УДК 004.032.26, 004.932.4

АНИМАЦИЯ ИЗОБРАЖЕНИЙ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

Климович А.Н., студент гр.150501

Белорусский государственный университет информатики и радиоэлектроники[1]
г. Минск, Республика Беларусь

Куприянова Д.В. – магистр техн. наук

Аннотация. Проанализированы современные методы анимации изображений на основе нейронных сетей. В качестве основной архитектуры выбрана First Order Motion Model (FOMM). Выполнен обзор и сравнительный анализ других методов, таких как X2Face и Monkey-Net. Проведена оптимизация FOMM для работы в реальном времени. Выполнено тестирование модели и собраны метрики ее работы.

Ключевые слова. Генеративно-состязательная сеть, анимация изображений, First Order Motion Model, перенос движений, генерация видео, X2Face, Monkey-Net, ключевые точки, обработка изображений, сравнение методов анимации.

Введение. Современные цифровые технологии оказывают заметное влияние на визуальный контент, трансформируя его в различных прикладных сферах – от кинематографа и видеоигр до AR/VR-приложений и социальных медиа. Одной из основных задач в этой области является реалистичная анимация статичных изображений, позволяющая «оживлять» портреты, создавать цифровые аватары или автоматизировать производство мультимедийного контента. Традиционные методы, основанные на ручной анимации или алгоритмах морфинга, требуют значительных временных затрат и не обеспечивают естественности движений. Также для данного подхода требуются дополнительные сведения об объекте анимации, например, его 3D-модель, необходимая для более корректной генерации движения, что стимулирует развитие нейросетевых методов, способных обходиться без явного 3D-представления.

Прорыв в области анимации изображений связан с появлением глубокого обучения и соответствующих моделей. В результате обучения на множестве видеороликов, изображающих объекты одной категории (например, лица, человеческие тела), такие сети позволяют анимировать объекты, относящиеся к данной категории. Данная генеративная способность позволяет автоматически переносить мимику и движения с исходного видео на целевое изображение, сохраняя его детализацию и реалистичность. В отличие от классических методов, нейросетевые решения обладают следующими преимуществами:

– минимизируют необходимость ручного вмешательства;
– обеспечивают плавность анимации даже для сложных сцен;
– адаптируются к различным объектам (лица, тела, животные).

Целью данной работы является проведение сравнительного анализа современных методов анимации изображений на основе нейронных сетей, оценив их качество результатов на тестовых данных, а также выявить преимущества First Order Motion Model (FOMM) [1] в задачах сохранения деталей и естественности движений.

Обзор методов анимации изображений. Большинство последних решений рассматриваемой задачи основываются на глубоком обучении моделей, в основе которых лежат генеративно-состязательные нейросети (Generative Adversarial Network, GAN) и вариационные автоэнкодеры (Variational Autoencoder, VAE). Данные модели обычно используют предобученные модули для поиска ключевых точек объектов на изображении. Главная проблема такого подхода – данные модули способны распознавать только объекты, на которых они были обучены.

X2Face [2] – нейросеть с самообучением для создания дипфейков (deepfake), которую разработали исследователи из Оксфорда. Она использует лицо другого человека для управления позой и выражение целевого лица. Модель решает задачу переноса движений, разделяя ее на две подзадачи, реализуемые двумя подсетями. Встраиваемая сеть изучает фронтальное изображение лица, а управляющая наделяет это лицо позой и выражением человека, анимирующего изображение. Однако результирующее изображение нередко содержит визуальные артефакты – искажения формы лица, размытость или некорректную передачу мимики, что делает подмену заметной для наблюдателя.

MonkeyNet [3] – открытая нейросеть, разработанная итальянскими исследователями, которая состоит из следующих модулей:

– детектор, который обучается без учителя и извлекает ключевые точки объекта;
– сеть прогнозирования (Dense Motion) для создания тепловых карт и кодирования информации о движении;
– сеть передачи движения (Motion Transfer Generator), которая синтезирует выходные кадры на основе тепловых карт движения и входного изображения.

