ПСИХОФИЗИОЛОГИЯ

Научный обзор
УДК: 159.9
https://doi.org/10.21702/rpj.2025.3.11

Зрительная салиентность: от теоретических предпосылок к современным высокопроизводительным моделям

Денис В. Явна[1]

[1]Южный федеральный университет, Ростов-на-Дону, Россиийская Федерация

yavna@fortran.su

Аннотация

Введение. Зрительная салиентность – термин, обозначающий перцептивное качество фрагмента зрительной сцены, субъективно проявляющееся в его привлекательности для наблюдателя, а объективно описываемое вероятностью переключения на него фокуса внимания и/или глазодвигательной фиксации на нём. Это качество первоначально возникает благодаря работе механизма интеграции карт зрительных признаков и модулируется рядом центральных механизмов. Важно различать термины «салиентность» и «заметность» – в теоретическом контексте это не одно и то же. **Теоретическое обоснование.** Впервые в формате обзора вместе с результатами компьютерного моделирования зрительной салиентости подробно представлены теоретические предпосылки создания таких моделей. Подробно рассматривается теория интеграции признаков А. М. Трейсман, её достоинства и ограничения, благодаря которым возникла трёхуровневая модель зрительного внимания К. Коха и Ш. Улльмана. Согласно ей, управление переключениями фокальным вниманием осуществляется специальным механизмом («победитель получает всё») на основании данных, хранящихся в карте салиентности, кодирующей степень привлекательности каждого фрагмента зрительной сцены. Механизм формирования карты салиентности не был описан создателями теории и является предметом исследований, которые проводятся методом компьютерного моделирования. **Обсуждение результатов.** Рассматриваются результаты работ по моделированию зрительной салиентности. Подробно описывается ранняя компьютерная модель Л. Итти, К. Коха и Э. Нейбура, заложившая основы множества последующих разработок. Раскрываются особенности подходов к моделированию, возникших

до появления высокопроизводительных нейросетевых моделей. Описывается ряд современных высокопроизводительных моделей, основанных на технологиях сетей глубокого обучения, перечисляются их характерные особенности. Обзор моделей салиентности на русском языке делается впервые. **Заключение.** К настоящему времени созданы модели, имеющие практическую ценность. Обсуждаются возможности практического использования моделей зрительной салиентности и возможные перспективы их применения в психологических исследованиях.

Ключевые слова

зрительная система, внимание, движения глаз, зрительный поиск, салиентность, айтрекинг, компьютерное зрение, моделирование

Финансирование

Исследование выполнено при финансовой поддержке Российского научного фонда (РНФ) в рамках научного проекта № 25-18-00377

Для цитирования

Явна, Д. В. (2025). Зрительная салиентность: от теоретических предпосылок к современным высокопроизводительным моделям. Российский психологический журнал, том(номер), 190–225. https://doi.org/10.21702/rpj.2025.3.11

Введение

Почему мы воспринимаем окружающий нас видимый мир таким образом, что часто замечаем мелкие детали обстановки, но порой в упор не видим вещь, которую уже давно и безуспешно ищем? И почему мы можем периодически обращать или не обращать внимание на один и тот же объект в разное время и при различных обстоятельствах? Дадим пока формальный ответ: обычно мы обращаем внимание на те объекты, которые наделены качеством *салиентности*. Можно было бы также ответить, что мы замечаем *заметные* объекты…Такие ответы выглядят немного странно и могут быть восприняты как основанные на первобытной пралогике; однако следует учесть, что под заметностью и салиентностью здесь понимаются достаточно хорошо формализованные конструкты, наполненные специальным содержанием и используемые в ряде направлений исследований зрительного восприятия. Более того, эти конструкты не являются умозрительными и обязаны своим появлением и содержательным наполнением в первую очередь *экспериментальным* работам когнитивных психологов рубежа 70 – 80 годов прошлого века. Важно также

отметить, что вне научного контекста (используется и в филологических науках) слово «салиентность» в русском языке практически не применяется; обычно однокоренные с ним слова западных языков, происходящие от лат. salio – прыжок, скачок, – переводятся как заметность, значимость, выраженность и т д.; однако как специальные термины эти слова имеют разный смысл.

Термин «зрительная салиентность» сравнительно недавно вошёл в русский язык (Кочурко и др., 2015; Мартынова & Балаев, 2015), употребляется довольно узким кругом специалистов и поэтому нуждается в пояснении. Под зрительной салиентностью (англ. visual saliency) в общем случае понимают свойство некоторой области изображения, характеризующее её «способность» притягивать внимание наблюдателя. Однако из такого понимания не следует, что салиентность является свойством, присущим исключительно объекту наблюдения. У салиентности есть и субъективная сторона.

Различают восходящую (bottom-up) и нисходящую (top-down) салиентность. Восходящая салиентность определяется прежде всего физическими свойствами фрагмента зрительной сцены и обрабатывается стимул-управляемыми (stimulus-driven) механизмами непроизвольного внимания. Например, красная вертикальная линия среди множества синих линий будет характеризоваться высокой степенью восходящей салиентности (рис. 1а) (Строго говоря, данный пример представляет крайний случай, когда обнаружение может быть теоретически объяснено в терминах карт признаков и заметности, без использования понятийного аппарата моделей салиентности; однако он хорошо иллюстрирует феноменальную сторону обсуждаемого вопроса.). Восприятие таких стимулов часто сопровождается эффектом выскакивания (pop out, примерно соответствует «бросаться в глаза»), объективно выражающимся в отсутствии временны***х*** затрат на поиск, а субъективно – в лёгкости и непроизвольности обнаружения. Важно отметить, что ранние исследования рассматривали преимущественно восходящую салиентность, и термин «нисходящая салиентность» в прошлом мог бы звучать странно**.**

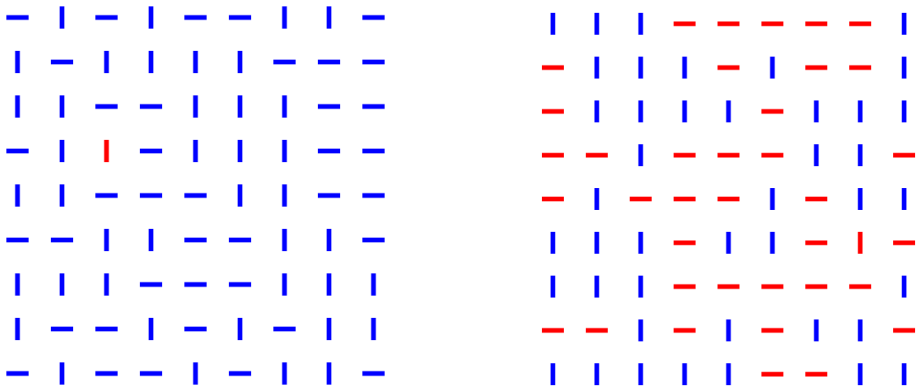
Нисходящая салиентность фрагмента сцены определяется в первую очередь перцептивной задачей, стоящей перед наблюдателем. Такая салиентность «присваивается» определённым объектам, признакам или их сочетаниям самим субъектом и адресуется в первую очередь механизмам произвольного внимания, осуществляющим поиск цели (goal-directed attention). Так, красная вертикальная линия среди красных горизонтальных и синих вертикальных (рис. 2б) будет характеризоваться довольно низкой восходящей салиентностью, но если в эксперименте она будет произвольно назначена целью, её нисходящая салиентность повысится, и в ходе последовательного зрительного поиска эта линия рано или поздно станет объектом внимания. В классических экспериментах с регистрацией движений глаз влияние задачи на управление вниманием было продемонстрировано А. Л. Ярбусом (Ярбус, 1965). Так, анализируя траектории осмотра различных изображений, записанных как при свободном просмотре, так

и в условиях, определявшихся достаточно специфичными инструкциями, Ярбус приходит к общему выводу, что «распределение точек фиксации на объекте, последовательность, в какой взор наблюдателя переходит от одной точки фиксации к другой, продолжительность фиксаций, своеобразная цикличность в рассматривании и т. д. определяются содержанием объекта и задачами, которые стоят в момент восприятия перед наблюдателем» (Ярбус, 1965, с. 148). По мнению исследователя, значительное влияние может оказывать также профессиональный опыт наблюдателя, его культурный уровень (значительную часть предъявляемых изображений составили произведения русской живописной классики). В ходе анализа траекторий осмотра он также неоднократно отмечает, что движения глаз отражают и процесс мышления. Отметим, что Ярбус вполне разделяет переключения внимания и движения глаз; и те, и другие могут быть как произвольными, так и непроизвольными. Именно смены фокуса внимания, а не точек фиксации, остаются в нашей памяти.

Таким образом, салиентность того или иного участка изображения может меняться в зависимости от перцептивной задачи, стоящей перед субъектом. Безусловно, определённую роль в формировании салиентности играют и особенности внимания самого субъекта, которые вносят дополнительный «шум» при обучении компьютерных моделей.

Рисунок 1

Пример массивов линий, в которых красная вертикальная линия отыскивается параллельно (а) или последовательно (б).



1а 1б

Моделирование зрительной салиентности, будучи непосредственно связанным с такими фундаментальными психологическими проблемами, как соотношение фокуса внимания и движений глаз, имеет большую предысторию. Если И. М. Сеченов ещё прямо отождествлял зрительное внимание со «сведением зрительных

осей глаз на рассматриваемое тело» (Сеченов, 1942 с. 80, переиздание текста 1866 г.), то Г. фон Гельмгольц (Helmholtz, 1896) показал существование механизма пространственного перемещения внимания, не зависящего от движений глаз (цит. по Рожкова и др., 2019). В настоящее время внимание, связанное с перемещением взора, принято называть явным (overt) в противоположность скрытому (covert), открытому Гельмгольцем (Подладчикова и др., 2017; Рожкова и др., 2019). Представления об этих видах внимания в когнитивной психологии были существенно развиты М. Познером (напр. Posner, 1980), позже предложившим трёхкомпонентную модель внимания (Posner & Petersen, 1990). Эта модель в значительной степени основывается на нейрофизиологических данных и описывает три подсистемы внимания: возбуждения-бдительности (alerting), собственно ориентировки (orienting) и экзекутивного контроля (executive control) (пер. по Величковскому, 2006). Мозговые механизмы и связи скрытой и явной ориентировки внимания до конца не ясны и являются актуальным предметом нейрофизиологических исследований (Petersen & Posner, 2012). Трудности объективной регистрации перемещений скрытого внимания с одной стороны и существенный прогресс в области методов окулографии с другой привели к тому, что в настоящее время при проверке моделей зрительной салиентности учитываются прежде всего акты явного внимания, причём объективным маркером этих актов выступают глазодвигательные фиксации; считается, что именно во время фиксаций мозг считывает основной объём информации, необходимой для решения перцептивных задач (Rayner, 2009). Тем не менее, первые работы по моделированию салиентности рассматривали в первую очередь скрытое внимание. Кажущееся противоречие может быть объяснено тем, что и скрытое, и явное внимание действует в пределах одной и той же карты салиентности, т. е. посещают примерно одни и те же локации, хотя как время существования фокуса, так и последовательность переключений для этих видов внимания могут отличаться; так, «переключения пространственного внимания обычно (но не обязательно) сопровождаются движениями глаз» (Theeuwes, 2013, p. 1). Движения глаз «часто рассматривают как средство (proxy) переключения внимания» (Borji & Itti, 2013, p. 186).

Теоретическое обоснование

Определяющее влияние на формирование представлений о механизмах зрительной салиентности оказала теория интеграции признаков А. Трейсман и Г. Джелэйда. На основании результатов ряда ранних работ авторы выдвигают положения, в «крайней форме» (Treisman & Gelade, 1980, p. 99) представляющие их теорию. Признавая представления о гештальте соответствующими нормальному субъективному опыту восприятия (его внутренней, субъективной картине), авторы тем не менее не считают их полезными для исследования ранних стадий обработки информации; как раз признаки (свойства, features) должны стоять в восприятии на первом месте.

