

Sep 8th, 2009

MULTIMODAL CONTENT ANALYSIS

IN SUPPORT OF

NEURAL-NETWORK BASED DETECTION AND RATING OF OFFENSIVE MEDIA

(PATENT PENDING)

White Paper

Abstract

vRate's approach to offensive content detection and rating is driven by a neural network trained on human assessments and feedback from a multimodal feature extraction engine. Businesses with existing feature extraction methods and technologies can take full advantage of vRate's open architecture by leveraging a repository of ratings by human raters and complementary nudity and profanity detection methods. This paper describes the vRateLite© implementation of the proposed approach.

vRate

This is a preliminary document and may be changed substantially prior to final commercial release of the software described herein. The information contained in this document represents the current view of vRate Corporation on the issues discussed as of the date of publication. Because vRate must respond to changing market conditions, it should not be interpreted to be a commitment on the part of VRaters, and VRaters cannot guarantee the accuracy of any information presented after the date of publication.

This white paper is for informational purposes only. vRate MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS DOCUMENT.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in, or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of vRate Corporation.

vRate may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from vRate, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

© 2009 vRate Corporation. All rights reserved.

Contents

1. Introduction	v
2. Problem Statement	v
3. Prior Art.....	v
4. Approach	vi
Multimodal Content Analysis.....	vii
Synthesis	vii
Optimized Nudity Detection	viii
5. System Architecture	ix
6. vRateLite© Implementation.....	xi
Scope	xi
vRateLite© at Work	xi
7. Summary.....	xii
References	xii
Appendix A	xiii
vRate_Nudity_Detection_Lite Probability Scale.....	xiii
Appendix B.....	xiv
vRateLite: Sample Report	xiv
vRateLite: Skin Blob Detection Results	xv
Appendix C.....	xvi
ICRA Rating Scale	xvi

1. Introduction

In the transparent culture built around social networking sites, media sharing and hosting tools, disturbing new trends involving, often unintentional, spread of private and X-Rated elements occur as a result of advances in technology. vRate is a rating service which employs a multimodal feature extraction engine, a repository of human rater assessments and a neural network mining model to detect and rate offensive elements embedded in media. vRate's approach incorporates multiple rating strategies to avoid the shortcomings of reliance on a single methodology and to improve the accuracy and effectiveness of salient feature detection. This paper describes our basic approach and a vRateLite© implementation illustrating the core vRate functionality.

2. Problem Statement

New reports are increasingly documenting legal repercussions after indecent media appear online. While there are many unanswered questions about the legal implications of sharing private media including prosecution for obscenity or child pornography, roughly 20 percent of the teens admit to participating in "sexting", according to a nationwide survey [Sex and Tech] by the National Campaign to Support Teen and Unplanned Pregnancy.

Similarly, hosting sites such as YouTube made it possible for anyone to post a video that could reach millions of viewers within minutes. According to data published by comScore*, YouTube as the dominant provider of online video in United States with a market share of around 43 percent and more than six billion videos viewed in January 2009. It is estimated that 20 hours of new videos are uploaded to the site every minute, and that around three quarters of the material comes from outside of the United States.

Considering the heavy traffic, it is inevitable for video hosting sites like YouTube to increasingly face criticism over the offensive content in some of its videos. Although most Terms of Service agreements forbid the uploading of material likely to be considered inappropriate, the inability to check all the videos and flag threats in a timely manner before they get published leads to occasional lapses.

There have been several debates calling online content providers to proactively review & police for offensive elements before making content available to the public. Large online content providers including Yahoo!, Facebook, YouTube & others rely on their users to flag the content as inappropriate. A member of staff then manually examines a flagged video to determine whether it violates the site's terms of service. Today's online media rating strategy is best characterized as reactive not proactive.

3. Prior Art

In the field of Multimodal Analysis there has been a plethora of research work focusing on the associations between low-level information extracted from the media (i.e. salient features) and the observer's need to interact with it semantically. Most ideas attempt to establish a mapping from low-level features to high-level semantic concepts e.g. rating assessment; often leading to complex,

* http://www.comscore.com/Press_Events/Press_Releases/2009/3/YouTube_Surpasses_100_Million_US_Viewers

unstable computation and poor performance. Such associations tend to be effective especially when they target a specific application domain [Calic et.al.]. Nevertheless, majority of the academic research employ multimodal processing for the purposes of generic content-based indexing, searching and retrieval of videos ([Informedia],[TRECVID-2003], [Li et.al.]).

