INTERNATIONAL WORKSHOP ON INDUSTRIAL APPLICATIONS OF
MACHINE INTELLIGENCE AND VISION (MIV-89), Tokyo, April 10–12, 1989

IMAGE SEGMENTATION BY LOCAL FEATURE BASED CLUSTERING FOR UNDERSTANDING NATURAL SCENE

Mutsuhiro Terauchi, Mitsuo Nagamachi, Koji Ito and Toshio Tsuji

Faculty of Engineering, Hiroshima University
Shitami, Saijo-cho, Higashi-hiroshima 724 Japan

Abstract

In this paper an approach to segment a gray level image is presented in order to reconstruct the 3-D shape of the unconstrained objects. In analyzing a natural image, we can not utilize any heuristic rules that constrain the degree of freedom (DOF) of reconstruction of the scene. Therefore we use local feature based clustering (LFBC) which utilizes the local distribution of the features. This clustering is based on an image itself and is considered as *object oriented processing*. The edge detection problem is also discussed as an opposite problem to clustering. Finally the clustering which utilizes both lowest features (for all pixels) and the features a little higher up are discussed for their ability for exact segmentation.

1. Introduction

The reconstruction of a scene from a projected image is ill-posed problem. Various approaches are suggested for computer vision to restore the scene projected onto an image. One of these is shape from contour. This method is more effective than shape from anything else, although there do not seem to be any mathematical reasons for it [1]. But several psychophysical demonstrations show its significance [2]. In interpretation of contour one way to constrain is to treat partial space of the world, for example block world, where there are only polyhedra. If we restrict the variety of objects, it is possible to obtain a unique interpretation of the scene. We can not, however, apply this method for reconstructing natural scenes. For such a scene the constraints or rules should be based on the image of the scene itself. Our purpose is to reconstruct the 3D scene from an image using contour. But in edge extraction using a differentiated operator as before, it is difficult to obtain perfect contours. Especially when we treat a natural image, it is even more difficult to obtain them. One possibility is that if there is additional information, then we could obtain more accurate boundary lines. Although these processes may be considered time-consuming and redundant, we would be able to obtain cues which might be useful at a later stage. Here we set the gradient vector which reflects one feature of the texture as additional information for our implementation.

Humans can easily recognize the shape of an imaged surface. It seems that textured image could contain cues for the recognition of the original surface. A shape from texture theory uses the same information as a human to reconstruct the shape of the surface. To recover shape, the distortion effects of the projection must be distinguished from the properties of the texture on which the distortion acts. This means that assumptions must be made about the properties of the texture. The problem of shape from texture in the case of planes has also been studied. In general, the distortions introduced in the projection can be considered as coming from the following three effects: the effect of distance, the effect of position and the effect of foreshortening. It is clear that the orthographic projection model is only affected by the foreshortened effect and not by the rests. In this area the hypotheses of properties of the texture or texels' distribution on the surface in the real 3D world, imposes of constraints. Anyway, it is one of the key points in finding constraints to obtain a unique solution.

On the other hand if we have dual cues, one from texture and another from contour, then we may obtain a unique solution. We focus on both shape from texture and from contour, and also on the integration of their cues. To segment an image we need a boundary line or the internal region of the boundary line or both of them. For this purpose we consider image segmentation using these cues.

2. Integration of multi cues

The shape from one is computed from a particular cue previously. Fig. 1 shows the approaches of shape from one using single image. As mentioned above, shape from one cue result ill-posed problem. To integrate these cues, we must make explicit that what information is available in other processes. If there are commonly usable characteristic among the methods, it is easy to use it of course.

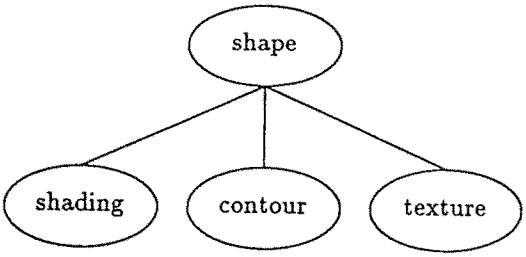


Fig. 1 Shape from an image.

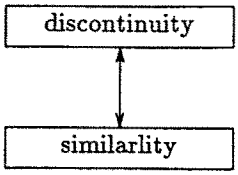


Fig. 2 Opposite features.

When we use a texture to segment an image and a contour to define the boundary, we will use these same (and opposite) features (Fig. 2).

In our approach some feature on pixels and boundary was used to segment a image into intrinsic regions. The schematic flow diagram is shown in Fig. 3.

