УДК 621.391

М. Н. Фаворская

НАХОЖДЕНИЕ ДВИЖУЩИХСЯ ВИДЕООБЪЕКТОВ С ПРИМЕНЕНИЕМ ЛОКАЛЬНЫХ 3D-СТРУКТУРНЫХ ТЕНЗОРОВ

Приведен анализ методов нахождения видеообъектов на основе пространственно-временного подхода. Введено понятие 3D-структурного тензора для эффективного определения локальной ориентации пространственно-временного движения. Построены подробные схемы этапов предсегментации сцены, пространственно-временной сегментации, постсегментации с дальнейшим распознаванием видеообъектов с применением локальных 3D-структурных тензоров.

Ключевые слова: пространственно-временная сегментация, 3D-структурный тензор, оценка движения.

Задача нахождения объектов из видеопоследовательностей с целью их дальнейшего распознавания и предсказания поведения востребована во многих приложений цифрового видео, таких как мультимедиа, виртуальная реальность, компьютерное зрение, искусственный интеллект. Существуют различные методы нахождения видеообъектов, которые объединяют сегментацию изображения (пространственные методы) и сегментацию на основе признаков движения (временные методы) с целью повышения точности нахождения этих объектов. Обычно подходы к сегментации видеообъектов классифицируются по трем категориям: методы, основанные на поиске регионов, методы, основанные на определении границ, и вероятностные методы [1–3].

Методы, основанные на поиске регионов, используют информацию о кластеризации или операциях расщепления и выращивания регионов в пространстве признаков. Такая информация обычно формируется из векторов движения и некоторых пространственных признаков: цветности, текстуры, взаимного расположения. Недостатками такого подхода являются проблемы появления, перекрытия и исчезновения регионов из кадра, а также низкая точность определения границ регионов.

Методы определения границ обычно используют угловые детекторы или активные контуры в сочетании с информацией о полях движения видеообъектов. Такие методы основаны на принципах когнитивной психологии, однако они имеют низкую помехоустойчивость, а активные контуры – еще и сильную зависимость от выбора начальных параметров.

Вероятностные методы для нахождения движущихся объектов используют байесовский подход, алгоритм максимизации-ожидания, минимизации расстояний в различных метрических пространствах.

Указанные подходы обладают высокой вычислительной сложностью, причем некоторые методы требуют предварительного задания количества объектов-регионов в качестве входного параметра, что ограничивает их использование на практике.

Примем, что рассматривается сложная сцена с движением нескольких объектов. Объекты могут двигаться с различной скоростью, ускорением, в разных направлениях и иметь различные локальные размеры. Изображение объекта $\mathbf{O}_i$ может включать набор регионов $\{R_j\}$, обладающих разделимыми цветовыми ColorParams(R_j) и текстурными TextureParams(R_j) показателями. Интерпретация регионов как единого объекта производится на основе анализа показателей движения MotionParams(R_j) и повторяемости поведения регионов (в простейшем случае это отслеживание траектории перемещений регионов). Поскольку методы определения цветовых и текстурных показателей достаточно хорошо изучены, остановимся на нахождении показателей движения.

Известны два основных метода оценки движения, используемых для временной сегментации: метод оптического потока и метод соответствия блоков. В обоих подходах информация о движении формируется путем обнаружения изменения интенсивности пикселей между последовательными кадрами. При этом метод оптического потока обеспечивает более точное нахождение границ регионов, так как он использует информацию на уровне пикселей и лучше выделяет регионы со сложной текстурой, имеющие большие значения градиентов.

