УДК 004.932.72'1

ОБНАРУЖЕНИЕ ОБЪЕКТОВ СИСТЕМАМИ КОМПЬЮТЕРНОГО ЗРЕНИЯ: ПОДХОД НА ОСНОВЕ ВИЗУАЛЬНОЙ САЛИЕНТНОСТИ

В.А. КОЧУРКО[1, 2], К. МАДАНИ[2], К. САБУРАН[2], В.А. ГОЛОВКО[1], П.А. КОЧУРКО[1]

[1] *Брестский государственный технический университет*
Московская, 267, Брест, 224017, Беларусь

[2] *Университет Пари-Эст Кретей, технологический институт Сенарт-Фонтенбло*
рю Жорж Шарпак, 36-37, Льюсан, 77127, Франция

Поступила в редакцию 23 марта 2015

Представлен комбинированный алгоритм выделения объекта на изображении и расчета карты вероятности фиксации взгляда, который можно использовать для прикладных задач автономного обнаружения. Экспериментальные результаты показывают жизнеспособность алгоритма и его эффективность в сравнении с несколькими state-of-art алгоритмами, предполагая его применимость в более широком классе задач – прикладных вариациях задачи прогноза фиксации взгляда.

Ключевые слова: визуальная салиентность, центрально-периферический антагонизм, эволюционная оптимизация, задачи прогноза фиксации взгляда, автономное обнаружение объектов.

Введение

Задачи обнаружения в области компьютерного зрения представляют собой задачи на поиск тех или иных объектов во входных визуальных данных (видеопотоке, наборе изображений) с помощью определенного условия. Сюда можно отнести как прикладные задачи (анализ видеопотока с шоссе на обнаружение проезжающих автомобилей), так и более фундаментальные (автономный поиск роботом конкретно заданного объекта в определенном пространстве).

Одно из направлений исследований в данной области (в том числе и применительно к задачам обнаружения) – механизм поиска визуальной салиентности, моделирующий один из аспектов человеческого когнитивного восприятия.

Салиентность – свойство объекта быть более заметным, чем окружающие его объекты (калька с английского слова saliency). В русскоязычной литературе этот термин употребляется в основном в рамках когнитивной лингвистики (например, [1]), но также используется и в трудах в области компьютерного зрения (КЗ) (например, [2]). Существуют также варианты «выделенность», «заметность» или «значимость», но они несут несколько иное семантическое значение и употребляются и как аналоги английского термина «saliency», так и как термины с иными смыслами даже в области КЗ (например, [2–4]). Визуальная салиентность в таком контексте представляет собой свойство объектов (пикселей, областей изображения) быть более заметными на изображении (на кадре в видеопотоке), чем окружающие его на этом изображении объекты (пиксели, области изображения).

В мировой литературе в связи с ростом вычислительных мощностей и общего интереса к области искусственного интеллекта в последние 30 лет возрастал интерес к моделированию когнитивного восприятия в целом и к механизму поиска визуальной салиентности в частности. Так, в обзорной работе [5] Borji et Itti приводят подход к классификации существующих моделей по 13 признакам и описывают результаты классификации для 63 моделей. Кроме того,

существует и классификация задач, к которым применимы те или иные модели, и эти задачи можно условно разделить на две основных области: задачи обнаружения объекта – в том числе и его распознавание, – и задачи определения точки фиксации взгляда (eye fixation problem) – точки, на которой с наибольшей вероятностью человек остановит свой взгляд. Подавляющее большинство существующих моделей используются для решения задач только одной из этих областей.

Однако некоторые комплексные прикладные задачи требуют двойной трактовки. Например, к таким задачам можно отнести задачу автономного поиска малых возгораний в лесных массивах, разрабатываемую совместно лабораториями LISSI (Университет Пари-Эст, Париж, Франция) и SPE (Корсиканский Университет, Корте, Франция) при участии лаборатории искусственных нейронный сетей (Брестский государственный технический университет, Брест, Беларусь). Подобная задача включает в себя подзадачу автономного выбора направления поиска и подзадачу определения задымленности и/или возгорания.

