Grósz, Tamás; Kallioniemi, Noora; Kiiskinen, Harri; Laine, Kimmo; Moisio, Anssi; Römpötti, Tommi; Virkkunen, Anja; Salmi, Hannu; Kurimo, Mikko; Laaksonen, Jorma

Tracing Signs of Urbanity in the Finnish Fiction Film of the 1950s: Toward a Multimodal Analysis of Audiovisual Data

Published in:
Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Long Papers

Published: 01/01/2022

Document Version
Publisher's PDF, also known as Version of record

Tracing Signs of Urbanity in the Finnish Fiction Film of the 1950s: Toward a Multimodal Analysis of Audiovisual Data

Tamás Grósz[1], Noora Kallioniemi[2], Harri Kiiskinen[2], Kimmo Laine[3], Anssi Moisio[1], Tommi Römpötti[3], Anja Virkkunen[1], Hannu Salmi[2], Mikko Kurimo[1], Jorma Laaksonen[4]

[1] *Department of Signal Processing and Acoustics, Aalto University, Finland*
[2] *Department of Cultural History, University of Turku, Finland*
[3] *Department of Media Studies, University of Turku, Finland*
[4] *Department of Computer Science, Aalto University, Finland*

Abstract

This paper traces signs of urban culture in Finnish fiction films from the 1950s by drawing on a multimodal analysis of audiovisual content. The Finnish National Filmography includes 208 feature films released between 1950–1959. Our approach to the automatic analysis of media content includes aural and visual object recognition and speech recognition. We concentrate on features that epitomize urbanity, including visual objects, such as forms of transportation (cars, horses) and sounds (rural and urban sounds, speech). Based on the scores and frequencies of these recognitions, we observe quantitative changes that took place during the 1950s. The paper demonstrates that aural and visual object recognition, as well as speech recognition, can successfully be applied in film historical analysis. The overall results support the idea that Finnish filmmakers fueled the imagination of urban life in the 1950s, paving the way for modern technologies and gradually pushing aside the signs of rural life.

Keywords

film history, modernization, media analysis, multimodal analysis, automatic speech recognition, computer vision, object detection

1. Introduction

This paper aims to trace signs of urban culture in Finnish fiction films from the 1950s through computational methods and by drawing on multimodal analysis of audiovisual content. In recent years, *distant viewing* [1] has increasingly been employed as a complementary concept to Franco Moretti's *distant reading* [2]. One of the problems in the field is that many of the digital tools developed for the multimodal analysis of audiovisual content are based on modern or present-day videos. The state-of-the-art tools used for distant viewing do not necessarily, as in our case, directly apply to old film material, but it is important to explore the possibilities, and limitations, of a computational analysis.

In cinema studies, there are two major links with the digital humanities. One is an emerging research field called digital cinema studies, which concentrates on phenomena such as movie theater networks and production and consumption indicators [3]. The other one is Cinemetrics, which draws on digital methods in approaching film style. Cinemetrics involves open-source film measurement software, developed for counting the number of shots, shot-scales, etc., in a film [4]. Containing more than 20,000 entries, the Cinemetrics database is extremely useful for studying historical variations in stylistic features. This article, however, suggests a third perspective by concentrating more on film content and less on style, and we found it more useful to draw on existing research into audio events and object detection.

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022
EMAIL: tamas.grosz@aalto.fi (T. Grósz); nmmkal@utu.fi (N. Kallioniemi); harri.kiiskinen@utu.fi (H. Kiiskinen); kimmo.laine@utu.fi (K. Laine); anssi.moisio@aalto.fi (A. Moisio); tomrom@utu.fi (T. Römpötti); anja.virkkunen@aalto.fi (A. Virkkunen); hansalmi@utu.fi (H. Salmi); mikko.kurimo@aalto.fi (M. Kurimo); jorma.laaksonen@aalto.fi (J. Laaksonen).
ORCID: 0000-0001-7918-9579 (T. Grósz); 0000-0001-9934-1478 (N. Kallioniemi); 0000-0003-4187-5551 (H. Kiiskinen); 0000-0001-5385-2184 (K. Laine); 0000-0002-2018-3524 (A. Moisio); 0000-0002-3623-0014 (T. Römpötti); 0000-0001-8607-6126 (H. Salmi); 0000-0001-5278-7974 (M. Kurimo); 0000-0001-7218-3131 (J. Laaksonen).

The 1950s offer a fruitful opportunity for testing the possibilities of computational analysis in film history. More feature films were produced annually than at any other time, despite the decreasing attendance and constantly rising production costs [5]. During the years 1950–1959, on average, 21 full-length feature films were released every year. The decade was characterized by rivalry between three big companies: Suomi-Filmi (Su-Fi), Suomen Filmiteollisuus (SF), and Fennada-Filmi (Fennada). Impacting SF in particular, the commercial success of *The Unknown Soldier* (Tuntematon Sotilas) in 1955 brought capital for new productions.

From the historical perspective, the 1950s were also significant in other ways. At the end of the 1940s, in the aftermath of World War II, Finnish society shifted from war-time conditions to an era of peace, and the next decade saw an economic rise and a boom in internationalization. Finland was still a rural country, but urbanization was already progressing strongly. This article is interested in the transformation of Finland from a rural, agrarian society into an industrial, urbanized country. Numerous studies consider the cultural and social history of Finnish modernization [6][7], but we concentrate on how cinema contributed to this process by circulating images of urbanity. This article asks: How can computational tools, especially those of aural and visual object recognition and speech recognition, be employed in the study of audiovisual heritage, in this case, the 1950s Finnish fiction films, to make visible the cinematic markers of urbanization? The number of films for this analysis is rather limited, but if these tools can be meaningfully employed, they will have wider relevance and applicability to the study of longer time spans and larger datasets.

In the study of modernization and the transformation of Finnish society, it is crucial to note that Finland urbanized rather late on the European scale. Its most intensive decade was the 1960s, but the 1950s paved the way for this development. For our analysis, we chose particular signs of urbanity that we could follow computationally. We concentrated on the forms of traffic, and especially, on the occurrence of horses vs. cars and other motorized vehicles.

At the end of 1945, after the war, there were 26,310 registered cars in Finland. Only 6230, that is, less than a quarter of all cars, were privately owned. Furthermore, there were only a few kilometers of asphalted roads, and the car traffic, in its entirety, was dependent on imports. The outbreak of the Finnish-Soviet Winter War in November 1939 was followed by rationing and expropriation. Lorries and coaches were expropriated by the defense force, and almost all traffic-related resources were reserved for the defense force and commercial traffic. Nearly all private car use had already come to a halt with the control of petrol, which was initiated at the start of World War II [8, 9, 10].

In the 1950s, Finnish vehicular traffic was dominated by lorries and coaches. The number of automobiles on the roads had reached the pre-war level in 1949, but private cars only did so in 1951 [9]. A major reason for this was the poor condition of the roads. However, the number of registered private cars climbed by over 50% during 1952, the year of the Helsinki Summer Olympics. Compared to the other Nordic countries, the modernization of traffic progressed slowly in Finland. At the beginning of the decade, there were 61,256 cars in Finland, but only 26,814 of those were private cars (6/1000 inhabitants). This was only one-sixth of all cars on the roads in Sweden.

In this study, we focused on developing automatic methods to detect the presence of automobiles, horses, and other typical urban and rural objects appearing in the movies. We used these methods to track how the urbanization process affected Finnish cinema.

2. Film Dataset

The Finnish National Filmography includes 208 fiction films premiered between January 1, 1950, and December 31, 1959. For copyright reasons, only 186 of the films were available for examination. We secured research permission to access all Su-Fi, SF, and Fennada productions, as well as Teuvo Tulio productions, but it was not possible to obtain digital copies of films made by the many small production companies. However, 186 films represent 89% of all fiction films released during the 1950s, which means that our dataset covered the period sufficiently well.

