СИСТЕМА ОТСЛЕЖИВАНИЯ КОНТУРА ГУБ ГОВОРЯЩЕГО

Губочкин Иван Вадимович,
Нижегородский государственный лингвистический университет им. Н.А. Добролюбова
05.13.17
jhng@yandex.ru

Аннотация. *В статье приводится описание архитектуры и особенностей реализации системы отслеживания контура губ говорящего в реальном масштабе времени. Даны результаты экспериментальной оценки ее быстродействия в различных режимах. Показано, что при выполнении части вычислительных операций с использованием графического процессора разработанная система позволяет существенно сократить время обработки анализируемых видеоданных в расчете на один кадр.*

Ключевые слова: речь, активный контур, активная контурная модель, аудиовизуальное распознавание речи.

THE SYSTEM FOR SPEAKER'S LIPS CONTOUR TRACKING

Gubochkin Ivan Vadimovich
Linguistics University of Nizhny Novgorod

Abstract. *The article describes the architecture and implementing features of a speaker's lips contour tracking system that work in real time. It is given results of experimental evaluation of its performance in different modes. It is shown that using a graphic processor unit dramatically reduces video data processing time.*

Keywords: speech, active contour, active contour model, audio-visual speech recognition.

Введение. В последнее время системы автоматического распознавания речи (АРР) получают все большее распространение. Они применяются для управления сложными технологическими процессами и системами, для организации удобного интерфейса с поисковыми и информационными системами, управления мобильными телефонами, роботами и т.д. Однако, несмотря на существенный прогресс, достигнутый в данном направлении, системы АРР имеют ряд недостатков, которые ограничивают область их применения. Один из наиболее существенных – недостаточная помехоустойчивость. Для его преодоления разработано множество методов и подходов. Среди них можно отметить двухуровневые марковские модели, методы, основанные на генетических алгоритмах, гибридные модели и другие [1]. Кроме того, одним из самых эффективных подходов к повышению помехоустойчивости системы АРР является применение аудиовизуального распознавания речи [2]. При данном подходе на вход системы распознавания поступает кроме речевого сигнала также видеоизображение говорящего. Таким образом, для распознавания речи используются сразу два информационных канала. Подобный поход позволяет существенно (более чем на 20%) сократить величину ошибки перепутывания слов в условиях высокого уровня фоновых шумов [3].

Системы аудиовизуального распознавания речи строятся по принципам, которые сходны с принципами построения обычных систем АРР. Различие заключается главным образом в том, что требуется, во-первых, производить вычисление векторов признаков еще и в видеоканале, а, во-вторых, объединять каким-либо образом информацию, поступающую из двух источников данных перед проведением процедуры распознавания. Для решения последней задачи в настоящее время широко применяются сдвоенные скрытые марковские модели [3, 4]. Что касается вычисления вектора видео-признаков, то наиболее информативной его составляющей является траектория движения губ говорящего. Для практической реализации операции отслеживания этой траектории в настоящее

время широко используются активные контуры или активные контурные модели [5]. В этом случае, как правило, требуется осуществить первичную подготовку анализируемого фрагмента изображения (выравнивание яркости, выделение интересующих фрагментов, бинаризацию), а затем непосредственно провести адаптацию активного контура к полученной бинарной маске. Все это требует существенных вычислительных ресурсов, что затрудняет реализацию систем аудиовизуального распознавания речи, работающих в реальном времени. Поэтому задача построения быстродействующей системы слежения и оценки параметров контура губ с использованием современных вычислительных средств является весьма актуальной. Ее решению и посвящена данная статья.

Постановка задачи. На рисунке 1 показана структурная схема, на которой отражены основные этапы обработки изображения, полученного из какого-либо источника (например, от видеокамеры или из видео-файла), для определения расположения губ говорящего и связанных с ними параметров.