MonkeyNet является первой моделью для анимации изображений неизвестных объектов. Она извлекает целевые ключевые точки на изображении с помощью самообучающегося детектора ключевых точек и генерирует плотную тепловую карту из разреженных ключевых точек. После этого синтезируется входной кадр, который использует тепловую карту движения и информацию о внешнем виде, извлеченную из входного изображения. Однако для MonkeyNet сложно моделировать преобразование внешнего вида объектов вблизи ключевых точек, что приводит к плохому качеству генерации, когда масштаб изменения объекта довольно велик.

Для поддержки более сложного движения предлагается применение FOMM, чтобы использовать неконтролируемые ключевые точки обучения и локальное аффинное преобразование для имитации сложного движения.

FOMM – это нейросетевая модель, предназначенная для анимации статичных изображений, которая основана на декомпозиции движения и представления. В отличие от методов, требующих парных данных (например, видео с одинаковым объектом в разных позах), FOMM работает в unsupervised-режиме, используя лишь исходное изображение и видео с референсными движениями.

Для обучения модели используется большая коллекция видеопоследовательностей, содержащих объекты одной и той же категории объектов. Модель реконструирует видео путем объединения одного кадра и изученного скрытого представления движения в видео. В процессе наблюдения пары кадров (исходного и движущегося), каждый из которых извлечен из одного и того же видео, модель учится кодировать движение в виде комбинации смещений ключевых точек, специфичных для движения, и локальных аффинных преобразований. Во время тестирования модель применяется к парам, состоящим из исходного изображения и каждого кадра движущегося видео, на основе чего выполняется анимация изображения исходного объекта.

Архитектура First Order Motion Model. Структура FOMM состоит из двух основных модулей: модуля оценки движения и модуля генерации изображения. Обзор и архитектура данного подхода представлены на рисунке 1 [1].