Видеопоследовательность $I(\boldsymbol{x})$, представленную в виде набора кадров, где $\boldsymbol{x} = [x, y, t]^T$, здесь x и y – пространственные координаты пикселей кадра по осям OX и OY соответственно, t – временная координата, учитывающая последовательность появления кадров, можно интерпретировать как трехмерный объем данных. Известно, что 3D-структурный тензор $\boldsymbol{J}$ позволяет эффективно определять локальную ориентацию пространственно-временного движения видеообъектов и определяется следующим образом:

$$\boldsymbol{J}(\boldsymbol{x}) = \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} = \begin{bmatrix} I_x^2 & I_x I_y & I_x I_t \\ I_y I_x & I_y^2 & I_y I_t \\ I_t I_x & I_t I_y & I_t^2 \end{bmatrix} = \nabla I(\boldsymbol{x}) \cdot \nabla I(\boldsymbol{x})^T,$$

где C – пространственно-временной градиент, вычисляемый по частным производным

$$\nabla I(\boldsymbol{x}) = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I}{\partial t} \right).$$

Собственные векторы $\boldsymbol{e}_k$ (k = 1, 2, 3) симметричной ковариационной матрицы $\boldsymbol{J}$ размером 3×3 можно определить по локальным смещениям интенсивностей изображений соседних кадров и использовать для оценки ло-

кальных ориентаций движущихся сегментов. При этом, в силу особенностей видеонаблюдения, собственные значения λ_k векторов $\boldsymbol{e}_k$ указывают на локальные отклонения яркости по трем направлениям и могут быть отсортированы в следующем порядке: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$. Выражение $\nabla I(\boldsymbol{x}) \cdot \nabla I(\boldsymbol{x})^T$ можно рассматривать как корреляционную матрицу, составленную из векторов градиентов в пространственно-временном объеме. В соответствии с методом главных компонент собственные векторы корреляционной матрицы сортируются в порядке убывания. Первый собственный вектор, соответствующий наибольшему собственному значению, указывает направление наибольшего изменения данных. Отношение каждого собственного значения к сумме трех собственных значений характеризует концентрацию энергии по соответствующему направлению. Таким образом, собственные значения собственных векторов локального 3D-структурного тензора можно использовать для обнаружения локальных изменений в последовательности кадров. Наименьшее собственное значение можно использовать для определения различий в кадрах, оно является более устойчивым к шуму и низко контрастным объектам фона по сравнению с простейшим методом яркостной разницы кадров.

На основе собственных значений $\lambda_1(x, y, t)$, $\lambda_2(x, y, t)$, $\lambda_3(x, y, t)$ можно построить карты $\lambda_1(I)$, $\lambda_2(I)$, $\lambda_3(I)$ локального 3D-структурного тензора. При этом карта собственных значений $\lambda_1(I)$ фиксирует как движущиеся объекты, так и некоторые изолированные текстурные регионы фона, карта $\lambda_2(I)$ является менее информативной для сегментации, а карта $\lambda_3(I)$ генерирует небольшие разрывы внутри масок видеообъектов. Поэтому при обнаружении движения основное внимание следует уделять первому собственному вектору корреляционной матрицы $\lambda_1(I)$.

Рассмотрим процесс нахождения объектов из видеопоследовательностей, который представляет собой сложную совокупность действий, особенно если сцена имеет несколько разнонаправленных движущихся объектов, а начальные условия наблюдения отсутствуют. Целесообразно принять, что имеются три этапа получения, обработки пространственно-временной информации и принятия решения:

– первый этап – предварительный анализ сцены (предсегментация);

– второй этап – пространственно-временная сегментация сцены;

– третий этап – постсегментация (с учетом многоуровневого движения) и распознавание объектов интереса сцены.

Этап предсегментации предназначен для грубой оценки пространственно-временных характеристик сцены (рис. 1). Предлагается использовать не исходное изображение опорного кадра Fr_0, а преобразованный с помощью гауссовой пирамиды слой с малым разрешением, что позволяет снизить количество обрабатываемых пикселей в 2^K раз, где K – номер слоя пирамиды Гаусса, без значительной потери качества. Гауссова пирамида строится как для опорного кадра Fr_0, так и для последовательности кадров $Fr_1, Fr_2, \ldots, Fr_i$. Полученные таким образом изображения низкого разрешения обрабатываются для получения информации о предварительной пространственной сегментации сцены (вычисление признаков цветности, интенсивности, текстурных характеристик регионов и формирование гомогенных регионов) и предварительной временной сегментации (использование локальных 3D-структурных тензоров и вычисление коэффициентов корреляции между соседними кадрами).