On inspection, we identified 46 of the movies as *historical films*, set in a time clearly prior to the

production of the film. There are many definitions of historical film [11], but we made a practical decision that, from the perspective of the 1950s, films that portrayed a time clearly prior to the production of the film were historical, since one would not expect to see cars and other signs of urbanization in periods when they did not yet exist. The boundary was by no means clear cut, since there were cars, for example, in films set in the 1920s or 1930s, which may have influenced the results. Most of the historical films were, however, set in the nineteenth century. The remaining 140 films, called *non-historical* films in this article, thus depicted contemporary time, the 1950s. The films were distributed over these years, as shown in the histogram of Figure 1. Of course, our definition of historical films is ambiguous from the perspective of our research problem. As pointed out, films portraying, for example, the 1920s can also show urban imagery and thus contribute to the construction of urban imagination [12]. Therefore, we present results for both the whole dataset and the subset of non-historical films. The first Finnish color film, *Juha* (SF), premiered in November 1956. In all, only five of the 186 films examined were color films.

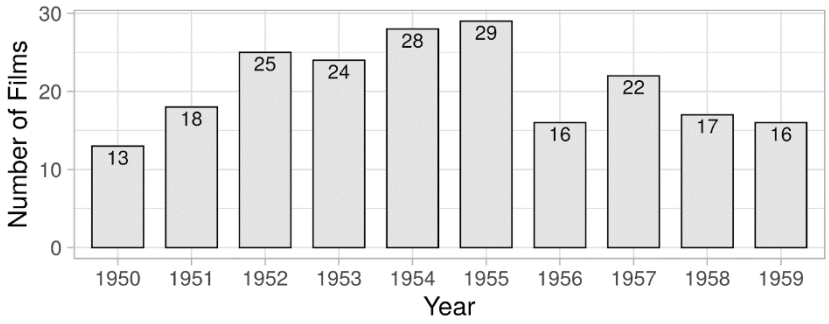


Figure 1: The number of films per year.

In addition to the audiovisual sources, we had 44 subtitled files at our disposal, which contributed to the study of automatic speech recognition (ASR) and the audio events. Furthermore, we were able to draw on the metadata for all the films analyzed, which was made available through the Elonet database (https://elonet.finna.fi/), maintained by the National Audiovisual Institute of Finland. The metadata includes textual movie descriptions that describe the events and locations in the films. However, these descriptions were produced considerably later, in the 1980s and 1990s, when the national filmography dataset was created. One has to ask whether these content descriptions are more reflective of their own times than of the times in the films they describe. Whereas a human reader can (possibly) see through the temporal lens of the later descriptions, using this data with any machine learning tool could be problematic. This problem was studied by another group within the same project. (See the paper "Deep learning and film history: Model explanation techniques in the analysis of temporality in Finnish fiction film metadata" by Ginter et al. in this same collection.)

3. Methods

Our approaches to the automatic analysis of media content included aural sound and speech recognition and visual object detection. We concentrated on particular features that epitomized urbanity, including visual objects, such as forms of transportation (cars, motorcycles, trains), and sounds (traffic and animal sounds, speech). One of the methodological problems we faced was related to the fact that most state-of-the-art tools have been trained with modern audiovisual materials. Old films may include, for example, car models that are difficult to identify as cars without historically sensitive reference data. The same can be argued for the auditive elements: the sounds of the past may be different from the sounds of the present. Furthermore, in film production, various stock sounds and effects were used, and visual and aural objects were not always in a realistic relationship with each other. For example, a motorboat sound may well have originally been any motor sound that was at hand.

3.1. Audio processing

To automatically analyze the sounds of films, we created a simple pipeline (see Figure 2), consisting of two essential modules. First, we processed the movies with a neural network, trained for automatic sound event detection, to find the sounds of relevant objects such as cars, horses, and trains. These predicted audio tags were then used to track the urbanization/modernization process presented in the films. Additionally, we wanted to take advantage of conversations in the movies. Unfortunately, only about 25% of the films had subtitles, so we resorted to employing ASR-based transcriptions. We tried out multiple ASR models and selected the best one for generating the transcriptions needed to analyze the content and language of the films. The available subtitles were used for the validation and selection of the best ASR system.

The first component of our audio processing system was the audio pattern recognition model. After our preliminary experiments, we picked a pre-trained audio neural network (PANNs) presented by Kong et al. in 2019 [13]. The CNN14 model that we selected was trained on the AudioSet corpus [14] to recognize 527 sound classes, and it demonstrated high performance on several other datasets, including our initial small validation set. In our application, we opted to generate audio tags for one-second-long, non-overlapping parts to avoid sudden changes due to mistakes (caused mainly by audio quality).

The audio tagging model was used for two purposes: First, we could analyze the predicted tags to track for the appearance of different objects. Additionally, we used the speech tag predictions from the model to identify the parts of the films where we would execute our speech recognition system. This allowed us to reduce the amount of data that we needed to process, and at the same time, it split the spoken parts into shorter conversations.

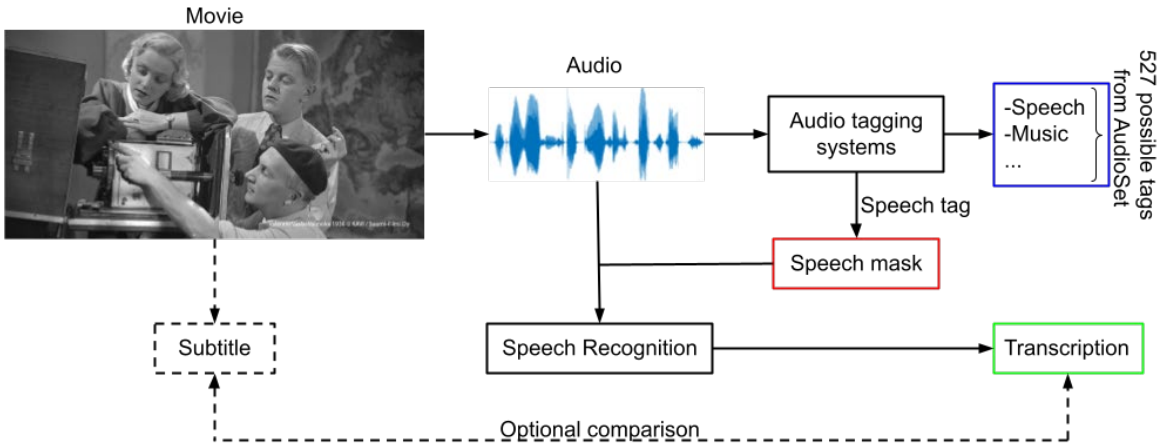


Figure 2: Pipeline developed for processing the soundtrack of movies.

Using the audio tagging system, we first investigated two tags, namely, the sounds of cars and horses. Figure 3 shows the yearly average frequencies of these two labels. To get these statistics, we first calculated the relative frequencies per movie (i.e., the proportion of the film where these sounds were detected), we then aggregated the values by averaging the statistics for movies released during the same year. We could see a steep increase in the detected car sounds, especially during the latter half of the 1950s, suggesting that they had become an increasingly common object to be heard in films. At the same time, the frequency of horse sounds remained relatively low and stable.

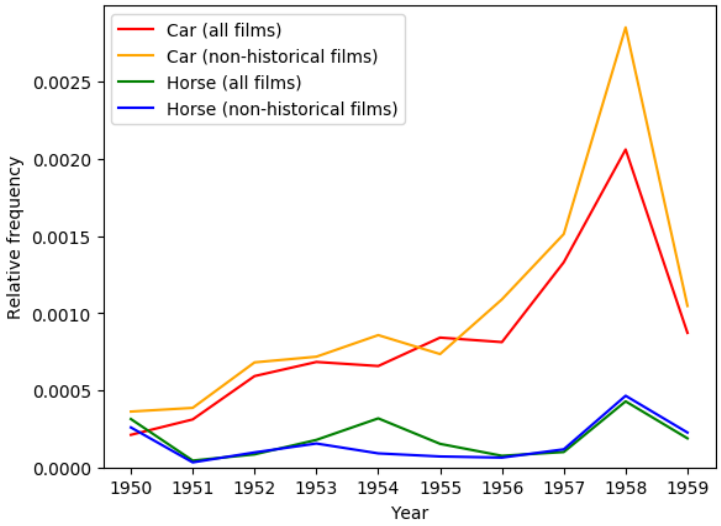


Figure 3: Average frequencies of the detected horse and car sounds.

