J. Pharm. Pharmac., 1969, **21**, 224–231 *Received October* 28, 1968

A simple infrared spectrum retrieval system

A. S. CURRY, J. F. READ AND C. BROWN

Home Office Central Research Establishment, Aldermaston, Reading, Berkshire, England

The retrieval of an unknown spectrum from a medium-sized collection (up to 10,000) of infrared curves has been studied and a system based on a film feature card using microfilm storage developed. Tests of alternative methods of coding resulted in the choice of the six most intense absorption peaks in defined areas of the spectrum. Correct identification within 1–2 min is achieved in over 95% of searches. The system is compact, easy to operate, relatively inexpensive and easily perused.

Infrared spectroscopy is a valuable aid to the forensic scientist for the unequivocal identification of a wide variety of drugs and other "scene of crime" materials. This method of qualitative analysis requires comparison of the unknown spectrum with that of an authentic sample of the material and confirmation that the two spectra are identical. Frequently, the forensic scientist has to identify unknown material and thus it is necessary to select similar curves for visual comparison by searching through reference spectra of a large number of compounds, a laborious and time-consuming task. By utilizing modern information techniques this effort can be minimized (Curry, D. R., 1963; Thomas, 1968).

Rapid identification of unknown compounds by infrared spectroscopy requires the availability of a comprehensive collection of reference spectra and means for the retrieval of individual spectra from the collection by name and by the spectral characteristics of the curves themselves. Every curve must therefore be reduced to a series of parameters or codes and the retrieval system must be capable of rapidly searching these data to select the spectrum identical to that of the unknown under investigation. The chosen method of coding the spectra should give the maximum spread of information, reducing bunching of the curves about major spectral characteristics so that a minimum number of spectra are retrieved. The reference spectra should be in a form which is compact for storage, which can be readily inspected for comparison purposes and which is flexible enough to cope with the continuing stream of new additions to the collection.

The system found by us to be the most suitable for use in a routine forensic science laboratory is described. For our work on retrieval the "Harrogate" infrared collection of the North-East Forensic Science Laboratory (Haddon Lodge, 32 Rutland Drive, Harrogate, Yorkshire) which contains the curves of 1100 compounds of importance to the toxicologist, was used. These spectra had been recorded as Nujol mulls or liquid films using the Perkin-Elmer Infracord 137 instrument.

Infrared spectrum retrieval systems

For the rapid identification of unknowns it is essential to be able to retrieve individual spectra by predetermined spectral characteristics (e.g. by position of major absorption bands) as well as by chemical name. The manual methods for retrieving spectra from reference collections by band position data are (1) absorption band

indexing and (2) punched card systems, including edge punched, body punched and feature card systems. Recently, large libraries of spectroscopic data have been stored on magnetic tape and the collections searched by computer (Anderson & Covert, 1967; Cross, Shields & Stanier, 1966; Erley, 1968). The alternative MIRACODE system for large collections of spectra requires the reference curves with their corresponding coding details to be photographed onto 16 mm film. The resulting library is then automatically scanned by a high speed photoelectric retrieval system. The cost of large systems limits their applicability to large libraries of data with a correspondingly large number of requests for searches and therefore to the central laboratory of an organization, thus restricting immediate availability.

Absorption band indexing

Small laboratory collections of spectra are frequently arranged in order of the peak having the maximum absorbance. If the wavelengths considered in coding are in the "fingerprint region" of the spectrum, grouping of large numbers of curves about a single wavelength is reduced to a minimum (Curry, A. S., 1963). This classification can be further improved by including two major peaks in the coding instead of one. Such simple manual methods of selecting curves are only efficient for infrared collections containing at the most a few hundred curves.

An extension of the major band method of classification is the commercially available Sadtler Spec-finder indexing system (Heyden and Son Ltd., Spectrum House, Alderton Crescent, London, N.W.4, England) for the identification of unknown compounds. In this system the wavelengths of the strongest bands in thirteen 1 μm intervals of the spectrum are arranged in groups depending on the most intense band, all wavelengths being recorded to the nearest 0·1 μm. Identification is by comparing the code of the unknown with numerical lists in the index and retrieving similarly coded curves by spectrum number for final visual comparison. The disadvantage of peak position indexing systems is that the bulky indices are difficult to maintain in the case of a randomly expanding collection, as updating requires the indices to be completely retabulated. This is only practical for the smallest laboratory collection or for large commerical collections where the revised numerical lists can be automatically prepared by computer.