When the context is narrowed down to detection and analysis of specific aspects of the media (e.g. nudity, profanity, violence), there are fewer algorithms, tools and services implemented as part of academic research and more for commercially available software with varying degrees of success and efficiency. Commercial examples in this application context include TRYNT Heavy Technologies API [TRYNT] which provides a free web service where in one can pass a URI of the picture to be rated and receive a % probability of nudity in the picture. There are standalone services one by Yang's Scientific Research Institute [YangSky] and other by Internet Safety Software [InternetSafety]. The former is a standalone tool that analyzes a video stream and/or a image for potential nudity by pixel comparing image/video to be rated against a library of images that are deemed potentially offensive, while the latter relies online users to voluntarily flag videos as unsuitable for under-aged viewers (under 18), which are then used by the tool to block access to a user based on the profile set by a parent or teacher.

It could be said that these methods or strategies are not conclusive or effective standalone – their success rate is particularly low for ambiguous cases where the offensiveness is implied, or undetectable due to poor media quality, or requires substantial contextual information to be derived. Most of the commercial and academic tools available today, rely on algorithms specializing on a single aspect of the rating exercise yielding false positives under several scenarios.

Hence the tasks involving detection and flagging of offensive attributes predominantly rely on human decision makers who are capable of combining many different aspects of the media and the contextual information to rate. For human raters, the degree of offensiveness varies considerably on certain aspects (e.g. violence, sexually explicit content in a video clip) due to the subjective nature of such rating exercise and the level of sensitivity of the human decision maker based on culture, age, gender and personal experiences.

Our research indicates that there is no apparatus marrying a human rater's ability to cope with ambiguities inherent in the media content, and the tractability, precision and impartiality of feature extraction methods to be able to arrive at a rating decision.

4. Approach

To address the challenges and shortcomings identified in the previous sections, vRate orchestrates concurrent rating strategies in the detection of offensive content:

- Multiple feature extraction methods which specialize in detecting salient features of the media in question (e.g. image detection to seek nudity in frames, speech recognition for capturing profanity in the audio track, voice recognition for content involving minors).

- Statistical data from human raters in the form of media rating repository through surveys or user feedback on rated samples. For example, vRate devises a media repository which incorporates samples with implied, ambiguous or straightforward depiction of offensive content, as well non-offensive media as control samples. vRate associates each sample with a computed average of a rating scale e.g. ICRA Rating System[†] for Nudity, Profanity, Violence, and Sex based on the statistics of human rater assessments.

The benefits of adopting these concurrent strategies are explored in the following sections.

Multimodal Content Analysis

vRate implements multimodal content analysis by incorporating numerous feature extraction methods:

- *Detecting the same aspect of the media in question.* For example, vRate allows for more than one nudity detection algorithms to run in parallel to assess the likelihood of nudity for an image. In addition to the computation of two potentially distinct likelihood percentages, methods can extract additional information e.g. existence of full frontal nudity, using skin exposure and template match indicating exposure of genitalia; both unique to the algorithm they implement. vRate collects and maintains all the algorithm specific or common findings.
- *Detecting different aspects of the media in question.* For example, vRate allows a feature extraction method employing speech recognition to detect profanity in the audio while other methods employ computer vision algorithms to detect nudity in the video stream. vRate collects and combines all these distinct aspects (i.e. high probability of profanity, nudity) to derive an assessment of the offensiveness inherent in the media.

The extraction results for these methods are brokered for improved accuracy in the detection of salient features and their associated probability.

Synthesis

vRate consolidates results from these distinct rating assessment strategies with the help of a neural network based mining model which is defined in terms of:

1. *Predictor input variables:* vRate applies the incorporated feature extraction methods to each media in the sample base to detect the potential offensiveness indicators (e.g. presence of nudity as assessed by various feature extraction methods, profanity, and relevant information available regarding the origin/host of the media). The application of these methods would yield a number of disparate results even when the methods work on the same indicators. For ambiguous cases where audio visual clues are missing and offensive content is implied (e.g. a shooting that is implied but not depicted explicitly) or where there are audio visual indicators yet the content is not offensive (e.g. exposure of naked breast as part of an educational breast examination video clip) the contextual information might prove helpful in making an improved assessment of

[†] <http://www.fosi.org/icra/>

offensiveness. vRate employs feature extraction methods that gather contextual information when possible by looking at:

