УДК 004.522

А.А. КАРПОВ, Л. АКАРУН, Ал.Л. РОНЖИН

МНОГОМОДАЛЬНЫЕ АССИСТИВНЫЕ СИСТЕМЫ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО ЖИЛОГО ПРОСТРАНСТВА

Карпов А.А., Акарун Л., Ронжин Ал.Л **Многомодальные ассистивные системы для интеллектуального жилого пространства**.

Аннотация. В статье представлен обзор систем, применяемых для ассистивного интеллектуального пространства. Также описывается разработанная многомодальная ассистивная система для интеллектуального жилого пространства, которая состоит из двух комплексов средств. Первый комплекс выполняет обработку видеопотоков для определения положения пользователя и слежения за его перемещением, а также анализа его действий. Ко второму комплексу относится система обработки аудиопотоков, предназначенная для автоматического распознавания речевых команд и акустических событий. Разработанная система автоматического распознавания речи многоязычна и позволяет распознавать слова, произнесенные на английском или русском. В процессе проведения экспериментов было записано 2811 аудиофайлов, содержащих речь и акустические события, средняя точность распознавания составила 96,5% и 93,8% соответственно.

Ключевые слова: системы видеонаблюдения, сегментация переднего фона, слежение за ключевыми точками, обработка аудиосигнала, распознавание речи, определение акустических событий.

Karpov A.A., Akarun L., Ronzhin Al.L **Multimodal assistive systems for a smart living environment.**

Abstract. The paper proposes a survey of assistive smart spaces and ambient assisted living environments. Also design of a multimodal assistive system for a smart living environment is presented. The system consists of two software complexes. The first one provides video signal processing and surveillance for detecting and tracking a user as well as analysis of his/her activity. The second software complex provides audio signal processing for automatic recognition of speech messages and non-speech acoustic events. The developed automatic speech recognition system is multilingual one and is able to recognize words both in English and in Russian. At the experiments, 2811 wave files with speech commands and simulated acoustic events have been recorded in total. Recognition rate for speech commands and non-speech acoustic events was 96.5% and 93.8%, respectively.

Keywords: video surveillance systems, elderly healthcare, foreground segmentation, keypoint tracking, audiosignal processing, speech recognition, acoustic event detection.

1. Введение. Интеллектуальные жилые пространства, комнаты и дома, оснащенные видео- и аудиосенсорами, а также «умными» бытовыми приборами и окружением, определяют будущее систем помощи, поддержки и реабилитации людей с ограниченными возможностями и пожилых людей. В последние годы развивается концепция ассистивных интеллектуальных пространств, и интерес к ним постоянно растет. Помимо чисто научных исследований [1, 2, 3] существуют и

коммерческие приложения, например, QuietCare Systems [4], Philips LifeLine [5] и The Intel Health Guide [6]. Исследования в этой области могут быть разделены по четырем категориям:

1) Работы по классификации активности и деятельности человека [7];

2) Исследования, направленные на поддержание активной деятельности пользователя в повседневной жизни [8];

3) Исследования, направленные на анализ движения тела человека и на определение падения/предупреждение падения [9];

4) Исследования, анализирующие медицинские и жизненно важные показатели человека [10].

Исследования, которые пытаются классифицировать деятельность человека, направлены на анализ активности человека, определение отклонений от обычного распорядка дня и обнаружение аномального поведения. Например, если пожилой человек спит больше, чем обычно, и не встает, это может являться сигналом о возможной болезни. Приложения по поддержанию деятельности в повседневной жизни могут помогать пожилому человеку в выполнении его обычных обязанностей, например, принимать предписанные лекарства, соблюдать диету или выполнять предписанные врачом физические упражнения. Устройства, отслеживающие местоположение, например, могут предупредить членов семьи, когда человек с болезнью Альцгеймера уходит из дома. Приложения по анализу положения тела и обнаружению падения на пол направлены на предупреждение физических травм, которые являются большим риском для здоровья пожилого человека. Приложения, которые отслеживают данные о состоянии здоровья, анализируют собранные данные электрокардиограммы, пульса и артериального давления.

