Multicue Visual Attention

H. Hügli, N. Ouerhani, P.-Y. Burgi

Visual attention is the ability to select within a given scene specific parts deemed important, or salient, by the observer. In human vision, visual attention is intimately linked to the rapid eye movements that tend to place the fovea onto the parts of the scene where rapid motion, bright color, etc. occur. In machine vision, there is a clear interest to use this principle for rapidly detecting salient locations of a scene so as to reduce computational complexity by focusing onto them for further analysis.
Computational models of visual attention can be thought of as processes that transform the input image flow into a sequence of output saliency maps providing spatially and temporally an objective measure for the intensity of visual attention. Figure 1 illustrates such a process, which consists in three main stages:

1. Sensed visual data is transformed into n multicue visual features, each being represented spatially by the so-called feature maps.

2. Each feature map is in turn transformed into the so-called conspicuity map that measures the degree by which the specific feature spatially differs from its surrounding (saliency measurement).

3. In the last stage, the conspicuity maps are fused together by a competitive integration process to finally yield the saliency map.

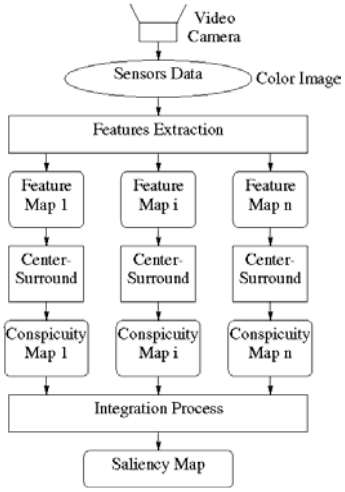

Figure 1
Multicue visual attention system

The visual attention system developed in this project takes advantage of cues available from conventional video sources (that is, motion, color and contrast), in addition to depth[1)]. As each feature contributes differently to the saliency map, they must be selected in a way that optimizes a global criterion based on a specific quality/cost ratio measure. This criterion should weigh more the motion channel as this cue strongly contributes to attracting attention in a visual scene. Because motion cues play such an important role in the proposed system, special care must be provided to robustly extract this key feature. For instance, methods simply based on image differences are not recommended because they cannot provide the motion's direction. Conversely, refined methods based on relaxation principles are also rejected because of their complexity. As a good compromise, a multiresolution scheme based on standard motion-detection methods has been selected for recovering a wide range of velocities.
Another aspect of main concern is related to the natural capacity of the human visual system to process visual information in channels with different spatial sensitivity. This calls for an extension of the visual attention model to treat multiple spatial scales. Practically, this is realized by representing each visual feature by a vector, where each component represents the feature at a given scale.

Figure 2
Three main spots of attention of a dynamic real scene

The software version of the multicue visual attention system runs in near real-time conditions on color video sequences. It typically delivers sequences of saliency maps. For visualization purposes, the saliency map is transformed into an ordered list of saliency spots overlaid on the video images (see Figure 2). Most experiments performed on various visual scenes have confirmed the importance of a multi-cue attention system based on motion, color and contrast. The continuation of this work will mainly involve the integration of this simulated visual attention system on a dedicated VLSI platform so as to extend its performances in real-time applications.

---

1) N. Ouerhani and H. Hügli, "Computing visual attention from scene depth", Proc. ICPR 2000, IEEE Computer Society Press, pp. 375-378, Sept. 2000