Современные проблемы дистанционного зондирования Земли из космоса. 2013. Т. 10. № 3. С. 75–84

Алгоритм двухэтапной классификации гиперспектральных данных в пространстве коэффициентов спектральной яркости по результатам авиационной съемки

В.Н. Остриков, С.И. Смирнов, В.В. Михайлов

Санкт-Петербургский филиал ОАО КБ «Луч»
E-mail: luchmail@spb.vega.su

Рассматривается проблема классификации данных гиперспектральной съемки в пространстве коэффициентов спектральной яркости. Предполагается, что каждый объект на снимке принадлежит какому-либо спектральному классу, содержащемуся в базе данных, сформированной посредством наземных измерений спектральных характеристик объектов. Входными данными алгоритма являются гиперспектральные снимки, прошедшие предварительную обработку (калибровка, фильтрация регулярного и случайного шумов, геометрическая коррекция). Классификация осуществляется в два этапа. На первом производится отсев классов, к которым заведомо не может принадлежать объект, с использованием «грубой» метрики. Процедура позволяет существенно снизить конечный объем обработки. На втором этапе из множества «оставшихся» классов производится селекция близких к объекту элементов базы с учетом чувствительной к спектральным различиям метрики (метрики Теребижа, декартовой). Алгоритм опробован на данных гиперспектральной съемки, полученных с авиационного носителя при различных условиях наблюдения. Выявлена робастность алгоритма в широком диапазоне отношений сигнала к шуму. Проведено сравнение результатов классификации.

Ключевые слова: классификация гиперспектральных данных, метрика Теребижа, кластеризация**.**

Введение

В настоящее время возрастает интерес к дистанционному зондированию с использованием аппаратуры гиперспектральной съемки. Такого рода данные находят применение в различных отраслях народного хозяйства, таких как детектирование очагов лесных пожаров, областей загрязнения окружающей среды, зараженных вредителями сельскохозяйственных угодий и многих других.

Одной из наиболее важных задач тематической обработки информации является классификация – сопоставление данных с авиационного носителя с базой данных (БД) измерений в автоматическом режиме. В большинстве отечественных работ по данной тематике классификация рассматривается в пространстве спектральной плотности энергетической яркости (СПЭЯ), тогда как в рассматриваемой работе используется пространство коэффициентов спектральной яркости (КСЯ). Выбор обусловлен тем, что вектор КСЯ является более гладкой и менее вариабельной функцией длины волны, нежели СПЭЯ, что позволяет корректно выполнять сравнения наземных и дистанционных измерений в условиях должного качества калибровки.

Как показывает практика, сравнение спектральных характеристик авиационного снимка с характеристиками БД в стандартной евклидовой метрике зачастую может давать «неадекватный» результат ввиду сильной зашумленности и вариабельности. В связи с этим возникает вопрос о необходимости предварительной обработки снимков (устранение случайного и полосового шума, нормировка и т.д.) и выборе метрики, соответствующей требуемому качеству классификации.

В работе делается предположение, что на обрабатываемом снимке может присутствовать любая характеристика из БД. Поскольку в практических задачах БД может иметь колоссальные размеры, простой перебор является дорогостоящей процедурой. Наличие такой проблемы приводит к необходимости применения двухэтапного алгоритма классификации, состоящего из предварительного отсева тех классов, к которым искомый объект точно не относится, на основе «грубой» метрики с дальнейшей селекцией близких к объекту элементов с использованием «качественной», но более ресурсозатратной метрики.

Постановка задачи и описание алгоритма

Исходными данными являются результаты гиперспектральной съемки (ГСС) с авиационного носителя (в дальнейшем – гиперкуб), результаты наземных измерений, собранные в БД (библиотеку типов) спектральных характеристик. На *рис. 1* отображены особенности представления данных ГСС.

