Системы обнаружения, сопровождения и кластеризации объектов на основе нейроноподобного кодирования

Беллюстин Н.С., Калафати Ю.Д., Ковальчук А.В., Тельных А.А.,
Шемагина О.В., Яхно В.Г.

В статье описываются алгоритмы, основанные на нейроноподобном кодировании, позволяющие решать как задачу обнаружения в видеопотоке произвольного объекта на сложном фоне, так и задачу его последующего сопровождения. Решение этих задач позволяет создать систему видеонаблюдения, обладающую способностью не только обнаруживать объекты, но и анализировать полученную видеоинформацию, формировать обобщенный информативный отчет о наблюдаемых событиях.

1. Построение системы видеонаблюдения с расширенными возможностями.

Под расширенными возможностями системы видеонаблюдения будем понимать программно-аппаратный комплекс, позволяющий вести видеонаблюдение на объектах с небольшим (до 10 человек) числом наблюдаемых лиц, предназначенный для оперативного оповещения пользователя с помощью SMS о появлении в поле зрения камер движения или лица. Анализ видеоинформации позволяет оповещать пользователя только о качественно новых событиях (например, появление в поле зрения лица другого человека), что повышает информативность оповещений. Комплекс состоит из нескольких видеокамер и блока обработки видеоинформации. Блок обработки видеоинформации включает в себя описанную ниже систему обнаружения объектов на сложном фоне, систему сопровождения найденных объектов, а также систему принятия решений относительно идентичности найденных на разных кадрах объектов (лиц одного и того же человека) (Рис. 1).

Такие системы видеонаблюдения могут быть применимы, например, в области создания охранных систем небольшого масштаба.

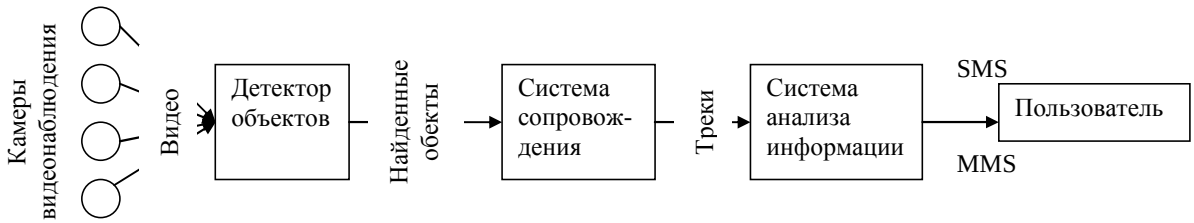

Рис.1 Принципиальная схема системы видеонаблюдения с расширенными возможностями.

2. Использование модели однородной двумерной нейроноподобной среды для обнаружения местоположения объектов на сложном фоне.

Основным этапом обнаружения объектов на сложном фоне является принятие решения о наличии в рассматриваемом фрагменте изображения искомого объекта. В общем случае это бинарная функция, которая переводит произвольный фрагмент $S_{w,h}:\begin{cases}-w/2 \le x-x_0 \le w/2\\ -h/2 \le y-y_0 \le h/2\end{cases}$ в пространство состояний $\{0,1\}$.

$$\Xi : S_{w,h} \to \{0,1\} \qquad (1),$$

где 1 – соответствует факту обнаружения искомого объекта, а 0 - факту его отсутствия в рассматриваемой области. Принципы построения подобной функции широко освещены в литературе, поскольку данное направление изучается уже несколько десятков лет. В качестве примеров можно привести принятие решения на основе геометрических характеристик исследуемого фрагмента изображения [1-3], на основе методов PCA [4,5], построение оптимальных шаблонов с использованием модифицированной меры Хаусдорфа[6,7], использования нейросетевых подходов для принятия решения[8,9], на основе простых признаков Хаара [10] и так далее.

