Proc. 2013 NEM Summit, Nantes, France, October 2013.

Think Before You Link — Meeting Content Constraints when Linking Television to the Web

Daniel Stein[1], Stefan Eickeler[1], Rolf Bardeli[1], Evlampios Apostolidis[2], Vasileios Mezaris[2], Meinard Müller[3]

[1]Fraunhofer Institute IAIS, Sankt Augustin, Germany [2]Information Technologies Institute CERTH, Thermi-Thessaloniki, Greece [3]International Audio Laboratories Erlangen, Germany

Abstract: **With a constantly rising demand for interactive videos, automatic enrichment of static videos with web links offers seemingly endless possibilities. Given the content that can be found on the web, however, this can be a rather mixed blessing. In this paper, we investigate web sites with contents of extreme physical violence, political extremism and self-harm and argue that their topic can be quite close to seed videos such as, e.g., news shows. We discuss ways to detect such problematic content and present first solutions to specific challenges.**

Keywords: LinkedTV, OCR, Cover Song Identification, Concept Detection

1 Introduction

Many recent projects focus on automatic enrichment of linear (i.e., static) videos with links to other text resources, images and videos. While offering seemingly endless possibilities, multimedia content that is automatically interlinked with various kinds of data can at best be partially checked for its new content. Thus, special care has to be taken that this data meets legal and moral constraints. This problem is not only inherit to video interlinking but also applies to other fields like, e.g., advertisement placement, which is why some research in this area already exists. However, the main focus of existing applications is the identification of pornographic content, a topic where simple keywords can already identify a large percentage of possible web sites.

The aim of this paper is two-fold. First, we want to give an overview of types of problematic web sites featuring other content than pornography, namely (a) depiction of extreme physical violence, (b) political extremism, and (c) self-harm, e.g., self-cutting and anorexia. Second, we identify challenges for classification algorithms arising from the nature of this material and conduct a series of proof-of-concept and large-scale experiments to evaluate their performance.

This paper is structured as follows. First, we describe the "Linking Television to the Web" (LinkedTV)[1] project and describe why each of the three topics inherit relevant arguments against unconstrained linking (Section 2). We proceed by reviewing related work with a focus on content constraint identification (Section 3). Then, we describe relevant content and indicate challenges for automatic classification techniques (Section 4). Then, we offer proof-of-concepts and thorough experiments on the following classification tasks that have been identified to be potential challenges: audio fingerprints on violent movies (Section 5), song identification on right-extremistic rock records (Section 6), colored text localization on self-harm photos (Section 7), and concept detection on self-injury images (Section 8). Finally, we conclude this paper with a summary (Section 9).

2 LinkedTV

The aim of LinkedTV is to provide an interactive multimedia service for non-professional end-users. To achieve this, linear videos are analyzed by various (semi-)automatic methods, both on the acoustic and the visual level. The raw data obtained is then used to interlink the videos with other multimedia information, using knowledge acquired from, e.g., web mining. The enriched videos are finally shown to the end-user as an interactive video with many links to further web content. Especially for news, which forms the basis of one of LinkedTV's scenarios, a fully automated process can lead to undesirable content, as the following examples from our seed data illustrate:

- A critical report about the former lawyer Horst Mahler, who has been imprisoned several times for right-wing utterances. The Wikipedia page of Horst Mahler links to uncommented speeches given by Mahler, the first 15 hits in one of the largest online video sites feature uncommented interviews where he states his beliefs – among them the denial of the Holocaust, a statement forbidden in many countries, including Germany.
- A report on anorexia nervosa, which follows the story of a young girl who almost died from undernourishment. In the interviews, she mentions several terms specific to the Pro-Ana movement, which glorifies emaciation. Searching these terms results in few educational web sites but for the most part Pro-Ana blogs and web sites which are full of "trigger" images, i.e., depicting extremely anorectic women (in the Pro-Ana movement, these are often called "thinspos", a portmanteau of "thin" and "inspiration").
- After incidents where a man was beaten to death close to a Berlin railway station, a state secretary discusses with the moderator whether there is a trend towards brutalization in our society, the role of computer games, and the phenomenon of "happy slapping" (where somebody is battered, sometimes even worse, in front of a camera). Again, these search terms in an online video portal produce only few educational and critical reports (and never as first results) but mostly actual instances of battering or extremely violent video games which have an age limit "R"[2] or higher.

