# Билет 6. Проблема понимания изображений. Задача распознавания изображений в сравнении с задачей их сопоставления. Проблема интерпретации сцен

Системы компьютерного зрения часто выступают в качестве компонента системы управления или принятия решения, предназначенной для решения некоторой задачи в реальном мире, которая обычно ставится в терминах объектов и ситуаций. Восстановление трехмерных форм и движения объектов, описанных в терминах совокупности поверхностей, оказывается недостаточным.

Представим, к примеру, систему управления транспортным средством. Информация о трехмерной структуре сцены является весьма полезной, так как позволяет устанавливать направление возможного движения и обнаруживать препятствия. Однако по трехмерным характеристикам обочина или тротуар могут мало отличаться от проезжей части. Также в качестве препятствий не будут различаться пешеход и куст на разделительной полосе. Для принятия правильных решений, особенно, в аварийно опасных ситуациях крайне важно такое различие устанавливать.

Таким образом, от систем компьютерного зрения может требоваться не только восстановление физических характеристик объектов, но и построение осмысленного описания сцен. Семантическая интерпретация изображений называется *пониманием изображений*.

Проблемы понимания изображений могут варьироваться от распознавания или обнаружения единичных объектов по их изображениям до построения полного описания сцены в рамках некоторой системы представления знаний. Иногда под системами понимания изображений подразумевают системы обработки изображений, включенные в системы искусственного интеллекта, однако такое определение (даже если его и считать верным) не говорит о содержательной стороне проблемы понимания изображений.

Следует также отметить, что в контексте проблемы понимания изображений термин «понимание» обычно используется в смысле *процедурной семантики*: считается, что система «понимает» смысл изображений или языковых высказываний, если может правильно действовать в соответствии с ними. К примеру, если система может выполнить команду «принеси зеленую пирамиду, расположенную на красном кубике», то считается, что система понимает смысл этого высказывания, равно как понимает и содержание изображений, на основе которых выполняет команду.

*Распознавание объектов*

Часто для «понимания» содержания изображения в рамках решаемой задачи системе компьютерного зрения может быть достаточно детектировать или распознавать отдельные объекты. К примеру, если ее рассматривать в качестве компоненты системы экологического мониторинга, то требуется обнаруживать возникновение пожаров, наводнений или других стихийных бедствий без составления полного описания сцены (рис. 4.13).

Рис. 4.13. Пример обнаружения дыма на изображении, что можно трактовать как простейшего «понимание» системой компьютерного зрения содержания изображения в рамках решаемой задачи



Распознавание объектов по их изображениям опирается на методы сопоставления изображений и может также выполняться в рамках различных представлений изображений. Однако при распознавании производится сравнение текущего изображения с несколькими эталонными изображениями (вернее, с результатом их отображения в выбранное представление). Таким образом, одним из ключевых моментов распознавания оказывается определение меры сходства изображений, на основе которой можно осуществить отнесение изображения к одному из классов.

В качестве такой меры сходства достаточно часто используется та же мера сходства, которая привлекается в процессе сопоставления пары изображений. Однако к мере сходства в задачах распознавания предъявляются более жесткие требования. К примеру, величина корреляции двух изображений

,

на основе которой может выполняться сопоставление изображений, не подходит для задачи распознавания. Действительно, простое повышение яркости одного из изображений приведет к повышению величины корреляции, хотя сходство изображений при этом не увеличится (повышением яркости можно добиться большого значения корреляции даже для несходных изображений). Более подходящим при распознавании будет коэффициент корреляции, который нормирован на среднеквадратичную яркость обоих изображений.

При чамферном сопоставлении меру Хаусдорфа тоже оказывается необходимым нормировать, поскольку на ее значение оказывают влияние не только сходство и различие двух изображений, но также и характеристики каждого из изображений в отдельности. Например, для небольшого контура на большом изображении легко найти фрагмент, на котором мера Хаусдорфа будет равна нулю, в то время как наложить два больших изображения с достижением малого значения этой меры значительно сложнее.

В случае структурного сопоставления мера сходства двух изображений, как правило, основывается на числе сопоставленных структурных элементов и среднеквадратичном отклонении их положений от глобального пространственного преобразования. Однако, опять же, в задачах распознавания этого может быть недостаточно.

