Семантический анализ визуальной сцены

А.А. Харламов
Институт высшей нервной деятельности и нейрофизиологии РАН; Московский государственный лингвистический университет; Высшая школа экономики
Москва
kharlamov@analyst.ru

Р.М. Жаркой
Интеллектуальные системы безопасности
Москва
roman.jarkoi@iss.ru

***Abstract* — Переход от графического представления видеосцены к ее естественно-языковому описанию является естественным шагом в процедуре семантического анализа видеосцены, позволяющим снять вариативность изображений с одной стороны, а с другой – использовать уже сформировавшиеся механизмы анализа текстов для формирования семантических представлений в виде неоднородных семантических сетей. При этом видеокадры видеосцены переописываются из графического формата в представление в терминах характерных признаков, полученных на выходе сенсоров. Этим представлениям ставятся в соответствие шаблоны естественного языка, соответствующие сценариям, представленным на видеокадрах. Далее, последовательность шаблонов, полученных при анализе видеоряда анализируется статистически как квазитекст. И, наконец, связи пар вершин полученной в результате статистического анализа однородной семантической сети размечаются типами отношений с помощью лингвистических механизмов синтактико-семантического анализа предложений. Полученные в результате этих процедур квазитекст и его неоднородная семантическая сеть могут служить для анализа более высокого уровня.**

Keywords* — *видеоаналитика, характеристики видеокадра, лингвистический шаблон сцены, статистический анализ текста, однородная семантическая сеть, синтактико-семантический анализ предложений текста, неоднородная семантическая сеть

Анализ семантики видеосцен становится все более актуальным в связи с широким применением видеоаналитики в решении задач обеспечения безопасности. В настоящий момент эффективно решаются задачи нижнего уровня семантического анализа сцен (на примере работы видеоохранных систем). Это: детектирование оставленных предметов, детектирование движения, детектирование объектов, детектирование пересечения границ. Более сложные задачи семантического анализа сцены чаще всего не выходят за рамки статистического анализа потоков с выявлением нештатных треков. Все эти алгоритмы, будучи эмпирическими по своей природе, плохо масштабируются, и еще хуже интегрируются в цельное представление семантики сцены.

Основной задачей видеоаналитики является анализ видеопотока с использованием алгоритмов компьютерного зрения с целью автоматического получения той или иной информации без прямого участия человека.

Анализируемое видеоизображение можно условно разделить на две составляющие: фон (задний план) и объекты (передний план). При этом фон, как правило, является квазистатическим (медленно меняющимся), в то время как объекты переднего плана меняют свое положение во времени. В связи с этим, в решении большинства задач видеоаналитики можно выделить несколько основных этапов. (1) Выделение фона. Алгоритм выделения фона должен быть адаптивным. (2) Выделение объектов переднего плана. Основная цель заключается в обнаружении движущихся объектов на каждом кадре видеопоследовательности. (3) Трекинг движущихся объектов. Выделение на каждом кадре объектов и оценка их смещения от кадра к кадру позволяют построить траектории движения и получить оценки различных параметров (скорость, направление движения и т.д.). (4) Классификация. Каждый объект переднего плана может быть отнесен к определенному классу (человек, транспортное средство и т.д.) в соответствии с некоторым набором признаков.

Располагая информацией об объектах переднего плана, можно осуществлять анализ их поведения. В результате возможно, например, обнаружение фактов пересечения линии, прохода в запрещенную зону, резкого ускорения и т.д.

Анализ полученной информации позволяет принять решение о наличии или об отсутствии той или иной тревожной ситуации.

Типовой набор характеристик объекта включает: (1) уникальный идентификатор (ID); (2) траекторию объекта (набор координат положения объекта в координатах изображения); (3) линейные размеры объекта; (4) среднюю и мгновенную скорости; (5) среднее и мгновенное направления движения; (6) принадлежность к классу (человек, машина, неодушевленный предмет небольшого размера, и т.д.); (7) результаты работы тех или иных логических (эмпирических) алгоритмов анализа поведения (проход в запрещенную зону, пересечение линии контроля, резкое ускорение и т.д.).

