

Структурная оптимизация алгоритма распознавания на базе КЭКМ*

А.Д. Рябушева¹, Л.И. Лебедев²

lebedev@pmk.unn.ru | pmk@unn.ac.ru

¹Факультет вычислительной математики и кибернетики,

²НИИ прикладной математики и кибернетики,

Нижегородский государственный университет им. Н.И. Лобачевского

В работе описывается алгоритм распознавания с самообучением графических изображений на базе корреляционно-экстремального контурного метода (КЭКМ) с использованием иерархического классификатора древовидной структуры. Специфика построения классификатора состоит в том, что его структура формируется на основе кластерного анализа неопознанных объектов, пополняющих список эталонов, и ведется в автоматическом режиме в процессе распознавания. Построение классификационного дерева позволяет существенно понизить вычислительную сложность алгоритма распознавания с самообучением.

Ключевые слова: распознавание с самообучением, КЭКМ, иерархический классификатор, кластер, оценка сходства, объект изображения, признаки описания, образ, эталон.

Structural Optimization of Recognition Algorithm on the Basis CECM*

A.D. Rybusheva¹, L.I. Lebedev²

¹Faculty of Computational Mathematics and Cybernetics,

²Research Institute of Applied Mathematics and Cybernetics,

Lobachevsky State University of Nizhni Novgorod

The paper describes a recognition algorithm of graphic images with self-learning on the basis of correlation-extreme contour method (CECM) using a hierarchical tree structure of the classifier. Specifics of the classifier consists in the fact that its structure is formed on cluster analysis of unrecognized objects, replenishing the list of standards, and is conducted in automatic mode in the recognition process. Constructing a classification tree can significantly reduce the computational complexity of the algorithm recognition with self-learning.

Keywords: recognition of self-learning, CECM, hierarchical classifier, cluster, evaluation similarity, object of image, signs descriptions, pattern, reference.

1. Введение

Корреляционно-экстремальные контурные методы распознавания базируются на вычисление оценок сходства эталонов с объектом, инвариантных относительно ортогональных преобразований и масштабирования (ОПМ) [1, 2]. Так как КЭКМ относятся к методам сравнения объекта с эталонами, поэтому их быстродействие зависит, в том числе, и от количества предъявляемых для распознавания образцов. В условиях большого разнообразия объектов на изображении (количество классов сотни, тысячи и более) требование к снижению временной сложности алгоритма распознавания на базе КЭКМ по-прежнему остается актуальным. Поэтому, оптимизация временной сложности алгоритма распознавания предполагает либо возможность принятия решения на основе частично вычисленных оценок близости, либо использование технологий и средств параллельного программирования. Естественное решение этой задачи заключается в разбиении потока последовательного вычисления

оценок близости на несколько потоков, в каждом из которых используется только некоторая часть исходного состава эталонов. При этом решение о принадлежности объекта к тому или иному классу может приниматься как по результатам вычислений внутри одного потока, так и по совокупности результатов вычислений во всех потоках. Рассмотрим два направления в разбиении последовательного потока вычисления оценок близости.

Наиболее простой по исполнению является возможность использования многоядерных компьютеров для распараллеливания вычислений оценок близости объекта с эталонами. Этот способ распараллеливания алгоритма распознавания на базе КЭКМ является наиболее эффективным при распределении вычислений по выбранной группе шрифтов и использовании суперкомпьютеров (то есть без существенного ограничения числа ядер).

Второе направление связано с построением классификатора с древовидной структурой для некоторой группы эталонов, например, для группы, представляющей символы какого-либо шрифта. В этом случае сложность распознавания будет определяться только суммарной сложностью вычисления всех

Работа выполнена и опубликована при поддержке РФФИ, гранты 13-07-00521, 15-07-20347.

оценок сходства с эталонами данного листа классификационного дерева.