В данной работе мы не будем подробно рассматривать задачу поиска лица человека на изображении. Кроме того логично предположить, что при работе с системой аудиовизуального распознавания речи в диалоговом режиме пользователь не будет слишком активно двигать головой. А это значит, что поиск лица можно будет проводить сравнительно редко (порядка 10 кадров/сек. и даже менее). Подобный результат легко достигается с использованием существующих программных средств. В качестве примера здесь можно привести реализацию алгоритма обнаружения лиц в широко известной библиотеке алгоритмов компьютерного зрения и обработки изображений OpenCV [6]. Вместе с тем, движения губ говорящего во время разговора являются достаточно активными, а, следовательно, оценка их контура должна производиться с большей интенсивностью. Поэтому, с учетом того, что современные системы распознавания речи используют посегментную обработку данных с длительностью сегмента 10 – 20 мс [7], представляется целесообразным проводить анализ видео-потока синхронно с аудио-потоком. Т.е. частота обработки видео данных должна составлять порядка 50 кадров/сек. Следовательно, время, необходимое на оценку параметров контура губ, не должно превышать 20 мс. В общем же случае оно должно быть как можно меньше, чтобы иметь возможность эффективно использоваться оставшееся процессорное время для выполнения других операций, связанных с распознаванием речи (синтаксического разбора и семантического анализа, слежением за лицом диктора, вывода результатов распознавания) в реальном времени.

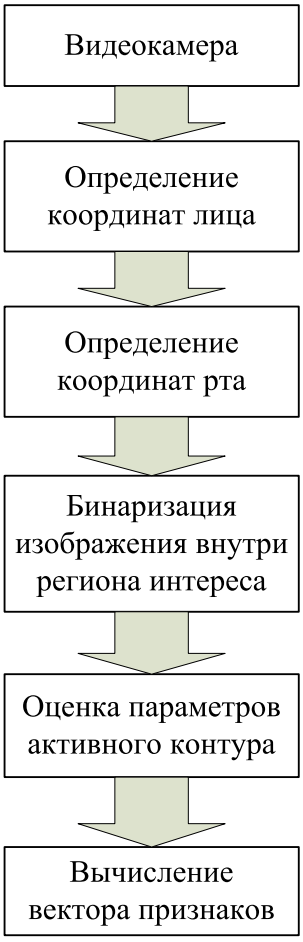

Рис. 1. Этапы обработки видеоизображения.

Описание программной архитектуры. Архитектура рассматриваемой системы строилась с учетом поставленных выше требований, а также основных направлений развития информационных технологий. Поэтому к ней были предъявлены следующие требования:

1. высокое быстродействие;
2. кроссплатформенность;
3. гибкое использование возможностей аппаратного ускорения работы.

Для достижения высокой скорости работы в качестве языка реализации системы был выбран C++. Кроссплатформенность достигается за счет постро-

ения системы на основе библиотек, реализованных для различных операционных систем и аппаратных платформ, что позволило минимизировать объем платформозависимого кода. В качестве поддерживаемой на данный момент программно-аппаратной платформы выбрана Intel x86 (как одна из наиболее распространенных) для операционных систем Windows и Linux. Также реализовано три режима использования аппаратного ускорения:

1. без ускорения;
2. ускорение с использованием инструкций центрального процессора (MMX и SSE инструкции);
3. ускорение с использованием графического процессора NVIDIA. Для реализации данного режима применялась технология NVIDIA CUDA [8].

Прежде чем описывать архитектуру разработанной системы, рассмотрим основные операции обработки изображения, показанные на рисунке 1, более подробно. Для простоты предположим, что мы уже каким-то образом выделили область рта диктора. В этом случае основные операции по обработке полученного изображения, которые нам необходимо выполнить, представлены на рисунке 2.

Рис. 2. Основные этапы обработки изображения.

