Распознавание эмоций на групповых фотографиях лиц на основе глубоких сверточных нейронных сетей

А. В. Тарасов
Национальный исследовательский университет Высшая школа экономики
Нижний Новгород, Россия
alxndrtarasov@gmail.com

А. В. Савченко
Национальный исследовательский университет Высшая школа экономики
Нижний Новгород, Россия
avsavchenko@hse.ru

***Аннотация*. В работе рассмотрена задача распознавания эмоций по фотографиям, содержащим большое количество людей. Предложен новый алгоритм, в котором на первом этапе для детектирования лиц небольшого размера применяется специальная сверточная нейронная сеть. Далее для каждого выделенного лица извлекаются векторы признаков с использованием глубокой нейронной сети, предварительно обученной для задачи распознавания эмоций по фотографии. Итоговое решение принимается с помощью классификации усредненного вектора признаков всех выделенных лиц. Проведено экспериментальное исследование традиционных классификаторов для набора изображений из конкурса EmotiW 2017. Показано, что наибольшая точность распознавания (75,5%) достигается при использовании ансамбля машины опорных векторов и случайного леса.**

Ключевые слова: распознавание образов; распознавание эмоций группового уровня; сверточные нейронные сети; распознавание лиц; вектор признаков изображения

I. Введение

Задача распознавания эмоций по изображениям может использоваться во многих практических приложениях, таких как образование [1], видеонаблюдение, взаимодействие человека с компьютером [2, 3] и т.д. Однако, в настоящее время точность известных решений этой задачи еще не достигло такого качества, которое наблюдается, например, в задачах распознавания лиц по фотографии. Связано это с тем, что даже человеку зачастую сложно определить запечатленную эмоцию. В приложении к распознаванию эмоций на фотографиях групп лиц [4] проблема усложняется тем, что во многих случаях на практике различные лица выражают разные эмоции. Такие конкурсы, как EmotiW (Emotion Recognition in the Wild), помогают привлечь внимание исследователей к подобным задачам. Так, в 2017 году рассматривалась задача распознавания эмоций групповых фотографий, сделанных в реальных условиях, разделенным по эмоциональному окрасу на позитивные, нейтральные и негативные [4, 5]. Участниками конкурса были представлены подходы, связанные с применением передачи знаний (transfer learning) [6], сверточных нейронных сетей (СНС) [7], анализа глобального контекста фотографий [8] и т.д.

В настоящей работе разработан новый алгоритм для решения задачи распознавания групповых эмоций, в котором для обнаружения лиц на фотографиях используется недавно появившийся метод Tiny Face Detector [9] далее для извлечения признаков отдельных лиц применяется СНС, предварительно обученная для классификации эмоций по фотографиям лица. Для распознавания групповых эмоций наборы признаки каждого выделенного лица агрегируются в один вектор признаков, и далее применяются стандартные методы классификации. Наконец, классификаторы объединяются в ансамбль с целью найти сочетание, приводящее к наивысшей точности распознавания.

Дальнейшая часть статьи организована следующим образов. В разделе 2 описан предлагаемый алгоритм. Результаты экспериментального исследования разных методов создания набора векторов приведены в разделе. В заключительном разделе сделаны выводы по итогам проведенной работы и указаны направления дальнейших исследований.

II. Предложенный подход

Рассматриваемая в настоящей работе задача состоит в классификации изображения группы лиц в одну из трех эмоциональных категорий («positive», «neutral», «negative»). Для проведения исследования использовался набор данных EmotiW 2017, взятый из базы данных Group Affect 2.0 [4]. Обучающая выборка состоит из 3630 фотографий, а тестовый набор для проверки точности распознавания – из 2065 фотографий. Можно заметить, что изображения в тестовом множестве распределены между эмоциями неравномерно. Часть позитивных изображений составляет 37% набора, а негативных – только 27%.