3 Related Work

While there is a vast scientific community focused on knowledge extraction from multimedia objects, little scientific attention is spent towards recognizing inappropriate material on the web, especially for non-textual data.

[1] http://www.linkedtv.eu

[2] cf. www.filmratings.com

There has been specific interest in detecting pornography (e.g., [4, 9, 11]), mainly for filtering search engine results. A special case is given by child pornography, where there is some effort to develop automatic means for detecting and thus support the fight against this crime [1]. Apart from that, the main investigations into detecting material problematic content have been seen in the context of the TRECVID and MediaEval evaluation campaigns, namely the category *physical violence* in TRECVID's feature extraction challenge [17] and the Violent Scenes Detection Task in MediaEval [6]. Beyond these specific domain, there is of course a lot of work in the multimedia analysis community giving a basis for supporting it.

4 Problematic content: overview and possible techniques

It is impossible to generalize over all relevant content to be found on such heterogeneous topics as violence, extremism and self-harm. In order to identify promising analysis techniques, we have asked a German bureau responsible for child protection in the Internet to compile lists on each of them, with special care for representative samples.[3] As base data, we manually scanned 3 000 web sites (1 000 per topic) with material not suitable for children of age 17 or below. Some web pages were extremely disturbing even for adults, e.g., featuring real suicides or decapitation, which required security measures such as psychological supervision and restricted data access when handling the data.

In this section, we look for common patterns that can be captured with classification algorithms with the goal of raising warning flags to an editor who manually checks outgoing links for their appropriateness. We make two assumptions: first, the interlinking is targeted for multimedia objects, i.e., audio, image or video, and second, we have no prior knowledge of its content via the surrounding web site or incoming links.

4.1 Physical Violence

Web sites with extreme violence typically feature videos or images. The two largest categories are filmed/fotographed content and violent computer games.

Filmed/Fotographed Content. Web sites containing violence images and videos are either (a) commercial web sites of horror movies, (b) shock sites that have the sole purpose of disturbing the viewer or (c) sites that want to provoke emotional reaction for political purposes (e.g., news sites against torture, or anti-abortion organizations). The nature of the violence depicted is very heterogeneous and ranges from small quarries to decapitation or other forms of murder (both authentic and fictitious). Fictitious movies are mostly from the genre of horror movies, which range from monster hunts to sexually-sadistic revenge movies. "Best-of" compilations, trailers, single scenes or self-made commentaries are frequent. These often share a logo/url/watermark of a dedicated collection forum/blog.

Violence Computer Games. All surveyed material containing violence games are shooters, with either first-person view or third-person view over the shoulder. For first-person games, the lower part of the screen typically shows the selected weapon, either military (e.g., machine gun), futuristic (e.g., laser pistol) or rough-and-ready (e.g. a simple stick). Some game videos are official trailer, but most are self-made fan videos ("let's play").

Conclusion. Among the three topics violence, extremism and self-harm, violence seems to be the topic with the highest heterogeneity of its material, which poses a huge challenge for analysis techniques. Extremely brutal videos such as decapitation are not as frequent but produce a huge number of re-mixes and samples. In our first approach trying to address the problem of violence detection in videos, we tested an algorithm that performs audio fingerprinting on a horror movie in Section 5.

4.2 Extremism

While there are many forms of political and religious extremism, the sites given to us were mostly from right-wing extremism, which is why we will focus on this content in the following – by no means a belittlement of other forms of dangerous extremism. Note that some videos from religious extremism for example were rather aligned to the category "violence" because they featured decapitation.

The majority of the web content in right-wing web sites falls into the categories shop sites, propaganda videos and music videos.

Shop Sites. The web pages in this category offer right-wing souvenir articles such as apparels and media articles. To attract attention, these sites feature a lot of obvious keywords and palpable images.

Propaganda Videos. The videos shown in the resort of propaganda were mostly focused on history revisionist point of views. The language was either in German or English, sometimes both languages were spoken simultaneously. The quality of the videos is, especially for older looking videos presumably digitized from old video cassettes or for war footage, rather mixed. Filming at historic sites, e.g., concentration camps in Germany, appear to have taken place without proper filming equipment, and the speaking person has no dedicated microphone. The name of the people speaking is sometimes shown via banners. Sometimes, extremism symbols like the German Swastika are shown.

