Exploring Computer Vision for Film Analysis: A Case Study for Five Canonical Movies

1. Introduction

Quantitative methods have a long tradition in film analysis going back to the predigital era [26, 27]. Nowadays, multiple projects explore movies via computational methods to investigate colors [4, 5, 12, 20, 22, 23], shot lengths [3, 10] or annotation possibilities [14, 19]. Recent research has also led to the definition of the term Distant Viewing [2] to describe large-scale digital movie analysis. A lot of the current research is focused on the analysis of text via scripts or subtitles [6, 15, 16, 18]. However, recent developments in computer vision have led to novel methods for the image channel of movies and are already applied in computer science to develop recommender systems [9, 28] but also in Digital Humanities to analyze movies [17, 24, 31]. We argue that these methods are beneficial for digital film studies and give new perspectives.

We present a small-scale exploratory study for the methods: Object detection, emotion recognition, gender and age prediction. We apply state-of-the-art models on a subset of frames of five different movies of varied decades and genres. We apply the exploratory research approach defined by Wulff (1998) [30] for traditional film analysis in this study for computational approaches. Our goals are to inspect the benefits and problems of the methods, explore if the methods uncover specific characteristics of the movies and what research questions seem promising to follow in further large-scale studies.

2. Material

We limited the analysis on 5 movies. Table 1 presents the movies and metadata. For all movies except *Avengers*, we use a digitally restored version. All movies have a 720x576 resolution, 25 frames per second and 32 bits per sample. We focus on canonical work and Hollywood productions.

Title	Release Date	Running time in seconds	Important Attributes	Genre
Metropolis	1927	8636	Silent film, black and white	Science-Fiction/Drama
Wizard of Oz	1939	5856	Color	Fantasy/Musical
Some Like It Hot	1959	6990	Black and white	Comedy
Breakfast at Tiffany's	1961	6406	Color	Romantic Comedy
Marvels The Avengers	2012	8224	Color, including CGI effects	Action/Fantasy

Table 1. Movies and metadata

3. Methods

All of our analysis was performed in Python 3. We extracted the frames of every movie, since all of the applied methods are image-based. However, we take one frame per second of a movie and regard this as the sample of a movie. We decided to employ this approach because using all frames makes the data processing very performance/resource-intensive and we argue that one frame per second offers sufficient information for our first explorations.

To perform the object detection, we use *Detectron2* [29] which offers state-of-the-art object detection models by Facebook AI Research[1]. We use a pretrained masked RCCN-model trained on the well-known *COCO*-Dataset [21], which can predict 80 object classes including vehicles, animals and sports objects. Applying this prediction model on an image, we receive the number of predicted objects, the locations and the prediction confidence (0-100%). As threshold for the detection, we select 50% which is usually very low but fits our exploratory approach.

For the emotion prediction we use the Python module *FER*[2] [13]. The module first performs face detection via a MTCNN Face Detector[3] [32] and then predicts the emotion via a convolutional neural network (CNN) trained on over 35,000 images. The model predicts the seven classes *anger, disgust, fear, happiness, sadness, surprise* and *neutral* on a scale from 0 to 1. All values sum up to 1 for one face.

We perform gender and age prediction via the module *py-agender*[4] which is also a CNN trained on the IMDB-Wiki dataset [25] consisting of over 500,000 faces. The model achieves a mean average error of 4.08 on standardized datasets [1]. For the gender prediction the model produces a value between 0 and 1, with values below 0.5 being male and above being female faces.

4. Results

4.1. Object Detection

We summarize the results of the object detection by looking at the 10 most frequent objects overall and per movie. Table 2 and 3 show the objects starting with the most frequent per unit. *Freq* is the absolute number of detected instances while *%* is the percentage of frames at least one of the specific objects was detected.

