An Introduction to Vectorization and Segmentation

David S. Doermann

Language and Media Processing Laboratory
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742-3275
doermann@cfar.umd.edu

Abstract. In this text, we briefly overview some of the basic issues and trends in the vectorization and segmentation of graphics, and provide a short list of relevant references on these topics. It is intended as an introduction to various papers on related topics, elsewhere in this collection.

1 Introduction

The field of graphics understanding has developed in recent years as a viable and thriving sub-discipline to document image understanding. Research progress is constantly being reported in major journals and at tutorials, workshops conferences internationally.

A special track on vectorization and segmentation was held as part of the recent Second IAPR Workshop on Graphics Recognition. The problems of vectorization and segmentation are fundamental to the field of graphics processing because they are used to provide the basic representation on which understanding algorithms operate. A scanned image of a drawing is initially an ordered array of pixels representing an average intensity over a small region. The ability to build up a representation from these individual pixels which exploits relationships such as local proximity and highlights the structure of the underlying components is important for the extraction of features during interpretation and recognition. Although the nature of this representation is application dependent, it is clear that deriving a representation of linear features (vectorization) and providing a way to divide components into basic graphic primitives (segmentation) is essential.

Vectorization is a task that is applied to perhaps a much larger domain then it was originally intended. Although vectors are most appropriate for the representation of lines and line-like components, authors have also used vectorization for other types of components, primarily as an intermediate step in the process of converting to a more expressive representation. For structures such as continuous curves and filled regions, vectorization is not ideal, but the vector representation is efficient and easy to work with, so researchers have relied on it as a basic tool for analysis and used the configuration of vectors to recognize primitives.
The problem of graphics segmentation is equally important. Segmentation is a broad term which is associated with many tasks, but in general is used to describe the separation of drawings into fundamental components. This includes, for example, the separation of lines and curves which interact, the separation of text from graphics and the division of higher level composite objects into basic structures such as lines and arcs.

In this text, we briefly overview some of the basic issues and trends in the vectorization and segmentation of graphics, and provide a short list of relevant references on these topics.

2 Vectorization

Classical vectorization methods attempt to convert raster images into a collection of vectors, each represented by either a two points or by a single point, a direction and a length. As mentioned previously, a vector based representation is appropriate for line features, but often inappropriate for representing non-linear components such as arcs or regions. For this reason, vector representations are rarely used on an entire image. One exception is in graphic image compression, when a precise representation is not necessary and further automatic processing is not anticipated. Polygonal estimates of arcs are often acceptable. Another exception is when the vectorization is simply an intermediate step in deriving a more appropriate representation.

The ability to derive a usable vector representation is often hampered by a number of factors, one of the most important of which is the quality of the image. As the image degrades, assumptions which could originally be made about the local behavior of features become less reliable. For example, as the image begins to degrade and lines become broken, we are not able to provide a continuous vector representation. Digitization can introduce noise along edges and cause problems for contour methods or methods which attempt line following.

Humans have an impressive ability to abstract and provide high quality vectorization in the presence of noise, in part because we can perform recognition in conjunction with feature extraction and interpolate beyond missing or corrupted data. In automated systems, care must be taken to decide if we should perform vectorization to precisely represent the image or if we should attempt to approximate the original features which most likely gave rise to the image at the risk of introducing more distortion. As systems mature, they will be able to incorporate higher level interpretation to deal more robustly with degradation.

In addition to degradation, there are factors inherent in the data which must be considered as well. The intersection of lines cause difficulty for any approach which tries to estimate lines from small areas of the image, because locally, these regions do not look like line segments. The goal is typically to recover the original line trajectory. Scale is another problem. Having features at different scales can result in one linear component being interpreted as a line, and other larger line components being interpreted as filled regions. Care must be taken to consider both local properties and global context.