Creating an application for automatic annotation of images and video

Adrian Matellanes, Alyson Evans, Burcak Erdal

MOTOROLA LABS,
Jays Close, Basingstoke,
RG22 4PD, UK
{adrian.matellanes, alyson.evans, burcak.erdal}@motorola.com

Abstract. This paper presents our experiences creating an application for automatic annotation of images and video. The paper focuses on how we have gathered user requirements and developed an application including state-of-the-art technologies for content analysis and automatic annotation, while making it highly relevant for the user.

"One of the best ways to persuade others is with your ears—by listening to them"
Dean Rusk

Introduction

In this paper we present our experiences in creating an application for automatic annotation of images and video. The work utilises user studies to focus technology development. We do not intend to give details of the internal workings of the application or its architecture; instead, we explain how we have gathered user requirements, how we have analyzed them and how we have transformed them to an application relevant for the user. We call this application Personal Content Services (PCS) application.

It is well known that consumers are gathering more and more digital multimedia content. Consumers capture content using their digital cameras, digital camcorders and mobile phones and store it on different devices. Consumers are beginning to store content in such quantity that it is becoming increasingly difficult for them to manage, find and, in the end, enjoy the videos and images they create.

For many years now research has been conducted in multimedia content analysis to help in these tasks of storage, search and retrieval of images and videos. Technology is advancing rapidly but there is still a long way to go before a fully automated solution is developed.

The PCS application not only contributes to solving the complex problem of multimedia content annotation but also addresses other important issues of multimedia content management such as content adaptation, facilitation of content flow between devices, content delivery, and privacy aspects of content.

The application presented deals with personal content; although work is ongoing within aceMedia to develop automatic annotation for a commercial content management application too.

While technology is not mature enough to provide a complete automatic annotation solution for consumers, and there are still many open research questions, it is important to address user needs early in the research and development process. This way the most relevant problems (for the user) are studied with higher priority and/or workarounds are put in place.

New technologies can be combined and integrated in many different ways but only some of them are meaningful for the users or give them sufficient benefit.

In our particular problem of automatic annotation of images and videos, user-centred design brings us users' perceptions of accuracy issues within annotation, i.e, which inaccuracies they can tolerate and, which they cannot and to what extent the user is willing to help improve annotations.

The structure of the paper is as follows. First we explain what we did to understand users and gather requirements from them. Then we give an overview of the application and how we addressed the user requirements. User evaluation results are presented before the conclusion and next steps.

Understanding users to gather their requirements

User-centred design is particularly important in the creation of consumer products. People have specific motivations associated with their activities, their behaviour is often goal driven and they have firm views about what they are prepared to use to help them to achieve something. To design well for people it is important to understand their motivations and the context within which they operate. User requirements can be gathered through semi-structured interviews with the relevant user group – in this case people who already have a digital photo and/ or video collection [11].

The semi-structured interview begins with an exploration of users' current habits and then moves to a discussion about unmet needs and the potential that they see a new technology has to offer them.

There are a number of reasons to find out about the current habits of users:

a. To uncover what they are trying to achieve - their goals and their motivations. Current motivations and goals are likely to remain stable and should be accommodated by the new application.

b. To find out if what they are currently trying to achieve could be made easier with the new technology in question.

c. To identify unmet needs that require new functionality that could be supported by the new technology.

In the second stage of the interview the potential of new technology is described to users and, with their minds fresh with the examination of their current habits, they are encouraged to be creative and to imagine how they could use the new technology to make their activity more efficient and more exiting.

From the interviews with users that were carried out in the aceMedia project it became clear that each user is now taking 500-2000 digital photos each year. These users currently use a basic form of annotation - adding titles to the most important items in their collection and storing them in folders with meaningful names. Tracking time is also very important to the users and they will include date in their folder labeling to support their current method of search and retrieval within their collection, which begins with estimating when they took an image or a video and remembering the occasion when it was taken. There are two problems in the current situation: users are not naming the majority of their photos and once the user names an image it is grouped separately in the folder from those that remain un-named, as per the automatic naming convention of current software. This changes the chronological order in a folder that users can refer to when they are trying to find an image. As a result when users are trying to find an image they still have to rely on examination of a number of individual thumbnails to view the contents of a folder until they find the photo they are looking for. From what we learned about the current situation and what users indicated they were prepared to do the following user requirements relating to annotation were identified.