Для грубого нахождения движущихся регионов целесообразно вычислять коэффициент R корреляции между кадрами, используя наименьшие собственные значения $\lambda_3(I_t)$ и $\lambda_3(I_{t+1})$ кадров Fr_t и Fr_{t+1} соответственно:

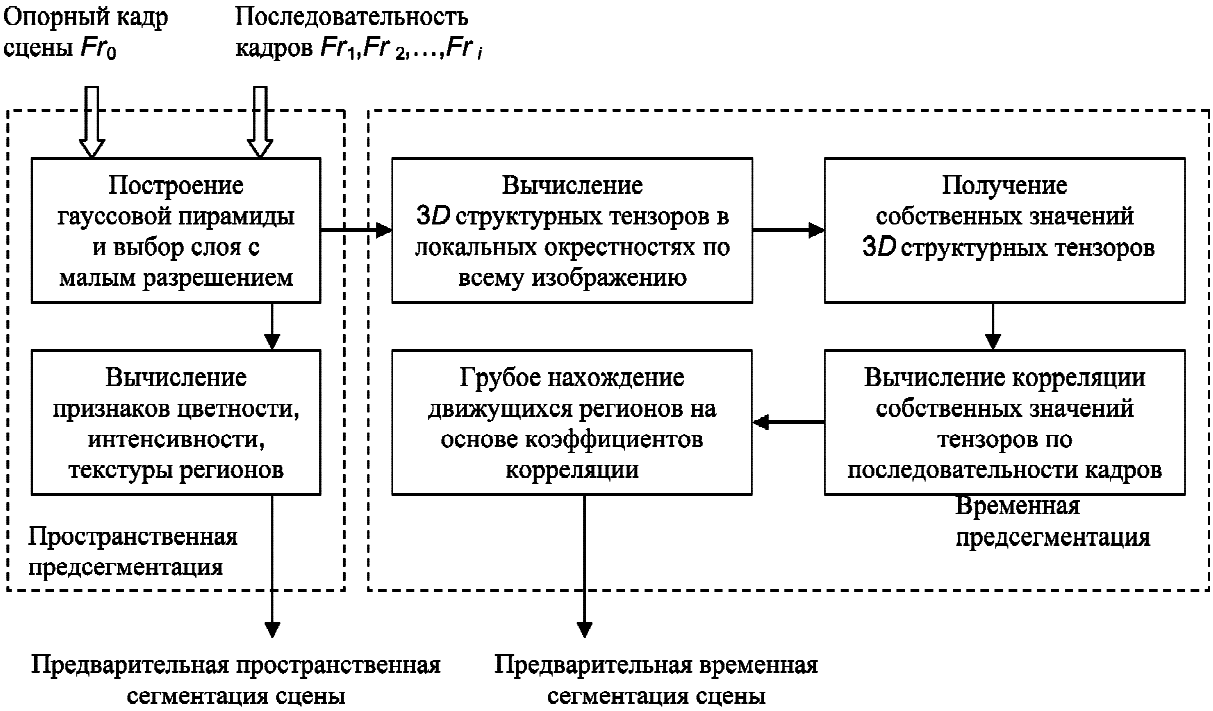

Рис. 1. Этап предварительного анализа сцены (предсегментация)

$$R = \frac{\sum_{i=1}^{N}(v_i \cdot u_i) - \left(\sum_{i=1}^{N} v_i \cdot \sum_{i=1}^{N} u_i\right)}{\sqrt{\left(\sum_{i=1}^{N} v_i^2 - \left(\sum_{i=1}^{N} v_i\right)^2\right)\left(\sum_{i=1}^{N} u_i^2 - \left(\sum_{i=1}^{N} u_i\right)^2\right)}},$$

где $v_i \in \lambda_3(I_t)$ и $u_i \in \lambda_3(I_{t+1})$; N – общее количество пикселей в кадре.

Разброс коэффициентов корреляции кадров сцены позволяет оценить степень изменчивости формы движущихся объектов. Так, для мало изменяемых по форме объектов он будет значительно меньше, чем для объектов, характеризующихся значительными изменениями их положений в пространстве. Величину такого разброса можно вычислить, используя стандартную формулу среднеквадратического отклонения:

$$S = \sqrt{\frac{1}{1-n}\sum_{i=1}^{n}\left(R_i - \overline{R}\right)^2},$$

где n – количество кадров в сцене; $\overline{R}$ – среднее значение величин R_i. Если значение величины S превышает установленное пороговое значение, то считается, что видеообъект претерпевает значительные геометрические изменения. В этом случае процедура сегментации усложняется.

