Computer Vision: the Last Fifty Years

Linda G. Shapiro
University of Washington

Computer vision began just over fifty years ago with the work of Larry Roberts at MIT in the early 1960s, published in his dissertation and in a landmark article in 1965. This work covered, in some sense, all aspects of computer recognition of three-dimensional objects from image capture to edge finding (the first edge operator), line fitting, and model-based object recognition. By the time I entered the field, reading the literature in about 1972, quite a lot of work had been done. I was greatly impressed by the work of Adolfo Guzman, also from MIT, who developed a follow-on system that input line drawings extracted from images of polyhedral scenes with any number of objects in all kinds of spatial relationships and could list the separate objects in the scene. I also read the work in syntactic pattern recognition that influenced my early work in computer vision, for example, work by Rangaswamy Narasimhan in 1966 and Max Clowes in 1969 on syntactic approaches to recognizing patterns. My own dissertation was on what is called *structural pattern recognition*, putting me in the same class as King Sun Fu and his students at Purdue who worked on grammar- or graph-based approaches to recognizing objects. Much of this work still concentrated on line drawings. In fact, it has been said that the M.I.T. focus on line drawings actually kept back the field of computer vision from realistic image analysis tasks. Many low-level operators had been developed by this time, including multiple edge operators, the most commonly used being the Sobel-Feldman, and the well-known co-occurrence texture operator of Dinstein and Haralick, which is still heavily used today.

Other groups were developing, of course. Thomas Binford started a computer vision group at Stanford in 1970 and supervised many Ph.Ds there, some of whom did basic and important work in the field. Binford is best known for inventing the generalized cylinder representation for 3D objects, a model-based approach that did not rely on just line segments and led to the well-known work of Ram Nevatia and others who followed him on how to describe and recognize 3D objects in terms of this representation. Azriel Rosenfeld, sometimes called the Father of Computer Vision, started a computer vision group at University of Maryland in about 1963, wrote the first computer vision text in 1969, and initiated the first journal in 1972. He initially called the area picture processing and wrote numerous articles with his students. He was willing to take on any aspect of this area from low-level image processing to high-level recognition and constraint analysis. I visited the University of Maryland for a week in about 1977 and worked with him on the latter.

Another early and active group was that of Edward Riseman and Alan Hanson at the University of Massachusetts (about 1969). They called their system VISIONS, and its job was to interpret color images of complex outdoor scenes that contained houses, trees, bushes, grass, and so on. They designed a hierarchical representation that went from schemas and objects at the top levels down to regions, segments, and vertices at the bottom. The system was to be model-based and was to take into account all sorts of physical phenomena, such as occlusion, perspective, lighting and shadows. I was especially impressed with this full AI approach to the problem. While they never produced a single working system, they did produce many Ph.D. students with excellent dissertations on all aspects that brought out the kinds of work in segmentation and analysis that needed to be done to produce a full system. Their 1978 edited volume "Computer Vision Systems" included chapters by all of the early workers in the field. Some of the ones that stand out now as most important over time are Barrow and Tenenbaum's paper on recovering intrinsic scene characteristics from images, Davis and Rosenfeld's work on relaxation, and Marr's work on representing visual information including his famous *primal sketch*.

In the 1980s, we concentrated on solving useful problems related to real domains such as the military, space science, and industry. Robot vision became an important thrust, especially at such institutions as Carnegie Mellon, Purdue, and Stanford Research Institute (SRI International). The goal became to recognize parts for robots to pick up, manipulate and inspect. The concept of CAD-based vision was developed to describe model-based vision that took its input from real CAD models of manufactured objects, instead of expecting the vision researchers to model the objects they wanted to recognize. The difficulty here was in converting real CAD models to computer models with features that would actually show up on an image. The actual CAD models we obtained from Boeing for airplane parts had points and splines to represent the curves around each part. Furthermore, they came from a proprietary CAD system and had to be converted before we could even look at them. So we got very few of these real CAD models and ended up looking for parts on the floors of the Mechanical and Civil Engineering buildings across the street and creating our own models.

