Image Motion Analysis Made Simple and Fast, One Component at a Time

Peter J. Burt
David Sarnoff Research Center
Subsidiary of SRI International
Princeton, NJ 08543-5300 USA

Abstract

Real-time vision can be organised as a sequence of discrete observations, or *focal probes*, that gather scene information critical to the vision task. Such selective analysis simplifies computations by isolating signal components, and reduces the data load by avoiding unimportant image detail.

I outline a sequential approach to the analysis of image motion. The approach uses selection mechanisms analogous to foveation and eye tracking in human vision to isolate motion components one at a time. Each observation estimates motion of a single patch undergoing simple translation. But a sequence of observations can interpret complex patterns containing discontinuities and transparency.

Computations are implemented within an image pyramid to provide direct selection of signal components in space, time, resolution, and velocity.

1 Introduction

It has often been observed that systems capable of performing challenging vision tasks in real time will need to perform a great many computations in parallel. Only through parallelism will it be possible to complete required processing in time to respond to events in a constantly changing visual world.

But real-time vision also requires focus-of-attention strategies that are inherently sequential. Objects and events of interest tend to be localised in space and time. To be efficient, a system must direct its sensing and computing resources to just those regions of a scene that are critical to its task, while avoiding unimportant detail. To achieve such selective processing, complex tasks must be performed in small steps. Then partial results can be used at each moment in time to direct subsequent observations and analysis.

While focus-of-attention analysis avoids wasting effort on unimportant image regions, it also simplifies the processing that must be performed in regions of interest. Processing steps are performed on isolated segments of the image signal that are much less complex than the signal as a whole. Relatively simple algorithms can often obtain precise results very fast.

In this paper I outline a dynamic, focus-of-attention, approach to the analysis of image motion. The system builds and modifies its interpretation of motion in the scene through a sequence of observations, or *focal probes*. Each observation is directed to a region of the scene that is expected to contain important information. A basic *selective stabilisation* algorithm is applied in that

BMVC 1991 doi:10.5244/C.5.1

region to determine the major component of motion. Subsequent observations provide additional components and these are assembled into an interpretation of scene motion.

Selection mechanisms analogous to eye movements in humans are used to isolate signal components in space, time, resolution, and velocity. Control of these signal parameters is achieved by implementing the basic motion estimator within a pyramid structure.

2 A System for Dynamic (Focus-of-Attention) Analysis

The overall organisation of a dynamic vision system is suggested in Figure 1. Again, interpretation of the scene is organised as a sequence of observations. Each observation itself entails a sequence of steps that are indicated by boxes in the diagram. Inputs to the system include both camera signals and a specification of the vision task to be performed. The system builds and maintains an internal world model that contains its current knowledge of the observed scene. As events unfold in the world, and as the needs of the system evolve, the system seeks to update its world model by gathering additional information. This is done through further directed observations.

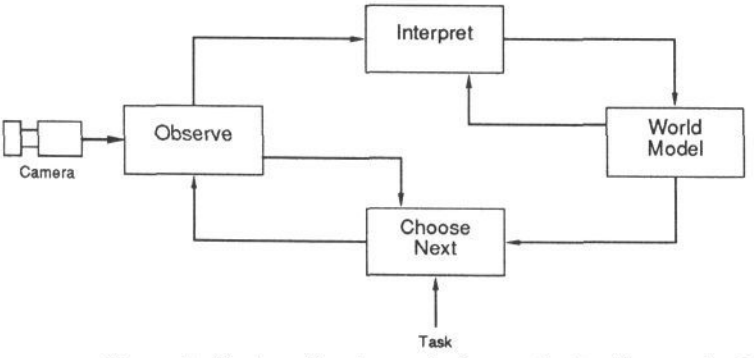

Figure 1: System for dynamic focus-of-attention analysis.

The cycle begins with the "choose next" box. Here a decision is made on what observation is most important for the system to undertake at the current moment in time. The decision is based on the task requirements, the system's world model, and alerting signals.

The designated observation is then carried out. This may entail manipulation of the sensor to shift its direction of gaze or to track an object. It also entails selective processing of the sensory signal. Here the basic computation performed is motion estimation, but more generally an observation may entail any of a number of early vision functions that require signal level computations.

Results of the observation are then interpreted in terms of objects and events in the physical world. Interpretation is based in large part on the current contents of the system's world model, and results of the interpretation are used to update that model.

Once new information has been entered into the model the process begins again, with the determination of the next observation to be made.