В данной работе авторы предлагают комбинированный подход, применимый к обеим подзадачам, на основе поиска визуальной салиентности, сравнивают его с существующими аналогами и приводят экспериментальные результаты.

Описание алгоритма

Алгоритм был получен комбинацией алгоритма, описанного Ramik et al. в [6, 7] и некоторых методик, описанных в [5, 8]. Алгоритм состоит из четырех этапов, обязательным среди которых является только первый: расчет карт(ы) визуальной салиентности; расчет карты вероятностей фиксации взгляда; сегментация изображения; выделение салиентного объекта.

В качестве входных данных для первого и третьего этапов берется изображение в цветовом пространстве RGB, в качестве выходных данных могут рассматриваться результаты каждого этапа: карта визуальной салиентности, карта вероятностей фиксации взгляда, сегментированное изображение или выделенный на изображении салиентный объект. В связи с тем, что основным этапом алгоритма является первый, ниже мы его рассмотрим более подробно.

Карта (map) – изображение в оттенках серого, представляющее распределение некоторой характеристики пикселей, где максимальное значение интенсивности $I(x) = 255$ соответствует максимальному значению характеристики пикселя x (здесь и далее $x \in \aleph^2$ представляет собой координаты пикселя в 2D-пространстве изображения) входного изображения, а минимальное значение $I(x) = 0$ – минимальному (нулевому). Так, для карты салиентности интенсивность каждого пикселя показывает относительную салиентность соответствующего пикселя входного изображения, а для карты вероятностей фиксации взгляда – относительную вероятность остановки взгляда на соответствующем пикселе исходного изображения.

Пусть нам дано изображение размерностью $w \times h$ пикселей. Обозначим через $\Omega_R(x), \Omega_G(x)$ и $\Omega_B(x)$ значения R, G и B-каналов интенсивности каждого пикселя x. Кроме того, введем сферическую RGB-модель (SiRGB), получаемую из обычной RGB-модели путем преобразования, аналогичного преобразованию декартовой трехмерной системы координат в сферическую. Эти характеристики – азимутальный угол φ, зенитный угол θ и интенсивность l, представляющая расстояние до начала координат соответственно. Тогда можно обозначить $\Omega_l(x), \Omega_\phi(x)$ и $\Omega_\theta(x)$ как значения соответствующих величин для пикселя x.

Расчет карты визуальной салиентности

Для получения карты визуальной салиентности необходимо предварительно получить карту глобальной салиентности и карты локальной салиентности. Под глобальной салиентностью в данном случае понимается салиентность относительно всего изображения, а под локальной – относительно небольшого окружающего участка изображения.

Для расчета карты глобальной салиентности используются следующие формулы:

$$M_l(x) = \left\| \Omega_{\mu l} - \Omega_l(x) \right\|,\ M_{\phi\theta}(x) = \sqrt{\left(\Omega_{\mu\phi} - \Omega_\phi(x)\right)^2 + \left(\Omega_{\mu\theta} - \Omega_\theta(x)\right)^2},$$

$$M(x) = \frac{1}{1 - e^{-10(C_c - 0.5)}} M_{\phi\theta}(x) + \left(1 - \frac{1}{1 - e^{-10(C_c - 0.5)}}\right) M_l(x),$$

где $M_l(x)$ и $M_{\phi\theta}(x)$ – компоненты карты, вычисляемые по интенсивности и по цветности соответственно; $\Omega_{\mu l}$, $\Omega_{\mu\theta}$, $\Omega_{\mu\phi}$ – средние по изображению значения величин $\Omega_l(x), \Omega_\phi(x)$ и $\Omega_\theta(x)$, а коэффициент C_c – «показатель насыщенности», вычисляемый как средняя по изображению нормализованная разность между максимальным и минимальным значениями RGB-модели:

$$C(x) = \max\left(\Omega_R(x), \Omega_G(x), \Omega_B(x)\right) - \min\left(\Omega_R(x), \Omega_G(x), \Omega_B(x)\right),$$

$$C_c = \frac{\sum_x^{w \cdot h} C(x)}{255 \cdot w \cdot h}.$$

Коэффициенты при компонентах $M_l(x)$ и $M_{\phi\theta}(x)$ отвечают за учет цветности и насыщенности всего изображения в качестве весовых коэффициентов при соответствующих значениях компонент. $M(x)$ – финальная карта глобальной салиентности пикселей.

