Адаптивный метод распознавания динамических жестов

Бизюкин Г.А., МГТУ им. Баумана
uashir@mail.ru
Майков К.А., МГТУ им. Баумана
maikov@bmstu.ru

Аннотация

Разработаны требования к жестовым интерфейсам общего назначения. Проведен анализ текущего состояния вопроса в области распознавания динамических жестов. Описаны этапы распознавания. Предложена модификация метода условных случайных полей со скрытыми состояниями, позволяющая повысить дискриминативные способности данной модели. Проведено исследование модели на алгоритмическом уровне. Установлено, что данная модификация алгоритмически реализуема и обладает сублинейной временной сложностью.

1 Введение

В связи с интенсивным развитием в последние годы технологий виртуальной реальности, а также с исследованиями в области естественных человеко-машинных интерфейсов, возрос интерес разработчиков и исследователей к жестовым способам системами обработки информации.

В качестве генератора жестов может выступать та или иная часть человеческого тела, изменяющая свою конфигурацию в пространстве. Чаще всего отслеживаются движения рук, в частности кистей или пальцев.

В контексте человеко-машинного взаимодействия жесты разделяют на статические (позы) и динамические. Поза — это одномоментная конфигурацию генератора жестов. Динамический жест является последовательностью поз, т.е. последовательностью сменяющих друг друга во времени конфигураций.

Задача распознавания статических жестов хорошо разработана и имеет ряд практических решений, которые обладают достаточным уровнем качества и могут быть использованы для построения жестовых интерфейсов [1-3].

Задача распознавания динамических жестов не решена в полном объеме и является перспективной и актуальной научно-технической задачей, потому что динамика обладает большей информационной емкостью. Интерфейсы на основе динамических жестов обладают большей интерактивностью и скоростью получения информации от пользователя.

Специфика современных интерфейсов общего назначения на основе динамических жестов для систем виртуальной реальности и персональных компьютеров определяет ряд требований к их реализации:

- способность учитывать частоту изменения конфигурации генератора жестов, превышающую 3 Гц;
- работа в режиме реального времени;
- возможность создания базы полностью персонализированных жестов для каждого пользователя;
- сохранение требуемой точности распознавания для наборов, содержащих более чем 30 динамических жестов.

В связи с тем, что существующие методы обладают теми или иными функциональными ограничениями, данная область требует исследований и разработки новых методов.

Целью данной работы является разработка метода распознавания динамических жестов, способного соответствовать поставленным требованиям.

В задачи исследования входит анализ существующих методов, выбор наиболее подходящего прототипа, его модификация на теоретическом уровне и алгоритмическое исследование.

2 Этапы распознавания

Задача распознавания динамических жестов относиться к области распознавания образов. Каждый жест несет информацию и передает некоторое сообщение, таким образом, он является сигналом.

Задачу распознавания можно разбить на три основных этапа:

- регистрация сигнала;
- обработка и выделение признаков;
- распознавание (классификация).

Каждый этап использует разные классы методов и технологий для того, чтобы получить некоторое промежуточное представление и передать его на следующий этап.

3 Регистрация жеста

В зависимости от того, какая аппаратная платформа используется, способы регистрации жеста можно разделить на три категории.

Видео камеры. Используется одна или несколько видеокамер в видимом или инфракрасном диапазоне. По типу используемых камер выделяют следующие подкатегории:

- монокулярные камеры. Одна, чаще всего цветная камера с одним объективом;
- бинокулярные камеры. Применяют два объектива для получения стереоизображения;
- инфракрасные камеры. Используются совместно с инфракрасным проектором, который проецирует когерентный свет определенной структуры на область совершения жеста. Камера воспринимает отраженный свет и с помощью метода триангуляции строит картину глубины;
- времяпролетные камеры. Измеряют дальность через скорость света, замеряя время пролета светового сигнала, испускаемого камерой, и отраженного каждой точкой получаемого изображения;
- смешанный подход. Использует одновременно несколько типов камер.

Датчики и сенсоры. Применяется ряд датчиков или сенсоров [2], которые требуют непосредственного контакта с генератором жестов. Выделяют четыре подкатегории:

- сенсорные экраны. Поверхности, которые способны распознать место, продолжительность и силу касания;
- механические перчатки и датчики; Устройства, чаще всего напоминающие перчатки и фиксирующие жест за счет своей механической конфигурации;
- гироскопы и акселерометры. Регистрирую жест путем реагирования на изменение положения датчика в пространстве.
- биомеханические сенсоры. Считывают биомеханические показатели тела для того, чтобы получить информацию о жесте.