In the present paper I am concerned primarily with computations performed in the "observation" box that can serve motion analysis. It is appropriate, nonethe-less, to show how this module fits into an overall vision system. The basic computation is one that isolates and estimates a single motion component at a time. A larger system is assumed that will assemble these components into an interpretation of scene motion.

3 Selection Mechanisms in Motion Analysis

A basic requirement for selective image analysis is that it be possible to isolate segments of the incoming image signal that may contain essential information. In the case of motion analysis there are at least four mechanisms for signal selection: isolation in space, isolation in time, isolation in resolution (or scale), and isolation in velocity.

These may be understood in part by analogy to human vision. A human selects the spatial regions of the visual world he observes, and the time intervals for these observations, through a sequence of saccadic eye movements. Similarly, a human can isolate an object moving at a particular velocity by tracking it with his eyes. The graded resolution of the eye itself, as well as the spatial-frequency tuned "channels" within the brain provide mechanisms for controlling resolution.

These same selection mechanisms can be implemented in computer vision both through mechanical and optical control of the sensor and electronic processing of the resulting signal. Electronic processing provides flexibility that is not available to humans. Multiple *focal analysis regions* can be defined at any moment in time, and these may be moved independently of one another and changed in size and resolution from one moment to the next.

The pyramid data structure provides a natural framework for selective signal processing. Focal analysis regions are defined and shifted within the visual scene and resolution is controlled simply by shifting analysis to corresponding regions-of-interest within a pyramid representation of the image. This is suggested in Figure 2. Here a sequence of observations made of a road scene are shown on the left. Motion analysis of a typical observation sequence would begin at low resolution within a large analysis region and then move to high resolution within small analysis regions. The corresponding data within a pyramid is shown on the right.

The pyramid data structure also provides a natural means for electronically controlling the range of velocities selected for analysis. Motion estimators, such as the one that will be described in the next section, typically can detect image motion only when frame to frame displacement is not too large. This velocity limit depends on the wavelength, or spatial frequency content, of the image signal being processed.

This relationship is shown in the spatial/temporal frequency domain in Figure 3. A pattern moving at uniform velocity v appears as a tilted plane with slope $-v$ in this diagram. Two such patterns are indicated in the diagram, moving at different velocities.

If the estimator is implemented within a pyramid, then this input signal

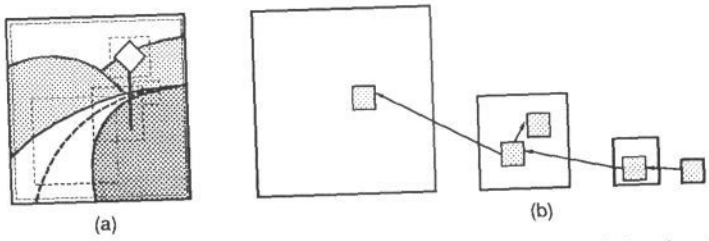

Figure 2: Control of the position, extent, and resolution of the focal analysis region within a pyramid structure.

will be restricted to a band of frequencies corresponding to the pyramid level used. These bands are indicated by vertical lines in the diagram. Temporal frequencies are also limited by the computation to be within a band bounded by $w = 0$ and $|w| < \frac{1}{2\tau}$ (τ is the time between successive frames). The intersection of the spatial and the temporal pass bands defines a *selection band* shown as the shaded region in the figure. The motion estimate is based on that portion of the signal that falls within the selection band. Thus no estimate is obtained if a particular pattern is moving so fast that its spectrum falls outside the band. However, performing analysis at lower resolution levels of the pyramid moves the selection band towards the origin, so that it will accommodate faster motion.

Motion selection occurs when there are two differently moving patterns in the analysis region, one of which falls within the selection band while the other falls outside the band. This is the case illustrated in the figure.

The coarse-fine motion estimation algorithm described below automatically finds conditions that separate two components when they occur within the analysis region.

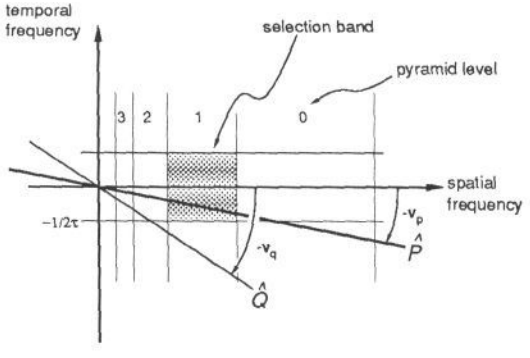

Figure 3: Velocity selection within a pyramid structure.