Music Videos. The extremism music videos that we watched were seldom professional videos like those shown in television, but mostly just the audio plus some images of the CD cover or some band photography. A notable exception were live videos from concerts. Most images contained extremism symbols, such as the triskele, the swastika, or logos of organizations or groups associated with extremism. While some of these objects are depicted on banners, others are sprayed in graffiti, painted on tissue or tattooed. The music itself can be practically any genre such as singer-songwriter or instrumental, but has a strong tendency towards hard rock or metal, with the interpreters mostly shout-singing.

Conclusion. Based on the impression of the 1,000 right-extremism web sites surveyed, the detection of this content seems to be the most easiest in direct comparison to violence and self-harm. Many individuals and groups in

[3] jugendschutz.net

this area use re-occurring keywords, named entities and symbols. For music identification, however, fingerprinting alone will only cover a fraction of the material, since fan-based covers and samplers appear frequent (cf. Section 6).

4.3 Self-Harm

The given web pages about self-harm included the topics self-injury, eating disorders, substance abuse and suicide. The nature of the web pages ranged from distress calls to tutorials (sometimes disguised as preventional education). Substance abuse was a very broad topic and contained virtually no clues for analysis techniques, while the topic suicide only had a few examples, as most content such as forum discussions is supposedly well-hid in private areas of the web pages (which, of course, holds true for the other topics as well, but to a lesser extend). In the following we will investigate the topics Pro-Ana and Pro-Mia as well as self-injury.

Pro-Ana/Pro-Mia. In web sites that glorify eating disorders such as anorexia and bulimia, there are often catch phrases (e.g., "angels have no hunger") or longer text collections (e.g., "Ana's letter", a fictional conversation of a young girl with an incarnate anorexia). While as text, they can be easily spotted, in videos they are often depicted in heavily stylized fonts. Moreover, the letters are often hard to read even for a human eye since often colored letters are put on top of a photo.

According to a survey in [3], 85% of Pro-Ana web sites contain thinspo images. We thus manually analyzed images with thinspos made available on a freely accessible server. In total, the server contained 84 163 images (5.3 GB). Based on this manual survey, we draw the following conclusions: the server mostly contains images that are in relation with fashion shows and catwalk models. Roughly as frequent are images from celebrities. The third largest collection of images features thin women at the beach, unknown people as well as celebrities. Most people depicted seem to have pathological body-mass-indexes, but the images without their context are not obvious eating disorder glorifications.

A second, albeit far smaller group of images, contains material that clearly is related to anorectic context: (a) images that show extreme signs of anorexia, with bones clearly visible all over the body, (b) before-after compilations, both fake and real, of people that starved, and (c) depictions of extremely adipose people ("anti-thinspos"), often accompanied by sarcastic text included in the image.

Self-Injury. Web sites which aim at attracting attention to persons who regularly inflict damages to themselves are, in principle, similar to Pro-Ana/Pro-Mia sites in the sense that they contain many pictures and catch-phrases on the topic of self-injury. Often, the difference to eating disorder websites include: (a) self-shots, which do not appear as much in Pro-Ana/Pro-Mia sites where disdain of the own body is often part of the disorder, (b) more general cries for help ("Why won't anybody listen?") that do not exclusively circle around the disorder itself. However, the images are mostly self-explanatory, making a specific search a helpful feature in the identification of this topic.

Conclusion. The web sites on self-harm are centered around few topics, but the high amount of user generated content, often from under-age persons, creates a lot of individualized content where fingerprinting is of little help. In the following, we will take a closer look at two challenges: colored fonts on colored ground (Section 7) and high-level concept detection on a proportion of the imagery (Section 8).

5 Audio Fingerprints on Violence

Given a small audio fragment as query, the task of audio identification consists in identifying the particular audio recording that is the source of the fragment. In this section, we investigate whether audio fingerprints can be employed on horror movies, where spoken dialogue is only a fraction of the audio signal and noisy fight scenes are more common. For a proof-of-concept, we use the horror movie "Braindead" (1992, director: Peter Jackson) as seed video. The story follows a suburban town drifting into annihilation by some mutant virus which turns the infected into blood-thirsty monsters. The un-cut version is confiscated by law in Germany, a cut version (by 16 minutes) exists which has an age restriction of "16".