Edge-punched card systems

Edge-punched cards are the simplest of the card systems and have been used extensively (Thompson, 1955; Muir & Hardie, 1962; McArdle & Skew 1965; and others). In an edge-punched retrieval system, coding details which describe each spectrum are represented by a series of holes along the edges of the card. The spectral parameters are incorporated by converting the holes assigned to these characteristics into deep slots. All other necessary information is included in the body of the card.

To search the collection, the complete set of cards is aligned and sorting needles inserted through the holes corresponding to the coding details being searched. All cards which contain the characteristics required fall away from the main collection. This procedure is awkward and tedious with a large collection and machine sorting methods must then be used. The edge-punched card system, however, has one big advantage; each card can be produced and used independently of any other. The disadvantage is the rapid deterioration of the cards with use and punching additional sets is laborious and costly.

Body-punched card systems

Body-punched cards operate on the same principle as the edge-punched system, but coding details of the spectrum are stored by punching holes on the body of the card (Casey, Perry & others, 1958). This enables the card to accommodate more punched data but only at the expense of other scientific information. Frequently, therefore, this system can only refer the searcher to the original spectrum by compound name or serial number. Retrieval using body-punched cards necessitates the examination of all the holes simultaneously and for this electromechanical sorting techniques must be used.

Feature card systems

Feature cards (commonly known as term entry, optical coincidence or peek-a-boo cards) use the principle of assigning one search characteristic to each card and punching holes into the card, corresponding to the number of each spectrum to which the characteristic applies (Curry & Moore, 1963; Schlichter & Wallace, 1963; Kaiser, 1965). The collection of data is searched by selecting the cards having the coding details required and aligning them before a light source and reading off the numbers of the coincident holes through which light appears (Fig. 1). These numbers correspond to spectra possessing all the features selected. Like body-punched cards, feature card retrieval systems require the actual spectrum to be retrieved from a subsidiary storage system.

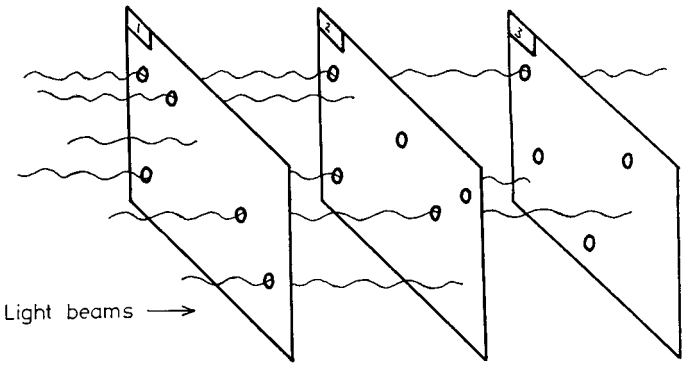

FIG. 1. Principle of operation of the feature card retrieval system.

Comparison of retrieval systems

Small laboratory collections of spectra (<2000 items) to which reference will be made infrequently are most conveniently served by manually operated edge-punched card systems, but the search time increases in proportion to the number of items. Where large numbers of cards are involved a mechanical means of sorting is needed. The cost of electro-mechanical sorters, however, can seldom be justified in small laboratories and for this reason both body-punched and edge-punched cards are not ideally suited to a collection of many thousands of cards. If feature cards are used then simple manual methods of retrieval are sufficient for systems containing up to 10,000 items. Feature card systems are compact, easy to operate, readily updated, easily perused and are undoubtedly superior for medium sized collections of spectroscopic data in small laboratories. For the very large collections of data fully automatic information processing techniques which require the use of computers become economically attractive.

Coding of infrared spectra

The basic data for any infrared retrieval system are derived from the accurately measured wavelengths of the major absorption bands which are characteristic of each spectrum. Frequently, negative search terms such as the absence of absorption bands in specific wavelength regions are included to eliminate unwanted data. The system devised in our laboratory makes use of the major peaks in the 5·0–15·0 μm spectrum range and these are recorded to the nearest 0·1 μm. Those bands appearing in the 6·7–7·6 μm range that are strong enough to be included in the coding are omitted to avoid complications in the examination of the spectra obtained from Nujol mulls. To ensure simplicity in coding procedures we have defined the strongest bands as those bands whose peaks are nearest to 0% transmittance, irrespective of the shape of the background. Shoulders are only counted as bands if they are completely resolved and the point of maximum absorption easily determined. In cases where two or more bands of equal intensity tie for the final coding detail, all these peaks are included in the coding of the spectrum. If the major peak is a hump rather than a sharp band and the point of maximum absorbance not easily determined, or the band too intense as a result of the sample being too thick, the wavelength of the peak mid-point is taken as the coding detail.

