Gang Hua
Stevens Institute of Technology

Visual Media: History and Perspectives

Thomas Huang, Vuong Le, Thomas Paine, Pooya Khorrami, and Usman Tariq

University of Illinois at Urbana-Champaign

In the early days of multimedia research, the first image dataset collected consisted of only four still grayscale images captured by a drum scanner. At the time, digital imaging was only available in laboratories, and digital videos barely existed. When more visual data become available, the problem of automatic image understanding emerged. In 1966, Marvin Minsky, the father of artificial intelligence, was assigned "computer vision" as a summer project.

Half a century later, the amount of visual data has exploded at an unprecedented rate. Images and videos are now created, stored, and used by the majority of the population. Consequently, image analysis has been transformed into a sophisticated and powerful research field, providing services to all aspects of people's lives.

From the early days to now, a major mission of multimedia research has been providing humans with visual information about the world. This includes capturing the scene's content into a computing system, enhancing the image's appearance, and delivering it to people in the most compelling way. However, sometimes the underlying metadata is arguably even more important than the content itself.[1] Visual data understanding research concentrates on either extracting the semantic meaning of the scene useful to user or assisting the user in interaction with computers.

In this historical overview, we will follow the great journey that visual media research has embarked upon by looking at the fundamental scientific and engineering inventions. Through this lens, we will see that all three aspects of media capturing, delivery, and understanding are developed surrounding the interaction with humans, making visual data processing a particular human-centric field of computing.[2]

Early Days of Visual Media: The Analog Era

The first visual media was captured almost two centuries ago when analog images were generated using cameras that recorded light on papers or plates with light-sensitive chemicals and stored with negative films, starting the long history of capturing methods. Figure 1 depicts the milestones during this era.

Together with the initial acquisition devices, delivery techniques also started to emerge. Analog images were then enhanced by optical processes and were printed on chemically sensitized paper. Analog optical instruments were also used for early image analysis methods such as frequency domain representation of images. With sinusoidal function basis, the Fourier transform offered a new perspective on how to observe and modify a visual signal. Based on space-frequency analysis and corresponding linear filters, algorithms were developed for applications such as compression, restoration, and edge detection.

Soon after early imaging was born, pioneers in the field realized that discretization of the analog visual signal could preserve most of the perceptible information while making operations much more convenient and efficient. This opened a new era of visual media processing: the digital era.

When Visual Media Became Mainstream: The Digital Era

You could say that the rise of digital visual media began in the 1950s when the first drum scanners digitized images. These scanners did not directly capture a photograph but instead copied preexisting photos by picking up the different intensities in a picture and saving them as a string of binary bits. Since the drum scanner, other image digitization devices/methods were created with marked improvements in quality and efficiency such as charge-coupled device (CCD) scanners and early TV cameras.

In the 1970s, digital color images attracted attention. The famous Lena image was scanned and cropped from the centerfold of the

1070-986X/14/$31.00 © 2014 IEEE

Published by the IEEE Computer Society

Multimedia Efficient Storage

Unfortunately, if someone tried to store the multimedia content naively it would require an extremely large amount of storage space. For instance, a two-hour standard definition (SD) video with a 720 × 480 pixel resolution and 24-bit color depth at 30 frames per second would take approximately 224 Gbytes in its original form.[3] The good news is that there is a lot of redundancy in the data, so we can store the same or a similar amount of information in much less storage space. Several influential works have proposed highly efficient compression techniques that make it feasible to process large numbers of images and videos computationally. Some examples include the discrete cosine transform (DCT), discrete wavelet transform (DWT), and motion compensation, which are used to compress JPEG files, JPEG-2000 images, and MPEG videos, respectively.