Существует множество технологий, содержащих различные модальности: специальный дизайн наручных часов с установленными в них сенсорами для получения информации от тела человека; специальные кровати, определяющие лежит ли на них человек; инфракрасные камеры или камеры с цветовой матрицей, установленные на стенах, и аудиосенсоры для определения запросов о помощи [11, 12, 13]. Среди перечисленных средств видеокамеры и микрофоны чаще всего устанавливаются в интеллектуальных пространствах, кроме того модальности учитываются в исследованиях для мониторинга состояния пожилых людей и людей с инвалидностью.

Визуальные данные, главным образом, используются для слежения за перемещением человека и для автоматического распознавания

активности в ассистивном пространстве. При распознавании активности видеоданные сегментируются во времени. Предполагается, что в видеоданных присутствует активность только одного человека. Кроме того, активность определена упорядоченным набором действий. Например, приготовление пищи является активностью, поскольку перемешивание является действием. Понимание человеческих действий содержит множество областей применения, таких как безопасность, наблюдение, бытовая помощь и даже развлекательные мероприятия. В технологии определения активности выделяют два этапа, вначале определяется время активности и место, далее помечается объект или цель применения активности. С модулем определения активности появляется возможность определения некоторого набора действий, содержащих критические состояния, например, падение и тревога пользователей системы в связи с экстраординарными ситуациями [14]. Современные технологии распознавания человеческой активности совмещают видеослежение за человеком и описание его движения посредством визуальных методов, таких как ключевые точки и классификация выполненных действий [15, 16]. Определение и слежение за людьми основано на низкоуровневых технологиях сегментации заднего и переднего фона, например модели Гауссовых смесей [17, 18] и модели кодового словаря [19]. Объект, найденный после сегментации переднего фона, используется для выделения признаков, применяемых для описания человека. В работе [20] представлен детальный обзор широко используемых методов определения точек интереса и выделения признаков. Данная работа показала, что дескрипторы точки интереса, в основном, используются в камерах с низким коэффициентом искажения.

Исследования, основанные только на визуальной информации, имеют некоторые недостатки, например, появление ложных объектов при изменении освещенности, могут возникнуть ложные объекты, появление которых может привести к тому, что действие может быть не классифицировано или классифицировано неправильно. Использование технологий обработки аудиосигнала позволяет создать более робастную и устойчивую систему. Аудиосигнал состоит из речи человека и звуков окружающего пространства, таких как стук в дверь или льющаяся вода в раковине для мытья рук. Человеческая речь относится к процессу, связанному с воспроизведением и пониманием звуков, используемых в разговорном языке, кроме того автоматическое распознавание речи является процессом конвертации речевого сигнала в последовательность слов. Существуют несколько типов речи: побук-

венная речь (с паузой между буквами), изолированная речь (с паузами между словами), слитная речь [21] (когда диктор не разделяет слова паузами) и спонтанная естественная речь. Современные системы автоматического распознавания речи используют математические технологии, такие как скрытые Марковские модели, искусственные нейронные сети, байесовские сети или методы динамического изменения временной шкалы (динамическое программирование). Самые популярные модули автоматического распознавания речи выполняют дикторонезависимое распознавание речи, хотя в некоторых случаях (например, персональные системы для распознавания только голоса владельца) применение дикторозависимых систем является более адекватным.

Несмотря на то, что речь, безусловно, является самым информативным акустическим событием, другие типы звуков могут также нести полезную информацию. Следовательно, определение или классификация акустических событий может помочь в определении людей и активности, которая может проявляться в помещении [22]. Например, аплодисменты или смех в процессе беседы; громкий зевок в середине лекции; перемещение кресла или звук открываемой двери и т.д. Кроме того, робастность систем автоматического распознавания речи может быть увеличена при помощи начального определения неречевых звуков в записанных сигналах.

Классификация и/или определение акустических событий является новой областью анализа слуховых сцен [23], которая выполняет обработку акустических сигналов и преобразование их в символьное описание, соответствующее восприятию слушателя различных звуковых событий, которые представлены в сигналах и их источниках.

Определение акустических событий может быть применено в различных окружающих пространствах [24], таких как: госпиталь [25], кухонное помещение [26], или даже ванная комната [27]. Для помещений, предназначенных для проведения мероприятий, задача определения акустических событий достаточно нова, однако уже была оценена в работах двух международных оценочных компаний: в проекте CLEAR 2006 тремя участниками и CLEAR 2007 шестью участниками. В большинстве представленных систем применяется стандартная комбинация кепстральных коэффициентов и скрытых Марковских моделей, широко используемых при распознавании речи.