- Available tags, flags, ratings explicitly associated with the source media (e.g. Rating header in MPEG4 files, YouTube videos with associated flags, tags, popularity, media file name)
- Available information concerning the host of the media (owner of the domain, estimated physical location for the host IP, host domain name)

2. *Predicted Input Variables*: Statistically computed averages of sample media ratings that are used in training the model to mimic human assessment for a candidate to be rated (e.g. Computed averages of ICRA's 1-to-5 rating scale for Nudity, Profanity, Sex and Violence[‡])

vRate is designed to reach a consensus on the scale of offensiveness for a candidate media file by relying on:

- An ever extending collection of feature extraction methods which work on detecting various aspects of the media with impartiality towards the subject matter
- Human rater statistics incorporating assessments on highly ambiguous cases characterized by implicit (or hard-to-detect) offensive content, and the need for complex synthesis of contextual information encapsulating the creation and delivery process of the media.

Consequently, vRate's approach is uniquely effective in addressing the rating problem as defined earlier and provides an advance over conventional systems and methods for rating offensive media.

Optimized Nudity Detection

One of the feature extraction methods implemented in vRate is a Nudity Detection algorithm designed to compute the probability of nudity in an image frame. Primary design goal for this algorithm is to speed up the real-time processing for vRate's Media Analysis Webservice queries, while maintaining a high accuracy rate, and minimizing false negatives. [Figure 1] places various algorithms for nudity detection on scale indicating speed, accuracy, scope and complexity. For scenarios where accuracy is of primary concern, template-matching is a robust alternative to other faster approaches such as simple skin detection.

vRateLite© implements an intermediate method [vRate_Nudity_Detection_Lite], for nudity detection. vRate_Nudity_Detection_Lite introduces simple heuristics to compute a probability[§] by aggregating skin percentage values for blobs detected in the foreground of the image frame. This eliminates requirements for lengthy comparisons between the target image and a library of templates with favorable processing speeds. Next generation of vRate Nudity Detection algorithms incorporates template-matching to handle frames with high risk of skin exposure as well as grayscale or monochromatic images.

[‡] Please refer to Appendix C for ICRA Rating Scale

[§] Please refer to Appendix A for how to interpret the computed probability

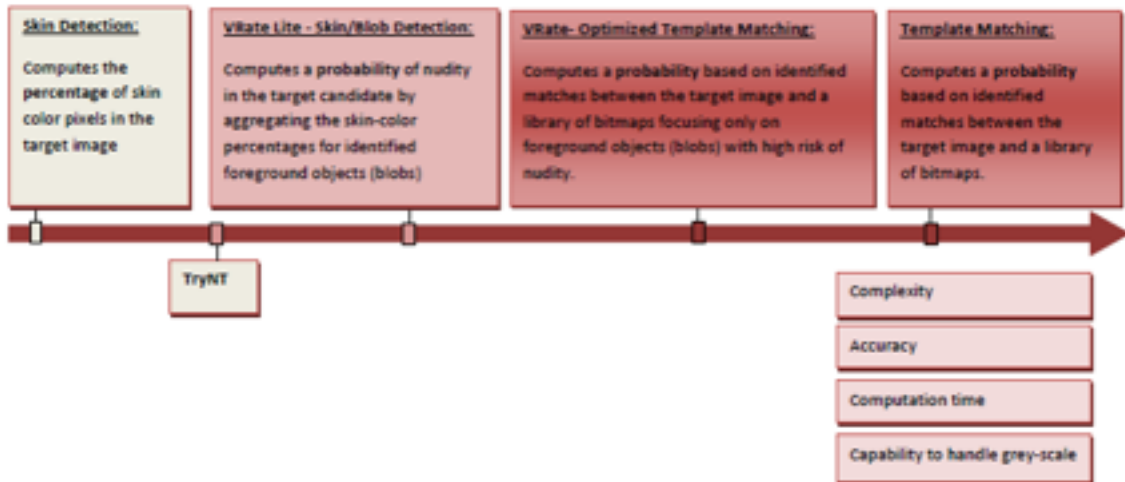


Figure 1: vRate Nudity Detection Algorithm

5. System Architecture

vRate system architecture enables the proposed approach by clustering distributed services around the dual rating strategies i.e. Neural Network Broker [I] and Feature Extraction Service and Engine [E], [F] and [G] and by providing infrastructure in support of Media Analysis Webservice query interface [C], media processing utilities [D], and data persistence and management [H].