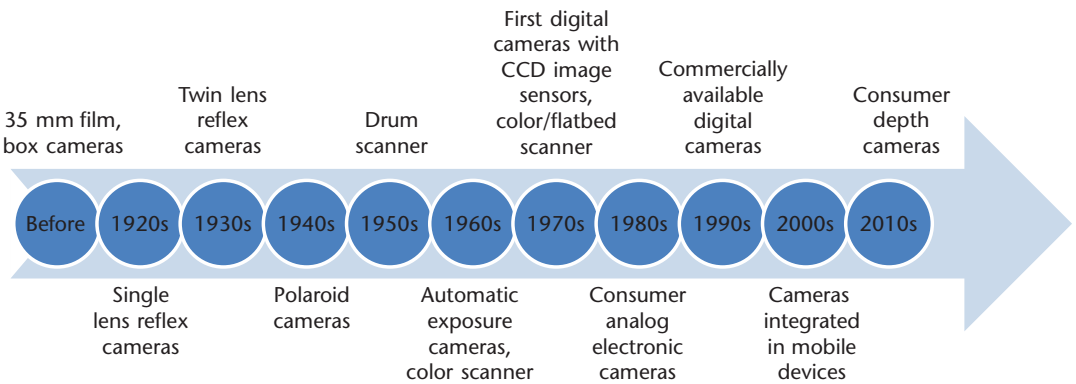

Figure 1. The evolution of multimedia acquisition over time.

November 1972 issue of *Playboy* magazine. It has since become widely used as a test object for evaluating image processing algorithms.

The next step in acquiring images came in 1990 with the introduction of the first consumer digital cameras. With the advances in image capture, and the ability to compress the images so they could be stored, it was suddenly possible for anyone to build photo collections of several hundred images. This is when consumer imaging devices found their way into people's everyday life.

In keeping up with the revolution of visual content capture, multimedia storage has also come a long way; from the magnetic tapes in the early 20th century to Blu-ray discs in the previous decade to holographic storage. Figure 2 gives a timeline of the evolution of multimedia content storage. These advances in multimedia capture and storage were first steps that would eventually facilitate large-scale multimedia data collection. (See the sidebar for more details.)

With more data efficiently created and stored, visual data understanding evolved to derive contextual information from visual data. For example, one may want to know if a particular object is present in an image or the identity of a suspect in a given mugshot. Most of these methods required constructing image models defined via machine learning. These models were traditionally obtained in a supervised manner where the labels of the training samples were given to a classifier or in an unsupervised manner using clustering algorithms. Some high-level inference tasks included multimedia retrieval, search, recommendation, and others for problems in human-computer interaction systems, such as gesture control, biometrics-based access control, and facial expression recognition.

Unfortunately, many of these data models worked poorly on raw image data resulting in the need for more sophisticated data representations. Image features were introduced as ways of extracting distinctive information from the image data and forming a compact vector or descriptor. Oftentimes, the most effective features used information gathered at the low-level such as edges or corners. In particular, scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG) feature descriptors helped construct summaries of the distribution of edges in the images and have led