Наиболее разработанными методами распознавания являются дискриминантные методы распознавания образов, работающие в рамках признаковых представлений. В этих методах каждый класс объектов отделяется от других совокупностью поверхностей в пространстве признаков, либо строятся статистические модели, на основе которых можно установить вероятность принадлежности любого объекта тому или иному классу. К сожалению, описание изображений в терминах векторов признаков сопряжено с потерей большого объема информации, что существенно сужает сферу применения этих методов.

Наиболее существенной проблема выбора адекватной меры сходства оказывается в задачах извлечения изображений из баз данных. В этих задачах количество изображений, среди которых требуется выбрать наиболее приемлемое, является чрезвычайно большим, но, к счастью, при этом в качестве решения может выступать не одно, а множество изображений. Тем не менее, на настоящий момент для решения этой проблемы привлекают, как правило, текстовые аннотации к изображениям.

*Интерпретация сцен*

Для интерпретации сцен требуется осуществлять выделение на них объектов с последующим распознаванием. Большинство естественных сцен содержит множество самых разнообразных объектов, поэтому распознавание всех объектов на сцене на современном уровне развития систем компьютерного зрения вряд ли возможно и во многих прикладных задач не нужно.

Обычно для решения задачи достаточно ограничиться распознаванием лишь значимых элементов сцены. К примеру, для системы управления транспортным средством описание сцены может включать выделенные на изображении области, соответствующие дороге, обочине, небу, пешеходам, машинам и прочим препятствиям (без их детального распознавания). Для бытового робота первоочередную значимость может представлять выделение на изображении областей пола, потолка, стен, дверей, статических и подвижных объектов (не разделяемых на классы). При интерпретации аэрокосмических снимков семантические метки должны назначаться областям леса, водоемов, полей, дорог и других антропогенных объектов (возможно, с разделением на несколько крупных классов).

Интерпретация сцен отличается от распознавания совокупности объектов не только необходимостью выделения этих объектов на изображении, но и необходимостью учета допустимого взаимного расположения объектов на сцене. Действительно, вряд ли небо на изображении может располагаться между дорогой и обочиной или пол – между дверью и потолком. Подобные отношения часто описываются в рамках некоторой системы представления знаний.

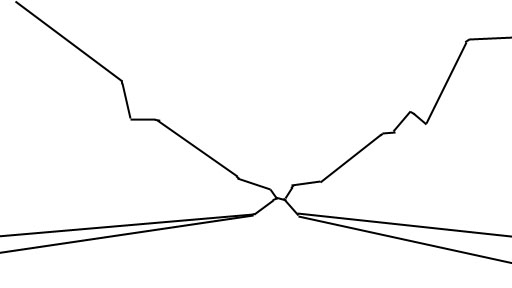
Рассмотрим пример такого описания в рамках логики предикатов. На рис. 4.14 приведено изображение. В качестве одного из первых этапов интерпретации обычно выступает сегментация изображения на области (рис. 4.15). Конечным результатом интерпретации сцены является назначение семантических меток каждой из областей.

Рис. 4.14. Пример интерпретируемого изображения дороги



Области на изображении могут быть описаны совокупностью признаков (например, текстурных), и на основе этих признаков каждой области могут быть назначены вероятность ее соответствия одному из допустимых классов.

Рис. 4.15. Результат сегментации изображения 4.14 на области



Region-1

Region-2

Region-3

Region-4

Region-5

Region-6

В данном случае представленные результаты сегментации являются весьма точными, однако на практике они обычно более фрагментарны, что усложняет задачу назначения меток. Привлечение знаний о предметной области может помочь снять неоднозначности интерпретации.

