# СУЧАСНІ ІНСТРУМЕНТИ АВТОМАТИЗОВАНОЇ КОНВЕРСІЇ ВІДЕОКОНТЕНТУ В СТАТИЧНІ ЗОБРАЖЕННЯ ТА ТЕКСТОВУ ІНФОРМАЦІЮ

***Левикін І.В.**, професор, кафедра МСТ, ХНУРЕ*
***Шимко Д.І.**, аспірант, кафедра МСТ, ХНУРЕ*

***Abstract.** This paper explores the conversion of video content into text and images using AI tools, focusing on preserving the visual and auditory context. It examines current transcription services, such as OpenAI's Whisper, and discusses the challenges of maintaining content accuracy while automating the conversion process.*

Відео є одним із найбільш повних та вичерпних видів медіаматеріалів, оскільки воно одночасно дозволяє сприймати інформацію у візуальному та аудіо форматах, що розподіляє потік інформації між кількома органами відчуття. У порівнянні з текстовими матеріалами, де для сприйняття матеріалу необхідно його повністю прочитати, що не завжди дає змогу адекватно охопити об'єкт вивчення, відео має значну перевагу. Однак збереження матеріалів у відеоформаті також має свої недоліки: велика вага файлів, низька якість відео чи аудіозапису, надмірна кількість зайвих кадрів, які можна пропустити під час вивчення матеріалу (наприклад, заставки та тематичні вставки між частинами відео), відсутність можливості швидкого пошуку в межах матеріалу та залежність від технічних характеристик пристроїв для відтворення відео.

Текстові матеріали, натомість, дозволяють переходити до будь-якої частини через швидкий пошук чи перехід за змістом, що створюється відповідно до структури матеріалу. Текст займає мінімум місця порівняно з іншими форматами медіаматеріалів і дозволяє зберігатися та відтворюватися на будь-якому пристрої, що додає зручності та підвищує доступність.

Велика кількість медіаматеріалів, таких як матеріали ЗМІ, подкасти або типові розважальні та навчальні ресурси, мають початкову форму відео, але їх конверсія в текстовий формат потребує перегляду та ручного вводу тексту, а також підбору релевантних зображень, що відповідають кожній частині тексту. Це підвищує потребу у можливості автоматизованої конверсії відео в текст з максимально можливим збереженням тих якостей, які були властиві початковому відеоформату, але не можуть бути відображені лише текстом.

Замість порівняння переваг та недоліків різних типів медіа форматів і вибору між ними, створення кросплатформенного мультимедійного контенту дозволяє охопити більшу кількість аудиторії шляхом публікації в різних форматах на відповідних медіаресурсах та соціальних платформах. Це робить автоматизацію процесу конверсії відеоматеріалів у текст та зображення актуальним елементом для збільшення охоплення.

Використання текстових матеріалів, доповнених візуальними елементами, є більш сприятливим для опанування під час навчання, ніж

матеріали, які складаються виключно з тексту без додаткових візуальних компонентів [1].

При звичайній транскрипції відео в текст втрачається значна частина змісту, як візуальної, так і аудіо складової. Огляд конкретного об'єкта у вигляді тексту втрачає важливу інформацію, якщо відео частково або повністю покладається на візуальну подачу.

Однією з ключових особливостей якісної трансформації відео в статичні текстові матеріали є збереження візуальної та контекстної складової, яку не можна передати звичайним текстом. Зображення є звичайним елементом, що доповнює текстові матеріали, але оскільки візуальні матеріали у відео можуть бути динамічними, при автоматизованій конверсії важливо точно відображати зображення в тих місцях тексту, які відповідають контексту. Якщо у відео змінюються доповідачі чи тема, навіть невелика затримка між текстом і скріншотом зображення може бути некоректною.

Метою дослідження є тестування можливостей сучасних інструментів для конверсії відеоматеріалів у текст та зображення з доповненням контексту шляхом структуризації тексту та відповідного відображення релевантних зображень до частин тексту.

Сучасні сервіси, що використовують штучний інтелект для транскрипції аудіо, як модель "Whisper" від OpenAI, дозволяють отримувати часові мітки для кожної визначеної частини тексту, що дозволяє поєднати їх із зображенням, яке було відображене у відео в той самий момент часу [2]. Також деякі сервіси, як моделі від AssemblyAI, здатні визначати різні голоси в аудіо та відокремлювати промовлені фрази при кожній зміні доповідача, що значно покращує сприйняття тексту завдяки його структуризації [3].

Аналіз існуючих інструментів та порівняння результатів найбільш популярних постачальників на реальних прикладах конверсії відео у структурований текст, доповнений зображеннями, дозволить розробити практичні методи для створення систем автоматизованої конверсії медіаматеріалів та визначити обмеження, що виникають через універсальність підходу до відеоматеріалів.

**Список літератури**

1. Mayer, R.E., & Moreno, R. (1998). A cognitive theory of multimedia learning: Implications for design principles. https://www.webcitation.org/670fh49gW?url=http://www.unm.edu/~moreno/PDFS/chi.pdf.

2. OpenAI. Speech to text – OpenAI API. https://platform.openai.com/docs/guides/speech-to-text.

3. AssemblyAI. Speaker diarization | AssemblyAI | Documentation. https://www.assemblyai.com/docs/speech-to-text/pre-recorded-audio/speaker-diarization.