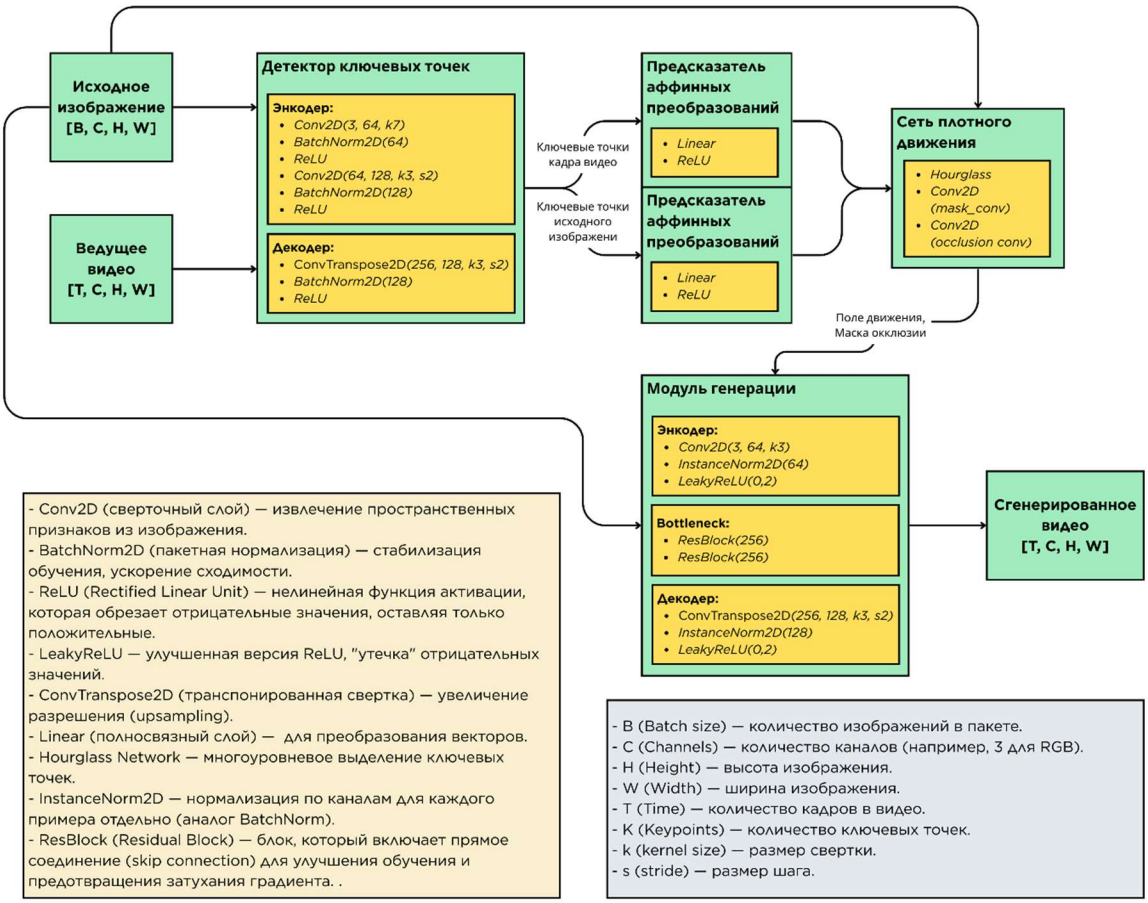

Рисунок 1 – Схема архитектуры First Order Motion Model

На первом этапе аппроксимируются оба преобразования из наборов разреженных траекторий, полученных с использованием ключевых точек, изученных самоконтролируемым способом. Моделируется движение в окрестности каждой ключевой точки с помощью локальных аффинных

преобразований. По сравнению с использованием только смещений ключевых точек, локальные аффинные преобразования позволяют моделировать большее семейство преобразований.

На втором этапе плотная сеть движения объединяет локальные аппроксимации для получения результирующего плотного поля движения. В дополнение к этому сеть выводит маску окклюзии, которая указывает, какие части ведущего изображения могут быть реконструированы путем деформации исходного изображения, а какие части должны быть закрашены (выведены из контекста). Наконец, модуль генерации визуализирует изображение исходного объекта, движущегося так, как показано в видео для анимации. Здесь используется сеть генератора, которая деформирует исходное изображение в соответствии с плотным движением и восстанавливает те области, которые были скрыты или отсутствовали на исходном изображении, например, из-за поворота головы или частичной окклюзии.

Описание набора данных. Модель обучалась и тестировалась на наборе данных VoxCeleb [4]. VoxCeleb – набор данных, который представляет собой видеопоследовательности с лицами людей, извлеченных из 22496 видео YouTube.

Для предварительной обработки в первом кадре извлекается начальная ограничивающая рамка вокруг лица человека. Затем эта рамка отслеживается, пока лицо не окажется слишком далеко от исходного положения. Далее видеокадры обрезаются, используя наименьшее кадрирование, содержащее все ограничивающие рамки. Процесс повторяется до конца последовательности. Последовательности, которые имеют разрешение ниже 256 × 256, отфильтровываются, а разрешение оставшихся видео уменьшается до 256 × 256 с сохранением соотношения сторон.

В итоге было получено 12331 обучающее видео и 444 тестовых видео, длина которых варьируется от 64 до 1024 кадров.

Результаты работы. Для оценки реконструкции видео выбраны следующие метрики из Monkey-Net:

1. L1 – сообщается среднее расстояние L1 между сгенерированным и реальным видео.
2. AKD (Average Keypoint Distance) – среднее расстояние до ключевых точек. Ключевые точки вычисляются независимо для каждого кадра. AKD получается путем вычисления среднего расстояния между обнаруженными ключевыми точками реального видео и сгенерированного видео.
3. AED (Average Euclidean Distance) – среднее евклидово расстояние между истинным и сгенерированным представлением кадра. Используется встраивание признаков, аналогичное в Monkey-Net.

