Scene-Centered Description from Spatial Envelope Properties

Aude Oliva[1] and Antonio Torralba[2]

[1] Department of Psychology and Cognitive Science Program
Michigan State University, East Lansing, MI 48824, USA
aoliva@msu.edu
[2] Artificial Intelligence Laboratory, MIT
Cambridge, MA 02139, USA
torralba@ai.mit.edu
BMCV 2002,LNCS 2525, pp.263-272
Eds: H. H. Bulthoff, et al. Springer-Verlag, Berlin Heidelberg

Abstract. In this paper, we propose a scene-centered representation able to provide a meaningful description of real world images at multiple levels of categorization (from superordinate to subordinate levels). The scene-centered representation is based upon the estimation of spatial envelope properties describing the shape of a scene (e.g. size, perspective, mean depth) and the nature of its content. The approach is holistic and free of segmentation phase, grouping mechanisms, 3D construction and object-centered analysis.

1 Introduction

Fig. 1 illustrates the difference between object and scene-centered approaches for image recognition. The former is *content*-focused: the description of a scene is built from a list of objects (e.g. sky, road, buildings, people, cars [1, 4]). A scene-centered approach is *context*-focused: it refers to a collection of descriptors that apply to the whole image and not to a portion of it (the scene is man-made or natural, is an indoor or an outdoor place, etc.). Object and scene-centered approaches are clearly complementary.

Seminal models of scene recognition [2, 8] have proposed that the highest level of visual recognition, the identity of a real world scene, is mediated by the reconstruction of the input image from local measurements, successively integrated into decision layers of increasing complexity. In this chain, the role of low-level and medium levels of representation is to make available to the high-level a useful and segmented representation of the scene image. Following this approach, current computer vision models propose to render the process of "recognizing" by extracting a set of image-based features (e.g. color, orientation, texture) that are combined to form higher-level representations such as regions [4], elementary forms (e.g. geons, [3]) and objects [1]. Scene identity level is then achieved by the recognition of a set of objects or regions delivered by the medium level of processing. For instance, the medium-level visual representation

of a forest might be a "greenish and textured horizontal surface, connected to several textured greenish blobs, and brownish, vertically oriented shapes" ([4]). High level processes might interpret these surfaces as a grass land, bushes and trees [1]).

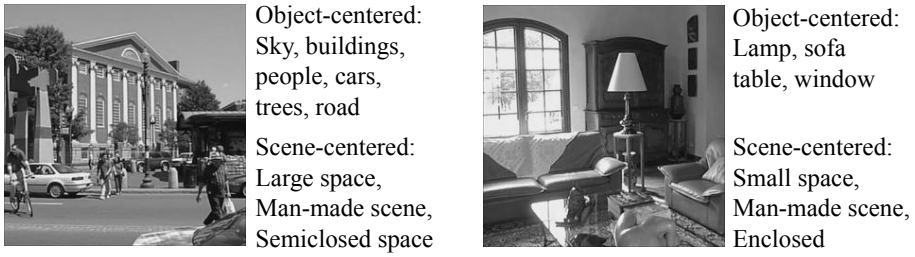

Fig. 1. Example of object-centered and space-centered descriptions.

On the other hand, some image understanding studies suggest that human observers can apprehend the meaning of a complex picture within a glance [11] without necessarily remembering important objects and their locations [7, 13]. Human observers are also able to recover scene identity under image conditions that have degraded the objects so much that they become unrecognizable in isolation. Within a glance, observers can still identify the category of a scene picture with spatial frequency as low as 4 to 8-cycles/image [15, 9]. Scene meaning may also be driven from the arrangement of simple volumetric forms, the "geons" [3]. In both cases (blobs or geons), detailed information about the local identity of the objects is not available in the image. All together, those studies suggest that the identity of a real world scene picture may also be recovered from scene-centered based features not related to object or region segmentation mechanisms.

