Adaptive Gesture Recognition System Integrating Multiple Inputs

Master Thesis - Colloquium

Tobias Staron

University of Hamburg
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Technical Aspects of Multimodal Systems

May 19, 2015
Table of contents

Introduction
Set-Up
Multiple Inputs & Adaptivity
Conclusion
Overview

Introduction
Set-Up
Multiple Inputs & Adaptivity
Conclusion
Overview

Introduction

Gesture Recognition in General

Motivation

Set-Up

Multiple Inputs & Adaptivity

Conclusion
Gesture Recognition in General

▶ several applications (more natural interaction with robots, way of communication, sign language, ...)

▶ Gesture recognition “is the process by which the gestures made by the user are recognized by the receiver.” (Mitra & Acharya, 2007 [3])

▶ static vs. dynamic gestures
Overview

Introduction
  Gesture Recognition in General

Motivation

Set-Up

Multiple Inputs & Adaptivity

Conclusion
Previous Work

- TAMS - Master Project “Intelligent Robotics” (2013-2014)
- vision-based system (Microsoft Kinect) for recognizing static gestures
- depth images and Support Vector Machines (SVMs)
- project paper (*Paetzel & Staron, 2014* [4])
Problems in Gesture Recognition

- recognition results in general
- context-depended applications
- changed circumstances, e.g. new users / users with different figures, changed environments, changed camera properties (position, calibration, ...), light changes, ...

⇒ exploiting features of Robotics (a robot might have more than one sensor; possible interaction between user and robot)
Hypotheses

- use of multiple inputs ⇒ improved recognition results (& context-independent systems)
- use of multiple inputs ⇒ robustness
- possible interaction between user and robot ⇒ ability of the system to adapt to changed circumstances
- possible interaction between user and robot ⇒ omitting of preliminary training

⇒ development of an adaptive gesture recognition system that makes use of multiple inputs
Overview

Introduction
Set-Up
Multiple Inputs & Adaptivity
Conclusion
Overview

Introduction
Set-Up
 Inputs
 Data
 Evaluation Criteria
 k-Nearest Neighbor (k-NN) Classifiers
 Multiple Inputs & Adaptivity
 Conclusion
Depth Images

- gray value images
- information about distances to the camera
- preprocessing (noise reduction, foreground separation, histogram equalization, grid) \((\text{Biswas & Basu, 2011 [2]})\)
- gray value binning in grid cells \(\Rightarrow 520\) features
Exemplary Preprocessing of a Depth Image

RGB image of an exemplary gesture.
Exemplary Preprocessing of a Depth Image

The corresponding depth image prior to preprocessing.
Exemplary Preprocessing of a Depth Image

The depth image but with reduced noise.
Exemplary Preprocessing of a Depth Image

Only the foreground of the depth image.
Exemplary Preprocessing of a Depth Image

The foreground of the depth image after histogram equalization.
Exemplary Preprocessing of a Depth Image

The equalized foreground of the depth image with a grid put on it.
Skeletal Information

- OpenNI tracker
- position and orientation of several joints of the human skeleton
- a coordinate frame for each joint \(\Rightarrow\) transformations into target frame
- 8 joints \(\Rightarrow\) 56 features
Overview

Introduction

Set-Up

Inputs

Data

Evaluation Criteria

k-Nearest Neighbor (k-NN) Classifiers

Multiple Inputs & Adaptivity

Conclusion
Collecting Training and Test Data

- 12 gestures
- 10 test users ⇒ different groups (users with similar/differing figures)
- different poses and positions (to the left or right)
- but no different distances to the camera
- different environments
- camera calibration and illumination remained unchanged
Different Environments
Overview

Introduction

Set-Up

Inputs
Data

Evaluation Criteria

k-Nearest Neighbor (k-NN) Classifiers

Multiple Inputs & Adaptivity

Conclusion
Evaluation Criteria

- precision: proportion of test instances classified correctly
- recall: proportion of instances that should have been classified as a certain gesture that have actually got the respective label
- $F_1$-score $= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
- average classification and (initial) training time
- nr. of training instances
Overview

Introduction

Set-Up

Inputs
Data
Evaluation Criteria

k-Nearest Neighbor (k-NN) Classifiers

Multiple Inputs & Adaptivity

Conclusion
k-Nearest Neighbor (k-NN) Classifier

- supervised learning method
- arbitrary number of dimensions
- no explicit training (computations during classification)
- label that occurred most among the k-nearest neighbors of a query instances is chosen
- distance measure (e.g. Euclidean distance)
Exemplary Dataset in the 2-Dimensional Space

Three classes, represented by blue squares, magenta diamonds and yellow
Weighted k-NN Classifier

- if a training example matches the query instance, its label will be chosen ⇒ Generalization
- the nearer one of its k-nearest neighbors lies by the query instance, the higher the probability that its label is the result
Training of Classifiers

- classifiers for each kind of input, for each group of users and for each environment
- the same amount of training (and test) data for each classifier
Overview

Introduction
Set-Up
Multiple Inputs & Adaptivity
Conclusion
Overview

Introduction
Set-Up
Multiple Inputs & Adaptivity
  Combining Multiple Inputs
  Adaptivity
  Integrated System
Conclusion
Sensor Fusion

- low-level sensor fusion: fusion at signal level, one classifier
- high-level sensor fusion: fusion at a more symbolic level, one classifier per input, classification results are fused
- low-level sensor fusion does not allow for variations regarding the chosen inputs (e.g. adding or removing of sensors) without omitting previous data / retraining everything
- ⇒ high-level sensor fusion was chosen
Hypotheses Verification