На этапе пространственно-временной сегментации сцены (рис. 2) анализируется полное изображение с уче-

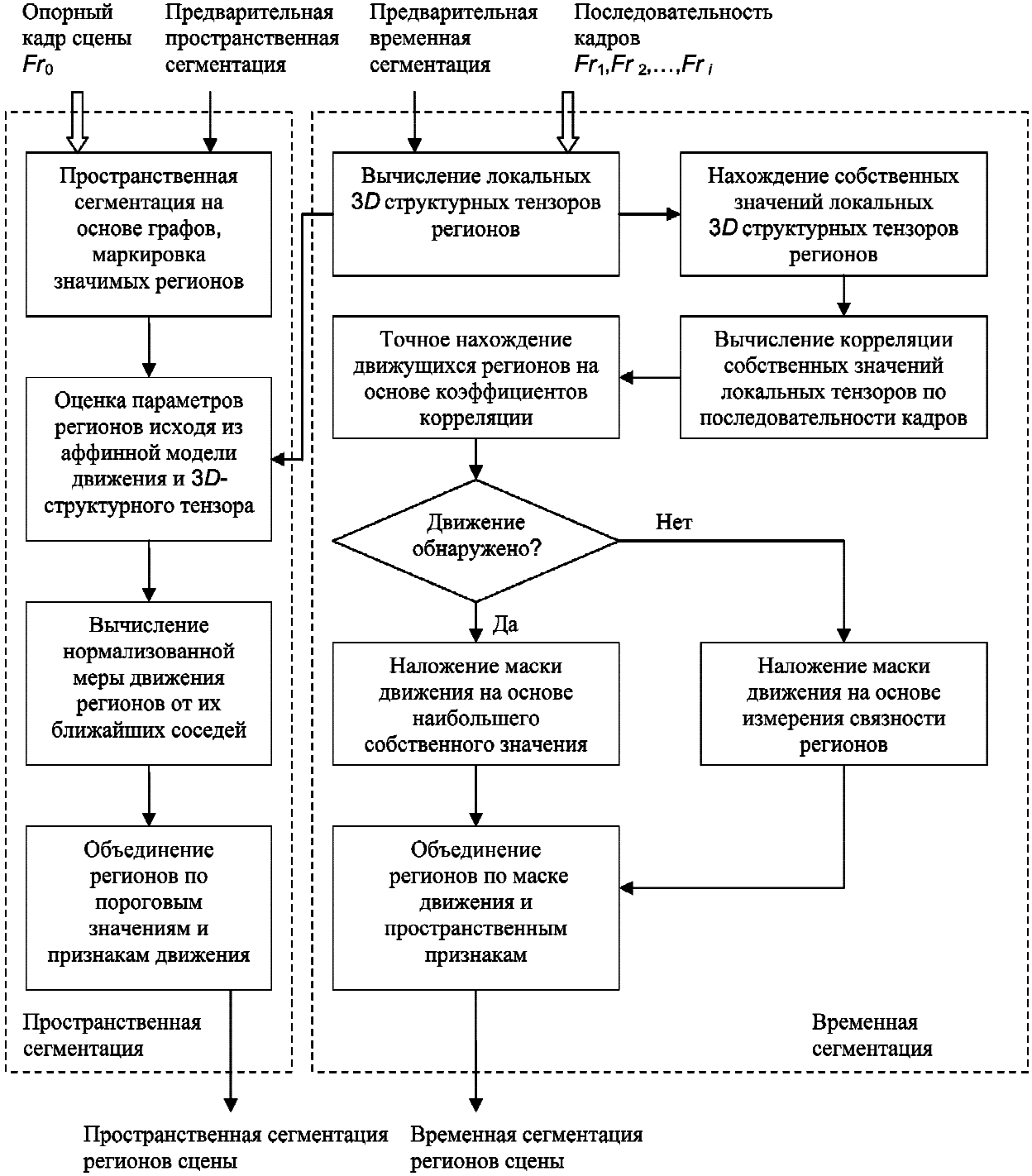

Рис. 2. Этап пространственно-временной сегментации сцены

том информации о регионах, полученной на этапе предсегментации. Можно предложить несколько стратегий сегментации в зависимости от дальнейших целей распознавания. Если обработка и распознавание должны происходить в реальном режиме времени и, например, поставлена цель слежения за движущимися объектами, то схему, приведенную на рис. 2, целесообразно применять не ко всем регионам, выявленным на этапе предсегментации, а только к тем, у которых имеются признаки движения. Если поставлена цель распознавания сцены с неизвестным заранее сюжетом, то следует воспользоваться принципами человеческого визуального восприятия: анализом больших по размерам статистически однородных регионов и движущихся регионов с последующим анализом более мелких деталей изображения. Если же происходит поиск какого-либо неподвижного объекта в видеобиблиотеках, то основное значение приобретает пространственная сегментация. Остановимся более подробно на анализе движения регионов.

Для классификации полей движения недеформируемых по геометрическим параметрам и деформируемых регионов можно ввести меры сглаженности, вычисляемые с помощью трех собственных значений $\lambda_k(I)$ регионов. Если все три собственных значения равны нулю (ранг матрицы rang($\boldsymbol{J}$) = 0), то перемещения по трем осям (x, y, t) отсутствуют, т. е. регион неподвижен. Если $\lambda_1(I) > 0$ и $\lambda_2(I) = \lambda_3(I) = 0$, то rang($\boldsymbol{J}$) = 1, что говорит о наличии изменений интенсивностей в нормальном направлении, т. е. о движении линии. Если $\lambda_1(I) > 0$, $\lambda_2(I) > 0$ и $\lambda_3(I) = 0$, то