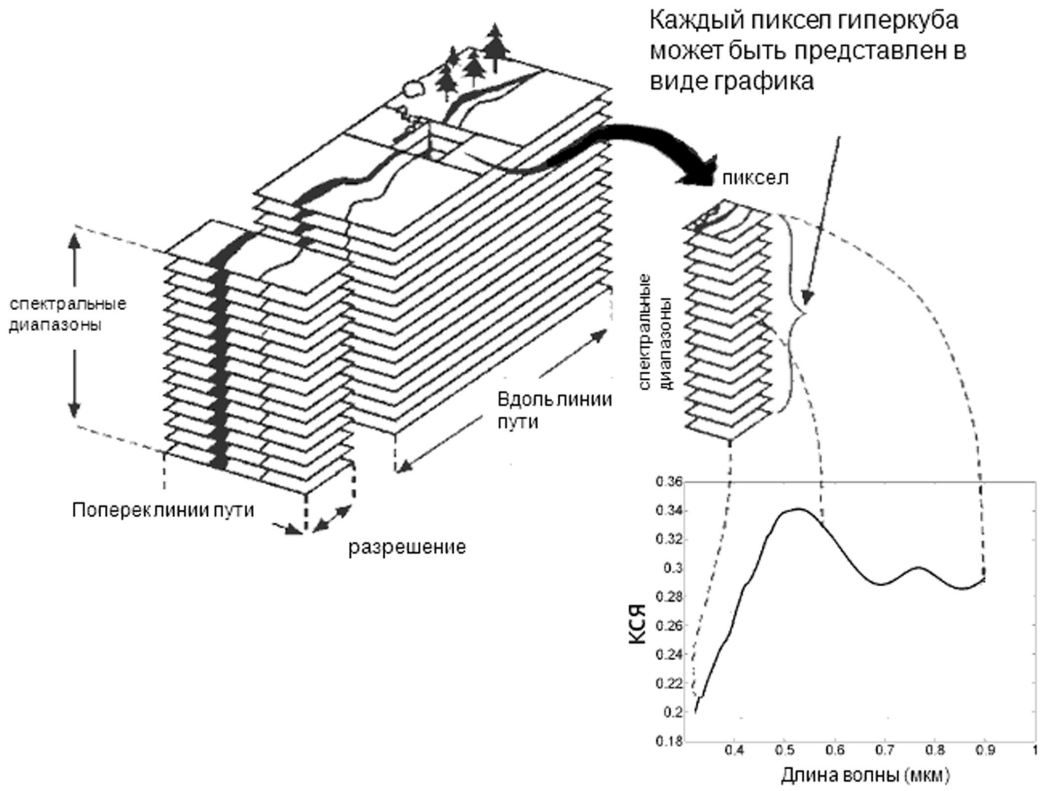

Рис. 1. Особенности представления данных гиперспектральной съемки

Перед применением алгоритмов тематической обработки следует провести предварительную обработк – ряд процедур, улучшающих качество исходных данных (Остриков В.Н. и др., 2012). В частности, качество тематической обработки ГСС существенно зависит от таких факторов, как:

- качество калибровки;
- качество выполнения геометрической коррекции;

– качество радиометрической коррекции;
– качество фильтрации случайного шума.

Следует отметить, что некачественная калибровка делает практически невозможным обнаружение объектов на основе анализа ГСС по той причине, что при этом не удается корректно сопоставить дистанционные данные по эталонам с их измерениями, полученными наземной съемкой. Кроме того, происходит существенное перепутывание спектральных типов.

