Let's Weave the Visual Web

Ramesh Jain
University of California, Irvine

Today, more than 75 percent of the people on our planet have mobile phones with cameras, which they frequently use to capture the moment, creating new documents (photos and videos). A major disruption is thus taking place in terms of how photographs are captured and the role they play in modern society. As the saying goes, "a picture is worth a thousand words," and this rise in digital photography offers the tantalizing possibility of joining each image's "words" with those of others, creating visual connections and conversations beyond anything we've yet seen or imagined.

Here, I take a look at the evolution of visual documentation and where we're headed. We've achieved nonlinear navigation for browsing text documents; now we need to do the same for photos and videos by building the Visual Web.

Revisiting History

Documents have been central to the sharing, storage, and dissemination of experiences, information, and knowledge.

The earliest form of knowledge representation was cave paintings, demonstrating the visual nature of knowledge representation. Then, as people started using languages, knowledge sharing moved to oral traditions. The next invention in this area was writing, followed by printed text.

Gutenberg's moveable printing press revolutionized the way people created, stored, and shared information and experiences.[1] This was arguably one of the most influential innovations of the last millennium. It was so impactful that many equate knowledge with literacy: "In democratic nations that subscribe to meritocratic principles, it is generally assumed that 'knowledge is power,' and that, to a large extent, knowledge is based on literacy."[2] Textual documents dominated the last millennium and resulted in enormous growth in science and technology.

Just a few decades ago, most of these documents resided in folders on various computers. People could share documents using tools such as FTP, but only if they knew the document's format, its location, and the path to that location. Then, around 1988, Tim Berners-Lee had the following thought: "Suppose all the information stored on computers everywhere were linked. Suppose I could program my computer to create a space in which anything could be linked to anything."[3] He went on to implement the World Wide Web, changing our world by redefining documents, data, information, and knowledge.

The Web grew so fast that discovering the most relevant documents, among too many seemingly related documents, became a solemn problem. Bookmarks were good for remembering but not for discovering or organizing documents, and with growth, they soon became too many to manage. Taxonomy-based organization of the list of documents was also overwhelming. By introducing page-rank-based search engines, Google solved that problem, and it's now relatively effortless to discover text-based information on the Web.

However, traditional text-dominant documents have limitations. Information is only discovered if it is available in text form. Furthermore, a specific language script makes it very difficult to share experiences among people trained in different languages. A text document in Chinese is not useful to an English or Hindi reader. Also, it is not useful to those who are (or close to) illiterate. Fortunately, recent progress in technology is likely to be transformative, leading to a new form of documents that is more visual and less dependent on language and literacy. This is an exciting time in defining how people share experiences, use information, and create knowledge.

The Transformation to Visual Information

Visual processing is dominant in the human sensory system. Most of our sensory receptors are in our eyes, and approximately 90 percent of information communicated to the brain is visual.[4] Although visual information and

1070-986X/15/$31.00 © 2015 IEEE

experiences have always been important for humans, technology to capture visual experiences in the form of photographs was invented only in 1826. Until recently, capturing photos was expensive and time consuming, so it played a secondary role in communicating information. That was until the end of the last century.

Figure 1. Smart cameras are universally used as a device to capture "visual documents." As these images exemplify, the user group is extremely diverse, ranging from young to old, urban to rural, and tech-savvy to tech-novice users, all over the world.

Creating and Consuming Photos

The 21st century started out differently for information and communication technology. Digital cameras were already making digital photos easy to capture, and they cost almost nothing. And then came the wave of phone cameras. With this, now most humans carry a camera ready to capture photos of anything even remotely interesting. Capturing, storing, and sharing experiences have now become easier than the corresponding operations using text.

This is a major change in the way that information gets created and consumed. At one time, people said that the "pen is mightier than the sword." Recent events have repeatedly shown that the mobile phone camera is mightier than all. Photos and videos are becoming the new documents. People use their smartphone camera to take and share notes more than they use pen and paper. And this phenomenon is global, as shown in Figure 1. This has resulted in many new approaches to the creation, storage, and use of these visual documents. Some popular applications and systems include Facebook, Flickr, Instagram, Snapchat, and GoPro, to name just a few, and for most social media systems, photos are the dominate presentation method for sharing experiences related to events.