Расчет карты локальной салиентности производится исходя из идеи о центрально-периферическом антагонизме (center-surround antagonism, см., напр., [9]), которая в рамках КЗ может быть представлена как разница гистограмм SiRGB-значений пикселей внутри некоей области и пикселей снаружи этой области. Так, обозначим некоторый участок изображения размером (w_p, h_p) с пикселем x в центре как $P(x)$, а окружающий его участок площадью $2w_p h_p$ – как $Q(x)$. Размер участка $P(x)$ связан с размером изображения через коэффициент WSC (Window size coefficient), значение которого будет рассмотрено далее: $(w_p, h_p) = WSC \times (w, h)$.

Для участков $P(x)$ и $Q(x)$ определим гистограммы H_P и H_Q как гистограммы распределения значений (0…255) по пикселям участков $P(x)$ и $Q(x)$ для каждого канала SiRGB-модели. Тогда мы можем рассчитать для каждого пикселя x три величины $d(x)$ – по каждому каналу SiRGB-модели:

$$d(x) = \sum_{i=0}^{255} \left(\frac{H_P(i)}{H_P} - \frac{H_Q(i)}{H_Q} \right),\ C_{\mu P}(x) = \frac{\sum_{k \in P(x)}^{w_p h_p} C(k)}{w_p h_p},$$

$$D(x) = \frac{1}{1 - e^{-C_{\mu P}(x)}} d_l(x) + \left(1 - \frac{1}{1 - e^{-C(x)}}\right) \max\left(d_\phi(x), d_\theta(x)\right),$$

где $D(x)$ – карта локальной салиентности, собираемая из компонент аналогично карте глобальной салиентности.

Очевидно, что основным настраиваемым параметром на данном этапе алгоритма является коэффициент размера окна WSC, значение которого согласно [7] для задач обнаружения объектов оптимально в границах (0,35…0,45) при небольших (до 800×600) размерах входных изображений. Карта визуальной салиентности получается из карт локальной и глобальной салиентностей по принципу локального приоритета:

$$M_{final}(x) = \begin{cases} D(x), \text{ if } D(x) > M(x) \\ \sqrt{D(x)M(x)}, \text{ if } D(x) \le M(x) \end{cases}.$$

Расчет карты вероятностей фиксации взгляда

В качестве входных данных рассматриваются карты глобальной и локальной салиентностей, рассчитанные с разными значениями *WSC*. Несколько карт локальной салиентности суммируются между собой и с картой глобальной салиентности:

$$E(x) = \sum_{i=1}^{N} k_{di} \cdot D_i(x) + k_m \cdot M(x),$$

где весовые коэффициенты k_{di}, k_m и количество локальных карт N – настраиваемые параметры. Затем, согласно принципу центральной фиксации, на карту накладывается 2D-массив вероятности, распределенной по закону Гаусса, и, применяя к результату пороговую функцию, получаем:

$$E'(x) = w_g \cdot A \cdot \exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2} \right) + w_{ng} \cdot E(x),$$

$$EF(x) = \begin{cases} E'(x), & E'(x) \geq FT \\ 0, & \text{иначе} \end{cases},$$

где амплитуда гауссианы A, среднеквадратичные отклонения σ_x, σ_y, весовые коэффициенты w_g, w_{ng} и значение финального порога FT – настраиваемые параметры. Здесь $EF(x)$ – карта вероятностей фиксации взгляда. Такое большое количество настраиваемых параметров вызвано основной проблемой задач прогноза фиксации взгляда – реальный человеческий взгляд зависит от большого количества факторов [10], как, например, свободный ли это взгляд или он ищет конкретный объект (т.е., восходящий – bottom-up – или нисходящий – top-down – процесс внимания [5]); количество времени, отведенное на осмотр; существование априорных установок и т.д. Для устранения подобной проблемы существуют протоколы, на основании которых создаются наборы данных и тесты производительности алгоритмов прогноза фиксации взгляда (напр., MIT1003 [11], Toronto [12]). В зависимости от прикладной задачи выбираются те или иные тесты, на которых и настраиваются параметры.

