Веселовский Владимир Владимирович,
магистрант, Финансовый Университет
Veselovskiy Vladimir Vladimirovich, Financial University

Фазлыев Руслан Фанузович,
магистрант, Финансовый Университет
Fazlyev Ruslan Fanuzovich, Financial University

АНСАМБЛЕВЫЙ ПОДХОД К АНАЛИЗУ ВИЗУАЛЬНО НАБЛЮДАЕМЫХ ХАРАКТЕРИСТИК ЛИЧНОСТИ И ИНТЕРЕСОВ ПО ИЗОБРАЖЕНИЯМ ЧЕЛОВЕКА
RESEARCH AND DEVELOPMENT OF METHODS FOR RECOGNIZING PERSONALITY TRAITS AND INTERESTS FROM IMAGES CONTAINING HUMANS

Аннотация. В статье рассматривается подход к анализу изображений с присутствием человека для выявления визуально наблюдаемых характеристик и предполагаемых интересов пользователя. Показано, что ансамблевое использование моделей компьютерного зрения позволяет объединять признаки внешности, эмоций, стиля и контекста сцены. Особое внимание уделяется интерпретируемости результатов и ограничению выводов рамками воспринимаемого визуального образа.

Abstract. The article presents a condensed version of a study devoted to the development of an image analysis system for images containing humans. The rejection of direct psychodiagnostic interpretation of photographs is justified, and an ensemble-based approach is proposed, in which personality is described through visually observable attributes, scene context, emotional state, and perceived impression. The system integrates branches for attribute analysis, face analysis, emotion recognition, object detection, context analysis, and text description generation. Experimental evaluation on 15 images demonstrated moderately good agreement with an external multimodal model: the overall similarity score reached 0.656, while the soft F1-score for the complete feature set was 0.743. The obtained results confirm the перспективность of the modular architecture while preserving validity limitations and maintaining ethical caution.

Ключевые слова: Компьютерное зрение, анализ изображений, воспринимаемая личность, визуальные атрибуты, Big Five, CLIP, мультимодальные модели.

Keywords: Computer vision, image analysis, perceived personality, visual attributes, Big Five, CLIP, multimodal models.

Введение

В условиях цифровизации социальные сети и мультимедийные сервисы формируют значительный массив визуальных данных, связанных с самопрезентацией человека. Фотография пользователя содержит не только признаки внешности, но и информацию о позе, мимике, одежде, предметном окружении, типе сцены и вероятном сценарии деятельности. Поэтому визуальный контент может рассматриваться как источник сведений о наблюдаемых интересах и поведенческих паттернах. Вместе с тем изображение не позволяет достоверно установить внутренние личностные свойства человека: оно фиксирует конкретный момент и зависит от условий съемки, выбора кадра и культурного контекста.

Классические модели личности, включая пятифакторную модель Big Five, создавались для психометрического описания устойчивых индивидуальных различий и требуют применения опросников или специально организованного наблюдения [1]. Исследования цифровых следов и изображений профиля показывают, что онлайн-активность и визуальная самопрезентация могут статистически соотноситься с пользовательскими характеристиками [2; 3]. Однако в задачах компьютерного зрения корректнее говорить не о диагностике личности, а о apparent personality – воспринимаемом образе, который складывается из внешних сигналов [4].

Материалы и методы

Исходная задача была представлена как комплекс взаимосвязанных подзадач компьютерного зрения. Отказ от единой универсальной модели обусловлен тем, что разные признаки имеют различную природу: эмоция определяется по лицу, стиль зависит от одежды и композиции, контекст связан с окружением, а интересы часто проявляются через объекты и повторяющиеся сценарии. Поэтому система реализована как ансамбль аналитических ветвей, результаты которых затем агрегируются в единое визуально-семантическое представление.

На вход системы подается изображение с присутствием человека. После проверки формата выполняются преобразование к RGB, нормализация и подготовка нескольких представлений кадра: полного изображения, области лица и версии для общего анализа. Основная ветвь решает задачу многометочного распознавания визуальных и воспринимаемых атрибутов. Для ее разработки использовался датасет изображений с описаниями первого впечатления; текстовые описания были преобразованы в структурированную разметку с использованием языковой модели Qwen2.5-3B-Instruct. После фильтрации демографических и школьно-возрастных категорий пространство основной модели составило 40 недемографических атрибутов.

В ходе разработки сравнивались ResNet50, EfficientNetV2-S, CLIP-like модель и частично дообученная CLIP-модель. Использование визуально-языковой логики связано с тем, что модели семейства CLIP позволяют сопоставлять изображения и текстовые описания в едином семантическом пространстве [5]. Дополнительные ветви системы включают эмоциональный анализ на основе лицевых изображений, определение базовых лицевых и демографических признаков, классификацию формы лица, выявление контекста и объектов, а также генерацию текстового описания. Эмоция рассматривается только как состояние в момент съемки, а не как устойчивая черта личности; демографические признаки используются как элементы визуального профиля с учетом риска смещений обучающих данных.