Как показано в работах [11-14], использование модели однородной среды, состоящей из нейроноподобных элементов, позволяет производить простейшее кодирование изображения. Для решения задачи распознавания фрагмента изображения, сформулированной в виде (1), рассмотрим модель рецептивного поля, записанную в виде

$$u'_{i,j} = F\left[-T + \alpha \sum_{k=-M}^{M} \sum_{l=-M}^{M} \Phi_{k,l} u_{i+k,j+l} + u^{ex}_{i,j}\right] \quad (2),$$

где $u'_{i,j}$ - активность рецептивного поля, F - нелинейная функция активации рецептивного поля; $\Phi_{k,l}$ - функция связи между элементами нейроноподобной системы, $u_{i+k,j+l}$ -

входной стимул на элементы нейроноподобной системы, $u_{i,j}^{ex}$ - дополнительные внешние сигналы, которые приходят на элементы нейроноподобной системы; α, T - нормировочные коэффициенты; $_{i,j}$ - местоположение в пространстве нейроноподобного элемента; k,l - расстояние от рассматриваемого элемента рецептивного поля до взаимодействующего с ним нейроноподобного элемента окрестности; M - радиус окрестности, в которой происходит взаимодействие между нейроноподобными элементами.

Из вида модели следует, что она формирует активность одного единственного рецептивного поля, собирая информацию с некоторой его окрестности, которая определяется видом функции связи Φ и ее параметрами. В рамках нашего подхода мы будем использовать различные виды функции пространственной связи Φ (Рис.2) для формирования откликов на простые стимулы, такие как линии заданных направлений, светящиеся или наоборот темные объекты различных масштабов.

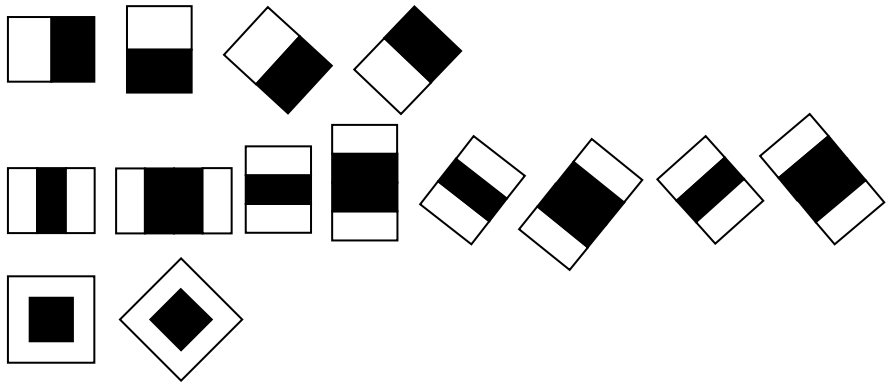

Рис.2 Виды функции пространственной связи Φ

Таким образом, для моделирования процесса первичной обработки информации выражение (2) перепишем в следующем виде:

$$\left\{u'_{i,j}\right\}^K = F^K\left[-T + \alpha \sum_{k=-M}^{M} \sum_{l=-M}^{M} \Phi^K{}_{k,l} u_{i+k,j+l} + u_{i,j}^{ex}\right], K = 1....N \qquad (3),$$

смысл которого заключается в следующем. Для данной точки исходного изображения формируется вектор активности рецептивных полей различного типа. Количество всевозможных рецептивных полей определяется параметром K. В самом простом случае вектор образуется путем изменения какого-нибудь параметра функции связи, например, ее пространственного масштаба M. В общем случае вектор образуется путем изменения масштаба пространственной связи и ее типа. В этом случае образуется целый набор карт активности рецептивных полей различного типа.

Если мы возьмем фрагмент размером 24x24 пикселя, то количество рецептивных полей, участвующих в процессе кодирования такого фрагмента, будет исчисляться сотнями тысяч. Но, подав на вход системы (3) стимул в виде лица человека (Рис.3а), можно увидеть, что не все рецептивные поля одинаково важны для принятия решения о наличии в рассматриваемой области искомого объекта.

а б

Рис. 3 Карта активности рецептивных полей (б) для входного стимула в виде лица (а); белый цвет соответствует активности, серый цвет соответствует отсутствию всякой активности, а черный - активности инвертированной функции связи Ф.