- Users want to add a title and some notes to images of their choice – they would like to add time and date stamp, location and author/ creator
- Users want to annotate images with the names of people in the photo attached to the correct person
- Users want to apply one annotation to a number of images at once
- Users are interested in trying out automatic annotation of images
- Users would like the application to use words they have already used to label a small number of images to annotate a group of images
- Users want to be able to accept or reject automatic annotations

Currently in searching their collections users cannot make use of the automatically assigned filenames generated by the current software, as they are meaningless. Users expressed their ideas about how they would ideally like to be able to search within their collections that resulted in the following requirements.

- Users would like to search their collections using keywords
- Users would like to enter a sentence during their search if the results of the search will be accurate
- Users would like to search using ‘person’, ‘date’, ‘time’ and ‘event’ (then a non-event can also be discriminated).

Most users have a ‘visual memory’ and they do remember the visual content of certain images, especially their most memorable images.

- Allow the user to search for further images based on a nominated image (e.g. take one from a collection and ask the application for ‘more like this’)

Users also discriminate between landscapes and photos containing people, and they already try to search their collections for landscape photos vs. ‘people’ photos for a variety of reasons, e.g. one user wanted to search for landscape photos to print and put on his wall, another user wanted to find images of a specific person to prepare a montage for their wedding day.

- Allow the user to search for ‘landscapes’ i.e. no people or ‘small people’. Support the user in finding photos that contain people and those that do not contain people
- Users want to be able to search their collection for all photos of a named person.

Users think this would give a more comprehensive search result than is currently possible as they have to rely entirely on their memory to remember which folders had landscapes or pictures of a specific person in them.

This initial user requirement gathering activity revealed some very specific user requirements that have been used to focus the technology development within the project.

The Personal Content Services application

Integrating technologies: aceMedia

The aceMedia project [1] aims to create a framework for multimedia content management that combines state-of-the-art technologies in image and video processing, knowledge and semantics. The realization of the aceMedia content management framework is based on the Autonomous Content Entity (ACE), an object made of three layers: the content layer, the metadata layer that includes manual annotations as well as automatically created semantic metadata and the intelligence layer, a programmable layer that enables ACE pro-active behavior 14].

Within the system, content management is translated into ACE management. ACEs are created, manually and automatically annotated (updating its metadata layer), stored in the content repository, transferred from one device to another, and deleted.

Application description

The Personal Content Services (PCS) application presented here refers to the first version of a two cycle development. After performing user evaluation of the first version we will update user requirements and subsequently system requirements and application specification for the second and final version.

The PCS application implements the complete personal digital content lifecycle (with the exception of content creation, usually done with a digital camera, camcorder or mobile phone): creation of items and collections, manual annotation, automatic annotation, content storage, visual, textual and hybrid content search, transmission and deletion. See figure 1.

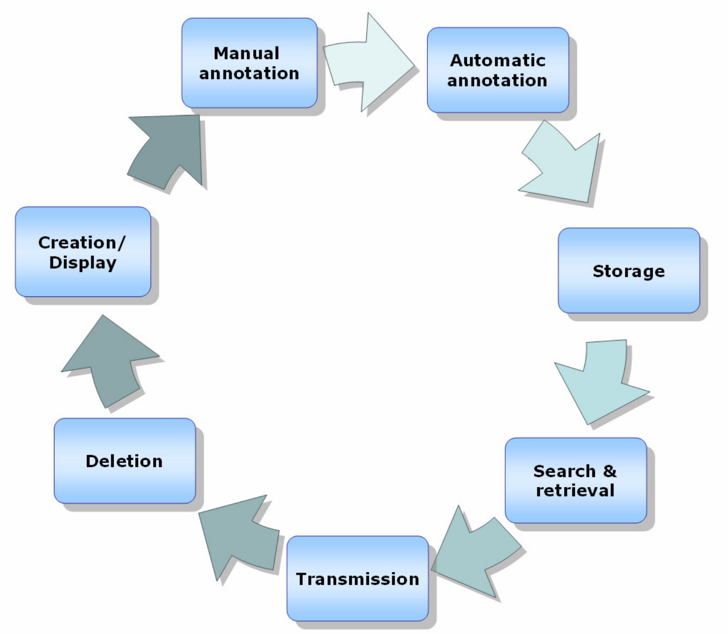

Fig. 1. Content lifecycle implemented by the PCS Application

The application is targeted at the personal computer (PC), mobile handset and set-top-box platforms. Not all requirements will be implemented on all targeted platforms, mainly for usability and performance reasons, and we describe in this paper the PC version.