В данной работе описывается многомодальное ассистивное жилое пространство, где слежение за перемещением нескольких людей осуществляется с помощью обработки видеопотока, поступающего с двух

всенаправленных камер. Кроме того, в данной системе применяется набор микрофонов для автоматической обработки аудиосигналов.

В следующих разделах более детально описаны методы, используемые в представленной системе, далее представлены классификация аудиособытий и база данных, собранная в процессе исследований, а также результаты, полученные в процессе проведения экспериментов. В заключении представлены основные достигнутые результаты.

2. Методы обработки аудиовизуальных потоков. В данной системе применены видео- и аудиосигналы. Вначале рассматриваются методы обработки видеопотока для слежения за перемещением пользователя. Затем описаны методы многоканальной обработки аудиосигналов для автоматического распознавания речи и акустических событий. Многомодальные пользовательские интерфейсы, обрабатывающие естественные для человека способы коммуникации, позволяют организовать удобный и интуитивно понятный процесс взаимодействия между пользователем и окружающим интеллектуальным пространством [28, 29].

2.1. Методы определения и слежения за перемещением пользователя. Для создания ассистивного жилого пространства требуется видеокамера для мониторинга помещения, что позволяет получить необходимую информацию для определения людей в положении лежа и действий пользователей при помощи анализа нестабильных регионов сцены, которые называются передним фоном. Метод моделей Гауссовых смесей (GMMs) используется для сегментации объектов переднего фона и в дополнение для устранения теней, что применяется для улучшения результатов определения при помощи удаления артефактов, соответствующих малым изменениям в цвете или интенсивности [17, 18]. В данном методе каждый пиксель обрабатывается индивидуально и моделируется гауссианами. Каждый пиксель помечается как пиксель переднего или заднего фона, с применением статистического подхода, основанного на многократном гауссовом распределении.

Далее рассмотрим метод определения больших бинарных объектов. Большими бинарными объектами называют группу пикселей, использованных для хранения информации высокого уровня, описывающей такие объекты, как внешний вид или выполняемое действие. Вместе с данным методом применено морфологическое смыкание для заполнения дыр и коррекции ближайших пикселей для формирования единого соединенного компонента, после чего модуль сегментации переднего фона обрабатывает индивидуальные пиксели. Учитывая то,

что на сцене может появиться только один объект, используется пороговое значение, осуществляющее фильтрацию компонентов с площадью. Когда найден и выбран самый большой объект, система выполняет слежение за его перемещением. Далее рассматривается метод сравнения больших бинарных объектов, основанный на признаках, выделенных из точек интереса объектов.

Для слежения за перемещением пользователя рассмотрим методы определения и выделения признаков. В соответствии с запросом предоставления своевременного отклика в ассистивном жилом пространстве был выбран метод «быстрого» определения точек признаков в отличие от более медленных аналогов, например, метода Харриса или метода определения различий в гуассианах [30]. Ключевые точки выделяются в каждом кадре только из границ прямоугольников найденных частей переднего фона. Совместное применение границ прямоугольников и сегментации позволяет улучшить результаты определения объектов переднего фона. После определения ключевых точек выполняется выделение соответствующих признаков (Бинарные Робастные Независимые Элементарные Признаки) BRIEF из кадра [31]. Признаки BRIEF определяются при помощи сравнительно простых тестов различия интенсивности и являются хорошо различимыми даже при использовании нескольких бит. Ещё одним преимуществом является использование бинарного представления, что делает сравнения эффективными при использовании расстояния Хемминга.

2.2. Методы автоматического распознавания речи и акустических событий. При разработке ассистивного многомодального интеллектуального пространства использовалось многоканальное оборудование записи и обработки видео- и аудиосигналов, представленное на рисунке 1.