In an effort to by-pass the segmentation step and object recognition mechanisms, a few studies in computer vision have proposed an alternative scene-centered representation, based on low-level features extraction [16, 20], semantic axes [10, 14] or space descriptors [10, 18]. Common to these studies is the goal to find the basic-level category of the image (e.g. street, living room) directly from scene descriptors bypassing object recognition as a first step. As an illustration, a scene-centered representation based on space descriptors [10, 18] could resemble the description provided in Figure 1: the picture of the street represents a man-made, urban environment, a busy and large space, with a semi-closed volume (because of the facades of the building). Such a scene-centered representation could be built upon space properties correlated with the scene's origin (natural, outdoor, indoor), its volume (its mean depth, its perspective, its size), the occupancy of the volume (complexity, roughness), etc. These spatial scene-centered characteristics happened to be meaningful descriptors highly correlated with the semantic category of the image [10, 18].

This paper is dedicated to the description of a scene-centered representation, at the medium level of visual processing. Our goal is to generate a meaningful description of real world scene, based on the identification of simple spatial properties. In the following sections, we describe how to compute a set of volumetric properties of a scene image, as they would appear in the 3D world, based on computations made on 2D images. The resulting space-centered scheme is independent of the complexity of the scene image (e.g. number of objects and regions) and able to provide multiple level of scene description (from superordinate to subordinate levels).

2 Scene space descriptors

In order to identify the candidate descriptors of real world scenes, we ran a behavioral experiment similar to the procedures used by [12, 6] for classifying textures. For example, the perceptual properties suitable for representing the texture of a forest may be its orientation, its roughness, and its homogeneity [12, 6]. These properties are meaningful to an observer that may use them for comparing similarities between two texture images. As a scene is inherently a 3D environment, in [10], we asked observers to classify images of scenes according to spatial characteristics. Seventeen observers were asked to split 81 pictures into groups. They were told that scenes belonging to the same group should have similar global aspect, similar spatial structure or similar elements. They were explicitly told not to use a criteria related to the objects (e.g. cars vs. no cars, people vs. no people) or a scene semantic groups (e.g. street, beach). At the end of each step, subjects were asked to explain the criteria they used with few words (see [10] for a detailed explanation).

The initial list of space descriptors described in [10] is given below. In this paper, we propose to use these perceptual properties as the vocabulary used to build a scene-centered description. We split the descriptors used by the observers in two sets:

The descriptors that refer to the volume of the scene image are:

- *Depth Range* is related to the size of the space. It refers to the average distance between the observer and the boundaries of the scene [18].
- *Openness* refers to the sense of enclosure of the space. It opposes indoor scenes (enclosed spaces) to open landscapes. Openness characterizes places and it is not relevant for describing single objects or close-up views.
- *Expansion* represents the degree of perspective of the space. The degree of expansion of a view is a combination of the organization of forms in the scene and the point of view of the observer. It is relevant to man-made outdoors and large indoors.
- *Ruggedness* refers to the deviation of the ground with respect to the horizon. It describes natural scenes, opposing flat to mountainous landscapes.
- *Verticalness*: It refers to the orientation of the "walls" of the space, whenever applicable. It opposes scenes organized horizontally (e.g. highways, ocean views), to scenes with vertical structures (buildings, trees).

The descriptors that refer to the scene content are:

- *Naturalness* refers to the origin of the components used to build the scene (man-made or natural). It is a general property that applies to any picture.
- *Busyness* is mostly relevant for man-made scenes. Busyness represents the sense of cluttering of the space. It opposes empty to crowded scenes.
- *Roughness* refers to the size of the main components (for man-made scenes, from big to small) or the grain of the dominant texture (for natural scenes, from coarse to fine) in the image.

We proposed to refer to the *spatial envelope* of a real world scene and a scene image, as a combination of space descriptors, in reference to the architectural notion of "envelope" of urban, landscape and indoor environments. In the next section, we defined the image-based representation that allows extracting spatial envelope descriptors from raw image, in order to generate the scene-centered description (cf. Fig. 4 and 5).