На первом шаге производится переход в новое цветовое пространство. Это необходимо в первую очередь для того, чтобы избежать влияния интенсивности освещения. В качестве цветового пространства в рассматриваемой системе выбрано $\{ R, G, Cb, Cr \}$, где $R = \frac{r}{r+g+b}$, $G = \frac{g}{r+g+b}$, а Cb и Cr – соответствующие компоненты цветового пространства YCbCr. На следующем шаге происходит бинаризация изображения в новом цветовом пространстве для того, чтобы провести разделения области кожи лица и области кожи губ. Одним из наиболее простых и, вместе с тем, надежных способов провести такое разделение является классификация всех точек изображения в цветовом пространстве $\{ R, G, Cb, Cr \}$ с помощью гауссовых смесей. Данную задачу можно свести к простой проверке гипотез:

$$W_{\nu}(X): L_{X,\nu} = \\ = \sum_{m} c_{mr} N(X, \mathbf{M}_{mr}, \mathbf{K}_{mr})\big|_{\nu=r} \to \max \ \ r = \overline{0,1} \quad (1)$$

Здесь X – анализируемая точка изображения, $L_{X,\nu}$ – функция правдоподобия, $N(\bullet)$ – нормальный закон распределения плотности вероятностей с математическим ожиданием $\mathbf{M}_{mr}$ и автокорреляционной матрицей $\mathbf{K}_{mr}$, c_{mr} – вес m-й компоненты смеси r-го класса. В этом случае точкам, относящемся к коже лица мы можем назначить цвет 0, а точкам, относящемся к коже губ – цвет 1. Таким образом мы получаем бинарную маску анализируемого изображения и переходим к следующему этапу – оценке параметров активного контура.

Согласно [9] активный контур – это деформируемая модель, шаблон которой задан в форме параметрической кривой, инициализированной вручную набором контрольных точек, лежащих на открытой или замкнутой кривой на входном изображении.

В общем случае активный контур задается параметрической кривой $\mathbf{v}(s) = [x(s), y(s)]$, $s \in [0,1]$ На практике кривая обычно является замкнутой, т.е. $\mathbf{v}(0) = \mathbf{v}(1)$. Задача состоит в нахождении такого расположения точек этой кривой, при которой энергия контура минимальна. Обычно энергия задается следующим образом:

$$E = \int_0^1 E_{\text{int}}[\mathbf{v}(s)] + E_{ext}[\mathbf{v}(s)]\, ds, \quad (2)$$

где E_{int} – внутренняя энергия контура, определяющая гладкость кривой, а E_{ext} – внешняя энергия контура, которая определяется внешними силами,