For our second analysis, we manually selected a list of urban and rural sounds that our system could recognize. While the urban list primarily focused on transportation objects, such as cars and trains, it also contained tags such as telephone and television sounds. The rural category mostly consisted of animal sounds (pig, horse, cow, etc.) and the sounds associated with boats (rowboat, canoe, kayak). Once the lists were finalized, it was possible to compare the frequencies. Figure 4 shows that the urban category exhibited an upward trend while the frequency of rural sounds, with some variations, remained on roughly the same level.

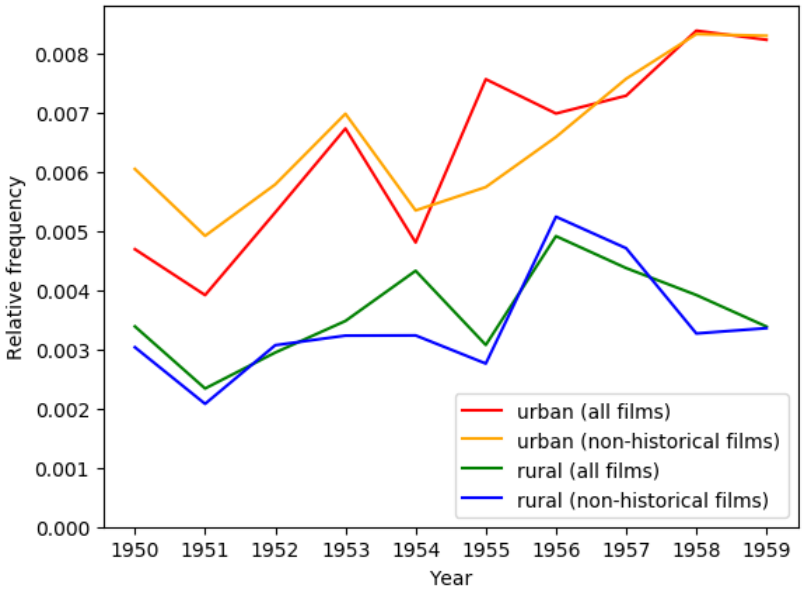


Figure 4: Trends of urban and rural sounds detected in the films.

Next, we used the ASR system to transcribe those parts of the movie audio where the audio tagging system detected speech. Our ASR system was a hybrid system trained with the Kaldi toolkit [15]. The training dataset contains about 1600 hours of transcribed speech from the Lahjoita Puhetta corpus. The corpus and the ASR training process are described in detail by Moisio et al. [16]. The language model (LM) used for the ASR was a 4-gram model trained for conversational Finnish using the SRILM toolkit. The LM training corpus includes 76M word tokens from the WEBCON [17] corpus pooled with 166M word tokens from the Open Subtitles corpus. The latter allowed us to prepare the LM for the style of language used in movies (and TV), although the Open Subtitles corpus did not match well with the time period or the genres of the films examined in this paper.

From the ASR transcripts, we searched for two sets of keywords that would indicate either rural or

urban environments in the movie. The basis for these lists were the audio tag labels, but we modified and extended these lists to include words that would be more likely to occur in conversation (e.g., “train horn” is a good audio tag class but not a good keyword to search for in a conversation). The initial lists were each about twenty words long, but to increase the sample size for the keywords, we obtained more words using a Word2Vec model [18]. Using the Word2Vec model, we extracted about 150 words that were semantically most similar to the two initial lists. We pooled these with the initial lists, and then searched for these keywords in the transcripts. Figure 5 shows the number of rural and urban keywords detected in the movies from each year, divided by the total number of words in the movies from each year, to give the relative frequency. No clear upward or downward trend can be seen for either of the classes in this figure. As expected, however, the historical films included more rural and fewer urban keywords, which was visible in the differences between the curves for all films and for the non-historical films.

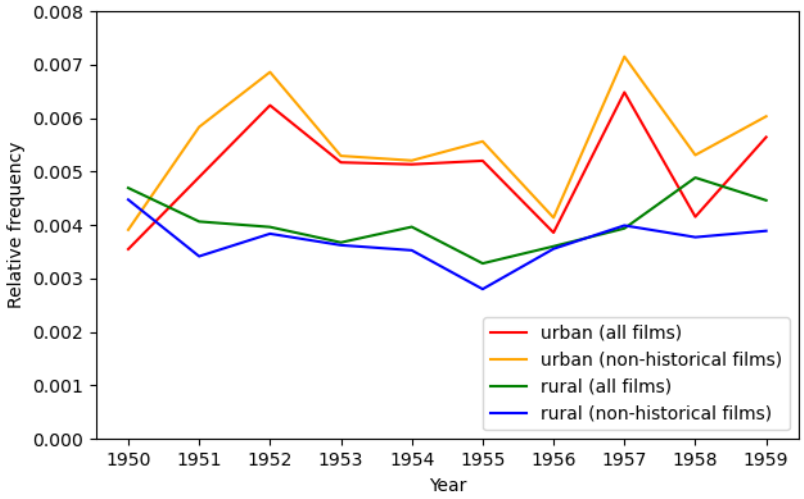


Figure 5: Trends for urban and rural keywords detected in the speech recognition transcripts of the movie conversations.

3.1.1. Effects of sound quality on film

Naturally, the old movies had markedly different sound quality from the modern data used to train the models. In fact, the 1950s was the decade when a gradual shift from optical sound to magnetic sound occurred in Finland [19]. However, this was quite different from later sound systems. This discrepancy raises the question: How much degradation could we expect due to the difference in quality?

First, we needed to know how well the model worked with modern data. From the paper by Kong et al. [13], we could see that, for the most common classes, such as speech and music, we could expect a performance of around 85%. For common sounds, the average precision (AP) varies between 20% and 60%, with some exceptions, and most sounds can be recognized with at least 20% AP. Looking at some important classes, we could see that trains are detected with 70% AP, vehicles with 50% AP and cars with approximately 40% AP. In the rural category, horse sounds have an AP of 50%.

We utilized two resources to re-evaluate the performances: first, we had the subtitles that were available for some movies and second, we had the annotations for several, randomly selected, short movie parts. In total, 44 movies had subtitles that we could use to estimate the accuracy of the ASR model and the audio tagging system. First, we extracted the time information from the subtitle files, assuming that if a subtitle is shown, then there must be speech, and we compared it with the audio tagging system’s speech tags. We saw that the average overlap between the two was around 85%. In general, the tagging system managed to find most parts where subtitles were shown, but it usually predicted a longer duration (some additional time before and after the subtitles). We must note that the time information contained in subtitles is not always accurate, as sometimes the subtitles are displayed for longer than the actual conversation. The false positives could largely be explained by the fact that the tagging system worked with one-second long segments, while the subtitles had exact timestamps. Overall, we can say that the 85% overlap meant that the system performed quite well in detecting the speech parts, despite the adverse quality of the soundtracks.

For further validation experiments, 50 movies were selected randomly, and short clips from those movies were annotated (both the spoken text and the sounds of objects). This process gave us approximately 100 minutes of manually annotated evaluation data. The validity of all automatically obtained findings was manually verified using this data.

We focused on three categories to evaluate the automatically generated tags, namely music, modern/urban, and rural sounds. Music, being the second most common tag in the data and an integral part of the movies, was recognized with extremely high accuracy: the recall was 100%, and the precision reached about 85%. This meant that every time the annotator marked an audio clip with the music tag, the system also produced that tag; furthermore, it produced only a few false positives.