rang($\boldsymbol{J}$) = 2 и наблюдается движение с постоянной скоростью в двух направлениях пространственно-временной структуры, т. е. движение точки. В этом случае оценить параметры движения возможно. Если же все три собственных значения больше нуля, то rang($\boldsymbol{J}$) = 3. Это означает, что локальная зона находится на границе двух полей движения и оценить параметры движения невозможно.

Однако показатель ранга матрицы $\boldsymbol{J}$ нельзя напрямую применить для оценки различных типов движения, поскольку он не является нормализованной мерой для оценки поступательного, вращательного или возвратно-поступательного движения. Следует использовать понятие «мера сглаженности», на основе которого и строятся маски обнаружения движения.

Собственные значения 3D-структурных тензоров можно использовать для нахождения локальных движущихся структур, таких как границы, углы, статистически однородные регионы и т. д. В работе [4] предложены следующие выражения для определения меры сглаженности недеформируемого региона:

$$C_t = ((\mathrm{l}_1(I) - \mathrm{l}_3(I))/(\mathrm{l}_1(I) + \mathrm{l}_3(I)))^2,$$

границ деформируемого региона

$$C_s = ((\mathrm{l}_1(I) - \mathrm{l}_2(I))/(\mathrm{l}_1(I) + \mathrm{l}_2(I)))^2,$$

угловых точек деформируемого региона

$$C_C = C_t - C_s =$$

$$= \frac{4\lambda_1(I)\left(\lambda_2(I) - \lambda_3(I)\right)\left(\left(\lambda_1(I)\right)^2 - \lambda_2(I)\lambda_3(I)\right)}{\left(\lambda_1(I) + \lambda_3(I)\right)^2\left(\lambda_1(I) + \lambda_2(I)\right)^2}.$$

Как правило, в реальной сцене найденные объекты группируются по уровням движения. В общем случае необходимо отнести объект к тому или иному уровню движения в зависимости от локальных размеров регионов объекта, значений модулей скорости и ускорения и гипотез, хранящихся в базе знаний.