Таким образом, решающую функцию (1) можно искать в следующем виде:

$$\Xi = \begin{cases} 1, \sum_{i=1}^{N} \omega_i h_i(u_i') \geq \Theta \\ 0, в\ противном\ случае \end{cases} \quad (4)$$

где N – общее число всех рецептивных полей всех типов, ω_i - вес рецептивного поля, h_i - функция активации данного рецептивного поля, в случае если внутри него оказывается искомый объект, u_i' - активность рецептивного поля (1), Θ - порог принятия решения. Будем считать, что h_i - бинарная функция следующего вида:

$$h_i = \begin{cases} 1, \Psi_i(u_i') \in \Omega \\ 0, в\ противном\ случае \end{cases}, \quad (5)$$

где Ω - область значений u_i', в которой функция Ψ_i принимает значение, равное 1. Анализируя (3) можно заметить, что N – общее число рецептивных полей всех типов, которые только возможны в заданной апертуре. Например, апертура 24x24 пикселя содержит порядка 100000 рецептивных полей Хаара, очевидно, что такое количество полей не может быть проанализировано в реальном времени на существующих вычислительных мощностях. С другой стороны, представляет немалую трудность процесс оптимизации параметров выражения (3), в котором надо найти порядка 100000 функций активации h и их весов, минимизируя, например, ошибку распознавания по большой базе данных, состоящей из миллиардов фрагментов. Все эти соображения приводят к следующему выводу: следует ограничить число используемых в выражении (3) рецептивных полей, при этом добившись высоких результатов распознавания на

относительно небольшом числе обучающих примеров. Полученная в результате подобной оптимизации структура H будет называться сильным классификатором[10] и обладать следующими свойствами:

- Пропускать все объекты заданного типа (полезный сигнал)
- Подавлять некоторые объекты фона (шумовой сигнал) с высокой эффективностью.

Множество подобных сильных классификаторов $\{H\}$, состоящее из N элементов, каждый из которых обладает вышеперечисленными свойствами, назовем «каскадом сильных классификаторов» и сформулируем следующее правило принятия решения: *если все сильные классификаторы из рассматриваемого множества были активированы исходным сигналом, то данный сигнал содержит искомый объект*.

$$\Xi = \mathrm{H}_0 \wedge \mathrm{H}_1 \wedge ... \wedge \mathrm{H}_{N,} \quad (6),$$ где N- число сильных классификаторов.

Подробно процедура построения каскада сильных классификаторов описана в [10,15].

В результате нами был получен детектор лиц с ошибкой ложного обнаружения (FAR) $10^{-6} - 10^{-7}$ на фрагмент. Размер прецедентной базы, на которой производилась оптимизация параметров, составлял 3000 фронтальных изображений лица человека.

Описанный выше каскад сильных классификаторов был нами использован в качестве детектора лиц людей, попадающих в поле зрения камер видеонаблюдения.

На первом этапе детектор лица работает только в режиме обнаружения: сканирование проводится по всему изображению, детектор в бинарном режиме определяет лицо или не лицо на каждом фрагменте, с последующей кластеризацией накладывающихся друг на друга фрагментов с объектом [15]. Области с найденным объектом «передаются» системе сопровождения [16], позволяющей спрогнозировать положение объектов на последующих кадрах. В этих областях детектор работает не в бинарном режиме (лицо – не лицо), а в количественном. Результатом работы детектора в этом режиме является значение «качества», вычисляемое по промежуточным результатам обработки - суммируются рейтинговые веса всех слабых классификаторов [10] во всех каскадах детектора, проголосовавшие за рассматриваемый фрагмент. Начиная с того кадра, где был найден первый объект, детектор работает одновременно в двух режимах: количественном в областях с прогнозируемым местоположением лица и в бинарном для всех остальных фрагментов изображения, что необходимо для поиска новых лиц.