Для решения задачи классификации некоторого множества эталонов необходимо сформировать признаковое пространство для их описания. Очевидно, что получение компонент этого пространства должно быть быстродействующей процедурой, значительно менее затратной, чем вычисление оценки сходства эталона с объектом. В нашем случае такая система признаков может быть сформирована на основании параметров, полученных от согласования описаний эталона и объекта, необходимого для вычисления оценки сходства. Для согласования описаний используются методы относительных смещений и метод парабол [3]. Метод относительных смещений основан на вычислении оценок близости $\varepsilon_m(\mathbf{B}, \mathbf{E}_j, s = s_i)$ базового эталона \mathbf{B} с эталоном \mathbf{E}_j при различных положениях начальной точки описания контура $s = s_i$. Местоположения начальных точек берутся в порядке обхода контура эталона с шагом приращения ΔS по длине контура. Значение находится ΔS из выражения $\Delta S = S^j/n$, где S^j - длина контура, n - число точек для получения различных описаний контура эталона. Тогда описание эталона \mathbf{E}_j можно задать системой признаков $c_{ij} = \varepsilon_m(\mathbf{B}, \mathbf{E}_j, s_i)$, $i = 1, 2, \dots, n$, где $s_i = i \cdot \Delta S$. Для того, чтобы предлагаемая система признаков была инвариантна не только относительно ОПМ, но относительно исходного описания эталона (объекта) циклической перестановкой сделаем так, чтобы первый признак был наименьшей величиной, то есть $c_{j1} = \min c_{ji}$, $i = 1, 2, \dots, n$. На рисунке 1 представлены графики оценок близости с базовым эталоном $\varepsilon_m(\mathbf{B}, \mathbf{E}_j, s)$ некоторых символов топографического шрифта Т-132, полученных при малых значениях шага ΔS . Из соображений быстродействия для вычисления оценок близости $\varepsilon_m(\mathbf{B}, \mathbf{E}_j, s = s_i)$ в качестве базового эталона был определен треугольник Пифагора (только три точки в исходном описании контура). На этом же рисунке отмечены значения четырех аргументов $s = \nu_q$, в которых вычисленные оценки близости $\varepsilon_m(\mathbf{B}, \mathbf{E}_j, s = \nu_q)$ обеспечивают решение задачи классификации шести представленных символов.

2. Постановка задачи

Отметим требования, которые должны выполняться при решении задачи построения классификатора. Во-первых, число признаков описания каждого из эталонов определяется количеством вычисляемых оценок близости в методе относительных смещений, то есть $n = 20$. Во-вторых, должны быть использованы только методы автоматической классификации, так как состав эталонов заранее неизвестен и определяется в процессе распознавания с самообучением. В третьих, по мере уве-