Metropolis			**Wizard of Oz**			**Some Like It Hot**		
Object	**freq**	**%**	**Object**	**freq**	**%**	**Object**	**freq**	**%**
person	26,844	73.9	**person**	17587	87.0	**person**	23,027	95.0
tie	1,704	14.8	**dog**	1049	17.0	**tie**	3,280	27.6
book	1,616	2.6	**handbag**	842	13.1	**chair**	1,138	12.1
chair	382	3.5	**chair**	722	10.9	**wine glass**	708	6.0
clock	366	3.0	**potted plant**	566	7.2	**handbag**	417	5.6
umbrella	265	1.2	**tie**	541	7.5	**cup**	405	4.9
horse	124	1.0	**vase**	354	5.1	**bottle**	351	3.7
dog	113	1.3	**horse**	344	4.8	**cell phone**	348	4.4
dining table	108	1.2	**cat**	306	5.0	**suitcase**	330	3.6
cup	103	1.0	**bottle**	226	2.9	**vase**	286	3.2

Table 2. Detected objects per movie and overall (part 1)

[1] More Information: https://github.com/facebookresearch/detectron2

[2] More Information: https://pypi.org/project/fer/

[3] More Information: https://github.com/ipazc/mtcnn

[4] More Information: https://github.com/yu4u/age-gender-estimation

Breakfast At Tiffany's			Avengers			Overall		
Object	**freq**	**%**	**Object**	**freq**	**%**	**Object**	**freq**	**%**
person	17,357	96.2	**person**	15,556	77.7	**person**	100,371	84.9
tie	4,057	44.4	**chair**	1,476	11.9	**tie**	101,84	19.3
book	2,248	5.4	**car**	917	6.1	**chair**	4,985	10.1
chair	1,217	13.6	**tie**	602	5.8	**book**	4,237	2.3
wine glass	731	7.6	**tv**	476	4.6	**handbag**	1,876	4.7
car	720	4.6	**bottle**	286	1.9	**car**	1,809	2.5
bottle	708	6.9	**airplane**	208	1.6	**wine glass**	1,662	3.0
cup	536	6.9	**cell phone**	205	2.3	**bottle**	1,653	3.0
dining table	413	5.5	**book**	189	1.3	**dog**	1,460	3.8
handbag	367	4.7	**backpack**	184	2.0	**cup**	1,320	3.1

Table 3. Detected objects per movie and overall (part 2)

Persons are the most frequently detected "objects" (figure 1). Other frequent objects are mostly furniture (book, chair), clothes (tie, handbag) and drinking objects (cup, wine glass).

Figure 1. Frame with the most detected person (Metropolis)

Comparing the movies, we identified that movies below 90% of frames with persons are indeed the more action-oriented movies (Avengers, Metropolis) or include fantasy/animal-like characters (Wizard of Oz). Many modern objects (e.g cell phones and airplanes) are more frequent in the contemporary movie *Avengers* (figure 2). One outlier we identified is the clock-object in Metropolis, which is not a frequent object in the other movies but represents a well-studied reoccurring motif of this specific movie (figure 3; cf. [8]).

Figure 2. Detected airplanes in Avengers

Figure 3. Clocks as a reoccurring motif in Metropolis

While we did not perform a systematic evaluation, we identified a lot of mistakes in the prediction e.g. guns were predicted as handbags or the character “Cowardly Lion” in Wizard of Oz was oftentimes predicted as dog (figure 4).

Figure 4. The „Cowardly Lion“ in Wizard of Oz detected as „Dog“

Nevertheless, we see potential in the method of object detection to explore specifics of the mise-en-scène as well as motif-like reoccurring objects in movies [31]. Furthermore, as object classes of the COCO dataset are not necessarily fitting for movies, we recommend exploring the possibilities of post-training via Detectron to analyze objects that are not part of the pretrained models.

4.2. Emotion Recognition

For the emotion recognition we decided create an average for a frame if multiple faces are detected. If no face is detected, we mark the frame with missing values. Table 4 summarizes the results. Maximums and minimums are marked in bold.