Given the un-cut version of this movie as reference and a possibly modified version of the same movie (e.g., cut version, samples only collecting the most violent scenes, different language version, ...), we want to find out which portion of the un-cut version these scenes relate to, only based on the audio signal. We employ an audio fingerprint method as described in [2].

On the whole movie, we created a fingerprint every ten seconds, resulting in 589 fingerprints total, of which 13 are discarded since they contained no audio. Of the remaining 576 fingerprints, 78.4% are detected in the German cut version, while the remaining sections most probably have been cut due to their brutal nature. Figure 1 illustrates the results: The minutes marked yellow above the top bar indicate gory scenes that are cited in the German confiscation enactment. Red sections are scenes where no fingerprints could be found in the cut version, i.e., these are deleted with respect to the un-cut version because of their violent content. All analyzed video snippets from an online video portal feature scenes of the un-cut version.

For this example, the algorithm also works on versions in other language (possibly due to the high number of fighting scenes where no-one is speaking and the audio signal remains unchanged). The English version featured 18.4% fingerprint matches, and the Spanish version even 22.6% matches.

6 Identification on Extremistic Songs

In this section, we look at fingerprinting performance and cover song identification of right-wing music recordings, where a large proportion is executed under bad recording conditions (bootlegs are quite frequent) and quite a few singers are not very melodic. On a data set of 523 albums and 6,631 records with a total length of $\approx$ 400 hours material, we used Echoprint [7] for a indexing and retrieval experiment. Of 6,000 songs, we indexed 3,000 of them in a Echoprint database and checked whether they were found correctly and whether the remaining 3,000 were correctly marked as unknowns. We thus obtained 86% true positives compared to 90% true negatives. These numbers are misleading, however, meaning that the performance is actually much higher, as the albums contain duplicates (one third

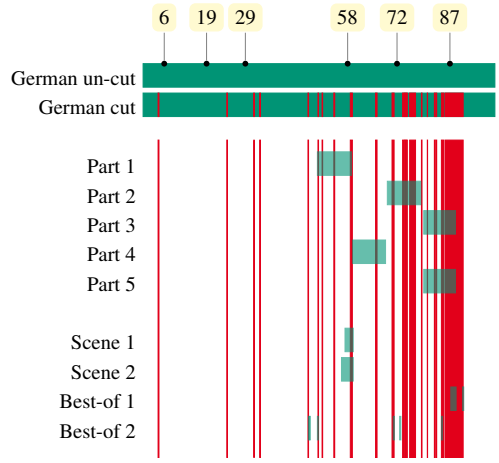


Figure 1: Fingerprints taken from the un-cut German version of "Braindead" that are identified in video snippets found in an online video portal.

of the false positives feature the words "tribute" and "sampler" in either song title or album title, which means that they should be evaluated as true positives since the song is probably indeed already known).

Another peculiarity of this data collection was the high amount of cover songs, especially for notorious right-wing bands like "Skrewdriver". Since fingerprinting will reject cover songs as completely different music, we proceed The goal of cover song identification is to identify different versions of the same piece of music within a database (opposed to an audio recording of a specific version as in audio identification). In the cover song scenario, one has to deal with changes in instrumentation, tempo, and tonality, as well as with more extreme variations concerning the musical structure, key, or melody [10]. This requires document-level similarity measures to globally compare entire documents.

The overall procedure crucially depends on the used feature variant, the type of score matrix, and a number of other parameters. In our implementation, we use a chroma variant supplied by [13] and apply various enhancement strategies to improve structural properties of the score matrices, see also [10, 12, 15].

For a proof-of-concept, we used the album version of the song "Hail the new dawn" from "Skrewdriver", plus seven covers, where two cover songs are from the same band and five songs are from different bands. Then, we compiled data sets containing other songs plus their cover versions, containing (a) 112, (b) 340 and (c) roughly 2,000 songs and measured for each Skrewdriver song how close it was assigned to the songs of its own group. For each query, the result is ranked with respect to decreasing similarity (see Table 1).

Overall, one may say that current systems for cover song identification yield reasonable results as long as the versions to be detected roughly follow the harmonic progression of the original song—at least in passages that have a duration of at least 40 to 60 seconds. This is, for example, clearly the case for the "Skrewdriver" cover group. When there is no dominating harmonic content in the song and if there are lots of variations in the melody (as may be the case for punk music or hard rock), however, the usage of chroma-based audio features becomes problematic and identification systems are likely to fail.