3. Система сопровождения найденных объектов

Задачей системы сопровождения объектов является повышение качества обнаружения объекта, а также повышение скорости обработки видеоданных.

Это означает, что в рамках задачи построения «интеллектуальной» системы видеонаблюдения к системе сопровождения предъявляются следующие требования:

- объект, найденный детектором, должен удерживаться в любом ракурсе - потеря объекта происходит только в случае его выхода за пределы кадра;
- скорость обработки видео размером 320x240 пикселей должна быть не меньше 20-25 кадров в минуту.

Для решения задачи сопровождения нами был использован следующий алгоритм.

Результатом работы детектора лиц является фрагмент или набор фрагментов исходного изображения (видеокадра), для которых принято решение, что это искомый объект - лицо человека.

Каждый из этих фрагментов разбивается на 64 равные части, как показано на рис.2

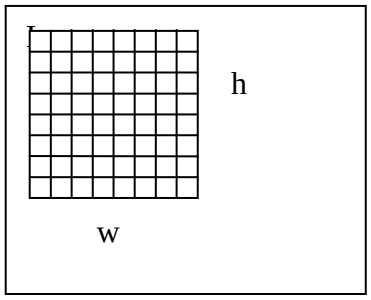

Рис.2 Способ разбиения фрагмента изображения

В качестве вектора признаков, описывающего фрагмент, выберем вектор средних яркостей, вычисленных в каждой из 64 ячеек. Среднюю яркость ячейки определим как $< I_{яч} >= \frac{64}{w \cdot h} \sum_{x=x_o}^{x=x+w/8} \sum_{y=y}^{y=y+h/8} I(x, y)$ (7), где w – ширина фрагмента, h – высота фрагмента, $I(x, y)$ - яркость пикселя с координатами x,y. Начальный вектор признаков запоминается и хранится в качестве эталона.

Для предсказания местоположения объекта на следующем кадре нами использовался алгоритм LucasKanadeFeature Tracker, предложенный корпорацией Интел (библиотека OpenCV).

Для каждого нового местоположения объекта вычисляем набор векторов признаков в некоторой его окрестности $D_{m,n}$.

$$D_{m,n}:\begin{cases} -i*w/32 - w \le x - x_0 \le w + i*w/32, i = 1,2,...,m \\ -j*h/32 - h \le y - y_0 \le h + j*h/32, j = 1,2,...,n \end{cases} \quad (8),$$

где x_0, y_0 - координаты центра исходного прямоугольника.

Понятно, что при уменьшении m и n ошибка системы слежения будет возрастать, но при этом время вычислений будет уменьшаться, что немаловажно в условиях обработки видео в режиме реального времени. Поэтому значения m и n выбираются из соображений оптимального соотношения уровня ошибки и времени вычислений. В наших экспериментах $m = n = 10$. Сравниваем полученный набор векторов признаков с эталонным, в качестве меры близости используем эвклидово расстояние между векторами в данном пространстве признаков. В результате в качестве нового местоположения объекта выбирается область, расстояние от вектора признаков в которой минимально по отношению к эталонному вектору.

Результатом работы системы слежения является набор треков. Трек представляет собой непрерывную последовательность кадров с выделенной на них областью лица человека.Эта область характеризуется следующими параметрами: x, y, w, h - определяют местоположение лица в кадре, $rate$ - оценка «качества», то есть степени уверенности при детектировании лица, ID - уникальный идентификатор лица, в пределах одного трека идентификатор одинаков для всех объектов (лиц), $Frame\_number$ - номер кадра в видеопоследовательности, на котором было обнаружено данное лицо.

4. Система анализа видеоинформации

Задачей системы анализа видеоинформации является повышение информативности посылаемых пользователю уведомлений. Так, если наблюдаемый объект пропал из поля зрения камеры, а затем вновь появился, то, при условии, что сообщение высылается при каждом обнаружении нового объекта, пользователь получит 2 сообщения о нахождении двух разных людей в поле зрения камер независимо от временного интервала ΔT между этими событиями. Но если этот интервал достаточно мал, то можно предположить, что на двух соседних треках присутствует один и тот же объект. Проверка этого предположения поможет увеличить информативность получаемых пользователем сообщений.