In the last few years, the Web has become increasingly visual. Just a few years ago, people shared their status on the Web in a few words. Once photo capture became easy, these updates started including photos, and now text-based personal reports are being transformed into photo-based reports. Furthermore, a completely new style of capturing and reporting has become popular with the introduction of *selfies*. To add to all this, to capture the dynamics of events, cameras are emerging that continuously capture actions that people are likely to share. The trend is to share not only carefully produced video but also spontaneously captured short video. Videos of 5 to 30 seconds are being used for personal reporting.

In addition to the volume of photos, two other dimensions are important in the current transformation. The first dimension relates to what's being captured. The analog camera, as well as early digital cameras, captured only intensity values and thus were truly only a "visual capture" device. Increasingly, cameras use sensors such as GPS, accelerometers, and Gyroscopes, and they save information about the focal length, aperture, use of flash, distance to objects, and so on. So a camera is no longer just a photo-capture device; it is now a moment-capture device that can gather the photographer's intent as well. Your photo knows where it was taken, when it was taken, how the objects were captured, and more.[5]

The second dimension relates to recognition technologies. Computer vision has been slowly developing techniques to recognize objects and activities. The availability of a large volume of photos has helped develop better recognition techniques. Although these techniques are still in the early stages, their accuracy is now sufficient for the development of applications. Thus, when you capture a photo, your camera (via your smartphone) can use other sensors, knowledge from the Web, and its own computing power to really understand the photo and assign those 1,000 words to describe it.

Disrupting the Collection Model

Photo albums are "made for preserving impressions and launching memories."[6] An album thus has two important elements: collection and presentation. “Hard copy” photos of days past resulted in relatively static collections of photos. Once we created an album, it was used and distributed as a physical object. Typically, such albums contained relevant photos presented in a specific order. This was a successful model for albums, scrapbooks, and even traditional books.

The digital age is now disrupting that collection model. Digital photography is breaking the boundaries of the traditional album and revolutionizing it to make it more relevant to modern photography and social habits.

In his 1945 *Atlantic Weekly* article,[7] “As We May Think,” Vannevar Bush described how our memory works on context-dependent associations or connections. Revisiting photos can create *memory trails*: looking at one photo can evoke a thought that leads to another photo, which in turn leads to other photos, and so on. This process is typically triggered by some prominent attribute in a photo, which leads us to another photo with perhaps another kind of attribute that leads us further still. For example, I might be looking at a group photo that includes my friend John, which reminds me of the trip John and I took to Singapore; a photo from that Singapore trip might then remind of a wedding I attended; a photo from that wedding might feature a particular style of clothing that evokes my memory of another photo; and so on.

Given this emerging development of photos becoming the new document, it’s natural to think about how we might link all photos and videos and other types of information sources. What if I could create a space in which all visual data was linked and could also be linked to textual information?

Linking Photos

Hyperlinking might let us jump from one photo to another—perhaps located across the planet—based on different associations. A new paradigm will connect visual and other documents, connecting photos, videos, text, and all other information sources.

People have experience in linking textual documents. However, photos are fundamentally different in the following ways:

- Photos are inherently two-dimensional (representing a 3D space) and thus essentially nonsequential, such that we can traverse in any direction in the photo’s plane.

- Photos are comprised of pixels that could be grouped in an infinite number of ways; effective grouping lets us understand or interpret photos easily.

- Photos typically capture real, immediate moments in the real world; text describes such moments from the author’s viewpoint.

- Photos present objects, scenery, and people instantaneously; to represent the same in text often requires numerous words, which often fail to do justice to the image and the relationships within it.

- Photo content is significantly more subjective than text; people see what they want to see in pictures. Even the same person might see the same picture differently over time, depending on changes in his or her real-world situations and events.