Статья подготовлена в результате проведения исследования (No 17-05-0007) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2017 г. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100". Работа А.В. Савченко выполнена при поддержке гранта президента РФ для молодых ученых – докторов наук No МД-306.2017.9.

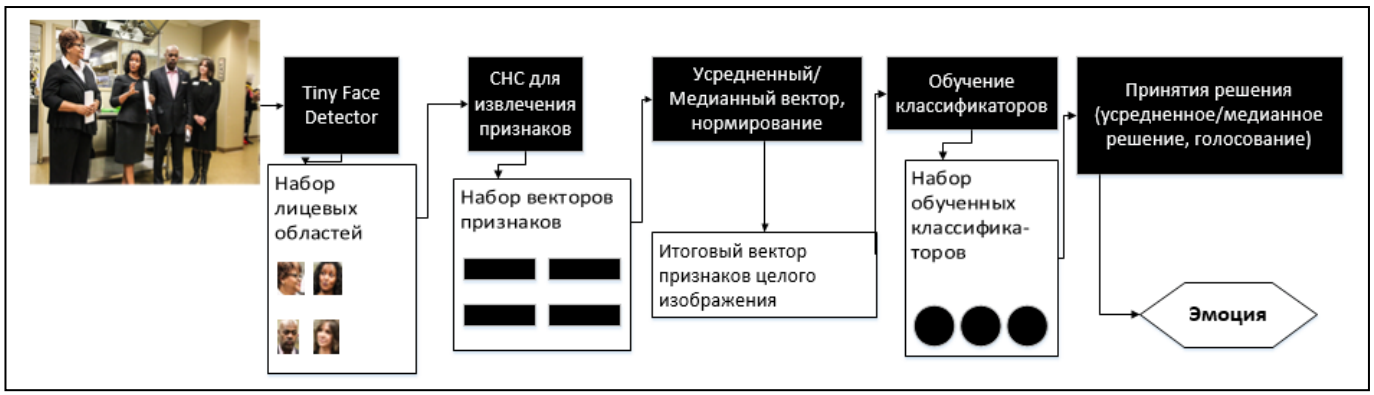

Рис. 1. Окончательная схема описываемого подхода

Групповые фотографии могут быть сделаны во время позитивных событий, таких как свадьбы, вечеринки, нейтральных (например, политические встречи) и событий негативного характера, таких как протесты. Особый интерес представляет то, что изображения могут включать лица разных размеров, при этом присутствуют лица всех возрастных групп.

Предложенный подход представлен на рис. 1. Далее рассмотрим основные шаги алгоритма подробнее.

A. Обнаружение лиц

Ожидается, что большая часть информации об эмоциональной окраске фотографии заключается в эмоциях присутствующих лиц. Поэтому для распознавания эмоций в первую очередь необходимо выделить лица на изображении. Традиционный алгоритм решения этой задачи состоит в линейной классификации пирамид гистограмм ориентированных градиентов [10] из библиотеки DLIB. Применив его к фотографиям из наборов для обучения и проверки, удалось выделить только 19291 лицевых изображений. Для более надежного детектирования лиц применялся недавно предложенный метод Tiny Face Detector [9], который реализует Resnet-101 CHC [11], специально обученную на наборе данных WIDER так, чтобы выделять лица очень малого размера. Этот детектор смог обнаружить уже 54053 лиц на тех же наборах фотографий. При использовании обоих методов распознавания лиц процент правильно выделенных лицевых областей был одинаковым, однако Tiny Face Detector обнаружил хотя бы одно лицо на 98% фотографий, а традиционный подход – лишь на 92%. Именно поэтому Tiny Face Detector был выбран как более предпочтительный метод.

B. Извлечение признаков лиц

Извлечение характерных признаков каждого выделенного лица осуществлялось с помощью технологий адаптации предметной области [11] с помощью глубокой CHC. Была выбрана традиционная архитектура нейронной сети, состоящая из 9 сверточных уровней, ReLUs, Batch Normalization и Global Average Pooling. CHC была предварительно обучена для распознавания эмоций на фотографии лица размером 64x64 с использованием внешнего набора данных FER-2013, в котором 35887 лицевых изображений разделены на 7 групп («angry», «disgust», «fear», «happy», «sad», «surprise», «neutral»). Модель достигла точности распознавания эмоций в 66% на наборе FER-2013.