Использование для анализа сцен подмножества естественного языка, обладающего по своей природе иерархической структурой, позволяет снять ограничения на масштабирование представлений по степени сложности, и, кроме того, позволяет реализовать автоматическое формирование визуализации семантики сцены в виде семантической сети. Для этого требуется осуществить

переход от динамики видеорядов в терминах представленных выше характеристик к статике семантических представлений, включающий:

- анализ сцен по предметным областям (классификация типов видеокадров);
- соотнесение типов видеокадров с шаблонами подмножества естественного языка;
- интерпретация видеоряда (также в терминах упомянутых характеристик) как квазитекста: формирование на основе этого квазитекста (возможно, с использованием других квазитекстов, описывающих заданную предметную область) однородной семантической сети, где вершины соответствуют объектам сцены, а связи однотипны (только ассоциативные);
- переход к семантическому представлению в виде неоднородной семантической сети с применением лингвистических правил выявления расширенных предикатных структур предложений – шаблонов – в которых отношения между элементами сцены поименованы.

При формировании шаблонов описаний кадров видеоряда необходимо учитывать два условия: одно – относящееся к зрительной модальности, другое – к текстовой. Выполнение первого условия возможно за счет принятия во внимание вариативности видеокадров видеорядов, относящихся к заданной предметной области. Выполнение второго условия обеспечивается учетом всех вариантов расширенных предикатных структур, включающих те или иные конкретные предикаты, описывающие видеокадр.

I. Классификация кадров видеоряда визуальной сцены в заданной предметной области

На множестве видеорядов (скорее треков), описывающих стандартные и нестандартные ситуации, представляющие реальные видеосцены, осуществляется выявление типов видеокадров (а, точнее, треков с их начальными и конечными состояниями), участвующих в видеорядах выбранного множества. Эти видеокадры относятся как к базовым элементам динамики сцены: начало движения, перемещение в заданном направлении, остановка, объединение треков объектов, разделение треков, и подобное; так и к нештатным (или тревожным) элементам: появление оставленных предметов, пересечение запрещенных границ, и подобное. Разные типы поведения объектов на сцене могут включаться как в штатные, так и в нештатные ситуации, а отнесение их к тем, или другим возможно лишь их соотнесением с более широким контекстом всей динамики сцены, которое в настоящий момент невозможно из-за использования для выявления отдельных типов поведения эмпирических алгоритмов, которые плохо интегрируются в комплексную сцену. Плохая интеграция может быть объяснена использованием в разных алгоритмах разных систем параметров объектов и разных систем правил для их анализа. С целью преодоления противоречия необходимо использование единого языка представления видеосцены на всех этапах анализа, каковым, например, является естественный язык. Дополнительный бонус использования естественного языка – легкость понимания человеком-оператором описаний сцены, представленных на этом языке.

Для представления видеорядов динамичной видеосцены с этой целью разумно использовать некие стандартные предложения естественного языка (шаблоны), характеризующие отдельные кадры видеоряда. Тогда весь видеоряд (вся динамичная видеосцена) может быть описан последовательностью таких шаблонов – текстом естественного языка. Такое описание не ново: это стандартный способ комментирования отдельных сцен видеофильмов. Нова сама постановка вопроса: заменить комментатора автоматическим устройством. Тогда нахождение нужного видеокадра в массиве видеорядов большого объема оказывается достаточно просто выполнимой задачей.

Для того чтобы сформировать систему шаблонов, необходимо предварительно сформировать систему классов кадров видеоряда, описывающих все возможные ситуации в заданной предметной области. Часть классов хорошо и давно известна. Это: переход из динамического состояния в статическое (оставление предмета), перемещение (движение) в некотором направлении, наличие объекта, переход преграды – условной, или физической (пересечение), появление, исчезновение. Наверное, есть и другие подобные классы. Им всем соответствуют определенные естественно-языковые шаблоны (предложения из некоторого подмножества естественного языка).

