, C., Smeaton, A. F., & Jones, G. J. F. (2007). Bluetooth familiarity: Methods of calculation, applications and limitations. In *Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services Mobile Interaction with the Real World*, Singapore. Retrieved from http://www.medien.ifi.lmu.de/mirw2007/ papers/MIRW2007_Lavelle.pdf

Lieberman, H., Rosenzweig, E., & Singh, P. (2001). Aria: An agent for annotating and retrieving images. *IEEE Computer*, *34*(7), 5762. doi:10.1109/2.933504

Lindley, S. E., Durrant, A., Kirk, D., & Taylor, A. (2009). Collocated social practices surrounding photos. *International Journal of Human-Computer Studies*, *67*(12), 995–1004. Retrieved from http://research.microsoft.com/pubs/115365/IJHCS_Editorial.pdfdoi:10.1016/j.ijhcs.2009.08.004

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press. Retrieved from http://nlp.stanford.edu/IR-book/pdf/irbookonline-reading.pdf

Mitchell, J. (2011). *How many photos are uploaded to Facebook each day?* Retrieved June 26, 2011, from http://www.quora.com/ How-many-photos-are-uploaded-to-Facebook-each-day

Moghaddam, B., Jebara, T., & Pentland, A. (2000). Bayesian face recognition. *Pattern Recognition*, *33*(11), 1771–1782. doi:10.1016/S0031-3203(99)00179-X

Monaghan, F. (2008). *Context-aware photograph annotation on the social Semantic Web* (Doctoral dissertation, National University of Ireland). Retrieved from http://sw.deri.org/~ferg/publications/thesis.pdf

Monaghan, F., & O'Sullivan, D. (2007). Leveraging ontologies, context and social networks to automate photo annotation. In B. Falcidieno, M. Spagnuolo, Y. S. Avrithis, I. Kompatsiaris, & P. Buitelaar (Eds.), *Proceedings of the 2nd International Conference on Semantics and Digital Media Technologies*, Genoa, Italy (LNCS 4816, pp. 252-255).

Naaman, M. (2005). *Leveraging geo-referenced digital photographs* (Doctoral dissertation, Stanford University). Retrieved from http://infolab.stanford.edu/~mor/research/ naamanthesis.pdf

Naaman, M., Harada, S., Wangy, Q., Garcia-Molina, H., & Paepcke, A. (2004). Context data in georeferenced digital photo collections. In *Proceedings of the 12th International Conference on Multimedia* (pp. 196-203). New York, NY: ACM.

Naaman, M., & Nair, R. (2008). Zonetag's collaborative tag suggestions: What is this person doing in my phone? *IEEE MultiMedia*, *15*(3), 34–40. doi:10.1109/MMUL.2008.69

Naaman, M., Yeh, R. B., Garcia-Molina, H., & Paepcke, A. (2005). Leveraging context to resolve identity in photo albums. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*. (pp. 178-187). New York, NY: ACM.

O'Hare, N., & Smeaton, A. F. (2009). Context-aware person identification in personal photo collections. *IEEE Transactions on Multimedia*, *11*(2).

O'Toole, A. J., Phillips, P. J., Jiang, F., Ayyad, J., Penard, N., & Abdi, H. (2007). Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(9), 1642–1646. doi:10.1109/TPAMI.2007.1107

Pingdom. (2011). *Internet 2010 in numbers* [Web log post]. Retrieved June 26, 2011, from http://royal.pingdom.com/2011/01/12/ internet-2010-in-numbers/

Rodden, K., & Wood, K. R. (2003). How do people manage their digital photographs? In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Fort Lauderdale, FL (pp. 409-416). New York, NY: ACM. http://www.rodden.org/kerry/chi2003.pdf

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1349–1380. doi:10.1109/34.895972

Suh, B., & Bederson, B. B. (2007). Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition. *Interacting with Computers*, *19*(4), 524–544. doi:10.1016/j.intcom.2007.02.002