- A photo’s semantics depend not only on its pixel values, but also on the context in which the picture was taken. Content and context are yang and yin; if content is king, then context is queen. Either content or context alone represents incomplete, often misleading—or simply wrong—semantics. Complete semantics emerge when the two elements are combined.[5]

Given these differences, associations and linking among photos will be different from those in text documents. In text, one can introduce a link simply by highlighting well-defined text representation. Also, in the original Web, there was only one type of link—a reference to other documents. In photos, links might have more diversity. Links might be to the complete photo, from and to objects in the photo, or even to the context of the photo.

Implicit links between photos might be created automatically using sensor-based technologies that analyze and capture context and content. For example, in a smartphone, sensors can automatically create links (or tags) based on location (GPS technology), time (the phone’s clock), or the photo’s subject (facial recognition). Advances in content analysis might soon allow recognizing and linking specific objects in photos to those in other photos or information sources. Explicit links between photos might be created when the user assigns additional information at capture time, such as his or her emotion, or the significance of the photo or its relevance to other photos. As a result, different types of implicit links (captured by sensors) and explicit links (assigned by photographers) will synergistically co-exist for both the complete photo and for individual objects within it. And all of these links will become an integral part of the photo to be used when and as desired.

Visual Navigation and Search

When hyperlinks were introduced in documents to create the Web, browsers were designed to go through a document. Once the number of documents increased, it was important to develop search engines to find most relevant documents. We are already facing a similar problem with photos.

Searching for a Photo

The problem of searching for a photo has been a topic of research for quite some time, and now the problem is even more important, given the large number of photos captured and stored by people—talk about searching for a needle in a haystack! Much of the research and development efforts in computer vision and multimedia content analysis have focused on developing techniques for object and concept recognition with the goal of finding, for example, "photos that contain a dog," "photos of a sunset," or "photos taken at Times Square."

Lately, recognition systems have been performing well enough to build impressive systems, such as Google Photos, which was released just as I finished writing this department. Google Photos is essentially a database of photos that makes it easier to search for photos based on queries using

- people (who is in the photo),
- other objects (what is in the photo),
- when the photo was taken (time), and
- where the photo was taken (location).

Clearly, progress in computer vision technology has made the first two items feasible, and the next two items are possible thanks to the availability of timestamps and GPS information for each photo. Powerful search environments can now be implemented by converting these items into tags and associating those tags with photos in the database. Recent announcements about search by Flickr and Google are establishing a clear trend in this direction.

However, photos are more than these four Ws (who, what, when, and where). Each photo is the result of an implied intent of the person taking the photo. This intent results in selecting the viewpoint that captures the four Ws and is clearly the result of a fifth W—that is, "why." In fact, all five Ws are essential to understanding the photo. The "who, what, when, and where" are explicit features of the photo, while the "why" is an implicit result of the intent and thus must be inferred from the context of the photo. Given the current state of technology, this is a difficult task. However, the person creating the document—capturing the photo—might also provide this information, which is what authoring environments did when linking textual documents. This is where explicit manual links might play an important role in linking visual information effectively.

Searching from a Photo

As discussed, associations are important in creating memories and building knowledge.[3,7] Suppose that I am looking at a photo, such as the one shown in Figure 2. The following questions might come to mind:

- Where was this photo taken?
- What was the event?
- Do I have other photos from the event?
- Who is the first person on the left in the photo, and who are the people sitting next to me?
- Do I have other photos of these people?
- I vaguely recall doing something interesting after the event … what did I do? Do I have photos of that event?
- Was the beer good? Where can I buy it in US?
- Those are nice name tags. Where were they made?

These questions are not related to the typical search in a browser. Rather, they are related to finding more information about interesting elements in the photo. When reading an online document, you might use hyperlinks to go to relevant documents, or you might perform a multiple-step Google search to answer certain questions. The problem is not finding the answer but finding it easily and effortlessly. The

Figure 2. When looking at a photo, many thoughts come to mind related to people, places, and events in the photo. These may be the associative links.

real issue is the associative search that was so eloquently championed by Bush.[7] The Web addressed this issue for text documents by hyperlinking some words with relevant documents using standard citation mechanism. Similarly, on the Visual Web, it should be possible to use nonlinear navigation to go to other documents of interest.