Обычно в сцене имеется несколько движущихся объектов. Если для сегментации регионов достаточно ввести одно пороговое значение, то к классификации многоуровневого движения следует отнестись более тщательно, учитывая не только модули значений скорости и ускорений, но и направления движений, тем самым группируя регионы по уровням движений (рис. 3). Простейшим способом является метод попиксельного вычитания кадров $D(N)$ с применением лапласиана и сравнением с заранее выбранными пороговыми значениями. При этом возможные ошибки не влияют на окончательную сегментацию, поскольку этот метод используется только для выбора масок движения.

Маски движения необходимы для устранения небольших перемещений объектов фона. Они определяются как процентное соотношение площадей движущихся регионов по методу вычитания кадров M_{fr} с использованием собственных значений локальных 3D-структурных тензоров для недеформируемых $\mathrm{M}_{\mathrm{eigen}}$ и деформируемых $\mathrm{M}_{\mathrm{corner}}$ регионов соответственно:

$$R_{\mathrm{c}} = \frac{M_{\mathrm{fr}}}{M_{\mathrm{eigen}}} 100\ \%,$$

$$R_{\mathrm{c}} = \frac{M_{\mathrm{fr}}}{M_{\mathrm{corner}}} 100\ \%.$$

Как показали эксперименты, после применения масок движения границы движущихся объектов остаются нечеткими, а внутри изображений видеообъектов имеются разрывы. Устранить эти недостатки можно с привлечением результатов пространственной сегментации. Каждый пространственный регион маркируется, и находится процентное отношение площадей движущихся регионов M_{move} и пространственных регионов M_{seg}:

$$R_m = \frac{M_{\mathrm{move}}}{M_{\mathrm{seg}}} 100\ \%.$$

Если значение R_m превышает заранее установленное пороговое значение (например, 50 %), то весь пространственный регион M_{seg} считается движущимся регионом. Однако сложные видеообъекты, как правило, состоят из нескольких движущихся регионов. В условиях отсутствия априорной информации формирование видеообъекта можно осуществлять только с использованием характеристик движения соседних регионов.

Известны два подхода к объединению регионов: непараметрический и параметрический. Непараметрический подход, основанный на слиянии границ, нельзя использовать совместно с полями движения, так как для неоднородных полей движения трудно определить точные границы. Более целесообразен параметрический подход, основанный на объединении регионов по функции минимизации энергии или по моделям движения в плоских проекциях аффинной или проективной групп. Задача ставится таким образом, чтобы, используя характеристики локальных 3D-структурных тензоров и параметров аффинной модели движения, можно было бы оценить расстояние между двумя соседними регионами.

Два региона объединяются, если вычисленное расстояние является достаточно малым, чтобы воспринимать объединенный регион как единый объект. Аффинная модель движения содержит шесть параметров:

$$v_x(x, y) = ax + by + c,$$

$$v_y(x, y) = dx + ey + f,$$

где v_x и v_y – проекции скорости $\boldsymbol{v}$ на оси OX и OY соответственно; a, b, c, d, e, f – параметры аффинной модели. Тогда проективная модель описывается восемью параметрами:

$$v_x(x, y) = a_1 + a_2x + a_3y + a_7x^2 + a_8xy,$$

$$v_y(x, y) = a_4 + a_5x + a_6y + a_7xy + a_8y^2.$$

Для дальнейших рассуждений будем пользоваться более простой аффинной моделью движения. Вектор скорости на плоскости расширим до 3D-направленного вектора $\boldsymbol{v}$ с учетом временной составляющей [4]:

