This paper presents a study of bilevel image similarity, including new objective metrics intended to quantify similarity consistent with human perception, and a subjective experiment to obtain ground truth for judging the performance of the objective similarity metrics. The focus is on scenic bilevel images, which are complex, natural or hand-drawn images, such as landscapes or portraits. The ground truth was obtained from ratings by 77 subjects of 44 distorted versions of seven scenic images, using a modified version of the SDSCE testing methodology. Based on hypotheses about human perception of bilevel images, several new metrics are proposed that outperform existing ones in the sense of attaining significantly higher Pearson and Spearman-rank correlation coefficients with respect to the ground truth from the subjective experiment. The new metrics include adjusted percentage error, bilevel local direction, and connected components comparison. Combinations of these metrics are also proposed, which exploit their complementarity to attain even better performance. These metrics and the ground truth are then used to assess the relative severity of various kinds of distortion and the performance of several lossy bilevel compression methods.