В представленном случае фрагмент базы знаний (общих правил) на языке предикатов может принять следующий вид.

region(X) & grass(X) => green(X)

region(X) & sky(X) & time(Y) => blue(X) & day(Y) V dark(X) & night(Y)

region(X) & region(Y) & grass(X) & sky(Y) => above(Y, X) & below(X, Y)

region(X) & region(Y) & sky(X) & sun(Y) => inside(Y, X)

В результате логического вывода должны быть получены частные факты

forest(Region-3) & forest(Region-4)

road(Region-1)

ground(Region-5) & ground(Region-6)

top(Region-2)

above(Region-3, Region-5)

to-left(Region-3, Region-4)

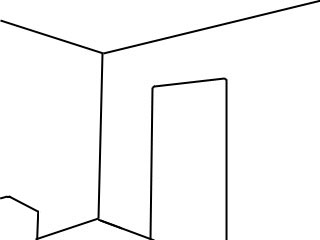
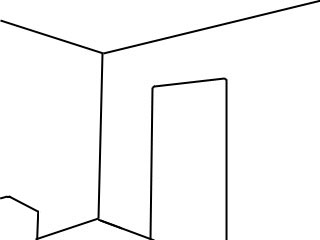
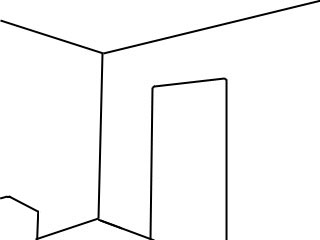
Здесь в качестве переменных могут выступать как элементы изображения (области, структурные элементы, значения каких-либо признаков), так и неизобразительные данные (время суток и года, текущие географические координаты, скорость движения и т.д.). Для некоторых из них могут быть заданы частные факты, полученные в результате обработки изображения (цвет той или иной области и т.д.) или из внешних источников (например, текущее время).

Иногда все элементы системы компьютерного зрения представляют в виде компонентов некоторой системы представления знаний. При этом выделяют три типа знаний: *перцепционное*, *семантическое* и *функциональное* знание. Перцепционное знание позволяет интерпретировать изображения в терминах линий, областей и т.д. (по сути, перцепционное знание реализуется в форме процедур анализа изображений). Семантическое знание описывает определенные абстрактные понятия, такие как форма или конкретные объекты и отношения между ними (приведенный выше пример относится именно к семантическому знанию). Функциональное знание предназначено для регулирования процесса интерпретации изображения в зависимости от предметной области и поставленной цели, то есть включает подсистему манипулирования знаниями (например, логического вывода).

Существуют следующие подходы к представлению знаний в системах интерпретации изображений: семантические сети, объектно-ориентированные представления и фреймы, продукционные системы, мультиагентный подход, представления, основанные на логике предикатов, и некоторые другие. Часто для записи знаний различного типа в одной системе может привлекаться несколько представлений.

Рассуждения в системах интерпретации сцен сопряжены с существенной долей априорной неопределенности, поэтому классические процедуры логического вывода здесь используются нечасто. Вместо этого может быть использован какой-либо вариант байесовских рассуждений или нечеткий вывод с постепенным выбором наиболее согласованной интерпретации (рис. 4.15).

Рис. 4.15. Пример последовательного устранения неопределенности интерпретации на основе знаний



CFWDT

CFWDT

CFWDT

CFWDT

CFWDT

CFWDT

C

WDT

WDT

WDT

F

CFWDT

C

D

W

T

F

W

Создание баз знаний является крайне трудозатратной операцией. Еще более сложной проблемой оказывается создание баз знаний, в которые интегрированы системы компьютерного зрения, поскольку оказывается необходимым в явном виде указывать взаимосвязи между перцептивным и семантическим (а также функциональным) знанием, чего не требовалось в классических экспертных системах. Не удивительно, что системы понимания изображений, основанные на знаниях, создаются, как правило, крайне узкоспециализированными.

Как и в случае классических экспертных систем, ключевой проблемой здесь оказывается автоматическое приобретение знаний. Таким образом, для создания систем понимания изображений общего назначения необходимо решение проблемы машинного обучения на основе перцептивного опыта. Существующие подходы в данной области заключаются в имитации процесса усвоения языка младенцем, который исходно не обладает никакими знаниями о свойствах различных объектов, однако способен научиться строить словесные описания сцен. Рассмотрим в качестве одного из типичных примеров данного подхода серию работ, выполняемых в Media Laboratory Массачусетского Технологического Института.

Развитие упомянутой серии работ началось с системы, получившей название CELL (Cross-channel Early Lexical Learning), которая впоследствии той же группой исследователей была развита в системы Describer, Newt, Ripley, Fuse. Далее мы кратко рассмотрим базовые принципы развиваемого в перечисленных выше работах подхода на примере системы CELL.