3 Scene-Space centered representation

Opposed to object-centered image representation, we describe here an image representation [10, 18] that encodes the distribution of textures and edges in the image and their coarse spatial arrangement without segmentation stage (see also [16, 20]). The resulting "sketchy" representation, illustrated in fig. 2, is not adequate for representing regions or objects within an image, but it captures enough of the image structure to reliably estimate structural and textural attributes of the scene (see sections 5-7). The sketch is based on a low-resolution encoding of the output magnitude of multiscale oriented Gabor filters:

$$A_M^2(\mathbf{x}, k) = \left\{ |i(\mathbf{x}) * g_k(\mathbf{x})|^2 \downarrow M \right\} \tag{1}$$

$i(\mathbf{x})$ is the input image and $g_k(\mathbf{x})$ is the impulse response of a Gabor filter. The index k indexes filters tuned to different scales and orientations. The notation $\downarrow M$ represents the operation of downsampling in the spatial domain until the resulting representation $A_M(\mathbf{x}, k)$ has a spatial resolution of M^2 pixels. Therefore, $A_M(\mathbf{x}, k)$ has a dimensionality M^2K where K is the total number of filters used in the image decomposition. Fig. 2 illustrates the information preserved by this representation. It shows synthesized texture-like images that are constrained to have the same features $A_M(\mathbf{x}, k)$ ($M = 4$, $K = 20$) than the original image. This scene-centered representation contains a coarse description of the structures of the image and their spatial arrangement.

Each scene picture is represented by a features vector $\mathbf{v}$ with the set of measurements $A_M(\mathbf{x}, k)$ rearranged into a column vector. Note that the dimensionality of the vector is independent of the scene complexity (e.g. number of regions). Applying a PCA further reduces the dimensionality of $\mathbf{v}$ while preserving most of the information that accounts for the variability among pictures. The principal components PCs are the eigenvectors of the covariance matrix

Fig. 2. Examples of sketch images obtained by coercing noise to have the same features than the original image. This scene-centered representation encodes the spatial arrangement of structures (here, at 2 cycles/image) without a previous step of segmentation. Each sketch has been reconstructed from 320 features (see [18]).

$\mathbf{C} = E[(\mathbf{v} - \mathbf{m})(\mathbf{v} - \mathbf{m})^T]$ where $\mathbf{v}$ is a column vector composed by the image features, and $\mathbf{m} = E[\mathbf{v}]$. We will refer to the column vector $\mathbf{v}$ as the L dimensional vector obtained by projection of the image features onto the first L PCs with the largest eigenvalues.

4 From image features to spatial envelope descriptors

The system described below is designed to learn the relationship between the sketch representation (image features $\mathbf{v}$) and the space descriptors. Each picture has two values associated to each space descriptor $\{R_j, \alpha_j\}$. The first parameter, R_j is the relevance of a specific space descriptor for a given picture, scaled between 0 and 1. To annotate the database (3,000 pictures), three observers selected, for each picture, the set of space descriptors that were appropriate. For example, a street image can be described in terms of its *mean depth*, its degree of *openness* and *expansion* and how cluttered it is. But for describing an object or an animal, *expansion* and *busyness* are not relevant descriptors. *Verticalness*, on the other hand, may apply specifically to some objects (e.g. a bottle), indoors (e.g. a stairs view) urban places (e.g. a skyscraper) and natural scenes (e.g. forest). *Naturalness* and *mean depth* are relevant descriptors of any image. The second parameter, α_j is the value of a specific space descriptor, normalized between 0 and 1. For instance, a city sky-line will have a large value of mean depth and openness; a perspective view of a street will have a high value of expansion and an intermediate value of depth and openness. To calibrate α_j, each picture was ranked among a set of pictures already organized from the lowest to the highest value of each descriptor (e.g., from the more open to the less open space). The position index for each picture corresponded to α_j.

The system learnt to predict the *relevance* and the *value* of each space descriptor, given a set of image features $\mathbf{v}$. Three parameters are estimated for each new picture:

– *Relevance*. Relevance of a space descriptor is the likelihood that an observer uses it for describing volumetric aspects of the scene. The Relevance may be approximated as: $P(R_j = 1|\mathbf{v})$.

- *Value*. Value of a space descriptor estimates which verbal label would best apply to a scene. It can be estimated from the image features as $\hat{\alpha_j} = E\,[\,\alpha_j \,|\, \mathbf{v}]$.
- *Confidence*. Confidence value gives how reliable is the estimation of each space descriptor provided the image features $\mathbf{v}$. It corresponds to $\sigma_j^2 = E[(\hat{\alpha_j} - \alpha_j)^2|\mathbf{v}]$. The higher the variance σ_j^2 the less reliable is the estimation of the property α_j given the image features $\mathbf{v}$.