Утверждается, что зрительная сцена первоначально кодируется по ряду отдельных признаков, таких как цвет, ориентация, пространственная частота, яркость, направление движения. Чтобы осуществить их правильный синтез для каждого объекта, содержащегося в сложном изображении, фокальное внимание должно последовательно обработать соответствующие местоположения (локации); оно выступает в качестве «клея» (Treisman & Gelade, 1980, p. 98), соединяя изначально разделённые признаки в единый объект. После того, как этот составной объект правильно воспринят, он сохраняется в памяти и в будущем может восприниматься уже как таковой. Однако при определённых обстоятельствах (ухудшение памяти и др.) склеенные признаки могут распадаться и снова «свободно плавать» (“float free”) или, возможно, рекомбинировать, образуя «иллюзорные сочетания» (“illusory conjunctions”) (Treisman & Gelade, 1980, p. 98). Если признаки свободно плавают вне фокуса внимания, то для отдельных признаков локализация и идентификация могут протекать как независимые процессы; однако для идентификации сочетаний признаков сначала нужно их локализовать, чтобы направить внимание к выявленной локации и обеспечить возможность их интеграции. Стимулы вне фокуса внимания могут влиять на выполнение задачи только на уровне содержащихся в них признаков, но не сочетаний.

Предсказания, сделанные авторами относительно ряда характеристик процесса восприятия на основании представленных положений, проверялись в девяти экспериментах; их результаты и соответствующие теоретические обобщения были опубликованы в 1980 г. (Treisman & Gelade, 1980). Несмотря на то, что впоследствии теория получила существенное развитие (так, были сформулированы представления о «файле» (Трейсман, 1987) или «досье» (по Фаликман, 2001)) распознаваемого объекта), именно эта работа оказала важнейшее влияние на развитие моделей салиентности.

Первая группа предсказаний заключалась в том, что если базовые признаки (features) могут быть обнаружены параллельно, без ограничений со стороны внимания, то на поиск целей, определяемых такими признаками (например, красным цветом или вертикальной ориентацией), изменения количества одновременно предъявляемых дистракторов должны влиять незначительно. Напротив, если для обнаружения целей, которые определяются совокупностью или сочетанием (conjunction) признаков (например, красная вертикальная линия среди красных горизонтальных и синих вертикальных, рис. 1б), необходимо фокальное внимание, такие цели возможно обнаруживать только после последовательного сканирования множества предъявленных элементов.

Вторая группа предсказаний касалась разделения текстур и фигуро-фонового группирования: если это параллельные преаттентивные процессы, то они должны определяться только пространственными разрывами между группами стимулов, различающихся по отдельным признакам, а не по их сочетаниям.

Третья группа предсказаний связана с возможностью иллюзорных сочетаний признаков, «свободно плавающих» вне фокуса внимания.

Четвёртая группа предсказаний касается вопроса о взаимосвязи идентификации и локализации признаков и их сочетаний. Если признаки вне фокуса внимания могут свободно плавать, причём наличие этих признаков можно установить, не определив их точное местоположение, то идентификация и локализация являются независимыми процессами. В случае поиска отдельного признака идентификация может предшествовать локализации; в случае поиска сочетаний локализация предшествует идентификации, так как внимание привлекается к определённому местоположению.

Пятая группа предсказаний связана с возможным влиянием объектов вне фокуса внимания на эффективность поиска: облегчать или затруднять его должны только признаки, но не их сочетания.

Проверка предсказаний, важнейшей частью которой были эксперименты со зрительным поиском, в основном подтвердила их правильность. Б. М. Величковский отмечает, что теория интеграции признаков «удивительно хорошо выдержала 20-летнюю экспериментальную проверку» (Величковский, 2006 с. 295), однако с объяснением достаточно плоских (10 – 20 мс на дистрактор) функций зависимости времени поиска от числа элементов у неё возникли затруднения. Напомним, что при последовательном поиске наклон составляет примерно 60 мс на элемент в отсутствие цели; если же цель присутствует, наклон сокращается примерно вдвое, и это может свидетельствовать о поисковой стратегии полного перебора: матожидание числа изученных вниманием элементов до её обнаружения как раз и равно половине числа элементов, если цель расположена в случайной позиции. Однако, по точному замечанию Величковского, малый наклон функций означал бы просмотр до 100 элементов в секунду, что не соответствует экспериментальным данным о переключениях скрытого внимания (напр. (Saarinen & Julesz, 1991)). Объяснение может быть предложено в рамках теории (модели) *управляемого* (более распространённый перевод исходного guided) или *ведомого* (по Величковскому) зрительного поиска, разработанная Дж. Вольфом с соавт. (Wolfe et al., 1989). Актуальное на момент написания настоящего обзора изложение теории (в её шестой версии) представлено в работе (Wolfe, 2021).

Подробное рассмотрение теории управляемого поиска не входит в задачи настоящего обзора, тем более что она достаточно известна в нашей стране и нередко выступает в качестве теоретической основы исследований, выполненных целым рядом отечественных авторов (напр. (Горбунова, 2023; Дренёва, 2020; Крускоп и др., 2023; Сапронов & Горбунова, 2025; Фаликман, 2015; Фаликман и др., 2019)). Однако представляется целесообразным дать её краткое изложение, чтобы показать общность задач, решаемых в рамках теорий управляемого поиска и салиентности, а также сходство их понятийного аппарата.

Когда мы смотрим на сцену, мы можем видеть что-либо в любой её локации, но не можем распознавать больше нескольких элементов одновременно; это своего рода «бутылочное горло» (bottleneck). Как и у Трейсман, локации выбираются вниманием, чтобы содержащиеся в них признаки могли бы быть склеены в распознаваемые объекты. Но чтобы порядок выбора был рациональным (intelligent), внимание, обеспечивающее доступ к «бутылочному горлу», управляется (guided) на основании пяти разных источников преаттентивной информации, а именно:

1. нисходящего (top-down);
2. восходящего (bottom-up), признакового;
3. предшествующей истории (например, благодаря прайммингу);
4. вознаграждения, подкрепления (reward);
5. синтаксиса и семантики сцены.

Эти источники управления формируют пространственную «карту приоритетов» (Serences & Yantis, 2006) – динамический ландшафт внимания, который развивается в ходе поиска. Избирательное внимание направляется к наиболее активной локации на карте приоритетов примерно 20 раз в секунду, т. е. каждые 50 мс. Наведение осуществляется неравномерно, предпочтение отдаётся элементам вблизи точки фиксации. Природа фовеальной «предвзятости» (bias) к локациям вблизи точки фиксации описывается тремя типами функциональных полей зрения (ФПЗ): разрешающей (resolution FVF), управляющей поисковыми (exploratory) движениями глаз и управляющей скрытым развёртыванием внимания. Несколько забегая вперёд, отметим, что в части описания способа переключения фокуса внимания теория управляемого поиска явным образом (Wolfe, 2021, p. 1068) опирается на представления К. Коха и Ш. Улльмана (Koch & Ullman, 1985) о механизме WTA, который будет подробно рассмотрен ниже.

Выбранный вниманием элемент помещается в рабочую память, которая также содержит *направляющий шаблон* и может задавать последующий маршрут. Например, при поиске банана внимание направляется к целевым атрибутам при помощи шаблонов «жёлтый» и «изогнутый» (Wolfe, 2021, p. 1064).

Чтобы быть идентифицированными как цели или отклонёнными как дистракторы, выбранные вниманием объекты должны сравниваться с *целевыми шаблонами,* хранящимися в *активированном* текущей задачей фрагменте долговременной памяти (ALTM). Сравнение помогает установить, что объект является не просто жёлтым и изогнутым, но действительно тем самым бананом, который нужно найти. Если руководящих шаблонов в рабочей памяти всего несколько, то целевых шаблонов может быть очень много; в качестве примера Вольф приводит т. н. *гибридный поиск* ((Wolfe, 2012), см. также (Ангельгардт и др., 2021; Сапронов и др., 2023 ; Rubtsova & Gorbunova, 2022)). Эти шаблоны могут быть как конкретными (спелый банан), так и гораздо более общими (фрукт).

Связывание и распознавание объекта внимания моделируется как процесс *диффузии* (Ratcliff, 1978; Воронин и др., 2020), осуществляющийся со скоростью

> 150 мс/элемент. Выбор может происходить и чаще, если несколько элементов подвергаются распознаванию одновременно, хотя и асинхронно; это делает управляемый поиск гибридом последовательных и параллельных процессов. Для каждого целевого шаблона, хранящегося в ALTM, существует один *диффузор* (канал диффузии), накапливающий данные (в т. ч. шум), приближающиеся к порогу выхода. Когда данные достигают порога, поиск прекращается и даётся либо истинный, либо ложноположительный ответ. Поиск может прекратиться и по достижении порога накопления *сигнала выхода*, что приведёт либо к истинно-, либо к ложноотрицательному ответу.

Установка порога накопленного сигнала выхода является адаптивной, что позволяет обратной связи о результатах предыдущих предъявлений программировать последующий поиск. Моделирование показывает, что комбинирование асинхронной диффузии с сигналом выхода позволяет воспроизвести основные паттерны времени реакции и ошибок, полученные в ряде экспериментов со зрительным поиском.

Таким образом, теория управляемого поиска явным образом описывает алгоритм аттентивного отбора, сильно сближаясь с теорией салиентности. Благодаря этому она успешно преодолевает ограничения теории интеграции признаков. Кроме того, она существенно расширяет последнюю в части описания алгоритмов принятия решения наблюдателем. Теория управляемого поиска развивается преимущественно в рамках теоретико-информационного подхода и традиционной экспериментально-психологической парадигмы когнитивных исследований. Теория салиентности находится на стыке когнитивных и технических наук и в основном описывает ранние этапы зрительной обработки, связанные с развёртыванием внимания; важной её частью является моделирование.

Теоретические основы математического и компьютерного моделирования салиентности были заложены более 40 лет назад работой К. Коха и Ш. Ullмана, где рассматриваются пространственные сдвиги внимания и их возможные нейрональные механизмы (Koch & Ullman, 1985). Нельзя не отметить, что термин «салиентность» использовался в психологии и ранее, но как более общее понятие, не отражающее специфику работы конкретной сенсорной системы. Так, ещё в 1977 году А. Тверски опубликовал значимую теоретическую работу, дающую формализацию понятия «сходство» (Tversky, 1977) в теоретико-множественных терминах. Передавая её содержание кратко, можно сказать, что каждый объект характеризуется множеством признаков, некоторые из которых являются общими с другими объектами, а некоторые – разграничительными, уникальными. Салиентность (скорее в значении «заметность, значимость») у Тверски является свойством *признака*; она зависит как от его физических характеристик (яркость и т. д.), так и от т. н. диагностических факторов – релевантности контексту и важности этого признака для решения конкретной задачи. Салиентность занимает важное место в теоретических построениях Тверски: так, более *салиентный объект* скорее станет референтым в человеческих суждениях о сходстве. Степень сходства объектов a и b может оцениваться по шкале S как:

$$S(a,b) = \phi f(A \cap B) - \alpha f(A - B) - \beta f(B - A),$$

$$\phi, \alpha, \beta \geq 0,$$

где ***A*** и ***B*** – множества свойств ***a*** и ***b*** соответственно, ***f*** – мера салиентности, которая, как и параметры ϕ, α и β зависит от контекста и решаемой задачи. Таким образом, *салиентность объекта* может быть определена в рамках проблематики оценки сходства объектов. Достаточно простую интерпретацию идей Тверски приводит в своей работе (Julesz, 1986) Б. Юлеш: салиентность может быть определена как функция (например, отношение) числа уникальных и общих признаков или же как функция числа уникальных признаков к их общему числу.