Радары. Используют радио сигнал для регистрации изменений в конфигурации генератора жестов.

Подходы на основе датчиков и сенсоров позволяют получить наибольшую точность регистрации сигнала, тем не менее, они накладывают жесткое ограничение, которое требует непосредственного контакта с генератором жестов.

Методы на основе видео камер и радаров позволяют работать с интерфейсом на расстоянии, но требуют использования более сложных методов, которые способны выделить жест на общем фоне, учесть и компенсировать возможные помехи и искажения.

Наиболее распространенным жестовым интерфейсом в настоящий момент является интерфейс мобильных устройств, который использует сенсорные экраны совместно с акселерометрами и гироскопами. Наиболее перспективным — на основе камер и радаров.

4 Обработка и выделение признаков

Из анализа работ [2, 3], следует, что задача выделения характеристических признаков хорошо разработана и для её решения существует ряд готовых программно-технических реализаций.

Наиболее практически развитые и популярные реализации построены на безе собственной аппаратной платформы.

Среди технологий на основе видео камер выделяют: Leap Motion [4], RealSense [5], Kinect 2 [6], Orbbec [7]. В основу реализации заложены различные методы фильтрации и логической обработки изображения. Обзор и классификацию данных групп методов можно найти в работах [1, 2].

Технологии Mio [8] и Gest [9] применяют подход на базе датчиков и сенсоров.

Project Soli [10] использует радио сигнал для захвата информации о характеристиках и динамике совершаемого жеста.

Каждая реализация предоставляет готовый интерфейс прикладного программирования, который позволяет получить в режиме реального времени численные характеристики, описывающие текущую конфигурацию генератора жестов с достаточной точностью.

5 Распознавания (классификация)

Основная сложность распознавания заключается в выборе подходящего классификатора, который соответствовал бы всем требованиям решаемой задачи.

В работах [2, 3, 11] показано, что подход с использованием моделей машинного обучения является наиболее подходящим при распознавании динамических жестов.

В настоящий момент с точки зрения точности распознавания и возможности работы в режиме реального времени, можно выделить генеративные и дискриминативные вероятностные модели.

Генеративные модели получили широкое применение как в распознавании динамических жестов, так и в других областях, таких как автоматическое распознавание речи и анализ естественных языков. Тем не менее, они обладают существенным недостатком, который заключается в предположении о том, что наблюдения независимы между собой. При распознавании жестов это предположение нельзя считать корректным [12].

Дискриминативные модели лишены этого недостатка и позволяют учитывать контекст совершения жеста. Помимо этого, они обладают преимуществом, которое позволяет обеспечить более высокую точность, чем генеративные модели, при меньшем объеме обучающей выборке выборки, что является актуальным для создания персонализированных баз жестов [12].

Среди генеративных моделей распознавания, как наиболее популярный, выделяют метод скрытых марковских моделей. Среди дискриминативных моделей лучше всего себя показывает метод условных случайных полей и его модификации.

Из работ [11, 12, 13] следует, что условные случайные поля превосходят скрытые марковские модели и являются наиболее перспективным методом распознавания динамических жестов. В частности, наибольшую точность распознавания удалось получить, используя модифицированные за счет срытых состояний условные случайные поля [13].

Как показано в [12], при работе с большими базами баз жестов имеет место потеряет точность распознавания.

Данный недостаток можно устранить путем добавления нелинейной сигмоидальной функции активации.

6 Теоретическое описание и алгоритмическое исследование модифицированной модели

В качестве входных значений выступает последовательность $X=\{x_1, ..., x_T\}$ длины T, которая упорядочена во времени. Длина может сильно варьироваться в зависимости от исходных данных. Каждый элемент последовательности представляет собой характеристический вектор $x_t \in R^D$ размерности D, который описывает одномоментную конфигурацию генератора жестов.

Каждая последовательность отображается в метку класса $y \in Y$, где Y — множество определенных ранее жестов.