Building the Visual Web

Creating and growing a Web requires two essential user-centric operations: link creation and nonlinear navigation from the document. However, these operations must be reconsidered for photos. To understand and explore the related issues, a small company that I cofounded with Pinaki Sinha and Neil Jain has implemented an early prototype of the Visual Web in a smartphone app called Krumbs (http://krumbs.net).

The Krumbs App

In Krumbs, a photo is considered a visual experience of a moment, which is characterized by other contextual bits, called "krumbs," related to the moment. These krumbs are either directly captured by sensors in the smartphone or can be derived by using some sensory information with available knowledge sources on the Web or on the personal smartphone. Using these sources, a good amount of relevant information related to the moment can be inferred and associated with the moment.

As shown in Figure 3a, Krumbs lets a person express his or her intention or reaction to the moment using an emoji that acts as the camera button to capture the photo. Using the emoji and context, Krumbs creates an informative caption for the moment that can be used in sharing the experience. Each emoji is associated with an annotation and is used as a mechanism for the photographer to express his or her intent.

For each photo, the system creates four types of links: events, places, emojis, and people. Eventually, more classes of links will be added, such as objects. These links appear on each photo (see Figure 3b) and also through a steering wheel for searching the gallery of photos (see Figure 3c). These links are the associative mechanisms that result in nonlinear navigation of photos. This is the essence of the Web. Touching any of the links leads to the gallery containing all photos related to the link (see Figure 3c).

In this resulting gallery, the photos can be presented in chronological order or in the order of their K-score (compellingness score) as computed by the modified version of the photo-ranking algorithm.[8] The "events" links are semantic and represent real-life events, such as a wedding, dinner, meeting, hike, or trip. Using time, location, and other sources of information, the event links are inferred, sometimes with some help from the photographer. The "place" link offers more than latitude-longitude data, and it uses reverse geocoding combined with other sources of information. The "emoji" links represent the photographer's intent, while the "people" links relate to the people in a photo.

The search environment in Krumbs presents related events, places, people, and emoji information, enhanced with ontological information for places and events, as shown in upper right corner of Figure 3c. Events, places, people, and objects are suggested automatically by recognition algorithms but can be modified by the photo creator to express this as a link rather than an automatic classification. The current version of the system has links at the photo level, but links at the object level will soon be introduced as well.

Research Challenges

Many research and development challenges need to be addressed in building a Web that links all visual documents connected to other visual and nonvisual documents. Here, I focus on some of the key multimedia-related challenges (as opposed to the Web-systems issues, which are equally important but out of scope for this discussion).

Visual-document address. As each photo or video is captured, it should be automatically assigned an address, which will serve as the