Next, we investigated the tags for rural sounds (mostly animals). Based on this validation set, we estimated that the model had an accuracy of 62.5% in recognizing those sounds. For some classes, such as the horse sounds, the system managed to achieve an astonishing 100% recognition rate.

Moving on to the modern and urban sounds, we faced the issue of data sparsity, as most of the labels were present in only one or two clips. Thus, estimating individual accuracy became quite problematic, as we did not have enough data to predict a representative performance value. We saw a wide variation in individual performance: on average, labels in this category were recognized with approximately 50% precision, but many classes had lower performances than expected.

Gunshot sounds were recognized with approximately 50% accuracy and were often mislabeled as firecrackers. Train sounds were recognized with 25% accuracy, which was a huge drop compared to the original performance on the AudioSet data. Similarly, boats were recognized with low precision, and they were often misclassified as cars. Other common mistakes included a confusion between a timecard machine and a cash register's sound and using a general vehicle tag instead of the specific class (motorboat, train, or car).

In summary, common things like speech, music, and animal sounds were recognized well, but the man-made technological objects had low precision recognition due to the sound quality. Fortunately, the low performance was counterbalanced by the fact that the modern sounds were generally confused with other modern sounds. Thus, we concluded that the model could be used to track the sounds of urbanization.

We calculated the speech recognition accuracy against the 44 subtitled films that were available. The word error rate (WER) was about 62.3%, and the character error rate (CER) was 40.4%, considerably worse than the 23.8% WER and 9.5% CER achieved on the Lahjoita Puhetta test set. An important factor that could partially explain this large gap is that subtitles are usually not written as exact transcriptions of what is uttered: they summarize the speech to fit the limited space available, and sometimes convert colloquial, informal word forms into a more formal style. Therefore, these numbers were somewhat higher than the true WER and CER.[1] To test this, we also performed the evaluation against the ~100 minutes of manually annotated audio, which included the transcripts of conversations. As expected, the numbers were lower for these manual evaluation data: about 52.9% and 27.5% for the WER and CER, respectively. Furthermore, we calculated the error rates for the manual transcripts against the subtitles to get an idea of how much the subtitles differed from the exact transcripts. The WER in this case was 42.2%, and the CER was 22.6%, showing that subtitles are most definitely not exact transcripts.

Last, we calculated the accuracy of the keyword detection by comparing it against the subtitles. Unlike the movie dialogue in general, we assumed that the keywords were accurately transcribed in the subtitles. This is because keywords are content words that cannot usually be summarized away, and because we used word stems (e.g., "hevonen" was truncated to "hevo"), so the difference between the formal and colloquial word endings (e.g., "hevosia" and "hevosii") would not affect the results. When combined, the two lists of keywords included 350 words that appeared in total 1804 times in the subtitles. The ASR system's overall recall for the keywords was 0.617, and the precision was 0.698 (F1 score 0.655).

3.2. Visual object detection

For the visual content, we applied visual object detectors that were able to recognize 80 common everyday objects that typically appear in the photos and videos that people upload to photo-sharing

[1] In Finnish, we typically measure both WER and CER, because words in agglutinative languages are often long, and the CER reflects better the number of completely wrong words compared to minor errors in some subwords.

services such as Flickr. These object classes include *airplane, apple, backpack, banana, ...* We used the readily-available Detectron2 object detector model [20] of Meta Research, which has been trained with the 120,000 image COCO dataset [21] of Microsoft Research. For each frame of each processed movie, the object detector first extracts the bounding boxes of the assumed object locations [22]. These box areas are then classified as belonging to the best matching of the 80 classes, and those classes that exceed a generally applicable preset threshold of 0.25 as the *detection score* are retained.

As a postprocessing step for the object detection, we formed *trajectories* of the detected objects. In this process, we required that a detected object should be visible in mutually overlapping locations on the screen for a minimum of 50 frames (i.e., approximately two seconds), while simultaneously allowing gaps of maximally ten consecutive frames of missed detection in the intermediate frames. Such gaps typically result from short occlusions and detector failures due to atypical viewing angles. This processing reduces the number of erroneous detections or, more precisely, lowers the *false-positive* detection rate or the *sensitivity* of the detector. Simultaneously, it also merges short object detection trajectories into longer, more meaningful trajectories.

Object detection and trajectory postprocessing were applied to all 186 movies in our test set. For each film, we observed the fraction of frames where each object class appeared at least once, relative to the total number of frames in the movie. This relative frequency measure is thus a non-negative value bound above by value one, and the larger the value is, the more frequently that object class appears in the film. Due to earlier human inspection, we knew that the object class car appeared in 124 of the 186 films in our data, but our annotation was limited to the *existence* of cars in the movies and did not reveal their *frequency*. Based on this annotation, we were able to get a rough estimate of the performance of our detections by means of the *car* detector's *receiver operating characteristic* (ROC) curve. The ROC curve showed how true positive and false positive *car* detections were distributed in the films, sorted according to decreasing detection scores. In Figure 6, we can see the ROC curve and how the true positive detections correctly dominate the bottom-left head of the detections scores, whereas false positives are found predominantly in the tail. The area under the ROC curve (ROC AUC), that is, the part of the plot that resides between the ROC curve and the x-axis, was measured to be 0.894, which can be considered a good detection result.

Similarly to the sound detection, we investigated the relative frequencies of the visual detection of *horses* and *cars* in the study material. Based on the preceding analysis of the existence of cars in the films, it was clear that the predicted frequencies contained errors caused by both false positive and false negative results. According to our limited manual evaluation, however, it seemed that these two types of error balanced each other out, and the obtained relative frequencies could thus be considered indicative of the true measures, with moderate additive noise. Figure 7 shows, separately for all films and for the non-historical ones, for each year, the relative frequencies of these two object classes for individual films (stars), the median and two middle-most quartals (boxes), and the 1.5 times longer extreme ranges (whiskers) of the values. Furthermore, the means of the yearly observations and their trends over the decade are shown. Overall, the trends in the *horse* and *car* detections are clear and will be addressed in more detail in the following section.

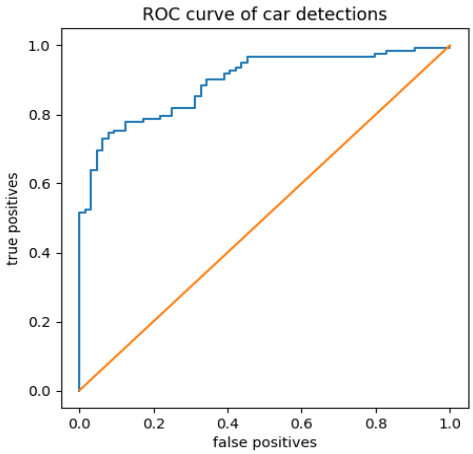


Figure 6: Receiver operating characteristic (ROC) curve of *car* detections in the full set of 186 1950s movies. The area under the ROC curve (ROC AUC) is 0.894 units.

It should be noted that the visual object detector and classifier model had been trained with color images of modern contents, whereas the majority of movies where they were applied were black and white, and the style of the cars was notably different from that of the contemporary ones. On the other hand, horses had not changed much since the filming of the 1950s movies, but their surroundings could be expected to be different. In our limited visual verification of the detection results, we were unable to find any systematic errors that could be explained by the colors of the depicted horses and cars or by the shapes of the latter.

4. Changes in transportation in the 1950s

We studied how the frequencies of the various aural and visual objects detected varied during the decade of the 1950s. Our first assumption was that such changes would be observed, for example, for *horses* and *cars*. In particular, the frequency of the latter was expected to increase. We studied this behavior for all 186 movies in the sample and also, separately, for the subset of non-historical films.