Понятие собственно зрительной салиентности вводилось Кохом и Ульманом (Koch & Ullman, 1985) как обозначение фундаментального звена организации зрительного внимания, объединяющего информацию из отдельных карт признаков в общую карту, содержащую меры «заметности» (conspicuity). Работа имела теоретический характер и во многом опиралась на положения, высказанные Трейсман и Джелэйдом (Treisman & Gelade, 1980), расширяя её в части объяснения алгоритма переключения фокального внимания. Рассмотрим эту статью более подробно, так как она оказала определяющее влияние на целое направление исследований внимания, оставаясь практически неизвестной в нашей стране.

Авторы начинают свою статью с доводов в пользу двухуровневой теории зрительного восприятия человека, предполагающей существование преаттентивного, на котором простые признаки обрабатываются быстро и параллельно по всему полю зрения, и аттентивного уровней. На втором уровне специализированный фокус обработки, т. е. фокус внимания, направлен на определённую локацию в поле зрения, причём анализ сложных форм и распознавание объектов связаны именно с этим уровнем. Если бы специфические алгоритмы, решающие задачи наподобие анализа формы или распознавания объекта в определённой локации, выполнялись параллельно, это привело бы к комбинаторному взрыву объёма требуемых вычислений и нехватке соответствующих ресурсов. Авторы ссылаются в частности и на критику возможностей перцептронов, представленную М. Мински и С. Пейпертом в хорошо известной книге (Мински & Пейперт, 1971), что представляет отдельный исторический интерес. Действительно, параллельная обработка в современных свёрточных сетях вряд ли могла бы служить метафорой ограниченных возможностей параллельной стадии обработки информации у человека; однако неглубокие полносвязные перцептроны тех лет вполне подходили на эту роль. В итоге авторы приходят к тому, что после определённой (параллельной) стадии предобработки анализ зрительной информации продолжается в последовательности операций, каждая из которых применяется к выбранной локации или локациям.

Приводя экспериментальные свидетельства существования селективного внимания, Кох и Ульман опираются как на «психофизические» (sic!), так

и на физиологические данные. Существование перемещающегося специализированного фокуса обработки, связанного с фовеальными проекциями, но не идентичного им, подтверждается двумя классами психофизических экспериментов. Во-первых, это исследования Трейсман с коллегами, в которых «поиск цели, заданной единичным признаком ..., оказывается параллельным ..., тогда как поиск конъюнктивной цели, определённой в терминах нескольких признаков ..., требует последовательного, произвольно прерываемого сканирования среди предъявленных дистракторов» (Koch & Ullman, 1985, p. 219). К этому же классу подтверждений относится и ряд исследований, посвящённых определению параллельно обнаруживаемых зрительных признаков. Так, в своих исследованиях различения текстур Юлеш с соавт. показали, что только ограниченный набор признаков-*текстонов* может обнаруживаться параллельно (Bergen & Julesz, 1983–29C.E.; Julesz, 1984). Во-вторых, это ряд ранних исследований, использующих парадигму пространственных подсказок (Bashinski & Bacharach, 1980; Eriksen & Hoffman, 1972; Posner, 1980; Remington & Pierce, 1984). В настоящее время существует несколько устоявшихся названий задач такого типа: Posner cueing task, spatial cueing, парадигма Познера, метод подсказки и др. (см. напр. (Гусев & Уточкин, 2012; Шевель & Фаликман, 2022)). Физиологические данные также свидетельствуют в пользу избирательной обработки зрительной информации. Излагая ряд исследований с регистрацией клеточной активности, авторы заключают, что «отдельные клетки в определённых частях зрительной системы по-разному отвечают на одинаковые физические стимулы, увеличивая свой ответ как функцию от решаемой зрительной задачи» (Koch & Ullman, 1985, p. 220).

В результате проведённого анализа авторы формулируют ряд принципиальных вопросов о механизмах селективной обработки. Их интересует, какие операции могут применяться к отобранным локациям, как осуществляется этот отбор и, в частности, как осуществляется смена локаций.

Переходя к теоретическим построениям, авторы в первую очередь вводят понятие *ранней репрезентации* – это набор топографических корковых карт, кодирующих зрительную информацию на уровне разных элементарных признаков, таких как ориентация границ, цвет, диспарантность и направление движения; каждая локация в таких картах имеет множественную признаковую размерность. Вероятно, в соответствии со свидетельствами в пользу существования пространственно-частотных каналов зрительной системы (напр. (Campbell & Robson, 1968; Wilson & Bergen, 1979)), для каждого отдельного признака могут иметься наборы карт разного разрешения. В картах присутствуют отношения соседства и локальные тормозные связи (латеральное торможение), благодаря которым локации, существенно отличающиеся от окружения, могут обнаруживаться уже на этом раннем уровне анализа. Таким образом, карты «сигнализируют» о *заметности* (conspicuity) участка зрительной сцены.

Здесь нужно сделать важную оговорку. Речь пока идёт именно о заметности,

а не о салиентности. Салиентность возникает на следующем этапе обработки как отдельный перцептивный механизм. Именно этим объясняется необходимость прямого переноса термина «салиентность» в русский язык; попытка его перевода может привести к путанице при именовании уровней обработки.

Когда внимание уделяется отдельной локации, присутствующие в ней признаки должны быть переданы на вышележащий более абстрактный и нетопографический уровень представления. Авторы отмечают, что такая постановка вопроса не противоречит идее иерархической обработки информации в коре; заметим также, что она хорошо согласуется с основными положениями теории интеграции признаков. Каким же образом осуществляется выбор локации для внимания? Как обрабатывается информация большой признаковой размерности, представленная в ранней репрезентации?

Авторы допускают, что заметность локации в зрительной сцене определяет уровень активности соответствующих элементов в различных картах признаков, при этом разные карты кодируют заметность внутри определённой признаковой размерности. Объединение всей этой разнородной информации осуществляется благодаря *карте салиентности*, представляющей собой единую глобальную оценку (measure) заметности, имеющую, как и карты признаков, топографическую структуру. Точную природу процесса объединения признаковых карт авторы не описывают, делая предположение, что она, всё ещё являясь частью ранней зрительной системы, «кодирует заметность объектов в терминах простых свойств, таких как цвет, направление движения, глубина и ориентация» (Koch & Ullman, 1985, p. 221). Именно эта неопределённость послужила точкой роста целого направления исследований в будущем. Отметим, что авторы допускали также возможность модулирующих влияний на карту салиентности со стороны вышележащих корковых центров; в будущем такие влияния начнут реализовываться в моделях нисходящей салиентности.

Центральное место в теоретических построениях Коха и Ульмана занимает основное звено аттентивного отбора, в явном виде отсутствавшее в теории интеграции признаков – сеть WTA («победитель получает все» (Feldman, 1982)), отвечающее за выбор локации для фокального внимания, свойства которой затем передаются в «центральную репрезентацию»; она работает с картой салиентности.

Механизм WTA может рассматриваться как эквивалент оператора поиска максимума, работающего над элементами карты салиентности x_i ; в нейронной сети x_i может интерпретироваться как электрическая активность элемента в локации i . WTA отображает множество входных элементов на эквивалентное множество выходов y_i по следующему правилу:

$$y_i = 0 \; if x_i < \max_j x_i$$
$$y_i = f(x_i) \; if x_i = \max_j x_j,$$

где f – любая возрастающая функция от x_i или константа. Таким образом, все выходные элементы кроме одного, соответствующего наиболее активному входному элементу, устанавливаются в 0.

Если не учитывать «аппаратные» особенности мозгового субстрата вычислений, построение сети WTA представляется достаточно простой задачей. Авторы рассматривают ряд возможных реализаций сети, как полностью последовательную, неприемлемую по причине крайне медленной работы, так и в высокой степени параллельные, характеризующиеся слишком большим числом связей между процессинговыми элементами и невозможностью обрабатывать произвольное число входов. Исходя из этого, авторы формулируют два биологически обоснованных допущения, строя на них возможные реализации WTA:

1. «За исключением некоторых дальних возбудительных связей, большинство их, как возбудительных, так и тормозных, является локальными» (Koch & Ullman, 1985, p. 222).
2. «Каждый элементарный процессинговый элемент выполняет только простые, вполне определённые операции, такие как сложение или умножение. В частности, базовые процессинговые элементы неспособны обрабатывать какую-либо символическую информацию, такую как адреса» (там же).

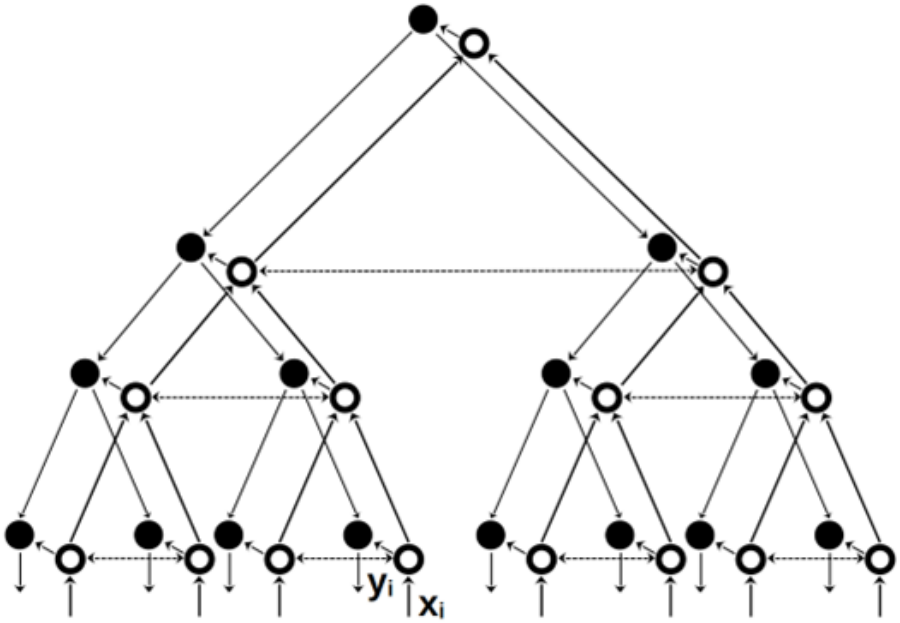
Всего таких реализаций две, причём авторы явным образом отдают предпочтение второй. Рассмотрим её более подробно.

Данная реализация WTA имеет иерархическую пирамидальную структуру и работает в высокопараллельном режиме. Сначала вычисляются максимумы среди m элементов из входного набора размером n. На следующем уровне иерархии процесс повторяется уже для n/m входных элементов; так происходит до тех пор, пока пирамида сравнений не сомкнётся на последнем элементе, в который отображается глобальный максимум. Однако для процесса селекции важно как абсолютное значение максимума, так и его локация. Она определяется с использованием второй пирамиды дополнительных элементов, в которой информация распространяется в обратном порядке. Каждый дополнительный элемент ассоциирован с элементом главной пирамиды и активируется только тогда, когда получает одновременное возбуждение от своего главного элемента и от дополнительного элемента, расположенного на вышележащем уровне. «Так как на каждом уровне наиболее активированный элемент главной пирамиды при локальном сравнении подавляет активность других $m-1$ главных элементов, ассоциированные дополнительные элементы так же, как и все дополнительные элементы в нижележащих ветвях, никогда не будут активированы» (Koch & Ullman,

1985, p. 223). На рис. 2 показан возможный пример реализации сети WTA с $n = 8$ входами и $m = 2$ сравниваемыми элементами. Число восходящих и нисходящих временны*ы*х вычислительных шагов для такой сети не должно превышать $2\log_m n$, сеть содержит не более чем $2nm/(m-1)$ элементов. Предполагается, что значения входов не должны быть в точности одинаковыми.

Рисунок 2

Вторая реализация сети WTA с $n = 8$ *входными элементами.*



Локальное сравнение производится между $m = 2$ элементами. Основные элементы показаны светлым, дополнительные – чёрным; x_i соответствует максимуму на входе в сеть, y_i – ответу сети на обнаруженный максимум. По (Koch & Ullman, 1985).