The *Relevance* is calculated as the likelihood:

$$P(R_j = 1|\mathbf{v}) = \frac{p(\mathbf{v}|R_j = 1)p(R_j = 1)}{p(\mathbf{v}|R_j = 0)P(R_j = 0) + p(\mathbf{v}|R_j = 1)P(R_j = 1)} \tag{2}$$

The PDFs $p(\mathbf{v}|R_j = 1)$ and $p(\mathbf{v}|R_j = 0)$ are modeled as mixture of gaussians: $p(\mathbf{v}|R_j = 1) = \sum_{i=1}^{N_c} g(\mathbf{v}, c_i)\, p(c_i)$. The parameters of the mixtures are then estimated with the EM algorithm [5]. The prior $P(R_j = 1)$ is approximated by the frequency of use of the attribute j within the training set.

Estimation of the value of each descriptor can be performed as the conditional expectation $\hat{\alpha_j} = E\,[\,\alpha_j \,|\, \mathbf{v}] = \int \alpha_j f(\alpha_j \,|\, \mathbf{v}) d\alpha_j$. The function can be evaluated by estimating the join distribution between that values of the attribute and the image features $f(\alpha_j, \mathbf{v})$. This function is modeled by a mixture of gaussians: $f(\alpha_j, \mathbf{v}) = \sum_{i=1}^{N_c} g(\alpha \,|\, \mathbf{v}, c_i)\, g(\mathbf{v} \,|\, c_i)\, p(c_i)$ with $g(\alpha \,|\, \mathbf{v}, c_i)$ being a gaussian with mean $a_i + \mathbf{v}^T \boldsymbol{b}_i$ and variance σ_i. The learning of the model parameters for each property is estimated with the EM algorithm and the training database [5, 18]. Once the learning is completed, the conditional PDF of the attribute value α_j, given the image features, is:

$$f(\alpha_j \,|\, \mathbf{v}) = \frac{\sum_{i=1}^{N_c} g(\alpha_j \,|\, \mathbf{v}, c_i)\, g(\mathbf{v} \,|\, c_i)\, p(c_i)}{\sum_{i=1}^{N_c} g(\mathbf{v} \,|\, c_i)\, p(c_i)} \tag{3}$$

Therefore, given a new scene picture, the attribute value is estimated from the image features as a mixture of local linear regressions:

$$\hat{\alpha_j} = \frac{\sum_{i=1}^{N_c} (a_i + \mathbf{v}^T \boldsymbol{b}_i)\, g(\mathbf{v} \,|\, c_i)\, p(c_i)}{\sum_{i=1}^{N_c} g(\mathbf{v} \,|\, c_i)\, p(c_i)} \tag{4}$$

We can also estimate the scene attribute using the maximum likelihood: $\hat{\alpha_j} = \max_{\alpha_j}\{f(\alpha_j \,|\, \mathbf{v})\}$. The estimation of the PDF $f(\alpha_j|\mathbf{v}, S)$ provides a method to measure the confidence of the estimation provided by eq. (4) for each picture:

$$\sigma_j^2 = E[(\hat{\alpha_j} - \alpha_j)^2|\mathbf{v}] = \frac{\sum_{i=1}^{N_c} \sigma_i^2\, g(\mathbf{v} \,|\, c_i)\, p(c_i)}{\sum_{i=1}^{N_c} g(\mathbf{v} \,|\, c_i)\, p(c_i)} \tag{5}$$

The confidence measure allows rejecting estimations that are not expected to be reliable. The bigger the value of the variance σ_j^2 the less reliable is the estimation $\hat{\alpha_j}$.