The CAD-model-based-vision work led to how to best represent 3D objects for rapid recognition by computers. The concept of an aspect graph was defined; it was a graph that would show every possible view of an object in terms of the visible features, which were usually line segments or surfaces. Several well known people, such as Kevin Bowyer and Jitendra Malik, worked on this problem. While this led to a number of theoretical papers, the graphs were too big to use in practice, and not all features showed up in the images. We called our version of aspect graphs *view classes* and defined them to be the major views of the object; the features that showed up in a single view class were similar but not identical. This led to our *relational distance* measure that could determine how similar were two views and allowed clustering of the views of an object into its major classes. When an image was analyzed, a voting procedure developed by Mauro Costa could be applied to find the correct view class of the correct object to recognize and localize it. Early work was extended to probablistic versions and parallel versions; locations of light sources were also taken into account. Other well-known work in CAD-model based vision was done at CMU by Katsushi Ikeuchi and at Purdue by Avi Kak. This work took vision into robotics and led into 3D vision from range data.

Parallel algorithms for computer vision had been active for some time from Steve Tanimoto's pyramids to the massively parallel but low-level MasPar Computer (company founded in 1987) and Danny Hillis's Connection Machines (company founded in 1983) which could be programmed in Lisp. We developed systems for the latter two and postulated algorithms for such systems in general and for a dataflow machine that our lab actually built. But all of these went away, because every year Intel put out faster and faster chips, and the big parallel machines were too expensive

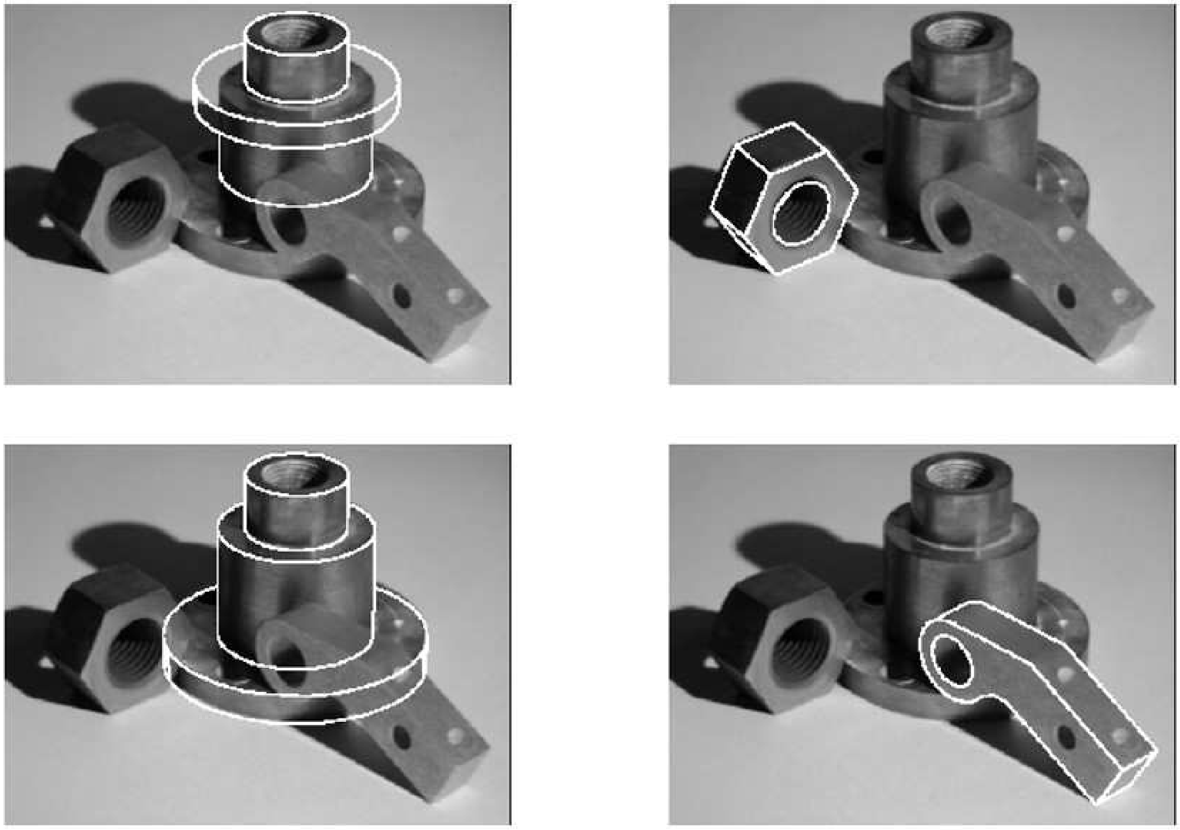

Figure 1: Images showing results from the RIO system in which the outlines of hypotheses for 3D object models are projected onto the 2D images. The projection in the upper left is an incorrect hypothesis that will later be ruled out in the verification step.

when the small new ones could do just as well.