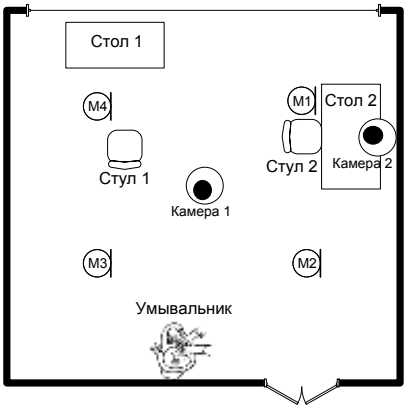

Рис. 1. Схема оборудования ассистивного многомодального пространства.

Лексический словарь системы распознавания речи состоит из пяти английских и пяти русских слов, кроме того, включает двенадцать типов акустических событий для различных типов активностей. Стандартная архитектура автоматической системы распознавания изолированной речи показана на рисунке 2. Представленная система может работать в двух режимах: обучение моделей и декодирование речи.

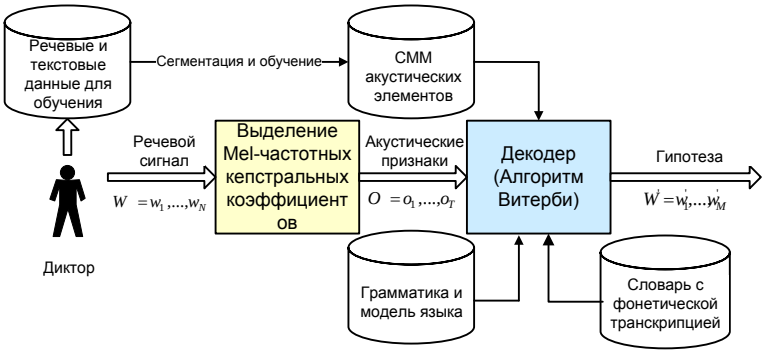

Рис. 2 Архитектура автоматической системы распознавания речи.

На рисунке 3 представлена система распознавания речи, использующая модифицированный алгоритм Витерби, так называемый метод передачи маркеров [32]. Синтаксис фразы описан простой грамматикой, позволяющей распознавать один элемент лексикона в гипотезе. Распознавание речевого сигнала происходит достаточно быстро (менее

чем 0,1 x Real-Time), поэтому результат доступен сразу же после окончания определения границ речи в аудиосигнале.

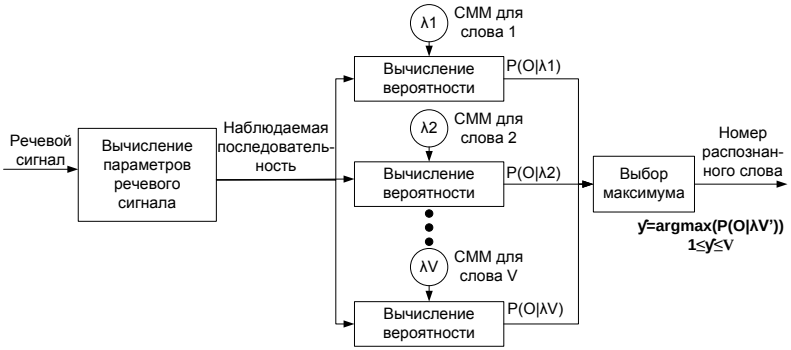

Рис. 3 Алгоритм распознавания слова в речевом сигнале.

Речевой корпус для обучения был записан в течение семинара eNTERFACE'2011 в условиях незначительных внешних шумов. В общей сложности было записано 75 минут речевых данных и сигналы, содержащие акустические события.

3. Классификация речевых команд и акустических событий.

Поскольку разрабатываемые технологии ассистивного интеллектуального пространства будут использоваться ограниченным кругом пользователей, то модуль обработки речевых сигналов может быть основан на дикторозависимом принципе. Более сложные дикторонезависимые приложения могут быть реализованы на основе методов, описанных в [33]. Рассмотрим классификацию речевых и акустических событий, включенных в базу данных и представленных на рисунке 4.

При создании базы данных для обучения в помещении были выбраны 5 позиций, в которых пользователь произносил речевую команду или моделировал акустическое событие. Четыре из пяти позиций были расположены под микрофонами, а пятая находилась в центре помещения. Каждая голосовая команда и каждое акустическое событие были имитированы и записаны 100 и 200 раз соответственно. Обозначения речевых команд и акустических событий представлены в таблицах 1 и 2 соответственно.