Авторы приводят оценки, в соответствии с которыми для реализации сети WTA в живой системе (приматы, кошки) достаточно лишь небольшой части имеющихся зрительных нейронов. Предположительно, на роль субстрата WTA хорошо подходят крупноклеточные системы, такие как Y-путь у кошек.

Как же происходит смена локаций, захваченных вниманием, по полю зрения? Здесь возможны два механизма, локальный и центральный, действующие через модификацию карты салиентности. Локальный механизм может реализовываться через адаптацию и ослабление активной локации в карте салиентности со временем; наиболее активный элемент локально тормозится, например, по истечении

определённого интервала времени. Центральный же задействует тормозящий сигнал из центральной репрезентации, куда ранее поступила информация. Между существованием этих механизмов нет противоречия, и они могут действовать одновременно; вероятно, локальный механизм постоянно включён в работу, а центральный задействуется, когда возникает побуждение к произвольному сдвигу внимания (Posner, 1980). Оба эти механизма осуществляют долгосрочное торможение выбранного элемента карты салиентности, предупреждающего на определённый временно́й период повторное посещение соответствующей локации – т. н. торможение возврата внимания (Posner et al., 1982; Уточкин & Фаликман, 2006).

Предложенные Кохом и Ульманом механизмы аттентивного отбора, основанные на карте салиентности и WTA, дают им возможность предложить свою интерпретацию эффектов параллельного и последовательного поиска, а также маскирования (camouflage) определённого объекта другими (Treisman, 1982). Если цель обладает салиентным признаком, отличающим её от соседей, WTA сразу определит её локацию, и цель будет обнаружена за время, не зависящее от числа дистракторов. Если же цель определяется сочетанием признаков, карта салиентности будет иметь множество локальных пиков, «в самом плохом случае даже столько, сколько объектов предъявляется» (Koch & Ullman, 1985, p. 224). Если не будет применена дополнительная оптимизирующая стратегия, WTA будет перебирать их; таким образом, для успешного завершения поиска потребуется просмотреть в среднем $n/2$ предъявленных объектов. Таким образом, объект «бросается в глаза» (pops out), потому что из-за салиентности он первый, который нужно посетить, а параллельный и последовательный поиск не являются принципиально разными процессами. Что касается маскирования, то тут возможны две различные стратегии: можно снизить заметность объекта, смешав его с окружением (примерно так работает военный камуфляж), а можно поместить его среди очень заметных объектов. В обоих случаях активность карты салиентности в точке, соответствующей целевому объекту, снизится *относительно* окружения.

В чём же заключается дополнительная оптимизирующая стратегия, позволяющая в значительном ряде случаев избежать необходимости полного перебора объектов в зрительной сцене? Авторы полагают, что такая стратегия может основываться на правилах приоритетов *близости* и *сходства*, примерно соответствующих феноменам перцептивного группирования и одноимённым принципам гештальта. Так, поиск цели вокруг выбранной локации выиграет, если предпочтения механизма селекции будут смещёны к соседним локациям. В качестве экспериментального подтверждения приоритета близости приводятся работы, демонстрирующие зависимость вероятности обнаружения цели от близости к локации, на которую направлено внимание (Engel, 1971, 1974). Поиск объектов с общим отличительным признаком улучшится, если локации со свойствами, сходными с представленными в текущей локации, станут предпочтительными. Это частично подтверждается

результатами, находившимися на момент написания авторами статьи в печати (Geiger & Lettvin, 1986): демонстрация фигуры в точке фиксации делает салиентной такую же фигуру, появляющуюся в другом месте поля зрения в том же предъявлении.

Самым простым способом реализации приоритета близости внутри механизма WTA является усиление всех элементов в карте салиентности, соседствующих с выбранным в данный момент. «Выход механизма WTA, ассоциированный с выбранной локацией, увеличивает заметность близлежащих элементов в карте салиентности на величину, зависящую от расстояния между данной локацией и окружением, тем самым упрощая сдвиг фокуса обработки к близлежащим локациям», что «эквивалентно утверждению о существовании аттрактивного потенциала вокруг каждой выбранной локации» (Koch & Ullman, 1985, p. 224).

Реализация приоритета сходства возможна следующим образом. Механизм WTA, срабатывая, запускает взаимодействия *внутри* отдельных карт признаков на уровне ранней репрезентации, благодаря которым в картах, содержащих выбранные в данный момент признаки, заметность их в окружении выбранной локации увеличивается. Такой процесс не предполагает взаимодействия *между* картами признаков, их точной топографической привязки друг к другу. Если будет выбран объект с красной горизонтальной линией, то соседние локации в картах признаков «красный» и «горизонталь» будут усилены; фокус внимания с б***о***льшей вероятностью сместится к ним. Процесс, обеспечивающий приоритет сходства, действует противоположно первоначальному приоритету заметных локаций, возникающему благодаря латеральному торможению внутри карт признаков; возможны различные варианты взаимодействия этих процессов.

Таковы в общих чертах основные теоретические положения, высказанные Кохом и Ульманом в 1985 году. Первые вычислительные модели салиентности появились значительно позже, в середине 90-х годов прошлого столетия (Baluja & Pomerleau, 1994; Itti et al., 1998; Milanese, 1993; Tsotsos et al., 1995); по мере своего совершенствования они начали обретать практическую значимость. Рассмотрим теперь основные результаты, полученные в рамках различных подходов к моделированию.

Обсуждение результатов

Традиционные модели салиентности

Подходы к моделированию салиентности можно условно разделить на традиционные и нейросетевые. Благодаря использованию современных нейросетевых архитектур, прежде всего свёрточных, в последние годы были побиты все рекорды качества обучения моделей (Borji, 2019). Не в последнюю очередь успеху нейросетевых моделей способствует увеличение объёма данных с результатами

окулографических исследований, находящихся в открытом доступе, и появление стандартизированных и относительно простых в использовании инструментов нейросетевого моделирования. Рассмотрим названные подходы более подробно, начав с традиционных и оказавших наибольшее влияние на последующее развитие направления.

Модель Лаурента Итти, Кристофа Коха и Эрнста Нейбура послужила основой для многих последующих моделей; она также выступает в качестве эталона при их сравнении (Borji & Itti, 2013). Модель осуществляет анализ интенсивности, цвета и ориентации. На первом этапе входное цветное (r, g, b) изображение 640x480 в каждом из соответствующих каналов представляется в виде гауссовой пирамиды (9 масштабов от 1:1 до 1:256 с шагом в октаву). Интенсивностное представление изображения $I = (r + g + b)/3$ используется для создания пирамиды $I(\sigma)$, где $\sigma \in [0..8]$ – масштаб. Оно же используется для нормализации первичных цветовых каналов r, g и b, применяемой для того, чтобы отделить цветовой оттенок от интенсивности. Так как изменения оттенка при низкой яркости не воспринимаются, нормализация применяется только там, где I больше $1/10$ своего максимума по всему изображению; в остальных локациях значения пикселей обнуляются.

Вычисление карт локальных характеристик осуществляется набором линейных центрально-периферических операторов, которые реализованы в модели как поточечная разность между тонким высокочастотным и грубым низкочастотным масштабными представлениями (обозначается $\ominus$): центр представляют пиксели в масштабе $c \in \{2,3,4\}$, а окружение – соответствующие пиксели в масштабе $s = c + d$, где $d \in \{3,4\}$. Шесть интенсивностных карт расчитываются как

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)|.$$

На основании первичных нормализованных цветовых каналов создаются четыре новых широкополосных:

- красный: $R = r - (g + b)/2$
- зелёный: $G = g - (r + b)/2$
- синий: $B = b - (r + g)/2$
- желтый: $Y = (r + g)/2 - |r - g|/2 - b$.

Отрицательные значения обнуляются. Из этих каналов создаются пирамиды $R(\sigma)$, $G(\sigma)$, $B(\sigma)$ и $Y(\sigma)$.

Наборы карт для цветовых каналов создаются подобно интенсивностным картам, при этом моделируются каналы с двойной цветооппонентностью (Engel et al., 1997; Хохлова, 2012): центры рецептивных полей нейронов возбуждаются одним цветом (например, красным) и тормозятся другим (например, зеленым), тогда как на периферии происходит обратное. Карты, моделирующие двойную цветооппонентность в первичной зрительной коре человека (зелёный/красный

($\mathcal{RG}$) и синий/жёлтый ($\mathcal{BY}$)), рассчитываются по формулам:

$$\mathcal{RG}(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|,$$

$$\mathcal{BY}(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|.$$

Локальная информация об ориентации извлекается из I с использованием ориентированной габоровской пирамиды $O(\sigma,\theta)$, где $\theta \in \{0°, 45°, 90°, 135°\}$. Ориентационные карты признаков $\mathcal{O}(c,s,\theta)$ кодируют локальные различия в ориентации между центром и периферией, представленных разными масштабами:

$$\mathcal{O}(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)|.$$

Таким образом, всего создаётся 42 карты признаков: 6 для интенсивности, 12 для цвета и 24 для ориентации.

Объединение карт признаков в карты *заметности* (conspicuity) и *салиентности* представляет собой проблему: разные «модальности» имеют разный динамический диапазон, для них используются разные механизмы извлечения признаков, вследствие чего их сложно сопоставить между собой. Кроме того, салиентные объекты, представленные только лишь на нескольких картах признаков, могут маскироваться шумом или менее салиентными объектами, представленными в большем числе карт. В условиях отсутствия в модели механизма, обеспечивающего нисходящий контроль, авторы предлагают применять оператор нормализации карт $\mathcal{N}(.)$, который бы повышал глобальную роль тех из них, что содержат небольшое число сильных пиков активности, и понижал бы её для тех, которые содержат большое количество сопоставимых по силе пиков. Применение $\mathcal{N}(.)$ предполагает:

- приведение значений карт к единому фиксированному диапазону $[0..M]$, чтобы избавиться от модально-специфичных амплитудных различий;
- поиск глобального максимума карты M и вычисление среднего $\overline{m}$ всех её локальных максимумов;
- глобальное перемножение карты на $(M - \overline{m})^2$.

Объясняя работу оператора, авторы ссылаются на модель корковых механизмов латерального торможения (Cannon & Fullenkamp, 1996): когда $M - \overline{m}$ достаточно велика, наиболее активная локация резко выделяется, и карта делается более важной; если же разница мала, карта не содержит ничего уникального и оказывается малозначащей.

Карты признаков объединяются в три карты заметности $\overline{I}$, $\overline{C}$ и $\overline{O}$ соответственно для интенсивности, цвета и ориентации. Карты заметности создаются путём сложения после приведения всех карт гауссовой пирамиды к единому масштабу с $\sigma = 4$; данная операция обозначена авторами $\bigoplus$:

$$\bar{I} = \bigoplus_{c=2s=c+3}^{4} \bigoplus^{c=4} \mathcal{N}(\mathcal{I}(c,s)),$$

$$\bar{C} = \bigoplus_{c=2s=c+3}^{4} \bigoplus^{c=4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))],$$

$$\bar{O} = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}\left(\bigoplus_{c=2s=c+3}^{4} \bigoplus^{c=4} \mathcal{N}(\mathcal{O}(c,s,\theta))\right).$$

При вычислении $\bar{O}$ сначала создаются четыре промежуточные карты путем объединения шести карт признаков для каждого θ, затем они объединяются в единую карту заметности.