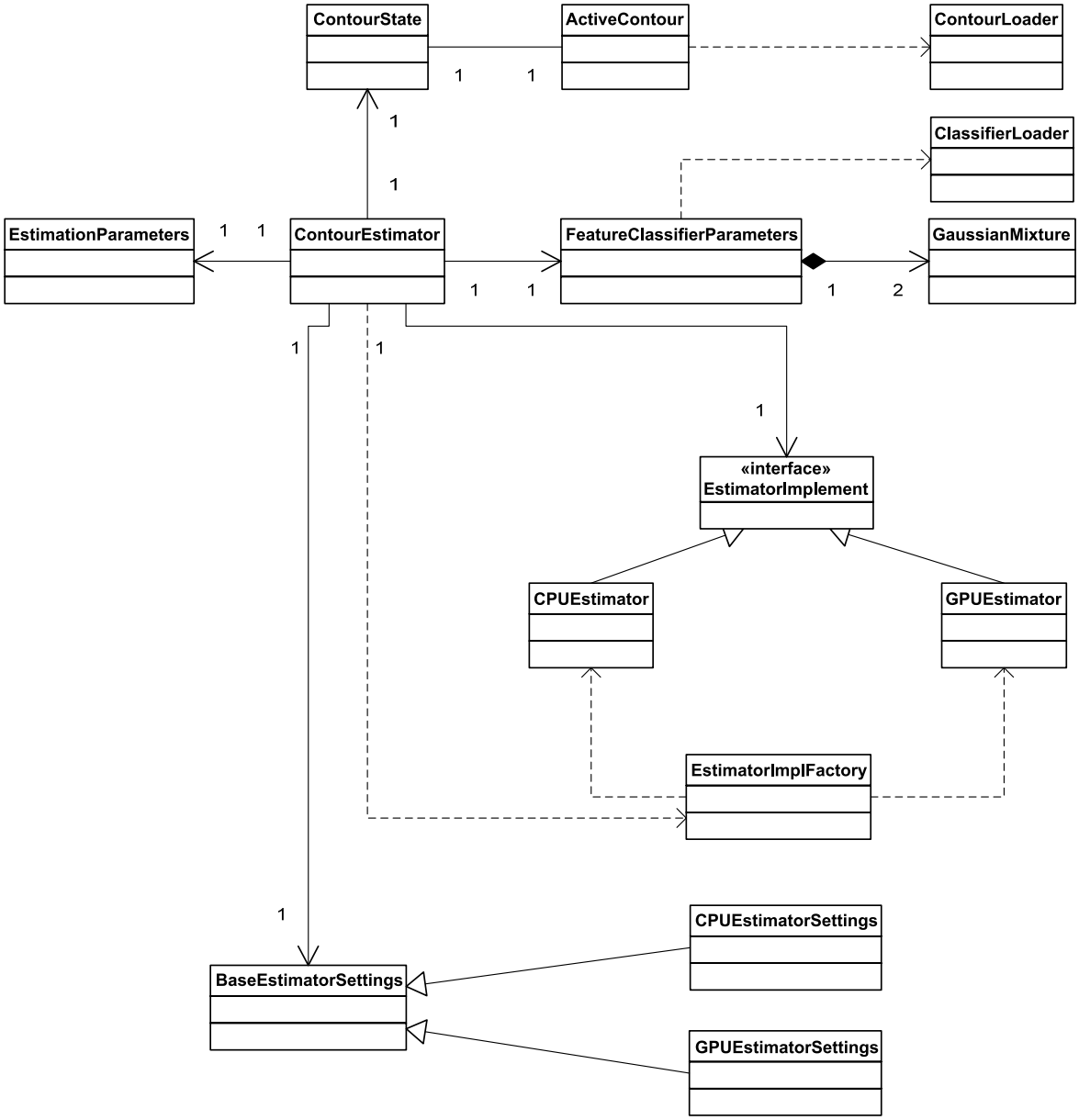

Рис. 3. Диаграмма классов системы.

действующими на контур. На практике обычно применяется дискретная аппроксимация кривой контура, а для минимизации выражения (2) используют различные способы итерационной численной оптимизации, например, градиентные методы.

С учетом вышеизложенного возникает задача программной реализации рассмотренных выше этапов обработки и анализа изображения. При этом необходимо соблюсти поставленные требования по быстродействию и гибкости реализации.

Для решения поставленной задачи была разработана программная система под названием ACLD (Active contour lips detection) [10], основной задачей которой является локализация губ на изображении при помощи активного контура и вычисления вектора признаков на основе полученных данных. Такой вектор признаков в дальнейшем может использоваться в работе систем компьютерного зрения, распознавания речи, биометрических систем и т.д.

В процессе разработки проводился анализ возможностей аппаратного ускорения вычислений. Выяснилось, что такое ускорение (как с использованием центрального, так и графического процессора) может быть легко организовано при преобразовании цветового пространства и бинаризации изображения с использованием классификатора (1), поскольку на этих этапах проводится независимая обработка большого числа однотипных объектов (точек).

Сложнее дело обстоит с ускорением вычислений при адаптации активного контура. Основная проблема здесь заключается в том, что для практического применения контур, как правило, представляется небольшим количеством точек (порядка 10 – 20). Это приводит к тому, что на скорость вычислений начинают оказывать существенное влияние накладные расходы, связанные с аппаратным ускорением. Так, при использовании технологии CUDA временные затраты, связанные с обменом данными, запуском на выполнение вычислительных процедур на графической подсистеме составили порядка 5 – 20 мкс в расчете на одну итерацию, что в конечном счете приводило к замедлению работы алгоритма адаптации активного контура. Поэтому алгоритмы работы с активными контурами небольшого размера эффективней реализовывать с помощью обычных вычислительных средств.

На рисунке 3 показана диаграмма классов, реализующих архитектуру разработанной системы, в которой учтены рассмотренные выше особенности.