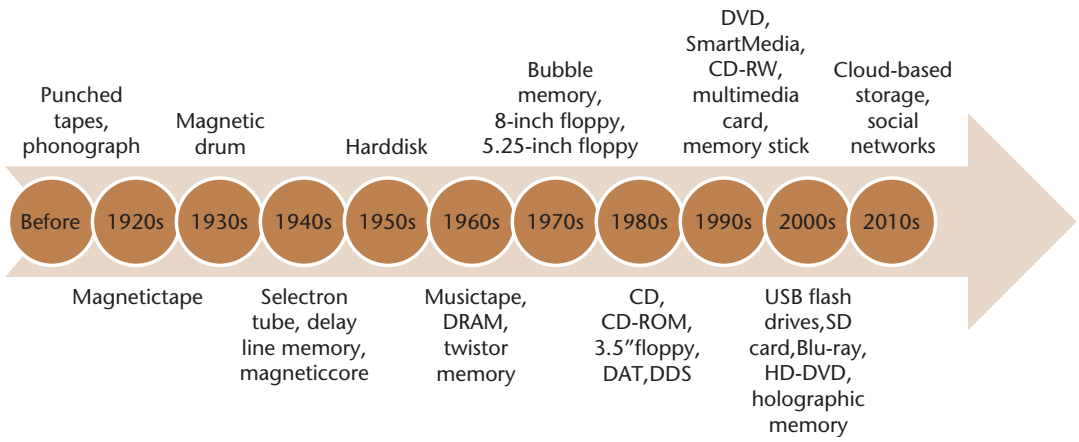

Figure 2. The evolution of multimedia storage over time.

to state-of-art performance in object detection and recognition. Slowly but surely, researchers began to learn which information is relevant in image data.

For the groundbreakers of digital visual understanding, model overfitting was a major problem. This issue came from the fact that the more sophisticated and expressive a statistical model became, the more training data it needed to reliably compute the model parameters. Luckily, in the late 2000s, the revolution in multimedia technology increased the ease of access for visual data, leading to an estimated 2.5 billion people around the globe owning digital cameras.[4] This number is predicted to soon surpass the world population. These seemingly unlimited sources of naturally captured and annotated data from everyday Internet users offers a potential remedy for data-lacking issues and leads the way to the Internet era of visual media research.

Ubiquitous Visual Media: The Internet Era

The Internet has made the process of collecting images convenient. Previously, the most expansive of photo sets (such as the US Library of Congress) were physically limited to thousands of images from hundreds of photographers. In contrast, a single social networking site such as Facebook can collect images from a billion active users, resulting in hundreds of billions of images in total. In 2012, Facebook had more than 300 million images uploaded daily, which is equivalent to 821 people taking 1,000 images daily for an entire year. Users upload images to these sites to share them with their friends and family. This builds upon the image capturing advances from decades before, making images easier to deliver. Thus, images are no longer artifacts you keep in your house or carry in your wallet. They are instantly sent and aggregated.

Media Understanding: Closing the Semantic Gap

Another benefit of social networking sites is that users often label their data. It is common for users to tag photos of their friends, describe the subject of videos, and curate their photos into albums of related events. This additional semantic information is another major difference between datasets used by research labs in previous decades and the resources available to media understanding researchers today.

Armed with this massive amount of media data and semantic labels, researchers can try new approaches to media understanding. But how? The secret is more parameters. Before, researchers had to limit the number of parameters in their statistical models due to overfitting. They favored dimensionality-reduction and linear models. But as datasets increased in size, overfitting became less of an issue, and bias became the limiting factor.

Now researchers are using increasingly more parameters in their media understanding algorithms, which can leverage larger datasets. A common way to increase the number of parameters is to extract features, learn an overcomplete mid-level representation, and then apply a linear classifier. This includes methods that discover object parts templates or that learn sparse dictionaries. These methods improved results in object recognition and object detection. Deep neural networks continue this trend with many layers of parameters that can model image statistics at multiple scales. Having more parameters to learn makes