Для распознавания требуется сопоставить наблюдаемую последовательность с меткой, соответствующего ему класса. Согласно определению условных случайных полей со скрытыми состояниями [13, 14] условная вероятность данного события определяется как

$$p(y \mid X; W)=\frac{1}{Z(X; W)}\sum_h e^{F(y, h, X; W)}.$$

W — вектор параметров модели, $h \in H$, где H есть множество скрытых состояний, $Z(X;W)$ — коэффициент нормализации, а $F(y,h,x;w)$ — функция признаков.

Коэффициент нормализации рассчитывается как

$$Z(x;W)=\sum_{y'}\sum_h e^{F(y,h,x;w)}.$$

В свою очередь, функция признаков определяется в виде

$$F(y,h,x;w)=\sum_t f^1(h,x,t;w)+\sum_t f^2(y,h,t;w)+$$

$$+\sum_t f^3(y,h,t,t+1;w).$$

Пусть *I* — индикаторная функция, такая, что $I(x)=\begin{cases}1, \text{если x — истина}\\0, \text{если x — ложь}\end{cases}$, тогда второе и третье выражение в функции признаков можно выразить как

$f^2(h,x,t;w)=w_{y,h}I(y=y')I(h_t=h')$ и

$f^3(h, X,t;W)=w_{y,h,h}I(y=y')I(h_t=h')I(h_t=h'')$, где $y' \in Y$, h' и $h'' \in H$, а $w_{y,h}, w_{y,h,h}$ — соответствующие веса.

Для выделения в пространстве признаков сложной формы, введем множество функций активации G и функцию $\psi_g(x,t;W)$ для того, чтобы определить f^1 как

$$\psi_g(x,t;W)=\frac{1}{|r(x_t)|}\sum_{x'\in r(x_t)} g\left(\sum_d w_{g,d}x'_d\right).$$

В качестве функции активации выступает сигмоидальная функция активации $g(x) = 1/(1+\exp(-x))$. Таким образом,

$$f^1(h,x,t;w)=\sum_{g\in G} w_{g,h} I\left(h_t=h'\right)\psi_g(x,t;W).$$

Функции активации создают дополнительный слой между слоем скрытых переменных и наблюдаемых данных, тем самым абстрагируют наблюдаемые данные.

Преимущество данной модификации заметно при сравнении с обычным определением функции:

$$f^1(h,x,t;w)=\frac{1}{|r(x_t)|}\sum_{x'}\sum_{g\in G} w_{h,d} I\left(h_t=h'\right)x'_d.$$

Как видно из определения, этап обучения отсутствует и вместо него применяется линейная комбинация признаков X'_d и весов $w_{h,d}$. В работе [15] показано, что использование нелинейности в модели позволяет повысит дискриминационные способности и тем самым увеличить точность распознавания сложных пространственно-временных закономерностей, таких как динамические жесты.

Параметрический вектор в модели представлен, как $W=\{w_{g,h},w_{g,d},w_{y,h},w_{y,h.h}\}$. Его размерность составляет $GH+GD+YH+YHH$, где G — количество функций активации, H — число скрытых состояний, D — размерность признаков и Y — число меток классов. Для последовательности X длиной T при использовании алгоритма распространения доверия вычислительная сложность составит $O(TYH^2)$.

В зависимости от конфигурации модели, число используемых в ней слоев может быть различно, поэтому для удобства и дальнейшего обозначения введем индекс для их описания $X^l=\{x_1^l, ..., x_T^l\}$ и ссылочный оператор $r(x_t^l)$, который возвращает группу наблюдений предыдущего слоя. Для $l=1$ $r(x_t^l)=x_t$.

Для обобщения последовательности X^l на следующий слой X^{l+1}, используется алгоритм, описанный в работе [16] с модифицированной метрикой подобия.