Also very popular in the 1980s was the area of video analysis or motion. Hans-Hellmut Nagel (with Ramesh Jain) actually led the way in 1977 with his work on analyzing a live video stream from a TV camera that pointed down from his third-floor office to the street below in Hamburg. The goal of the work was to separate non-stationary from stationary scene components. A seminal paper by John Roach and Jake Aggarwal in 1980 tackled the problem of determining the three-dimensional model and movement of an object from a sequence of two-dimensional images by solving a system of nonlinear equations. Video analysis has gone a long way since these humble beginnings. The main work today is on recognizing objects, structure (3D) from motion, and tracking objects for surveillance. Depth cameras are now available as inexpensive units, such as the Microsoft Kinect system.

Another aspect of computer vision that began its rise in the 1990s and is still important today is content-based image retrieval. The first well-known system was IBM's QBIC: Query by Image and Video Content. It was a full research system, but usable through a web demo, that could retrieve images by color, texture, and shape through a graphical query language. The color histograms employed by QBIC are still the most common means of image retrieval today and probably the most useful for just appearance similarity. Content-based retrieval morphed into multimedia retrieval in which multiple different modalities could be retrieved. My former student Andy Berman designed a general purpose indexing methodology for such systems that used key images instead of key words as the index objects and could handle multiple distance measures. One of our systems could retrieve such disparate objects as eye images, brain images, and 3D skulls.

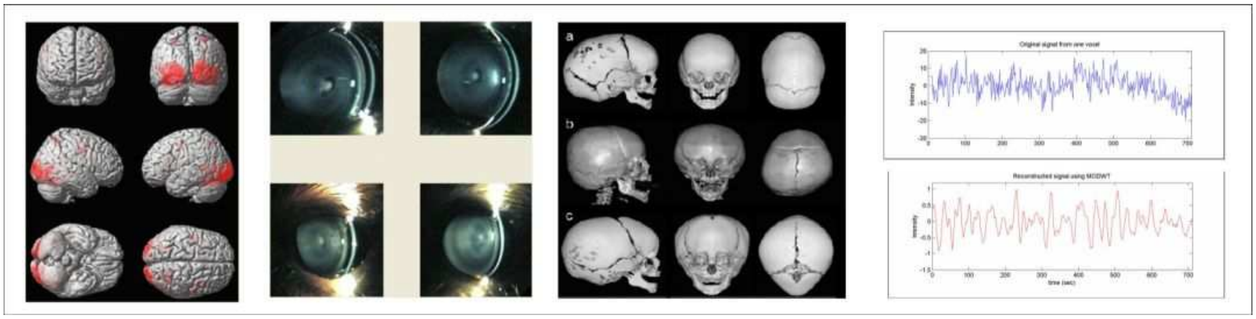

Figure 2: Objects retrieved by our multimedia retrieval system.

The use of 3D data, original referred to as *range data*, became popular in the 1990s as the physical sensors and 3D reconstruction algorithms improved. Patrick Flynn and Anil K. Jain at Michigan State were pioneers in this field and developed databases of range images for others to try. Some of the beginning problems were in just segmenting such images into the objects and their different surfaces for later use in recognition. Work on both 3D reconstruction and 3D recognition went on in parallel; my own students did both. 3D reconstruction work from space carving was pioneered by my colleagues Steve Seitz and Brian Curless (before they joined me at UW), while a nice piece of work was done by my own former student Kari Pulli who went on to a postdoc at Stanford and went with Brian and his adviser Marc Levoy to digitize the entire Michelangelo statue of David in Florence, Italy. Meanwhile, we started working on recognizing objects from their 3D shape. There was prior work in this area by Johnson and Hebert at CMU; they developed a representation called a *spin image* that could describe the surface structure about each point on a range image (or full 3D model) of an object. Matching such descriptors could be used for model-based object recognition or for finding matching points of the same object in multiple images. We started with this descriptor at a low-level, and my former student Salvador Ruiz-Correa developed a machine-learning-based system for learning the shape structure of regions on 3D objects that allowed them to be distinguished from one another. Regions with similar spin images became primitives and were then described by their spatial relationships for classification.