Общая структура разработанного алгоритма представлена на *рис. 2.*

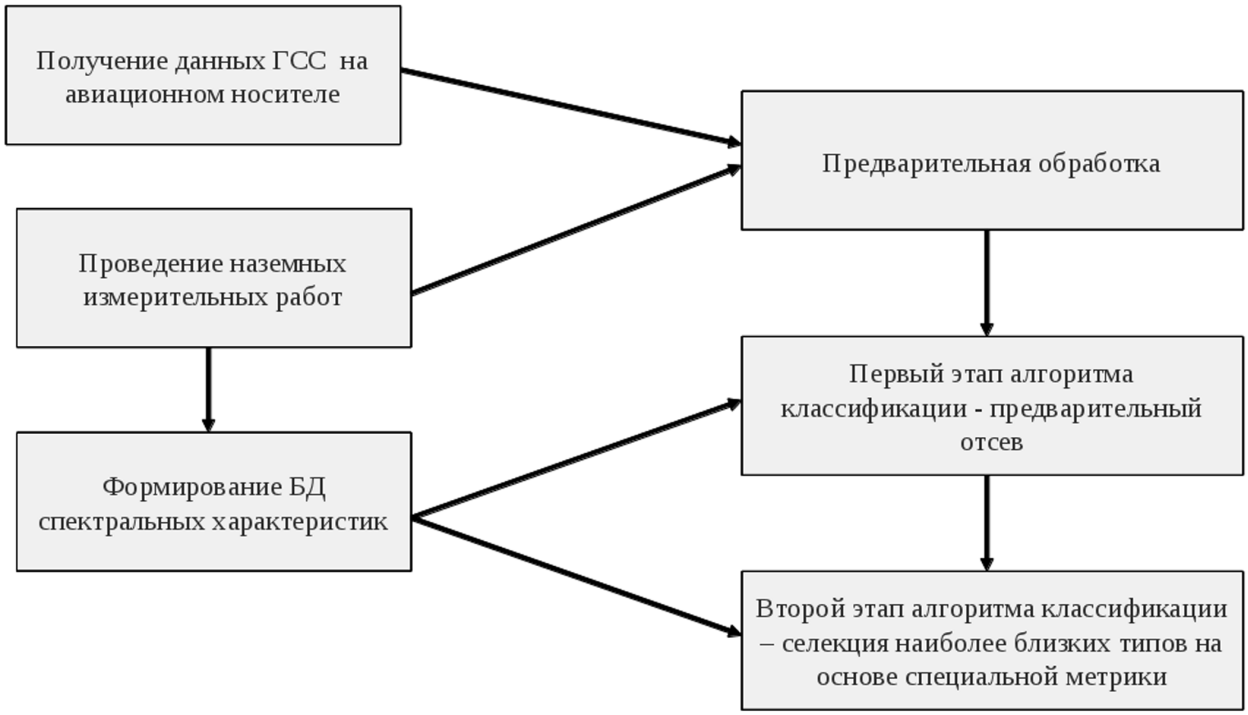

Рис. 2. Схема получения и обработки гиперспектральных данных

Поскольку изменчивость измеряемых значений коэффициентов спектральной яркости в разных условиях наблюдения является существенной, сопоставление с ранее полученными результатами (выбранными из базы данных) проблематично с точки зрения получения удовлетворительной точности. Подобного типа эффекты, а также сложность учета таких условий (наличие случайным образом распределенной облачности разного типа, состава атмосферной трассы и т.д.) приводят к достаточно высокой вариабельности спектральных характеристик. Отсюда следует, что эталонные данные о КСЯ должны удовлетворять требованию статистической обусловленности на значительной выборке измерений. Кроме того, для получения устойчивых к условиям наблюдениям результатов их целесообразно приводить к некоторому стандарту.

В этом качестве применяется следующее условие:

$$\int f^2(\lambda) d\lambda = 1, \qquad (1)$$

где $f(\lambda)$ – приведенные отсчеты КСЯ.

Поскольку регистрируемый участок спектра определен на дискретной совокупности линий $\lambda_1,\ldots,\lambda_L$, ($L$ – общее число всех спектральных линий, на которых проводятся измерения ГСС) заданных, в рассматриваемом случае, в единицах длины волны (нм), выражение (1) можно преобразовать:

$$\sum_{i=1}^{L} f^2(\lambda_i)\Delta_i = 1\ , \tag{2}$$

где Δ_i – ширина спектральной линии λ_i, интегрирование ведется только в пределах чувствительности приемников.

Предположим, что в библиотеке имеется N типов поверхностей с измеренными средними $\rho_i(\lambda_k)$, $i = 1\ldots N$, $k = 1\ldots L$ и среди них нет повторяющихся (спектральные характеристики БД также проходят предварительную обработку для снижения случайного шума). Ставится задача построить алгоритм, который в автоматическом режиме производит отождествление пикселей гиперкуба с одним типом (или с группой типов) из библиотеки. Стандартный подход можно сформулировать следующим образом: вводится некая метрика и на основе значений этой метрики для совокупности типов выбираются наиболее близкие. Недостатком такого подхода является необходимость полного перебора библиотеки спектральных типов.