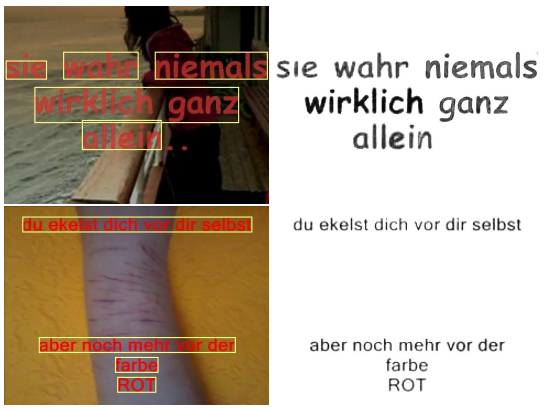


Figure 2: Text localization and text extraction from screen-shots of self-cutting videos

7 Colored Text Localization

OCR algorithms typically require that the text portion within a video is automatically detected and separated from the background. We noted that in self-harm videos the text is often highly stylized and colored, and also placed on top of colored background pictures. State of the art text detection methods [5, 8] use the gray-scale image and cannot cope with this situation. In this section, we introduce a new text detection algorithm.

The text detection is based on color segmentation using the statistical region merging (SRM) [14]. The algorithm determines uniform colored connected components (CC), which represent the characters and other objects/-parts in the image. In a subsequent step the extension of SMR for handling occlusions is used to merge the characters to words. Additionally to the color information of the original SRM, the difference of the height of the CC and the distance between the CC is used.

In order to refine the text separation for the OCR, a Gaussian model is calculated for the CC and for the background of the words. An up-scaled version of the image is used to create an up-scale gray-scale image. Each pixel is assigned a gray-scale value which is proportional to the probability of the text model given the RGB value. The last step is to use Tesseract for the recognition based on the gray-scale image and the information about the found text regions.

Figure 2 shows two example screen-shots where conventional text localization and text separation failed for conventional models, and where our method produces the correct results.

8 Concept Detection on Self-Injury

In this section, we want to investigate whether a visual concept detection algorithm can identify images of self-injury. On a non-protected server with the topic self-injury we had access to a collection of 5 383 images (625 MB), which mainly features self-cutting with sharp items, e.g., razors or scissors. The cuts appear all over the body, but mainly on the lower arm. The majority shows bleeding wounds with or without visible fat tissue, others are in the state of healing and thus show coagulated blood tissue or scars. Other images contained self amputation or blunt injuries.

As a comparison class, we reduced these images to cuts on the forearm, and used material from a web server on fore arm tattoos as counter class. See Figure 3 for some

(a) *Average precision* (AP, s. [15] for definition) values for the 8 queries and the three datasets

Idx	Band	Remark	AP_{112}	AP_{340}	AP_{2000}
1	Skrewdriver	original	0.982	0.924	0.883
2	Skrewdriver	demo tape	0.948	0.929	0.878
3	Skrewdriver	live version	0.891	0.875	0.860
4	English band	very “bawly”	0.916	0.849	0.777
5	American band		1	1	1
6	Swedish band	female singer	1	1	1
7	Swedish band	melodic female singer	0.831	0.789	0.620
8	British band	quite close to original	1	1	1

(b) Ranking matrix for each song, showing pairwise document-level similarities between the query and all documents contained in the dataset 112

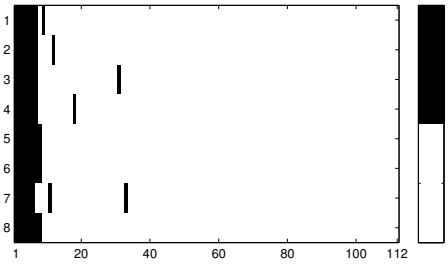


Table 1: Cover song identification on eight selected cover songs for the band Skrewdriver

examples from the set and the similarity of the classes. In order to increase the challenge for the classification algorithm, we further discarded tattoos in any other color than black or red. In total, 400 images for each class remained for training, and we used 67 (cutting) and 86 (tattoo) images for testing.

