Nevenka Dimitrova
Philips Research

Human-Centered Multimedia: Culture, Deployment, and Access

Alejandro Jaimes
FXPAL Japan, Fuji Xerox

The main human-centered activities in multimedia include content production, annotation, organization, archival, retrieval, sharing, analysis, and multimedia communication. Within this group of activities I identify three key factors in the development of future computing systems:

- culture,
- integration of sensors and multiple media, and
- access outside the desktop by a wide range of users.

When someone enters any establishment in Japan, where I live, there's an immediate *irashaimase* greeting ("Welcome!") by store employees. But it's also often automatic: When I enter an elevator, approach an ATM, a photo booth, or a metro ticket vending machine, a sensor activates (see Figures 1 and 2), and I am greeted by multimedia cartoon characters that speak to me: "Going down." "All information will be displayed in English." The characters do not speak like computers—they speak like Japanese sales clerks (high-pitched voices with specific characteristics). They even bow.

It's interesting to consider these systems while thinking about multimedia. Although the interfaces are primitive and some of them are not really multimedia computing systems, they exhibit several important characteristics:

- they act according to the cultural context in which they're deployed;
- they integrate different types of sensors for input and communicate through a combination of media; and
- they're deployed outside the desktop and they're meant to be accessed by a diversity of individuals.

These examples highlight that ubiquitous computing (see the "Definitions" sidebar on p. 14 for a clarification of what I mean when I refer to the term ubiquitous computing and other terms in this article) is becoming a reality, and that the distinctions between the physical and the digital world are blurring, as are the distinctions between multimedia computing and computing. Consider the process for taking the train: I approach a machine that greets me. By pressing buttons I get either a paper ticket or an electronic card, which is then inserted into another computer that processes it and allows me to enter the system. What part of the process was digital, analog, or multimedia? It becomes clear why companies like Google want to organize the world's information without making any explicit reference to digital information.

Multimedia technologies are key in accessing the world's resources, particularly if we extend our notion of multimedia to what it really is—a combination of digital, analog, spatial, and sensory inputs and outputs. Perhaps the ultimate example of this is the automatic toilet (see Figure 3), a device installed in most restaurants and homes throughout Japan. Such toilets have sensors and produce music and various other

Editor's Note

Cultural setting is an intrinsic part of what we're trying to capture and use in multimedia systems. However, being in our own culture (both everyday culture and professional culture) we forget that multimedia interfaces and communication are culture-specific. This article gives some great insights that stem from diversity in countries (and cultures) as well as inside the interdisciplinary multimedia community.

—Nevenka Dimitrova

1070-986X/06/$20.00 © 2006 IEEE

sounds to aesthetically improve time spent in the bathroom.

The toilets are available in several makes and models. They include a shower spray, and some next-generation toilets offer to give users a personalized health analysis, sent directly over the Internet to their doctor for monitoring and before their scheduled check-up.

Cultural factors

Culture plays an important role in human–human communication because the way we generate signals and interpret symbols depends entirely on our cultural background. Multimedia systems should therefore use cultural cues during interaction (such as a cartoon character bowing when a user initiates a transaction at an ATM), as well as during analysis (such as algorithms to automatically analyze news broadcasts from different countries, meeting videos, or any other content).

The majority of work in multimedia analysis and interaction, however, assumes a one-size-fits-all model, in which the only difference between systems deployed in different parts of the world (or using different input data) is language. The spread of computing under the language-only difference model means people are expected to adapt to the technologies, imposed arbitrarily using Western thought models.

The problem is that this approach discourages creativity and leads not only to a technology divide, but also to a content gap between rich and poor countries (and/or upper and lower classes). Although our research community is international, the majority of international conference organizers and researchers that publish internationally work in developed countries, so it's rare to see highly innovative technical work that addresses local cultural issues. Most researchers in multimedia end up following a single model.

With such a narrow view, we lose robustness and more importantly, limit the access of technology and content to a small percentage of the world's population.[1,2] Technology is increasingly acting as the gateway to all basic resources, giving those that have it a real competitive advantage—for instance, what would happen to our food, water, healthcare, security, and communications if the computing infrastructure suddenly collapsed? Thus, the problem with a narrow view is not only limiting access to technology, but limiting access to all the benefits associated with it.