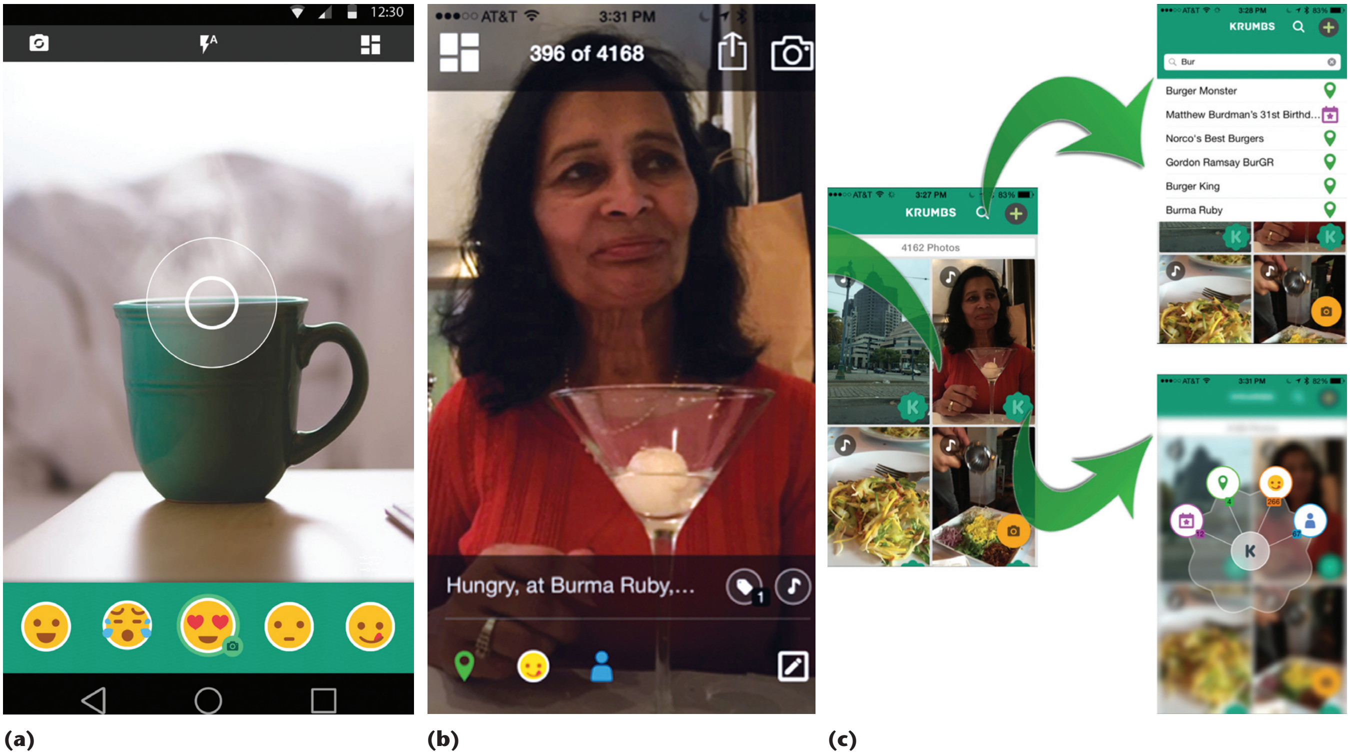

Figure 3. The Krumbs app: (a) Moment capture is done using an emoji that associates with the moment the intent/reaction of the person. (b) Each moment captured has the photo and its associated links. (c) The gallery resulting from each association includes a search icon that takes you to the search environment for Krumbs (top green arrow) and a steering wheel for finding how many moments are associated with each link (lower green arrow).

universal photo (resource) identifier. During its edit and transfers to other places, it is not clear when a visual document should be considered a new document and be assigned a new address. Also, how to maintain a link between the original captured document and its offspring's might be a tricky issue.

Transfer protocol. Once photos and videos have links, how should the hyperlinked documents be represented and transferred? The linking structure is likely to be very different, and will require its own protocol. Similarly, people might want to present the same photo using different linguistic or application information. The protocol should address these issues.

Authoring and presentation. HTML was designed to deal with the presentation of documents, and XML was designed to manage metadata effectively. Can they be modified for photos and videos, or do we need new tools?

Ranking. Page-ranking resulted in a powerful mechanism to assign importance to documents and present them in a form that people find convenient—a rank-ordered list. Given that millions of photos are taken at a place or of an object, how do we rank them based on their perceived relevance to people? What should go into a photo-ranking algorithm? When should this rank be computed? What is the role of the creator? Should the viewer be also considered in the ranking?

Contextual processing. Until recently, people referred to content as "the king," but now context is becoming more dominant through different sensors. (See, for example, *Age of Context,* by Robert Scoble and Shel Israel.[9]) Moreover, this trend seems to just be in its early stage. With the emergence of smartphones and now several wearable sensors that keep getting better as well as smaller, measuring anything and everything you want, context is increasingly becoming more powerful and important, determining both the relevance and role of content. Mobile phones have clearly established that context plays the primary role in searches. On desktops, the location was not important, but on mobile devices, it is a top priority.

