**Олейник Андрей Леонидович Методы и алгоритмы взаимной реконструкции лицевых и голосовых данных**

ВВЕДЕНИЕ ДИССЕРТАЦИИ (ЧАСТЬ АВТОРЕФЕРАТА)

на тему «Методы и алгоритмы взаимной реконструкции лицевых и голосовых данных»

Общая характеристика работы

Актуальность темы. Человеческое общение включает в себя множество составляющих: голос и его интонацию, мимику, жесты. Поэтому при решении различных задач человекомашинного общения и построения биометрических систем распознавания личности прибегают к использованию нескольких модальностей1, основными из которых являются визуальная и звуковая. Один из подходов к задействованию этих модальностей — использование лицевых и голосовых данных, для чего существует ряд технических средств. При этом многие устройства (в основном мобильные) оснащены не только микрофонами и камерами, но и дополнительными сенсорами: датчиками глубины, а также камерами, работающими в ближнем и тепловом инфракрасных спектрах. Это положило начало исследованиям, связанным с анализом и обработкой мультисенсорных изображений лиц: фотографий в видимом, ближнем инфракрасном и тепловом спектрах, карт глубины («range image», «depth map») и скетчей2.

Совместное использование мультисенсорных лицевых и голосовых данных позволяет расширять класс решаемых задач и разрабатывать принципиально новые решения в области мультимедийных систем, человеко-машинного общения, анимации, робототехники, лицевой и голосовой биометрии. С другой стороны, из-за различной природы исходных данных (физической и сенсорной) трудно обеспечить согласованность процессов их получения, анализа, обработки, распознавания и синтеза. Преодоление связанных с этим проблем и трудностей требует разработки и исследования новых методов, алгоритмов, моделей и программных средств.

Актуальность темы также подтверждается множеством докладов на международных конференциях по проблемам компьютерного зрения и обработки речи и статей в научных журналах. Примерами тому служат журналы «Труды СПИИРАН», «Автоматика и телемеханика», «IEEE Transactions on Pattern Analysis and Machine Intelligence», «ACM Transactions on Intelligent Systems and Technology», «IEEE Transactions on Multimedia», «Pattern Recognition Letters». Многие передовые исследования размещены в виде препринтов на arXiv.org.

Степень разработанности темы. Различные аспекты проблем обработки, распознавания, анализа и синтеза бимодальных и мультисенсорных данных рассмотрены в работах А.А. Карпова, А.Л. Ронжина, О.О. Басова, И.А. Саитова, N. de Freitas, Stan Z. Li, A.K. Jain, B. Klare, K. Nandakumar, A. Ross, K. Wang, M. Zelezny, R. Chellappa, H. Kaya, X. Tang, X. Wang, X. Gao. Исследования в этой области обработки проводят ведущие отечественные и зарубежные научные и научно-технические организации: СПИИРАН, ТУСУР, Академия ФСО России, Oxford University, Chinese Academy of Sciences, University of Maryland, Michigan State University, Massachusetts Institute of Technology, University of West Bohemia in Pilsen, Namik Kemal University, Chinese University of Hong Kong, Xidian

1 Модальность — способ представления информации, пригодный для восприятия человеком.

2Скетчи — фотороботы и рисунки, выполненные художником вручную или на компьютере.

University, The Hong Kong Polytechnic University, Google DeepMind, Microsoft Corporation, Rank One Computing.

Рассматриваемая область включает широкий спектр задач: от распознавания речи по голосу и лицу до поиска преступников по фотороботам. Под каждую из этих задач разработан ряд узкоспециализированных решений, позволяющих решать их достаточно эффективно. При этом на сегодняшний день не предложена технология, позволяющая решать вышеупомянутые задачи на основе одних и тех же принципов, моделей, методов и алгоритмов. Такая технология позволила бы реализовать операции по анализу и обработке бимодальных и мультисенсор-ных данных в рамках универсального набора программных модулей.

При разработке такой технологии возникают два аспекта, первый из которых — получение, предобработка и подготовка исходных данных. В связи с этим решаются задачи детектирования (поиска) лиц на изображении, трекинга (отслеживания) лиц в видеопотоке, удаления шума из аудиозаписей, синхронизации аудио- и видеопотоков. Несмотря на обилие разнообразных решений в данной области, класс исходных данных настолько широк, что порождает потребность в новых методах и алгоритмах.

Второй аспект — объединение бимодальных и мультисенсорных данных, выполняемое на основе слияния3, взаимной реконструкции (двунаправленного преобразования) и генерации популяций (наборов подобных, но не идентичных данных). Один из путей реализации этих операций — использование глубоких нейронных сетей, которые позволяют достичь высокой точности и надёжности. Однако это требует больших обучающих выборок, а в случае бимодальных и мультисенсорных данных такие выборки зачастую недоступны.

Альтернативой глубоким нейронным сетям являются методы проекции в собственные пространства4 (или проекционные методы): анализ главных компонент, линейный дискриминантный анализ, канонический корреляционный анализ и метод частичных наименьших квадратов5. Основная идея этих методов, предложенная ещё в первой половине XX века, заключается в построении собственных базисов. Свойства эти базисов позволяют легко решать задачи взаимной реконструкции данных и генерации популяций. Проекционные методы универсальны, математически обоснованны, интерпретируемы и при этом не требуют объёмных обучающих выборок (как методы глубокого обучения).

Проблемой является то, что проекционные методы не были разработаны для обработки объектов, представленных в матричной форме. К таким объектам относятся изображения и видеозаписи. В этом случае проявляется проблема малой выборки, а при увеличении объёма выборок необходимо решать задачи на

3англ. fusion, в русскоязычной литературе иногда используют термин «фузирование»

4Под собственным пространством понимается линейное пространство, порождённое собственным базисом. В свою очередь, собственный базис — это система базисных векторов, являющаяся решением некоторой задачи на собственные значения

5Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Canonical Correlation Analysis (CCA), Partial Least Squares (PLS).

собственные значения высокой размерности. Это приводит к большим вычислительным затратам и неустойчивости получаемых решений. Преодолеть эти трудности можно с помощью двумерных проекционных методов6. Они были предложены недавно, а возможности их применения к задачам обработки бимодальных и мультисенсорных данных изучены недостаточно глубоко.

Таким образом, актуальна задача разработки и исследования моделей, методов и алгоритмов взаимной реконструкции мультисенсорных лицевых и голосовых данных на основе проекционных методов (в том числе двумерных). При этом важно учитывать особенности получения, предобработки и подготовки исходных данных.

Целью диссертации является исследование и развитие проекционных методов и разработка на их базе решений для анализа и обработки мультисенсор-ных лицевых и голосовых данных в широком классе сценариев с высокой вычислительной эффективностью и при малых обучающих выборках.

Для достижения поставленной цели решались следующие задачи:

1. Исследование различных путей развития проекционных методов для анализа и обработки мультисенсорных лицевых и голосовых данных и выбор подходов для решения задачи их взаимной реконструкции.

2. Определение состава процедур предобработки и подготовки исходных данных (изображений, аудио- и видеозаписей), а также разработка в рамках этих процедур алгоритма трекинга лиц в видеопотоке.

3. Разработка методов взаимной реконструкции мультисенсорных лицевых и голосовых данных, учитывающих специфические особенности различных практических задач и сценариев.

4. Алгоритмическая и программная реализация разработанных на предыдущих этапах решений.

5. Экспериментальная оценка предложенных решений, выработка рекомендаций их по применению и дальнейшему развитию.

Объект исследования. Наборы бимодальных и мультисенсорных данных, модели, методы и алгоритмы их анализа, распознавания и обработки.