CHAPTER 7

Conclusion and further work

In this thesis we have proposed a user-independent gesture recognition system, which is capable of operating in real-time under real-world conditions, and as such could be useful as a building block in a more sophisticated human-computer interface. In the proposed system, efficient and robust extraction/representation of gestural motion patterns has been achieved by utilizing the relative-motion dependent primitive features introduced in chapter 3. The bottom-up approach for primitive features extraction in the early processing stage of our system, is combined with top-down based statistical learning (described in chapter 4) and noise-filtering and online segmentation abilities (the Dynamic Buffer Structures described in chapter 5) at the later stages. Even though having been trained with only a few samples, the proposed system has shown good generalization abilities. In test experiments conducted on several different data sets (described in chapter 6) it was observed that system's performance is robust to changes in the background and in the illumination conditions, to subjects' external appearance (body size, texture and color of the clothes), to shifts both in the horizontal plane and in depth, capable to cope with the non-uniformity in the performance speed of the gestures, etc., which are important conditions for successful application in human-computer interfaces. Also, gesture recognition in the present system is not restricted to hand gestures only – larger-scale whole body gestures can be processed simultaneously with the relatively smaller-scale hand or head gestures. No manual segmentation of any kind or use of markers, sensors, etc. is necessary. The method requires no special conditions, e.g. it is suitable for use in ordinary office environments. Also, since no domain knowledge is utilized, the method can be easily adapted and applied to other problems involving motion recognition.

Although encouraging results have been obtained for a limited number of gestures (which might be enough for certain applications), the system's performance has yet to be extensively tested with a larger gesture vocabulary, such as the one necessary for more complicated tasks like sign language, etc. Apart from the experiments mentioned in the previous section, we have also conducted some preliminary experiments involving recognition of facial expressions and obtained promising results. In this case, however, only expressions involving motion changes (i.e. not static expressions) can be recognized, due to the fact that the system in its present form is essentially blind to anything which is not moving. To overcome this limitation it might be necessary to introduce into the present system (or combine it with) some static object recognition abilities.

Presently, the method works under the assumption of a single user at a time, and the possibility for multiple users performing different gestures at the same time should also be investigated. For that purpose it might be appropriate to incorporate some form of attentive faculties into the system, or some other method should be used to pre-segment the figures of the different users in the original images, after which the present method can be applied to each of the segmented areas separately. In this case, possible mutual occlusions of the users would further aggravate the problem, and also would have to be taken into consideration. Another way to further improve the performance of the system would be to utilize some additional sources of information which can be easily provided by already existing hardware solutions, e.g. by integrating 3D-information (input from two or more cameras) for certain gestures which might be difficult to handle in 2D, or using depth-segmented images for eliminating the influence of objects moving in the background. In a similar manner (as already suggested in chapter 3), an alternative way to ensure size-invariance of the proposed method would be to use video camera(s) with zoom control.