Content analysis. Most research in computer vision and multimedia has addressed recognition using pixel values. With advances in computing and the availability of big data, much better tools are emerging for recognition. Learning techniques require training data, and the quality of the data determines the quality of the model that the learning machine will use. Furthermore, the model's design must account for the context in which the particular learning machine will operate—each machine works effectively in that context only. Consequently, during the training phase, the context is carefully coded in the learning algorithm.[10] How can these algorithms be made more general so they can work in a wider context?

Knowledge in the Visual Web. Researchers need to be aware of the photo as a linked knowledge element that could be very effectively associated with other knowledge elements to provide a more holistic view of events and experiences A photo thus captured might be, by itself, worth only a thousand words—but as a strongly multi-linked element that is part of thousands of thousand-word elements, it could be worth millions of words or more. How to harvest this knowledge from the Visual Web will be a very rewarding and challenging problem.

It will be interesting to see what kind of applications emerge for the Visual Web. As mentioned, searching from a picture (associative search) for objects, concepts, people, opinions, and such will be very useful. This is the primary mode of learning from and enjoying photos. Traditional image-search problems will be more effectively solved in this environment. But the more exciting applications will emerge from visual knowledge that transcends traditional language-dependent knowledge mechanisms. A farmer in India could contribute effectively to research in agriculture in collaboration with farmers in China and Brazil using only visual knowledge captured, collected, and shared with them. One of the simplest example applications, even without the linked photo mechanism, is PhotoSynth (https://photosynth.net), a huge collection of photos that could be used to virtually explore a monument.[11]

With the Visual Web, people will be able to explore everything from monuments to diseases, and from objects to activities, by collecting visual data from lots of people, in the proper context, helping them understand subtle and sophisticated concepts. This Visual Web will let people use their most important sensory processing ability to effectively address next-generation challenges. It will be the basis for universal experience sharing and knowledge creation, independent of traditional literacy levels. **MM**

Acknowledgments

This article benefited from discussions with several people, particularly those on my research team. Laleh Jalali was especially helpful. I appreciate all help that I received. The Web of People is possibly the most important web one can weave.

References

1. W. Ong, *Orality and Literacy: The Technologizing of the Word*, Routledge, 2012.
2. T.G. Sticht, C.R. Hofstetter, and C.H. Hofstetter, “Knowledge, Literacy, and Power,” San Diego Consortium for Workforce Education & Lifelong Learning (CWELL), Mar. 1997; www.coreknowledge.org/mimik/mimik_uploads/documents/40/KnowledgeLiteracyPower_1997.pdf.
3. T. Berners-Lee and M. Fischetti, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*, Orion, 1999.
4. R. Valdueza, “We Are 90% Visual Beings,” blog, 11 Jan. 2013; https://ernestoolivares.com/we-are-90-visuals-beings.
5. R. Jain and P. Sinha, “Content without Context Is Meaningless,” *Proc. Int'l Conf. Multimedia*, 2010, pp. 1259–1268.
6. V.P Curtis, *Photographic Memory: The Album in the Age of Photography*, Aperture and the Library of Congress, 2011.
7. V. Bush, “As We May Think,” *The Atlantic*, July 1945, p. 101 (reprinted in *Life* magazine, 10 Sept. 1945).
8. P. Sinha, S. Mehrotra, and R. Jain, “Effective Summarization of Large Collections of Personal Photos,” *Proc. Int'l World Wide Web Conf.* (WWW), 2011.
9. R. Scoble and S. Israel, *Age of Context: Mobile, Sensors, Data and the Future of Privacy*, Patrick Brewster Press, 2014.
10. V. Ramachandran and S. Anstis, “The Perception of Apparent Motion,” *Scientific American*, vol. 254, no. 6, 1986.
11. N. Snavely, S.M. Seitz, and R. Szeliski, “Photo Tourism: Exploring Photo Collections in 3D,” *ACM Trans. Graphics*, vol. 25, no. 3, 2006, pp. 835–846.

Ramesh Jain is a Bren Professor at the University of California, Irvine. Contact him at jain@ics.uci.edu.