		Metropolis	Wizard of Oz	Some Like it Hot	Breakfast At Tiffany's	Avengers	Overall
Angry	**M**	0.22	**0.23**	0.17	**0.13**	0.21	0.19
	Max	0.98	0.97	0.95	0.85	0.93	0.98
	Sd	0.17	0.17	0.13	0.12	0.15	0.16
Disgust	**M**	0.00	0.00	0.00	0.00	0.00	0.00
	Max	0.12	0.18	0.18	0.19	0.22	0.22
	Sd	0.01	0.01	0.01	0.01	0.01	0.01
Fear	**M**	**0.16**	0.11	0.11	**0.08**	0.12	0.11
	Max	0.88	0.94	0.73	0.8	0.65	0.94
	Sd	0.13	0.09	0.09	0.08	0.09	0.1
Happy	**M**	0.1	**0.13**	**0.13**	0.09	**0.07**	0.11
	Max	1.0	1.0	1.0	1.0	1.0	1.0
	Sd	0.16	0.17	0.18	0.18	0.14	0.17
Neutral	**M**	0.22	**0.16**	0.23	**0.37**	0.23	0.24
	Max	0.96	0.96	0.9	0.99	0.93	0.99
	Sd	0.19	0.15	0.18	0.25	0.2	0.21
Sad	**M**	0.29	**0.32**	**0.27**	0.28	0.28	0.29
	Max	0.93	0.95	0.88	0.91	0.9	0.95
	Sd	0.18	0.19	0.17	0.18	0.17	0.18
Surprise	**M**	0.04	0.04	**0.09**	0.04	0.08	0.06
	Max	0.87	0.93	0.95	0.83	0.95	0.95
	Sd	0.1	0.09	0.14	0.08	0.13	0.11

Table 4. Emotion values per movie and overall (M=mean, Max=maximum, Sd=standard deviation)

Overall, highest averages for emotions are the neutral (*M*=0.24) and the sad class (*M*=0.29). Surprise (*M*=0.11) and disgust (*M*=0.00) are rather rare among the movies. The two comedies in the movie corpus (Wizard of Oz, Some Like it Hot) do indeed have the highest happy averages (*M*=0.13) (figure 5).

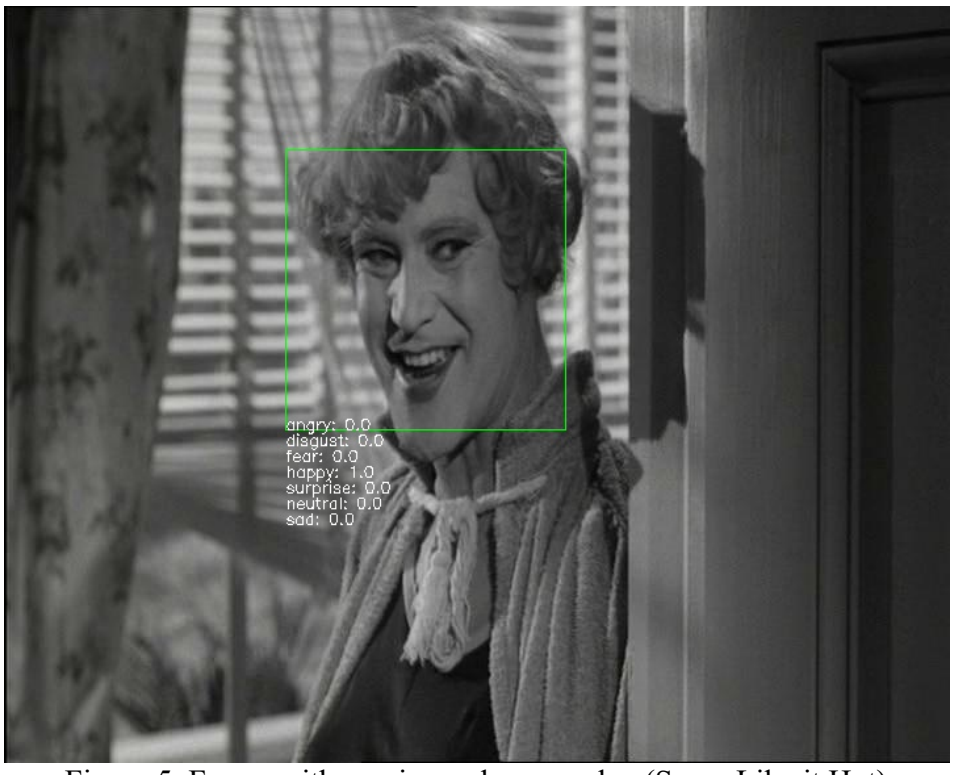

Figure 5. Frame with maximum happy value (Some Like it Hot)

However, the results are rather inconsistent since Wizard of Oz has also the highest sad and angry averages and therefore is the movie with generally the strongest emotion expressions. Breakfast at Tiffany's on the contrast is the most neutral movie (*M*=0.37; figure 6).