Представленная диаграмма, однако, лишь реализует высокоуровневый интерфейс взаимодействия с внешними компонентами. Это необходимо для исключения необходимости управления параметрами и настройками, зависящими от выбора режима работы, платформы, под управлением которой работает система и т.д. Основная логика предварительной обработки анализируемого изображения и оценки параметров активного контура реализована в двух дополнительных компонентах, которые предоставляют возможности по аппаратному ускорению вычислений с использованием центрального или графического процессора.

Результаты экспериментальных исследований. Основной целью экспериментальных исследований являлась оценка быстродействия разработанной системы отслеживания контура губ. При этом в качестве такой оценки использовалось время адаптации активного контура к анализируемому изображению. Для проведения экспериментов использовался видео-файл с записью диктора произносящего фразу «*joe tooks fathers green bench out*». Объем видеозаписи составлял 103 цветных кадра форматом 720x576 точек. Применявшийся для моделирования компьютер имел следующую конфигурацию: операционная система Ubuntu Linux 10.04, центральный процессор Athlon X245 3 ГГц, 4096 МБ оперативной памяти, видеокарта NVIDIA GTS450 с 1024 MB встроенной памяти.

Для уменьшения объема выполняемых вычислений и повышения точности работы системы внутри кадра предварительно в автоматическом режиме выделялась область губ. Выделенная область изображения передавалась для анализа с помощью разработанной программной системы. В качестве примера на рисунке 4 показан результат обработки одного из кадров.

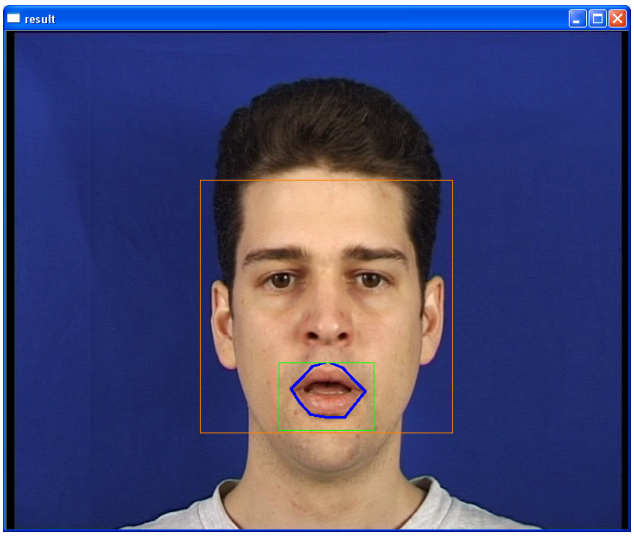

Рис. 4. Результат обработки видеокадра.

Здесь зеленой линией обозначен анализируемый фрагмент изображения, а синей линией обведен контур губ, найденный системой. Из рисунка видно, что он достаточно хорошо совпадает с реальными границами кожи губ.

Ниже на рисунке 5 показано среднее время адаптации активного контура к анализируемому фрагменту изображения губ человека в расчете на один кадр для различных режимов аппаратного ускорения вычислений в зависимости от количества смесей классификатора (1). Измерение проводилось при следующих установленных параметрах:

- количество точек активного контура – 12;
- максимальное количество итераций алгоритма минимизации энергии контура (2) установлено равным 50;
- кривая контура параметризована бикубическим сплайном.

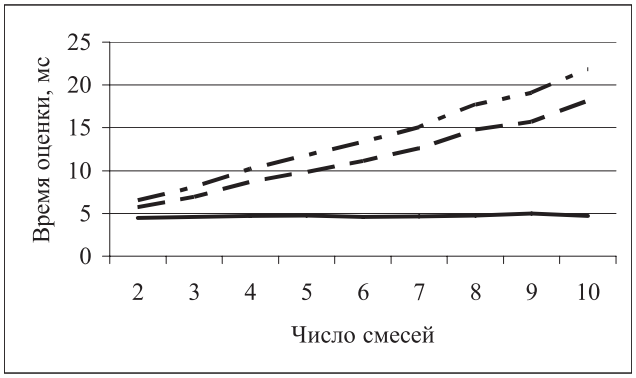

Рис. 5. Время обработки видеокадра.