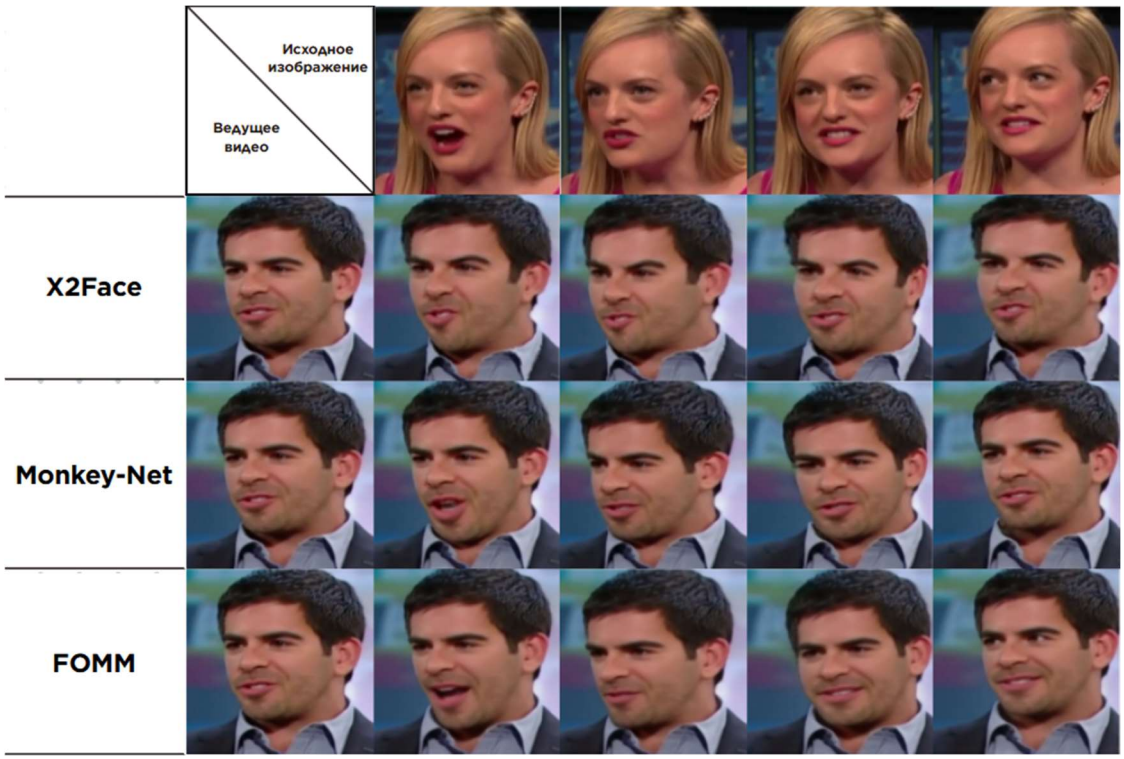

Рисунок 2 – Примеры анимации изображений на основе X2Face, Monkey-Net и FOMM

На рисунке 2 представлено несколько примеров результатов генерации анимации изображений для визуального сравнения. Визуальное сравнение демонстрирует высокую степень соответствия мимики и позы между оригинальными и синтезированными кадрами, особенно в случае сложных

движений лица и головы. В отличие от более ранних моделей, таких как X2Face и Monkey-Net, FOMM обеспечивает более стабильную и детализированную передачу движений.

В таблице 1 приведены метрики сравнения алгоритмов анимации изображений.

Таблица 1 – Метрики сравнения алгоритмов анимации изображений

Набор данных	Метрика	X2Face	Monkey-Net	FOMM
VoxCeleb	L1	0,078	0,049	0,043
	AKD	7,687	1,878	1,294
	AED	0,405	0,199	0,140

L1 измеряется в диапазоне от 0 до 1, где 0 – идеальное совпадение по пикселям. Значение 0,043 у модели FOMM означает, что среднее отклонение интенсивности каждого пикселя между сгенерированным и реальным изображением составляет всего 4,3% от возможного максимума, что говорит о высокой точности генерации.