Размер каждого выделенного лица был изменен до 64x64, после чего изображения подавались на вход CHC. Для извлечения признаков использовались два варианта [11, 12]: 1) выход последнего слоя softmax CHC, который состоит из 7 оценок апостериорных вероятностей принадлежности лица одному из 7 классов; 2) 112 выходов предпоследнего слоя CHC (далее «embedding»).

C. Классификация эмоций группы лиц

После извлечения признаков каждого лица необходимо их агрегировать итоговый вектор признаков фотографии. Исследовались такие методы, как усреднение вектора признаков, вычисление вектора медиан [6], а также комбинации вышеперечисленных методов с L2-нормировкой.

Итоговые признаки подавались на вход различных методов классификации для принятия окончательного решения об эмоциональной окраске групповой фотографии. В настоящей работе использовались классификаторы из библиотеки scikit-learn: AdaBoost classifier, Bagging, Extra-Trees, Gradient Boosting, Random Forest, C-Support Vector Machine (SVM) и Linear SVM.

Параметры классификаторов были дополнительно настроены с помощью многопараметрической оптимизации grid search. Наконец, для повышения точности итогового решения классификаторы были объединены в несколько различных ансамблей [6, 11]. Исследовались следующие подходы к реализации ансамбля моделей: 1) усредненное решение отдельных классификаторов; 2) медиана решений; 3) взвешенное голосование, в котором выходы индивидуальных классификаторов взвешиваются на основе их точности.

ТАБЛИЦА I ОЦЕНКА ТОЧНОСТИ КЛАССИФИКАТОРОВ В ЗАВИСИМОСТИ ОТ СПОСОБОВ ИЗВЛЕЧЕНИЯ ПРИЗНАКОВ

Классификатор	Softmax		Embeddings	
	Среднее арифметическое	*Медиана*	*Среднее арифметическое*	*Медиана*
Extra Trees	0.604	0.561	0.61	0.606
Bagging	0.606	0.591	0.63	0.605
RandomForest	0.6	0.596	0.636	0.58
RBF SVM	0.626	0.589	0.646	0.608
Ada Boost	0.62	0.613	0.646	0.626
Linear SVM	0.64	0.614	0.65	0.659
GradientBoosting	0.654	0.622	0.695	0.655

III. ДОСТИГНУТЫЕ РЕЗУЛЬТАТЫ

В рамках экспериментального исследования сопоставлялись два варианта извлечения признаков с различными методами их агрегации. В таблице 1 представлены оценки точности различных методов классификации. Как видно, использование выхода предпоследнего слоя CHC («embeddings») в комбинации с усредненным вектором признаков приводило к наиболее высокой точности классификации все методы, за исключением Linear SVM. Поэтому именно такой способ извлечения признаков групповой фотографии использовался на последующих этапах экспериментов.

К сожалению, точность всех тестируемых классификаторов оказалась не выше 66%. Для снижения вероятности ошибки последовательно применялось несколько алгоритмов предварительной обработки, такие как L2-нормализация векторов признаков, а также извлечение главных компонент. Кроме того, использовалась многопараметрическая оптимизация grid search. Результаты, достигнутые на этом этапе, приведены в табл. 2.