In our retrieval system, each 0·1 μm of the wavelength region considered in coding represents a possible coding detail or characteristic feature of the spectrum. This gives a total of 92 features to describe each spectrum. To allow for small variations in band positions resulting from differences in instrumentation and analytical conditions and techniques, it is necessary to include a tolerance of $\pm$ 0·1 μm during searching. In practice this is achieved by coding features 0·1 μm to each side of the coded spectral peak.

Our initial studies employing hand punched Carter-Parrett feature cards (J. L. Jolly and Partners Ltd., Orchard Road, Sutton, Surrey, England) indicated that a minimum of five or six spectral characteristics would have to be coded if the number of false (but equally probable) retrievals per search is to be acceptable for a collection containing several thousand curves. These five or six absorption bands may be selected in two ways. The selection of the six most intense bands in the spectrum is straightforward (six peak method of coding). For some spectra this may restrict the coding to the intense bands originating from the functional groups at the expense of relatively less intense bands in the fingerprint regions. These "fingerprint bands" are important because they exhibit differences between closely related compounds. Alternatively, the most intense bands in five specific intervals of the spectrum can be coded. This method called the five range method of coding ensures that details of the fingerprint spectrum range are included in the coded description of the compound. In our studies the following five ranges were used for this coding method (1) 5·0–6·6, (2) 7·5–8·9, (3) 9·0–9·9, (4) 10·0–10·9 and (5) 11·0–15·0 μm.

To determine the most satisfactory method of coding, the complete "Harrogate" collection of spectra was coded using both the five range and six peak methods. To eliminate the laborious task of punching cards on this scale the coding data were stored on magnetic tape and the spectroscopic library searched by computer (I.C.T. model 1301) in the manner of a retrieval system. The efficiencies of the two coding methods were determined by coding fifty test spectra and the spectroscopic library searched to recover these curves using the procedure which would be employed for the identification of unknown compounds. All the test spectra were coded from written instructions

by an operator not directly involved with infrared work. The first twenty-five of these spectra were copies of reference spectra in the "Harrogate" collection. The remaining test spectra were of compounds for which reference curves were available in the reference library but which had been recorded under different conditions and on different spectrometers. The results of these tests are indicated in Table 1.

Table 1. *Retrieval efficiencies obtained for the 5 range and 6 peak coding methods*

Source of test spectra	Method of coding	No. of features used	No. of spectra correctly retrieved	Average No. of spectra retrieved per search
25 "Harrogate" ..	5 range	4	23	Very large
		5	21	6
	6 peak	5	25	20
		6	25	7
25 "Unknowns" ..	5 range	4	21	Very large
		5	21	8
	6 peak	5	25	15
		6	25	4

The results indicate that of the two methods, the six peak is superior and it is encouraging that this method gave no failure to retrieve the spectrum of interest. The results also emphasize that coding a minimum of six peaks is essential for efficient retrieval from a medium sized collection when this method of coding is used. Detailed examination of the spectra and coding details indicated that the failure of the five range method to select the correct curve was in many cases due to the inclusion (or exclusion) of unreliable peaks of low absorption in the final code. These doubtful bands would not qualify for inclusion with the six peak method. It is of interest that the overall distribution of coding details obtained for the complete "Harrogate" collection by the two methods are not as markedly different as might have been anticipated (Fig. 2A, B).

Spectrum storage systems

All manual methods of spectrum retrieval, excluding edge-punched card systems, yield a serial number which directs the searcher to the location of the standard curve in an independent storage system. For small laboratory collections, the reference spectra are normally stored either numerically or alphabetically in suitable loose leaf folders. This arrangement requires ample storage facilities for the larger collections and suffers from the disadvantage that day to day removal and subsequent replacement frequently results in damage and loss of individual curves. It is, therefore, undesirable that collections should be stored in this manner.

A more suitable and compact system of storing spectroscopic records is microfilm filing. With such a system printed copies of individual curves can be obtained immediately for detailed study using a reader-printer facility while the original microfilmed collection is retained intact. A further advantage is the ready availability of inexpensive microfilm copies of the complete reference collection for distribution to other laboratories.