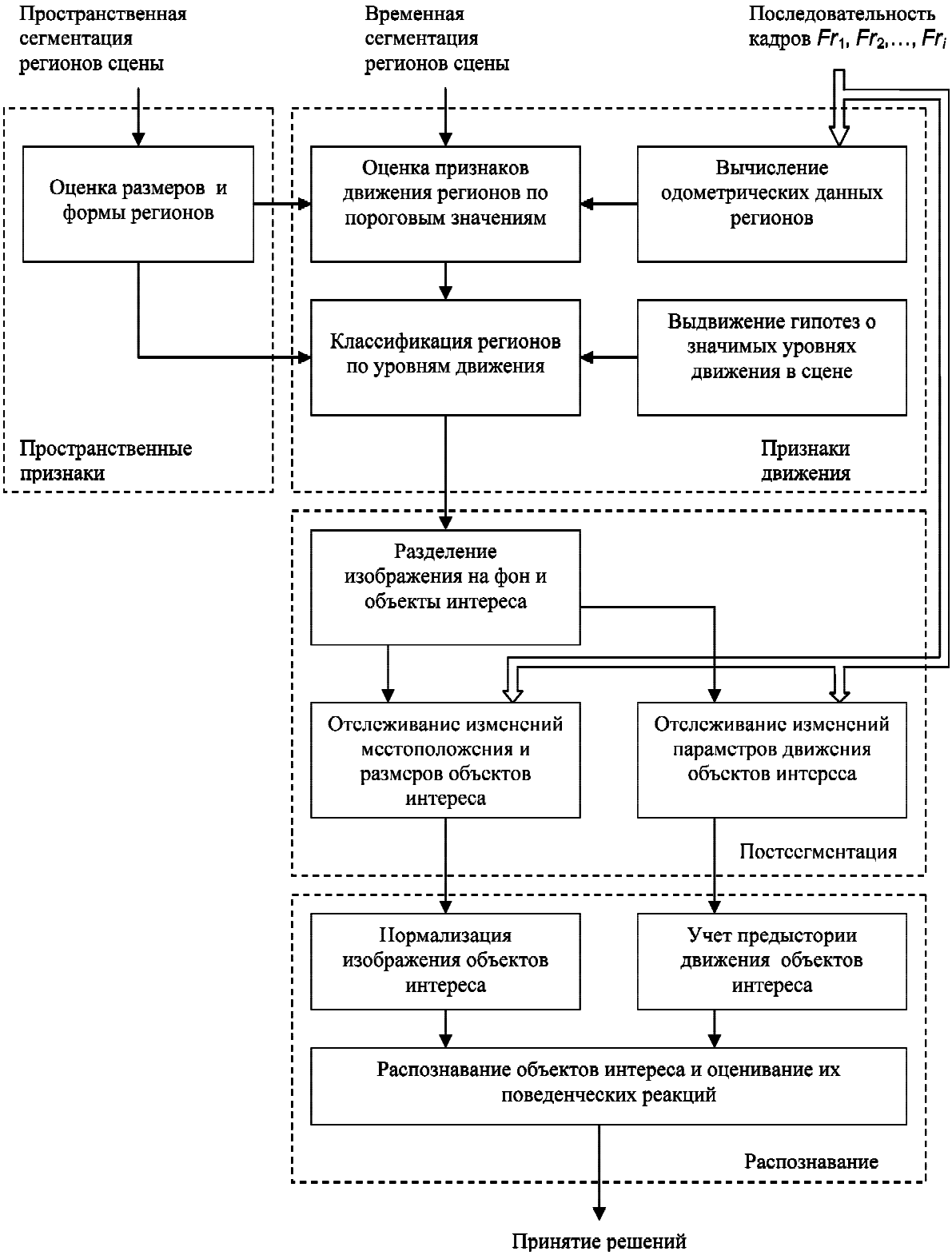

Рис. 3. Этап постсегментации (с учетом многоуровневого движения) и распознавания объектов интереса сцены

$$\boldsymbol{v} = \begin{bmatrix} v_x \\ v_y \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} x & y & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x & y & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \\ 1 \end{bmatrix} = \boldsymbol{Sp}.$$

Для объединения регионов используем функции расстояний трех видов:

$$d_1(\boldsymbol{v}_i, \boldsymbol{J}_i) = \boldsymbol{v}_i^T \boldsymbol{J}_i \boldsymbol{v}_i,$$

$$d_2(\boldsymbol{v}_i, \boldsymbol{J}_i) = \frac{d_1(\boldsymbol{v}_i, \boldsymbol{J}_i)}{|\boldsymbol{v}_i|^2},$$

$$d_3(\boldsymbol{v}_i, \boldsymbol{J}_i) = \frac{d_1(\boldsymbol{v}_i, \boldsymbol{J}_i)}{|\boldsymbol{v}_i|^2 \, tr(\boldsymbol{J}_i)}.$$

