Computer Vision: Evolution and Promise

T. S. Huang

University of Illinois at Urbana-Champaign
Urbana, IL 61801, U. S. A.
E-mail: huang@ifp.uiuc.edu

Abstract

In this paper we give a somewhat personal and perhaps biased overview of the field of Computer Vision. First, we define computer vision and give a very brief history of it. Then, we outline some of the reasons why computer vision is a very difficult research field. Finally, we discuss past, present, and future applications of computer vision. Especially, we give some examples of future applications which we think are very promising.

1 What is Computer Vision?

Computer Vision has a dual goal. From the biological science point of view, computer vision aims to come up with computational models of the human visual system. From the engineering point of view, computer vision aims to build autonomous systems which could perform some of the tasks which the human visual system can perform (and even surpass it in many cases). Many vision tasks are related to the extraction of 3D and temporal information from time-varying 2D data such as obtained by one or more television cameras, and more generally the understanding of such dynamic scenes.

Of course, the two goals are intimately related. The properties and characteristics of the human visual system often give inspiration to engineers who are designing computer vision systems. Conversely, computer vision algorithms can offer insights into how the human visual system works. In this paper we shall adopt the engineering point of view.

2 History of Computer Vision

It is commonly accepted that the father of Computer Vision is Larry Roberts, who in his Ph.D. thesis (cir. 1960) at MIT discussed the possibilities of extracting 3D geometrical information from 2D perspective views of blocks (polyhedra) [1]. Many researchers, at MIT and elsewhere, in Artificial Intelligence, followed this work and studied computer vision in the context of the blocks world.

Later, researchers realized that it was necessary to tackle images from the real world. Thus, much research was needed in the so called ``low-level” vision tasks such as edge detection and segmentation. A major milestone was the framework proposed by David Marr (cir. 1978) at MIT, who took a bottom-up approach to scene understanding [2].

Low-level image processing algorithms are applied to 2D images to obtain the ``primal sketch” (directed edge segments, etc.), from which a 2.5 D sketch of the scene is obtained using binocular stereo. Finally, high-level (structural analysis, *a priori* knowledge) techniques are used to get 3D model representations of the objects in the scene. This is probably the single most influential work in computer vision ever. Many researchers cried: ``From the paradigm created for us by Marr, no one can drive us out.”

Nonetheless, more recently a number of computer vision researchers realized some of the limitation of Marr's paradigm, and advocated a more top-down and heterogeneous approach. Basically, the program of Marr is extremely difficult to carry out, but more important, for many if not most computer vision applications, it is not necessary to get complete 3D object models. For example, in autonomous vehicle navigation using computer vision, it may be necessary to find out only whether an object is moving away from or toward your vehicle, but not the exact 3D motion of the object. This new paradigm is sometimes called ``Purposive Vision" implying that the algorithms should be goal driven and in many cases could be qualitative [3]. One of the main advocates of this new paradigm is Yiannis Aloimonos, University of Maryland.

Looking over the history of computer vision, it is important to note that because of the broad spectrum of potential applications, the trend has been the merge of computer vision with other closely related fields. These include: Image processing (the raw images have to be processed before further analysis). Photogrammetry (cameras used for imaging have to be calibrated. Determining object poses in 3D is important in both computer vision and photogrammetry). Computer graphics (3D modeling is central to both computer vision and computer graphics. Many exciting applications need both computer vision and computer graphics - see Section 4).

3 Why is Computer Vision Difficult?

Computer Vision as a field of research is notoriously difficult. Almost no research problem has been satisfactorily solved. One main reason for this difficulty is that the human visual system is simply too good for many tasks (e.g., face recognition), so that computer vision systems suffer by comparison. A human can recognize faces under all kinds of variations in illumination, viewpoint, expression, etc. In most cases we have no difficulty in recognizing a friend in a photograph taken many years ago. Also, there appears to be no limit on how many faces we can store in our brains for future recognition. There appears no hope in building an autonomous system with such stellar performance.

Two major related difficulties in computer vision can be identified:

1. How do we distill and represent the vast amount of human knowledge in a computer in such a way that retrieval is easy?

2. How do we carry out (in both hardware and software) the vast amount of computation that is often required in such a way that the task (such as face recognition) can be done in real time?

4 Application of Computer Vision: Past, Present, and Future

Past and present applications of computer vision include: Autonomous navigation, robotic assembly, and industrial inspections. At best, the results have been mixed. (I am excluding industrial inspection applications which involve only 2D image processing and pattern. recognition.) The main difficulty is that computer vision algorithms are almost all brittle; an algorithm may work in some cases but not in others. My opinion is that in order for a computer vision application to be potentially successful, it has to satisfy two criteria: 1)Possibility of human interaction. 2) Forgiving (i.e., some mistakes are tolerable). It also needs to be emphasized that in many applications vision should be combined with other modalities (such as audio) to achieve the goals.

Measured against these two criteria, some of the exciting computer vision applications which can be potentially very successful include:

Image/video databases-Image content-based indexing and retrieval.

