Katsushi Ikeuchi
Editor

Computer Vision
A Reference Guide

With 433 Figures and 16 Tables
Gesture Recognition

Matthew Turk
Computer Science Department and Media Arts and Technology Graduate Program, University of California, Santa Barbara, CA, USA

Synonyms
Human motion classification

Definition
Vision-based gesture recognition is the process of recognizing meaningful human movements from image sequences that contain information useful in human-human interaction or human-computer interaction. This is distinguished from other forms of gesture recognition based on input from a computer mouse, pen or stylus, sensor-based gloves, touch screens, etc.

Background
Automatic image-based gesture recognition is an area of computer vision motivated by a range of application areas, including the analysis of...
human-human communication, sign language interpretation, human-robot interaction, multimodal human-computer interaction, and gaming. Human gesture has a long history of interdisciplinary study by psychologists, linguists, anthropologists, and others in the context of human communication [9], exploring the role of gesture in face-to-face conversation, universal and cultural aspects of gesture, the influence of gesture in human evolution and in child development, and other topics, going back at least to the work of Charles Darwin with *The Expression of the Emotions in Man and Animals* (1872). Research in computer vision-based gesture recognition began primarily in the 1990s as computers began to be capable of supporting real-time (or interactive time, fast enough to support human interaction) processing and recognition of video streams.

Several gesture taxonomies or categorizations have been developed by different researchers that underscore the breadth of the problem in general. Cadoz [4] described three functional roles of human gesture: semiotic (gestures to communicate meaningful information), ergodic (gestures to manipulate the environment), and epistemic (gestures to discover the environment through tactile experience). Most work in automated gesture recognition is concerned with the first role (semiotic gestures), whereas the area of human activity analysis tends to focus on the latter two.

Kendon [11] described a gesture continuum, defining five types of gestures: gesticulation, language-like gestures, pantomimes, emblems, and sign languages. Each of these has a varying association with verbal speech, language properties, spontaneity, and social regulation, indicating that human gesture is indeed a complex phenomenon. Gesticulation, defined as spontaneous, speech-associated gesture, makes up a large portion of human gesture and is further characterized by McNeill [14] into four types:

- **Iconic** – representational gestures depicting some feature of the object, action, or event being described
- **Metaphoric** – gestures that represent a common metaphor, rather than the object or event directly
- **Beat** – small, formless gestures, often associated with word emphasis
- **Deictic** – pointing gestures that refer to people, objects, or events in space or time

These gesture types modify the content of accompanying speech and often help to disambiguate speech, similar to the role of spoken intonation. Cassell et al. [5] described early research in conversational agents that models the relationship between speech and gesture and generates interactive dialogs between three-dimensional animated characters that gesture as they speak.

**Vision-Based Gesture Recognition**

Vision-based gesture recognition must detect human movements from image sequences, ideally in real time and independent of the specific user, the imaging condition, the camera viewpoint, clothing and other confusing factors, and the significant variation in how people gesture. Aspects of a gesture that may be critical to its interpretation include spatial information (where the gesture occurs and/or refers to), pathic information (the path a gesture takes), symbolic information (sign(s) made during a gesture), and affective information (the emotional quality of a gesture, which may be related to the speed and magnitude of a gestural act, as well as to facial expression).

A gesture may be considered as a continuous set of movements or as a sequence of discrete poses or postures. Gestures are inherently dynamic and time varying, while postures are specific – and static – configurations; recognizing specific configurations (such as making a “victory sign”) should properly be referred to as posture recognition. Analyzing movement (such as dance or general behaviors in a social situation) is generally referred to as activity analysis or activity recognition [1].

Unless the gestures are constrained to a particular point in time (e.g., with a “push to gesture” functionality), it is necessary to determine when a dynamic gesture begins and ends. This temporal segmentation/detection of gesture is a challenging problem, particularly in less constrained environments where several kinds of spontaneous gestures are possible amidst other non-gestural movement. While temporal detection and segmentation of gestures may be attempted as a first step, other approaches combine spatial (or spatiotemporal) segmentation with recognition [2, 12].