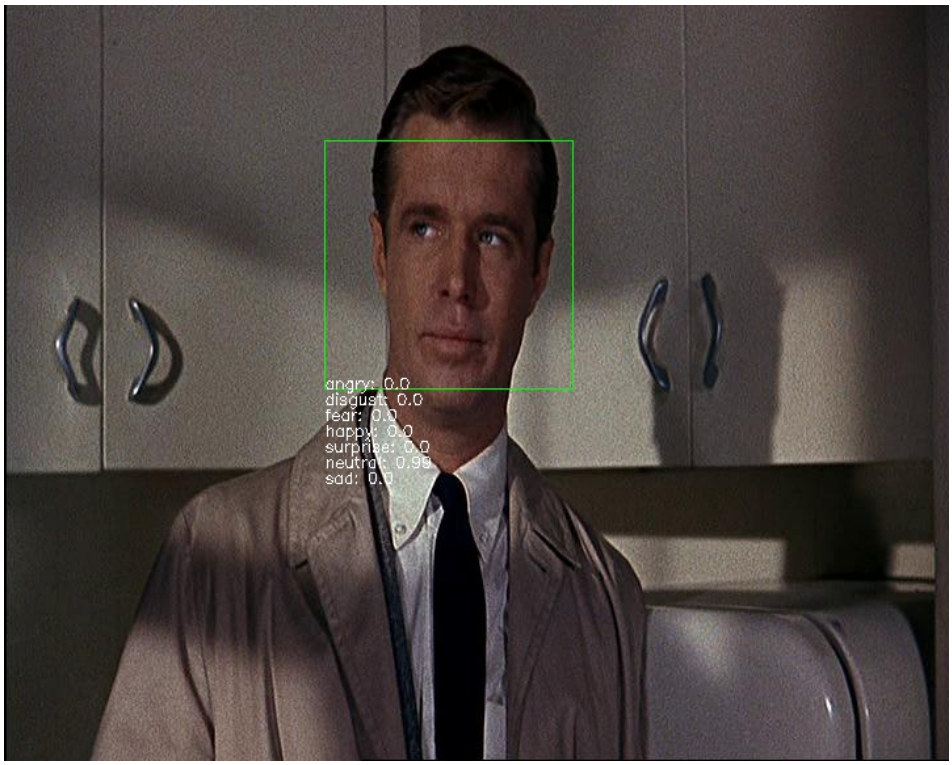

Figure 6. Frame with highest neutrality value in the corpus (Breakfast at Tiffany's)

Additionally, we performed a Welch-ANOVA to investigate if the movies differ to each other significantly (all requirements for the test are met [11]). Indeed, we do find significant differences (p<0.05) for all emotion categories but rather small effects according to Cohen (1988) [7] defining η^2<0.01 as weak, <0.06 as moderate and <.14 as strong effect. We report the p-, F- and η^2-value (table 5).

	p-value	F-value	η^2
angry	<0.001	302.11	0.06
disgust	<0.001	46.38	0.02
fear	<0.001	124.75	0.03
happy	<0.001	75.00	0.02
neutral	<0.001	431.98	0.12
sad	<0.001	32.21	0.01
surprise	<0.001	112.78	0.03

Table 5. Results of Welch-ANOVA-Tests for all emotion categories

The strongest effect can be seen for neutral. Performing post-hoc tests and inspecting a box-plots graph (figure 7) we identified *Breakfast at Tiffany's* as interesting outlier. This might be due to the fact that the main characters of the movie try to stay rather "unaffected" up until the ending of the movie while Wizard of Oz, as a musical, consist of strong emotional outbursts.