Создание базы данных для обучения, а также проведение экспериментов проводилось по сценарию, представленному в следующем разделе.

Рис. 4 Схема классификации речевых команд и акустических событий.

4. Эксперименты. Оценка работы системы обработки аудиосигнала производилась в ассистивном многомодальном пространстве в условиях незначительных внешних шумов [34]. Для проведения экспериментов было разработано несколько сценариев.

Для проведения первого эксперимента в помещении было выбрано пять контрольных точек. Данные точки располагались под четырьмя установленными микрофонами и в центре зала. Рассмотрим сценарий эксперимента и действий участника: (1) занять первую установленную контрольную точку; (2) воспроизвести выбранное событие или речевую команду пять раз с паузой в одну секунду между каждым моделированием; (3) перейти на следующую установленную контрольную точку.

В двух других сценариях также участвовал один человек, действия, выполняемые им, были максимально приближенными к условиям жилого помещения.

Таблица 1. **Обозначение речевых команд**

Речевая команда	Обозначение
Answer phone	Aph
Help	Hlp
No	N
Problem	Pro
Yes	Y
Да	D
Нет	nt
Ответить	ot
Помогите	pmg
Проблема	pbm

Таблица 2. **Обозначение акустических событий**

Акустическое событие	Обозначение
Аплодисменты	ap
Перемещение кресла	cm
Кашель	co
Стон	cr
Хлопанье дверью	ds
Падение	fa
Падение ключей	kd
Звон ключей	kj
Шелест бумаги	pw
Звонок телефона	pr
Шаг	st
Прочищение глотки	th

Рассмотрим план проведения первого сценария: (1) войти в помещение через дверь; (2) дойти до стола №1; (3) взять стакан воды со стола; (4) подойти и сесть на стул №1; (5) выпить воду; (6) воспроизвести кашель после питья воды; (7) встать, вернуться к столу №1 и поставить стакан обратно; (8) подойти к раковине и помыть руки; (9) выйти из комнаты.

Во втором сценарии участник выполняет следующие действия: (1) войти в помещение через дверь; (2) подойти и сесть на стул №2; (3) звонит телефон на столе №2; (4) произнести голосовую команду «Ответить»; (5) поговорить по телефону с вызывающим; (6) закончить диалог и подойти к столу №1; (7) взять металлическую миску и начать с ней движение в противоположной конец помещения; (8) уронить миску, но продолжать движение; (9) упасть; (10) воспроизвести акустическое событие «Стон».

В процессе проведения экспериментов в общей сложности было записано 2811 аудиофайлов, 1226 из которых содержали акустические события, а 1585 — речевые команды. Точность распознавания речевых команд и акустических событий представлена в таблицах 3 и 4 соответственно.

Результаты, представленные в таблице 3, показали, что большинство речевых команд распознавалось с точностью более 90%, однако присутствуют некоторые ошибки. Например, при произнесении команды «Problem» в некоторых случаях результатом распознавания яв-

лялось русское слово «Проблема», такая ошибка связана с наличием в обоих словах нескольких одинаковых фонем, но сами команды имеют одинаковые значения.

Таблица 3. **Точность распознавания речевых команд**

Количество речевых команд в тестовой базе данных		**Результат распознавания, %**									
		aph	**hlp**	**n**	**pro**	**y**	**d**	**nt**	**ot**	**pmg**	**pbm**
aph	388	100									
hlp	100		85								15
n	120			100							
pro	249				94						6
y	104					100					
d	100						100				
nt	100							100			
ot	118								100		
pmg	155				2					95	3
pbm	151				9						91

Таблица 4. **Точность распознавания акустических событий**

Общее число событий в тестовой базе данных		**Результат распознавания, %**											
		ap	**cm**	**co**	**cr**	**ds**	**fa**	**kd**	**kj**	**pw**	**pr**	**St**	**th**
ap	112	100											
cm	152		100										
co	72			100									
cr	108				100								
ds	111				2	98							
fa	108			1			63					36	
kd	129	15						75		10			
kj	44								100				
pw	68	4								96			
pr	32										100		
st	166			6								94	
th	124												100

Результаты, представленные в таблице 4, показали, что самый низкий процент распознавания 63% у события «Падение». Более чем в 30% случаев данное событие было распознано как событие «Шаг». Такие значительные ошибки связаны с тем, что при имитации события «Падения» использовался искусственный предмет, например, сумка, заполненная мягкой одеждой.