Для сравнения двух треков нами был предложен следующий алгоритм.

Используя информацию о местоположении, области лица вырезаются на каждом кадре трека и приводятся к стандартизованному виду. В нашем случае изображения приводятся к размеру 32x32 пикселя, нормируются по освещенности и поворачиваются так, чтобы глаза были расположены горизонтально (Рис.3).

Рис.3 Пример стандартизованного изображения

С использованием метода главных компонент [4] и [9] на прецедентной базе, содержащей порядка 1000 фронтальных изображений лиц, было создано подпространство невысокой размерности L. Все фрагменты трека, содержащие область лица, проецируются на это пространство – после чего каждое лицо характеризуется вектором размерности L. Таким образом, теперь каждый трек представляет собой набор векторов размерности L:

$$\left\{\vec{k}_1, \vec{k}_2, ..., \vec{k}_N\right\} \qquad (9),$$

где $\vec{k}_i$ - проекция $i-го$ кадра трека, N - количество кадров в данном треке. Кроме того, как уже было сказано выше, каждый элемент трека характеризуется значением параметра $rate_i$ - оценкой «качества» $i-го$ кадра, вычисленной при помощи детектора лиц.

Для того чтобы ответить на вопрос, принадлежат ли изображения лиц, представленные на двух соседних по времени треках, одному и тому же человеку, необходимо ввести меру близости треков. Для этого вычислим среднее (10) и дисперсию (11) расстояний между элементами внутри трека, а также среднее (12) и дисперсию (13) расстояний между элементами разных треков:

$$\bar{d}_{in} = \frac{2}{N_{1q} * (N_{1q} - 1)} \sum_{i=1}^{N_{1q}} \sum_{j=i+1}^{N_{1q}} d_{ij} \qquad (10)$$

$$D[d_{in}] = \frac{2}{N_{1q} * (N_{1q} - 1)} \sum_{i=1}^{N_{1q}} \sum_{j=i+1}^{N_{1q}} d^2{}_{ij} - \bar{d}^2{}_{in} \qquad (11)$$

$$\bar{d}_{ex} = \frac{2}{N_{1q} * (N_{2q} - 1)} \sum_{i=1}^{N_{1q}} \sum_{j=i+1}^{N_{2q}} d_{ij} \qquad (12)$$

$$D[d_{ex}] = \frac{2}{N_{1q} * (N_{2q} - 1)} \sum_{i=1}^{N_{1q}} \sum_{j=i+1}^{N_{2q}} d^2{}_{ij} - \bar{d}^2{}_{ex} \qquad (13),$$

где d_{ij} - эвклидово расстояние между проекциями $i-го$ и $j-го$ кадров, N_{1q} и N_{2q} - количество элементов первого и второго треков соответственно, у которых оценка «качества» больше некоторого порогового значения $rate_q$. Высокое значение этой оценки означает, что ракурс найденного на этом кадре лица близок к фронтальному, поскольку

для построения каскада сильных классификаторов использовалась прецедентная база фронтальных изображений лица. Это означает, что в процессе сравнения треков участвуют только те элементы трека, на которых ракурс лица близок к фронтальному. Кроме того, треки, у которых хотя бы для одной пары элементов совпадает значение номера кадра $Frame\_number$ в видеопоследовательности, считаются принадлежащими разным людям.

Будем считать, что два трека содержат лицо одного и того же человека, если одновременно выполняются следующие условия:

$$\begin{cases} \left|\bar{d}_{ex} - \bar{d}_{in}\right| < K * \sqrt{D[d_{in}]} \\ \left|\bar{d}_{ex} - \bar{d}_{in}\right| < K * \sqrt{D[d_{ex}]} \end{cases} \qquad (14),$$

где K - коэффициент, позволяющий регулировать жесткость этого условия.