ТАБЛИЦА II ОЦЕНКА ТОЧНОСТИ ПОСЛЕ ПРИМЕНЕНИЯ НОРМАЛИЗАЦИИ, МНОГОПАРАМЕТРИЧЕСКОЙ ОПТИМИЗАЦИИ И МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

Классификатор	Точность
RBF SVM	0.693
GradientBoosting	0.701
Extra Trees	0.702
Linear SVM	0.703
Bagging	0.707
RandomForest	0.709
Ada Boost	0.709

ТАБЛИЦА III ОЦЕНКА ТОЧНОСТИ АНСАМБЛЕЙ КЛАССИФИКАТОРОВ

Наборы классификаторов в ансамбле	Точность		
	Усредненное решение	Медианное решение	Решение голосованием
AdaBoostClassifier RandomForestClassifier	0.702	0.702	0.713
AdaBoostClassifier RBF SVM LinearSVM BaggingClassifier RandomForestClassifier GradientBoostingClassifier ExtraTreesClassifier	0.712	0.724	0.723

ПРОДОЛЖЕНИЕ ТАБЛ. 3

Наборы классификаторов в ансамбле	Точность		
	Усредненное решение	Медианное решение	Решение голосованием
LinearSVC RandomForestClassifier	0.686	0.686	0.724
LinearSVM RandomForestClassifier BaggingClassifier RBF SVM	0.732	0.731	0.746
LinearSVM AdaBoostClassifier RBF SVM	0.718	0.743	0.746
LinearSVM RandomForestClassifier ExtraTreesClassifier RBF SVM	0.727	0.732	0.751
RandomForestClassifier RBF SVM	0.696	0.696	0.753
LinearSVM RandomForestClassifier RBF SVM	0.723	0.753	0.755

Как показали эксперименты, снижение размерности до 80 главных компонент в среднем позволяло повысить точность на 2%. После операций, описанных выше, точность тестируемых методов превысила 69%. Наибольшую точность 70.9% показали классификаторы RandomForest и AdaBoost.

На заключительном этапе были протестированы различные комбинации классификаторов, объединенных в ансамбль. Результаты приведены в табл. 3.

Как видно из этой таблицы, взвешенное голосование позволяло повысить точность по сравнению с усредненным решением и медианным решением. Наиболее высокую точность 75.5% достиг ансамбль, который включал в себя Random Forest, C-SVM (RBF) и Linear SVM. Предложенный подход оказался на 23%, 20% и 8% известных способов решения задачи по сравнению с базовым методом классификации специально подобранных признаков изображений [4] применении CHC на тепловых картах [13] и комбинации глубоких CHC с байесовскими классификаторами [14], соответственно.

IV. ЗАКЛЮЧЕНИЕ

В работе описывается новый алгоритм распознавания эмоций на фотографиях групп людей, который включает в себя обнаружение лиц с помощью Tiny Face Detector [9], извлечение характерных признаков лиц с помощью глубокой CHC, предварительно обученной для распознавания эмоций на изображениях лиц с низким разрешением, и классификацию усредненных признаков отдельных лиц. Окончательное решение принимается ансамблем из нескольких классификаторов путем голосования, когда решения отдельных классификаторов взвешиваются на основе их точности. Предлагаемый подход достигает 75,5% точности распознавания на валидационном множестве, что примерно на 9% выше по сравнению с точностью, продемонстрированной в нашем предыдущем исследовании [6] метода на основе Tiny Face

Detector. Причиной повышения точности стало то, что в последнем случае признаки извлекались с помощью СНС VGGFace [15], для обучения которой использовался набор фотографий лиц изображения высокого качества. В результате эмоции на лицах небольшого размера в большинстве случае были распознаны неверно. В настоящей работе указанная проблема была преодолена с использованием для обучения СНС набора изображений лиц FER-2013 с низким разрешением.

Предложенный подход может быть полезен для распознавания эмоций в системах видеоаналитики и таргетированной рекламы, предоставляющих изображения, которые содержат множество лиц небольшого размера. В дальнейших исследованиях планируется исследовать способы повышения точности распознавания эмоций с помощью комбинирования разработанного алгоритма с известными методами [8], основанными на расширении обучающего набора изображениями и использовании СНС для определения не только эмоций лиц, но и глобального контекста изображений [6]. Кроме того, планируется повысить среднее время принятия решений [2, 16] с помощью, например, сжатия СНС [17], для повышения практической значимости и обеспечения возможности реализации алгоритма на малопроизводительных мобильных устройствах в автономном режиме.