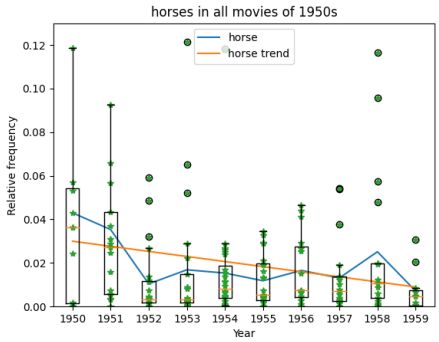
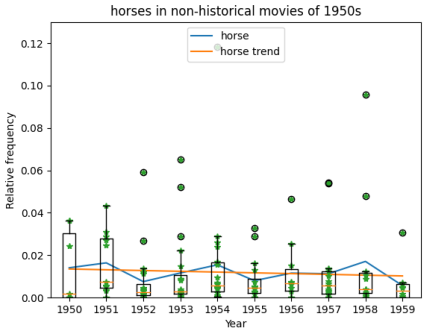
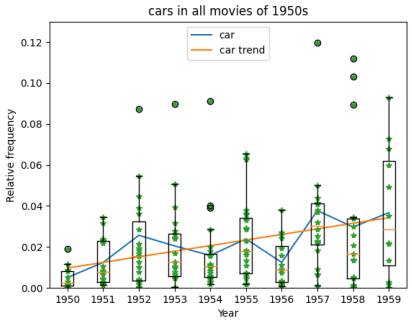
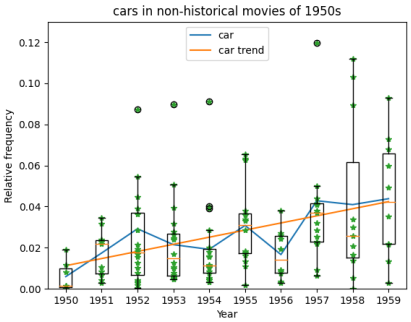
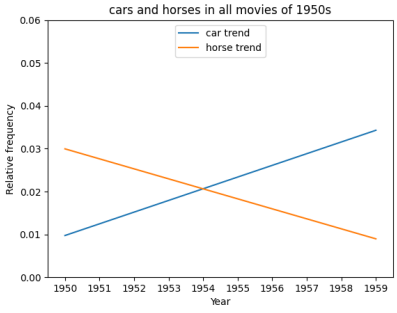
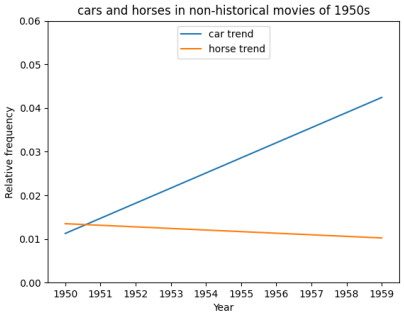
	All films	Non-historical films
Visually detected horses		
Visually detected cars		
Visually detected cars & horses		

Figure 7: The measured detection of horses and cars in the individual films, with their variation boxes, whiskers and means by year, and the trends over the decade.

Figure 7 shows the results of our visual object detection and indicates that in the dataset for all films, the visibility of horses declined throughout the 1950s, while the occurrence of cars increased. The latter increase is predictable since the number of cars in Finland was steadily growing. At the same time, as a small, closed, and intimate space, the interior of a car was increasingly used as a dramatic component on the screen, especially after the increase in private cars. During the 1950s, American influences and the car culture started to flow into Finland but, in relation to private cars, Finland was the promised land for eastern cars: between 1953 and 1957, approximately 60% of private cars on Finnish roads were manufactured in the Soviet Union and the Eastern bloc countries of Europe. When the standard of living began to rise in the latter half of the 1950s, the number of cars on the road increased faster than the number of new registrations, which indicates that there were also illegally imported cars on Finnish roads [23]. Figure 8 shows the increasing number of cars and the share of private cars during the 1950s. The increase in all cars matches quite accurately the rising trend in the relative frequency of visually detected cars as depicted in Figure 7.

YEAR	ALL CARS	PRIVATE	PRIVATE %
1950	61 256	26 814	43,8 %
1953	110 983	51 308	46,0 %
1956	171 547	102 371	59,5 %
1959	225 622	160 419	71,0 %

Figure 8: Registered cars on the road in Finland 1950–1959.

An interesting difference can be seen in Figure 7 between the whole dataset and the subset non-historical films. The trend in the visibility of cars seems to be about the same in both sets, but the role of horses differs. In the subset of non-historical films, horses appear in a rather stable manner throughout the 1950s. This reminds us of the fact that horses were still an everyday sight, both in the city and in the countryside. The number of private cars had increased, but their distribution in the country was not even. Most private cars were registered and driven in the south of the country, in the Tampere area, and in the Vaasa area on the west coast [24, 25]. The planning of the first highway project, the Turku–Helsinki motorway in the south, began in 1956, but the first part of it was not completed until 1962 [26]. This was the region where most of the films were produced.

It is, of course, to be expected that the number of horses used in traffic would decrease. However, because cars were unevenly distributed, horses were still used in the countryside. According to the traffic census carried out late in the summer of 1950, the number of horses on the roads had decreased to five percent of traffic. Traffic-wise, the pace of modernization seems to have been increasing. However, as the traffic census took place only on public roads, the number of horses observed was less than the reality: horses were certainly used on minor roads and, for example, in winter traffic and to draw log loads [27]. In 1955, there were still twice as many horses as there were cars in the country, but they were mostly in fields and forests, and on the back roads. Despite the large number of horses, changes in traffic patterns indicate that modernization was occurring: simultaneously, as the number of cars increased, the quantity of horse traffic decreased. This is also evident in the films.

In addition to the visual analysis, it was important to pay attention to the presence of horses and cars on the audio track. The average frequencies of horse and car sounds have already been presented in Figure 3. There was clearly a rise in car sounds toward the end of the decade. Due to false positives, this might include other modern motor sounds, but the overall trend is evident. We can also compare the visual and

aural presence of horses and cars. These are presented in Figures 9 and 10 (horses) and Figures 11 and 12 (cars), which represent the whole dataset, including the historical films. The figures also indicate which particular films emphasize horses and cars. Figure 9 shows that horses are very visible in historical films, such as *Tanssi yli hautojen* (SF, 1950), *Kuningas kulkureitten* (Fennada, 1953), and *Sven Tuuva* (SF, 1958). The horses are generally less aural than visible, with the exception of *Sven Tuuva*, in which horse sounds are heard to a considerable extent (Figure 10).