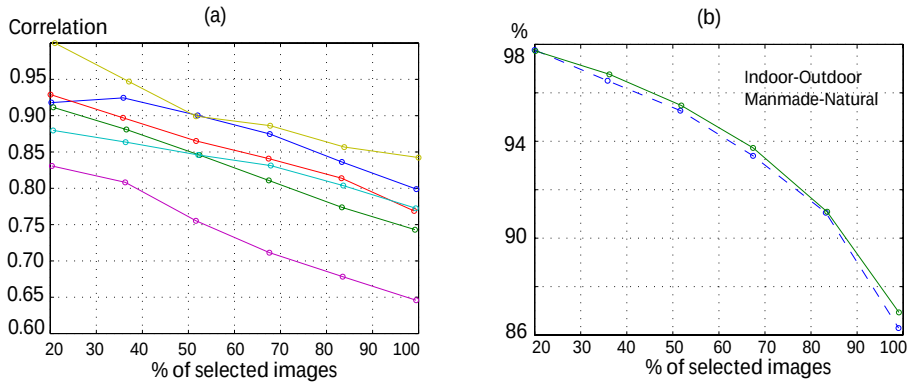

Fig. 3. a) correlation of picture ranking, for each space dimension, between observers and the model as a function of the percentage of images selected. The images are selected by putting a threshold to the confidence σ_j^2. From top to bottom: verticalness, openness, mean depth, ruggedness, expansion, busyness. b) Performance of correct classification in man-made vs. natural scenes and indoor vs. outdoor.

5 Performances of classification

Each space property is a one-dimensional axis along which pictures are continuously organized [10]. Fig. 3.a shows correlation values between the ranking made by human observers and the ranking computed by the model (from eq. 4). We asked two observers to perform 10 orderings, of 20 images each (images not used in the training), for each of the spatial envelope properties. Orderings were compared by measuring the Spearman rank correlation:

$$Sr = 1 - 6\frac{\sum_{i=1}^{n}(rx_i - ry_i)^2}{n(n^2-1)} \tag{6}$$

with $n = 20$. rx_i and ry_i are respectively the rank positions of the image i given by the algorithm and by one subject. A complete agreement corresponds to $Sr = 1$. When both orderings are independent, $Sr = 0$. When considering the entire image database, correlation values go from 0.65 to 0.84 for the different attributes of the spatial envelope. When considering a percentage of images with the highest level of confidence (σ_j^2, Eq. 5), performances improve.

6 Generating space-centered descriptions

The value α_j of a specific descriptor assigned to an image can be translated into a verbal description. Each space axis was split in a few subgroups (from 2 to 5). For example, the mean depth axis was separated in 4 groups: close-up view, small space, large space, and panoramic view. The openness descriptor was represented by 5 categories (open, semi-open, semi-closed, closed, and enclosed). All together, the verbal labels are expected to form a meaningful description

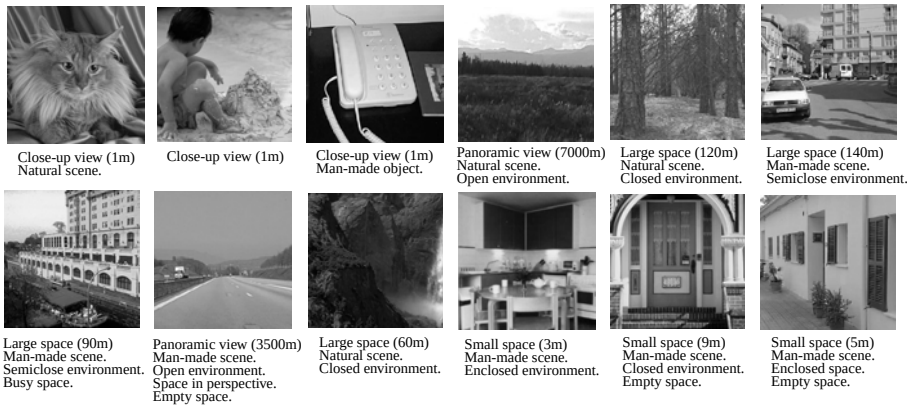

Fig. 4. Examples of space-centered descriptions automatically generated. For each image, the description contains only the space properties that are relevant and that provided high confidence estimates.

of the scene (see Fig. 4). Whenever a property was not relevant for a type of image ($R_j < threshold$) or the level of confidence (σ_j^2) was not high enough, the system did not use the corresponding verbal label. Therefore, instead of being wrong, the model provides a less precise description.

7 Computing space-centered similarities

In [10], we provided *space-features* based image similarities: pictures with similar spatial envelope values were closed together in a multi-dimensional space formed by the set of space descriptors. Within this space, neighbor images look alike.

The spatial envelope representation is able to generate descriptions at different levels of categorization (fig. 5): the super-ordinate level of the space (e.g. large scaled views), a more basic-level (e.g. open and expanded urban space), and a subordinate-level (e.g. open and expanded urban space, crowded) where pictures are more likely to look similar. To which extend a specific space property is relevant for a subordinate or super-ordinate level of description in regard to human observers, still need to be determined, but the general principle illustrated in fig. 5 shows the potentiality of the spatial envelope description for categorizing very different pictures at multiple levels of categorization, as human observers do.