II. Формирование естественно-языковых шаблонов классов кадров видеоряда

Для формирования перечня шаблонов необходимо проанализировать классы слов естественного языка (в первую очередь - предикатов), которые участвуют в описании видеокадров, как с точки зрения номенклатуры валентностей соответствующих слов, так и с точки зрения влияния синонимии на эту номенклатуру. Учет номенклатуры валентностей позволит максимально точно и подробно описывать видеокадры. Учет влияния синонимии позволит выявить различия в подробности описания одного и того же видеокадра различными вариантами шаблонов. В идеале необходим выбор перечня шаблонов, описывающих видеокадры максимально инвариантно. Такой перечень будет открытым, но постепенно процесс его заполнения должен сходиться.

Формирование перечня шаблонов будет идти параллельно с формированием перечня классов видеокадров. Эти два процесса будут влиять друг на друга: выявление новых типов шаблонов будет порождать новые классы видеокадров; появление новых классов видеокадров приведет к формированию новых типов шаблонов. И еще: отсутствие соответствия видеокадров, интерпретируемых с точки зрения имеющейся видеосенсорики, наличествующим шаблонам будет понуждать видеоаналитиков формировать новые механизмы видеосенсорики для интерпретации недостающих элементов видеокадра.

III. Сопоставление лингвистических шаблонов отдельным кадрам видеоряда визуальной сцены. Формирование квазитекста из шаблонов, соответствующих последовательным кадрам видеоряда визуальной сцены

Два параллельных процесса формирования перечня классов видеокадров и формирования перечня классов описывающих их шаблонов приводят к взаимнооднозначному и на (изоморфному) сопоставлению двух перечней. Такое сопоставление будет возможно в том случае, если видеосенсорика будет давать все необходимые данные для формирования естественно-языкового описания видеокадра, соответствующего его шаблону.

Сопоставление перечня шаблонов с перечнем описываемых шаблонами видеокадров позволяет осуществить автоматический переход от описываемого с помощью шаблонов видеоряда к тексту, описывающему динамичную сцену. Степень подробности описания может варьироваться во времени от учета наиболее крупных изменений событий на видеокадрах до протокольной покадровой записи. Степень подробности описания может варьироваться в пространстве признаков в зависимости от наличия тех или иных элементов видеосенсорики, дающих более или менее подробное представление видеокадра в соответствие с принятым шаблоном.

IV. Однородная семантическая сеть, описывающая динамичную видеосцену

Однородная семантическая сеть представляет собой циклический граф, вершины которого соответствуют ключевым объектам видеосцены, а дуги соответствуют отношениям совместной встречаемости ключевых объектов на видеокадрах видеосцены [4, 5].

Благодаря наличию специфических механизмов видеосенсорики, которые учитывают не только статику видеокадра, но и динамику сцены (объекты перемещаются, из-за чего возникает эта динамика), удается объединить видеокадры в видеоряд - квазитекст. То есть последовательное объединение соответствующих видеокадрам естественно-языковых шаблонов позволяет сформировать естественно-языковой текст, описывающий видеосцену – последовательность видеокадров.

Полученная в результате анализа такого текста однородная (ассоциативная) семантическая сеть, таким образом, позволяет показать взаимосвязи ключевых объектов на видеосцене. Цепочки вершин на ассоциативной сети позволяют увидеть зависимости между объектами.

V. Формирование однородной семантической сети из квазитекста, описывающего видеосцену

Как только мы получили естественно-языковой текст – последовательность шаблонов, описывающий видеоряд, соответствующий динамичной видеосцене, мы в состоянии построить однородную семантическую сеть, характеризующую эту видеосцену. Построение однородной семантической сети осуществляется стандартными механизмами технологии TextAnalyst [2] для автоматической смысловой обработки текстов [1].