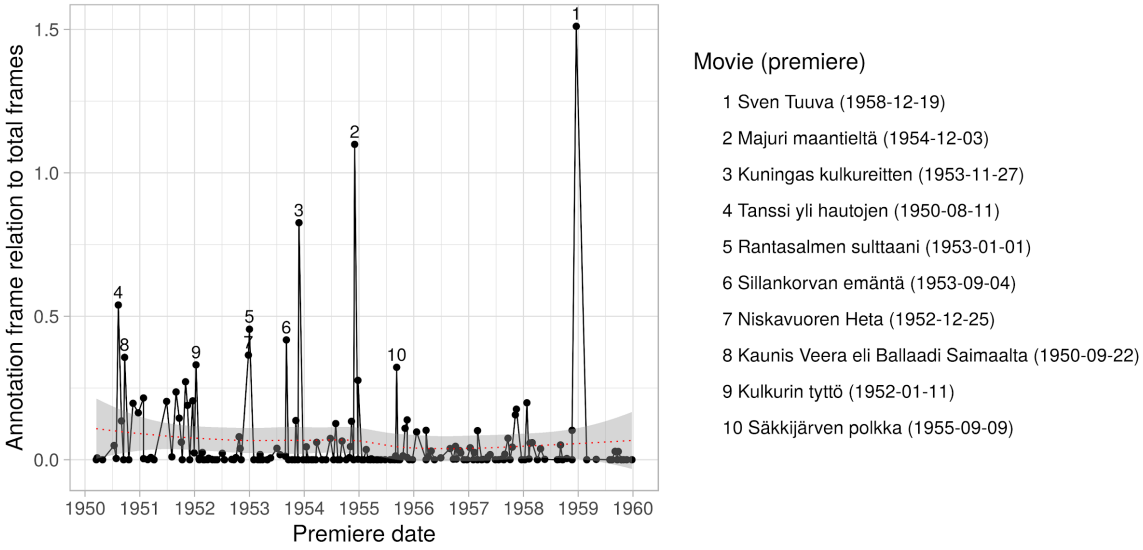


Figure 9: Horse objects in films

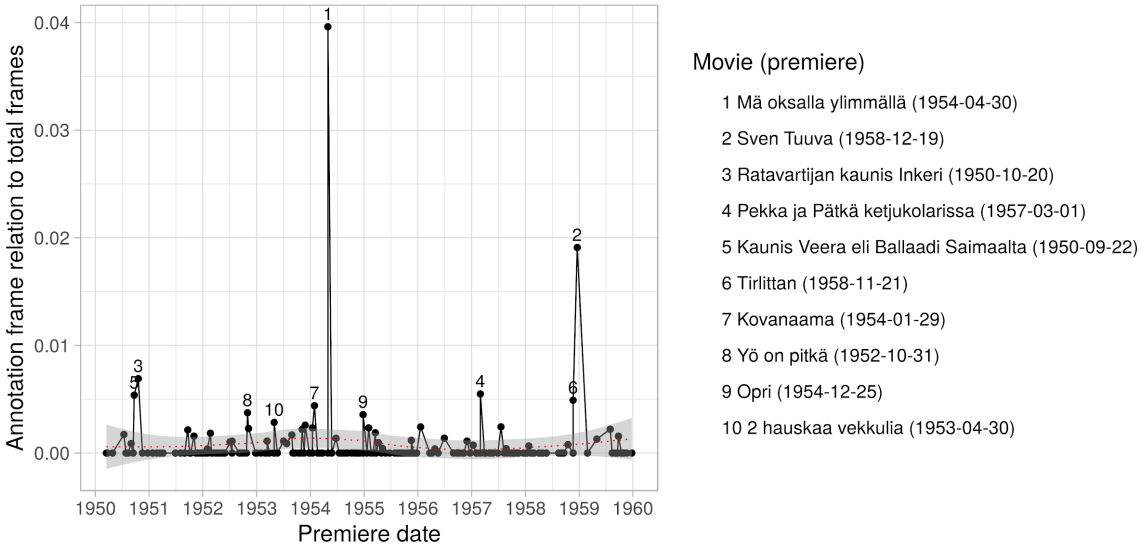


Figure 10: Horse sounds in films

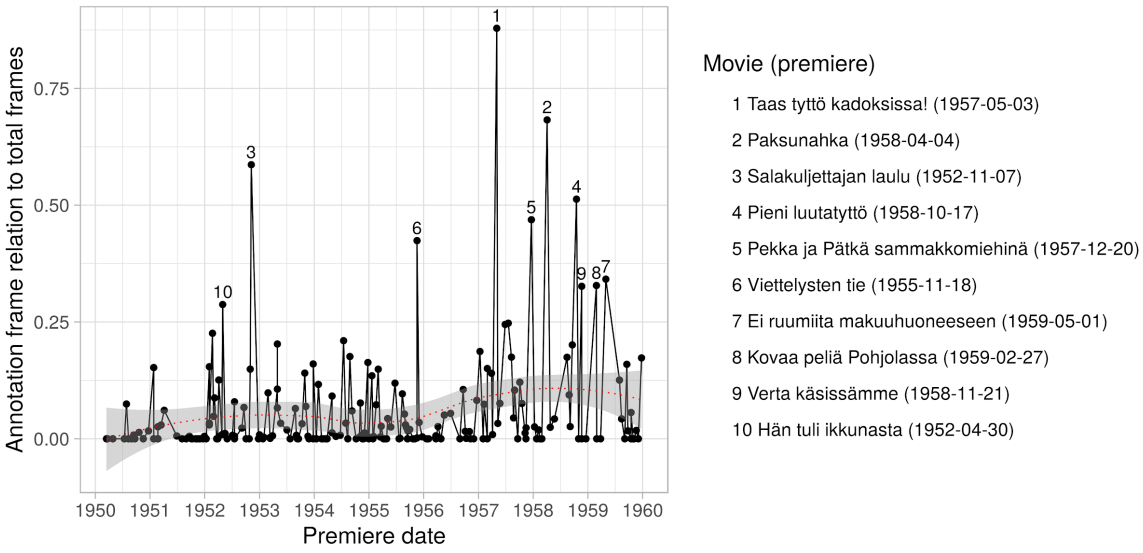


Figure 11: Car objects in films

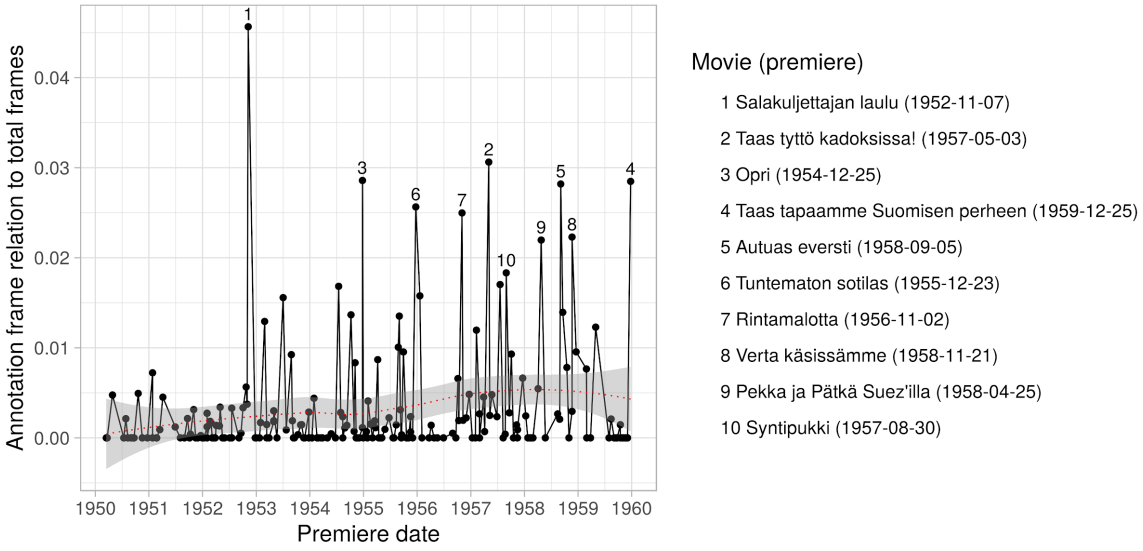


Figure 12: Car sounds in films.

Figures 11 and 12 show car objects and car sounds. Both figures confirm the rising trend toward the end of the decade. Also, the highest peaks, indicating films that foreground cars, are stressed more toward the end of the 1950s, as being a period of motorization.

5. From rural to urban

Our previous discussion on horses and cars serves the wider purpose of understanding the role of urban life in Finnish fiction films of the 1950s. The visibility and audibility of cars, and of motor traffic in general, can be argued to epitomize not only the pervasiveness of urban culture but also its increasing significance toward the end of the decade. During the next decade, this would become even more obvious: in the films of the 1960s, the presence of cars is inescapable, especially since the number of historical films decreased radically in this decade.