Авторы объясняют создание трёх независимых каналов $\bar{I}$, $\bar{C}$ и $\bar{O}$ и их отдельную нормализацию гипотезой, что схожие признаки сильно конкурируют за салиентность, тогда как различные модальности вносят независимый вклад в карту салиентности. Три карты заметности нормализуются и суммируются в финальный вход $\mathcal{S}$ карты салиентности SM:

$$\mathcal{S} = \frac{1}{3}(\mathcal{N}(\bar{I}) + \mathcal{N}(\bar{C}) + \mathcal{N}(\bar{O})).$$

В каждый момент времени максимум активации карты SM определяет наиболее салиентную локацию в изображении, на которую должен быть направлен фокус внимания. Для определения точки, в которую модель должна переключиться в следующий раз, можно было бы просто выбрать наиболее активное местоположение на карте. Однако авторы, исходя из соображений биологического правдоподобия, моделируют карту салиентности как двумерный слой нейронов-пороговых интеграторов с утечкой (leaky integrate-and-fire neuron) в масштабе $\sigma = 4$. Модель таких нейронов включает один «конденсатор», накапливающий заряд от синаптического входа, проводимость утечки и пороговое напряжение. Когда достигается порог, генерируется «потенциал действия» (прототипический спайк), и заряд «конденсатора» сбрасывается до нуля. Максимум активации карты поступает в биологически правдоподобную двумерную нейронную сеть WTA, в которой синаптические взаимодействия между элементами гарантируют, что остается только самое активное местоположение, в то время как все остальные подавляются (здесь авторы отсылают нас в т. ч. к ранее рассмотренной работе (Koch & Ullman, 1985)).

Нейроны в SM получают возбудительный вход от $\mathcal{S}$ и не зависят друг от друга; следовательно, их потенциал в более салиентных локациях увеличивается быстрее

(эти нейроны используются как чистые интеграторы и не спайкируют постоянно). Каждый нейрон SM возбуждает свой соответствующий нейрон WTA. Все нейроны WTA также изменяют своё состояние независимо друг от друга, пока один («победитель») первым не достигнет порога и не сработает. Это запускает три одновременных механизма:

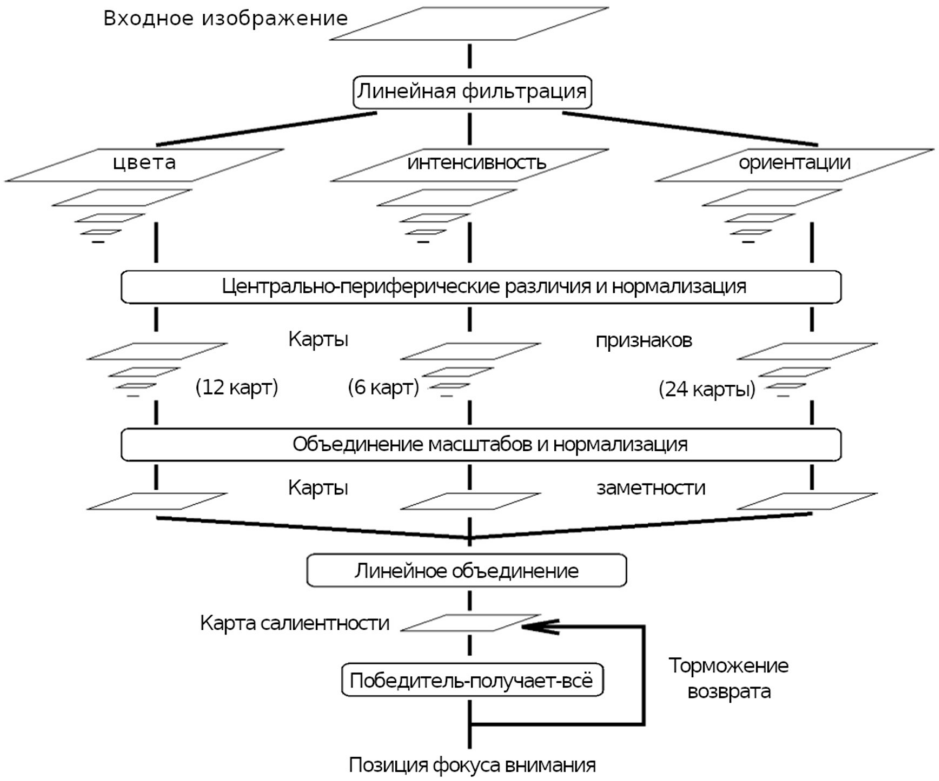
- фокус внимания смещается к локации нейрона-победителя;
- запускается глобальное торможение и полностью подавляет (сбрасывает) все нейроны WTA;
- в SM, в области, соответствующей положению и размеру нового фокуса внимания, временно активируется локальное торможение; это не только приводит к динамическим сдвигам фокуса, позволяя следующей наиболее салиентной локации впоследствии стать победителем, но и не даёт фокусу внимания немедленно вернуться в ранее посещённое место.

Такое «торможение возврата внимания» описано в исследованиях зрения человека (см. напр. (Уточкин & Фаликман, 2006)). Кроме того, моделируется правило «предпочтения близости» Коха и Ульмана (Koch & Ullman, 1985): чтобы слегка переориентировать модель на поиск следующей салиентной локации, близкой к ранее посещённой, в SM, в ближнем окружении текущего фокуса внимания, временно активируется небольшое возбуждение.

Поскольку данная модель салиентности не учитывает управление «сверху–вниз», фокус внимания представляет собой простой диск, радиус которого постоянен и равен $\frac{1}{6}\min(h,w)$, где h, w – соответственно высота и ширина входного изображения. Временны**ы**е константы, значения проводимости и пороги срабатывания моделируемых нейронов были выбраны таким образом, чтобы фокус переходил от одного салиентного места к другому примерно за 30 – 70 мс, а ранее посещённая локация была подавлена примерно на 500 – 900 мс, что соответствует психофизическим данным (Posner & Cohen, 1984). Разница в относительной величине этих задержек оказалась достаточной для обеспечения полного сканирования изображения и предотвращения зацикливания по ограниченному числу локаций. Все настроечные параметры зафиксированы в авторской реализации модели на C++, и с ними система демонстрирует временн**у**ю стабильность на всех тестовых изображениях. Обобщённая схема модели показана на рис. 3.

Рисунок 3

Общая схема модели салиентности Л. Итти, К. Коха и Э. Нейбура. Адаптировано из (Itti et al., 1998)



Обзор Али Борджи и Лаурент Итти (Borji & Itti, 2013), фактически подводящий итоги развития моделирования салиентности к моменту возникновения массового интереса к технологиям глубокого обучения, охватывает более полусотни моделей, опубликованных за период с 1998 по начало 2012 г. Так, авторы анализируют 52 собственно модели салиентности, рассматривающие прежде всего восходящее внимание, причём в этот анализ не попали известные им разработки (Baluja & Pomerleau, 1994; Milanese, 1993; Tsotsos et al., 1995), представленные ранее 1998 г., т. е. до опубликования «первой полной реализации и верификации модели Коха и Улльмана, предложенной Итти с соавт.» (Borji & Itti, 2013, p. 186). В обзоре также анализируются работы, представляющие более обобщённые модели внимания с нисходящим управлением – их насчитывается 11, две предложены до 1998 г. (McCallum, 1996; Rao et al., 2002). Вероятно, нет смысла перечислять все рассмотренные модели здесь; однако особенно интересными представляются теоретические обобщения, сделанные авторами в ходе проведённого анализа, краткое изложение которых представлено ниже. Авторы выделяют следующие свойства моделей, важные для категоризации и понимания их особенностей:

1. управление «снизу-вверх» (bottom-up) и «сверху-вниз» (top-down). Модели могут представлять преимущественно восходящие, основанные на некоторых характеристиках зрительной сцены, или нисходящие (знания, ожидания, подкрепление, текущие цели и др.) факторы управления вниманием, либо учитывать и те, и другие. При этом они отличаются по:

a. используемым признакам (features). Могут учитываться как отдельные низкоуровневые (цвет, ориентация и т. д.), так и достаточно сложные свойства объектов. В случаях, когда в модели присутствует управление «сверху-вниз», может использоваться механизм подстройки детекторов признаков. Модели, обрабатывающие признаки, тесно связаны с чисто вычислительными методами обнаружения объектов; когнитивное моделирование и компьютерное зрение взаимно обогащают друг друга;

b. степени учёта контекста сцены. Известно, что при очень коротких экспозициях (80 мс и менее) наблюдатель способен улавливать основное содержание ("gist") сцены. Её репрезентация не содержит большого числа деталей представленных в ней объектов, однако может дать информацию, достаточную для грубого (coarse) различения (например, внутри или вне помещения). Влияние контекста проявляется также в скорости обнаружения объектов и в особенностях движений глаз. Традиционные вычислительные модели, учитывающие основное содержание сцены, как правило используют фильтрацию (в т. ч. биологически обоснованную типа центрально-периферической, фильтров Габора) или спектральные методы для извлечения признаков, размерность пространства которых в дальнейшем снижается при помощи метода главных компонентов (МГК), анализа независимых компонентов (АНК) или кластерного анализа. В результате получают вектор значений ("gist vector"), характеризующих сцену. Авторы обзора отмечают, что на момент его написания популярность данного подхода в компьютерном зрении росла;

c. учёту требований задачи. Задача сильно влияет на распределение внимания, и сцены могут интерпретироваться на основе потребностей, возникающих для удовлетворения требований задачи. При решении сложных задач наблюдается сильная связь между визуальным познанием и движениями глаз. Так, при визуальном контроле большинство фиксаций направляется на релевантные задаче области. По движениям глаз часто можно понять и алгоритм решения, которого придерживается испытуемый. В частности, в задаче копирования блоков (парадигма Балларда, подробнее см. (Ballard et al., 1997, 1995; Hayhoe & Ballard, 2005)), предполагающей воспроизведение испытуемым конструкции из элементарных «строительных» блоков типа фигур разного цвета, испытуемые сначала выбирали целевой блок в исходной конструкции, удостоверяясь в его положении, а затем совершали фиксацию на рабочем пространстве, чтобы поместить соответствующий блок в правильное место. Авторы приводят также список работ, в которых подобным образом исследовалась и деятельность в естественных условиях.

Авторы обзора отмечают, что восходящее и нисходящее внимание объединяются для управления нашим вниманием, приводя несколько вариантов реализации правил интеграции этих процессов.

2. только пространство или пространство и время. Модели могут учитывать движение объектов, а также предсказывать переключения внимания между объектами в статической или динамической сцене;
3. явное (overt) и скрытое (covert) внимание. Модели могут описывать как явное, так и скрытое внимание, однако степень учёта ими скрытого внимания сложно оценить из-за сложности его измерения;
4. объекты или пространственные локации. Учитывая то, что имеются основания выделять внимание, основанное на признаках (feature-based), и внимание, имеющее дело с объектами (object-based), модели могут отдавать предпочтение какому-либо из его видов;
5. признаки, используемые в модели. Многие модели используют традиционные признаки, используемые в теории интеграции; однако встречается и множество других, таких как математически сконструированные (вейвлетные, основанные на МГК, АНК), геометрические и т. д.;
6. стимулы и тип задачи. Так как для проверки модели необходимы реальные эмпирические данные, авторы выделяют два основания для различения моделей по используемым при сборе данных стимулам: статические/динамические и искусственные/натуральные. Важен и тип задачи, решаемой испытуемым. Это может быть свободный просмотр, зрительный поиск или интерактивная задача;
7. метрики, используемые для оценивания. При оценивании модели выдаваемый ею прогноз обычно сравнивается с эмпирически полученным результатом (ground-truth); часто в качестве такого результата выступают различные варианты карт фиксаций взора. В зависимости от карты и типа выдаваемого моделью результата (точки фиксаций, двумерное распределение вероятностей и т. д.) могут использоваться площадь под кривой в нескольких модификациях (далее AUC), нормализованная салиентность пути осмотра (далее NSS), метрика Кульбака-Лейблера (KL), коэффициент корреляции Пирсона и др. Подробное обсуждение различных метрик содержится в более свежей работе (Bylinskii et al., 2017);
8. используемые наборы данных о движениях глаз. На момент выхода обзора Итти и Борджи в свободном доступе уже имелись данные о движениях глаз, записанные как при просмотре статических изображений (Bruce & Tsotsos, 2005; Judd et al., 2009), так и видео (Marat et al., 2009). Многие авторы для обучения и проверки моделей использовали собственные данные, которые со временем становились доступны другим исследователям;
9. Модели могут быть классифицированы на основании того, каким именно образом вычисляется салиентность. Например, модель может строиться на нейроноподобных вычислениях, а может использовать формальные высокоуровневые подходы. Авторы отмечают, что некоторые модели попадают сразу в несколько категорий, но тем не менее в дальнейшем используют простую одноуровневую классификацию:

a. когнитивные модели. Почти все модели внимания создавались под влиянием когнитивных концепций. Однако к данному классу авторы относят те из них, которые сильнее связаны с психологией или нейрофизиологией; автору настоящего обзора представляется, что речь может идти о содержательной связанности, т. к. используемые в них алгоритмы так или иначе пересекаются с психологическими и/или нейрофизиологическими концепциями;

b. байесовские модели. «В этих моделях априорные знания (например, контекст сцены или её суть (gist)) и сенсорная информация (например, целевые признаки) вероятностно комбинируются в соответствии с правилом Байеса (например, для обнаружения объекта интереса)» (Borji & Itti, 2013, p. 194). Эти модели способны обучаться на данных и обобщать различные факторы;

c. модели, основанные на теории принятия решений. В основе таких моделей лежит идея о том, что зрительное внимание должно управляться в контексте текущей задачи оптимальным образом; при этом они могут основываться на очень разных алгоритмах (как биологически обоснованных, так и чисто вычислительных);

d. модели, основанные на теории информации. Эти модели строятся на положении, что салиентные области являются наиболее содержательными с точки зрения количества информации. Вычислительно эти модели основываются на сравнении различных статистических оценок участков изображения (энтропия, параметры распределения и др.);

e. графовые вероятностные модели. «Графовые модели можно рассматривать как обобщённую версию байесовских моделей» (Borji & Itti, 2013, p. 197). В таких моделях используются графы, отображающие структуру условной независимости случайных величин; движения глаз рассматриваются как временной ряд. Из-за существования скрытых переменных, влияющих на формирование движений глаз, в них могут применяться такие решения, как скрытые марковские модели (HMM), динамические байесовские сети (DBN) и условные случайные поля (CRF);

f. модели, основанные на спектральном анализе. Данная группа моделей строится на анализе свойств изображения, часто с масштабированием, представленного в частотной области (амплитудный и фазовый спектр);

g. модели, основанные на классификации паттернов. В этих моделях применяются методы машинного обучения, такие как метод опорных векторов (SVM), регрессия и др. Обучение проводится на специальным образом размеченных данных (например, с разбивкой на области, каждая из которых помечается как салиентная или несалиентная);

h. прочие модели. Довольно обширный и сильно размытый «класс» моделей, характеризующихся оригинальностью и основанных на самых разных вычислительных решениях.

Опираясь на эти свойства, авторы обзора составили чрезвычайно полезную сводную таблицу рассмотренных ими моделей (Borji & Itti, 2013, p. 201), позволяющую

читателю быстро ориентироваться в огромном массиве достаточно сложных разработок и найти необходимую библиографическую информацию. Каждое из перечисленных свойств представляет собой столбец таблицы, в строках перечисляются известные авторам модели; в ячейках расположены условные обозначения, которыми кодируется наличие у модели того или иного свойства. Так, пользуясь таблицей, можно быстро определить, что рассмотренная нами ранее классическая модель Итти, Коха и Нейбура (Itti et al., 1998) является восходящей, пространственной, а не пространственно-временной, статической; имеющей дело с натуральными стимулами и задачей свободного просмотра, основанной на пространственных локациях, а не на объектах, учитывающей только простые признаки (цвет, яркость, ориентация), когнитивной; данные для обучения модели не использовались.

Нейросетевые модели салиентности

Переходя к описанию моделей салиентности, основанных на методах глубокого обучения, нельзя не упомянуть о существовании замечательного обзора, опубликованного А. Борджи в 2021 (Borji, 2021), но доступного как препринт с 2019 (Borji, 2019). Хочется рекомендовать этот документ заинтересованному читателю как ценный источник справочной информации по нейросетевым моделям и датасетам, созданным в прошедшем десятилетии, по применяемым метрикам и методикам оценивания производительности моделей. Имея в виду существование этого высококлассного обзора, автор настоящего текста (Д. Я.) ставит перед собой две достаточно скромных задачи: познакомить читателя с историей и логикой развития направления на примере работ одной из самых успешных научных групп, работающих в области моделирования салиентности; рассмотреть модели, созданные после написания обзора Борджи, и попытаться выделить и обобщить их характерные особенности.

Работа А. Крижевского, И. Суцкевера и Дж. Э. Хинтона (Krizhevsky et al., 2012) произвела очередную революцию в исследованиях искусственного интеллекта, возродив массовый интерес к нейронным сетям глубокого обучения, несколько угасший вследствие бурного развития на рубеже веков таких направлений машинного обучения, как ядерные методы и деревья решений (см. напр. (Шолле, 2023)). Модель, в дальнейшем получившая название AlexNET, в 2012 одержала уверенную победу на ежегодном соревновании ImageNet, достигнув рекордной производительности в 83% при классификации 1000 категорий объектов. Использование новой тогда многослойной свёрточной архитектуры и графических ускорителей вычислений позволили исследователям в ближайшие годы добиться впечатляющих результатов, в том числе и при моделировании зрительной салиентности.

Уже в 2014 группа исследователей из университета Тюбингена (Bethge Lab) разработала модель DeepGaze I (Kümmerer et al., 2015), при создании которой

использовались веса из нейросети А. Крижевского с соавт. Использование технологии переноса знаний (transfer learning) позволило авторам достичь существенного прироста производительности по сравнению с ранее созданными моделями. Так, корреляция между прогнозами и трекинговыми данными на наборе данных MIT300 составляет 0,6144. Модель использовала выходы слоёв свёрточной части AlexNET, которые линейно комбинировались с разными весами. Получившийся слой подвергался фильтрации (свёртка с гауссовым ядром), затем к нему поэлементно прибавлялась матрица весов, реализующих поправку на центральное смещение (center bias). В таком виде результат поступал на слой softmax, на выходе которого формировалось распределение вероятностей фиксаций. Чтобы стимулировать разреженность, в модели применялась l1-регуляризация весов.

В 2017 появилась новая версия модели, DeepGaze II (Kümmerer et al., 2017). В ней в качестве базовой части использовалась уже свёрточная часть VGG-19 (Simonyan & Zisserman, 2015); информация извлекалась из слоёв conv5_1, relu5_1, relu5_2, conv5_3, relu5_4. Обучаемая часть была усложнена (4 свёрточных слоя 1x1), но в остальном модель напоминала предыдущую. Модель продемонстрировала очень высокую на тот момент производительность: так, корреляция эмпирических данных MIT300 с прогнозом составила 0,7703.

Параллельно с ней была создана модель DeepGaze ICF, в которой вместо базовой части в виде слоёв сети, которую предварительно обучали распознаванию объектов, использовались операции выделения исключительно низкоуровневых признаков. Вычисления проводились для яркостного и двух цветоразностных компонентов в пяти масштабах (гауссова пирамида) для соответственно яркости и контраста; таким образом, на выходе формировалось 30 карт низкоуровневых признаков. Эта модель достигала лучшей производительности (корреляция 0,5876 на MIT300), чем все модели, не использующие признаки из нейросетей, предварительно обученных распознаванию объектов, что, по мнению авторов, делает её надежной базой для оценки полезности высокоуровневых признаков. Благодаря этой модели авторы обнаружили, что часть фиксаций значительно лучше предсказывается по низкоуровневым признакам.

Модель DeepGaze IIE (Linardos et al., 2021), представленная в 2021 году, является улучшенной версией DeepGaze II. Обучаемая часть сети сделана более глубокой, а активации ReLU заменены на norm и softplus. Обучение производилось на датасетах Salicon, а затем MIT1003. Главное изменение коснулось базовой сети: оригинальная VGG-19 могла заменяться на другие глубокие сети, обученные на датасете ImageNet (ResNet50 (He et al., 2015), EfficientNet85 (Tan & Le, 2020) и т. д.). По данным MIT/Tübingen Saliency Benchmark) (https://saliency.tuebingen.ai/results.html), наивысшая корреляция между прогнозом и эмпирическими картами фиксаций составила 0,8242; фактически это наилучшая модель из протестированных на данный момент и представленных на сайте. Однако авторы продолжают создание новых версий модели.

В 2022 была представлена DeepGaze III (Kümmerer, Bethge, et al., 2022; Kümmerer, Wallis, et al., 2022), включающая модуль пространственного прогнозирования,

который учитывает влияние содержимого сцены на положение фиксации, и модуль истории траекторий сканирования, который выявляет влияние более ранних фиксаций и, следовательно, динамику траектории перемещения взора. Первый модуль в основных чертах повторяет ранее созданные пространственные модели; второй использует информацию о четырёх или менее предыдущих фиксациях для прогноза текущей фиксации, которая представляется в виде карт трёх признаков – расстояния до данной фиксации, а также смещения по x и по y. Информация о предыдущих фиксациях, сделанных субъектом, обрабатывается в этом модуле, а затем объединяется с пространственной картой в сети выбора фиксаций. Финальное предсказание размывается, объединяется с весами поправки на центральное смещение и преобразуется в распределение вероятностей с помощью softmax. Судя по приводимым авторами значениям AUC = 0,906 и NSS = 2,957, полученным на MIT300 (величина корреляции не приводится), модель демонстрирует наивысшую производительность из ранее представленных, однако данные о ней на MIT/Tübingen Saliency Benchmark пока отсутствуют. Использованный авторами подход позволяет исследовать влияние на перцептивную заметность не только физических свойств изображения и задачи, но и ранее произведённых фиксаций.

Идею обработки признаков, извлекаемых из слоёв свёрточной сети, обученной распознаванию объектов, используют и авторы модели TranSalNet (Lou et al., 2022). При разработке модели они ставили перед собой не только задачу получения максимального результата, но и стремились приблизить архитектуру искусственной сети к перцептивной системе человека. Сначала изображение подается в свёрточный кодировщик. Для получения многомасштабных представлений из кодировщика извлекаются три набора карт признаков с различными пространственными размерами. Из-за присущих свёрточным архитектурам индуктивных искажений извлеченные представления изображений не содержат контекстной информации крупного плана, что потенциально делает модель салиентности менее похожей на человеческую – авторы обращают внимание читателя на то, зрительная система человека способна улавливать как локальную, так и глобальную информацию. Поэтому для получения прогноза, более релевантного с точки зрения восприятия, эти карты признаков пропускаются через три трансформера-кодировщика (Vaswani et al., 2023), что позволяет получить карты признаков глобального характера с улучшенной передачей контекстной информации. Трансформеры-кодировщики содержат многоголовый слой самовнимания (multi-head self-attention) и многослойный перцептрон. Затем свёрточный декодировщик объединяет карты признаков для построения прогноза салиентности. Модель демонстрирует производительность, сопоставимую с DeepGaze: при использовании DenseNet-161 (Huang et al., 2018) в качестве базовой сети корреляция между прогнозом и данными MIT300 составляет 0,8070; с ResNet-50 корреляция незначительно снижается (0,7991).

Прямосвязные свёрточные нейросети, несмотря на их значительные возможности по формированию репрезентаций элементов изображения, могут

игнорировать их внутренние связи и лишены потенциальных преимуществ, обеспечиваемых использованием обратной связи в зрительных задачах. Это относится и к моделированию салиентности. Учитывая это обстоятельство, авторы модели SalFBNet (Ding et al., 2022) предлагают сверточную архитектуру с обратной связью и рекурсией. Предлагаемая модель может формировать множественные контекстные представления, используя рекурсивный путь от блоков признаков более высокого уровня к низкоуровневым слоям.