Построение однородной семантической сети включает несколько этапов: первичную обработку, формирование частотной сети, и переранжирование частотных характеристик вершин сети в их смысловые веса. Первичная обработка включает: сегментацию предложения, удаление стоп-слов, рабочих и общеупотребимых слов, морфологическую обработку (точнее стемминг) с целью устранения информационного шума. Формирование частотного портрета текста включает вычисление частоты встречаемости оставшихся корневых основ $< c_i >$ в тексте (соответствующих вершинам сети), полученных в результате стемминга, вычисление частоты попарной встречаемости корневых основ $< c_i c_j >$ в тексте, формирование первичной ассоциативной сети, а также выявление устойчивых словосочетений (которые также ставятся в соответствие вершинам сети). И наконец – этап переренжирования – перевычисления частотных характеристик вершин сети в смысловые характеристики ключевых понятий.

В результате получается однородная семантическая сеть как множество несимметричных пар понятий.

Определение. Под ассоциативной (однородной) семантической сетью понимается двойка $< c_i c_j >$, где c_i и c_j - несимметричная пара понятий, связанных между собой отношением ассоциативности (совместной встречаемости в предложении текста). Или, что то же самое, семантическую сеть можно представить в виде множества звездочек $< c_i < c_j >>$, где c_j – множество ближайших ассоциантов ключевого понятия c_i:

$$N \cong < c_i < c_j >> \quad (1)$$

VI. Формирование неоднородной семантической сети из однородной семантической сети

Сформированная ранее однородная семантическая сеть может быть преобразована в неоднородную путем добавления в нее информации о типах отношений между парами понятий, которая (информация) может быть получена из исходного текста (корпуса текстов) привлечением лингвистических методов анализа – методов выявления расширенной предикатной структуры отдельного предложения [3].

Расширенная предикатная структура предложения представляет собой ту же звездочку (1), только строится она не на основе пар понятий, а на основе пар понятий, отношения между которыми поименованы их типами.

Под расширенной предикатной структурой будем понимать тройку $P \cong < c_i < r_{ij} c_j >>$, где c_i - субъект, $r_{ij} \cong r_p$, где $j = 1,$ - предикативное отношение между субъектом и главным объектом, а r_{ij}, где $j > 1$– связи субъекта с другими актантами предиката.

После выявления расширенной предикатной структуры отношения, выявленные в процессе формирования расширенной предикатной структуры, могут быть использованы для разметки связей между вершинами однородной семантической сети. И тогда семантическая сеть превращается в неоднородную семантическую сеть:

$$N \cong < c_i < r_{ij} c_j >> \quad (2)$$

VII. Пример формирования однородной семантической сети на основе анализа квазитекста

В качестве примера формирования однородной семантической сети можно взять произвольный видеоряд, описывающий конкретную видеосцену. Для простоты восприятия текста в данной работе формальные термины «Объект А», «Объект В», «Неподвижный объект» заменены на «Вася», «Петя», «Кейс» и подобное.

> **«Вася** появляется в проеме **Входной двери**. **Вася** движется от **Входной двери** в **Направлении Киоска**. **Вася** движется от **Входной двери** в **Направление Киоска**. **Вася** движется от **Входной двери** в **Направление Киоска**. **Вася** движется от **Входной двери** в **Направление Киоска**. **Вася** прекращает движение в **Направление Киоска**. У **Киоска Вася** оставляет **Кейс**. **Вася** начинает движение в **Направление Внутренней двери**. **Петя** появляется в проеме **Внутренней двери**. **Петя** движется от **Внутренней двери** в **Направление Киоска. Петя** движется от **Внутренней двери** в **Направление Киоска. Петя** движется от **Внутренней двери** в **Направление Киоска. Петя** движется от **Внутренней двери** в **Направление Киоска. Петя** проходит мимо неподвижного **Кейса».**

В результате автоматического анализа текста формируется ассоциативная сеть. Ее фрагмент представлен в Таблице I.