A typical approach to human gesture recognition involves detecting and tracking component body parts, such as hands, arms, head, torso, legs, and feet, based on an articulated body model, and subsequently classifying the movement into one of a set of known
gestures (e.g., [19]). The output of the tracking stage is a time-varying sequence of parameters describing (2D or 3D) positions, velocities, and angles of the relevant body parts and features, possibly including a representation of uncertainty that indicates limitations of the sensor and the algorithms. An alternative is to take a view-based approach, which computes parameters directly from image motion, generally bypassing human body modeling (e.g., [6, 7]).

Hand gestures have received particular attention in gesture recognition, as hands provide the opportunity for a wide range of meaningful gestures, as evidenced by the rich history of human sign languages such as American Sign Language (ASL) (e.g., [18, 20]), and may be convenient for quickly and naturally conveying information in vision-based interfaces (e.g., [3, 8]). Video-only approaches have had limited success, however, due to the complexities of highly articulated hands and skin-on-skin occlusions.

Recently, there has been a significant amount of work in gesture recognition from depth imagery or combinations of video (RGB) and depth data, largely driven by the availability of the Microsoft Kinect sensor (and SDK/toolkit) and the use of body modeling, tracking, and gesture recognition in consumer applications using the Kinect [16]. Other companies are developing new technologies for gesture recognition (e.g., [17] and [13]), as well as for spatial operating environments that leverage tracking and gesture technologies (e.g., [15]).

Open Problems

Gesture recognition is a broadly defined set of problems and challenges, for which there are some domain-specific solutions that are adequate for commercial use; however, the general problems are largely unsolved. At the low level, there are limitations to any choice of sensor type, and work remains to be done on integrating data from multiple sensors. There is no agreement on how to best represent the sensed spatial and temporal information and its relationship to human movement. Temporal segmentation of natural dynamic gestures is unlikely to be solved without a deep understanding of the gesture semantics – i.e., the high-level context in which the gestures take place. Despite the recent impact of depth sensors on this area, the field is still wide open for solutions that can provide precise and robust gesture recognition in a wide range of environments.

Research in vision-based gesture recognition can be stimulated by the creation and sharing of thorough, annotated data sets that capture a wide range of spontaneous gestures and imaging conditions and by apples-to-apples comparisons such as the recent ChaLearn Gesture Challenges [10].

References

Gradient Vector Flow

Chenyang Xu¹ and Jerry L. Prince²
¹Siemens Technology-To-Business Center, Berkeley, CA, USA
²Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

Synonyms
GVF

Related Concepts
Edge Detection

Definition
Gradient vector flow is the vector field that is produced by a process that smooths and diffuses an input vector field and is usually used to create a vector field that points to object edges from a distance.

Background
Finding objects or homogeneous regions in images is a process known as image segmentation. In many applications, the locations of object edges can be estimated using local operators that yield a new image called an edge map. The edge map can then be used to guide a deformable model, sometimes called an active contour or a snake, so that it passes through the edge map in a smooth way, therefore defining the object itself.

A common way to encourage a deformable model to move toward the edge map is to take the spatial gradient of the edge map, yielding a vector field. Since the edge map has its highest intensities directly on the edge and drops to zero away from the edge, these gradient vectors provide directions for the active contour to move. When the gradient vectors are zero, the active contour will not move, and this is the correct behavior when the contour rests on the peak of the edge map itself. However, because the edge itself is defined by local operators, these gradient vectors will also be zero far away from the edge, and therefore the active contour will not move toward the edge when initialized far away from the edge.

Gradient vector flow (GVF) is the process that spatially extends the edge map gradient vectors, yielding a new vector field that contains information about the location of object edges throughout the entire image domain. GVF is defined as a diffusion process operating on the components of the input vector field. It is designed to balance the fidelity of the original vector field, so it is not changed too much, with a regularization that is intended to produce a smooth field on its output.

Although GVF was designed originally for the purpose of segmenting objects using active contours attracted to edges, it has been since adapted and used for many alternative purposes. Some newer purposes including defining continuous medial axis representation [1], extracting scale-invariant image features [2], regularizing image anisotropic diffusion algorithms [3], finding the centers of ribbon-like objects [4], and much more.

Theory
The theory of GVF was originally described in [5]. Let \( f(x,y) \) be an edge map defined on the image domain. For uniformity of results, it is important to restrict the intensities to lie between 0 and 1, and by convention \( f(x,y) \) takes on larger values (close to 1) on the object edges. The gradient vector flow (GVF) field is given by the vector field \( \mathbf{v}(x,y) = [u(x,y), v(x,y)] \) that minimizes the energy functional