Выразим расстояние $d_1(\boldsymbol{v}_i, \boldsymbol{J}_i)$ следующим образом:

$$d_1(\boldsymbol{v}_i, \boldsymbol{J}_i) = \boldsymbol{v}_i^T \boldsymbol{J}_i \boldsymbol{v}_i = \boldsymbol{p}^T \boldsymbol{S}_i^T \boldsymbol{J}_i \boldsymbol{S}_i \boldsymbol{p} = \boldsymbol{p}^T \boldsymbol{Q}_i \boldsymbol{p},$$

где $\boldsymbol{Q}_i = \boldsymbol{S}_i^T \boldsymbol{J}_i \boldsymbol{S}_i$ – положительно определенная матрица. Сумма расстояний внутри заданного пространственного сегмента, содержащего N пикселей, определяется как

$$d_{\text{seg}}(\boldsymbol{P}) = \sum_{i=1}^{N} d_1(\boldsymbol{v}_i, \boldsymbol{J}_i) = \boldsymbol{p}^T \left(\sum_{i=1}^{N} \boldsymbol{Q}_i \right) \times$$

$$\times \boldsymbol{p} = \boldsymbol{p}^T \boldsymbol{Q}_{\text{seg}} \boldsymbol{p}.$$

Поскольку величина $\boldsymbol{Q}_i$ при расчете расстояния $d_1(\boldsymbol{v}_i, \boldsymbol{J}_i)$ более критична к большим скоростям, чем к малым, то можно использовать нормированное расстояние $d_3(\boldsymbol{v}_i, \boldsymbol{J}_i)$.

После проведения процесса постсегментации в течение нескольких последовательных кадров видеопоследовательности можно приступать к нормализации изображения объекта (приведению к эталонному виду) и формированию гипотез о поведении объекта с целью его распознавания с использованием стандартных методов кластеризации.

Таким образом, в данной статье кратко рассмотрены существующие подходы к сегментации видеообъектов, которые классифицируются по трем категориям: методы, основанные на поиске регионов, методы, основанные на определении границ, и вероятностные методы. Показано, что при сегментации сложной сцены с многоуровневыми полями движения регионов целесообразен пространственно-временной подход с применением локальных 3D-структурных тензоров. Введено понятие меры сглаженности для деформируемых регионов, для границ и угловых точек недеформируемых регионов на основе собственных значений 3D-тензора. Построены подробные структурные схемы этапов предсегментации, пространственно-временной сегментации и постсегментации сложных сцен, характеризующихся аффинной моделью движения. Введены три функции расстояний на базе локальных структурных тензоров с целью объединения регионов в видеообъект и разделения видеообъектов сцены по уровням движения.

Библиографический список

1. Bresson, X. A Variational Model for Object Segmentation Using Boundary Information and Shape Prior Driven by the Mumford-Shah Functional / X. Bresson, P. Vandergheynst, J.-P. Thiran // Intern. J. of Computer Vision. 2006. Vol. 68, № 2. P. 145–162.

2. Cavallaro, A. Shadow-aware object-based video processing / A. Cavallaro, E. Salvador, T. Ebrahimi // IEEE Vision, Image and Signal Proc. 2005. Vol. 152, Iss. 4. P. 14–22.

3. Thirde, D. Spatio-Temporal Semantic Object Segmentation using Probabilistic Sub-Object Regions / D. Thirde, G. Jones, J. Flack // British Machine Vision Conf. Norwich, UK, 2003. P. 163–172.

4. Wang, H.-Y. Spatio-Temporal Video Object Segmentation via Scale-Adaptive 3D Structure Tensor / H.-Y. Wang, K.-K. Ma // EURASIP : J. on Applied Signal Proc. 2004. Vol. 6. P. 798–813.

M. N. Favorskaya

DETECTION OF MOVING VIDEO OBJECTS BASED ON LOCAL 3D-STRUCTURE TENSORS

The analysis of video objects detection based on spatio-temporal approach is given. It covers the term of 3D-structure tensor for effective definition of spatio-temporal motion local orientation. The detailed schemes of pre-segmentation, spatio-temporal segmentation, and post-segmentation with further video objects recognition using local 3D-structure tensors are built.

Keywords: spatio-temporal segmentation, 3D-structure tensor, moving estimation.