Paper:

Gestures Recognition Based on the Fusion of Hand Positioning and Arm Gestures

Didier Coquin[*], **Eric Benoit**[*], **Hideyuki Sawada**[**], **and Bogdan Ionescu**[*,***]

[*]LISTIC – University of Savoie, Domaine Universitaire, B.P. 806, 74016 Annecy-Cedex, France
E-mail: Didier.coquin@univ-savoie.fr
[**]Kagawa University, 2217-20 Hayashi-cho, Takamatsu, Kagawa 761-0396, Japan
[***]LAPI, University "Politehnica" Bucharest, 061071 Romania

[Received April 4, 2006; accepted August 11, 2006]

To improve the link between operators and equipment, communication systems have begun using natural (user-oriented) languages such as speech and gestures. Our goal is to present gesture recognition based on the fusion of measurements from different sources. Sensors must be able to capture at least the location and orientation of the hand, as is done by Dataglove and a video camera. Dataglove gives the hand position and the video camera gives the general arm gesture representing the gesture's physical and spatial properties based on the two-dimensional (2D) skeleton representation of the arm. Measurement is partly complementary and partly redundant. The application is distributed over intelligent cooperating sensors. We detail the measurement of hand positioning and arm gestures, fusion processes, and implementation.

Keywords: hand positioning, arm gestures, fusion process, gesture recognition

1. Introduction

A primary goal of gesture recognition research is to create systems that identify specific human gestures and to use them to convey information or control devices. General solutions for gesture analysis are divided into two categories – using mechanical devices, such as glove-based, to directly measure hand joint angles and spatial location and using computer-vision-based techniques –. Although using mechanical devices enables real-time processing and produces reliable information, it requires that users wear cumbersome devices and even carry cables connecting the device to a computer, with the larger number of requirements making the sensing of natural hand movement difficult. In contrast, using computer-vision-based techniques is suited for hand positioning determination because vision is non-invasive sensing.

In gesture recognition, a camera reads movements and communicates data to a computer that uses the gestures as input to control equipment or applications. Gesture recognition is the process by which gestures formed by users are made known to the system. In completely immersive Virtual Reality environments, keyboards are generally not included, necessitating other control over the environment. The goal is "natural" control or at least control that is easy to learn. Gestures with the fingers or hands provide this and are usually inexpensive and easy to implement. Gestures may be static, in which users assume poses or certain configurations, or dynamic, in which movement is the gesture itself. To make these gesture configurations accessible to the computer, sensing devices are either connected directly to the user or measure the configuration indirectly and remotely.

Attachments – gloves, datasuits, or 6-degree-of-freedom (DOF) trackers – generally provide information in all three spatial dimensions. This is not true in image-based approaches, in which only two-dimensional (2D) projection is possible. This presents problematic because a 3D configuration project different 2D views for different user and camera locations. In some cases, a 2D view may even correspond to different 3D configurations. Computer vision is used to either recognize gestures from 2D views or to reconstruct the third dimension from several views taken by more than one camera. These tasks are, however, difficult and computationally expensive, requiring dedicated hardware for real-time performance. To make recognition from images easier, gestures should be defined redundantly over several dimensions to avoid deterioration due to reduction from three to two dimensions.

Gesture measurement is complex, requiring several sensor techniques to improve the final result. Many approaches using Dataglove provide data on hand gestures but cannot access the hand location or orientation.

This necessitates adding other sensors that capture at least hand location and orientation, e.g., 6Dof sensors or video cameras. Video cameras provide information on the arm movement useful in dynamic hand gestures recognition. The objective of the dynamic hand gesture recognition we propose is to guide robots remotely using hand gestures. To guide robots, however, we must "send" them information, preferably "encapsulated" in dynamic hand gestures. Our task is thus to translate information into instructions to be sent to the robot.

Increasing the number of information sources is useful only if the measurement results can be combined. To simplify fusion, measurements are expressed as fuzzy descriptions or fuzzy distributions of possibility - representation widely used to enable heterogeneous sources to be combined [1, 2].

Redundant information is combined by a possibilistic aggregation of measurements, enabling the uses of sources having different degrees of confidences. Complementary information is fused by aggregation based on fuzzy rules, enabling the distribution of fusion over a distributed sensor network.

We define the fuzzy description of measurement and its application to the fuzzy representation of hand positioning and arm gestures, presenting the fusion process based on complementary and redundant information applied to driving a mobile robot.