Small toy objects such as snowmen, bunnies, and dogs were used in our experiments, which

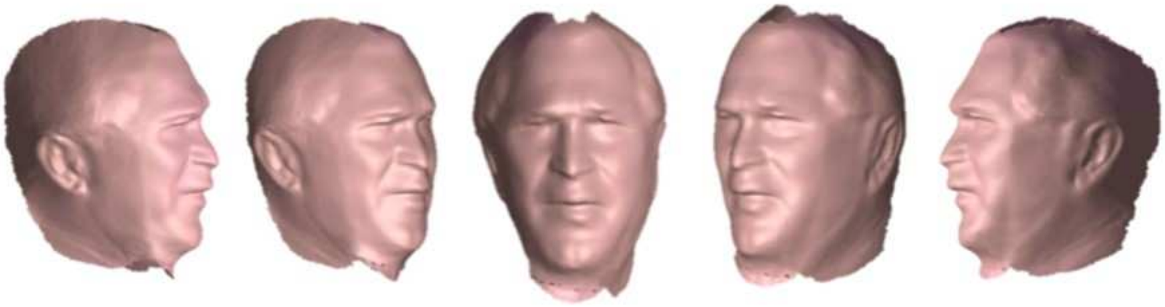

Figure 3: 3D Reconstruction of heads from uncalibrated internet photos.

showed that just the cheek areas could be used to differentiate them. Later, the technique was used in a medical application to recognize an abnormal condition of the skull called craniosynostosis. That was the start of period of work on abnormal craniofacial anatomy for my group in which we worked on fully 3D images of children with craniosynostosis, plagiocephaly, 22q11.2 deletion syndrome, and cleft lip and palate. This work involved both the development of new features and the use of machine learning for both classification and quantification of the severity of these disorders. One descriptor developed by my former student Indriyati Atmosukarto that worked particularly well for both plagiocephaly, which is concerned with flat parts on the back of a child's head, and 22q11.2 deletion syndrome, which causes multiple different abnormalities such as bulbous nasal tip and midface hypoplasia, was a 2D histogram of the azimuth and elevation angles of the 3D mesh of the head. For cleft lip and palate, the surgeons with whom we worked were mostly interested in symmetry measures, espcially in the areas of the nose and mouth. Jia Wu developed a whole suite of such measures, and we are still developing more. Another interesting approach was to use the error of reconstruction of the 3D face from principal components of a database of normal heads as a severity measure, which was pioneered by graduating student Shu Liang.

The 3D reconstruction work has continued to this day at the University of Washington. Working with Steve Seitz, Brian Curless, and Richard Szeliski (at Microsoft Research), Noah Snavely's 2008 dissertation on Scene Reconstruction and Visualization from Internet Photo Collections led to the Microsoft Photosynth product and to multiple other exciting papers and theses including "Reconstructing Rome", which was work done by Sameer Agarwal, then a postdoc at UW and now at Google. The addition of Ira Kemelmacher-Shlizerman to our team led to reconstructions of human faces and heads. Ira is well known for constructing 3D models from large collections of internet photos "in the wild" and has produced such papers as "What Makes Tom Hanks Look Like Tom Hanks" with her student Supasorn Suwajanakorn and Steve Seitz, "Head Reconstruction from Internet Photos" with our joint student Shu Liang and me, and "Transfiguring Portraits", in which she explores modifying 2D images of people to give them different hair styles, clothing, and very different appearances, while keeping the basic facial details of the person. Her most recent work with Suwajanakorn and Seitz was on learning lips sync from audio and could synthesize high quality video of a person (such as President Obama) speaking words that he never actually said. This produced some hubub from the press, since it could be used to create "fake news."

Object recognition has always been important in computer vision, but it died out for a while in the general sense in the 1980s and then returned in the late 1990s in much more powerful systems that used powerful machine learning techniques and large databases of training images, as well as modern image features. David Lowe's object recognition from his now famous SIFT features was one of the first of this kind of work. These features made it possible to recognize object classes from multiple different views. Rob Ferus's classic work then put together a machine learning framework for modeling object classes based on another type of new descriptor, the Kadir saliency operator. The exciting part of this work was that the system could learn what features best represented each object class; no more hand-constructed models. In our own group, my former student Yi Li developed his own new features called *abstract regions*, proposing the idea that any kind of region segmentation (color, texture, structure, or whatever) could be used together to learn to recognize classes of objects. In his final paper on the subject, he developed a generative/discriminative learning algorithm that first found the regions for each abstract type and extracted fixed length descriptors from each training image summarizing them and then trained a classifier to learn each particular class depending on these training vectors, concatenated for multiple types of abstract regions.

