Background. A vast number of artistic styles have been developed over the course of human history. Different styles take the same scene and show it in different ways. One way of thinking about this is to say that each of the different styles are presenting an abstracted version of the scene. The process of abstracting consists of emphasizing or de-emphasizing certain classes of information. Given that we use speciﬁc information to recognize facial expressions, it is reasonable to ask whether different artistic stylization methods might either enhance or obscure that information, thus affecting our ability to recognize expressions. The series of experiments in Wallraven et al. (2007), which are discussed in part here, provide a ﬁrst examination of the issues surrounding the what information that is emphasized in different stylistic renderings and how that relates to facial expressions.

Research question. How effective do people think these three stylization techniques are in conveying the intended expression? Do these opinions change as a function of the amount of detail preserved by the stylization technique?

Imagine, for example, that we are designing an expert system on a mobile computing platform. Space on the screen is very limited. Nonetheless, we want an avatar of a virtual expert to be present. How much of the screen space must we dedicate to the avatar and how much can we use for other items? To examine this, we need to decide what our avatar will be doing. In this case, the avatar will mostly be talking to a customer. Therefore, we can restrict ourselves to facial (or face and hand) animation and do not need to present the whole body. Likewise, it means that we are interested in conversational behaviors. Previous research and common experience have shown that a spoken statement of “surprised” that is accompanied by a neutral expression or a look of boredom has a very different meaning than one accompanied by a surprised expression. In fact, whenever spoken and facial statements conﬂict, the facial meaning tends to dominate (Carrera-Levillain and Fernandez-Dols, 1994; Fernandez-Dols et al., 1991; Mehrabian and Ferris, 1967). Before we can determine how much screen space our avatar needs in order for the facial expressions to be properly perceived and understood, we need to know how people in general recognize expressions. If we show the happy expression picture in Figure 2.9 to one person and ask that person to identify the expression (using a non-forced-choice task , see Chapter 6), we can say how well that one person is able to identify that one expression on that one trial. We would not be able to say anything about how well other people might identify the expression, let alone how well people can identify expressions in general. If we show that expression to a representative sample of the population, then we can begin to talk about how well people on average can identify that one expression. We still can say nothing about expressions in general. To do that, we would need to measure several expressions. To do this, we constructed a short list of expressions that were likely to show up in an expert system and recorded them from real people (note that videos and not static photographs were used— not only because dynamic expressions will be used in the ﬁnal expert system, but because dynamic expressions are the norm for people in the real world). This initial list had nine expressions (see Figure 2.9). This is near the lower limit of expressions necessary to be able to say something about expressions in general. Of course, if we were to use expressions only from one person, we would not be able to say anything about expressions in general. Thus, expressions were recorded from six individuals (for more on the database of expressions, see Section 5.3.3.6 or Cunningham et al., 2005).

The experimenter might choose to think about quality in very speciﬁc terms and postulate that what makes an avatar good in terms of quality is that its facial expressions are easy to recognize. That is, quality becomes recast in terms of recognition accuracy. The experimenter therefore has the different animations display a set of facial expressions, which participants have to recognize— for example, by entering the corresponding number from a list of given expressions. In such a design there is a right and a wrong answer, as the experimenter knows what expression was shown and can ﬁnd out whether the participant selected the correct expression from the list. After having gathered recognition performance data from a suitable number of participants, the experiment would yield a quality measure for each animation style, which is the average recognition accuracy for that particular style. Recognition accuracy as a variable behaves very much like any other measurable number, and therefore it is possible to calculate its average, its variance, and other properties. Another example for such a variable would be response time, or the time it took the participant to make a response after the stimulus has been shown. In the context of this experiment, response time would also be a viable quality measure, as it would surely be important how long participants take to make decisions about the facial expression that is being shown.