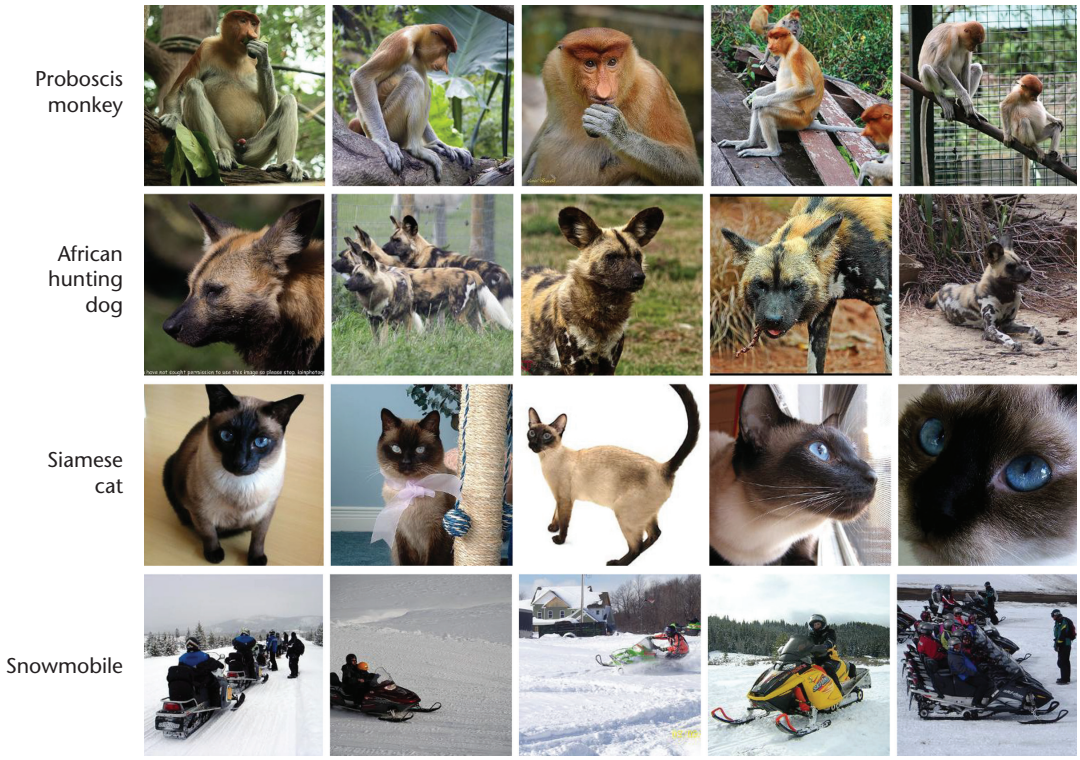

Figure 3. Examples of the variations among the images of each category in the ImageNet Large-Scale Vision Recognition Challenge dataset.

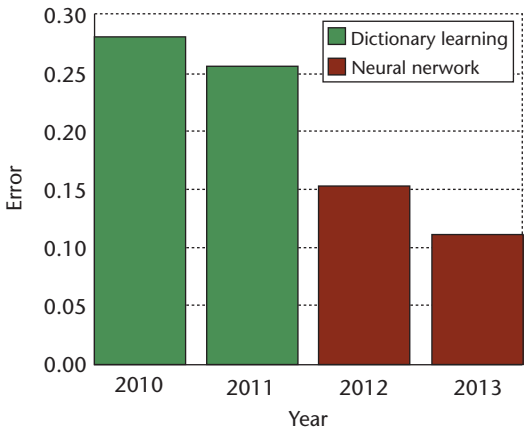

Figure 4. Improvement of recognition results for the ImageNet dataset over time and methods. The graph shows the best results per year.

these models flexible, allowing them to fit datasets with millions of images and generalize better to images in the wild.

One specific case of this is the 2012 ImageNet Large Scale Vision Recognition Challenge,[5] which resulted in the ImageNet dataset (see Figure 3 for examples from several categories). The dataset contains more than one million images in 1,000 object categories, with highly variated images in each category. Figure 4 shows the best recognition results on this dataset over recent years. With models consisting of millions of parameters and massive computational infrastructure, deep neural network models were able to get a 10 percent performance increase over dictionary learning methods.[6]

Future of Visual Media: What's Next?

Despite the rapid evolution of visual media research, it is hard to predict the future. In the following, we highlight some of the emerging trends and applications.