Infrared retrieval and storage system in use at the Central Research Establishment

This laboratory will soon have over 2,500 spectra of materials which are pertinent to forensic science comprising not only a large number of pharmaceuticals but also substances likely to be met in a routine forensic science laboratory such as butter,

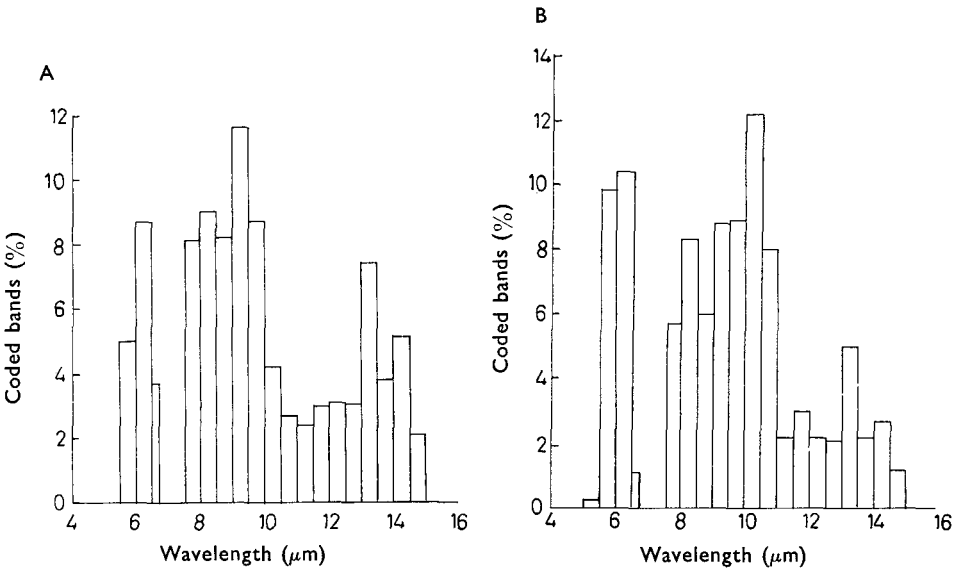

FIG. 2. Distribution of coded bands throughout the infrared spectrum, obtained (A) with the six peaks coding method and (B) with the five range method.

soap and Vaseline. This collection is steadily expanding as new items are encountered and these new additions have to be included in the system. The physical size of this spectroscopic library prohibits its publication for circulation to other workers and these curves are gradually being photographed on microfilm and copies of this film circulated to other forensic science laboratories in the British Isles. At the present time the complete "Harrogate" collection has been microfilmed. This photographic library of curves contains an index which lists the spectra in alphabetical order with an accession number. This accession number is also the same as that obtained from the infrared retrieval system described below.

The most suitable method of retrieving individual curves from a collection of this size is to use feature cards. The coincident card system manufactured by Materials Data Ltd. (19/27 South Street, Farnham, Surrey, England) has several features which make it more attractive than the alternative systems which are available. The MDL feature card can accommodate 3,600 items and the system is normally searched using six cards simultaneously. A larger card able to accommodate twice as many items is being developed. The chief advantage of the MDL system is that the cards are accurately lithographically reproduced on translucent polyester material from punched masters. This not only produces an extremely robust and hard wearing card but also eliminates the costly task of repeatedly punching separate sets of cards. Periodic updating is easily achieved by punching additional holes in the reserved masters with subsequent reproduction of replacement sets of cards.

The MDL system containing all the six peak coding details (see coding of infrared spectra) of the complete "Harrogate" collection of eleven hundred spectra has been in everyday use in our laboratory for several months. Our experience has shown that this collection can be sorted for six spectral characteristics and the spectrum of an unknown identified by matching with the corresponding reference curve in 3 to 5 min. The retrieval efficiency of this system is indicated in Figs 3 and 4. First searches have resulted in the successful identification of 96% of the unknowns which have reference

curves in the collection. The average number of spectra retrieved per search is 2·4. However, the number of false, but also equally probable, retrievals will increase with the population of the collection and it may be necessary to include additional coding features in the system at a later stage. In the few cases where the first search fails to select the correct reference curve, secondary searches are made using combinations of the coded bands. In this context the MDL system is advantageous in that "near-misses" are readily detected by the relative intensities of light appearing through the "holes" or light paths on the card. It is our experience that the correct reference curve

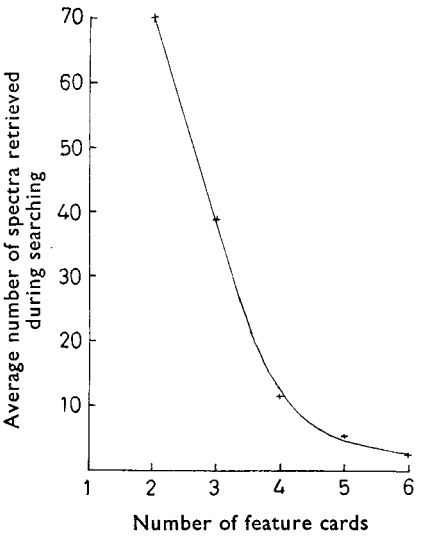

FIG. 3. Fall off in the number of retrievals with increasing number of feature cards for a collection containing eleven hundred items.