Vision-based human computer interface - e.g., using gesture (combined with speech) in interacting with virtual environments.

Virtual agent/actor - generating scenes of a synthetic person based on parameters extracted from video sequences of a real person.

It is heartening to see that a number of researchers in computer vision have already started to delve into these and related applications.

5 Characterizing Human Facial Expressions: Smile

To conclude this paper, we would like to give a very brief summary of a research project we are undertaking at our Institute which is relevant to two of the applications mentioned in the last Section, namely, vision-based human computer interface, and virtual agent/actors, as well as many other applications. Details of this project can be found in Ref. 4.

Different people usually express their emotional feelings in different ways. An interesting question is number of canonical facial expressions for a given emotion. This would lead to applications in human computer interface, virtual agent/actor, as well as model-based video compression scenarios, such as video-phone. Take smile as an example. Suppose, by facial motion analysis, there are 16 categories found among all smiles posed by different people. Smiles within each category can be approximately represented by a single smile which could be called a canonical smile. The facial movements associated with each canonical smile can be designed in advance. A new smile is recognized and replaced by the canonical smile at the transmitting side, only the index of that canonical smile needs to be transmitted. At the receiving sides, this canonical smile will be reconstructed to express that person's happiness.

We are using an approach to the characterization of facial expressions based on the principal component analysis of the facial motion parameters. Smile is used as an example, however, the methodology can be generalized to other facial expressions.

A database consisting of a number of different people's smiles is first collected. Two frames are chosen from each smile sequence, a neutral face image and an image where the smile reaches its apex. The motion vectors of a set of feature points are derived from these two images and a feature space is created. Each smile is designated by a point in this feature space. The principal component analysis technique is used for dimension reduction and some preliminary results of smile characterization are obtained. Some dynamic characteristics of smile are also studied.

For smiles, the most significant part on the face is the mouth. Therefore, four points around the mouth are chosen as the feature points for smile characterization: The two corners of the mouth and the mid-points of the upper and lower lip boundaries.

About 60 people volunteered to show their smiles. These four points are identified in the two end frames of each smiling sequence, i.e., the neutral face image and the one in which the smile reaches its apex. The two face images are first registered based on some fixed features, e.g., the eye corners and the nostrils. In this way, the global motion of the head can be compensated for since only the local facial motions during smiles are of interest. Thus, every smile is represented by four vectors which point from the feature points on the neutral face image to the corresponding feature points on the smiling face image. These motion vectors are further normalized according to the two mouth corner points. Then, each component of these vectors serves as one dimension of the ``smile feature space." In our experiments to date, these are 2D vectors. Thus, the dimensionality of the smile feature space is 8. Principal component analysis is applied to this 8D feature space.

In addition to looking at the two end frames of a smile sequence, it is also of interest to study the dynamic characteristics of smiles from real motion trajectories of the feature points. Using the same four feature points, their motions through the whole smile sequence are tracked and the trajectories are recorded.

- **Temporal uniformity**: Whether a point moves equally between two consecutive frames.
- **Spatial linearity:** Whether an overall trajectory can be approximated by a straight line.

The feature tracking procedure is applied to 20 smiling sequences. The motion trajectories are estimated and these two characteristics are further calculated. Preliminary results indicate that: 1) The motions of the mouth corners for most smiles are asymmetric. 2) The assumption of motion smoothness (both spatially linear and temporally uniform) is quite reasonable. 3) After principal component analysis, there are still no obvious multiple clusters in the feature space. This may be due to the fact that the database we use is still too small to cover the large variation of smiles.

One possible way to do smile clustering is to distinguish smiles by using qualitative criteria, for example, whether the mouth is open or closed during smile, whether the smile is symmetrical, etc. Another interesting thing to do is to perform subjective tests. The difference between different smiles is determined by human subjects. The purpose is to see how far two points in the feature space should be moved apart so that the smiles will differ from each other. The results can be used for smile clustering based on the motion vectors. Finally, motion vectors of more facial features (eyes, nose, cheek areas) should be used in constructing the smile feature space.

6 Concluding Remarks

Computer Vision is more than 30 years old. Although as a research field it has been offering many challenging and exciting problems, in terms of successful engineering applications it has been rather disappointing. However, more recently, several very exciting applications have appeared where computer vision I believe can make major contributions.

Acknowledgment:

This work was supported by U. S. Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0003.

References

1. Y. Aloimonos (ed.), Special Issue on Purposive and Qualitive Active Vision, *CVGIP B: Image Understanding*, **Vol. 56** (1992).
2. D. Marr, ``Vision: A Computational Investigation into the Human Representation and Processing of Visual Information'', Freeman, San Francisco (1982).
3. L. Roberts, ``Machine perception of 3D solids", Chapter 9 in J. T. Tippett, *et al*. (eds), Optical and Electro-Optical Information Processing, MIT Press, pp. 159-197 (1965).
4. L. Tang and T. S. Huang, ``Characterizing smiles in the context of video phone data compression", *Proceedings of Internation Conference on Pattern Recognition*, Vienna, Austria (1996).