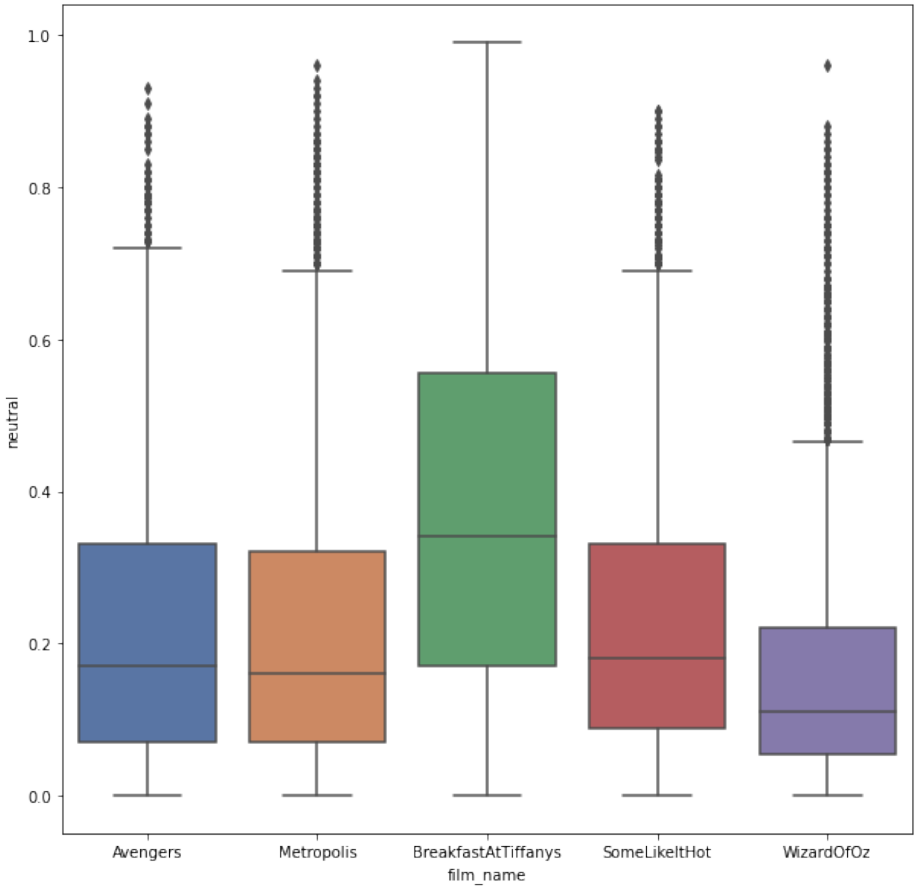

Figure 7. Box-plots graph for *neutral*

4.3. Gender- and Age-Recognition

Table 6 illustrates the descriptive statistics for the gender and age detection.

		Metropolis	Wizard of Oz	Some Like it Hot	Breakfast At Tiffany's	Avengers	Overall
Age	**M**	37.28	**35.73**	**40.55**	40.26	39.31	38.67
	Min	23.04	13.65	27.35	26.16	22.6	13.65
	Max	58.64	71.28	65.9	66.27	64.82	71.28
	Sd	3.81	7.67	5.77	5.88	4.96	6.01
Gender	**M**	0.35	**0.49**	0.41	0.38	**0.33**	0.39
	Min	0.01	0.02	0.02	0.01	0.01	0.01
	Max	0.91	0.96	0.94	0.97	0.98	0.98
	Sd	0.17	0.2	0.22	0.26	0.25	0.23

Table 6. Descriptive statistics for age and average gender

The average age is for most movies is around 40 which is a rather consistent over-estimation since most leading actors in the selected movies are around 30. Performing a Welch-ANOVA shows that the difference between the movies is significant (p<0.001, F=336.07, η^2=0.09) with a moderate effect. The strongest outlier movie, as shown with post hoc tests, is Wizard of Oz with a child/teenager as leading actor that gets correctly detected as around 14-16 years old (figure 8).

Figure 8. Lowest age in the corpus (Wizard of Oz)

An average score for gender below 0.5 points to more male detections and it is striking that all movies point below 0.5, thus a more frequent representation of males which is in line with the reality of the movies. There is a significant difference but with a smaller effect compared to age (p<0.001, F=251.36, η^2=0.06) and with the strongest differences concerning *Wizard of Oz*. The differences become apparent regarding the distribution of gender-classes (table 7). We assigned every frame with male if *average gender* > 0.6 and female if <0.4. We decided to include a class androgynous for in-between-values pointing to either multiple genders on one screen or uncertainty by the model.