8 Conclusion

The scene-centered representation based on spatial envelope descriptors show that the highest level of recognition, the identity of a scene, may be built from of a set of volumetric properties available in the scene image. It defines a general recognition framework within which complex image categorization may be

achieved free of segmentation stage, grouping mechanisms, 3D interpretation and object-centered analysis. The space-centered approach provides a meaningful description of scene images at multiple levels of description (from superordinate to subordinate levels) and independently of image complexity. The scene-centered scheme provides a novel approach to context modeling, and can be used to enhance object detection algorithms, by priming objects, their size and locations [17]).

Fig. 5. Scene categorization at different levels of description.

Acknowledgments

Many thanks to Whitman Richards, Bill Freeman, John Henderson, Fernanda Ferreira, Randy Birnkrant and two anonymous reviewers whose comments greatly helped improve the manuscript. Correspondence regarding this article may be sent to both authors.

References

1. Barnard, K., Forsyth, D.A.: Learning the semantics of words and pictures. Proceedings of the International Conference on Computer Vision, Vancouver, Canada (2001) 408–415
2. Barrow, H. G., Tannenbaum, J.M.: Recovering intrinsic scene characteristics from images. In: Hanson, A., Riseman, E. (eds.): Computer Vision Systems, New York, Academic press (1978) 3–26
3. Biederman, I.: Recognition-by-components: A theory of human image interpretation. Psychological Review. **94** (1987) 115–148
4. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using Expectation-Maximization and its Application to Image Querying. IEEE Transactions on Pattern Analysis and Machine Intelligence. **24** (2002) 1026–1038
5. Gershnfeld, N.: The nature of mathematical modeling. Cambridge university press (1999)
6. Heaps, C., Handel, S.: Similarity and features of natural textures. Journal of Experimental Psychology: Human Perception and Performance. **25** (1999) 299–320
7. Henderson, J.M., Hollingworth, A.: High level scene perception. Annual Review of Psychology. **50** (1999) 243–271.
8. Marr, D.: Vision. San Francisco, CA. WH Freeman (1982)
9. Oliva, A., Schyns, P. G.: Diagnostic color blobs mediate scene recognition. Cognitive Psychology. **41** (2000) 176–210.
10. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the Spatial Envelope. International Journal of Computer Vision. **42** (2001) 145–175
11. Potter, M. C.: Meaning in visual search. Science. **187** (1975) 965–966.
12. Rao, A.R., Lohse, G.L.: Identifying high level features of texture perception. Graphical Models and Image Processing. **55** (1993) 218–233
13. Rensink, R. A., O'Regan, J. K., Clark, J. J.: To see or not to see: the need for attention to perceive changes in scenes. Psychological Science. **8** (1997) 368–373
14. Rogowitz, B., Frese, T., Smith, J., Bouman, Kalin, E.: Perceptual image similarity experiments. Human Vision and Electronic Imaging, SPIE Vol 3299. (1998) 576–590
15. Schyns, P.G., Oliva, A.: From blobs to boundary edges: evidence for time- and spatial-scale dependent scene recognition. Psychological Science. **5** (1994) 195–200
16. Szummer, M., Picard, R. W.: Indoor-outdoor image classification. IEEE International Workshop on Content-based Access of Image and Video Databases, Bombay, India (1998)
17. Torralba, A.: Contextual Modulation of Target Saliency. In: Dietterich, T. G., Becker, S, Ghahramani, Z. (eds.): Advances in Neural Information Processing Systems, Vol. 14. MIT Press, Cambridge, MA (2002)
18. Torralba, A., Oliva, A.: Depth estimation from image structure. IEEE Transactions on Pattern Analysis and Machine Intelligence. **24** (2002) 1226–1238
19. Torralba, A., Sinha, P.: Statistical context priming for object detection: scale selection and focus of attention. Proceedings of the International Conference in Computer Vision, Vancouver, Canada (2001) 763–770.
20. Vailaya, A., Jain, A., Zhang, H. J.: On image classification: city images vs. landscapes. Pattern Recognition. **31** (1998) 1921–1935