Сегментация изображения и выделение объекта

Сегментация производит разделение всех пикселей на связные группы, пытаясь разбить единое изображение на изображения объектов, на основании метрики $d(x,y)$ для пикселей x, y:

$$d(x,y) = \bar{\alpha}(x,y) \cdot \left|\Omega_l(x) - \Omega_l(y)\right| + \alpha(x,y) \cdot \sqrt{\left(\Omega_\theta(x) - \Omega_\theta(y)\right)^2 + \left(\Omega_\phi(x) - \Omega_\phi(y)\right)^2},$$

где $\alpha(p,q)$ и $\bar{\alpha}(p,q)$ – комплементарные функции активации. Функция активации введена для учета насыщенности цветом всего изображения и является сигмоидальной [7].

На основе принципа 2-связности (пиксель может принадлежать группе, если минимум 2 ближайших соседних пикселя из 4 принадлежат группе) через сравнение метрик $d(x,y)$ (значение метрики для каждой пары пикселей внутри группы меньше, чем для каждой пары «пиксель из группы»-«пиксель не из группы») пиксели разбиваются на группы. Более детальное описание алгоритма на этапе сегментации см. в [7].

Выделение объекта производится следующим образом: если для пикселей одной группы среднее значение показателя салиентности выше порогового, а дисперсия этого показателя – ниже пороговой, то группа пикселей помечается как часть объекта. После рассмотрения всех групп при существовании хотя бы одной помеченной группы непомеченные удаляются; иначе пороговые значения понижаются и поиск производится заново. В [7] начальные пороговые значения определены как 128 и 20 соответственно.

Реализация и экспериментальный этап

Настройка параметров из этапа фиксации взгляда производилась с помощью стандартной эволюционной оптимизации, где в качестве генома каждой сущности

использовались значения параметров, а в качестве оценки пригодности – сумма трех оценок AUC_{judd} [13], AUC_{borji} [8] и $(1\text{-}KL_{div})$ [14], где каждая оценка вычисляется как среднее значение соответствующей метрики для каждого изображения теста производительности MIT1003. В связи с этим этапом было предложено именовать алгоритм EOA (evolutionary optimized algorithm).

После этого было проведено сравнение алгоритма с настроенными параметрами и трех лучших, согласно разным версиям, алгоритмов общего прогноза фиксации взгляда на двух наборах данных, MIT1003 и Toronto, используя те же метрики плюс дополнительную метрику «потраченное время», показывающую количество времени, потраченного для получения карт вероятности фиксации взгляда для 1003 изображений на одинаковой инфраструктуре (CPU Intel i7 3.3 GHz, RAM 16GB, реализация Python/Matlab).

Значения метрик для различных state-of-art алгоритмов в сравнении с EOA-алгоритмом

Алгоритм	AUC_{judd}, MIT1003	AUC_{Borji}, MIT1003	$1\text{-}KL_{div}$, MIT1003	AUC_{judd}, Toronto	AUC_{Borji}, Toronto	$1\text{-}KL_{div}$, Toronto	Потраченное время, сек
eDN [15]	**0,8651**	**0,7685**	0,3319	**0,8541**	0,6291	0,5192	3041
BMS [16]	0,7652	0,6103	0,3739	0,7461	0,5368	0,5799	**65**
RARE2012 [17]	0,7706	0,6171	**0,3948**	0,7688	0,5381	**0,609**	301
EOA	0,8422	0,7445	0,3611	0,8372	**0,6375**	0,5496	331

Как видно из таблицы выше, показатель AUC для алгоритма EOA в среднем меньше на 2 %, чем для алгоритма eDN, в то время как показатель 1-KLdiv – наоборот, несколько выше, при десятикратно меньших временных затратах. В сравнении же с двумя другими алгоритмами EOA показывает более высокое значение AUC и сопоставимое значение 1-KLdiv. Таким образом, невозможно однозначно определить, какой из алгоритмов лучше. Однако, соотношения значений показателей и количества затраченного времени, вкупе с фактором возможности использования алгоритма EOA для выделения объектов, позволяют считать представленный алгоритм эффективным и способным к работе в реальном времени.