		Metropolis	**Wizard of Oz**	**Some Like it Hot**	**Breakfast at Tiffany's**	**Avengers**
androgynous	# frames	689	**1,024**	838	638	**395**
	% frames	7.98	**17.49**	11.99	9.96	**4.8**
	% frames with faces	22.37	**40.54**	28.43	21.35	**18.3**
female	# frames	**290**	**694**	623	669	314
	% frames	**3.36**	**11.85**	8.91	10.44	3.82
	% frames with faces	**9.42**	**27.47**	21.14	22.39	14.6
male	# frames	**2,101**	808	**1,486**	1,681	1,441
	% frames	**24.33**	13.8	**21.26**	26.24	17.52
	% frames with faces	**68.21**	31.99	**21.26**	56.26	67.02

Table 7. Frequency distributions of gender classes

Wizard of Oz has the most frames classified as androgynous. In general, this means that female and male characters are equally on the frame but in this case the classification is due to the high number of human-like fantasy creatures for which the model is unsure to pick a gender (figure 9).

Figure 9. An “androgynous” face (Wizard of Oz)

5. Discussion

While this study war rather small and exploratory in the approach, we did gain important first insights for our future research. Overall, we find it promising that we were able to find significant results, even for this small set of movies. For object detection we see the most potential in adjusting pretrained models to objects that are of interest for a specific research question. We see a lot of potential for interesting diachronic but also genre-based emotion and gender analysis with larger corpora. For this case study, we did not find striking differences of method performance considering technical differences between the movies. We are planning systematic evaluations on a cross section of movies of different decades to get a better understanding on the performance of the methods before we move on to explore more concrete research questions. We see potential concerning research on the intercourse of gender and film studies. We plan to explore the relationship of gender representations with expressed emotions throughout the time to explore how the representation of gender roles developed.

References

1. Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., & Rothe, R. (2017). Apparent and Real Age Estimation in Still Images with Deep Residual Regressors on Appa-Real Database. 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 87–94. https://doi.org/10.1109/FG.2017.20

2. Arnold, T., & Tilton, L. (2019). Distant viewing: Analyzing large visual corpora. Digital Scholarship in the Humanities. https://doi.org/10.1093/digitalsh/fqz013

3. Baxter, M., Khitrova, D., & Tsivian, Y. (2017). Exploring cutting structure in film, with applications to the films of D. W. Griffith, Mack Sennett, and Charlie Chaplin. Digital Scholarship in the Humanities, 32(1), 1–16. https://doi.org/10.1093/llc/fqv035

4. Burghardt, M., Kao, M., & Walkowski, N. O. (2018). Scalable MovieBarcodes–An Exploratory Interface for the Analysis of Movies. In IEEE VIS Workshop on Visualization for the Digital Humanities (Vol. 2).

5. Burghardt, M., Kao, M., Wolff, C. (2016). Beyond Shot Lengths – Using Language Data and Color Information as Additional Parameters for Quantitative Movie Analysis. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 753-755.

6. Byszuk, J. (2020). The Voices of Doctor Who – How Stylometry Can be Useful in Revealing New Information About TV Series. Digital Humanities Quarterly, 014(4).

7. Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Academic press.

8. Cowan, Michael. (2007). The Heart Machine: „Rhythm“ and Body in Weimar Film and Fritz Lang's Metropolis. Modernism/Modernity, 14(2), 225–248. https://doi.org/10.1353/mod.2007.0030

9. Deldjoo, Y., Elahi, M., Cremonesi, P., Garzotto, F., & Piazzolla, P. (2016). Recommending Movies Based on Mise-en-Scene Design. Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 1540–1547. https://doi.org/10.1145/2851581.2892551

10. DeLong, J. (2015). Horseshoes, handgrenades, and model fitting: The lognormal distribution is a pretty good model for shot-length distribution of Hollywood films. Literary and Linguistic Computing, 30(1), 129–136. https://doi.org/10.1093/llc/fqt030

11. Field, A. P. (2009). Discovering statistics using SPSS: And sex, drugs and rock „n“ roll (3rd ed). SAGE Publications.