TABLE I. Ассоциативная звездочка – фрагмент однородной семантической сети

Родитель	Частота	Вес	Дочерняя вершина
…			
Вася	10	84	Входной двери
Вася	12	90	Направление
Вася	12	90	Киоска
…			

В фрагменте сети представлена только одна звездочка. Здесь главное слово звездочки – один из основных объектов, участвующих в описании видеосцены – «Вася» . А второстепенные – другие объекты – «Входная дверь», «Направление», «Киоск», описывающие видеосцену. Эта звездочка соответствует предложению текста «Вася движется от Входной двери в Направление Киоска».

Неоднородная сеть, полученная после разметки отношений с помощью лингвистического анализа [6], очень точно описывает видеосцену (см. Таблицу II), связность же ключевых понятий обеспечивается за счет статистического подхода к анализу всего текста.

TABLE II. Звездочка с типами отношений – фрагмент неоднородной семантической сети

Родитель	Тип отношения	Дочерняя вершина
…		
Вася	«откуда»	Входной двери
Вася	«движется»	Направление
Вася	«куда»	Киоска
…		

Conclusion

Переход от графического представления видеосцены как последовательности видеокадров через типовые характеристики объектов к квази-тексту как последовательности лингвистических шаблонов, соответствующих отдельным видеокадрам позволяет анализировать сцену, очищенную от информационного шума вариативности изображений. Статистическая, а далее, лингвистическая обработка последовательности шаблонов как текста естественного языка дает в руки эксперта мощный инструмент семантического анализа видеоматериала, а также удобный и компактный способ представления информации аналитику. Такое представление информации позволяет перейти к логическому и эмпирическому анализу информации верхних уровней абстракции.

References

[1] Харламов А.А. Нейросетевая технология представления и обработки информации (естественное представление знаний). - М.: Радиотехника, 2006. - 89 с.

[2] Харламов, А.А. Автоматический структурный анализ текстов [Электронный ресурс] / А.А. Харламов // Открытые системы. – 2002 – № 10. – Режим доступа: http://www.osp.ru/os/2002/10/182010/. – 20.12.2011 г.

[3] Харламов А.А., Ермоленко Т.В. Автоматическое формирование неоднородной семантической сети на основе выявления ключевых предикатных структур предложений текста // Труды Международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» (OSTIS'2012), - Минск: 2012. С. 385 - 390

[4] Харламов А.А., Ермоленко Т.В. Нейросетевая среда (нейроморфная ассоциативная память) для преодоления информационной сложности. Поиск смысла в слабоструктурированных массивах информации. Часть I. Структурная обработка информации в коре / Информационные технологии, N 11, 2015. – Стр. 814—820.

[5] Харламов А.А., Ермоленко Т.В. Нейросетевая среда (нейроморфная ассоциативная память) для преодоления информационной сложности. Поиск смысла в слабоструктурированных массивах информации. Часть II. Обработка информации в гиппокампе. Модель мира / Информационные технологии, N 12, 2015. – Стр. 883—889.

[6] Alexander A. Kharlamov, Tatyana V. Yermolenko, Andrey A. Zhonin Text Understanding as Interpretation of Predicative Structure Strings of Main Text's Sentences as Result of Pragmatic Analysis (Combination of Linguistic and Statistic Approaches)//Speech and Computer, Proceedings of the 15th International Conference SPECOM 2013, Pilsen, Czech Republic, September, 2013. Pp. 333-339.

SEMANTIC ANALYSIS OF VISUAL SCENE

Alexander A. Kharlamov
Roman M. Jarkoi

Transition from pictures of video to its nature language description is a natural step in the procedure of semantic analysis of video. The transition eliminates the pictures variability from one side and to use instruments of the semantic text analysis from the other one. A nonhomogeneous semantic network is builded during such an analysis. The scene pictures in the process are translated from graphical format into set of specific attributes, which are the result of scene censoring. Some natural language templates are put into accordance to these scene representations. The template is a scenarios of the scene pictures. Then the sequence of templates is analyzed as a quasytext by statistical methods to build a homogenous semantic network. And at last the relationships of nodes pares of homogenous semantic network are named of their types by linguistic mechanisms of syntactical-semantic sentence analysis. Such a quasytext and its nonhomogenous semantic network could then be used for semantic analysis of more complex level.