Wearable Gadgets and Moving to the Cloud

The word on the street is that wearable tech is the new chic. Wearable cameras will be the future trend. Some examples include, GoPro, camera watches, and Google Glass. GoPro has already positioned itself, especially in adventure sports, for example, where users attach a camera to their headgear and helmets. Camera watches, with data connection either through a cell phone or data network, will find applications in video telephony. Google Glass will enhance the user experience by bringing about new possibilities in communication and navigation. Such devices will once again change

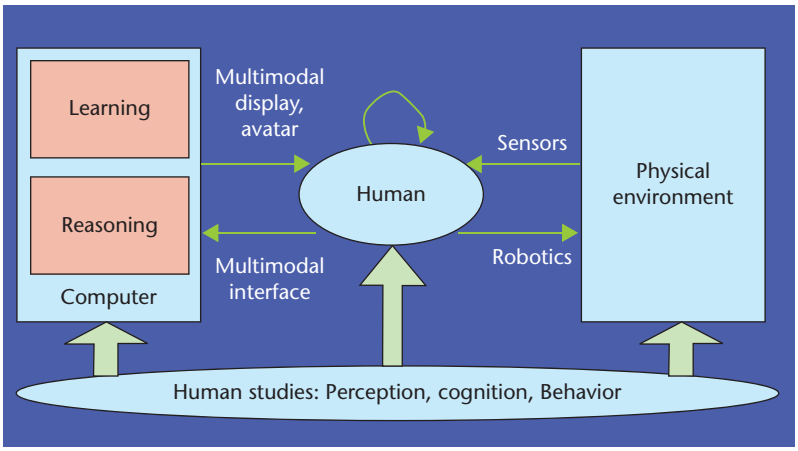

Figure 5. Overview of human-computer interaction systems.

how visual data is captured, delivered, and eventually understood. These devices will further swell the generation of multimedia content because users will begin recording their activities and surrounding environments, spontaneously or possibly even unintentionally.

In terms of storage, with the evolution of communications networks, another future trend in visual media seems to be moving to the cloud (cloud storage). In many cases, users record a movie or take a picture that may be then temporarily stored in the capturing device and then transferred to remotely located storage locations using data networks. The users are seemingly unaware of the underlying processes or the locations of their data storage—hence the name "cloud." Some of the big names in cloud storage include Amazon Cloud Drive, iCloud, and Dropbox.

Media Understanding: Closing the Intention Gap

With cheaper and better sensors, the exponential generation of visual media, and cloud storage come new challenges and directions in media understanding. These may include human-computer interaction (HCI), aesthetics, and search. The HCI field tries to find new ways for a computer to map humans natural movements to their actual intent. Figure 5 illustrates the general framework for most of the HCI techniques being used today. They may take input from various sensors, such as hand gestures, eye tracking, and voice commands. Some examples of such systems are the Microsoft Kinect with gesture control, eye tracking for the disabled,[7] and Apple's SIRI (www.apple.com/ios/siri/).

These systems may work well for a range of commands, but as the intended actions become complex, limited sensory inputs lead to ambiguous models and mappings. This ambiguity is commonly referred to as the "intention gap." HCI systems of today also fall short of the expectations of being socially aware. Some HCI systems may adapt to the user's affective state by analyzing their facial expressions, although this is generally in research scenarios. However, whether these systems are good enough for practical use in different situations is open for debate. In the future, we may see a tangible reduction in the intention gap. Some future research directions may be the fusion of various sensory inputs, building user profiles, and learning from the history of commands passed to an HCI system. HCI systems may also become more socially aware by assessing the cognitive states of the users and may well serve as automated agents such as avatars in computer-aided learning, gaming, or sales.

Another interesting area of potential future impact is automatic analysis of aesthetics in visual media. With readily available cameras, users now generate tremendous amounts of visual media. Gone are the days when you would have to think twice before taking another shot. A birthday party, a trip to Miami, or a graduation party may each generate hundreds of images. It is time consuming to sift through them all at once, however. Research in this area can potentially yield to software that would be able to select the most aesthetically pleasing shots.[8] Will they be the best? The answer to this question would lie in further understanding the semantic content.