Список литературы

[1] D'Mello S., Picard R., Graesser A. Toward An Affect-Sensitive AutoTutor. *IEEE Intelligent Systems*, 2007, no. 4, pp.53-61. DOI: 10.1109/MIS.2007.79

[2] A.V. Savchenko. 2016. *Search Techniques in Intelligent Classification Systems*. Springer, ISBN: 978-3-319-30515-8

[3] Pantic M., Rothkrantz L. J. M. Toward an Affect-Sensitive Multimodal Human–Computer Interaction. *Proceedings of the IEEE*, 2003, vol. 91, vo. 9, pp. 1370-1390. DOI: 10.1109/JPROC.2003.817122

[4] Dhall A., Goecke R., Ghosh S., Joshi J., Hoey J.,Gedeon T. From individual to group-level emotion recognition: EmotiW 5.0. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Glasgow, United Kingdom, 2017, pp. 524-528. DOI: 10.1145/3136755.3143004

[5] Sun B., Wei Q., Li L., Xu Q., He J., Yu L. LSTM for dynamic emotion and group emotion recognition in the wild. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. Japan, 2016. pp. 451-457. DOI: 10.1145/2993148.2997640

[6] Rassadin A., Gruzdev A., Savchenko A. Group-level Emotion Recognition using Transfer Learning from Face Identification. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Glasgow, United Kingdom, 2017, pp. 544-548. DOI: 10.1145/3136755.3143007

[7] Tan L., Zhang K., Wang K., Zeng X., Peng X., Qiao Y. Group Emotion Recognition with Individual Facial Emotion Convolutional Neural Networks and Global Image Based Convolutional Neural Networks. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Glasgow, United Kingdom, 2017, pp. 549-552. DOI: 10.1145/3136755.3143008

[8] Abbas A., Chalup S. K. Group emotion recognition in the wild by combining deep neural networks for facial expression classification and scene-context analysis. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Glasgow, United Kingdom, 2017, pp. 561-568. DOI: 10.1145/3136755.3143010

[9] Hu P., Ramanan D. Finding Tiny Faces. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, USA, 2017, pp. 1522-1530. DOI: 10.1109/CVPR.2017.166

[10] Dalal N. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition*, San Diego, USA, 2005, pp. 886-893. DOI: 10.1109/CVPR.2005.177

[11] Goodfellow I., Bengio Y., Courville A. *Deep Learning*. MIT Press. 2016. 777 p.

[12] Savchenko A.V., Belova N.S., Savchenko L.V. Fuzzy Analysis and Deep Convolution Neural Networks in Still-to-video Recognition, *Optical Memory and Neural Networks (Information Optics)*, 2018, vol. 27, no. 1, pp. 23-31

[13] Shamsi S., Rawat B. P. S., Wadhwa M. Group Affect Prediction Using Multimodal Distributions. Available at: https://arxiv.org/pdf/1710.01216.pdf (accessed 20 March 2018)

[14] Surace L., Patacchiola M., Sönmez E. B., Spataro W., Cangelosi A. Emotion Recognition in the Wild using Deep Neural Networks and Bayesian Classifiers. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Glasgow, United Kingdom, 2017, pp. 593-597. DOI: 10.1145/3136755.3143015

[15] O.M. Parkhi, A. Vedaldi, A. Zisserman. 2015. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, pp. 1-12

[16] A.V. Savchenko Maximum-likelihood approximate nearest neighbor method in real-time image recognition. *Pattern Recognition*, 2017, vol. 61, pp. 459-469

[17] Rassadin A.G., Savchenko A.V. Compressing deep convolutional neural networks in visual emotion recognition. *Proceedings of the International Conference on Information Technology and Nanotechnology (ITNT). Session Image Processing, Geoinformation Technology and Information Security Image Processing (IPGTIS)*, Samara, Russia, 2017, pp. 207-213.