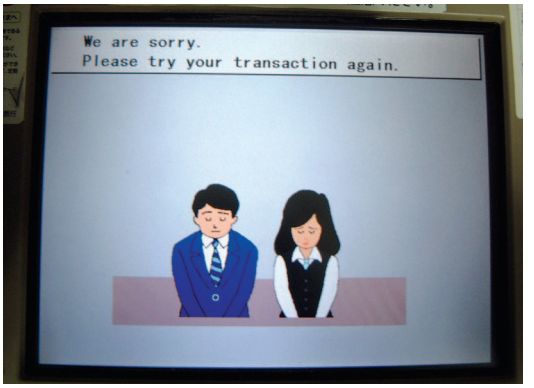

Figure 1. The ATM machine activates when it's approached. Unlike ATMs in the US or Europe, the customer first chooses the desired option and then inserts the card. Incidentally, people can buy tickets for concerts or movies at an ATM-like kiosk in any convenience store.

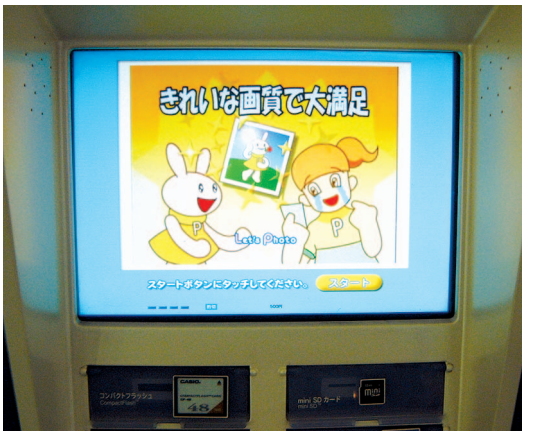

Figure 2. Japanese photo booths are popular. Users receive instructions (cartoon character that speaks) and can customize and modify their photos prior to printing. The screen turns on and the cartoon character bows as the machine is approached.

Cultural factors play a role in every aspect of computing, but let's examine two important areas in multimedia.

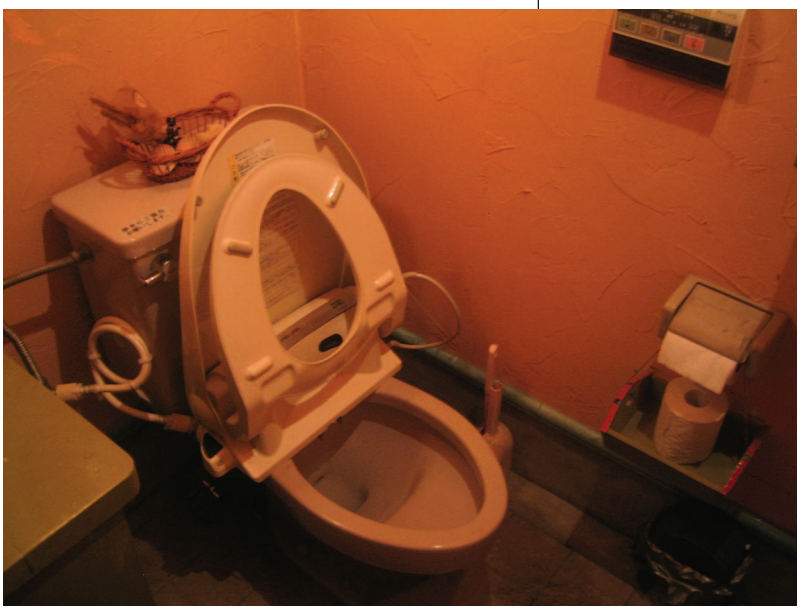

Figure 3. The toilet seat goes up automatically as the customer opens the door to the restroom. It closes when he closes the door. Many of these toilets warm the toilet seat, have a water jet spray, and play a range of melodies while in use.

Definitions

The following are definitions of some of the terminology reffered to in this article:

- *User-centered design*. This term has been used widely in the human–computer interaction (HCI) community for several years (a brief history is available elsewhere[1]). Although there's no general agreement in the field as to the focus areas of UCD (or the definition), the activities generally focus on understanding the needs of the user as a way to inform design.[1] Unfortunately, the UCD work in the HCI community has had little impact on the development of multimedia. I conjecture that this is in part because there isn't that much overlap between the two communities, and because in the short history of multimedia, interest has been mainly on analysis, retrieval, and production of multimedia content.