There has been much work in attempting to understand the semantic content of the images, but we are nowhere close to human performance. How far in the future will there be any huge strides in this direction is still an open question. Visual media search and retrieval may see breakthroughs in the upcoming years. This could be aided by the availability of massive image datasets and the associated metadata in the form of user annotations. As outlined earlier, large datasets, even if unlabelled, help in learning complex statistical models that may perform and adapt well for various applications.

Another promising future direction for visual media search involves the social aspect. This draws its intuition from human-in-the-loop architectures. The underlying assumption in such an architecture is that explicitly incorporating human information during learning can lead to algorithms with high quality results.

One of the most common settings for human-in-the-loop techniques is image retrieval, where a user queries a certain topic and the computer returns a list of images it considers relevant. Then a feedback mechanism allows humans, or "workers," to iteratively refine the algorithm's results by choosing which of the retrieved images were correctly associated with a given concept.

With the evolution social media such as Facebook, we can now model social connections among various users. This development has led to the introduction of a social network component to both the construction and application of multimedia systems, as Figure 6 shows.

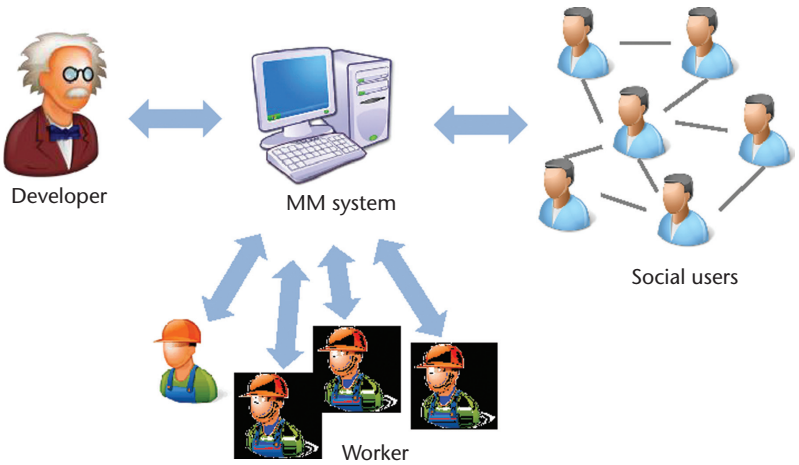

Figure 6. Modern human-in-the-loop system architecture. Media understanding now incorporates social media connections into multimedia systems.

Upcoming Applications

Visual media will have its tangible impact on various areas such as healthcare, HCI, and security in the upcoming years. In healthcare, we may see its applications in psychology and see screening tools being developed for autism, depression, or attention deficit disorders using multimodal automated affective analysis. We may also see the application of visual media for taking vital signs. One such recent success has been CardioCam, which finds a user's heart rate with a webcam by monitoring minute changes in skin color that correlate with blood circulation.[9] We may also see applications of visual media understanding in automated nursing. For example, if we are able to build a system that can accurately track body movements, then we could monitor if the exercises prescribed by a physiotherapist are being followed correctly.

In security, we may find intelligent systems being developed for anomaly detection and to track entities across surveillance cameras. The number of surveillance cameras installed in the public is ever increasing. For instance, the British Security Authority estimates that there are up to 5.9 million security cameras in the UK alone.[10] It is impossible to monitor every camera at every instant, but is it possible to find anomalies automatically? Humans are very good at finding abnormal events in a particular situation. However, event detection depends strongly on context. For instance, running may be normal on a beach but abnormal inside a bank. Automating anomaly detection is an open research problem and is related to semantic understanding. Apart from this, another interesting problem is tracking entities across multiple cameras so that law enforcement agencies can follow suspects or vehicles. We hope that the research community will continue to make strides in these directions while addressing future challenges.

