We present an approach for monitoring and interpreting human activities based on a novel multimodal vision-based interface, aiming at improving the efficiency of human-robot interaction (HRI) in industrial environments. Multi-modality is an important concept in this design, where we combine inputs from several state-of-the-art sensors to provide a variety of information, e.g. skeleton and fingertip poses. Based on typical industrial workflows, we derived multiple levels of human activity labels, including large-scale activities (e.g. assembly) and simpler sub-activities (e.g. hand gestures), creating a duration- and complexity-based hierarchy. We train supervised generative classifiers for each activity level and combine the output of this stage with a trained Hierarchical Hidden Markov Model (HHMM), which models not only the temporal aspects between the activities on the same level, but also the hierarchical relationships between the levels.