4 Estimating Component Motion

While motion in the visual world can be quite complex, when viewed locally within a small window it generally appears as simple, uniform translation. Exceptions occur at motion boundaries and in regions of transparency. There the scene can generally be modelled as the sum of two differently moving patterns. We define a *motion component* as an image pattern that can be modelled as undergoing uniform translation when viewed within an appropriately sized window in space and time, and at an appropriate resolution. The complex motion within the scene as a whole is modelled as a patchwork of these basic components.

Motion components defined in this way may overlap within the image. The extent of a given component depends on the resolution at which an image is observed. For example, a tree observed at coarse resolution from a moving car may appear to be translating uniformly past the observer. But when the tree is observed at higher resolution, individual branches may appear to move relative to one another. Components also overlap in the case of transparency, where one pattern, such as a reflection, appears superimposed on another.

Motion analysis may be formulated as a sequence of observations in which estimates are obtained for one component of motion at a time. The motion estimator, a process implemented in the "observation" box in Figure 1, determines the translational component of motion within the region designated by the "choose next" box. The motion estimation process may also provide an indication of deviations from uniform translation within the analysis regions.

As an example, component motion estimates can be estimated effectively through a *selective stabilisation* procedure [2]. This type of motion estimation algorithm determines motion by finding that displacement of the first image of a pair that brings it into alignment with the second. The shifted first image is then "stable" with respect to the second.

In the stabilisation algorithm, a motion estimate is obtained as a sequence of refinement steps. At step k the first image is shifted towards the second by a displacement corresponding to estimated velocity v_{k-1} obtained at the previous step. A computation is then performed to determine residual motion, Δv_k. The velocity estimate is updated, $v_k = v_{k-1} + \Delta v$, and the cycle is repeated.

An estimator that seeks to minimise the mean squared difference between frames is given (in one spatial dimension) by [5]:

$$\Delta v_k = -\frac{\int_R \frac{\Delta I}{\Delta x}\frac{\Delta I}{\Delta t}}{\int_R \frac{\Delta I}{\Delta x}\frac{\Delta I}{\Delta x}}.$$

This estimator can be represented in the frequency domain by [4]:

$$\Delta \hat{v}_k = -\frac{s}{2\tau}\frac{\int sin(\pi s u) sin(2\pi\tau w)|\hat{I}(u)|^2 du}{\int sin(\pi s u)^2 |\hat{I}(u)|^2 du}.$$

This representation reveals the dependance of the estimate on spatial and temporal frequency characteristics of the signal, as shown in Figure 3. If analysis is performed at pyramid level ℓ, then the power spectrum $|\hat{I}(u)|^2$ is confined to a corresponding frequency band. The term $sin(2\pi\tau w)$ may be interpreted

Figure 4: Selective motion analysis used to detect a moving person from a moving vehicle.

roughly as a temporal filter with pass band between $w = 0$ and $|w| = \frac{1}{2\tau}$. Spectral energy sums incoherently for $|w| > \frac{1}{2\tau}$.

Again, it is expedient to implement the motion estimator within a pyramid structure. The computation normally starts at a relatively low resolution level where the estimator can accommodate a wide range of velocities, then proceeds in steps toward high resolution where precise motion estimates are possible, but only for a very narrow range of velocities.

An important property of this motion estimation procedure is that it tends to select just one motion component even when two differently moving patterns occur within the analysis region. Once one motion has be determined, the corresponding pattern can be largely removed from the signal through a shift and subtract procedure. The second motion is obtained by applying the estimator to a sequence of difference images [1].

5 Examples

Two examples will illustrate selective analysis of image motion. The first is an application to vehicle navigation where motion can be used to detect other moving vehicles as well as obstacles in the road [3]. Precise motion analysis need only be performed in the portion of the camera's field of view that contains the road. However, this analysis can be quite challenging because small differences in the motion of an object and its background must be detected while the object and background appear to move rapidly due to the observer's own motion. Sensitivity for differential motion is achieved through selective stabilisation: the scene is first stabilised within the analysis region, then detailed analysis reveals small relative motions of interest.

Figure 4 is one frame of a sequence taken from a car moving down a rough country road. The dominant motion in the resulting video is due to camera bounce. In addition, there is significant parallax motion of the trees along the side of the road. Finally, there is motion of a person walking across the road in the distance. The task of the vision system is to detect potential hazards to driving, such as the person in the road.