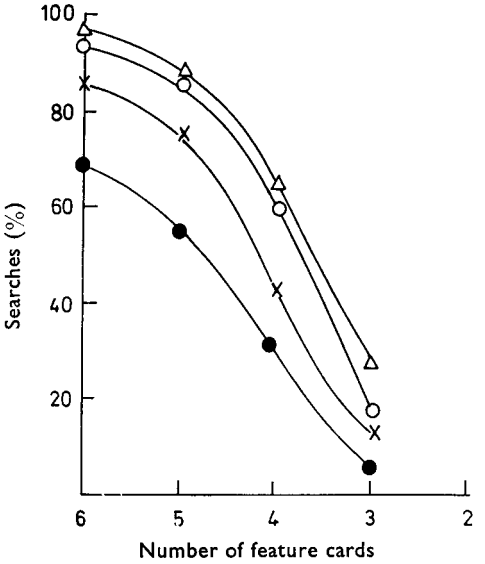

FIG. 4. Percentage of searches yielding x or less retrievals at different stages of searching a collection of eleven hundred items. △ x = 10. ○ x = 8. X x = 4. ● x = 2.

is always selected within two searches. The failure to obtain complete coincidence of coding between the unknown and reference spectra can be attributed to the following factors:

1. Impurities in the unknown or reference samples may introduce additional bands which are strong enough to be included in the coding.

2. Variations in the levels of background absorptions may effect the relative intensities of the absorption bands in the two spectra.

3. Polymorphism in organic compounds may result in the inversion of peak intensities (Mesley & Johnson, 1965; Mesley & Houghton, 1967; Mesley & Clements, 1968 and Mesley, Clements & others, 1968).

4. Grating instruments with high resolution may resolve peaks which appear as shoulders with instruments incorporating only prism optical systems.

It will be obvious that these factors will cause some difficulties with all peak position infrared retrieval systems.

The system described has aroused considerable interest in other laboratories concerned with the problem of drug identification and it is intended to make this system commercially available. We are at the present time considering other possible forensic applications of this type of feature card retrieval system. One such application is an aid to the rapid identification of solid dosage forms for use in hospitals as well as forensic science laboratories.

Acknowledgements

The authors wish to thank the Treasury and the O and M Branch of the Home Office for their interest and encouragement without which this work would have not been possible. They further thank Mr. Quinney of the Home Office and Metropolitan Police Joint Automatic Data Processing Unit who carried out the computer analyses and all the staff of the Central Research Establishment who participated in the coding of the spectra.

REFERENCES

ANDERSON, D. H. & COVERT, G. L. (1967). *Analyt. Chem.*, **39**, 1288–1294.

CASEY, R. S., PERRY, J. W., BERRY, M. M. & KENT, A. (1958). *Punched Cards.* New York: Reinhold.

CROSS, L. H., SHIELDS, D. J. & STANIER, H. M. (1966). Imperial Chemical Industries Report No. MD 4804/A.

CURRY, A. S. (1963). "The Coding of Analytical Data", 3rd International Meeting in Forensic Immunology, Medicine, Pathology and Toxicology, London.

CURRY, D. R. (1963). *Analyst*, **88**, 829–834.

CURRY, D. R. & MOORE, P. J. (1963). *Overseas Geology and Mineral Resources*, **9**, 61.

ERLEY, D. S. (1968). *Analyt. Chem.*, **40**, 894–898.

KAISER, H. (1965). *Hilger Journal, Nov*, 64–73.

MCARDLE, C. & SKEW, E. A. (1965). *Symposium on Identification of Drugs and Poisons.* London: Pharmaceutical Press.

MESLEY, R. J. & CLEMENTS, R. L. (1968). *J. Pharm. Pharmac.*, **20**, 341–347.

MESLEY, R. J., CLEMENTS, R. L., FLAHERTY, B. & GOODHEAD, K. (1968). *Ibid.*, **20**, 329–340.

MESLEY, R. J. & HOUGHTON, E. E. (1967). *Ibid.*, **19**, 295–304.

MESLEY, R. J. & JOHNSON, C. A. (1965). *Ibid.*, **17**, 329–340.

MUIR, J. W. & HARDIE, G. G. M. (1962). *J. Soil Sci.*, **13**, 249–254.

SCHLICHTER, N. E. & WALLACE, E. (1963). *Appl.*, **17**, 98–101.

THOMAS, I. C. (1968). *A New Chemical Structure Code for Data Storage and Retrieval in Molecular Spectroscopy.* London: Heyden and Sons.

THOMPSON, H. W. (1955). *J. chem. Soc.*, 4501–4509.