Tuffield, M. M., Harris, S., Dupplaw, D. P., Chakravarthy, A., Brewster, C., & Gibbins, N. …Wilks, Y. (2006). Image annotation with photocopain. In *Proceedings of the First International Workshop on Semantic Web Annotations for Multimedia*, Edinburgh, UK. Retrieved from http://www.image.ntua.gr/swamm2006/ resources/paper09.pdf

Turk, M. A., & Pentland, A. P. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86. doi:10.1162/jocn.1991.3.1.71

US Census Bureau. (2010). *International Data Base (IDB)*. Retrieved January 19, 2010, from http://www.census.gov/ipc/www/idb/6.1.5

Veltkamp, R. C., & Tanase, M. (2002). *Content-based image retrieval systems: A survey* (Tech. Rep. No. TR UUCS-2000-34). Utrecht, The Netherlands: Utrecht University. Retrieved from http://www.aa-lab.cs.uu.nl/cbirsurvey/cbir-survey/

Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, *23*(176), 88–93.

Viola, P., & Jones, M. (2001). Robust real time object detection. In *Proceedings of the IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, Vancouver, BC, Canada. Washington, DC: IEEE Computer Society. Retrieved from http://research.microsoft.com/en-us/um/ people/viola/Pubs/Detect/violaJones_IJCV.pdf

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive Psychology*, *18*, 225–252. doi:10.1016/0010-0285(86)90013-7

Wunsch-Vincent, S., & Vickery, G. (2006). *Participative web: User-created content* (Report No. DSTI/ICCP/IE(2006)7/FINAL). Paris, France: OECD. Retrieved from http://www.oecd.org/dataoecd/57/14/38393115.pdf

Zhao, M., Teo, Y. W., Liu, S., Chua, T.-S., & Jain, R. (2006). Automatic person annotation of family photo album. In H. Sundaram, M. Naphade, J. R. Smith, & Y. Rui (Eds.), *Proceedings of the 5th International Conference on Image and Video Retrieval* (LNCS 4071, pp. 163-172).

ENDNOTES

1 (http://www.dataportability.org/)
2 (http://www.skyhookwireless.com/howitworks/xps.php)
3 (http://www.geonames.org/)
4 (http://acronym.deri.org/downloads/midlet/acronym.jar)
5 (http://acronym.deri.org/)
6 (http://acronym.deri.org/schema#)
7 (http://vasc.ri.cmu.edu/idb/html/face/index.html)
8 (http://acronym.deri.org/datasets/)
9 (http://www.itl.nist.gov/iad/humanid/feret/feret_master.html)
10 (http://linkeddata.org/)

Fergal Monaghan is a Research Team Lead at SAP Research within the Business Intelligence (BI) Practice. He holds an Honours Degree in Electronic & Computer Engineering and a PhD from the National University of Ireland, Galway (NUIG) received while a researcher at the Digital Enterprise Research Institute (DERI). His research interests include context-aware systems at the intersection of society, the Internet of Things and Web of Data. Most recently his work focuses on decision rationale capture, tracking and change management via argumentation mining and reasoning. He has patented his work and published several times.

Siegfried Handschuh is a Senior Lecturer at the National University of Ireland, Galway (NUIG) and leader of the Semantic Collaboration research stream, as well as of the Semantic Information System and Language Engineering Group (SmILE), at the Digital Enterprise Research Institute (DERI). Siegfried holds Honours Degrees in both Computer Science and Information Science and a PhD from the University of Karlsruhe. He published over 100 papers as books and journal, book chapters, conference, and workshop contributions, mainly in the areas of Annotation and Authoring for the Semantic Web, Knowledge Acquisition, Information Visualization and Social Semantic Collaboration.

David O'Sullivan (PhD) is a Director of Research at the School of Engineering and Informatics, National University of Ireland, Galway. His research interests are in innovation management where he directs a number of industry sponsored research projects. His most recent projects include innovation management within SMEs and distributed innovation management across extended enterprises. David has a number of publications including books – Applying Innovation (Sage); Manufacturing Systems Redesign (Prentice-Hall); Reengineering the Enterprise (Chapman & Hall); and The Handbook of IS Management (Auerbach). David also works with leading organizations where innovation is a core value including IBM, Thermo King, Fujisawa, Hewlett-Packard and Boston Scientific.