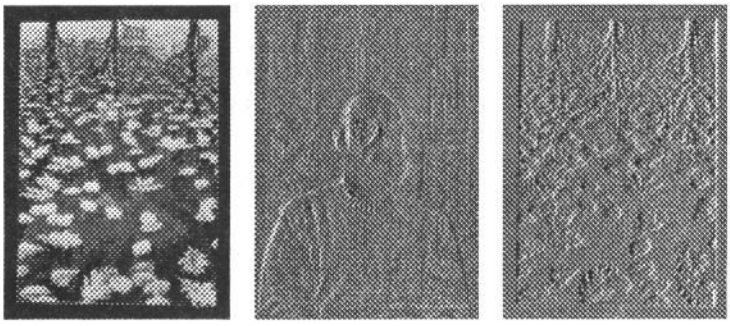

Figure 5: Selective motion analysis is used to separate the image of a picture on the wall from the reflected image of a person looking at the picture.

A three step procedure can be used to isolate the moving person. First, an observation of the entire scene gives the system an estimate of image translation due to camera bounce. The sequence is effectively stabilised based on this motion. Second, an observation made within an analysis region centred on the road in the distance stabilises this region in isolation from the differently moving foreground trees. Finally, difference images and "change energy" computed between frames of the stabilised road reveal motion of the person. This change energy image is shown in Figure 4b as an inset into the original image frame.

The second example shows selective motion analysis used to separate differently moving transparent patterns [1]. The source image sequence is of a picture hanging on a wall with the superimposed reflection of a person observing the picture. These two patterns are moving with respect to one another in the video sequence. One image in the original sequence is shown in Figure 5a. Application of the motion algorithm to this sequence first yields an estimate of motion of the hanging picture. If this estimate is used to construct a sequence of difference images, a second application of the motion algorithm, now to the difference sequence, yields an estimate of motion of the reflected image. (These steps can be repeated to further refine the estimates of both motions) Difference images are shown in Figure 5b and 5c. Note that the reflected person is hardly visible in the original image, but is clearly revealed when the video sequence is stabilised with respect to the picture and difference images are formed. This example is particularly interesting as it shows component selection in the velocity domain. The two patterns cannot be separated spatially.

6 Summary and Observations

I have outlined an approach to image motion analysis that interprets complex motion one component at a time. This approach is particularly suited for real-time vision because it provides a means for selecting the critical regions of a scene in which detailed analysis should be performed. The system maintains efficiency by directing its resources to just these regions. At the same time component selection reduces the complexity of data processed, so simplifies the computations that need to be performed at each observation. Although observations are made one at a time, the resulting analysis can be both simple

and fast.

Finally, it is interesting to note that there is growing interest, particularly within the "active vision" community, in camera systems that can achieve signal selection through mechanical and optical means. For example camera heads are being developed that allow a vision system to mechanically rotate the camera to shift gaze and track objects. Novel camera sensors are also being developed that have graded resolution from the centre to the periphery, as in the fovea of the human eye.

To date, electronic processing has provided the most successful means for rapidly shifting visual processing within a camera's field of view. A general purpose "vision front end" should combine these mechanical, optical, and electronic means of signal selection. Processing within the camera head could include a pyramid generating element, for example, to allow ready control of resolution and field of view. It could also include a basic motion estimation element, such as that described here, to allow effective isolation of signals in the velocity domain through selective stabilisation. Several components of such a general vision front end are now under development at David Sarnoff Research Center, including a chip to perform pyramid processing, and hardware for video rate motion analysis.

Acknowledgements

Many individuals have contributed to the ideas and results presented here. These include P. Anandan, James R. Bergen, Keith Hanna, Rajesh Hingorani, Raymond Kolczynski, and Jeffrey Lubin of David Sarnoff Research Center, and Shmuel Peleg of the Hebrew University.

References

[1] J. R. Bergen, P. J. Burt, Rajesh Hingorani, and Shmuel Peleg, Computing two motions from three frames, **Proc. 3rd International Conf. on Computer Vision**, pp. 27-32, 1990.

[2] P. J. Burt, J. R. Bergen, R. Hingorani, R. Kolczynski, W. A. Lee, A. Leung, J. Lubin, and H. Shvaytser. Object tracking with a moving camera, an application of dynamic motion analysis, **IEEE Workshop on Visual Motion**, pages 2-12, Irvine, CA, March 1989.

[3] P. J. Burt, J. R. Bergen, R. Hingorani, and P. Anandan. Dynamic Analysis of Image Motion for Vehicle Guidance, **Proc. IEEE International Workshop on Intelligent Motion Control**, pages IP-75–IP-82, Istanbul, Turkey, August 1990.

[4] P. J. Burt, R. Hingorani, R. Kolczynski. Mechanisms for Isolating Component Patterns in the Sequential Analysis of Multiple Motion, to appear in **IEEE Workshop on Visual Motion**, Princeton, NJ, October 1991.

[5] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision, **Image Understanding Workshop**, pp. 121-130, 1981.