Более рациональным методом решения является предварительная кластеризация библиотеки спектральных типов с последующим построением соответствия: пиксел-кластер БД. Такое соответствие сокращает вычислительную работу за счет того, что при проверке принадлежности пикселя к одному из типов отвергается гипотеза о его принадлежности ко всем типам, не входящим в тот кластер, с которым он был соотнесен.

В качестве отображения для перехода в пространство кластеризации БД в работе использовался вариант метрики отношений. Под метрикой отношения понимается отображение из пространства КСЯ во множество неотрицательных целых чисел в соответствии с выражением

$$L\big(\rho(\lambda)\big) = \sum_{i=1}^{d-1} \delta_i\big(\rho(\lambda)\big),$$

$$\delta_i\big(\rho(\lambda)\big) = \begin{cases} 1, & \textit{если} \quad \rho(\lambda_{i+1}) > \rho(\lambda_i) \\ 0, & \textit{если} \quad \rho(\lambda_{i+1}) \le \rho(\lambda_i). \end{cases} \tag{3}$$

Такое отображение строится для всех типов из БД (по спектральным кривым, прошедшим соответствующую фильтрацию случайного шума и нормировку) и для каждого пикселя гиперкуба, прошедшего предварительную обработку (калибровка, радиометрическая коррекция, фильтрация случайного шума, нормировка спектральных отчетов). В результате для каждого типа из БД и для каждого пикселя гиперкуба рассчитывается собственное значение метрики отношений.

Таким образом, в используемом алгоритме под кластеризацией понимается разбиение множества типов из БД на непересекающиеся подмножества, внутри которых типы будут близки по предложенной метрике; при этом должно выполняться условие существенного отличия между непересекающимися подмножествами – кластерами. В случае

кластеризации типов БД входными данными являются значения вспомогательной функции (в данном случае метрики отношений). На выходе получается дизъюнктное разбиение множества типов на кластеры; причем разбиение осуществляется с помощью алгоритма k-средних (k-means). Выбор данного алгоритма обусловлен тем, что данные для кластеризации представляют собой евклидово пространство и не требуется использовать более сложные алгоритмы кластеризации, такие как, например, метод FOREL (метод формального элемента) (Загоруйко, 1999).

Благодаря вычислительной простоте, метрика (3) значительно ускоряет обработку данных; кроме того, после построения соотношения пиксель-кластер БД сопоставлять пиксель гиперкуба необходимо лишь с очень небольшой частью типов (только с типами, входящими в данный кластер). Следует уточнить, что при заданном количестве кластеров и их размере можно оценить выигрыш в быстродействии алгоритма при использовании кластеризации. В частности, при допущении, что количество пикселей в гиперкубе на порядок превосходит количество типов БД, можно пренебречь затратами на кластеризацию БД и расчет метрики (3). Такое предположение дает оценку для выигрыша по произвольному пикселю по сравнению с алгоритмом без кластеризации в пределах от $N/N_{\max}$ до $N/N_{\min}$ раз, где $N_{\max}$ и $N_{\min}$ – количество элементов в самом большом и самом маленьком кластерах БД. Так как априори неизвестно распределение спектральных типов по пространству, то можно дать только грубую оценку: выигрыш от использования кластеризации будет в пределах от $N/N_{\max}$ до $N/N_{\min}$ раз.

Необходимо отметить, что близость спектральных типов и значительная вариабельность данных в рамках одного типа делают невозможным получение аналитической оценки для «оптимального» количество кластеров с точки зрения баланса качества классификации и уменьшения вычислительной сложности алгоритма. Однако приведенный выше принцип – брать наибольшее возможное количество кластеров, обеспечивающее «хорошую» различимость типов в терминах какой-либо релевантной спектральным характеристикам метрики, вполне применим на практике. В данной статье предлагается использовать метрику (3) и количество кластеров подбирать из эмпирических соображений в зависимости от отделимости кластеров БД и требуемой для решения конкретной задачи вероятности ошибочной классификации (например, если требуется очень быстрая грубая классификация, то можно увеличить количество кластеров).