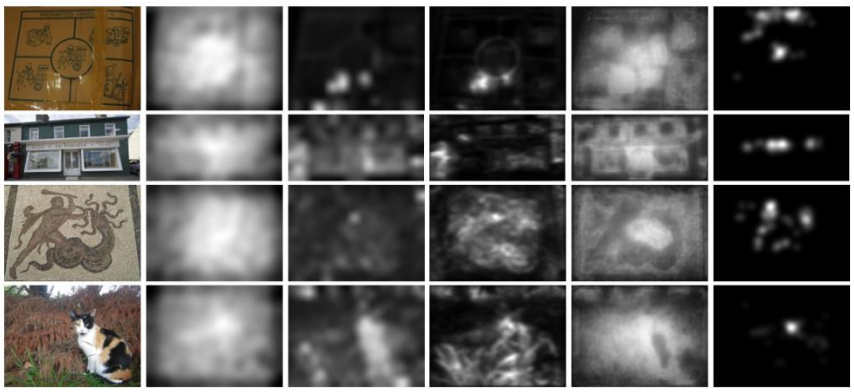

Рис. 1. Расчет карт вероятности фиксации взгляда разными алгоритмами на примере изображений из набора MIT1003 (слева направо): исходное изображение, eDN, BMS, RARE2012, EOA и данные, полученные в результате фиксации взглядов реальных людей (ground truth, эталонные данные)

Для проверки эффективности алгоритма при работе над обеими подзадачами сразу использовалась вышеупомянутая задача обнаружения задымления и малого возгорания, реализованная на платформе NEXTER Robotics Wifibot-M – 6-колесном мобильном роботе, оснащенном камерой WONWOO WCM-101, имеющей 3 степени свободы (Pan-Tilt-Zoom).

Протокол эксперимента: в экспериментальной зоне, представляющей собой газонную площадку размером 25×25 м, находятся: робот, малое возгорание, производящее небольшое количество дыма, а также набор «отвлекающих» предметов – листья, столбы, люди, деревья. Получая с камеры изображения, робот ориентируется на точку с наибольшей вероятностью фиксации взгляда, и с равной вероятностью либо производит малое движение в ту сторону, либо поворачивает камеру. В случае, если самая вероятная точка фиксации взгляда уже в центре, выделяется объект и запоминается его средний показатель салиентности, после чего

делается случайный поворот тела робота на 30–45° в и малое движение. Эксперимент проводится одну минуту и повторяется 30 раз.

В 26 случаях из 30 наиболее частым извлеченным объектом было задымление (в среднем 6 раз против 3 раз для отвлекающего объекта). Также в 26 случаях из 30 оно имело наибольший средний показатель салиентности среди всех выделенных объектов (в среднем 200 против 180). В оставшихся случаях наиболее салиентным оказывался кленовый лист. Остальные отвлекающие объекты были найдены не более 2 раз для каждой итерации эксперимента. Такие результаты показывают, что для каждой прикладной задачи автономного обнаружения использование вышеописанного алгоритма дает хорошие результаты при должной настройке параметров под соответствующую задачу.