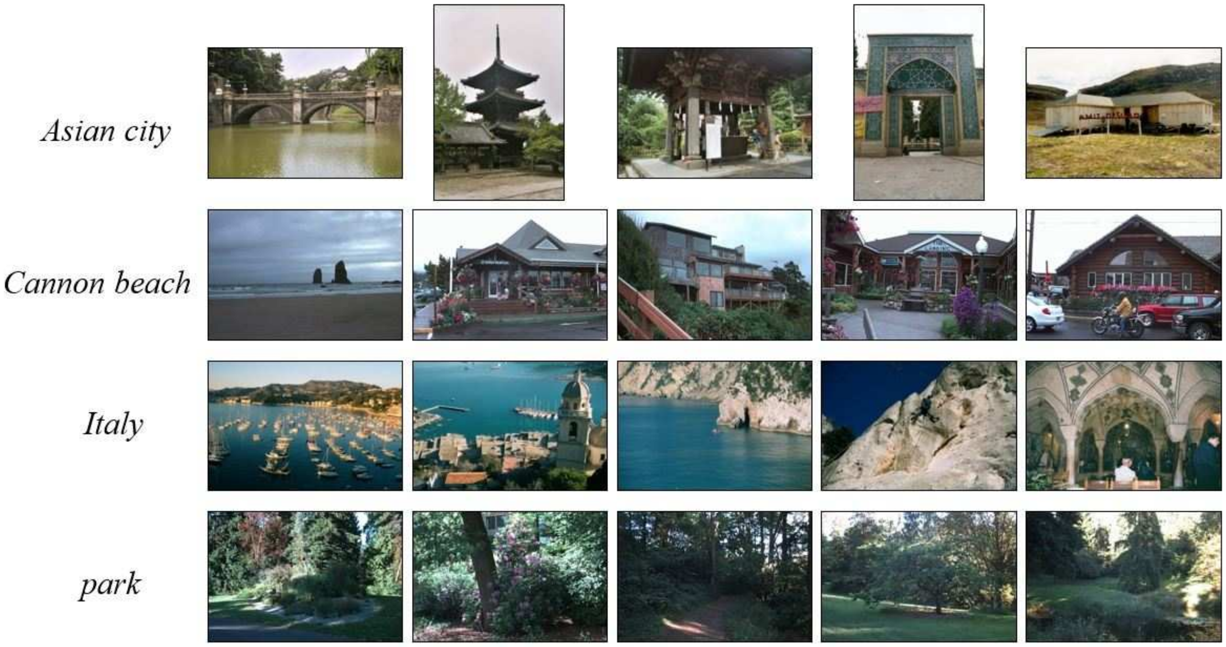

Figure 4: Image retrieval using generative/discriminative object recognition.

Object recognition has really taken off in the later 2000s. Such large benchmark datasets as the PASCAL VOC dataset and ImageNet allowed researchers to train and compare their algorithms on a consistent set of data with the same objects. Pedro Felzenszalb developed the *deformable parts model*, which represented objects via a root node and parts nodes that are detected by the HOG (histogram of gradients) feature detector. This model won the PASCAL VOC challenge 2007, achieving a mean Average Precision (mAP) score of 17%. It was then improved in various ways for several years, continuing to win and to improve to 23% in 2008, 28% in 2009, 37% in 2010, and 41% in both 2011 and 2012. Meanwhile deep neural nets were starting to take over. First used in practice by Yann LeCun in 1998 for a document recognition application and popularized by Geoff Hinton, convolutional neural networks (CNNs) were applied to the ImageNet object classification challenge (ILSVRC) in 2012 by Alex Krizhevsky. Their model, AlexNet, won the ILSVRC-2012 challenge, achieved a top-5 accuracy of 83%. It was then improved by building very deep convolutional neural networksd to 93.33% in 2014 (GoogLeNet) and 96.43% in 2015 (ResNet). Deep neural networks were applied to the the PASCAL VOC challenge in 2013 by Ross Girshick in a form called R-CNNs (region-based convolutional networks) in which he first used region proposals to locate regions potentially containing objects and then trained a CNN to recognize the objects in those regions. He won the PASCAL VOC challenge in 2013 with a mAP score of 53% and in 2014 with 62%. CNNs have continued to outperform other methods. My colleague Ali Farhadi and his student Joseph Redmon have recently developed YOLO, a neural network approach to object detection that is small and fast. Their newest product YOLO9000 is a real-time system that achieved a mAP of 76.8% on PASCAL VOC 2007. My own student Sachin Mehta is working on object detection in real world images for use in navigation for the blind and disabled. In this case, there is not necessarily a big benchmark database that has all objects compiled in advance, and it is important to know what the object is and approximately where it is with respect to the human navigator.