Пусть, $G=(V,E,W)$ представляет собой неориентированный граф с множеством вершин V, ребер E и весов W, определяющих схожесть между двумя вершинами.

```
Вход: Граф G
Выход: C = {c_1, ..., c_k}
C = V, c_t = c(x_t^{l+1}) = {x_t^l}, ∀t
O = сортировать(E, W)
от q = 1 до |O|:
   s, t = O_q
   если c_s != c_t && M_{s,t} ≤ MInt(C_s, C_t):
      C = объединить(C_s, C_t)
```

Рис. 1. Обобщение последовательности путем использования алгоритма группировки

Алгоритмы объединяет $r(x_s^{l+1})$ и $r(x_t^{l+1})$, если подобие между группами меньше, чем минимальное внутреннее различие, которое определено как

$$\mathrm{MInt}(C_s,C_t)=$$
$$=\max(\mathrm{Int}(C_s)+\tau(C_s),\ \mathrm{Int}(C_t)+\tau(C_t)).$$

В данной формуле Int — внутреннее различие, определяемое по формуле:

$\mathrm{Int}(C)=\max\limits_{(s,t)\in mst(C,E)} M_{s,t}$, где mst — минимальное остовное дерево. В свою очередь, $\tau(C)=\tau/|C|$ — пороговая функция.

Метрика подобия в контексте решаемой задачи высчитывается следующим образом:

$$M_{s,t}=\sum_{y,h'}\left|p\left(h_s=h'|y,X;W\right)-p\left(h_t=h'|y,X;W\right)\right|$$

Как показано в [16] вычислительная сложность алгоритма составляет $O(T\log T)$, где K — длина последовательности.

С учетом переменного числа слоев условная вероятность может быть определена как

$\prod_{l=1}^{L} p(y|X^l;W^l)$.

Так как W^l — вектор параметров для текущего слоя, то вектор параметров для всей модели $W=\{W^1,\ldots,W^l\}$.

Пусть дана обучающая выборка $D=(X_i, y_i)$, $X_i \in R^{D\times T_i}$, $y_i \in Y$, $i=1...N$. Стандартный подход к поиску оптимального решения W^* заключается в минимизации функции потерь:

$\min_W L(W)=\frac{1}{2\sigma^2}\|W\| - \sum_{i=1}^{N} \log p(y_i|X_i;W)$.

Для данной модели такой подход не применим, т.к. для получения обобщения последовательности на следующем слое X^{l+1} требуется знать $p(h^l|y,X^l;W^l)$.

Подходящим подходом будет последовательная оптимизация для каждого отдельного слоя l [17]:

$\min_{W^l} L(W^l)=\frac{1}{2\sigma^2}\|W^l\| - \sum_{i=1}^{N} \log p(y_i|X_i^l; W^l)$.

Данная задача решается с помощью квазиньютоновских методов, в частности L-BFGS [18]. Данный ряд методов был успешно применен для решения аналогичного класса задач [13].

Частные производные по параметру W^l для обучающей выборки D высчитываются как

$$\frac{\partial \log p(y_i \mid X_i^l; W^l)}{\partial W^l} = \sum_{h^l} p(h^l \mid y_i, X_i^l; W^l) \frac{\partial F}{\partial W^l} -$$

$- \sum_{y', h^l} p(y', h^l \mid X_i^l; W^l) \frac{\partial F}{\partial W^l}$.

Вид частных производных $\frac{\partial F}{\partial W^l}$ относительно $w_{y,h}$ и $w_{y,h.h}$ аналогичен, показанному в работе [13]:

$\frac{\partial f^2}{\partial w^l_{y,h}} = \sum_t I(y=y') I(h_t^l=h')$,

$\frac{\partial f^3}{\partial w^l_{y',y,h}} = \sum_t I(y=y') I(h_t^l=h') I(h_t^l=h'')$,

$\frac{\partial f^1}{\partial w^l_{g,h}} = \sum_t I(h_t^l=h') \psi_g(X^l, t; W^l)$,

$\frac{\partial f^1}{\partial w^l_{g,d}} = \sum_t w^l_{g,d} \frac{1}{|r(X_t^l)|} \sum_{x'} g(\partial w^l_{g,d} x') (1-g(\partial w^l_{g,d} X')$.

```
Вход: обучающая выборка D
Выход: оптимальное решение w*
от l = 1 до L:
  W* = argmin L(w^l)
  для каждого x_i ∈ D
     x_i^{l+1} = обобщить (x_i^l, w*^l)
```

Рис. 2. Алгоритмы обучения модели

```
Вход: последовательность x,
      оптимальное решение W*
Выход: метка класса y*
p(y|x; w*) = 0
от l = 1 до L:
  log p(y|x; w*) += p(y|x^l; w*^l)
  x^l = обобщить (x_i^l, w*^l)
y* = argmax log p(y|x; w*)
```

Рис. 3. Алгоритм проверки

Рис. 2 и рис. 3 показывают процедуру обучения и проверки.