В качестве критерия близости исследуемого пикселя к *i*-му спектральному типу соответствующего кластера использовалось видоизменение дискриминанта Теребижа (Теребиж, 2005), имеющее вид

$$\tau^i = \sum_{k=1}^{L} \frac{\left(\rho(\lambda_k) - \rho^i(\lambda_k)\right)^2}{\rho^i(\lambda_k)}, \qquad (4)$$

где τ^i – значения видоизмененного дискриминанта Теребижа для *i*-го опорного вектора из соответствующего кластера БД; $\rho(\lambda_k)$ – значения искомого спектрального вектора; $\rho^i(\lambda_k)$ – *i*-й опорный вектор из соответствующего кластера БД.

Для проверки качества функционирования алгоритма использованы две специальные тестовые поверхности, которые по своим спектральным характеристикам близки к травяному покрову и песку. Их положение на данных ГСС было известно, что позволило сравнить результат автоматической классификации с таким эталоном. Этот подход позволяет получить объективную численную оценку результата классификации.

Алгоритм тестировался на двух наборах данных ГСС:

– первый набор был получен при условиях отсутствия облачности;

– второй набор был получен в условиях переменной облачности.

Для тестов, помимо метрики Теребижа, рассматривалась ещё и евклидова метрика. На *рис. 3* приведены характеристики используемых тестовых объектов и их спектральных прототипов.

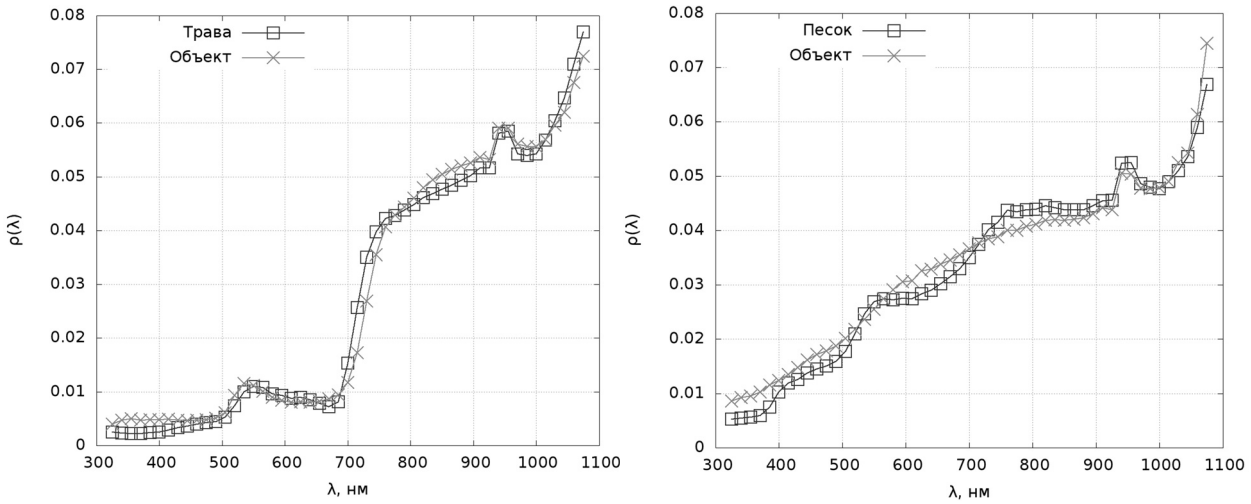

Рис. 3. Нормированные спектральные характеристики пар трава – тестовая поверхность 1 (слева) и песок – тестовая поверхность 2 (справа)

Процесс обработки данных, получения результатов и их визуализации осуществляется в общем аппаратно-программном комплексе. На *рис. 4* и *5* наглядно приведены результаты работы алгоритма классификации, а в *табл. 1* и *2* – численные результаты. Анализ показывает, что введенная в работе метрика имеет схожие результаты распознавания с евклидовой метрикой на первом наборе данных (вероятность распознавания чуть ниже при меньшем количестве ложных тревог), тогда как на втором наборе данных она значительно превосходит результаты евклидовой метрики, что связано с ее структурой, лучше учитывающей форму спектральной кривой.