Рис. 2. Примеры извлечения объектов из кадров видеопотока робота Wifibot-M во время эксперимента с огнем: человеческие ноги, огонь, задымление

Заключение

Представлен комбинированный алгоритм выделения объекта на изображении и расчета карты вероятности фиксации взгляда, который можно использовать для прикладных задач автономного обнаружения. Экспериментальные результаты показывают жизнеспособность алгоритма и его эффективность в сравнении с несколькими state-of-art алгоритмами, предполагая его применимость в том числе и в более широком классе задач – прикладных вариациях задачи прогноза фиксации взгляда (задачи композиции и дизайна, расположение рекламных элементов на Web-страницах и т.д.). Комбинирование данного подхода с существующими алгоритмами визуальной навигации может привести к повышению эффективности в задачах автономного обнаружения. Дальнейшая работа предполагает улучшение качества карты вероятности фиксации зрения (уменьшение количества ложно-позитивных ошибок) и добавление этапа распознавания извлеченных объектов на основании существующих современных методик.

OBJECT DETECTION IN COMPUTER VISION SYSTEMS: A VISUAL SALIENCY BASED APPROACH

V.A. KACHURKA, K. MADANI, C. SABOURIN, V.A. GOLOVKO, P.A. KACHURKA

Abstract

A combined approach of object detection in image and eye fixation probability map calculation is proposed. This approach can be used in applied tasks of autonomous object detection. Experimental results show viability and efficiency of this approach as compared with state-of-art algorithms, and predict its usability on the broader class of tasks – applied variations of eye fixation problem.

Список литературы

1. *Рахилина Е. В.* // МГУ. Семиотика и информатика. 1998. № 36. С. 274–323.
2. *Ахмадеева И.Р.* // Вестник НГУ. Серия Информационные технологии. 2014. № 3. С. 5–15.
3. *Андрианов А. И.* // МФТИ. Физико-математические науки. 2013. № 3. С. 47–50.
4. *Малахов К.А.* // Известия СПбГЭТУ ЛЭТИ. 2010. № 8. С. 7–11.
5. *Borji A., Itti L.* // IEEE Transactions on Pattern Analysis & Machine Intelligence. 2013. № 35–1. P. 185–207
6. *Ramík D.M., Sabourin C., Madani K.* // Proc. signal-image technology and internet-based systems. Dijon, 28 November–3 December. P. 438–446.
7. *Ramík D.M.* Contribution to complex visual information processing and autonomous knowledge extraction: application to autonomous robotics : Ph.D. dissertation. Paris, 2012.
8. *Borji A., Tavakoli H.R., Sihite D.N. et al.* // Proc. IEEE Computer Vision and Pattern Recognition. Sydney, 1–8 Dectmber 2013. P. 921–928.
9. *Liu T., Sun J., Zheng N.N. et al.* // Proc. IEEE Computer Vision and Pattern Recognition. Minneapolis, 18–23 June, 2007. P. 1605–1613.
10. Visual salience [Electronic resource] / ed. L. Itti. – Scholarpedia, 2007, rev. 2(9):3327. – Mode of access : http://www.scholarpedia.org/article/Visual_salience. – Date of access : 15.03.2015.
11. A Benchmark of Computational Models of Saliency to Predict Human Fixations [Electronic resource] / ed. T. Judd, F. Durand and A. Torralba. – MIT Technical Report, 2012. – Mode of access : http://saliency.mit.edu/. – Date of access : 15.03.2015.
12. *Bruce N., Tsotsos J.* // J. Vision. 2007. Vol. 7, №9. P. 950–957.
13. *Judd T., Ehinger K., Durand F. et al.* // Proc. IEEE International Conference on Computer Vision. Kyoto, 27 Sep.–4 Oct., 2009. P. 2106–2113.
14. *Contreras-Reyes J.E., Arellano-Valle R.B.* // Entropy. 2012. Vol. 14, № 9. P. 1606–1626.
15. *Vig E., Dorr M., Cox D.* // Proc. IEEE Computer Vision and Pattern Recognition. Columbus, 23–28 June, 2014. P. 2798–2805.
16. *Zhang J., Sclaroff S.* Saliency Detection Proc. IEEE Computer Vision and Pattern Recognition. Sydney, 1–8 Dec. 2013. P. 153–160.
17. *Riche N., Mancas M., Duvinage M. et al.* RARE2012 // Signal Processing: Image Communication. 2013. Vol. 28, № 6. P. 642–658.