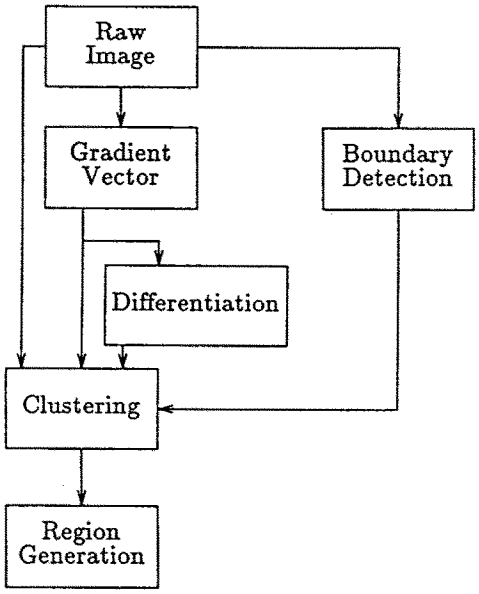


Fig. 3 Segmentation using boundary.

3. Hierarchical Feature Structure

In this section we outline the relationship between raw image intensity and various tokens. It is necessary when we utilize some tokens, for example edge or terminal, for image segmentation. Hierarchy of features and tokens in an image was first discussed by Marr [3]. He suggested that there are mainly three levels in visual information processing [3]. But they in turn can be divided into a further number of levels. They create a hierarchical feature structure in which features interact mutually. In our case we concentrate on low and intermediate levels of vision, so tokens are separated to 3 different levels as shown in Fig. 4.

3.1. Interpretation of Image Feature

Computer vision is a system of theory and techniques that represents our present understanding of how far competence in visual perception can be implemented in computing hardware. The dominant paradigm, signals-to-symbols, is one in which the raw sense data is transformed into a meaningful and explicit description of the corresponding scene by a series of inductive steps, employing progressively more abstract representation. These steps can be divided into three categories, which Marr had suggested, based on the nature of the modeling required to carry out the analysis.

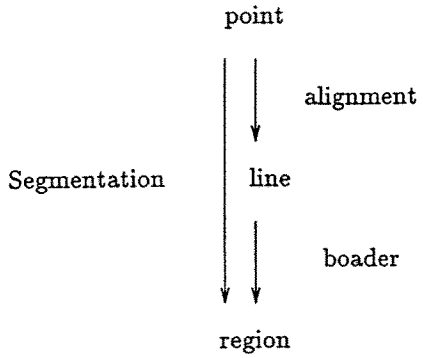


Fig. 4 Relationship betwen components and group.

In a monochrome image the whole scene is represented by intensities of pixels, from black to white. Therefore, all the features are computed from an array of pixel intensity. So, ideally, all the tokens which reflect features in a scene can be obtained from intensity level.

3.2. Resolution of Image

But how much resolution do we need for computation of image analysis? When we compute a microscopic image, we have to provide a high-resolution. When we compute only for contours or boundaries, there is no need to treat the objects accurately. It depends on the sensor's resolution or the size of the central mask for blurring and it also depends on the purpose of the computation. The pyramid of resolution is shown in Fig. 5. But does the clustering for a low resolution image facilitate the clustering for a high resolution image? If a fine image is obtained, then it is easy to transform it to coarse image by Gaussian filter. The fineness depends on filter's variance σ.

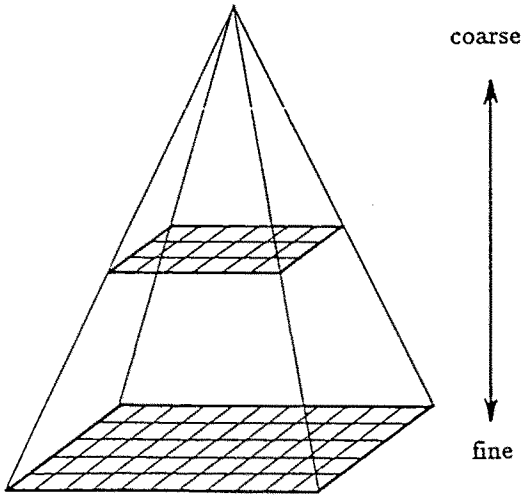


Fig. 5 Resolution of image.

4. Clustering as Segmentation

In a quantized image there are finite number of pixels. It means that the segmentation of image is realized by clustering pixels on it. When we see an image, we can recognize a group of points on the image as an object, which may exist physically, according to their similarity and the completeness of object. As the latter concerns with human knowledge about the real objects, we only take account on the former. We assumed that the pixels, which is represented by coefficient vector, are distributed on an array.