Здесь сплошной линией показано время обработки кадра с использованием графического процессора, штриховой линией – время обработки с использованием средств ускорения вычислений центрального процессора, штрихпунктирной линией – без ускорения. Из рисунка видно, что наибольший прирост быстродействия достигается при использовании графического процессора. При этом выигрыш в быстродействии растет с увеличением сложности классификатора. Например, для 10 смесей время обработки кадра сокращается более чем в 4 раза по сравнению с работой без какого-либо аппаратного ускорения вычислений.

В заключении на рис. 6 приведена зависимость вычисляемых на основе активного контура признаков от времени (в качестве основы использована работа [11]). На верхнем графике приведены зависимости значений 5 компонент вектора признаков, вычисленных по координатам точек активного контура с применением метода главных компонент. На среднем графике показано отношение ширины контура к его высоте. На нижнем графике приведена временная диаграмма речевого сигнала, сопровождающего видео.

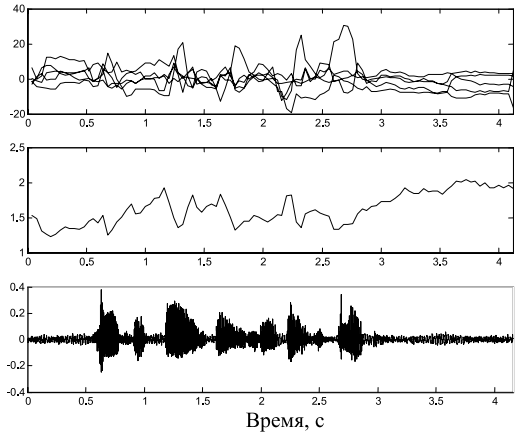

Рис. 6. Видео-признаки и речевой сигнал.

Приведенные графики позволяют сделать вывод о том, что получаемые на основе активного контура признаки имеют определенную зависимость от речевого сигнала. Следовательно, их можно использовать в задачах аудиовизуального распознавания и анализа речи.

Выводы. В работе предложена архитектура системы отслеживания в реальном времени контура губ и построения по нему вектора признаков. Показано, что использование графического процессора позволяет в несколько раз повысить быстродействие применяемых алгоритмов. Отсюда можно сделать вывод о целесообразности использования графических подсистем компьютера в решении рассмотренной задачи анализа изображения с использованием активных контуров.

Полученные результаты в перспективе могут позволить повысить точность работы систем распознавания речи за счет использования визуальной информации. Также предложенная система может найти применение и в других направлениях анализа и обработки аудиовизуальных сигналов.

Список литературы

1. Grimm M., Kroscel K. Robust speech recognition and understanding / I-Tech, 2008, 468 pp.
2. Neti C., Potamianos G., Luettin J. and oth. Audio-Visual speech recognition: An ovierview. / Issues in Visual and Audio-visual Speech Processing, 2004.
3. Bengio S. An asynchronous hidden markov model for audio-visual speech recognition / NIPS, 2003, URL: http://books.nips.cc/papers/files/nips15/SP07.pdf
4. Карпов А.А. Автоматическое распознавание аудиовизуальной русской речи с применением асинхронной модели // Информационно-измерительные и управляющие системы. №7, 2010, с. 91 – 96.
5. Blake A., Isard M. Active contours, Springer, 1998.
6. Bradski G., Kaehler A. Learning OpenCV: Computer vision with the OpenCV library. O'Reilly, 2008.
7. J. Picone, "Continuous Speech Recognition Using Hidden Markov Models," IEEE ASSP Magazine, vol. 7, no. 3, pp. 26-41, July 1990.
8. NVIDIA CUDA Compute Unified Device Architecture. // NVIDIA corp.
9. Дегтярева А.А. Деформируемые модели. // Компьютерная графика и мультимедиа. Вып. №3(2) / 2005. URL: http://cgm.computergraphics.ru/content/view/75
10. Губочкин И.В. Программа отслеживания контура губ в видеопотоке: программа для ЭВМ / Роспатент. Свидетельство о гос. регистрации №2011618786 по заявке №2011617355 от 27.09.2011.
11. Soldatov S. Lip reading: Preparing feature vectors / GraphiCon, 2003, pp. 254 – 256.