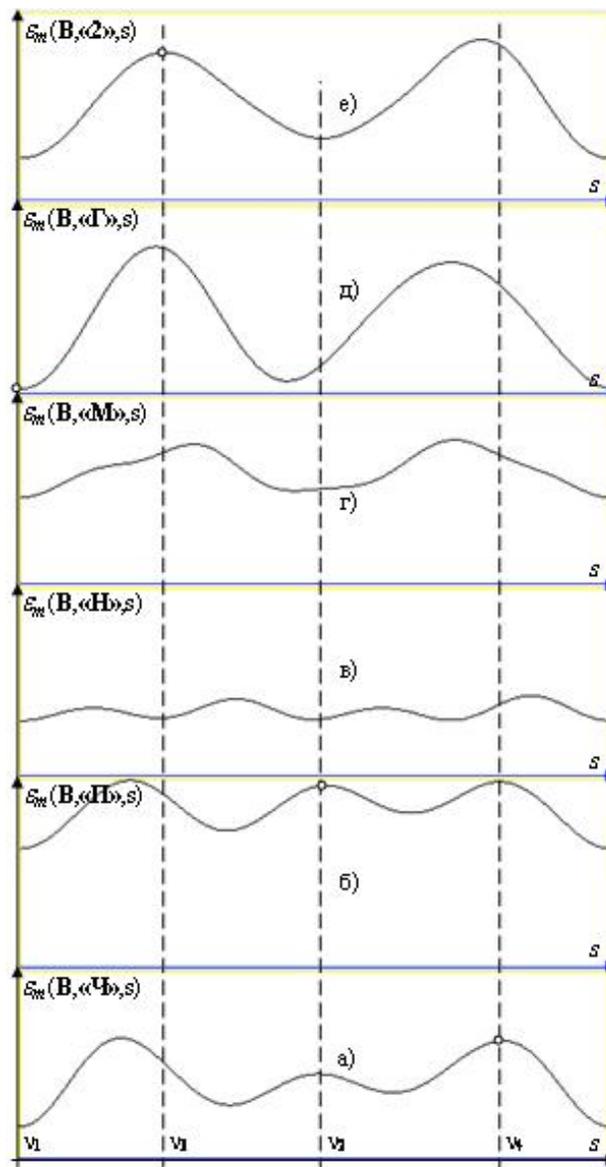


Рис. 1: Графики оценок близости с базовым эталоном.

личения числа эталонов для обеспечения быстродействия алгоритма распознавания должна иметься возможность перестроения (обновления) полученного классификатора в целях уменьшения числа представителей на листах дерева. В-четвертых, построение классификатора должно вестись одновременно с процедурой распознавания. Наконец, временная сложность построения классификатора должна быть ниже, чем сложность распознавания с использованием того множества эталонов, для которого строится или обновляется классификатор.

3. Методы решения

В результате исследования, проведенного в целях выбора алгоритма построения классификатора, удовлетворяющего сформулированным вы-

ше требованиям, был взята следующая технология разбиения множества эталонов на кластеры и построения решающих правил. В качестве способа разбиения множества на кластеры был взят алгоритм нечёткой кластеризации методом k ближайших соседей (k -means) [4]. В качестве входных параметров алгоритм k -means следующие данные:

- Тренировочная выборка.
- Количество кластеров.
- Максимальное количество итераций.
- Метод вычисления расстояния.

При разбиении заданной выборки эталонов на кластеры в любой вершине классификационного дерева их возможное количество задавалось равным трем. В этом случае сложность разбиения множества точек объемом v при ограничении на количество итераций числом l будет определяться выражением $O(3 \cdot v \cdot l)$.

Для построения решающих правил был выбран алгоритм SVM (support vector machine) - машина опорных векторов с полиномиальным ядром отображения в спрямляющее пространство признаков [5]. Входными данными для алгоритма SVM являются кластеры, полученные в результате применения алгоритма k -means. Алгоритм SVM при построении разделяющей гиперплоскости максимизирует расстояние между опорными векторами (ближайшими точками кластеров в спрямляющем пространстве). Построение решающих правил осуществляется независимо для каждой пары кластеров (дихотомия). Сложность построения решающих правил в каждой вершине зависит только от объема выборки в ней и оценивается величиной $O((3 \cdot v)^4)$.

На базе полученных алгоритмом SVM расстояний между кластерами в данной вершине классификатора могут быть созданы две или три дочерние вершины в зависимости от результатов проведенного разделения множеств:

- Расстояния между кластерами **A**, **B** и **C** удовлетворяют заданному порогу - образуются три дочерних вершины с формированием обучающих выборок на основе точек этих кластеров;
- Расстояние между одним из кластеров **A** и объединением двух других **B ∪ C** больше δ (расстояние между объединяемыми кластерами было меньше δ) - образуются две дочерних вершины с формированием обучающих выборок кластеров **A** и **B ∪ C**;
- Расстояние только между двумя из кластеров **A** и **B** больше δ - образуются две дочерних вершины с формированием обучающих выборок кластеров **A ∪ C** и **B ∪ C**;
- Все расстояния между кластерами меньше δ - вершина объявляется листовой.

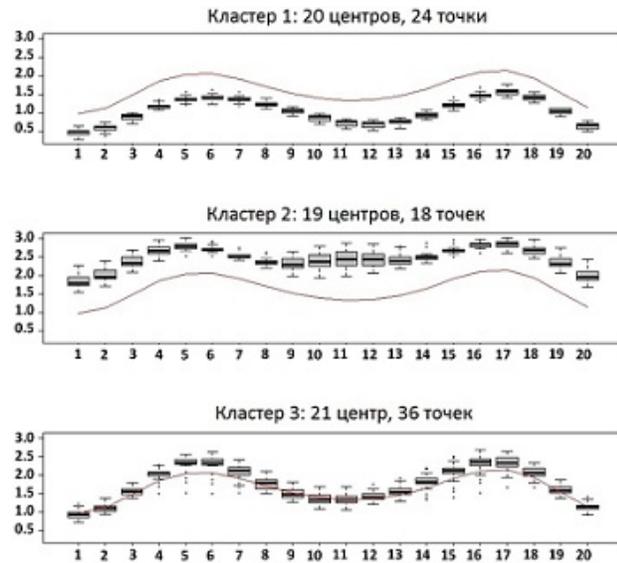


Рис. 2: Кластеры начальной классификации.