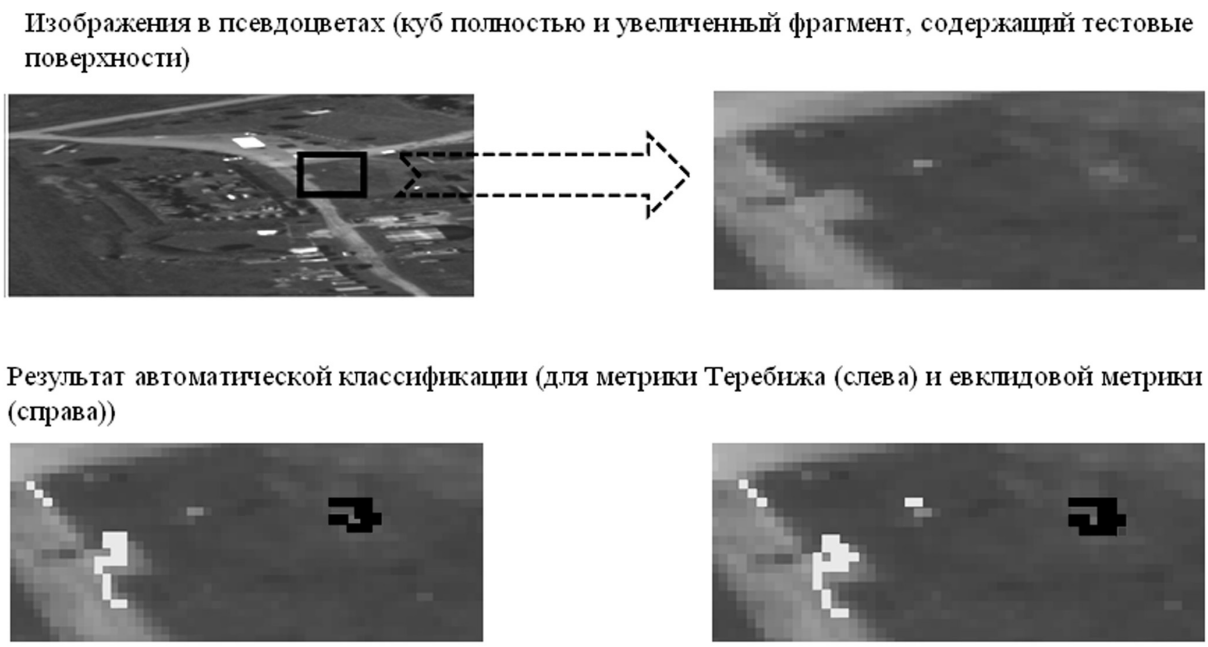

Рис. 4. Результаты автоматической классификации на реальных данных из первого набора при классификации с тремя кластерами

Таблица 1. Результаты автоматической классификации на реальных данных из первого набора

Тип поверхности	***Вероятность распознавания***	***Вероятность ложных тревог***
Тестовая поверхность 1 ■	Теребиж – 0,74 Евклид – 0,8	Теребиж – 0,0002 Евклид – 0,0008
Тестовая поверхность 2 ■	Теребиж – 0,73 Евклид – 0,79	Теребиж – 0,00004 Евклид – 0,005