12. Flueckiger, B. (2017). A Digital Humanities Approach to Film Colors. The Moving Image: The Journal of the Association of Moving Image Archivists, 17(2), 71–94. JSTOR. https://doi.org/10.5749/movingimage.17.2.0071

13. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., … Bengio, Y. (2013). Challenges in Representation Learning: A report on three machine learning contests. arXiv:1307.0414 [cs, stat]. http://arxiv.org/abs/1307.0414

14. Halter, G., Ballester-Ripoll, R., Flueckiger, B., & Pajarola, R. (2019). VIAN: A Visual Annotation Tool for Film Analysis. Computer Graphics Forum, 38(3), 119–129. https://doi.org/10.1111/cgf.13676

15. Hołobut, A., Rybicki, J., and Woźniak, M. "Stylometry on the Silver Screen: Authorial and Translatorial Signals in Film Dialogue." Book of Abstracts of the International Digital Humanities Conference (DH) (2016).

16. Hołobut, A., & Rybicki, J. (2020). The Stylometry of Film Dialogue: Pros and Pitfalls. Digital Humanities Quarterly, 014(4).

17. Howanitz, G., Bermeitinger, B., Radisch, E., Sebastian G., Rehbein, M., and Handschuh, S. "Deep Watching - Towards New Methods of Analyzing Visual Media in Cultural Studies." Book of Abstracts of the International Digital Humanities Conference (DH) (2019).

18. Hoyt, E., Ponto, K., & Roy, C. (2014). Visualizing and Analyzing the Hollywood Screenplay with ScripThreads. Digital Humanities Quarterly, 008(4).

19. Kuhn, V., Craig, A., Simeone, M., Satheesan, S. P., & Marini, L. (2015). The VAT: Enhanced video analysis. Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure, 1–4. https://doi.org/10.1145/2792745.2792756

20. Kurzhals, K., John, M., Heimerl, F., Kuznecov, P., & Weiskopf, D. (2016). Visual Movie Analytics. IEEE Transactions on Multimedia, 18(11), 2149–2160. https://doi.org/10.1109/TMM.2016.2614184

21. Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs]. http://arxiv.org/abs/1405.0312

22. Masson, E., Olesen, C. G., Noord, N. van, & Fossati, G. (2020). Exploring Digitised Moving Image Collections: The SEMIA Project, Visual Analysis and the Turn to Abstraction. Digital Humanities Quarterly, 014(4).

23. Pause, J., & Walkowski, N. O. (2018). Everything is illuminated. Zur numerischen Analyse von Farbigkeit in Filmen. Zeitschrift für digitale Geisteswissenschaften.

24. Pustu-Iren, K., Sittel, J., Mauer, R., Bulgakowa, O., & Ewerth, R. (2020). Automated Visual Content Analysis for Film Studies: Current Status and Challenges. Digital Humanities Quarterly, 014(4).

25. Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. International Journal of Computer Vision, 126(2), 144–157. https://doi.org/10.1007/s11263-016-0940-3

26. Salt, B. (1974). Statistical style analysis of motion pictures. Film Quarterly, 28(1), 13-22.

27. Vonderau, P. (2020). Quantitative Werkzeuge. In M. Hagener & V. Pantenburg (Hrsg.), Handbuch Filmanalyse (S. 399–413). Springer Fachmedien. https://doi.org/10.1007/978-3-658-13339-9_28

28. Wei, C. Y., Dimitrova, N., & Chang, S. F. (2004, June). Color-mood analysis of films based on syntactic and psychological models. In 2004 IEEE international conference on multimedia and expo (ICME)(IEEE Cat. No. 04TH8763) (Vol. 2, pp. 831-834). IEEE.

29. Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2

30. Wulff, H. J. (1998). Semiotik der Filmanalyse: Ein Beitrag zur Methodologie und Kritik filmischer Werkanalyse. Kodikas/Code, 21(1-2), 19-36.

31. Zaharieva, M., & Breiteneder, C. (2012). Recurring Element Detection in Movies. In K. Schoeffmann, B. Merialdo, A. G. Hauptmann, C.-W. Ngo, Y. Andreopoulos, & C. Breiteneder (Hrsg.), Advances in Multimedia Modeling (S. 222–232). Springer. https://doi.org/10.1007/978-3-642-27355-1_22

32. Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, 23(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342