For classification, we make use of high-level concept detection algorithms, where we follow the approach of [16] with a large sub-set of the base detectors described there. The results for tattoo images are shown in Figure 4(b), die results for self-cuts are shown in Figure 4(a). With an assumed threshold of 0.5, 8 out of 151 images are misclassified.

9 Conclusion

In this paper, we elaborated on the needs to identify problematic web content in applications that rely on automatic interlinking such as the ones foreseen in the LinkedTV project. We gave a brief overview of possible contents in the areas of physical violence, extremism and self-harm, by manually scanning through thousands of web sites deemed inappropriate for children and adolescents. On the four selected challenges — (a) audio fingerprints in a horror movie, (b) right-wing cover song identification, (c) colored font OCR and (d) self-cutting concept detection — we presented our approaches and gave both proof-of-concepts and evaluations regarding their performance.

Overall, whenever interlinked content is not constrained via white-lists, we believe that outgoing links should be checked by human editors whenever the seed videos feature vast topics such as news, and that automatic analysis techniques are capable of producing warning flags which will support this task to a large degree.

Acknowledgments

This work has been partly funded by the European Community’s Seventh Framework Programme (FP7-ICT) under grant agreement n° 287911 LinkedTV.

References

[1] Aldhous, P. (2011). The digital search for victims of child pornography. *New Scientist*, 210(2807):23–24.

[2] Bardeli, R., Schwenninger, J., and Stein, D. (2012). Audio fingerprinting for media synchronisation and duplicate detection. In *Proc. MediaSync*, pages 1–4, Berlin, Germany.

[3] Borzekowski, D., Schenk, S., Wilson, J., and Peebles, R. (2012). e-Ana and e-Mia: A Content Analysis of Pro-Eating Disorder Web Sites. *American Journal of Public Health*, 100 (8):1526–1534.

[4] Bosson, A., Cawley, G. C., Chan, Y., and Harvey, R. (2002). Non-retrieval: Blocking pornographic images. In Lew, M. S., Sebe, N., and Eakins, J. P., editors, *Image and Video Retrieval*, volume 2383 of *Lecture Notes in Computer Science*, pages 50–60. Springer Berlin Heidelberg.

[5] Chen, H., Tsai, S. S., Schroth, G., Chen, D. M., Grzeszczuk, R., and Girod, B. (2011). Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *2011 IEEE International Conference on Image Processing*, Brussels.

[6] Demarty, C. H., Penet, C., Gravier, G., and Soleymani, M. (2012). The MediaEval 2012 Affect Task : Violent Scenes Detection. In *MediaEval 2012 Workshop*, volume 927.

[7] Ellis, D. P., Whitman, B., and Porter, A. (2011). Echoprint: An open music identification service. In *Proc. ISMIR*.

[8] Epshtein, B., Ofek, E., and Wexler, Y. (2010). Detecting text in natural scenes with stroke width transform. In *CVPR*, pages 2963–2970. IEEE.

[9] Fleck, M. M., Forsyth, D. A., and Bregler, C. (1996). Finding naked people. In Buxton, B. F. and Cipolla, R., editors, *ECCV (2)*, volume 1065 of *Lecture Notes in Computer Science*, pages 593–602. Springer.

[10] Grosche, P., Müller, M., and Serrà, J. (2012). Audio content-based music retrieval. In *Multimodal Music Processing*, volume 3, pages 157–174. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany.

[11] Kim, M. J. and Kim, H. (2012). Audio-based objectionable content detection using discriminative transforms of time-frequency dynamics. *IEEE Transactions on Multimedia*, 14(5):1390–1400.

[12] Müller, M. and Clausen, M. (2007). Transposition-invariant self-similarity matrices. In *Proc. ISMIR*, pages 47–50, Vienna, Austria.

[13] Müller, M. and Ewert, S. (2011). Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proc. ISMIR*, pages 215–220, Miami, FL, USA.

[14] Nock, R. and Nielsen, F. (2004). Statistical region merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(11):1452–1458.

[15] Serrà, J., Serra, X., and Andrzejak, R. G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017.

[16] Sidiropoulos, P., Mezaris, V., and Kompatsiaris, I. (2013). Enhancing video concept detection with the use of tomographs. In *Proc. ICIP*.

[17] Smeaton, A. F., Over, P., and Kraaij, W. (2009). High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin.