2. Background of Gesture Recognition Studies

The literature on hand gesture recognition is growing. In the two principal overviews proposed by Pavlovic et al. [3] and by Wu and Huang [4], hand gesture recognition is divided into two main tasks – *feature extraction* in which low-level information from raw data is analyzed to produce higher-level semantic information and, once features are collected, classification in which collected data is used to detect hand gestures –.

Features are built on distance, velocity, and acceleration information, energy measurements, or angle information and based on techniques such as active shape models, principal component analysis, linear fingertip models, and spatiotemporal vector analysis [5–7].

In classification, we consider the template-based approach, statistical models (Hidden Markov Model [8, 9]), and miscellaneous algorithms (Neural Networks [8, 10], causal analysis). A template-based approach, e.g., *conventional template matching*, *instance-based learning* [11], *the linguistic approach* [12], and *appearance-based motion analysis* [13], generally creates data records for individual positioning and gestures in a set and uses them to classify new positioning and gestures.

Conventional template matching, the simplest approach, involves two steps. In the first step, templates are created by collecting data on individual gestures in a gesture set; repeating each gesture a number of times, and averaging raw data for each sensor (glove, camera) and storing it as a template. In the second step, current sensor readings are compared to a given set of templates to find the gesture template most closely matching the current data record. The dynamic gesture recognition we propose uses conventional template matching for classification.

We focus on emotional and intentional aspects of gestures in measurement, and we construct robust gesture recognition by integrating redundant information obtained from image processing and sensor-based measurement. Our objective is to use simple recognition via complementary or redundant data from different sources and to implement our approach in a way that requires no extra calculation.

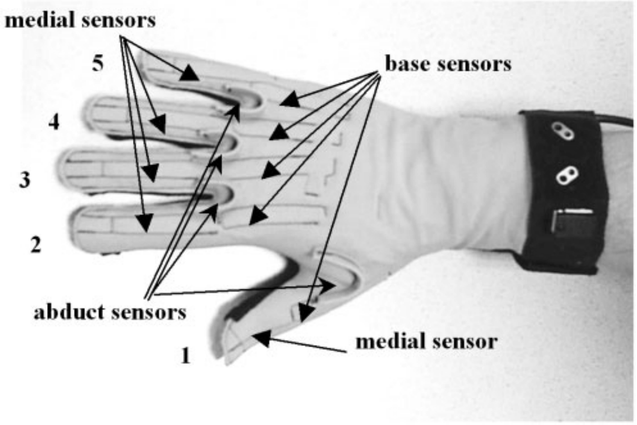

Fig. 1. Cyberglove and sensors.

3. Measurement

3.1. Fuzzy Description of Measurement

The link between a physical state and its linguistic representation is characterized by symbolism defined by triplet $< X, L, R >$ in which X is the set of physical states, L the lexical set representing measurement results, and R a relation on $X \times L$. The two mappings extracted from this relation – ***description mapping*** denoted ***d*** associating a subset of L with any item of X and meaning mapping denoted ***m*** associating a subset of X with any item of L – are linked to the following equation:

$$\forall x \in X,\ \forall a \in L,\ x \in m(a) \Leftrightarrow a \in d(x). \quad . \quad . \quad . \quad (1)$$

The R relation may be fuzzy. The translation of a physical state into its linguistic representation – called *fuzzy linguistic description mapping* or *fuzzy description mapping* – transforms object x of the set of physical states X into a fuzzy subset of linguistic terms called the *fuzzy description* of x. Dual mapping, called *fuzzy meaning* mapping, associates a fuzzy subset of X with each term a of lexical set L. This fuzzy subset is the *fuzzy meaning* of a. These two mappings are linked with the following equation:

$$\forall x \in X,\ \forall a \in L,\ \left(\mu_{m(a)}(x) = \mu_{d(x)}(a)\right). \quad . \quad . \quad . \quad (2)$$

3.2. Hand Gestures

As in this paper, the fuzzy description of hand positioning proposed in [14] gives finger configurations numerically via the Cyberglovec's bending sensors (**Fig.1**). Numerical-to-lexical conversion is made using a fuzzy partition of dataglove measurement spaces. The fuzzy glove then provides a fuzzy description of each finger configuration and its relative location.

The robustness and efficiency of recognition depends on the definition of partitions. A partition may be constructed several ways – empirically, by interpolation, by clustering, etc. – and the fuzzy glove is calibrated by modifying this partition.

A preliminary study showed that each finger takes five