In our study, we not only tried to detect horses and cars, which would have presented a rather narrow view on the issue, but we also detected other objects and categories. The aural analysis offered findings

that can be used to reflect on how rural or urban the soundscapes of Finnish films have been, at least in the realm of fiction. A critical comment must be made, however: the filmmakers did not always use the soundtrack to characterize the milieu of the scene if the impact on the spectator could be effectively achieved by visual means. Sometimes, the musical accompaniment captured the main intent but eclipsed the diegetic sound. In a film set in the countryside, for example, nature might be represented through musical instruments rather than through onsite recordings or real-life stock sounds. Accordingly, lively music might sometimes accompany a car-driving scene in a city film, without any engine sounds. A practical reason for this was that exterior scenes were often shot with a silent camera so that no sounds were recorded on location. Therefore, post-recorded music would dominate the soundtrack. The perceived importance of sound design in creating and characterizing cinematic story space has arguably increased since the period discussed here, and the possibilities for mixing various sounds and sound effects have developed considerably.

However, it can be argued that it is meaningful to analyze the nature of different audio events and to see which kinds of sound were offered to the ear of the moviegoer. We concentrated on modern sounds, such as auditive references to transportation, telephone, radio, and television, which can all be interpreted as signs of an urban culture. Radio, of course, was spread throughout the country, but it was part of a technological process that emphasized the role of urban centers in communication. We also identified sounds that could refer to rural life, such as animal sounds, like a pig or a cow, and sounds associated with waterways, such as rowboats. These findings are presented in Figure 4. It becomes evident that the presence of modern and urban sounds increased in the films of the 1950s. There were fewer rural sounds during that period, but, interestingly, their presence increased over the course of the decade. This possibly indicates the fact that there was a shift from optical to magnetic sound toward the end of the decade, upgrading the quality of sound mixing and giving emphasis to a more nuanced sonority, in general.

In addition to sounds, we also drew on ASR to trace urban and rural keywords from the spoken language in the films of the 1950s. Figure 5 shows the general trends, both in the whole dataset and in the subset of non-historical films. It is evident that our keywords for rural, such as ‘navetta’ (cowhouse), ‘karja’ (cattle), ‘lehmä’ (cow), ‘kana’ (chicken), ‘kukko’ (rooster), and ‘maaseutu’ (countryside), were steadily present throughout the decade. Urban keywords, such as ‘auto’ (car), ‘lentokone’ (airplane), ‘imuri’ (vacuum cleaner), ‘puhelin’ (telephone), ‘radio’ and ‘kamera’ (camera), were also steadily there, but with greater frequency than the rural keywords. Neither group shows any particular rise or decline, but the signs of urban culture were detectable throughout the decade.

Statistically, most of the Finnish population lived in the countryside in the 1950s. The balance between rural and urban life changed fundamentally during the 1960s, but clear signs of urbanization could already be detected in the earlier decade. It is important to add that the relationship between the country and city was a theme in the movies themselves as well, for example, in the melodrama *Kukonlaulusta kukonlauluun* (SF, 1955), where the city was a symbol of decay and the sinful life. Such overtly anti-urban films were rare, however, and in most of the films, ‘rural’ and ‘urban’ were not presented as opposites [12]. Many films indicate the co-existence of both aspects and include movement between rural and urban surroundings. This is exemplified in Figure 13, which shows an image from the film *Vieras mies* (Su-Fi, 1957). Here, the object detection has found both a horse and a car in the same shot. Both exist in the same story world, with no discordance between them. A closer look at the image confirms this observation: in the background, we see not only a traditional farmhouse but also power poles running through the fields.

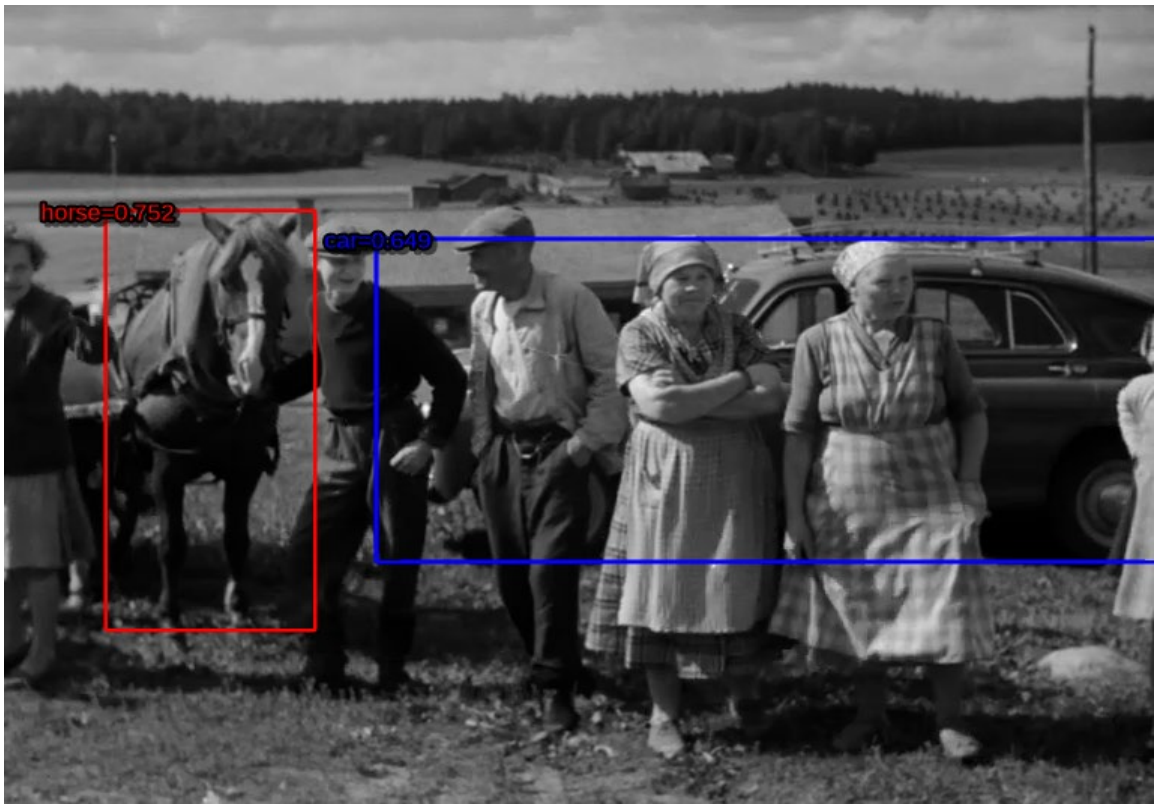


Figure 13: A rural setting in the film *Vieras mies* (Su-Fi, 1957), including both a horse and a car.

In Finland in the 1950s, there were many variations between the urban centers and in their relationship with the countryside. Many films, such as *Hilmanpäivät* (SF, 1954), were situated in smaller towns that still carried many characteristics of rural life but had also acquired some urban features. By representing a middle ground between the rural and urban milieus, such films challenged the assumed duality between them. Modern technology and modern ways of life were becoming visible outside of the bigger centers. Our findings support the idea that Finnish fiction films showed blurred and mixed environments. Rural and urban elements were simultaneously present. Finnish films had large audiences in the countryside. For a rural audience, these films portrayed urban life, but they also inserted modern technology into agrarian stories. For an urban audience, the films showed familiar city landscapes, often shot in Helsinki, but they also offered a route back to the countryside, from where many members of the audience had actually come.

6. Conclusion

The paper demonstrates that aural and visual object recognition, as well as speech recognition, can be applied successfully to historical film analysis. Obviously, fiction films were not direct reflections of social life; they were sometimes commentaries on contemporary issues, but they also include, for example, historical films that did not portray the period of their making, at least not directly. However, the overall results support the idea that Finnish filmmakers fueled the imagination regarding urban life in the 1950s, paving the way for modern technologies and gradually pushing signs of rural life aside. The results thus challenge the often-stated view that when modernization was sweeping through society in the latter half of the 1950s, Finnish cinema was not able to follow, and a rural mentality prevailed in Finnish films until the New Wave cinema of the 1960s [28, 29].