- *Multimedia*. This term, generally referred to as "a fusion of multiple types of data sources used to acquire, process, transmit, store, and utilize information,"[2] has been used in a wide variety of contexts. In fact, the term multimedia often includes many types of media, but not necessarily a combination of media. I view multimedia as a combination of digital, analog, spatial, and sensory inputs and outputs.

- *Human-centered computing*. A human-centered computing system is a system that involves any human activity (such as multimedia indexing for retrieval), or whose design parts from human models or gives special consideration to human abilities (such as human memory or subjectivity). This differs from user-centered computing, which assumes an explicit user (see also Flanagan et al.[3] for related definitions).

- *Multimodal system*. This refers to a system that responds to inputs in more than one modality or communication channel (for example, speech, gesture, writing, and others).

- *Ubiquitous computing and ambient intelligence.* There is much overlap between these two terms. Ubiquitous computing places emphasis on connecting devices to devices, people to devices, and people to people, anywhere and anytime. Ambient intelligence is based on similar principles, but seeks to make people's environments intelligent.[4]

References

1. J. Karat and C.M. Karat "The Evolution of User-Centered Focus in the Human-Computer Interaction Field," *IBM Systems J.*, vol. 42, no. 4, 2003, pp. 532-541.
2. L. Rowe and R. Jain, "ACM SIGMM Retreat Report," *ACM Trans. Multimedia Computing, Communications, and Applications,* vol. 1, no. 1, 2005, pp. 3-13.
3. J. Flanagan et al., eds., *Human-Centered Systems: Information, Interactivity, and Intelligence*, tech. report, Nat'l Science Foundation, 1997.
4. E. Arts, "Ambient Intelligence: A Multimedia Perspective," *IEEE MultiMedia*, vol. 11, no. 1, 2004, pp. 12-19.

Automatic analysis

The TREC Video Retrieval Evaluation 2005 (TRECVID; http://www-nlpir.nist.gov/projects/trecvid) set serves as a good example of how culture affects content production and automatic analysis techniques. In news programs in some Middle Eastern countries there are mini-soap segments between news stories. The direction of text banners differs depending on language, and the structure of the news itself varies from country to country, so a technique developed for news from the US is not likely to perform well in news from some of these countries. Thus, it's important to use varied data sets and consider cultural factors in multimedia production during automatic analysis (that is, we should use culture-specific computational models).

Without a doubt, the cultural differences in semantic content span every level of the multimedia content pyramid and the content production chain,[3,4] from low-level features (colors have strong cultural interpretations) to high-level semantics (consider the differences in communication styles between Japanese and American business people). The contrast is even greater in film: colors, music, and all kinds of cultural signals convey the elements of a story. Consider the differences between Bollywood and Hollywood movies (colors, music, story structure, and so on). Content is knowledge, and to make this knowledge widely accessible, we must develop culture-specific automatic analysis techniques.

Interaction

Significant research has been done in the field of human–computer interaction (HCI) on cultural factors. However, the general emphasis in HCI research has been on novel techniques or applications, so cultural issues haven't been explored in depth and little agreement exists on how or whether culture-specific techniques should be developed. Unfortunately, the multimedia community hasn't applied work in HCI and other fields where culture has been studied. Furthermore,

the majority of tools for content production follow a standard Western model, catering to a small percentage of the world's population and ignoring the content gap (see the World Summit Award—http://www.wsis-award.org—which is an initiative to create an awareness of this gap).

Our technical developments must be aligned with social and cultural developments, in a new model that embraces multiculturalism and sees the human impact of the technologies we're developing. One of the big research challenges, however, is how to translate what we learn about culture into technical approaches that can be applied computationally. There's a large gap between the work of sociologists, psychologists, and anthropologists, and the work of researchers and developers of multimedia technologies.

Although social scientists and nontechnical researchers have used technology for years, this use has been limited and has not been reciprocal (we start from the technology). The most promising research direction to address this problem is further integration between the development of multimedia systems and techniques and social studies. An excellent example is the work of Pentland[5] on social computing and attempts at benchmarking using culturally diverse data sets, such as the TRECVID 2005 set.