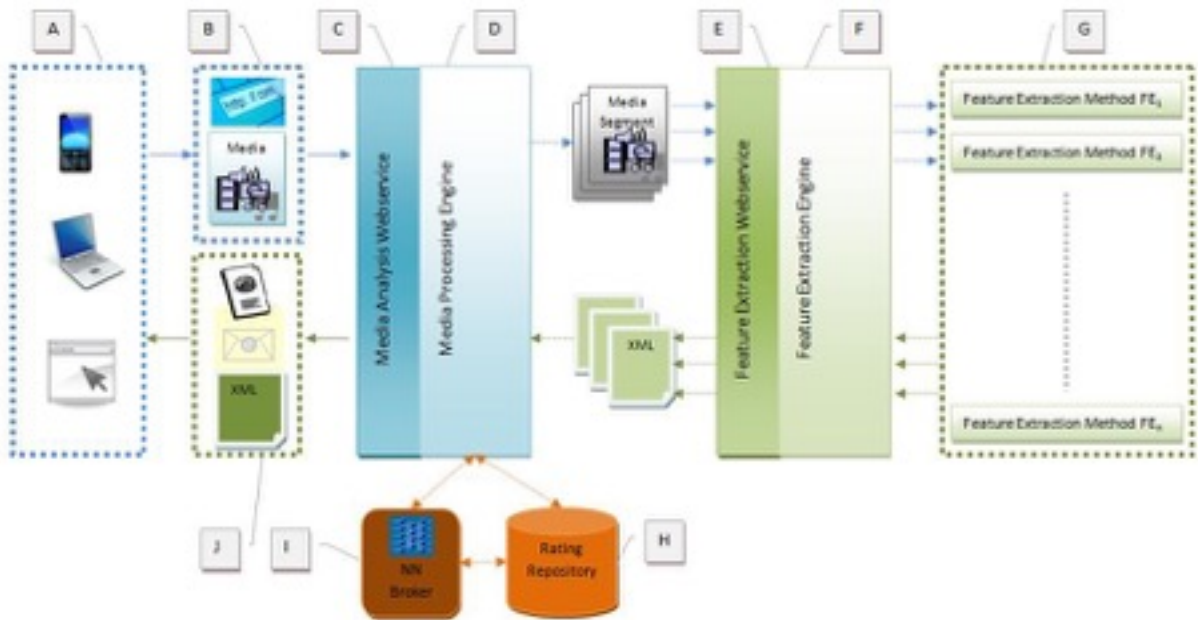


Figure 2: vRate System Architecture

Below are the brief definitions for each system architecture component, flow element, and sub-system.

[A] Possible end-user access points to the vRate service (e.g. Mobile device users through MMS messages, vRate Web application clients by uploading the actual media or providing the URL, External applications by calling methods on vRate's web-service API)

[B] URL media reference or the actual media file that was uploaded to the vRate domain by the end users.

[C] Media Analysis Web Service entry point exposing numerous media query methods for the detection and rating of offensive content.

[D] Media Processing Engine which segments the media in preparation for parallel and distributed feature extraction processes.

[E] Feature Extraction Engine Web Service entry point exposing numerous feature extraction methods for the detection and rating of offensive content.

[F] Feature Extraction Engine which consists of an ever growing collection of feature extraction algorithms which may be:

- a. External hence remotely accessed through Web Service APIs (e.g. TryNT web-service)
- b. Directly built in to the engine Assembly (e.g. vRate nudity detection, vRate profanity detection methods)

[G] Registry of remotely accessed or locally implemented feature extraction methods.

[H] Media rating repository which consists of a media samples, survey results and user feedback captured as ICRA ratings for the sample base, feature extraction method definitions, processed media resource definitions along with the associated:

- a. High-level feature extraction and rating results for the media resource
- b. Frame (sub-segment) level extraction and rating results for the media resource
- c. Rating session history

[I] Neural Network Broker which is a mining model built on top of data aggregated from the Media Rating Repository's Media Sample-Base that incorporates the associated feature extraction results and ICRA ratings.