5. Заключение. Разработанная многомодальная ассистивная система для жилого пространства состоит из двух комплексов средств. Первый выполняет обработку видеопотоков для определения положения пользователя и слежения за его перемещением, а также анализа его действий. Ко второму комплексу относится система обработки аудиопотоков, предназначенная для автоматического распознавания речевых команд и акустических событий. В ходе разработки была записана база аудиоданных, содержащая голосовые команды и акустические события, которые были смоделированы и записаны 100 и 200 раз, соответственно. В процессе проведения экспериментов в общей сложности было записано 2811 аудиофайлов, 1226 из которых содержали акустические события и 1585 содержащих речевые команды. Результаты показали, что самый низкий процент распознавания акустических событий (63%) выявлен у события «Падение», а среди речевых команд - «Help», данная ошибка может быть связана с тем, что запись базы данных для обучения велась в условиях реверберации и незначительных внешних шумов, кроме того все акустические события были моделированы искусственно. Средняя точность распознавания акустических событий и речевых команд составила 93.8% и 96.5% соответственно.

Литература

1. *Alemdar H. and Ersoy C.* A Survey on Wireless Sensor Technologies for Health-care. Computer Networks, 2010.
2. *Koch S. and HÄagglund M.* Health informatics and the delivery of care to older people. Maturitas, May 2009.
3. *Sneha S. and Varshney U.* Enabling ubiquitous patient monitoring: Model, decision protocols, opportunities and challenges. Decision Support Systems, vol. 46, February 2009. pp. 606-619.
4. QuietCare Systems https://www.quietcaresystems.com
5. Philips LifeLine http://www.lifelinesys.com/content/lifelineproducts/classic-pendant.jsp
6. Intel Health Guide http://www.intel.com/healthcare/ps/healthguide/index.htm
7. *Wood A.,Stankovic J., Virone G., Selavo L., Zhimin H., Qiuhua C., Thao D., Yafeng W., Lei F., and Stoleru R.*, Context-aware wireless sensor networks for assisted living and residential monitoring, Network, IEEE, vol. 22, no.4, 2008. pp. 26-33.