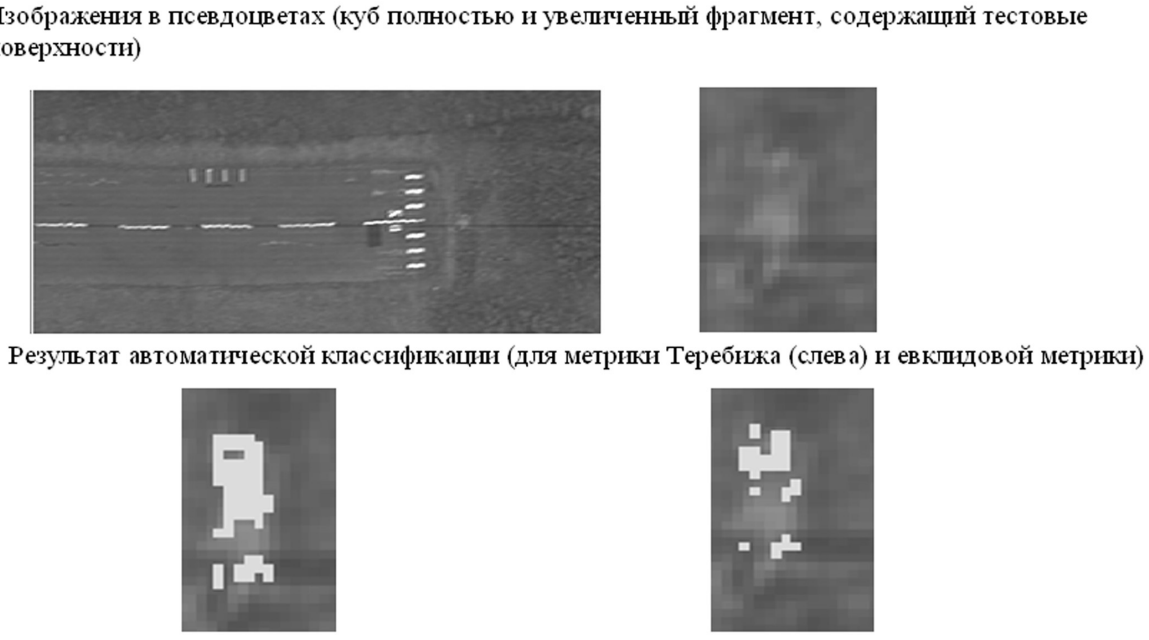

Рис. 5. Результаты автоматической классификации на реальных данных из второго набора при классификации с тремя кластерами

Таблица 2. Результаты автоматической классификации на реальных данных из второго набора

Тип поверхности	*Вероятность распознавания*	*Вероятность ложных тревог*
Тестовая поверхность 1	Теребиж – 0,81 Евклид – 0,34	Теребиж – 0,0008 Евклид – 0,002

Кроме того, была проведена классификация гиперспектральных данных из первого тестового набора, на которых приведены поля с сельскохозяйственными культурами, с кластеризацией (три кластера) и без (один кластер). Так как расположение полей и засеянные на них культуры были заранее известны, результат классификации интересен как с точки зрения верификации алгоритма, так и с точки зрения оценки выигрыша по скорости при его использовании. Результаты классификации приведены на *рис. 6*, время выполнения программы – в *табл. 3*. Следует отметить, что при достаточно схожих спектральных типах классификация была произведена в целом удовлетворительно (правильно определены поля с овсом и картофелем, выделены участки занятые травяным покровом, асфальтовая дорога), при этом время работы программы сократилось более чем в 2 раза за счет применения предварительной кластеризации ($N/N_{max} \approx 1{,}8$ и $N/N_{min} \approx 10{,}5$).