- inspired by Aldoma et al. (*Aldoma et al., 2013 [1]*)
- high-level sensor fusion approach
- one classifier per kind of input
- each classifier can generate an unspecified number of hypotheses
- each hypothesis is weighted
- hypothesis with the highest weight is chosen as recognition result
Weighting Cues

- each hypothesis is weighted by an unspecified number of weighting cues
- neighboring cue (in case of k-NN classifiers): all labels occurring among k-nearest neighbors as hypotheses; weights depend on nr. of examples with respective labels / on their distance to the query instance
- meta-features: e.g. reliability of classifiers
- summation of weights of a hypothesis
Evaluation (1)

- the same data were used as for testing the classifiers with depth respectively skeletal information individually
- k-NN classifier: best performance for the neighboring cue
- weighted k-NN classifier outperformed the standard one
- improved robustness
Evaluation (1)

Comparison of the individual inputs and their combination via $F_1$-score

- depth images
- skeletal information
- neighboring cue
Evaluation (1)

Comparison of the individual inputs and their combination via

- depth images
- skeletal information
- neighboring cue
Evaluation (1)

<table>
<thead>
<tr>
<th></th>
<th>depth images</th>
<th>skeletal data</th>
<th>combined inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$-score</td>
<td>0.027499</td>
<td>0.837523</td>
<td>0.805485</td>
</tr>
</tbody>
</table>

**Table:** Comparison of the individual inputs and their combination via neighboring cues for the weighted 5-NN classifier, trained on data from users with similar figures and tested on data from the same users, but in a different environment.
Overview

Introduction
Set-Up
Multiple Inputs & Adaptivity
  Combining Multiple Inputs
    Adaptivity
  Integrated System
Conclusion
Online Learning

- goal: recognition of gestures under changed circumstances
- classifiers try to recognize query instances and are told the correct label afterwards to update their model
- no online version for SVMs (they need to be retrained every time new training are added) ⇒ k-NN classifiers
- different points when to learn showed no apparent effects
Evaluation (2)

- 5-NN classifier
- trained on depth images from users with similar figures and tested on depth images from the same users, but in a different environment
- online learning after each misclassification
- training data and the test data of iteration 1 the same as for previous tests
- similar tests in the remaining iterations, but with newly sampled test data
Evaluation (2)

The graph shows the change in $F_1$-score and the number of training data and retraining iterations over iterations. The $F_1$-score improves with each iteration, indicating better performance of the gesture recognition system. The number of training data and retraining also increases, suggesting that the system is adapting and learning from more data.
Evaluation (2)

![Graph showing average classification time, number of training data, and number of retraining over iterations.](image)

- **Average classification time** increases with each iteration.
- **Number of training data** also increases with each iteration.
- **Number of retraining** shows a slight increase with each iteration.

The graph illustrates the performance of the gesture recognition system as it adapts over time, demonstrating improved efficiency with each iteration.
Overview

Introduction
Set-Up
Multiple Inputs & Adaptivity
  Combining Multiple Inputs
  Adaptivity
  Integrated System
Conclusion
Multiple Inputs

- Hypotheses Verification
- additional weighting cues are enabled by online learning: the experience of a classifier (the number of examples it has been trained with)
Adaptivity

- online Learning
- what examples to learn
- previously: all misclassified ones
- alternative: misclassified examples as soon as the fusion result is wrong, too
Evaluation (3)

- 5-NN classifier
- trained on data from users with similar figures and tested on data from the same users, but in a different environment
- depth images and skeletal data combined via neighboring cue
- online learning after each misclassification
- training data and the test data of iteration 1 the same as for previous tests
- similar tests in the remaining iterations, but with newly sampled test data
Evaluation (3)

- $F_1$-score for learning after false classifier prediction
- $F_1$-score for learning after false classifier and overall prediction

Graph showing $F_1$-score over iterations.
Evaluation (3)

![Graph showing the time (sec) vs iteration for learning after false classifier prediction and learning after false classifier and overall prediction.]

- Learning after false classifier prediction
- Learning after false classifier and overall prediction
Evaluation (4) - Final Test

- weighted 5-NN classifier
- depth images and skeletal data combined via neighboring cue
- online learning after each misclassification when fusion result false, too
- no preliminary training
- test data from users with similar figures (first ten and last ten iterations), data from users with varying figures (iteration 11 - 20), original users, but in a different environment (iteration 21 - 30) and the users with the varying figures in that environment (iteration 31 - 40)
Evaluation (4) - Final Test

![Graph showing F1-score and number of training data over iterations]

- **F1-score**
- Number of training data: depth images
- Number of training data: skeletal information

Tobias Staron
Overview

Introduction
Set-Up
Multiple Inputs & Adaptivity
Conclusion
Overview

Introduction
Set-Up
Multiple Inputs & Adaptivity
Conclusion
Summary
References
Hypotheses

- use of multiple inputs lead to improved recognition results as well as a more robust system
- system is able to adapt to changed circumstances due to online learning
- preliminary training can be omitted because of online learning

⇒ adaptive gesture recognition system that makes use of multiple inputs
Overview

Introduction
Set-Up
Multiple Inputs & Adaptivity
Conclusion
Summary
References
References 1


References II


The End

Thanks for Your Attention!