In academia, the importance of this has been recognized, and several programs integrate arts and engineering. In some universities graduate programs in human-centered computing have been established (for example, Georgia Tech). Collaboration with people in other fields is positive, but to really change the way we think about technology, it's necessary to create programs that integrate disciplines from start to finish. It's necessary to integrate approaches used in other human-centered fields and build new methodologies to incorporate cultural knowledge. Ontologies, multicultural knowledge bases, and techniques that use machine learning (because of their ability to be used with diverse training sets) can be possible starting points. Such integration will hopefully lead to new computational models and design methodologies.

To summarize, in our own interactions we recognize cultural differences and act accordingly (for example, company culture, social status, and so on), so why not develop highly adaptive or culture-specific systems? It's clear that using diverse content can help us gain a better understanding of how cultural differences are manifested in multimedia content production and that understanding how users of different cultures interact will make our systems more effective.

Figure 4. Some restaurants have wireless touch screens so customers can order when they're ready. The screen can be passed around the table just like a menu. The waiters only show up when the food is ready or when someone presses the waiter icon to call them.

Culture defines a large part of who we are, what we do, and how we interact with our environment, so it should be considered when designing multimedia systems, whether it's multimedia production, annotation, organization, retrieval, sharing, communication, or content analysis. Multimedia systems should include culture-specific models at every level of content and interaction in multimedia communication.

Integrating sensors and multiple media

New ATMs in Japan use biometric technology (palm and index readers) to verify identity, and some tour buses use GPS technology to automatically project tour-guide videos as the bus passes tourist attractions. Some restaurants use wireless touch screens so customers can order as soon as they're ready (see Figure 4).

What's interesting about these applications is how, even at primitive levels, information from networks and sensors is integrated with other types of inputs and outputs. Despite great efforts in the multimedia research community, integrating multiple media (in analysis and interaction) is still in its infancy. Our ability to communicate and interpret meanings depends entirely on how multiple media is combined (such as body pose, gestures, tone of voice, and choice of words), but most research on multimedia focuses on a single medium model.

Most of the systems I describe employ simple motion sensors that have been available for years (for example, in washrooms and for automatic

doors). Nonetheless, many novel applications have been developed integrating multiple sensors. In this context, multimodal interaction becomes a crucial part of a multimedia system. In the past, interaction concerns have been left to researchers in HCI—the scope of work on interaction within the multimedia community has focused mainly on image and video browsing. Multimedia, however, includes many types of media and, as evidenced by many projects developed in the arts, multimedia content is no longer limited to audiovisual materials. Thus, I see interaction with multimedia data not just as an HCI problem, but as a multimedia problem.

Our ability to interact with a multimedia collection depends on how the collection is indexed, so there is a tight integration between analysis and interaction. In fact, in many multimedia systems we actually interact with multimedia information and want to do it multimodally.

Two major research challenges are modeling the integration of multiple media in analysis, and in multimodal interaction techniques. Statistical techniques for modeling are a promising approach for certain types of problems. For instance, Hidden Markov Models have been successfully applied in a wide range of problems that have a time component, while sensor fusion and classifier integration in the artificial intelligence community have also been active areas of research. In terms of content production, we don't have a good understanding of the human interpretation of the messages that a system sends when multiple media are fused—there's much we can learn from the arts and communication psychologists.

Because of this lack of integration, existing approaches suit only a small subset of the problems and more research is needed, not only on the technical side, but also on understanding how humans actually fuse information for communication. This means making stronger links between fields like neuroscience, cognitive science, and multimedia development. For instance, exploring the application of Bayesian frameworks to integration,[6] investigating different modality fusion hypothesis[7] (discontinuity, appropriateness, information reliability, directed attention, and so on), or investigating stages of sensory integration[8] can potentially give us new insights that lead to new technical approaches.

Without theoretical frameworks on integrating multiple sensors and media, we're likely to continue working on each modality separately and ignoring the integration problem, which should be at the core of multimedia research. We need new mathematical models that truly integrate multiple sources and media, both in analysis and interaction, and a better understanding of how humans perceive and interpret multiple modalities.

Ubiquitous access

On one hand we can use mobile systems such as third-generation mobile phones to create and access multimedia content. On the other hand we have nonmobile systems (such as ATMs and ticket vending machines). An interesting characteristic of this second group of devices is that because they're deployed in public spaces, they're designed to be used by anyone (no need to read extensive manuals).