Значения AKD выражаются в пикселях. У FOMM значение равно 1,294 пикселя, то есть в среднем ключевые точки сдвинуты на 1,3 пикселя. X2Face имеет ошибку в 7,7 пикселя – это уже значительное смещение, заметное невооруженным глазом.

AED – мера, показывающая семантическую схожесть изображений. Чем меньше ее значение, тем более «узнаваемым» и реалистичным является результат. У FOMM получено значение 0,140, что означает высокую близость к реальному изображению в пространстве признаков. Для сравнения, значение у X2Face почти в 3 раза хуже.

Численные значения метрик ясно показывают, что модель FOMM превосходит другие подходы по всем показателям. Это делает ее предпочтительным выбором для решения задач анимации изображений на основе одного кадра.

Заключение. Методы анимации изображений, основанные на нейронных сетях, представляют собой мощный инструмент для создания высококачественных анимаций. Особенно это касается применения предварительно обученных моделей, зависящих от большого объема размеченных данных.

В то же время исследования в области неконтролируемой анимации изображений без использования маркированных данных достигли значительного прогресса. Хотя архитектура First Order Motion Model демонстрирует превосходные результаты, подобные методы все еще имеют свои ограничения. Существует неточность прогнозирования оптического потока, что приводит к некачественным сгенерированным кадрам. Кроме того, существующие алгоритмы могут неправильно определять ключевые точки, располагая их на фоне вместо движущихся объектов. Это приводит к тому, что деформация смещения между ключевыми точками не всегда правильно описывает движения жестких объектов, вызывая такие эффекты, как ложные тени.

Для дальнейшего совершенствования технологий анимации требуется повышение точности прогнозирования и минимизация искажений в сгенерированных кадрах. Необходимо проведение оптимизации архитектуры и алгоритмов, внедрение новых подходов к обучению. Это позволит преодолеть существующие ограничения и повысить доступность и эффективность анимации изображений в прикладных задачах.

Особое внимание следует уделить учету межкадровой корреляции. Для решения данной задачи требуется использовать такие нейронные сети, как LSTM (Long Short-Term Memory), способные сохранять результаты генерации предыдущих кадров и использовать их при формировании текущего. Учитывание временной взаимосвязи между кадрами способствует повышению согласованности и качества реконструкции.

Кроме того, необходимо задействовать несколько исходных изображений, полученных с разных ракурсов, и реализовать автоматическую настройку степени их вклада в итоговую генерацию посредством нейронных сетей. Такой подход обеспечит более точную и реалистичную генерацию анимации.

Список использованных источников:

1. First Order Motion Model for Image Animation [Электронный ресурс] – Электронные данные. – Режим доступа: https://papers.nips.cc/paper_files/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf – Дата доступа: 20.03.2025

2. X2Face: A network for controlling face generation using images, audio, and pose codes [Электронный ресурс] – Электронные данные. – Режим доступа: https://arxiv.org/pdf/1807.10550 – Дата доступа: 22.03.2025

3. Animating Arbitrary Objects via Deep Motion Transfer [Электронный ресурс] – Электронные данные. – Режим доступа: https://arxiv.org/pdf/1812.08861 – Дата доступа: 24.03.2025

4. VoxCeleb [Электронный ресурс] – Электронные данные. – Режим доступа: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/– Дата доступа: 30.03.2025

UDC 004.032.26, 004.932.4

IMAGE ANIMATION BASED ON NEURAL NETWORKS

Klimovich A.N.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Kupryianava D.V. – Master of Technical Sciences

Annotation. Modern neural network-based image animation methods have been analyzed. The First Order Motion Model (FOMM) was selected as the primary architecture. A review and comparative analysis of other approaches, such as X2Face and Monkey-Net, were conducted. FOMM was optimized for real-time performance. The model was tested, and performance metrics were collected.

Keywords: Generative Adversarial Network, Image Animation, First Order Motion Model, Motion Transfer, Video Generation, X2Face, Monkey-Net, Keypoints, Image Processing, Comparison of Animation Methods.