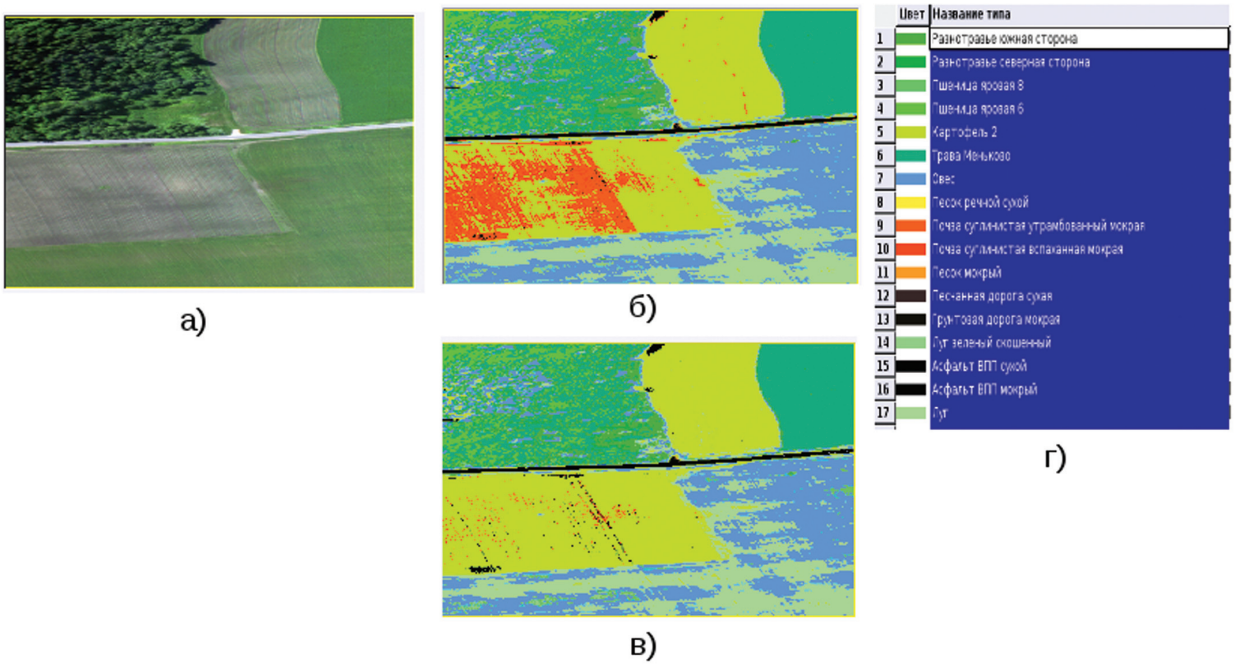

Рис. 6. Изображения в псевдоцветах (а) и результат классификации для одного (б) и трех кластеров (в), легенда (г)

Таблица 3. Время работы алгоритма классификации для одного и трех кластеров

Количество кластеров	*1 кластер*	*3 кластера*
Время выполнения программы	18 с	8 с

Заключение

В работе предложен двухэтапный алгоритм классификации гиперспектральных снимков с авиационного носителя. Алгоритм протестирован на наборе данных, полученных при различных условиях наблюдения (при постоянной и переменной облачности). Выявлена робастность алгоритма к отношению сигнал / шум. На модельных примерах показана эффективность процедуры кластеризации базы данных.

Следует отметить, что предложенный алгоритм использует только спектральные характеристики и никак не учитывает геометрические свойства искомых объектов, что делает перспективным его совместное применение с алгоритмами детектирования объектов.

Литература

1. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний // Новосибирск: ИМ СО РАН, 1999.
2. *Остриков В.Н., Плахотников О.В., Кириенко А.В.* Комплекс обработки гиперспектральных данных, получаемых с авиационных и космических носителей // Материалы десятой всероссийской открытой конференции «Современные проблемы дистанционного зондирования Земли из космоса». 2012.
3. *Теребиж В.Ю.* Введение в статистическую теорию обратных задач // М.: Физматлит, 2005.
4. *Шовенгердт Р.А.* Дистанционное зондирование. Модели и методы обработки изображений / Пер. с англ. М.: Техносфера, 2010. 556 с.
5. *Lloyd S.* Least square quantization in PCM's // Bell Telephone Laboratories Paper, 1957.
6. *MacQueen J.* Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. on Math. Statistics and Probability, 1967. P. 281–297.
7. *Steinhaus H.* Sur la division des corps matériels en parties // Bull. Acad. Polon. Sci., C1. III. Vol. IV. 1956. P. 801–804.

Two-step algorithm for the classification of hyperspectral data in the space of the spectral brightness coefficients of aerial photography

V.N. Ostrikov, S.I. Smirnov, V.V. Mikhailov

LUCH Construction Bureau, St.-Petersburg Division
E-mail: luchmail@spb.vega.su

In the paper the problem of the classification of hyperspectral data taken in the space of coefficients of the spectral brightness is observed. It is assumed that each object in the hypercube belongs to a spectral class contained in the database, formed by ground-based measurements of the spectral characteristics of objects. The inputs to the algorithm are the hyperspectral data, passed calibration, filtering of regular and random noise, geometric correction, and a database of measured spectral brightness. Classification is carried out in two steps. At the first step the classes that

obviously can not belong to the object are screened, using a rather “rough” measure of proximity. The procedure can significantly reduce the amount of data to be processed, which will positively affect the speed of the algorithm. In the second step from a set of classes, “remaining” after the pre-classification, the closest to the object elements of the database are selected using a more sensitive to the spectral difference metric (e.g., Terebizh’s metric, Euclid metric). The algorithm is tested on hyperspectral survey data obtained from aircraft carrier under different conditions of observation. The robustness of the algorithm was revealed in a wide range of signal-to-noise ratio. A comparison of the classification results on real images was held.

Keywords: classification of hyperspectral data, the Terebizh’s metric, clustering.