[J] vRate Web-Service API output brokered by [I] and compiled by [D] in the format requested by the end-user (e.g. XML, report, MMS header)

6. vRateLite[®] Implementation

Scope

The scope of the vRateLite[®] implementation is limited to the following core players in the vRate system architecture [Figure 2]:

- An Analyzer Web Service which takes a URL reference or an uploaded media, and returns a brokered rating assessment indicating degree of offensiveness [C].
- A distributed and parallel Media Processing Engine which prepares the media through segmentation in to video frames and sound clips and engages the feature extraction engine [D].
- A feature extraction engine coupled a web-service interface which exposes an open set of content analysis algorithms [G] for the detection of salient features which include nudity detection, profanity detection, media host risk assessment [E] and [F].
- A neural network mining model trained to consolidate feeds in the form of collected human ratings and feature extraction engine results to confer a degree of offensiveness [I] & [H].

vRateLite[®] at Work

A sample operational scenario is where vRate is queried to detect and rate offensive content in personal media captured via a cell phone camera and uploaded directly to vRate [1] or posted on a media sharing site such as YouTube [2]. Salient features of interest in this scenario would be:

- a) Nudity – likes of skin exposure, body part (or like) exposure
- b) Media file naming indicating potential obscenity
- c) In the case of shared media [2], self reported user ratings, content ratings provided by the hosting website, the hosting site's IP address(es)/website history/profile, hosting site's country of origin
- d) Obscene, explicit language

After the media itself [1] or its URI [2] reaches the Media Analysis Web Service, vRate's internal flow is as the following:

1. Media Processing Engine [D] creates a temporary local copy of the media for analysis purposes with a unique URI.
2. Media Processing Engine [D] pre-processes the media for feature extraction:
 - i. Persists the media related information and meta-data as a resource record in the Media Rating Repository [H].
 - ii. Identifies applicable feature extraction algorithms.
 - iii. Splits video in to image frames for image analysis.
 - iv. Saves the audio track in to one or more track segments for speech recognition.
 - v. Performs additional formatting as needed by the applicable feature extraction methods (e.g. scaling, noise reduction)

3. Media Processing Engine [D] processes the media by launching feature extraction processes in parallel and distributed fashion.
4. Feature Extraction Engine [F] receives processing requests via its Web Service methods [E]
5. Feature Extraction Engine [F] dispatches the media pre-processed by the Media Processing Engine [D] to specific feature extraction algorithm adapters [G].
6. Activated feature extraction methods [G] return extraction results back to the engine [F].
7. Feature Extraction Engine [F] through its Web Service Interface [E], passes the feature extraction process results back to Media Processing Engine [D].
8. Media Processing Engine [D] saves the results to the media rating repository [H]
9. Media Processing Engine [D] sends a mining query encapsulating the feature extraction results and media resource definition to Neural Net Broker [I].
10. Media Processing Engine [D] compiles a XML based query result that incorporates the rating provided by the Neural Net Broker [I]. If detailed rating information is requested in the original query, the compiled result also includes specifics on extracted features**.
11. In this scenario, Media Analysis Web Service [C] returns the compiled media rating as an XML based rating summary to be attached to the original message header [J].

7. Summary

vRate is a media rating service which employs a multi-faceted feature extraction engine, a repository of human rater assessments and a neural network mining model to detect offensive content inherent in a media file.

This core technology can be applied in various business scenarios. vRate's immediate focus areas include tagging ICRA type ratings to MMS, SMS, emails, voice mails, html pages and media scans for patterns including offensive elements or objects of interest. We believe that vRate's innovative approach, which incorporates multiple rating strategies to avoid the shortcomings of reliance on a single methodology, improves the accuracy and effectiveness of offensive content detection.

References

[Calic et.al.] Calic, J., Campbell, N., Dasiopoulou, S., and Kompatsiaris Y. (2005) *An Overview of Multimodal Video Representation for Semantic Analysis*, European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies (EWIMT 2005), IEE

[Informedia] Hauptmann, A. G., and Smith M. A. (1995) *Text, Speech and Vision for Video Segmentation: The Informedia Project*, AAAI Fall Symposium, Computational Models for Integrating Language and Vision

** [Please refer to **Appendix B** for a sample report]

[InternetSafety] Internet Safety Software, URL: <http://internetsafety.com> (8/20/2009)

[Li et.al.] Li, J., Wang, Z., Li, X., Xiao, T., Wang, D., Zheng, W., and Zhang, B. (2007) Video retrieval with multi-modal features, Conference On Image And Video Retrieval archive, Proceedings of the 6th ACM international conference on Image and video retrieval, Amsterdam, The Netherlands, pp: 652 - 652

[Sex and Tech] Results from a Survey of Teens and Young Adults, National Campaign to Support Teen and Unplanned Pregnancy URL: http://www.thenationalcampaign.org/sextech/PDF/SexTech_Summary.pdf (8/20/2009)

