Deep Watching: Towards New Methods of Analyzing Visual Media in Cultural Studies

Bernhard Bermeitinger (bernhard.bermeitinger@unisg.ch), U St. Gallen, Switzerland and Sebastian Gassner (sebastian.gassner@uni-passau.de), U of Passau, Germany and Siegfried Handschuh (siegfried.handschuh@unisg.ch), U St. Gallen, Switzerland and Gernot Howanitz (gernot.howanitz@uni-passau.de), U of Passau, Germany and Erik Radisch (erik.radisch@uni-passau.de), U of Passau, Germany and Malte Rehbein (malte.rehbein@uni-passau.de), U of Passau, Germany
XML

A large number of digital humanities projects focuses on text. This medial limitation may be attributed to the abundance of well-established quantitative methods applicable to text. Cultural Studies, however, analyse cultural expressions in a broad sense, including different non-textual media, physical artefacts, and performative actions. It is, to a certain extent, possible to transcribe these multi-medial phenomena in textual form; however, this transcription is difficult to automate and some information may be lost. Thus, quantitative approaches which directly access media-specific information are a desideratum for Cultural Studies.

Visual media constitute a significant part of cultural production. In our paper, we propose Deep Watching as a way to analyze visual media (films, photographs, and video clips) using cutting-edge machine learning and computer vision algorithms. Unlike previous approaches, which were based on generic information such as frame differences (Howanitz 2015), color distribution (Burghardt/Wolff 2016) or used manual annotation altogether (Dunst/Hartel 2016), Deep Watching allows to automatically identify visual information (symbols, objects, persons, body language, visual configuration of the scene) in large image and video corpora. To a certain extent, Tilton and Arnold's Distant-Viewing Toolkit uses a comparable approach (Tilton/Arnold 2018). However, by means of our customized training of state-of-the-art convolutional neural networks for object detection and face recognition we can, in comparison to this toolkit, automatically extract more information about individual frames and their contexts.

1. Research object

The focus of our project is Ukrainian nationalist Stepan Bandera, who during World War II collaborated with Germany and tried to forcefully establish a Ukrainian national state against Polish and Russian opposition, and his instrumentalisation in the recent Ukraine conflict by both the Ukrainian and the Russian side. In the Russian narrative, Bandera is used as an example for Ukrainian fascism, whereas for Ukrainian nationalists he symbolises the uncompromising fight for national independence. New media such as video clips uploaded to YouTube are used extensively to disseminate these contradicting interpretations of Bandera; a first study showed that this instrumentalization is present in all major digital media and was already immanent before 2014 (Fredheim et al. 2014). Our paper builds on this preliminary work and traces Bandera's image and position within cultural memory in Poland, Ukraine, and Russia from the Euromaidan in 2013 up until now.

2. Methodology

We use the first 200 Youtube search results for the terms "Stepan Bandera" and "Степан Бандера" as a corpus. Because of overlapping search results, our corpus comprises 274 videos, uploaded to Youtube between 2007 and 2017 with a total length of 3 days, 11 hours, 49 minutes, and 16 seconds. It should be noted that YouTube's search engine does not provide direct access to its database, but rather adapts the result list according to country, browser, and other details of the user.

The 274 videos are split into their individual frames. In order to analyse the corpus, we trained Detectron, an open source framework developed by Facebook AI Research (Girshik et al. 2018), to recognize 12 emotionally charged symbols, which help identify in which context Bandera is presented and thus, hint at different instrumentalization. Table 1 presents these symbols within their four different main classes and the corresponding numbers of manual training annotations used.

Ukrainian nationalist symbols	*German fascist symbols*	*Polish nationalist symbols*	*Russian / Soviet nationalist symbols*
Ukrainian coat of arms (208)	SS-Rune (106)	Polish coat of arms (38)	Hammer & Sickle (111)
Logo of Swoboda (129)	Swastika (187)	Falanga (91)	Ribbon of St. George (147)
Ukrainian flag (198)	Wolfsangel (96)		
Flag of UPA (212)			
OUN Symbol (57)			

Table 1: List of all 12 symbols within their 4 distinct classes

To create the training set we manually annotated 793 images with 1731 annotations, i.e. an average of 144 annotated objects per symbol. An annotation consists of point coordinates indicating the outline of the object and the corresponding name of the symbol. Between 1 and 13 annotations are assigned to an image, on average 2.2; the median is 1. For proper testing and evaluation, the corpus is randomly divided into training, testing, and evaluation data using a ratio of 70/15/15.