Чтобы решить проблему дефицита обучающих данных, авторы используют особый подход к переносу знаний, создавая крупномасштабный обучающий набор при помощи готовых моделей салиентности, перечисленных на сайте MIT/Tübingen Saliency Benchmark. Сначала они обучают предлагаемую модель на полученных таким образом искусственных данных, затем дообучают её на реальных фиксациях взора. Кроме того, чтобы облегчить обучение своей модели с обратной связью, авторы предлагают новую функцию потерь, названную ими sFNE (ошибка селективной фиксации и нефиксации). Многочисленные экспериментальные результаты показывают, что SalFBNet с меньшим количеством параметров достигает конкурентоспособных результатов в общедоступных тестах моделей салиентности, что говорит об эффективности как самой модели с обратными связями, так и использования искусственных данных для предварительного обучения. SalFBNet находится на втором месте по производительности после DeepGaze IIE (корреляция с данными MIT300 0,8141).

Модель Saliency TRansformer (SalTR) (Dahou Djilali et al., 2024) основывается на новом подходе к прогнозированию салиентности в изображениях, использующем параллельное декодирование в сетях-трансформерах для обучения сети исключительно на основе карт фиксаций. Чтобы преодолеть сложности оптимизации для дискретных карт, модели обычно обучаются на непрерывных картах. Разработчики SalTR осуществляют попытку построить экспериментальную вычислительную систему, которая генерирует наборы данных о салиентности. Авторский подход рассматривает оценивание салиентности как проблему прямого прогнозирования набора данных с помощью функции глобальной потери, которая обеспечивает прогнозирование отдельных фиксаций посредством двустороннего сопоставления и архитектуры трансформера – кодировщика-декодировщика, на входе которого располагается базовая сеть ResNet50. Используя фиксированный набор изученных запросов фиксаций, перекрестное внимание обрабатывает информацию о свойствах изображения для непосредственного вывода точек фиксации, что отличает данную разработку от других современных моделей. Авторы отмечают, что их подход позволяет достичь оценок, сравнимых с другими современными подходами, в тестах Salicon и MIT300. Так, реализация SalTR-Small обеспечивает корреляции прогнозов и исходных образцов на уровне 0,84 и 0,7 для Salicon и MIT300 соответственно, SalTR-Base – 0,87 и 0,75. Применение в моделях деформируемых свёрток увеличивает сходство до соответственно 0,86 и 0,76 (small)

и 0,89 и 0,8 (base). Таким образом, SalTR является действительно одной из лучших современных моделей зрительной салиентности.

Моделирование зрительной салиентности развивается и в направлении обработки видеопотока. В работе (Droste et al., 2020) авторы обращают внимание на то, что моделирование салиентности для изображений и видео рассматривается в современной литературе по компьютерному зрению как две независимые задачи. И если моделирование для изображений является хорошо разработанной проблемой, и прогресс в этой области замедляется, что видно по бенчмаркам SALICON и MIT300, модели салиентности для видео в последнее время показали быстрый рост в бенчмарке DHF1K (Wang et al., 2021). Авторы задаются вопросом – можно ли подойти к моделированию салиентности для изображений и видео с помощью единой модели с взаимной пользой? По их мнению, ключевые перспективы для совместного моделирования даёт применение сдвига домена (domain shift – адаптация системы ИИ к применению в новой области и/или к новым данным) как между данными о салиентности для изображений и для видео, так и между различными наборами видеоданных. В дополнение к улучшенному алгоритму создания обученных гауссовых приоров (корректировка на сдвиг взгляда к центру), для решения этой задачи предлагаются четыре новых метода адаптации домена: доменно-адаптивные априорные значения, доменно-адаптивное слияние, доменно-адаптивное сглаживание и обход рекуррентной сети. Эти методы интегрируются «в простую и легкую» (Droste et al., 2020, p. 1) сеть UNISAL, имеющую архитектуру «кодировщик – рекуррентный блок – декодировщик», обученную на данных о салиентности и для изображений, и для видео. Результаты обучения оцениваются на наборах видеоданных DHF1K, Hollywood-2 и UCF-Sports, а также на статических датасетах SALICON и MIT300. С одним и тем же набором параметров UNISAL достигает наивысших на момент публикации показателей на всех наборах данных о салиентности для видео и находится на одном уровне с лучшими моделями в тестах на данных для изображений (корреляция с данными MIT300 составляет 0,7851); при этом по сравнению со всеми конкурирующими моделями, использующими глубокое обучение, время выполнения сокращается 5–20 раз, а сама модель имеет меньший размер. Авторы также проводят ретроспективный анализ и абляционные исследования (ablation studies – исследования роли компонента ИИ-системы, проводимые путём его отключения), которые подтверждают важность сдвига домена при моделировании.

Таким образом, для современных подходов в моделировании салиентности, использующих методы глубокого обучения, характерны:

- использование модульных архитектур нейронных сетей с возможностью замены модулей;
- технологии переноса знаний – модульное подключение сетей, обученных распознаванию объектов, с целью извлечения репрезентаций из их слоёв, а также использование искусственных наборов данных для предобучения;

- тенденции к совершенствованию моделей через сдвиг области их применения (domain shift: например, изображения и видео);
- выход за пределы классических свёрточных архитектур, применение реккурентных вставок, слоёв самовнимания, обратных связей, трансформеров;
- возможность манипулирования отдельными модулями с целью изучения их вклада в работу системы (например, для проведения абляционных исследований).

Заключение

Между выходом статьи Коха и Улльмана (1985) и началом практической проверки и воплощения их идей прошёл довольно большой срок. Исследователей интересовал в первую очередь алгоритм формирования первоначальной карты салиентности, деталей реализации которого практически на касались авторы основополагающей работы. Первый, традиционный этап развития моделей салиентности, характеризовался большим разнообразием применяемых вычислительных методов и подходов; предлагались в том числе и решения, хорошо совместимые с психологическими и нейрофизиологическими данными. На этом этапе модели зрительной салиентности в основном были «прозрачными» с точки зрения их внутреннего устройства, что делало их особенно ценными по причине возможности сопоставления с теоретическими моделями когнитивных наук. Однако по мере развития методов машинного обучения (байесовские классификаторы, метод опорных векторов и др.), ставших особенно популярными в первом десятилетии 21 века, некоторые условно традиционные решения стали всё больше напоминать «чёрный ящик». С приходом очередной революции в технологиях нейронных сетей, случившейся в 2012 году, тенденция многократно усилилась, однако были достигнуты и впечатляющие результаты в плане производительности. Хочется надеяться, что по мере развития инструментов анализа специфических алгоритмов, вырабатываемых сетью в процессе обучения, содержимое «чёрного ящика» станет не так уж сложно прочесть. Внушает оптимизм также рост объёма общедоступных данных, используемых для обучения моделей салиентности, и наличие ясного понимания сообществом важности учёта типа (свободный просмотр, зрительный поиск и т. д.) и особенностей задачи, которую решает испытуемый при их сборе.

По мере формирования эффективных вычислительных подходов к моделированию в литературе стали обсуждаться возможности практического применения моделей салиентности в компьютерном зрении (Medioni & Mordohai, 2005), инженерной психологии и «юзабилити» (Sun et al., 2019), анализе медицинских снимков (Arun et al., 2020; Jampani et al., 2012), сжатии видео (Gitman et al., 2014; Lyudvichenko et al., 2017) и т. д. Появляются и первые коммерческие решения. Таким образом, моделирование зрительной салиентности приобрело к настоящему времени практическую значимость, позволяя как имитировать внимание в чисто технических целях, так и предсказывать его переключения у человека.

Литература

Ангельгардт, А. Н., Макаров, И. М., & Горбунова, Е. С. (2021). Роль уровня категории при решении задачи гибридного зрительного поиска. *Вопросы психологии*, 2, 148–158.

Величковский, Б. М. (2006). *Когнитивная наука. Основы психологии познания* (Т. 1). Смысл, Издательский центр «Академия».

Воронин, И. А., Захаров, И. М., Табуева, А. О., & Мерзон, Л. А. (2020). Диффузная модель принятия решения: оценка скорости и точности ответов в задачах выбора из двух альтернатив в исследованиях когнитивных процессов и способностей. *Теоретическая и экспериментальная психология, 13*(2), 6–24.

Горбунова, Е. С. (2023). Механизмы построения репрезентации в категориальном поиске: роль внимания и рабочей памяти. *Российский психологический журнал, 20*(3), 116–130. https://doi.org/10.21702/rpj.2023.3.6

Гусев, А. Н., & Уточкин, И. С. (2012). Влияние вероятности подсказки на эффективность пространственной локализации зрительного стимула. *Известия Иркутского государственного университета. Серия: Психология,* 1(1), 34–39.

Дренёва, А. А. (2020). Категориальный поиск трехмерных фигур испытуемыми с разным уровнем математической экспертизы. *Национальный психологический журнал, 1*(1 (37)), 57–65. https://doi.org/10.11621/npj.2020.0106

Кочурко, В. А., Мадани, К., Сабуран, К., Головко, В. А., & Кочурко, П. А. (2015). Обнаружение объектов системами компьютерного зрения: подход на основе визуальной салиентности. *Доклады Белорусского государственного университета информатики и радиоэлектроники*, 91(5), 47–53.

Крускоп, А. С., Лунякова, Е. Г., Дубровский, В. Е., & Гарусев, А. В. (2023). Особенности движений глаз в задаче зрительного поиска в зависимости от вербализуемости и симметричности стимулов. *Вестник Московского Университета. Серия 14: Психология,* 46(4), 88–111. https://doi.org/10.11621/LPJ-23-40

Мартынова, О. В., & Балаев, В. В. (2015). Возрастные изменения в функциональной связанности сетей состояния покоя. *Психология. Журнал Высшей школы экономики, 12*(4), 33–47.

Мински, М., & Пейперт, С. (1971). *Персептроны.* Мир.

Подладчикова, Л. Н., Самарин, А. И., Шапошников, Д. Г., Колтунова, Т. И., Петрушан, М. В., & Ломакина, О. В. (2017). *Современные представления о механизмах зрительного внимания.* Южный федеральный университет.

Рожкова, Г. И., Белокопытов, А. В., & Иомдина, Е. Н. (2019). Современные представления о специфике периферического зрения человека. *Сенсорные системы*, 33(4), 305–330. https://doi.org/10.1134/S0235009219040073

Сапронов, Ф. А., & Горбунова, Е. С. (2025). Сравнение сгенерированных ИИ стимулов и фото: исследование зрительного поиска. *Вестник Московского Университета. Серия 14: Психология*, 48(2), 109–131. https://doi.org/10.11621/LPJ-25-14

Сапронов, Ф. А., Макаров, И. М., & Горбунова, Е. С. (2023). Категоризация в гибридном поиске: исследование с использованием регистрации движений глаз. *Экспериментальная психология, 16*(3), 121–138. https://doi.org/10.17759/exppsy.2023160308

Сеченов, И. М. (1942). *Рефлексы головного мозга.* Издательство АН СССР.

Уточкин, И. С., & Фаликман, М. В. (2006). Торможение возврата внимания. Часть 1. Виды и свойства. *Психологический журнал,* 27(3), 42–48.

Трейсман, А. (1987). Объекты и их свойства в зрительном восприятии человека. *В мире науки*, 1, 68–78.

Фаликман, М. В. (2001). *Динамика внимания в условиях быстрого последовательного предъявления зрительных стимулов* (Дисс. … канд. психол. наук). МГУ имени М. В. Ломоносова, М.

Фаликман, М. В. (2015). *Структура и динамика зрительного внимания при решении перцептивных задач: конструктивно-деятельностный подход: дис. ... докт. психол. наук* [МГУ имени М. В. Ломоносова].

Фаликман, М. В., Уточкин, И. С., Марков, Ю. А., & Тюрина, Н. А. (2019). *Нисходящая регуляция зрительного поиска: есть ли она у детей?* В *Когнитивная наука в Москве: новые исследования:* Материалы конференции, Москва, 19 июня 2019 года (С. 513–517). БукиВеди. Институт практической психологии и психоанализа, 2019.