Our own current recognition work being done by students Deepali Aneja and Bindita Chaudhuri is on human facial expession recognition and conversion to stylized cartoon characters using convolutional neural networks. Our neural nets were initially trained on large databases of both human facial expressions and a limited number of character facial expressions. We have now moved into learning how to generate the 3D parameters of the characters and to be able to generalize from a single humanlike character to multiple different stylized characters. On the medical side, we are working on analysis of breast and melanoma biopsy slide images. In the breast domain, we have completed a five year study that examines both the whole slide image data and the tracking data from three expert and more than 200 community pathologists in order to better understand what they do during their diagnostic process. In this work, we have studied the characteristic patterns of the pathologists, discovering only so far that they tend to make more errors when zooming in more, and have developed automated systems for detecting regions of interest and for diagnosing the slide. Ezgi Mercan's Ph.D. dissertation has produced several high-quality papers on this topic. Her work on using the structure around a duct in diagnosis is particularly novel. Meanwhile, we have moved to another grant on melanoma biopsy diagnosis and have found that the melanoma whole slide images are even more challenging than the breast biopsies. We are currently work-

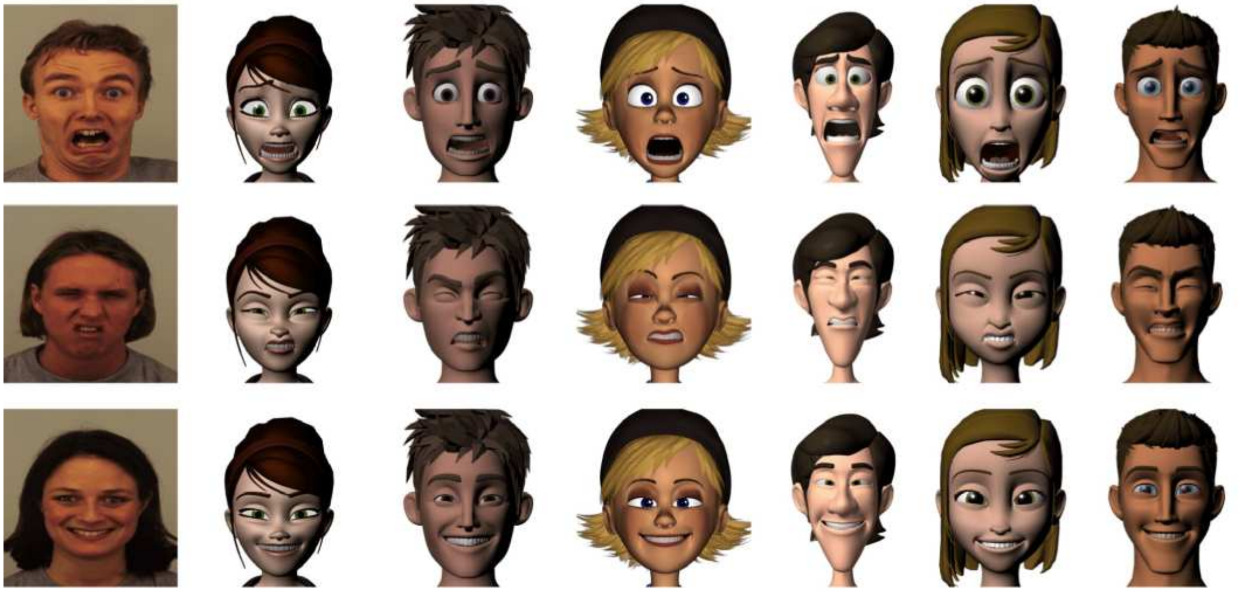

Figure 5: Use of deep learning to transfer facial expressions from human input to cartoon characters.

ing on detection of melanocytes and mitotic figures using CNNs and then will try to develop a structural pattern recognition approach.

Computer vision has gone from an experimental research area to a field that is in demand from multiple industrial concerts. While medical and part inspection applications will always be around, there are now whole new working systems developed and working in such areas as face recognition, face and head reconstruction, self-driving cars, robot navigation, and virtual and augmented reality. Students with skills in computer vision and machine learning are being snapped up by companies even before they are close to graduation. The present is bright, and the future is even brighter.