4.1. Clustering approach

One of the most powerful ways to segment an image is cluster analysis [4]. That is based on similarity between feature vectors, in which each feature is detected individually and is independent of any others. In clustering there are many way to define a similarity between clusters. It is considered as "distance" as well, for example,

- ☐ Euclidean distance
- ☐ Weighted Euclidean distance
- ☐ Minkovsky distance
- ☐ Mahalanobis distance

If we compute weighted value of a coefficient vector before clustering, then the process can be more valid, even when we adopt Euclidean distance. Here we use this Euclidean distance which define dissimilarity between clusters, and it can be considered to be physical distance in Euclidean space. Furthermore we define the offset which reflects the deviation of local feature variances. This offset is considered to act as a selector of intrinsic features in local image.

4.2. Preprocessing for Clustering

It is necessary that some of processing have finished before the clustering. At first, features (intensity, gradient vector, differentiated gradient vector) are extracted. The edge segment, which may amount to boundary line, is computed separately.

4.2.1. Feature Extraction

The main feature in an image is its intensity. It can be obtain easily directly from an image. In the next step the gradient vectors are computed using horizontal and vertical differential masks. The orientation and the scalar component of the vector are given as follows,

$$\theta = \tan^{-1}\frac{G_v(i,j)}{G_h(i,j)}$$

$$|\vec{G}| = \sqrt{G_v(i,j)^2+G_h(i,j)^2}$$

where $G_v(i,j)$ and $G_h(i,j)$ are the vertical and horizontal components of gradient vector obtained on pixel i,j. These features θ and $|\vec{G}|$ are treated separately. Furthermore we computed spatial differentiation of these θ and $|\vec{G}|$. Thus five features $f_{i,j}$, $\theta_{i,j}$, $|\vec{G}|_{i,j}$, $\theta_{i,j}'$ and $|\vec{G}|_{i,j}'$ are obtained from the image.

4.2.2. Use of Boundary

Edge detection, by finding a zero crossing or by scanning a ridge on a differentiated image, and the image segmentation, by clustering pixels can be called dualistic processing. One may generally use either of above processes, but it is desirable to utilize both processes. But how can we combine two methods into one? If edges are clearly defined, they will act as boundaries for pixel clustering. If the extracted edge is too short, then it is considered as a component of the texture. So a sufficient long edge is regarded as a boundary line. The edge is computed using the Laplacian operator. For our clustering, the boundary line is only used as additional information.

4.2.3. Normalization

The component of the coefficient vector must have been normalized in order to be even with each other.

The mean:

$$\bar{X}_n = \frac{1}{m}\sum_{k=1}^{m} X_{k,n} \longrightarrow 0$$

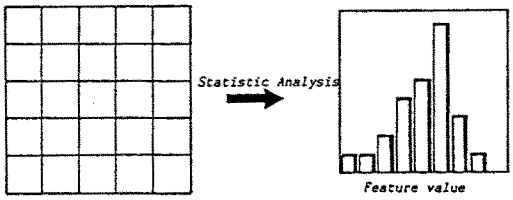
The variance:

$$s_n^2 = \frac{1}{m}\sum_{k=1}^{m}(X_{k,n} - \bar{X}_n) \longrightarrow 1$$

For whole feature the normalization procedure is done. Then we can treat all coefficient in the same order.

4.3. Detection of Local Feature Distribution

All pixels have their features as a feature vector which consists of n feature values. The value of each feature is added an offset value according to the variance of the local neighbor feature distribution around the pixel. The neighborhood is defined as 5*5 (25) pixels as shown in Fig. 6. For each feature



5*5 Neighborhood — Histogram for each feature

Compute variance

According to the variance the offset value is decided.

If $\sigma^2 \leq 0.1$ then offset = 1.0
If $0.1 < \sigma^2 \leq 0.2$ then offset = 0.6
If $0.2 < \sigma^2 \leq 0.4$ then offset = 0.3
If $0.4 < \sigma^2 \leq 0.8$ then offset = 0.1
If $0.8 < \sigma^2 \leq 1.6$ then offset = 0.05
If $1.6 < \sigma^2$ then offset = 0

Fig. 6 Enhancement of features.

the variance is computed, and to which an offset value is added according to the value of the variance. The process is performed for all pixels. This enhancement of features is considered as attention to local characteristics of feature variance. The operation enables us to realize object oriented processing.

4.4. Clustering in Feature Space

The features obtained from the image are mapped into the feature space according to the feature axis. An arbitrary pixel on the image is mapped into the coordinates space with n dimension which contains n features' axes. They are clustered to several groups in the multi-dimension space. In cluster analysis similarity, or distance between data or variance of data is used for segmentation of all data. However it does not guarantee whether the result of clustering is significant or not. Namely, whether the classification is valid or not depends on an interpretation of the result. It is the advantage of this method that which needs no criteria for classification.