Computing is migrating from the desktop, at the same time as the span of users is expanding dramatically to include people who wouldn't normally access computers. This is important because although in industrialized nations almost everyone has a computer, a small percentage of the world's population owns a multimedia device (millions still do not have phones). The future of multimedia, therefore, lies outside the desktop, and multimedia will become the main access mechanism to information and services across the globe.

Mobile devices

Everyone seems to own a mobile device (see Figure 5)—there's a new wave of portable computing, where a cell phone is no longer a cell phone but rather a fully functional computer that we can use to communicate, record, and access a wealth of information (such as location-based, images, video, personal finances, and contacts). Although important progress has been made, particularly in ambient intelligence applications[8] and in the use of metadata from mobile devices,[9,10] much work needs to be done and one of the technical challenges is dealing with large amounts of information effectively in real time.

Developing effective interaction techniques for small devices is one of our biggest challenges because strong physical limitations are in place. In the past, we assumed the desktop screen was the only output channel, so advances in mobile devices are completely redefining multimedia applications. But mobile devices are used for the entire range of human activities: production,

Figure 5. Maybe nowhere in the world more than in Shibuya, a crowded, young area of Tokyo, are pedestrians bombarded with videos, images, and sounds. It has being said that the Shibuya Crossing (pictured) is the place with the highest density of mobile phone use in the world.

annotation, organization, retrieval, sharing, communication, and content analysis.

Shared resources

While mobile phone sales are breaking all records, it's increasingly common for people to share computational resources across time and space. As in the examples previously discussed, public multimedia devices are becoming increasingly common. In addition, it's important to recognize that—particularly in developing countries—sharing of resources is often the only option. Many projects exist on using community resources, particularly in rural areas, for education and other important activities. One of the main technical research challenges here is constructing scalable methods of multimodal interaction that can quickly adapt to different types of users, irrespective of their particular communication abilities.

The technical challenges in these two cases seem significantly different: mobile devices should be personalized, while public systems should be general enough to be effective for many different kinds of users. Interestingly, however, they both fall under the umbrella of ubiquitous multimedia access: the ability to access information anywhere, anytime, on any device. Clearly, for these systems to succeed we need to consider cultural factors (for example, text messaging is widespread in Japan, but less popular in the US), integration of multiple sensors, and multimodal interaction techniques.

In either case, it's clear that new access paradigms will dominate the future of computing and ubiquitous multimedia will play a major role. Ubiquitous multimedia systems are the key in letting everyone access a wide range of resources critical to economic and social development.

Research agenda for human-centered multimedia

Human-centered multimedia systems should be multimodal (inputs and outputs in more than one modality or communication channel). They must also be proactive (understand cultural and social contexts and respond accordingly), and be easily accessible outside the desktop to a wide range of users.

A human-centered approach to multimedia departs from user models that consider how humans understand and interpret multimedia signals (feature, cognitive, and affective levels), and how humans interact naturally (the cultural and social contexts as well as personal factors such as emotion, mood, attitude, and attention).

Inevitably, this means considering some of the work in fields such as neuroscience, psychology, communications research, HCI, and others, and incorporating what's known in those fields in mathematical models that we can use to construct algorithms and computational frameworks that integrate different media.

Machine learning integrated with domain knowledge, automatic analysis of social networks, data mining, sensor fusion research, and multimodal interaction[11] will play a special role. More research into quantifying human-related knowledge is necessary, which means developing new theories (and mathematical models) of multimedia integration at multiple levels. In particular, I propose the following research agenda:

- Create new interdisciplinary academic and industrial programs, as well as workshops that tackle the issues discussed and involve researchers across disciplines.
- Use culturally diverse data sets for common benchmarking and evaluation (including behavioral and multisensory data).
- Make software tools available (for annotation, feature extraction, and so on).
- Work on new human-centered methodologies for the development of algorithms in each of these aforementioned areas (not just systems).
- Focus research efforts on the integration of multiple sensors and media, with the human as the starting point (based on studies, theories in neuroscience, and so on).
- Instead of just considering the impact of technology, consider the social, economic, and cultural context in which it might be deployed (see Bohn et al.[12] for an interesting discussion on implications of ubiquitous computing).