Conclusions

Visual data has evolved tremendously since the first set of pictures were digitized. This evolution has occurred in all aspects of storage, delivery, and understanding. Recent years in particular have witnessed unprecedented growth, and we expect exciting new breakthroughs in the near future. Moreover, these aspects are quickly converging via the ubiquity of devices and algorithms leading to stronger interactions with humans. In the near future, the human factor will continue to be at the center of the field's development and will be the source of inspiration for the major working fields for media researchers and engineers in the next era. **MM**

Acknowledgments

Although this article lists only five authors, it was in fact a team effort, written by Thomas Huang and a number of his graduate students including Le, Paine, Khorrami, and Tariq in the Image Formation and Processing (IFP) Group. The group has weekly meetings as well as more frequent subgroup meetings, where many things are discussed, so the team knows each other's ideas and views well. In addition to the authors listed on the title page, the following students contributed to this article: Xinqi Chu, Kai-Hsiang (Sean) Lin, Jiangping Wang, Zhaowen Wang, and Yingzhen Yang.

References

1. M. Slaney, "Web-Scale Multimedia Analysis: Does Content Matter?" *IEEE MultiMedia*, vol. 18, no. 2, 2011, pp. 12–15.
2. A. Jaimes et al., "Guest Editors' Introduction: Human-Centered Computing–Toward a Human Revolution," *Computer*, vol. 40, no. 5, 2007, pp. 30–34.
3. R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, 3rd ed., Prentice Hall, 2008.
4. "Samsung Announces Ultra-Connected SMART Camera and Camcorder Line-Up Throughout Range," Samsung, 9 Jan. 2012; www.samsung.com/us/news/20074.
5. J. Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), 2009, pp. 248–255.
6. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems 25*, P. Bartlett et al., eds., 2012, pp. 1106–1114.
7. J. Davis, "Eye-Tracking Devices Help Disabled Use Computers," Texas Tech Today, 21 Sept. 2011; http://today.ttu.edu/2011/09/eye-tracking-devices-help-disabled-use-computers/.
8. S. Dhar, V. Ordonez, and T.L. Berg, "High Level Describable Attributes for Predicting Aesthetics and Interestingness," *Proc. 2011 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), 2011, pp. 1657–1664.
9. M.-Z. Poh, D.J. McDuff, and R.W. Picard, "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam," *IEEE Trans. Biomedical Eng.*, vol. 58, no. 1, 2011, pp. 7–11.
10. D. Barrett, "One Surveillance Camera for Every 11 People in Britain, Says CCTV Survey," The Telegraph, 12 July 2013; www.telegraph.co.uk/technology/10172298/Onesurveillance-camera-for-every-11-peoplein-Britain-says-CCTV-survey.html.

Thomas Huang is a Swanlund Endowed Chair Professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana–Champaign. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. Huang has a ScD in electrical engineering from Massachusetts Institute of Technology. Contact him at t-huang1@illinois.edu.

Vuong Le is a doctoral candidate at the University of Illinois at Urbana–Champaign. His research interests include computer vision and machine learning, especially 3D shape modeling and image recognition. Le has an MS in computer engineering from University of Illinois at Urbana-Champaign. Contact him at vuongle2@gmail.com.

Thomas Paine is a graduate student in the informatics program at the University of Illinois at Urbana–Champaign. His research interests include computer vision and large-scale deep learning. Paine has an MS in bioengineering from the University of Illinois at Urbana–Champaign. Contact him at tom.le.paine@gmail.com.

Pooya Khorrami is a doctoral candidate at the University of Illinois at Urbana–Champaign. His research interests include facial expression recognition, eye gaze estimation, deep learning, and video surveillance. Khorrami has an MS in electrical and computer engineering from the University of Illinois at Urbana–Champaign. Contact him at pkhorra2@illinois.edu.

Usman Tariq is a research engineer at Xerox Research Center Europe, Meylan, France. His research interests include image feature learning, domain adaptation, and face analysis. Tariq has a PhD in electrical and computer engineering from the University of Illinois at Urbana–Champaign. Contact him at usman.tariq@xrce.xerox.com.