We use Intersection over Union (IoU) as the evaluation metric for symbol recognition. IoU covers the interval 0 to 1, 1 being a perfect match between proposed and predefined region. In our experiment, we reach an average IoU of 0.68. On closer inspection, our results show that objects which have not been recognized in one frame are likely to be recognized in a subsequent one. Hence, the recognition of symbols is even better then the test result suggests. A sample visualization of recognized symbols in a single frame can be seen in Figure 2.

Figure 2: Recognized symbols in a single video frame, taken from https://www.youtube.com/watch?v=axFz-SU8cIM (accessed 27 November 2018).

Unfortunately, our instance Detectron is not optimized for individual face recognition; trying to recognize distinct persons (in our case Bandera and Adolf Hitler) in individual frames led to a high error rate within this class. Hence, we decided to combine Detectron with OpenFace (Amos/Ludwiczuk/Satyanarayanan 2016), an implementation of the FaceNet algorithm (Schroff/Kalenichenko/Philbin 2015). We are currently evaluating recognition accuracy in a test corpus and will present the combined results of Detectron and Facenet at the conference.

3. Results

As Figure 3 shows, symbols related to Poland (the Polish eagle and the Falanga) and Russia (Ribbon of St. George and Hammer & Sickle) are seldom encountered in our corpus, whereas the flags of Ukraine and UPA are rather common, as is the Ukrainian coat of arms. The Ukrainian flag, for example, shows up in 2% of all video frames in the corpus (i.e. for 1 hour and

40 minutes). Also common, albeit less frequently occurring than their Ukrainian counterparts, are the symbols of the Third Reich. This distribution suggests that Bandera is presented in a Ukrainian nationalist context and his connections to the Nazis is underlined, whereas his position in the Polish and the Soviet context does not play a big role. This interpretation becomes even more clear when symbol co-occurrences (i.e. symbols showing up in the same frame) are plotted (Figure 4). Both the Ukrainian and the Nazi symbols not only co-occur within their group but also with the respective other groups. This finding hints at the dominance of a Russian nationalistic discourse on Youtube, which frames Bandera as an example of Ukrainian fascism.

Figure 3: Mean percentage of occurrence for each symbol

Figure 4: Symbol co-occurrences in 274 videos, adjusted for symbol frequency

Figure 5: Total symbol occurrences over time

The next step is to combine the detection results from Detectron with the appearances of Hitler and Bandera in our corpus. What is more, we plan to compare the results discussed above with a second video corpus about Bandera, which was collected in 2013 as part of previous work. This comparison may uncover how the Ukraine crisis changed the way Bandera is represented in Youtube videos and will be presented at the conference. A first glance at diachronic symbol occurrences is presented in figure 5; this visualisation suggests that specific symbolic discourses rise and fall in the course of time, and most symbols peak in 2014 when the conflict is in the most heated stage.

4. Discussion and outlook

Recognizing specific symbols allows for new ways to study large visual corpora. Nonetheless, this approach is tied to a specific research question because a RCNN has to be trained to recognize predefined symbols. This limiting factor led us to experiment with a more general approach which focuses on visual depictions of human bodies as embodied signs, a key question of Cultural Studies. We are currently evaluating additional algorithms to automatically recognize specific people and assess both their posture and mimic on a sample corpus of 1000 trading cards from the 1930s depicting German-American actress Marlene Dietrich.

Body postures can be analyzed on the basis of keypoints (Bourdev/Malik 2009) such as hands, feet, head, etc., which result in different postures. (Figure 6) These postures can be used to study symbolic meanings communicated through the body (Impett/Moretti 2017). The connection between postures and gender stereotypes is much discussed (Mühlen-Achs 1998); in the case of Marlene Dietrich's androgynous self-staging which relies on elements connoted as “male”, a quantitative analysis of postures and their changes over time allows new insights.

Figure 6: Posture detection with Detectron (left), Face analysis by OpenFace (right)

Generic information on Faces can be extracted by the algorithm OpenFace 2.0 (Baltrušaitis et al. 2018) which extracts three-dimensional orientation and keypoints such as eyes or nose from a given digital image. (Figure 6) Moreover, these keypoints are compared with a standard face model defined by the Facial Action Coding System (FACS, Ekman/Friesen 1978), which describes facial expressions. In the case of Marlene Dietrich, the expression outer brow raiser is often encountered, which can be explained by the makeup trends of the 1930s. Thus, our approach can to a certain extent assess fashion and style-related questions in a quantitative manner.

By means of combining various algorithms to automatically identify symbols, objects, faces, posture, and mimics, we propose a potent framework to study large corpora of visual media. We are convinced that Deep Watching will advance the quantitative methodology of Cultural (and Media) Studies significantly.