Human-centered approaches have been the concern of several disciplines,[13] and some of the initiatives I've mentioned have been undertaken in separate fields. The challenges and opportunities in the field of multimedia, however, are great, not only because so many of the activities in multimedia are human-centered, but because multimedia data itself is used to record and convey human activities and experiences. It's only natural, therefore, for the field to converge in this direction and play a key role in the transformation of technology and human livelihood.

Conclusions

Many technical challenges lie ahead and in some areas progress has been slow. With the cost of hardware continuing to drop and the increase in computational power, however, there have been many recent efforts to use multimedia technology in entirely new ways. One particular area of interest is new media art. Many universities around the world are creating new joint art and computer science programs in which technical researchers and artists create art that combines new technical approaches or makes novel use of existing technology with artistic concepts. In many new media art projects, technical novelty is introduced while many issues are considered: cultural and social context, integration of sensors, migration outside the desktop, and access.

Technical researchers need not venture into the arts to develop human-centered multimedia systems. In fact, in recent years many human-centered multimedia applications have been developed within the multimedia domain (such as smart homes and offices, medical informatics, computer-guided surgery, education, multimedia for visualization in biomedical applications, and so on). However, more efforts are needed and we must realize that multimedia research, except in specific applications, is meaningless if the user is not the starting point. The question is whether multimedia research will drive computing (with all its social impacts) in synergy with human needs, or be driven by technical developments alone. **MM**

Acknowledgments

I'd like to thank Nicu Sebe for fruitful discussions on this topic, and the attendees of our recent tutorials on this topic (at ACM Multimedia, International Conference on Computer Vision, International Conference on Multimedia and Expo, and the Pacific Rim Conference on Multimedia). I would also like to thank Nevenka Dimitrova for insightful comments on this article.

References

1. E. Brewer et al., "The Case for Technology for Developing Regions," *Computer*, vol. 38, no. 6, 2005, pp. 25-38.
2. R. Jain, "Folk Computing," *Comm. ACM*, vol. 46, no. 4, 2003, pp. 27-29.

3. A. Jaimes and S.-F. Chang, "A Conceptual Framework for Indexing Visual Information at Multiple Levels," *Proc. Soc. for Imaging Science and Technology and Int'l Soc. for Optical Eng.* (IS&T/SPIE) Internet Imaging 2000, IS&T/SPIE, vol. 3964, 2000, pp. 2-15.
4. N. Dimitrova, "Context and Memory in Multimedia Content Analysis," *IEEE MultiMedia*, vol. 11, no. 3, 2004, pp. 7-11.
5. A. Pentland, "Socially Aware Computation and Communication," *Computer*, vol. 38, no. 3, 2005, pp. 33-40.
6. S. Deneve and A. Pouget, "Bayesian Multisensory Integration and Cross-Modal Spatial Links," *J. Physiology Paris*, vol. 98, nos. 1–3, 2004, pp. 249-258.
7. T.S. Andersen, K. Tiippana, and M. Sams, "Factors Influencing Audiovisual Fission and Fusion Illusions," *Cognitive Brain Research 21,* 2004, pp. 301-308.
8. C.E. Schroeder and J. Foxe, "Multisensory Contributions to Low-level, 'Unisensory' Processing," *Current Opinion in Neurobiology,* vol. 15, 2005, pp. 454-458.
9. S. Boll, "Image and Video Retrieval from a User-Centered Mobile Multimedia Perspective," *Proc. Int'l Conf. Image and Video Retrieval*, Springer LNCS 3568, 2005, p. 18.
10. M. Davis and R. Sarvas, "Mobile Media Metadata for Mobile Imaging," *Proc. IEEE Int'l Conf. Multimedia and Expo*, IEEE CS Press, 2004, pp. 936-937.
11. A. Jaimes and N. Sebe, "Multimodal HCI: A Survey," *Proc. Int'l Conf. Computer Vision* (ICCV 2005), N. Sebe, M.S. Lew, and T.S. Wang, eds., Springer LNCS 3766, pp. 1-15.
12. J. Bohn et al., "Social, Economic, and Ethical Implications of Ambient Intelligence and Ubiquitous Computing," *Ambient Intelligence*, Springer-Verlag, 2005, pp. 5-29.
13. J. Flanagan et al., eds., *Human-Centered Systems: Information, Interactivity, and Intelligence*, tech. report, Nat'l Science Foundation, 1997.