Во время обучения для каждого слоя l высчитывается W^{*l} и создается обобщение X^{l+1} для каждого предшествующего слоя.

Во время проверки проводится суммирование $\log p(y_i \mid X_i^l; W^l)$ и определяется оптимальная метка класса.

Причем, если $X_i^{l+1}=X_i^l$, то процедура обучения и проверки прекращается для X_i. Т.е. X^{l+1} будет всегда короче, чем X^l.

Вычислительная сложность всей модели складывается из $O(TYH^2)$ и $O(T \log T)$. Для L слоев она составит $O(LTYH^2 + LT\log T)$. Так как T — длина каждого слоя, то она будет уменьшаться с увеличением порядка слоев т.к. X^{l+1} будет всегда короче, чем X^l. Следовательно, вычислительная сложность модели растет сублинейно относительно числа слоёв.

7 Заключение

К интерфейсам на основе динамических жестов предъявляется ряд требований: работа в режиме реального времени, способность регистрации изменения конфигурации генератора жестов с частотой более 3 Гц, возможность создания персонализированных баз жестов, сохранение высокой точности распознавания при увеличении жестового набора более чем за 30 жестов.

В результате проведенных анализов показано, что существующие методы распознавания обладают функциональными ограничениями и не могут в полном объеме реализовать практически необходимые требования. Поэтому была предложена и описана на теоретическом уровне модификация метода условных случайных полей со скрытыми состояниями, которая позволяет устранить ряд ограничений модели и повысить её дискриминативные способности.

Показано, что модификация метода алгоритмически реализуема и обладает сублинейной временной сложность.

Список литературы

[1] Ese, T. H. 2009. *Human Gesture Recognition*. INRIA Sophia Antipolis, France.

[2] Dominio, F., Donadeo, M., Marin, G., Zanuttigh, P., Cortelazzo, G. M. 2014. *Real-Time Hand Gesture Recognition*. University of Padova, Italy.

[3] Nowicki, M., Pilarczyk, O., Wąsikowski, J., Zjawin, K. 2014. *Gesture Recognition Library for Leap Motion Controller*. Poznan University of Technology, Poland.

[4] Технология *Leap Motion*. Режим доступа: https://leapmotion.com (дата обращения: 12.02.2017).

[5] Технология *RealSense*. Режим доступа: https://software.intel.com/realsense (дата обращения: 12.02.2017).

[6] Технология *Kinect*. Режим доступа: https://microsoft.com/kinect (дата обращения: 12.02.2017).

[7] Технология *Orbbec*. Режим доступа: https://orbbec3d.com (дата обращения: 12.02.2017).

[8] Технология *Myo*. Режим доступа: https://www.myo.com (дата обращения: 12.02.2017).

[9] Технология *Gest*. Режим доступа: https://gest.co (дата обращения: 12.02.2017).

[10] Технология *Project Soli*. Режим доступа: https://atap.google.com/soli (дата обращения: 12.02.2017).

[11] Souza, C. R., Pizzolato, E. B. 2013. *Sign language recognition with support vector machines and hidden conditional random fields*. Springer Berlin Heidelberg, Germany.

[12] Vail, D. L., Veloso, M. M., Lafferty, J. D. 2007. *Conditional random fields for activity recognition*. ACM, USA.

[13] Wang, S. B., Quattoni, A., Morency, Demirdjian, D., Darrell, T. 2006. *Hidden Conditional Random Fields for Gesture Recognition*. IEEE, USA.

[14] Blei, D. M., Ng, A. Y., Jordan, M. I., Wallach, H. M., Hinton, G. E., Osindero, S., 2004. *Conditional random fields: An introduction*. University of Pennsylvania, USA.

[15] Do, T. M. T., Artieres, T. 2010. *Neural conditional random fields*. Thirteenth International Conference on Artificial Intelligence and Statistics, USA.

[16] Felzenszwalb, P. F., Huttenlocher, D. P. 2004. *Efficient graph-based image segmentation*. International journal of computer vision, USA.

[17] Hartline, J. R. K. 2008. *Incremental optimization*. Cornell University, USA.

[18] Nocedal, J., Wright, S. J. 1999. *Numerical Optimization*. Springer, USA.