При поступлении нового эталона определяется листовая вершина классификатора, которой он будет принадлежать. Со временем, накопление эталонов, приводит к тому, что объект, поступивший на распознавание, будет попадать в лист с большим количеством эталонов. Это повлечет снижение быстродействия алгоритма распознавания, и следовательно, потребует необходимость перестройки классификатора. Для каждой обремененной листовой вершины возможны два варианта обновления классификатора. Если уровень данного листа и количество эталонов не входящих в его кластер допускают создание дочерних вершин, то вершина считается висячей и доработка классификатора производится обычным образом. Если количество объектов не входящих в кластер данного листа меньше требуемого, то будет рассматриваться предок этой вершины. В ней объединяются в кластер все эталоны, не принадлежащие кластерам листов всех ветвей классификатора. В итоге, всплытием вверх по дереву находится вершина, для которой имеется возможность построения решающих правил для выделения точек вновь образованного кластера. При этом из списка эталонов всех листов нижних уровней данной вершины удаляются точки нового кластера.

4. Полученные результаты

Описанный алгоритм получения иерархического классификатора с древовидной структурой был апробирован на информации о символах топологического шрифта T-132. Исходное количество эталонов включало 78 символов. В результате построенный классификатор содержит 35 листовых вершин с максимальной глубиной из 10-и уровней и минимальной глубиной поддеревя из 4-х уровней.

Среднестатистическая глубина составила 7.37 уровней. 50 эталонов в дереве принадлежат только одному листу, 15 эталонов отметились присутствием на двух листах, а для классификации 12 эталонов использовались три ветви дерева и наконец, один эталон присутствует на четырех листах дерева. На рисунке 2 показано разбиение исходного множества эталонов на три кластера.

5. Заключение

Проведенный анализ сложности алгоритма построения и перестройки классификатора показал, что они могут быть вписаны в процедуру распознавания при существенном увеличении его быстродействия. Если предположить, что каждый эталон используется для опознавания в среднем 40 объектов, то сравнение быстродействия алгоритма распознавания с алгоритмом построения и перестройки классификатора показало, что наибольшая эффективность от применения классификатора достигается, если обновление совершается, когда число новых эталонов будет в диапазоне от 150 до 200. При этом быстродействие распознавания увеличивается в среднем на порядок для графических изображений, содержащих около 45000 объектов при накоплении до 1100 эталонов.

Литература

- [1] *Васин Ю.Г., Лебедев Л.И., Пучкова О.В.* Контурные корреляционно-экстремальные методы обнаружения и совмещения объектов видеоинформации. // Автоматизация обработки сложной графической информации: Межвуз.темат.сб.науч.тр. // Под ред. Ю.Г.Васина.- Горьков.гос.ун-т, Горький, 1987. С.97-112.
- [2] *Васин Ю.Г., Лебедев Л.И.* Задача нахождения согласованных описаний в корреляционно-экстремальных контурных методах распознавания. // Математические методы распознавания образов (ММРО-15): 15-ая Всеросс. конф.: Сборник докладов. / М.: Изд-во ООО "МАКС Пресс 2011. – С.342-345.
- [3] *Лебедев Л.И.* Корреляционно - экстремальные контурные методы распознавания. Теоретические основы: Учебное пособие. – Нижний Новгород, Изд-во Нижегородского государственного университета, 2013. – 113 с.
- [4] http://www.basegroup.ru/glossary/definitions/k_means/. – BaseGroup Lab – 2015.
- [5] <http://www.basegroup.ru/glossary/definitions/svm/> – BaseGroup Lab – 2015.