8. *Huiyu Z., Hu H.* Human motion tracking for rehabilitation--A survey. In Biomedical Signal Processing and Control, Volume 3, Issue 1, January 2008. pages 1-18
9. *Iso-Ketola P., Karinsalo T., and Vanhala J.* HipGuard: A wearable measurement system for patients recovering from a hip operation, in Second International Conference on Pervasive Computing Technologies for Healthcare, 2008. pp. 196-199.
10. *Virone G., Wood A. D., Selavo L., Cao Q., Fang L., Doan T., He Z., and Stankovic J. A.* An Advanced Wireless Sensor Network for Health Monitoring, in Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare (D2H2), Arlington, VA, 2006
11. *Alemdar, H. Ö., Yavuz, G. R., Özen, M. O., Kara, Y. E., Incel, Ö. D., Akarun, L., & Ersoy, C.* Multi-modal fall detection within the WeCare framework. Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks – IPSN'10. 2010. pp. 436-437.
12. *Nakashima, H., Aghajan, H., Augusto, J. C., Nakashima, H., & Aghajan, H.* Handbook of Ambient Intelligence and Smart Environments. (H. Nakashima, H. Aghajan, & J. C. Augusto, Eds.). Boston, MA: Springer Verlag. 2009.
13. *Papadopoulos, A., Crump, C., & Wilson, B.* Comprehensive home monitoring system for the elderly. Wireless Health 2010 on - WH'10 2010. pp. 214-215.
14. *Kara, Y. E., & Akarun, L.* Human action recognition in videos using keypoint tracking. IEEE 19th Signal Processing and Communications Applications Conference (SIU) 2011. pp 1129-1132.
15. *Poppe, R., Elsevier B.V.* A survey on vision-based human action recognition. Image and Vision Computing, 28(6), 2010. pp.976-990.
16. *Weinland, D., Ronfard, R., & Boyer, E.* A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding, 115(2), 2011. pp.224-241.
17. *KaewTraKulPong, P., & Bowden, R.* An improved adaptive background mixture model for real-time tracking with shadow detection. Proc. European Workshop Advanced, 1(3), 2001. pp.1-5.
18. *Zivkovic, Z., & van der Heijden, F.* Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters, 27(7), 2006. pp.773-780.
19. *Kim, K., Chalidabhongse, T., Harwood, D., & Davis, L.* Real-time foreground–background segmentation using codebook model. Real-Time Imaging, 11(3), 2005. pp.172-185.
20. *Tuytelaars, T., & Mikolajczyk, K.* Local Invariant Feature Detectors: A Survey. Foundations and Trends in Computer Graphics and Vision, 3(3), 2007. pp.177-280.
21. *Кипяткова И. С., Карпов А. А..* Эксперименты по распознаванию слитной русской речи с использованием сверхбольшого словаря // Труды СПИИРАН. Вып. 12, СПб.: Наука, 2010, С. 63-74.
22. *Temko A., Malkin R., Zieger C., Macho D., Nadeu C.,* Acoustic event detection and classification in smart-room environments: Evaluation of child project systems // IV Jornadas en Tecnología del Habla, Zaragoza, Nov. 2006, pp. 5-11
23. *Wang D., Brown G.,* Computational Auditory Scene Analysis: Principles, Algorithms and Applications, Wiley-IEEE Press, 2006
24. *Temko A., Nadeu C.,* Acoustic event detection in meeting-room environments // Pattern Recognition Letters. Vol. 30. 2009. pp. 1281-1288
25. *Vacher, M, Istrate, D., Besacier, L., Castelli, E., Serignat, J.,* Smart audio sensor for telemedicine. In: Proc. Smart Object Conference 2003. pp.15-17
26. *Stäger, M., Lukowicz, P., Perera, N., Büren, T., Tröster, G., Starner, T.,* Sound button: Design of a low power wearable audio classification system. In: Proc. IEEE Int. Symp. on Wearable Computers, 2003. pp. 12–17

27. *Jianfeng, C., Jianmin, Z., Kam, A., Shue, L.,* An automatic acoustic bathroom monitoring system. In: Proc. IEEE Int. Symp. on Circuits and Systems 2005. 2. pp.1750-1753.
28. *Ронжин А.Л., Карпов А.А.*. Проектирование интерактивных приложений с многомодальным интерфейсом // Доклады ТУСУРа, № 1 (21), часть 1, 2010, С. 124-127.
29. *Ronzhin A.L., Karpov A.A.* Russian Voice Interface // MAIK Nauka/Interperiodica: Pattern Recognition and Image Analysis, 2007, Vol. 17, No. 2, pp. 321–336.
30. *Rosten, E., & Drummond, T.* Machine learning for high-speed corner detection. Computer Vision–ECCV 2006, 2006. 430–443.
31. *Calonder, M., Lepetit, V., & Fua, P.* BRIEF: Binary Robust Independent Elementary Features. (K. Daniilidis, P. Maragos, & N. Paragios, Eds.)Computer Vision – ECCV 2010. pp. 778-792
32. *Rabiner, L., & Juang, B.* Speech Recognition. In J. Benesty, M. M. Sondhi, & Y. Huang (Eds.), Springer Handbook of Speech Processing. Springer New York. 2008.
33. *Кипяткова И.С., Карпов А.А.* Автоматическая обработка и статистический анализ новостного текстового корпуса для модели языка системы распознавания русской речи // Информационно-управляющие системы. – СПб: СПбГУАП, № 4(47), 2010, С. 2-8.
34. *Ронжин Ал.Л.*. Способы оценивания систем аудиолокализации, выступающих в зале совещаний // Труды СПИИРАН. №2, Вып. 17, СПб.: Наука, 2011, С. 101-113.