From the architectural point of view, the PCS application sits on top of what is called the aceMedia framework or aceFramework [2]. This framework provides content management services to applications; specifically the aceFramework provides common networking and content repository interfaces (access to content and metadata) to applications as well as automatic annotation capabilities and search facilities.

Content analysis modules are in charge of extracting knowledge from the multimedia content. To fulfill user requirements and get appropriate images and video automatic annotation we must combine several content analysis modules. The aceMedia framework combines content analysis modules in two ways: on the software side they are integrated through an OSGi™ framework that helps homogenizing software interfaces and on the knowledge representation side they provide their results, i.e. automatically detected concepts, according to the aceMedia knowledge representation [4],[5],[6], allowing integration at the semantic level.

The aceFramework is also designed in a way that allows future content analysis modules to be plugged into it and enhance the automatic semantic annotation.

Having the content analysis and other basic services down in the framework, the application is in charge of implementing the User Interface and then background logic to support its functionality. This architecture decouples content analysis, storage and manipulation from the specific application targeted, e.g. another application for professional content management can be built onto the same aceFramework.

The PCS application is shown as the upper box in figure 2. In order to do so, the application must be deployed as an OSGi™ bundle [3]. Deploying an OSGi™ framework facilitates software components management as they can be installed, updated and removed on the fly; OSGi™ also reduces complexity by homogenizing component interfaces. This is particularly important in our case, as we integrate into the PCS application several automatic annotation modules. See figure 2.

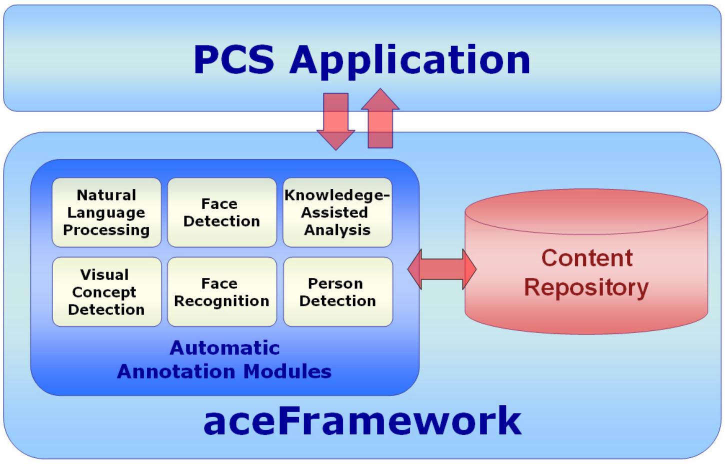

Fig. 2. Annotation modules of the PCS application

How the application addresses user requirements

aceMedia's approach to the complex problem of multimedia content annotation is to combine different automatic annotation tools and manual annotations.

In order to get richer and more robust metadata, we have combined several content analysis modules in the process of automatic annotation of images and video. The PCS application integrates the following modules: natural language processing, person and face detection, face recognition, visual content detection and knowledge-assisted multimedia analysis. These modules make use of the aceMedia knowledge infrastructure [4],[5] to automatically annotate images. As explained in [14], “the [aceMedia knowledge] infrastructure consists of three main modules, namely, the Core Ontology, the Multimedia Ontologies and the Domain Ontologies. The Core Ontology is an abstract ontology modeling basic concepts of the world, e.g. events and objects. The Multimedia Ontologies specify various aspects of multimedia content, e.g. its structure in terms of video frames or segments of images, and a basic set of relations to annotate the content, i.e. to relate concepts of the domain ontology to parts of the multimedia content. The domain ontologies finally model the domain of interest, such as beach holidays or motor sports, and contain the basic concepts and relations of the specific domain”.

Annotations are stored in the ACE metadata layer that resides in the content repository. The format used for these annotations is RDF/XML.

