УДК 004.932.2

МЕТОД ИНТЕРПРЕТИРУЕМОГО АУДИОВИЗУАЛЬНОГО АНАЛИЗА ПЕРСОНАЛЬНЫХ КАЧЕСТВ ЛИЧНОСТИ НА ОСНОВЕ ОБЪЕДИНЕННОГО ПРЕДСТАВЛЕНИЯ МОДАЛЬНОСТЕЙ

Гранкина Е. Г.[1]
Научный руководитель – канд. техн. наук Рюмин Д. А.[1]
[1] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН)
liza.grankina@gmail.com

Введение

Персональные качества личности отражают устойчивые индивидуальные различия в поведении и коммуникативных проявлениях человека. В задачах подбора персонала, дистанционного обучения и пользователь-ориентированных цифровых сервисов требуется быстрое, воспроизводимое и масштабируемое оценивание таких качеств, тогда как традиционная психологическая экспертиза характеризуется высокой трудоемкостью, субъективностью [1,2]. Современные методы глубокого обучения позволяют анализировать аудиовизуальные данные, однако их практическое применение ограничивается недостаточной интерпретируемостью и снижением устойчивости на независимых данных [1,2]. В связи с этим актуальной задачей является разработка метода интерпретируемого аудиовизуального анализа, который обеспечивает не только точную оценку персональных качеств личности, но и предоставляет понятные объяснения, пригодные для проверки и практического использования.

Основная часть

Цель работы заключается в разработке метода интерпретируемого аудиовизуального анализа персональных качеств личности по модели Big Five на основе объединенного представления модальностей и вычислительно эффективной обработки видеопоследовательности и речевого сигнала с последующим поздним слиянием модальностей.

В качестве целевого психометрического описания личности используется пятифакторная модель Big Five, включающая: открытость опыту (Openness), добросовестность (Conscientiousness), экстраверсию (Extraversion), доброжелательность (Agreeableness) и негативную эмоциональность (Neuroticism) [3]. Экспериментальная проверка выполняется на корпусе First Impressions V2, содержащем около 10 тыс. коротких видеоклипов и экспертные аннотации по указанным пяти качествам [1].

Для анализа видеопоследовательности используется вычислительно эффективный визуальный энкодер на основе state-space архитектур (в том числе VSSD), предназначенный для извлечения темпоральных признаков области лица из последовательности кадров при умеренных вычислительных затратах [4]. Речевой сигнал нормализуется и преобразуется в акустические признаки: лог-мел-спектрограммы или мел-частотные кепстральные коэффициенты, а также просодические характеристики (основная частота, энергия, темп речи). Далее признаки речевого сигнала агрегируются по времени с формированием фиксированного вектора признаков. Позднее слияние модальностей реализуется на уровне прогнозов, при этом для каждого из пяти качеств Big Five вычисляются частные оценки по визуальному и речевому каналам и формируется итоговый прогноз с учетом обучаемых весов их вкладов. Интерпретируемость обеспечивается раздельным анализом модальностей и построением карт значимости (Grad-CAM), локализующих области кадра, оказывающие наибольшее влияние на оценку конкретного качества [5]. Экспериментальная проверка проводится на корпусе First Impressions V2. Точность оценивания измеряется по средней абсолютной ошибке (*MAE*) для каждого качества, а для интегрального сравнения

используется сводная метрика средней точности (*mACC*), принятая для данного корпуса [1]. Дополнительно анализируются вычислительные характеристики (время обработки видеоклипа и потребление ресурсов), при этом выгодный баланс «точность/скорость» достигается за счет линейной по длине последовательности обработки видеоданных и вычислительно эффективной архитектуры визуального энкодера [4].

Выводы

Разработан метод интерпретируемого аудиовизуального анализа персональных качеств личности по модели Big Five, основанный на совместной обработке видеопоследовательности и речевого сигнала и последующем позднем слиянии модальностей. Для визуального анализа применен вычислительно эффективный энкодер класса state-space, обеспечивающий извлечение темпоральных признаков области лица при умеренных вычислительных затратах. Интерпретируемость результатов обеспечивается построением карт значимости Grad-CAM, позволяющих выявлять области кадра, наиболее влияющие на оценку каждого качества. Экспериментальная проверка на корпусе First Impressions V2 подтверждает работоспособность метода и позволяет оценивать точность по MAE для каждого качества и по сводной метрике mACC. Полученные результаты могут быть использованы в прикладных системах предварительного оценивания персональных качеств личности, включая задачи подбора персонала, дистанционного обучения и пользователь-ориентированных сервисов.

Литература

1. Ponce-López V., Chen B., Oliu M., Corneanu C., Clapés A., Guyon I., Baró X., Escalante H.J., Escalera S. ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results // Computer Vision - ECCV 2016 Workshops. 2016. P. 400-418. DOI: 10.1007/978-3-319-49409-8_32.
2. Jacques Junior J.C.S., Güçlütürk Y., Pérez M., Güçlü U., Andujar C., Baró X., Escalante H.J., Guyon I., van Gerven M.A.J., van Lier R., Escalera S. First Impressions: A Survey on Vision-Based Apparent Personality Trait Analysis // IEEE Transactions on Affective Computing. 2022. Vol. 13, No. 1. P. 75-95. DOI: 10.1109/TAFFC.2019.2930058.
3. McCrae R.R., John O.P. An Introduction to the Five-Factor Model and Its Applications // Journal of Personality. 1992. Vol. 60, No. 2. P. 175-215. DOI: 10.1111/j.1467-6494.1992.tb00970.x.
4. Shi Y., Li M., Dong M., Xu C. VSSD: Vision Mamba with Non-Causal State Space Duality // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2025. P. 10819–10829.
5. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017. P. 618–626. DOI: 10.1109/ICCV.2017.74.