Карпов Алексей Анатольевич — канд. техн. наук**,** старший научный сотрудник лаборатории речевых и многомодальных интерфейсов Учреждения Российской академии наук Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: автоматическое распознавание речи, многомодальные интерфейсы, аудиовизуальное распознавание и синтез речи. Число научных публикаций — 120. karpov@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Karpov Alexey Anatolyevich — PhD, senior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, multimodal interfaces, audio-visual speech recognition. The number of publications — 120. karpov@iias.spb.su; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Лале Акарун — доктор наук, профессор, зав. факультета вычислительной техники Богазичи Университета. Область научных интересов: обработка изображений, техническое зрение, компьютерная графика. Число научных публикаций — 200. akarun@boun.edu.tr; Богазичи Университет, 34342 Бебек, Стамбул, Турция; т. +90 (212) 3596858, факс +90 (212) 287 2461.

Lale Akarun — PhD, Professor, Chair of Computer Engineering Department at Boğaziçi University. Research interests: image processing, computer vision, and computer graphics. The number of publications — 200. akarun@boun.edu.tr; Boğaziçi University, 34342 Bebek, İstanbul, Turkey; office phone +90 (212) 3596858, fax +90 (212) 287 2461.

Ронжин Александр Леонидович — младший научный сотрудник лаборатории речевых и многомодальных интерфейсов Учреждения Российской академии наук Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: технологии интеллектуального пространства, аудиолокализации, техническое зрение. Число научных публикаций — 23. ronzhinal@iias.spb.su;

СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Ronzhin Alexander Leonidovich — junior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: smart space, sound source localization, computer vision. The number of publications — 23. ronzhinal@iias.spb.su; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Поддержка исследований. Данное исследование поддержано в рамках федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы» (ГК № 11.519.11.4025)

Рекомендовано лабораторией речевых и многомодальных интерфейсов, заведующий лабораторией Ронжин Ан.Л., д-р техн. наук, доцент.
Статья поступила в редакцию 01.11.2011.

РЕФЕРАТ

Карпов А.А., Акарун Л., Ронжин Ал.Л. **Многомодальные ассистивные системы для интеллектуального жилого пространства.**

В статье представлен обзор ассистивных и жилых интеллектуальных пространств. Также описывается разработанная многомодальная ассистивная система для интеллектуального жилого пространства, которая состоит из двух комплексов средств. Первый выполняет обработку видеопотоков для определения положения пользователя и слежения за его перемещением, а также анализа его действий при помощи метода моделей Гауссовых смесей. Данный метод выполняет сегментацию объектов переднего фона и в дополнение осуществляет устранение теней, что применяется для улучшения результатов определения при помощи удаления артефактов соответствующих малым изменениям в цвете или интенсивности. Ко второму комплексу относится система обработки аудиопотоков, предназначенная для автоматического распознавания речевых команд и акустических событий.

Разработанная система автоматического распознавания речи многоязычна и позволяет распознавать слова, произнесенные на английском или русском. Для обработки аудиосигнала в помещении были установлены четыре микрофона Oktava MK-012 и многоканальная аудиоплата M-Audio ProFire 2626. В каждом микрофоне установлен кардиоидный капсюль, позволяющий записывать звук в диапазоне углов от -60^0 до 60^0 с приблизительно равным усилением. В процессе проведения экспериментов, проводившихся на семинаре eNTERFACE'2011, было записано 2811 аудиофайлов, 1585 содержащих речевые команды и 1226 содержали акустические события, средняя точность распознавания составила 96.5% и 93.8% соответственно.

SUMMARY

Karpov A.A., Akarun L., Ronzhin Al.L. **Multimodal assistive systems for a smart living environment.**

The paper proposes a survey of assistive smart spaces and ambient assisted living environments. Also design of a multimodal assistive system for a smart living environment is presented. The system consists of two software complexes. The first one provides video signal processing and surveillance for detecting and tracking a user as well as analysis of his/her activity by the method of Gaussian mixture models. The second software complex provides audio signal processing for automatic recognition of speech messages and non-speech acoustic events.

The developed automatic speech recognition system is multilingual one and is able to recognize words both in English and in Russian. Four Oktava MK-012 condenser microphones and one multi-channel sound board M-Audio ProFire 2626 are used for audio signal capturing in the model of smart living environment. In the experiments made during eNTERFACE'2011 workshop 2811 wave files have been recorded, which consist of 1226 files with non-speech acoustic events and 1585 other files with speech commands of users. Recognition rate for speech commands and non-speech acoustic events was 96.5% and 93.8%, respectively.