Methodologically, our experiment could be further enhanced to be more sensitive to the objects, sounds, and speech of the past. In the past, people spoke differently from the way they do today. Living conditions were different, as were the everyday objects that people used in their daily lives. Our analysis was based on fiction films that also drew on constructed sets, painted backdrops, artificial props, and stock sounds, and not necessarily on real-life objects. In the 1950s, practically all the interior scenes, and even

part of the exterior scenes, of Finnish fiction films were shot and sound recorded inside the studios in Helsinki. Still, the results can illuminate our understanding of cinema as a realm of imagination and as an interpretation of a bygone world.

The use of old film material reveals challenges for modern data resources. For this study, we drew on existing tools. It became obvious that there were many misclassifications, for example, in object detection, an oil lamp was identified as a bottle, obviously because an oil lamp is not a category for current object detection tools. We also drew on a relatively small quantity of data. For the purposes of this article, the limited number of films made it possible for the film historians of the research group to verify that the results were plausible within the context of 1950s cinema. More work can be done to estimate the reliability of the results so that these methods can later be applied to a broader dataset. Considering the huge volume of archival collections of audiovisual heritage, footage, and recordings, there is a need for historically sensitive tools. In the future, this is something that waits to be done as a collaborative effort that crosses disciplinary borders.

7. Acknowledgements

This work was supported by the research consortium *Movie Making Finland: Finnish fiction films as audiovisual big data, 1907–2017* (MoMaF), funded by the Academy of Finland (329266). Film data and metadata were provided by the National Audiovisual Institute in Finland and computational resources by CSC—IT Center for Science, Espoo, Finland.

8. References

[1] Taylor Arnold, Lauren Tilton, Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*, 34 (Supplement_1): 13–116, December, 2019.

[2] Franco Moretti, *Distant Reading*. London: Verso, 2013.

[3] Daniel Biltereyst, Richard Maltby & Philippe Meers (eds), *The Routledge Companion to New Cinema History*. New York: Routledge, 2019.

[4] Yuri Tsivian, Cinemetrics, Part of the Humanities' Cyberinfrastructure. In Michael Ross, Manfred Grauer & Bernd Freisleben (eds), *Digital Tools in Media Studies. Analysis and Research. An Overview*. Bielefeld: transcript, 2009, 93–100.

[5] Kari Uusitalo, *Suomen Hollywood on kuollut. Kotimaisen elokuvan ahdinkovuodet 1956–1963*. Helsinki: Suomen elokuvasäätiö, 1981.

[6] Matti Peltonen, Vesa Kurkela & Visa Heinonen (eds), *Arkinen kumous – suomalaisen 1960-luvun toinen kuva.* Helsinki: Suomalaisen Kirjallisuuden Seura, 2003.

[7] Jaana Laine, Susanna Fellman, Matti Hannikainen & Jari Ojala, *Vaurastumisen vuodet. Suomen taloushistoria teollistumisen jälkeen.* Helsinki: Gaudeamus, 2019.

[8] Tapio Bergholm, Suomen autoistumisen yhteiskuntahistoriaa. In Kalle Toiskallio (ed.), *Viettelyksen vaunu: Autoilukulttuurin muutos Suomessa*. Helsinki: Suomalaisen Kirjallisuuden Seura, 2001, 65–92.

[9] Jarmo Peltola, Maatalous-Suomen liikenne. In Jaakko Masonen & Mauno Hänninen (eds), *Pikeä, hikeä, autoja. Tiet, liikenne ja yhteiskunta 1945–2005*. Helsinki: Tielaitos & Painatuskeskus, 1995, 26–36.

[10] U. E. Moisala, *Auto Suomessa: Auton kaupan, käytön ja korjaamotoiminnan historia vuoteen 1983*. Helsinki: Autoalan keskusliitto ry & Autotuojat ry., 1983.

[11] Hannu Salmi, *Elokuva ja historia.* Helsinki: Painatuskeskus, 1993.

[12] Kimmo Laine K. & Silja Laine, Näkemyksiä 1920-luvun elokuvan maaseudusta, kaupungista ja

niiden välisistä suhteista. *Lähikuva* 21, 1 (2008), 8–25.

[13] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M.D. Plumbley, Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2020 Oct 19;28:2880-94.

[14] J.F. Gemmeke, D.P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) 2017 Mar 5 (pp. 776–780). IEEE.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely; *The Kaldi speech recognition toolkit*. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, 2011, iEEE Catalog No.: CFP11SRW-USB

[16] Anssi Moisio, Dejan Porjazovski, Aku Rouhe, Yaroslav Getman, Anja Virkkunen, Ragheb AlGhezi, Mietta Lennes, Tamás Grósz, Krister Lindén, Mikko Kurimo*, Lahjoita puhetta -- a large-scale corpus of spoken Finnish with some benchmarks*, arXiv preprint 2203.12906.

[17] S. Enarvi, *Modeling conversational Finnish for automatic speech recognition*. PhD thesis, Aalto University, 2018.

[18] Noora Kallioniemi, Harri Kiiskinen, Mikko Kuutti, Kimmo Laine, Tommi Römpötti & Hannu Salmi, *Suomalaisen näytelmäelokuvan materiaaliset ulottuvuudet 1907–2017: metatietoanalyys*i. Movie Making Finland: Finnish fiction films as audiovisual big data, 1907–2017 (MoMaF), 2021. https://doi.org/10.5281/zenodo.4925899

[19] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo & Ross Girshick, *Detectron2*, 2019. Available https://github.com/facebookresearch/detectron2.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick & Piotr Dollár, *Microsoft COCO: Common Objects in Context.* arXiv preprint 1405.0312, 2015.

[21] Shaoqing Ren, Kaiming He, Ross Girshick & Jian Sun, *Faster R-CNN: Towards real-time object detection with region proposal networks*. arXiv preprint 1506.01497, 2016.

[22] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado & J. Dean, Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013, 26.

[23] Jarmo Peltola, Liikenne maassamuuton Suomessa. In Jaakko Masonen & Mauno Hänninen (eds), *Pikeä, hikeä, autoja. Tiet, liikenne ja yhteiskunta 1945–2005*. Helsinki: Tielaitos & Painatuskeskus, 1995, 36–42.

[24] Autotuojat ry. (1956) *Autokannan kokonaislaskenta 31.12.1956*. Mobilian eKokoelmat [Electronic collection of The National Road Traffic Museum]. mobilia.mediafiles.fi/mobilia.fi.

[25] Autotuojat ry. (1959) *Autot merkittäin ja valmistusvuosittain 31.12.1959*. Mobilian eKokoelmat [Electronic collection of The National Road Traffic Museum]. mobilia.mediafiles.fi/mobilia.fi.

[26] Marko Nenonen, Tienrakennuksen ja työllisyyspolitiikan yhteiselo päättyy. In Jaakko Masonen & Mauno Hänninen (eds), *Pikeä, hikeä, autoja. Tiet, liikenne ja yhteiskunta 1945–2005*. Helsinki: Tielaitos & Painatuskeskus, 1995, 144–190.

[27] Tapani Mauranen, Satavuotias autoilija: Suomalaisen autonkäytön historiaa. In Kalle Toiskallio (ed.), *Viettelyksen vaunu: Autoilukulttuurin muutos Suomessa*. Helsinki: Suomalaisen Kirjallisuuden Seura, 2001, 33–63.

[28] Antti Alanen, Suomen modernisoituminen: Suomalainen yhteiskunta 1950-luvun lopussa ja 1960-luvun alussa. In Kari Uusitalo et al. (eds), *Suomen kansallisfilmografia 6*. Helsinki: Valtion Painatuskeskus & Suomen elokuva-arkisto, 1991, 19–23.

[29] Laura Kolbe, Suomalainen kaupunki. Historian hierarkioita ja modernia lähiöelämää. In Markku Löytönen & Laura Kolbe (eds), *Suomi. Maa, kansa, kulttuurit*. Helsinki: SKS, 1999, 156–170.