The typical annotation process is as follows. First, if there are manual annotations, then they are analyzed by the natural language processing module. It is important to perform this step at the beginning as its output can be useful in the next steps of the content analysis. If the manual annotation is rich enough, the analysis can also give indications of the domain ontology the system should use. Second, the visual content detector is run, this module will produce annotations referring to different “global” aspects of the content as landscape, cityscape, outdoors, indoors, etc. When the manual annotation has not given enough information (or is not present), the visual content detector can help in the identification of the domain ontology, otherwise, the domain ontology will have to be manually selected.

These first two steps, processing of manual annotations and visual content detection, address fundamental user requirements, as shown in the previous section, namely, “Allow the user to search for ‘landscapes’”, and “Users would like to search their collections using keywords”.

Third, we run the person and face detection modules [9][10]. After that, the face recognition module is run. Annotation of images and videos containing standing persons and faces (known to the system or not) is of utmost importance for the users as they usually enjoy images and videos that contain their friends and relatives.

Then we perform the knowledge-assisted analysis [6], which will label regions of the scene and associate them to concepts in the domain ontology.

So far we have aggregated the automatic annotations produced by the aforementioned modules, natural language processing, visual content detection, person and face detection, face recognition and knowledge-assisted analysis. All these annotations, which form part of the ACE metadata layer and deployed using RDF/XML, are finally analyzed by a reasoning module [8] that will get rid of ambiguities, merge regions or eliminate duplicates if they are present.

The knowledge infrastructure used, i.e. visual descriptors and domain ontologies to represent knowledge [7] and a suitable knowledge base to access it, provide the appropriate mechanisms to combine automatic annotations coming from different visual content analysis modules as well from the analysis of manual annotations. We must emphasize the importance of combining manual and automatic annotations. As shown in the previous section, the user is aware of current systems' limitations and is willing to participate in the annotation process if that is going to significantly improve system’s performance. We are confident that by exploiting users' manual annotations, the overall content annotation will significantly improve.

We have seen that user requirements actually reinforce the presented approach of combining different content analysis technologies. To fulfill user requirements we require natural language processing to analyze manual annotations; person and face detection and identification to find users' friends and relatives; visual content detection (detection of concepts such as landscape, cityscape, outdoors, indoors, sunset) and visual search to help the users find content based on their “visual memories”; knowledge-assisted multimedia analysis (which will label regions with concepts within the domain ontology) to let the user perform complex textual queries that include specific objects or regions, e.g., “a sailing boat in the sea during sunset”.

User evaluation

The first PCS prototype was submitted to a thorough evaluation, first by usability experts and then by representatives of the typical PCS users.

The goals of this user evaluation were:

- To validate the aceMedia PCS application concept, to validate more scenarios, and to further research context of use and typical user tasks;
- To evaluate how far the PCS application functionality meets user needs;
- To evaluate usability in order to guide the development of the second PCS prototype;
- To review the user requirements in PCS application.

The user evaluation process had three different parts: The expert evaluation was performed by 5 usability experts applying established heuristics to identify usability issues [12]. The user evaluation was based on the PCS prototype, and was performed with 18 typical PCS users (7 female, 11 male; ages 20-66). Finally, KANO questionnaires [13], which measures in how far features of a product meet user needs, were completed by the users. The majority of the PCS application features, including “automatic annotation“, have been rated “highly attractive” or “attractive”.

The PCS prototype was a very helpful basis to discuss aceMedia concepts with users, in addition, innovative features have been discussed and validated.

The evaluation results have validated the PCS application concept and majority of conceptual requirements. We also gained a better understanding of context of use, typical user tasks and user interface requirements to support innovative features of aceMedia.

With regards to automatic annotations, user requirements which were gathered from previous studies were confirmed. From the results, it is clear that users are willing to use automatic annotation.

For input of semantically rich text annotations, the majority of users thought that voice recording in combination with speech recognition would be an attractive solution, since they found text entry time-consuming and uncomfortable, especially on the mobile interface. This is an important new user requirement that will form part of the next development stage.

It was understood that users would like to be able to annotate multiple images at the same time and ready to provide a limited input to the application during automatic annotation. They also would like to have control over annotations created automatically, such as accepting and rejecting annotations, updating annotations for a few images etc. Users think that automatic annotation and organization of content would also support their use of search functionality in the application.

Conclusion

We have shown how a user-centred approach has helped us create an application for automatic annotation of images and video. Conducting user interviews to gather user requirements has helped us identify which issues are of most importance to the user, and we were able to discover to what extent the user is willing to help, with their inputs and annotations, in the complex problem of automatic annotation of images and videos.