5. Выводы

В данной работе предложены наборы нейроноподобных алгоритмов обнаружения и сопровождения найденного объекта, которые позволяют удерживать объект даже при условии значительных искажений его характеристик. Применительно к сопровождению лица человека такой набор алгоритмов позволяет удерживать лицо в любом ракурсе.

Кроме того, показана возможность, повышения информативности системы видеонаблюдения за счет слияния последовательностей видеокадров.

Список литературы

1. *W.W. Bledsoe*, The Model Method in facial recognition. Panoramic Research Inc., Palo Alto, CA, Rep. PRI:15 Aug 1966
2. *T.Kanade*, Picture Processing System by computer complex and recognition of human faces. Dept. of information science, Kyoto University, Nov. 1973
3. *S.Carey and R.Diamond*, From piecemeal to configurational representation of faces. //Science, Vol.195. Jan. 21,1977, 312-13

4. *M.Turk and A. Pentland*. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71-86, 1991.
5. *B. Moghaddam and A. Pentland*. Face recognition using view-based and modular eigenspaces. In *Automatic Systems for the Identification of Humans, SPIE*, volume 2277, 1994
6. *Oliver Jesorsky, Klaus J. Kirchberg, and Robert W. Frischholz* Robust Face Detection Using the Hausdorff Distance, In Proc. Third International Conference on Audio- and Video-based Biometric Person Authentication, Springer, Lecture Notes in Computer Science, LNCS- 091, pp. 90–95, Halmstad, Sweden, 6–8 June 2001.
7. *M.P. Dubuisson and A.K. Jain*. A modified Hausdorff distance for object matching.In ICPR94, pages A:566–568, Jerusalem, Israel, 1994.
8. *Aizenberg I. N., Aizenberg N. N. and Krivosheev G.A.* Multi-valued and Universal Binary Neurons: Learning Algorithms, Applications to Image Processing and Recognition // Lecture Notes in Artificial Intelligence - Machine Learning and Data Mining in Pattern Recognition, 1999. P. 21-35.
9. *С.Хайкин*. Нейронные сети. Полный курс. //Издат. Дом «Вильямс» 2006
10. *Viola P & Jones M*. Rapid Object Detection using boosted cascade of simply classifiers. //2001
11. *Нуйдель И.В., Кузнецов С.О.*, Использование однородных нейроноподобных сред для обработки изображений. // Изв. Вузов. Радиофизика. 1994 Т.37. N8. Сю 1053-1061
12. *Kuzhetsov S.O. Nuildel I.V., Yakhno V.G*. Segmentation and Pattern Recognition of a Composite Image Product by System of Elements with Neural-network Architecture // In book: Neurocomputers and Attention. Ed. A. Holden, A. Kryuk. Manchester University Press, 1991. P591-596.
13. *Яхно В.Г., Беллюстин Н.С., Красильникова И.Г., Кузнецов С.О., Нуйдель И.В., Панфилов А. И., Перминов А.О., Шадрин А.В., Шевырев А.А.*, Исследовательская система принятия решений по фрагментам сложного изображения, использующая нейроноподобные алгоритмы. // Изв. Вузов. Радиофизика. 1994. Т. 37. N8, С.961-986
14. *Bellustin N.S., Kuznetsov S.O., Nuidel I.V., Yakhno V.G.*, Neural Networks with Close Nonlocal Coupling for Analyzing Composite Imges// Neurocomputing. V.3. 1991. P 231-246
15. *Беллюстин Н.С., Калафати Ю.Д., Ковальчук А.В., Тельных А.А., Шемагина О.В., Яхно В.Г.*, Нейроподобный детектор лица. Технические особенности реализации и

обучения, X Всероссийская научно-техническая конференция "Нейроинформатика-2008", Сборник научных трудов, часть 2 МИФИ, Москва, янв. 2008, с. 123-132

16. *J.- Y.Bouguet*, Pyramidal Implementation of the LucasKanadeFeature Tracker: Description of the algorithm, Technical Report, Intel Corporation: Microprocessor Research Labs, 2000, OpenCV documentation.