Итоговые данные приводятся к структурированному формату JSON. В нем признаки распределяются по смысловым блокам: базовые визуальные характеристики, внешность, стиль, воспринимаемое впечатление, контекст сцены, объекты, текстовое описание и дополнительные результаты. Такая организация позволяет сохранять различие между непосредственно наблюдаемыми признаками и интерпретационными выводами, а также делает систему расширяемой: отдельные модели могут заменяться без полной переработки конвейера.

Результаты и обсуждение

Оценка качества проводилась на наборе из 15 изображений людей. В качестве внешнего ориентира использовалась крупная мультимодальная модель Алиса AI. Она не рассматривалась как абсолютная эталонная разметка, поскольку для признаков стиля, впечатления и контекста часто отсутствует единственный правильный ответ. Цель эксперимента заключалась в определении степени согласованности двух систем при одинаковом изображении и едином структурированном формате результата.

Обе системы возвращали базовые визуальные признаки, описания внешности, стиль, впечатление, контекст, объекты и текстовые поля. Перед сравнением выполнялась нормализация: приведение к единому регистру, удаление лишних пробелов и унификация близких формулировок. Для оценки использовались точность, полнота, F1-мера, коэффициент Жаккара, а также мягкие метрики, учитывающие смысловую близость признаков. Текстовые описания сопоставлялись по семантической близости, поскольку разные модели могут описывать одну сцену различными словами.

Средний интегральный показатель близости разработанной системы к референсной модели составил 0,656, медианное значение – 0,650, стандартное отклонение – 0,060. Базовые визуальные признаки показали жесткое совпадение 0,617 и мягкое сходство 0,675. Строгое сравнение тегов оказалось ниже: средняя F1-мера составила 0,114, коэффициент Жаккара – 0,071. Однако мягкая F1-мера по общему объединенному списку признаков достигла 0,743.

Это показывает, что часть расхождений вызвана не отсутствием смысловой близости, а различиями словаря, уровня детализации и распределения признаков по блокам.

Наиболее устойчивыми оказались контекст сцены и объекты: их мягкая F1-мера составила 0,610 и 0,576 соответственно. Эти признаки имеют выраженную визуальную основу и поэтому определяются согласованнее. Признаки внешности показали средний уровень близости, тогда как стиль и воспринимаемое впечатление оказались наиболее сложными: мягкая F1-мера составила 0,178 и 0,088. Данный результат согласуется с теоретической природой задачи apparent personality: чем выше уровень интерпретации, тем сильнее влияние словаря модели, субъективности наблюдателя и культурных ожиданий [4].

Текстовые описания продемонстрировали высокую семантическую близость: среднее значение составило 0,828, а для краткого итогового описания – 0,934. Эти показатели подтверждают способность системы формировать связное описание изображения, близкое к результату крупной мультимодальной модели. Однако высокая текстовая близость не должна трактоваться как полная визуальная точность, поскольку на нее влияет единый формат ответа и шаблонность формулировок. Поэтому ключевым направлением развития остается нормализация словаря тегов, расширение тестовой выборки и привлечение независимой экспертной оценки.

Заключение

Предложенная система подтверждает перспективность ансамблевого подхода к анализу изображений человека. Она формирует многоуровневое описание наблюдаемых признаков, эмоционального состояния, контекста, объектов и осторожных впечатлений, не подменяя психологическое тестирование. Сравнение с внешней мультимодальной моделью показало интегральный показатель 0,656 и мягкую F1-меру общего списка 0,743. Дальнейшее развитие системы целесообразно связывать с расширением данных, экспертной разметкой и унификацией словаря признаков

Список литературы:

1. McCrae R.R., John O.P. An Introduction to the Five-Factor Model and Its Applications // Journal of Personality. – 1992. – Vol. 60. – № 2. – P. 175–215.

2. Kosinski M., Stillwell D., Graepel T. Private Traits and Attributes are Predictable from Digital Records of Human Behavior // Proceedings of the National Academy of Sciences. – 2013. – Vol. 110. – № 15. – P. 5802–5805.

3. Liu L., Preotiuc-Pietro D., Samani Z.R., Moghaddam M.E., Ungar L.H. Analyzing Personality through Social Media Profile Picture Choice // Proceedings of the International AAAI Conference on Web and Social Media. – 2016. – Vol. 10. – № 1. – P. 211–220.

4. Jacques Junior J.C.S., Güçlütürk Y., Pérez M. et al. First Impressions: A Survey on Vision-Based Apparent Personality Trait Analysis // IEEE Transactions on Affective Computing. – 2022. – Vol. 13. – № 1. – P. 75–91.

5. Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S. et al. Learning Transferable Visual Models From Natural Language Supervision // Proceedings of the 38th International Conference on Machine Learning. – 2021. – P. 8748–8763