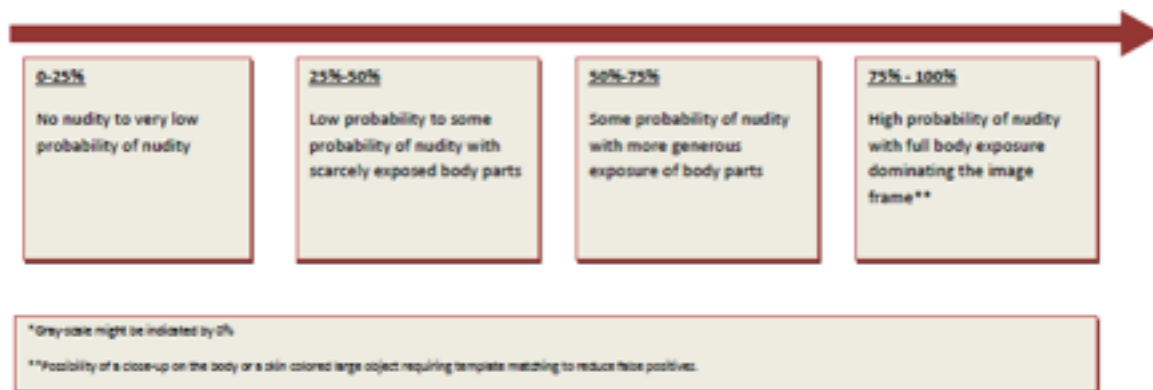
[TRECVID-2003] Amir, A., Berg, M., Chang, S., Iyengar, G., Lin, C., Natsev, A., Neti, C., Nock, H., Hsu, W., Smith, J. R., Tseng, B., Wu, Y., and Zhang, D. (2003), *IBM Research TRECVID-2003 Video Retrieval System*, URL:<http://www.ee.columbia.edu/ln/dvmm/publications/03/ibmcutrec03.pdf> (8/20/2009)

[TRYNT] TRYNT Heavy Technologies, URL: <http://www.trynt.com> (8/20/2009)

[YangSky] Yang's Scientific Research Institute, URL: <http://yangsky.com> (8/20/2009)

Appendix A

vRate_Nudity_Detection_Lite Probability Scale



Appendix B

vRateLite: Sample Report


VRATE ENGINE : REPORT PAGE - Windows Internet Explorer

http://192.168.1.3:8883/vrate/result.aspx

VRATE ENGINE : REPORT PAGE


AVG | Yahoo! Search | Search | Total Protection | AVG Info

REPORT



F:\vrate\upload\image\barbies1.jpg

VRATE ENGINE RESULTS

AnalysisImage	
FG-Skin-Blobs	23
ProcessedBlobs	5
Above-Skin-TH	5
Raw-Skin-Percentage	0.844441566666667
Adjusted-FG-Skin-Percentage	0.844441566666667
Adjusted-FG-Skin-Percentage-Desc	Highly Probable Nudity - Major body exposure in frame or template match

TRYNIT RESULTS

Filename	barbies1.jpg
http://71.191.191.45:8883/vRateOut/image/barbies1.jpg	
86bcae413724916328ee66d60ba896fe	
Score	53.26
Likely-Nude	High

Internet | Protected Mode: On | 100%

vRateLite: Skin Blob Detection Results



Maximal frame for blob with high skin exposure

Blob with high skin exposure already contained in maximal frame

Appendix C

ICRA Rating Scale

	Violence Rating Descriptor	Nudity Rating Descriptor	Sex Rating Descriptor	Language Rating Descriptor
Level 0:	Harmless conflict, some damage to objects	No nudity or revealing attire	Romance, no sex	Inoffensive slang; no profanity
Level 1:	Creatures injured or killed; damage to objects; fighting	Revealing attire	Passionate kissing	Mild expletives
Level 2:	Humans injured or with small amount of blood	Partial nudity	Clothed sexual touching	Expletives; non-sexual anatomical references
Level 3:	Humans injured or killed	Non-sexual frontal nudity	Non-explicit sexual activity	Strong, vulgar language; obscene gestures; Racial Epithets
Level 4:	Wanton and gratuitous violence; torture; rape	Provocative frontal nudity	Explicit sexual activity; sex crimes	Crude or explicit sexual references; Extreme Hate Speech