At the time of writing this article, and as part of our user-centred approach, user evaluation has been conducted on the PCS application. User evaluation of the first prototype will result in an updated set of requirements that will be addressed in the second version. In general, the user evaluation results confirmed our approach to annotation of images and video.

Regarding how the combination of manual and automatic annotations improves system's performance, no specific results have been obtained yet. First experiments are being conducted that will provide useful feedback for the second version of the PCS application.

One of the main problems that our application faces is that of automatic selection of domain ontologies and their incompleteness. As we have explained, on the one hand, we intend to make use of the manual annotations and the visual content detection module to help automatically select domain ontologies; on the other hand, domain ontology incompleteness is an issue currently under research and we must continue our investigations, identifying solutions and workarounds to minimize the impact on the user.

One requirement that has surfaced repeatedly is that of voice recognition capabilities to help users annotate content with their voice. Although aceMedia does not include any voice recognition module, and work in this area is not planned within the project, the system architecture on which the PCS application is based, has a mechanism enabling "pluggable" analysis modules to be added later. The potential voice recognition module could easily be linked with current natural language processing, thereby improving the user experience.

Acknowledgements

This research was supported by the European IST project aceMedia under contract FP6-001765

The work presented here is possible thanks to a collaborative research and development. Partners involved in this work are: Philips Research, Queen Mary University of London, Fraunhofer-FIT, Universidad Autónoma de Madrid, Fratelli Alinari, Telefónica I+D, Dublin City University, Centre for Research and Technology-Hellas, INRIA, France Télécom R&D and University of Koblenz-Landau. In particular, we would like to thank, Barbara Schmidt-Belz and Fraunhofer-FIT, for their research on user requirements and user evaluation.

References

1. aceMedia. http://www.acemedia.org
2. A. Matellanes, A.May, F.Snijder, E.O.Dijk, A.Kobzhev, P.Villegas: An architecture for multimedia content management. EWIMT, London 2005.
3. OSGi Alliance. http://www.osgi.org
4. aceMedia Visual Descriptor Ontology. http://www.acemedia.org/aceMedia/files/software/m-ontomat/acemedia-visual-descriptor-ontology-v09.rdfs
5. N. Simou, V. Tzouvaras, Y.Avrithis, G.Stamou, S.Kollias: A visual descriptor ontology for multimedia reasoning. WIAMIS 2005
6. T. Athanasiadis, V. Tzouvaras, K. Petridis, F. Precioso, Y. Avrithis and Y. Kompatsiaris: Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. SemAnnot 2005 at the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, Nov. 2005
7. K. Petridis, F. Precioso, T. Athanasiadis, Y. Avrithis and Y. Kompatsiaris: Combined Domain Specific and Multimedia Ontologies for Image Understanding. 28th German Conference on Artificial Intelligence, KI 2005, Koblenz, Germany, Sep. 2005
8. N. Simou, C. Saathoff, S. Dasiopoulou, E. Spyrou, N. Voisine, V. Tzouvaras, I. Kompatsiaris, Y. Avrithis and S. Staab: An Ontology Infrastructure for Multimedia Reasoning. International Workshop VLBV 2005, Sardinia, Italy, 15-16 September 2005
9. M. Visani, C. Garcia and J-M. Jolion: Bilinear Discriminant Analysis for Face Recognition. 3rd International Conference on Advances in Pattern Recognition, ICAPR 2005, Bath, UK.
10. S. Duffner and C. Garcia: A Connexionist Approach for Robust and Precise Facial Feature Detection in Complex Scenes. 4th International Symposium on Image and Signal Processing and Analysis (ISPA 2005), Zagreb, Croatia, September 2005
11. H Beyer and K Holtzblatt. Contextual Design: Defining Customer-Centred Systems. 1998. Morgan Kaufmann
12. Nielsen, J., Usability Engineering. 1993: Morgan Kaufmann.
13. Sauerwein, E., et al. The KANO Model: How to delight your customers. in IX. International Working Seminar on Porduction Economics. 1996. Insbruck, Austria.
14. aceMedia 2005 public report. http://acemedia.org/aceMedia/files/document/aceMedia-Annual-public-report-2005.pdf