(a) Straight scars

(b) Straight patterns in a tattoo

(c) Patterns from wounds

(d) Patterns in a complex tattoo

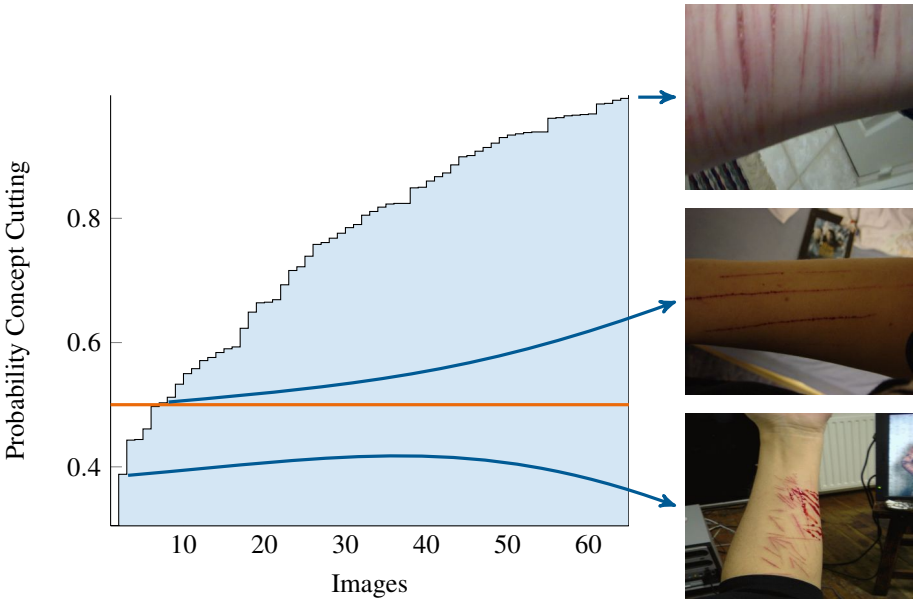
(e) Cut symbols

(f) Tattooed symbols

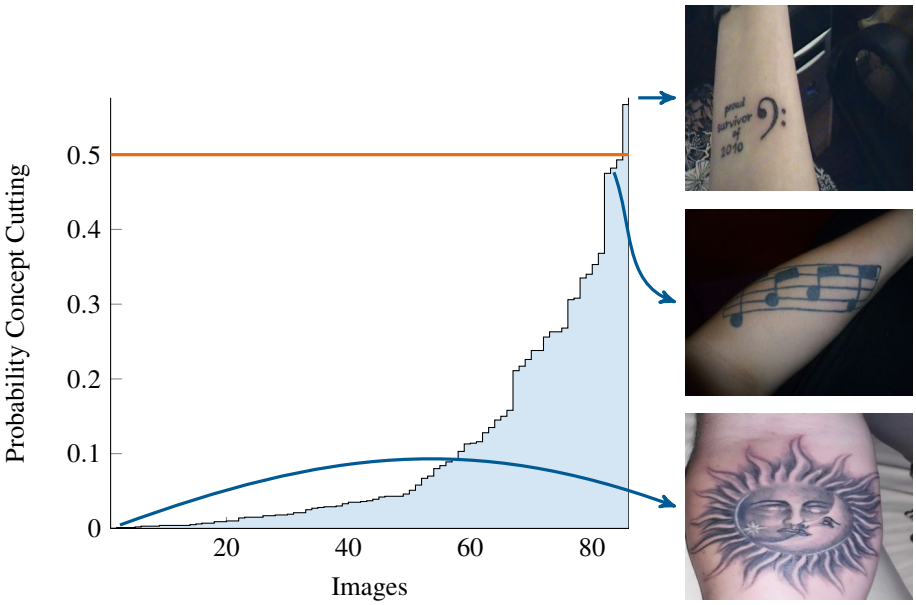
(g) Cut font

(h) Tattooed font

Figure 3: Comparison of pictures from the category self-injury and tattoo, both on the forearm



(a) Results of concept detection “Self-cuts”, for actual cuttings on the lower arm



(b) Results of concept detection “Self-cuts”, for tattoos on the lower arm

Figure 4: Overview for the probabilities of the concept detection “cutting vs. tattoo”, with an assumed threshold of 50%.