Хохлова, Т. В. (2012). Современные представления о зрении млекопитающих. *Журнал общей биологии,* 73(6), 418–434.

Шевель, Т. М., & Фаликман, М. В. (2022). «Подсказка взглядом» как ключ к механизмам совместного внимания:основные результаты исследований. *Культурно-историческая психология,* 18(1), 6–16. https://doi.org/10.17759/chp.2022180101

Шолле, Ф. (2023). *Глубокое обучение на Python* (2nd ed.). Питер.

Ярбус, А. Л. (1965). *Роль движений глаз в процессе зрения* (Н. Д. Нюберг, Ed.). Наука.

Arun, N. T., Gaw, N., Singh, P., Chang, K., Hoebel, K. V., Patel, J., Gidwani, M., & Kalpathy-Cramer, J. (2020, May 29). *Assessing the validity of saliency maps for abnormality localization in medical imaging.* http://arxiv.org/abs/2006.00063

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory Representations in Natural Tasks. *Journal of Cognitive Neuroscience*, 7(1), 66–80. https://doi.org/10.1162/jocn.1995.7.1.66

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*(4), 723–742. https://doi.org/10.1017/s0140525x97001611

Baluja, S., & Pomerleau, D. (1994). Using a Saliency Map for Active Spatial Selective Attention: Implementation & Initial Results. *Proc. Advances in Neural Information Processing Systems,* 451–458.

Bashinski, H. S., & Bacharach, V. R. (1980). Enhancement of perceptual sensitivity as the result of selectively attending to spatial locations. *Perception & Psychophysics*, 28(3), 241–248. https://doi.org/10.3758/bf03204380

Bergen, J. R., & Julesz, B. (1983–29C.E.). Parallel versus serial processing in rapid pattern discrimination. *Nature,* 303(5919), 696–698. https://doi.org/10.1038/303696a0

Borji, A. (2021). Saliency Prediction in the Deep Learning Era: Successes and Limitations. IEEE *Transactions on Pattern Analysis and Machine Intelligence,* 43(2), 679–700. https://doi.org/10.1109/TPAMI.2019.2935715

Borji, A. (2019, May 24). *Saliency Prediction in the Deep Learning Era: Successes, Limitations, and Future Challenges.* http://arxiv.org/abs/1810.03716

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207. https://doi.org/10.1109/TPAMI.2012.89

Bruce, N. D. B., & Tsotsos, J. K. (2005). Saliency Based on Information Maximization. In *NIPS'05:* Proceedings of the 18th International Conference on Neural Information Processing Systems (pp. 155–162).

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2017, April 6). *What do different evaluation metrics tell us about saliency models?* http://arxiv.org/abs/1604.03605

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology, 197*(3), 551–566. https://doi.org/10.1113/jphysiol.1968.sp008574

Cannon, M. W., & Fullenkamp, S. C. (1996). A model for inhibitory lateral interaction effects in perceived contrast. *Vision Research, 36*(8), 1115–1125. https://doi.org/10.1016/0042-6989(95)00180-8

Dahou Djilali, Y. A., McGuinness, K., & O'Connor, N. (2024). Learning Saliency From Fixations. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 382–392). https://doi.org/10.1109/WACV57701.2024.00045

Ding, G., İmamoğlu, N., Caglayan, A., Murakawa, M., & Nakamura, R. (2022). SalFBNet: Learning pseudo-saliency distribution via feedback convolutional networks. *Image and Vision Computing,* 120, 104395. https://doi.org/10.1016/j.imavis.2022.104395

Droste, R., Jiao, J., Noble, J.A. (2020). Unified Image and Video Saliency Modeling. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12350. https://doi.org/10.1007/978-3-030-58558-7_25

Engel, F. L. (1971). Visual conspicuity, directed attention and retinal locus. *Vision Research, 11*(6), 563–576. https://doi.org/10.1016/0042-6989(71)90077-0

Engel, F. L. (1974). Visual conspicuity and selective background interference in eccentric vision. *Vision Research*, 14(7), 459–471. https://doi.org/10.1016/0042-6989(74)90034-0

Engel, S., Zhang, X., & Wandell, B. (1997). Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature,* 388(6637), 68–71. https://doi.org/10.1038/40398

Eriksen, C. W., & Hoffman, J. E. (1972). Temporal and spatial characteristics of selective encoding from visual displays. *Perception & Psychophysics, 12*(2), 201–204. https://doi.org/10.3758/bf03212870

Feldman, J. A. (1982). Dynamic connections in neural networks. *Biological Cybernetics, 46*(1), 27–39. https://doi.org/10.1007/BF00335349

Geiger, G., & Lettvin, J. Y. (1986). Enhancing the Perception of Form in Peripheral Vision. *Perception, 15*(2), 119–130. https://doi.org/10.1068/p150119

Gitman, Y., Erofeev, M., Vatolin, D., Bolshakov, A., & Fedorov, A. (2014). Semiautomatic visual-attention modeling and its application to video compression. In *2014 IEEE International Conference on Image Processing (ICIP)* (pp. 1105–1109). https://doi.org/10.1109/ICIP.2014.7025220

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. In *Trends in Cognitive Sciences, 9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. https://doi.org/10.48550/arXiv.1512.03385

Helmholtz, H. von. (1896). *Handhuch der Physiologischen Optik (Zweite umgearbeitete Auflage).* Verlag von Leopold Voss.

Huang, G., Liu, Z., Maaten, L. van der, & Weinberger, K. Q. (2018, January 28). *Densely Connected Convolutional Networks.* https://doi.org/10.48550/arXiv.1608.06993

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Jampani, V., Ujjwal, Sivaswamy, J., & Vaidya, V. (2012). Assessment of computational visual attention models on medical images. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing - ICVGIP '12*, (pp. 1–8). https://doi.org/10.1145/2425333.2425413

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 2106–2113). https://doi.org/10.1109/ICCV.2009.5459462

Julesz, B. (1986). Texton gradients: The texton theory revisited. *Biological Cybernetics, 54*(4-5), 245–251. https://doi.org/10.1007/BF00318420

Julesz, B. (1984). A brief outline of the texton theory of human vision. *Trends in Neurosciences, 7*(2), 41–45. https://doi.org/10.1016/s0166-2236(84)80275-1

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems,* 25.

Kümmerer, M., Bethge, M., & Wallis, T. S. A. (2022). DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision, 22*(5), 7. https://doi.org/10.1167/jov.22.5.7

Kümmerer, M., Theis, L., & Bethge, M. (2015, April 9). Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. http://arxiv.org/abs/1411.1045

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2022). Using the DeepGaze III model to decompose spatial and dynamic contributions to fixation placement over time. *Journal of Vision, 22*(14), 3964. https://doi.org/10.1167/jov.22.14.3964

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 4789–4798).

Linardos, A., Kümmerer, M., Press, O., & Bethge, M. (2021). DeepGaze IIE: Calibrated Prediction in and Out-of-Domain for State-of-the-Art Saliency Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12919–12928).

Lou, J., Lin, H., Marshall, D., Saupe, D., & Liu, H. (2022). TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494, 455–467. https://doi.org/10.1016/j.neucom.2022.04.080

Lyudvichenko, V., Erofeev, M., Gitman, Y., & Vatolin, D. (2017). A semiautomatic saliency model and its application to video compression. In *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 403–410). https://doi.org/10.1109/ICCP.2017.8117038

Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. (2009). Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos. *International Journal of Computer Vision, 82*(3), 231–243. https://doi.org/10.1007/s11263-009-0215-3

McCallum, R. (1996). *Reinforcement learning with selective perception and hidden state: PhD thesis.* University of Rochester.

Medioni, G., & Mordohai, P. (2005). Saliency in Computer Vision. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 583–585). Academic Press. https://doi.org/10.1016/B978-012375731-9/50099-9

Milanese, R. (1993). *Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation: PhD thesis*. Univ. Geneva.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32*(1), 3–25. https://doi.org/10.1080/00335558008248231

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance* (Vol. 10, pp. 531–556). Erlbaum.

Posner, M. I., Cohen, Y., & Rafal, R. D. (1982). Neural systems control of spatial orienting. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 298*(1089), 187–198. https://doi.org/10.1098/rstb.1982.0081

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, 13, 25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. Annual Review of Neuroscience, 35, 73–89. https://doi.org/10.1146/annurev-neuro-062111-150525

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research, 42*(11), 1447–1463. https://doi.org/10.1016/s0042-6989(02)00040-8

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108. https://doi.org/10.1037/0033-295x.85.2.59

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention

in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Remington, R., & Pierce, L. (1984). Moving attention: Evidence for time-invariant shifts of visual selective attention. *Perception & Psychophysics, 35*(4), 393–399. https://doi.org/10.3758/bf03206344

Rubtsova, O. S., & Gorbunova, E. S. (2022). The Manifestation of Incidental Findings in Different Experimental Visual Search Paradigms. Psychology in Russia: State of the Art, 15(4), 140–158. doi: 10.11621/pir.2022.0409

Saarinen, J., & Julesz, B. (1991). The speed of attentional shifts in the visual field. *Proceedings of the National Academy of Sciences of the United States of America, 88*(5), 1812–1814. https://doi.org/10.1073/pnas.88.5.1812

Serences, J. T., & Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences, 10*(1), 38–45. https://doi.org/10.1016/j.tics.2005.11.008

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition.* http://arxiv.org/abs/1409.1556

Sun, X., Houssin, R., Renaud, J., & Gardoni, M. (2019). A review of methodologies for integrating human factors and ergonomics in engineering design. *International Journal of Production Research, 57*(15–16), 4961–4976. https://doi.org/10.1080/00207543.2018.1492161

Tan, M., & Le, Q. V. (2020). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. https://doi.org/10.48550/arXiv.1905.11946

Theeuwes, J. (2013). Feature-based attention: It is all bottom-up priming. *Philosophical Transactions of the Royal Society B: Biological Sciences, 368*(1628), 20130055. https://doi.org/10.1098/rstb.2013.0055

Treisman, A. M. (1982). Perceptual grouping and attention in visual search for features and for objects. Journal of Experimental Psychology. *Human Perception and Performance, 8*(2), 194–214. https://doi.org/10.1037//0096-1523.8.2.194

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*(1–2), 507–545. https://doi.org/10.1016/0004-3702(95)00025-9

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352. https://doi.org/10.1037/0033-295x.84.4.327

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 2). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Wang, W., Shen, J., Xie, J., Cheng, M.-M., Ling, H., & Borji, A. (2021). Revisiting Video Saliency Prediction in the Deep Learning Era. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(1), 220–237. https://doi.org/10.1109/TPAMI.2019.2924417

Wilson, H. R., & Bergen, J. R. (1979). A four mechanism model for threshold spatial vision. *Vision Research, 19*(1), 19–32. https://doi.org/10.1016/0042-6989(79)90117-2

Wolfe, J. M. (2012). Saved by a Log: How Do Humans Perform Hybrid Visual and Memory Search? *Psychological Science, 23*(7), 698–703. https://doi.org/10.1177/0956797612443968

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review, 28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology. Human Perception and Performance, 15*(3), 419–433. https://doi.org/10.1037//0096-1523.15.3.419

Информация об авторе

Денис Викторович Явна — кандидат психологических наук, доцент кафедры психофизиологии и клинической психологии, Федеральное государственное автономное образовательное учреждение высшего образования «Южный федеральный университет», г. Ростов-на-Дону, Россия; WoS Researcher ID: BB-1314-2013; Scopus ID: 56034231500; РИНЦ Author ID: 512495; SPIN-код РИНЦ: 3357-7716; ORCID ID: https://orcid.org/0000-0003-2895-5119; e-mail: yavna@fortran.su

Информация о конфликте интересов

Автор заявляет об отсутствии конфликта интересов.