Let us consider that the two data X'_i and X'_j, which are pixels of the image, have n features and they are represented as coefficient vectors as follows :

$$X'_i = (X_{i1}, X_{i2}, \cdots, X_{in})$$
$$X'_j = (X_{j1}, X_{j2}, \cdots, X_{jn}) .$$

It is assumed that each feature is continuous function and whole data are mapped in N dimensional Euclidean space. Then the distance between pixel I and J is given by

$$d_{ij}^2 = \sum_{k=1}^{n} (X_{ik} - X_{jk})^2 .$$

Each feature is normalized as described in Section 4.2.3. Euclidean distance in N dimensional space is used for the classification, and the furthest neighbor method is used for the calculation of the distance. In this method, the space of clustered group is not so different each other, and solitary group, which contains only one pixel, is seldom generated. The clustering procedure is iterated until the clustering have got to be valid. In our implementation the number of iteration is decided as $m^{0.5}$ heuristically, where m is the number of pixel. Then the clustered groups are returned back to the image plane. The process is shown in Fig. 7.

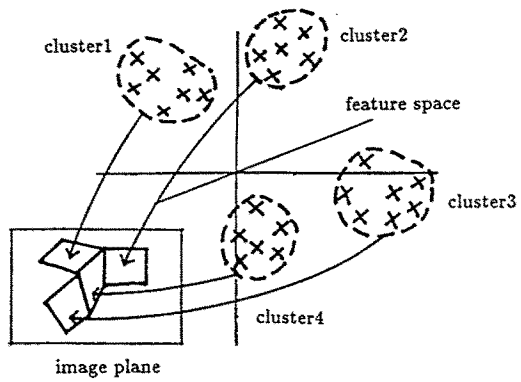


Fig. 7 Clustering in feature space.

5. Discussion

We are trying to find the intermediate level processing of vision. We have computed the segmentation of the image which amount to surfaces in the 3D scene. When we see the objects and try to know their structures, we obviously use not only the texture on surface but also its reflectance, i.e. a sense of materials. Therefore it is necessary to find the global feature of gray level images furthermore. It is also considered that we use something like a priori information as common knowledge.

The intermediate level of vision that Marr had introduced is considered as a process to extract depth from an image in order to obtain 2.5D sketch [3]. The extraction of depth from a two dimensional image generally needs not only physical law but some latent knowledge. We can not obtain the macro structures in an image only by using the bottom up processing [5]. The problem is how to combine the high level knowledge with the bottom up data efficiently in a top-down process. The schema which was proposed by Arbib is one of their solution as an idea [6].

It is popular to utilize the cluster analysis in the area of of computer vision. We computed clusters which are considered as independent region in 3D scene. A large amount of computation for the extracting various features from an image and for the computation distances between each cluster which is represented by n dimensional feature vector. Since the computation is based on local information processing, it could be executed by parallel distributed processing machine (PDP) [7]. In recent there are many research on special hardware for massively parallel or hierarchical machine [8]. If they are available, we can get more performance for it.

6. Conclusion

The clustering in which each area amounts to surfaces in 3D scene is studied. The segmentation is based on clustering in the feature space where features of pixel in an image are mapped. The additive information for clustering is the edge which is considered as contour in an image. In a natural image we can not make assumption for constraints on the shape in the scene or in the image. We would not make any assumption and any model for the scene. In the case when features in the image change smoothly, and when the image contains few texture, especially when the intensity change of the texture is smooth, we can obtain good segmentation. In our segmentation the edge that is sufficient long and is considered to be boundary line is only used. It is desired the edge which is contained in texture area is also utilized for further segmentation.

Acknowledgment

The authors wish to thank Ms. Burt for help to proofread this paper.

References

[1] Waltz D., *Understanding Line Drawings of Scenes with Shadows,* in P.H.Winston (eds.), The Psychology of Computer Vision, McGraw-Hill, New York (1975).

[2] Aloimonos J., *Visual Shape Computation,* Proc. IEEE, vol.76 No.8 (1988) pp.899-916.

[3] Marr D., *VISION,* W.H.Freeman, San Francisco, (1985) pp.19-38.

[4] Fukunaga K., *Introduction to Statistical Pattern Recognition,* Academic Press, San Diego, (1972) pp.323-354.

[5] Fischler M. and Firschein O., *Intelligence The Eye, the Brain, and the Computer,* Addison Wesley, Massachusetts, (1987) pp.239-280.

[6] Arbib M., *Schemas and Perception* in Pattern Recognition by Humans and Machines, Academic Press, New York, (1986) pp.121-157.

[7] Rumelhart D. et.al., *Parallel Distributed Processing,* MIT Press, Massachusetts, (1986) pp.323-354.

[8] Clippingdale S. et.al., *Motion Estimation for Video Bandwidth Compression using a Heterogeneous Pyramid Image Processing,* Proc. IAPR Workshop on CV, Tokyo, (1988) pp.24-27.